Lab - Comparing Health Risks

Name: Yiran Hu, netid: yiranhu3

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

Complete the pre-lab tutorial (Comparing Groups) for Lab 5 first: https://stat212-learnr.stat.illinois.edu/ (https://stat212-learnr.stat.illinois.edu/)

Load `tidyverse` package.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidy
verse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
me errors
```

Load the data

- Download the heart.csv file to your computer. Save it to the same folder as this RMarkdown file.
- Notice that this is a csv file–we will use the `read_csv` function to load it in! This function should be activated with the `readr` package, which should load with `tidyverse`.
- Make sure the name of the file *matches* what you input inside `read_csv`. If it's `heart_1` for example, be sure to adjust that!

```
library(readxl)
#heart = read_excel("heart.xlsx")
heart = read_csv("heart_new.csv")
```

```
## Rows: 303 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (3): fbs, exang, target
## dbl (11): age, sex, cp, trestbps, chol, restecg, thalach, oldpeak, slope, ca...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

View data. Run once below, but delete before knitting your markdown!

```
View(heart)
```

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?

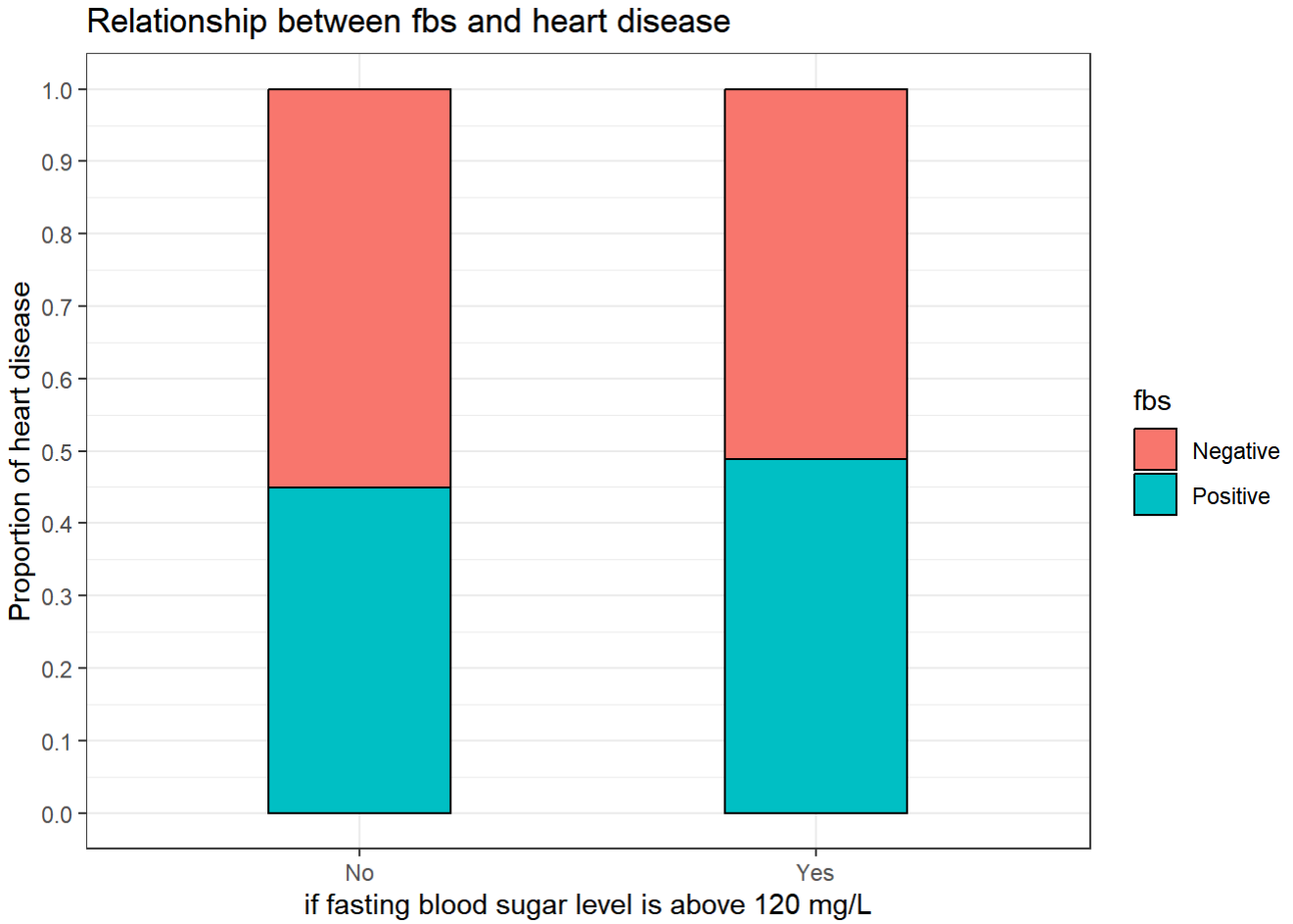
Question 1 (5pts)

