Lab 5 – Comparing Health Risks

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please submit your lab report as a **pdf** to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.
- Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- We'll be investigating the heart dataset, which collected data on the health factors of 303 patients being screened for heart disease. We'll use this data to address the following three research questions (one on each page):
    - Do people with fasting blood sugar levels above 120 mg/dL have a **higher** risk for heart disease?
    - Do people who have experienced an exercise induced angina have a **higher** risk for heart disease?
    - Do people who experience exercise induced anginas have **different** cholesterol levels on average?

STEP 0

- **Pre-lab work**
    - Complete the pre-lab tutorial (Comparing Groups) for Lab 5 first: https://stat212-learnr.stat.illinois.edu/
- **Download** the heart.csv file to your computer and then **import** into your RStudio session.
- Create a new **R script** (or use the **RMarkdown file** if you are using that option)
- Remember to **library(tidyverse)** so that you can use the ggplot function!

Variables

Each row of this dataset represents one patient being screened, and the following variables were documented for each patient:

- **age:** age in years
- **sex:** biological sex (0 if female, 1 if male)
- **cp**: chest pain type (0 if typical angina, 1 if atypical angina, 2 if non-anginal pain, 3 if asymptomatic)
- **exang:** binary variable documenting whether patient experienced exercise induced angina
- **trestbps:** resting systolic blood pressure (in mm/Hg on admission to hospital)
- **chol:** serum cholesterol (mg/dL)
- **fbs:** binary variable documenting whether fasting blood sugar was high ("yes" if > 120 mg/dL and "no" if <= 120 mg/dL)
- **restecg:** resting electrocardiographic results (0 if normal, 1 if having ST-T wave abnormality, 2 if showing probable or definite left ventricular hypertrophy)
- **thalach:** maximum heart rate achieved
- **oldpeak:** ST depression induced by exercise relative to rest
- **slope:** the slope of the peak exercise ST segment
- **ca:** number of major vessels (0-3) colored by flourosopy
- **target:** Whether patient was found to have angiographic disease status (heart disease) as determined by amount of blood vessel narrowing ("positive" if heart disease diagnosis, "negative" if no heart disease diagnosis)

Research Question 1: *Do people who are diabetic (fasting blood sugar levels above 120 mg/dL) have a* ***higher*** *risk for heart disease?*

Question 1 (5pts): Let's first investigate visually. **Create a 100% stacked barplot** to compare the proportion of patients with heart disease based on whether their fasting blood sugar level was above 120 mg/dL.

Include an image of your barplot in the report *and* **Include your R code**

- One bar should represent those who are diabetic, and the other should represent those who are not. The bar should be shaded to reflect what proportion in each group have heart disease.
- Give the bars a black border, and adjust the width to be between 0.2 and 0.5
- Add an appropriate x axis label, y axis label, and title.

Question 2 (5pts): Now, let's use a test for two proportions to make a statistical inference. Using the dplyr package, create a frequency table to get counts of how many people have or don't have heart disease based on whether they are diabetic or not.

Copy or screenshot the frequency table into your report *and* **Include your R code**

- If done correctly, this table will have 4 rows.
- You can display the table exactly as it appears in R output, or you can re-format it in your document if you wish to.

Run a proportions test to answer research question 1 *and* **Include your R code.**

- **Tip:** Is this a directional or non-directional test? Read the research question again!
- Remember that you need to enter two vectors into your code, the first vector includes the numbers in each group who have heart disease, and the second vector includes the totals for each group.
- Copy+paste or screenshot the summary output from your proportions test.

In your own words, interpret the results and make a conclusion in context. A full response should:

- Identify the proportion with heart disease in each group
- Identify the p-value
- Briefly summarize your answer to our first research question using these results.

Research Question 2: *Do people who have experienced an exercise induced angina have a* ***higher*** *risk for heart disease?*

Question 3 (5pts): Repeat the procedures for Question 1, but with this new predictor variable.

Include an image of your 100% stacked barplot in the report *and* **Include your R code**

Question 4 (5pts): Follow the same procedures in Question 2 to address our second research question statistically.

Copy or screenshot the frequency table into your report *and* **Include your R code**

Run a proportions test designed to answer your second research question *and* **Include your R code.**

In your own words, interpret the results and make a conclusion in context (same as Question 2).

Question 5 (5pts): Let's now report the odds ratio for heart disease for each set of two groups we're comparing. Rather than code this computationally in R, we will use an **online calculator**! *We'll talk more about this in Chapter 9, but and Odds Ratio is similar to Relative Risk, but is more appropriate for this design context.*

Report the odds ratio (and 95% confidence interval) for heart disease when patient is diabetic (fasting blood sugar is above 120 mg/dL) as compared to when they are not diabetic.

Report the odds ratio (and 95% confidence interval) for heart disease when the patient had experienced an exercise induced angina as compared to one who didn't.

"Simpler" calculator suggestion: https://www2.ccrb.cuhk.edu.hk/stat/confidence%20interval/CI%20for%20ratio.htm

- *"Feature Present" row should represent number of patients with causal factor (fbs above 120, or angina reported), and outcome positive column should represent number who experience adverse outcome (heart disease)*
- *Be sure to set the 1 – α = 0.95 for a 95% confidence interval*

Calculator that Kelly used in class: https://istats.shinyapps.io/Association_Categorical/

- *Choose contingency table setup*
- *Have your rows represent your explanatory variables (fbs above/below 120, or angina status), and columns represent number who experience adverse outcome (heart disease)*
- *Be sure to choose odds ratio from the drop-down and choose the "95% confidence interval" option below.*

Let's consider possible risk factors for high levels of cholesterols. Notice that cholesterol will be a *numeric* variable, so our approach to this question will be slightly different.

Research Question 3: *Do people who experience exercise-induced anginas have **different** cholesterol levels on average? Let's say the researchers believe either a drop or an increase in cholesterol is possible and noteworthy to report!*

Question 6 (5pts): **Create a jittered plot** to compare cholesterol levels between the angina and no angina groups.

Include an image of your jittered plot in the report *and* **Include your R code**

- Keep the width of your jitter small (like between 0.02 and 0.10)
- Color each group of points differently (one color for "No" and one color for "Yes")
- Add an appropriate x axis label, y axis label, and title

Question 7 (5pts): **Complete a t-test** to address the research question posed. *Even though we have enough observations such that a z-test would be fine, it's easier in R to just run a t-test, and the results will be approximately the same! We will **not** assume equal variances (software can handle this situation easier, and this is the "safer" testing option).*

Copy or screenshot the summary output from your t-test

In your own words, interpret the results and make a conclusion in context. A full response should:

- Identify the average cholesterol level for each group
- Identify the p-value
- Briefly summarize how this result helps you address research question 3.