Lab 5 – Comparing Health Risks

Assignment Overview

- We'll be investigating a dataset containing health factors for 303 patients being screened for heart disease. We'll use this data to address the following **two** research questions:
  - Among those getting screened for heart disease, do people who have experienced an exercise-induced angina have a higher risk for heart disease?
  - Among those getting screened for heart disease, do people who are diabetic (that is, they have fasting blood sugar levels above 120 mg/dL) have different cholesterol levels on average from those who are not diabetic?

Formatting Instructions

- Please write and save all code in an R script. Also be sure to save your written responses to upload to Gradescope when finished.

Step 0

- **Open RStudio** (or your RStudio workspace in Posit Cloud) to get started
- **Create a new script** to write and save your code and open/upload into your RStudio workspace. Click the small green plus sign in the top left corner of your RStudio window!
- **First steps**
  - Don't forget to run `library(tidyverse)` each time you open RStudio.
  - Upload the **heart_testing** dataset from the Canvas assignment page. Watch the Importing Data video for a quick glimpse of how to do that!
  - Each row represents one patient who was screened for heart disease.

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)
- **exang:** binary variable documenting whether patient experienced exercise induced angina (0 if no, 1 if yes)
- **trestbps:** resting systolic blood pressure (in mm/Hg on admission to hospital)
- **chol:** serum cholesterol (mg/dL)
- **fbs:** binary variable documenting whether patient is diabetic. Determined by checking whether fasting blood sugar was high (1 if > 120 mg/dL and 0 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 **(1** if heart disease diagnosis, **0** if no heart disease diagnosis)

Generative AI policy: You are **encouraged** to use **Github Copilot** to generate **code**. You may also use other generative AI tools for coding as needed. You should not use code directly from other people—you must be involved in the code generation.

Written responses should be in your own words unless otherwise stated—these questions are designed to get you thinking about interpretations and to develop your conceptual understanding to prepare you for the exams. **All questions in red** should ***not*** be fed to a generative AI tool and should be answered in ***your* own words *without* bullet points**.

Research Question 1: *Among those getting screened for heart disease, do people who have experienced an exercise-induced angina (pronounced an-JY-nah) have a higher risk for heart disease?*

Question 1 (6pts)

Part a) Open the data and notice that our binary variables for exercise-induced angina and heart disease status are 0's and 1's. We'd like to change these 0's and 1's to be more descriptive labels using the mutate function.

- Create a new variable with the name of your choice with the category names "Angina reported" and "No angina reported" using the appropriate variable.
- Create a new variable with the name of your choice with the category names "Heart disease" and "No heart disease" using the appropriate variable.
- You may do this in one step, or two separate steps!

Part b) Now let's examine visually. **Create a 100% stacked barplot** to compare the proportion of patients with heart disease based on whether they had experienced an exercise-induced angina. *Use the new variables you created above. If you were unsuccessful, no worries—just use the original ones with 0-1 outcomes!*

- One bar should represent those who experienced an angina, and the other should represent those who did 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, fill legend label, and title.
- You may optionally add other features!
- **Export** and save the image to **upload to Gradescope** with your submission

Question 2 (5pts): Now, let's test for a difference in our two proportions as the basis for the statistical inference we will make in the next question.

Part a) **Create a contingency table** to get counts of how many people have or don't have heart disease based on whether they experienced an angina or not (adding margin totals are optional). If done correctly, this table will have 2 rows and 2 columns of counts.

- You will **copy+paste this table output** into Gradescope.

Part b) **Run a proportions test**

- **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.
- **You will copy+paste the summary output** from your proportions test into Gradescope.

Part c) What proportion of the patients with an exercise-induced angina were diagnosed with heart disease? What proportion of the patients with no exercise-induced angina were diagnosed with heart disease?

Question 3 (6pts): Now, let's interpret your results from the hypothesis test in the previous question and make a conclusion in context.

Part a) In words, state the null and alternative hypotheses for this investigation.

Part b) Identify and interpret the p-value *in context* (We are **not** asking you to make a decision about rejecting/failing to reject the null here. We want you to interpret this as a value like we discuss on **page 28** and **page 88** of the notes. If drawing on these interpretational examples, be sure to incorporate that this situation is a comparison of proportions/risk!)

Part c) Summarize your answer to our first research question using these results (we recommend 1-2 sentences)

Question 4 (3pts): Let's now estimate the relative risk for heart disease when comparing these two groups. You are encouraged to use the **online calculator** we saw in class (https://www.medcalc.org/en/calc/tests.php)

…Or if you feel ambitious, you may use your Generative AI tool to do this!

Report the relative risk (and 95% confidence interval) for heart disease when the patient had experienced an exercise induced angina as compared to if they hadn't.

Research Question 2: Among those getting screened for heart disease, do people who are diabetic (that is, they have fasting blood sugar levels above 120 mg/dL) have different cholesterol levels on average from those who are not diabetic?

Question 5 (6pts)

Part a) Similar to what we did in Question 1, let's use the mutate function to create a new variable that indicates whether the patient meets the criteria for being diabetic or not.

- Create a new variable with the name of your choice with the category names "Diabetic" and "Not diabetic" using the appropriate variable.

Part b) **Create a jitter plot** to compare cholesterol levels between the diabetic and non-diabetic groups. *Use the new variable you just created. If you were unsuccessful, no worries—just use the original one!*

- Keep the width of your jitter small (like between 0.02 and 0.10)
- Color each group of points differently based on diabetic status
- Add an appropriate x axis label, y axis label, and title
- You may optionally add other features!
- **Export** and save the image to **upload to Gradescope** with your submission

Question 6 (3pts): Now, let's test for a difference in our two proportions as the basis for the statistical inference we will make in the next question. *We will **not** assume equal variances (software can handle this situation easier, and this is the "safer" testing option).*

Part a) **Run a t-test test**

- Consider the phrasing of the question and whether we should set this directionally or non-directionally
- We will **not** assume equal variances (software can handle this situation easier, and this is the "safer" testing option).

Part b) What is the mean cholesterol level for patients who are diabetic? What is the mean cholesterol level for the non-diabetic patients?

Question 7 (6pts): Now, let's interpret your results from the hypothesis test in the previous question and make a conclusion in context.

Part a) In words, state the null and alternative hypotheses for this investigation.

Part b) Identify and interpret the p-value *in context* (We are **not** asking you to make a decision about rejecting/failing to reject the null here. We want you to interpret this as a value like we discuss on **page 28** and **page 88** of the notes. If drawing on these interpretational examples, be sure to incorporate that this situation is a comparison of means!)

Part c) Summarize your answer to our first research question using these results (we recommend 1-2 sentences)