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.

```
ggplot(data = heart, aes(x = fbs, fill = target)) +
  geom_bar(color = "black", position = "fill", width = 0.4) +
  labs(x = "if fasting blood sugar level is above 120 mg/L",
       y = "Proportion of heart disease",
       fill = "fbs",
       title = "Relationship between fbs and heart disease")+
       theme_bw() +
       scale_y_continuous(breaks = seq(0,1,0.1))
```



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 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.

```
heart %>%
  group_by(fbs,target) %>%
  summarise(count=n())
```

```
## `summarise()` has grouped output by 'fbs'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 4 × 3
## # Groups:   fbs [2]
##   fbs   target   count
##   <chr> <chr>    <int>
## 1 No    Negative   142
## 2 No    Positive   116
## 3 Yes   Negative    23
## 4 Yes   Positive    22
```

```
prop.test(x=c(116,22),
          n=c(258,45),
          alternative="less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(116, 22) out of c(258, 45)
## X-squared = 0.10627, df = 1, p-value = 0.3722
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.1065069
## sample estimates:
##    prop 1    prop 2
## 0.4496124 0.4888889
```

The probability of not diabetic people having heart disease is about 44.96%, the probability of diabetic people having heart disease is about 48.89%. The p-value is 0.3722. I think people who are diabetic do not have a higher risk for heart disease because the p-value is high and we have strong evidence that the null hypothesis is true.

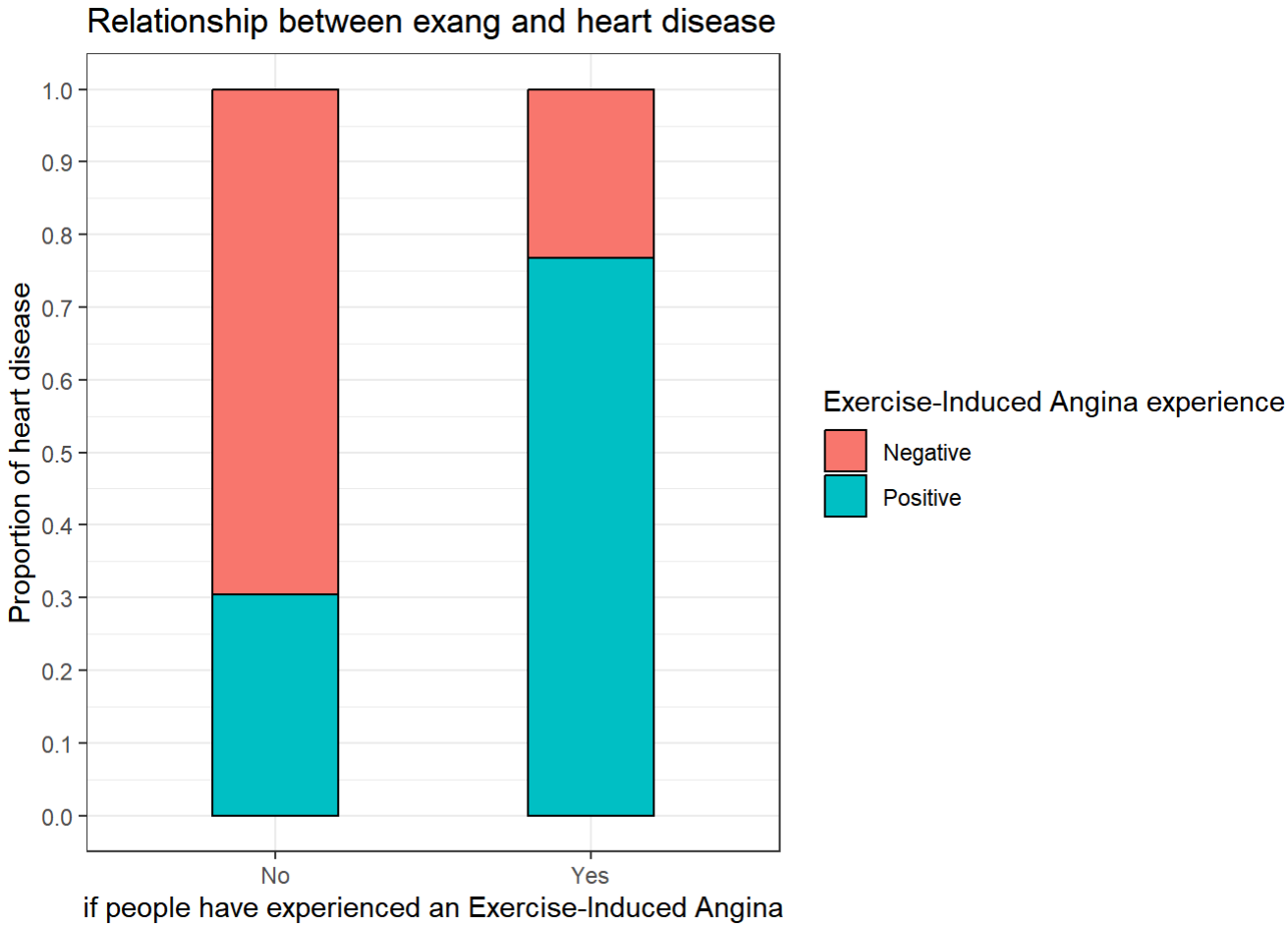
Question 3 (5pts)

