Lab 7 – Comparing Health Risks

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

- Be sure that all **group members** are **added** in your submission to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.

READ THIS!

Introduction: We will be using the heart.csv data set to examine health characteristics for 303 patients who were being screened for heart disease. We would like to use this data to explore possible health risks. Each row 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)

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

exang: binary variable documenting whether patient experienced exercise induced angina

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)

STEP 0

- Create an R Script (or RMarkdown) to get started! (Don't try to code in the console. It's harder!)
- *In this lab, we will use the dplyr package for the first time! If you have already installed tidyverse, you are good. You don't need to install this again since it is included inside.*
- You should **library tidyverse** when starting your session. Write it at the top of your script as a reminder for each time you re-enter RStudio and work on this.
- Upload the heart.csv data into your session of RStudio.

Question 1 (4pts): One of the reasons why this data was collected was to identify different possible risk factors for heart disease. One possible factor would be if participants have high fasting blood sugar levels.

Our first question: *Do people with fasting blood sugar levels above 120 mg/dL have a **higher** risk for heart disease?*

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 this plot in your report.

- Fasting blood sugar should be mapped to an axis, and the other axis should be proportion with heart disease
- Give the bars a black border, and adjust the width to be between 0.2 and 0.5
- Scale the numeric axis in increments of 0.1
- Add an appropriate x axis label, y axis label, and title.
- All other formatting (theme styles, color choices, etc.) optional . Keep the legend visible for this one!

Question 2 (4pts): Now, let's use a test for two proportions to make a statistical inference.

Using the dplyr package, create a contingency table to get counts of how many people have or don't have heart disease based on their fasting blood sugar status. If done correctly, this table will have 4 rows *(you do not have to include this in your report).*

Now, use these results to **run a proportions test.**

- Post the summary output from your proportions test.
- In your own words, identify the proportion with heart disease in each group, report, the p-value, and briefly summarize how this result helps you address our first question.

Question 3 (4pts): Our second question: *Do people who have experienced an exercise induced angina have a **higher** risk for heart disease?* **Create a plot** as you did in Question 1, but with this new predictor variable.

Question 4 (4pts): Follow the same procedures in Question 2 to address our second question statistically (**include proportions test summary output**, and **briefly summarize findings**).

Question 5 (6pts): This is an *observational study* (no one was being assigned to interventions—just recording their data as they lived their lives). Additionally, all of our data comes from people who were being screened for heart disease, likely because their doctor noticed risk factors for heart disease. Thus, this study *could* be functionally classified as a *case-control study*—observing certain outcomes first, then checking for heart disease as a possible cause.

Report the odds ratio (and 95% confidence intervals) for heart disease when fasting blood sugar is above 120 mg/dL as compared to when it is equal to or lower than 120. *Tip: Fill in the 4 cells carefully. Feature present would be fbs above 120.*

Report the odds ratio (and 95% confidence intervals) for heart disease when the patient had experienced an exercise induced angina as compared to one who didn't. *Tip: Fill in the 4 cells carefully. Feature present would be angina experienced.*

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

Question 6 (4pts): Let's switch gears. Since these participants also recorded their cholesterol levels, perhaps we can see if there are any risk factors to predict high cholesterol.

Our third question: *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!*

Create a jittered plot to compare cholesterol levels between the angina and no angina groups.

- Keep the width of your jitter small (like between 0.02 and 0.1)
- Scale the y axis in increments of 40 (and be sure your breaks range covers the entire range of data)
- 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
- Remove the legend this time!
- All other formatting (theme styles, color choices, etc.) optional

Question 7 (4pts): **Complete a t-test** to address the research question posed. *Even though we have enough observations to just do a z-test, 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).*

- Post the summary output from your t-test
- In your own words, identify the average cholesterol level for each group, report, the p-value, and briefly summarize how this result helps you address the research question.