Lab 7 – 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.
- 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**.

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:
    - 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

- **Download** the heart.csv file to your computer and then **import** into your RStudio session.
- Create a new **R script** (or **RMarkdown** 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)

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)

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.

- ***Research Question 1:*** *Do people with 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 this plot in your report.

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

- 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 (5pts): 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.

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 determine if there is evidence for a difference in proportions beyond random chance sampling variability *and* **Include your R code.**

- Remember that you need to enter two vectors into your code, the first vector is the number in each group who have heart disease, and the second vector representing the total number of people in each group
- Post the summary output from your proportions test.

In your own words, interpret the results and make a conclusion. 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 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 to determine if there is evidence for a difference in proportions beyond random chance sampling variability *and* **Include your R code.**

In your own words, interpret the results and make a conclusion (same questions from Question 2).

Question 5 (5pts): Let's now report the relative risk for heart disease for each set of two groups we're comparing.

Report the relative risk (and 95% confidence interval) 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. "Exposed" numbers represent patients with an fbs above 120.*

Report the relative risk (and 95% confidence interval) 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. "Exposed" numbers represent patients who experienced an angina.*

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

---

Now, 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 0.1)
- 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 (5pts): **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).*

Copy or screenshot the summary output from your t-test

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

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