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

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

```
ggplot(data = heart, aes(x = exang, fill = target)) +
  geom_bar(color = "black", position = "fill", width = 0.4) +
  labs(x = "if people have experienced an Exercise-Induced Angina",
       y = "Proportion of heart disease",
       fill = "Exercise-Induced Angina experience",
       title = "Relationship between exang and heart disease")+
       theme_bw() +
       scale_y_continuous(breaks = seq(0,1,0.1))
```



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 in context (same as Question 2).

```
heart %>%
  group_by(exang,target) %>%
  summarise(count=n())
```

```
## `summarise()` has grouped output by 'exang'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 × 3
## # Groups:   exang [2]
##   exang target   count
##   <chr> <chr>    <int>
## 1 No    Negative   142
## 2 No    Positive    62
## 3 Yes   Negative    23
## 4 Yes   Positive    76
```

```
prop.test(x=c(62,76),
          n=c(204,99),
          alternative="less")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(62, 76) out of c(204, 99)
## X-squared = 55.945, df = 1, p-value = 3.727e-14
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.3686194
## sample estimates:
##    prop 1    prop 2
## 0.3039216 0.7676768
```

People who have experienced an exercise included anginas have about 76.77% probability to have heart disease. While people who have never experienced exercise included angina have about 30.39% probability to have heart disease. I think people who have experienced exercise included anginas have a higher risk for heart disease because the p-value is very low.

Question 5 (5pts)

Let's now report the odds ratios for heart disease for each set of two groups we're comparing.

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

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. *Tip: Fill in the 4 cells carefully. "Feature Present" numbers should represent patients with an fbs above 120.*

Data Input: (Help) (Formula) (Example)

Test Group	Outcome Positive	Outcome Negative	Totals
Feature Present	22	23	45
Feature Absent	116	142	258
Totals	138	165	303
1-α		重置	Calculate

Result:

Group	Proportion of Outcome Positive	Proportion of Outcome Negative
Feature Present	0.48889	0.51111
Feature Absent	0.44961	0.55039

Result	Point Estimate	Lower C.I.	Upper C.I.
Odds Ratio	1.17091	1.17091	1.17092
Absolute Risk Reduction*	-0.03928	-0.03928	-0.03928
Relative Risk Reduction**	-0.08736	-0.08736	-0.08736
Number Needed to Treat***	-25.46053	-25.46033	-25.46073
Patient Expected Event Rate	0.44961		

The odds ratio: 1.17091

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. *Tip: Fill in the 4 cells carefully. "Feature Present" numbers should represent patients who experienced an angina.*

Data Input: (Help) (Formula) (Example)

Test Group	Outcome Positive	Outcome Negative	Totals
Feature Present	76	23	99
Feature Absent	62	142	204
Totals	138	165	303
1-α		重置	Calculate

Result:

Group	Proportion of Outcome Positive	Proportion of Outcome Negative
Feature Present	0.76768	0.23232
Feature Absent	0.30392	0.69608

Result	Point Estimate	Lower C.I.	Upper C.I.
Odds Ratio	7.56802	7.56801	7.56803
Absolute Risk Reduction*	-0.46376	-0.46376	-0.46375
Relative Risk Reduction**	-1.5259	-1.52591	-1.5259
Number Needed to Treat***	-2.15631	-2.15631	-2.15631
Patient Expected Event Rate	0.30392		

The odds ratio: 7.56802

Question 6 (5pts)

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!*

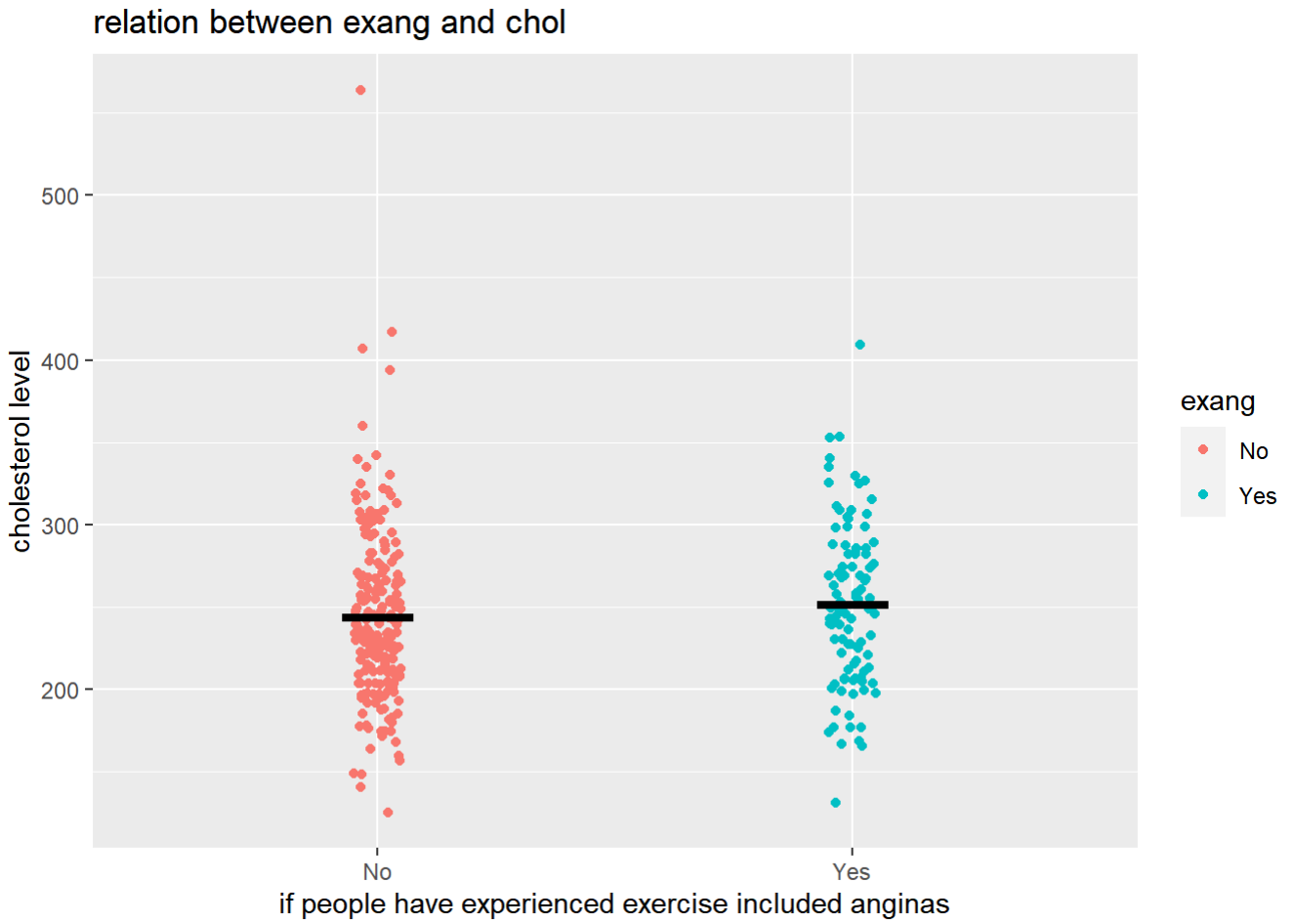
Create a **jittered** plot to compare cholesterol levels between the angina and no angina groups.

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
- Add an appropriate x axis label, y axis label, and title

```
ggplot(data = heart, aes(x = exang, y = chol, color = exang)) +
  geom_jitter(width = 0.05) +
  stat_summary(fun = mean, fun.min = mean, fun.max = mean, geom = "errorbar", color = "black",
width = 0.15, size = 1.5) +
  labs(x = "if people have experienced exercise included anginas", y = "cholesterol level", tit
le = "relation between exang and chol")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



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 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 the research question.

```
t.test(data = heart,
       chol ~ exang,
       var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  chol by exang
## t = -1.1929, df = 206.28, p-value = 0.2343
## alternative hypothesis: true difference in means between group No and group Yes is not equal
to 0
## 95 percent confidence interval:
##  -19.615616   4.826846
## sample estimates:
##  mean in group No mean in group Yes
##          243.8480          251.2424
```

The average cholesterol level for people who have experienced exercise-included anginas is 251.2424, and for people who have never experienced is 243.8480. The p-value is 0.2343. We don't have strong evidence that people who experience exercise-induced anginas have different cholesterol levels on average because p-value is high.