Lab - Comparing Health Risks

Varenya Jain

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 7 first: https://stat212-learnr.stat.illinois.edu/

Load `tidyverse` package.

```
cat("\014")
```

```
shell("cls")

library(rmarkdown)
#remove.packages("rlang")
#install.packages("devtools")
#install.packages("rlang")
#install.packages("tidyverse", dependencies = TRUE)
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.1.8
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

library(dplyr)
library(ggplot2)
tinytex::install_tinytex(force = TRUE)
```

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!

```
#heart = read_csv("heart.csv")
Heart <- read_csv("C:/Users/varen/Desktop/Illini+Uni/SP23/STAT_212/STAT_212-
Lab_07/heart-1.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)
#?(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.

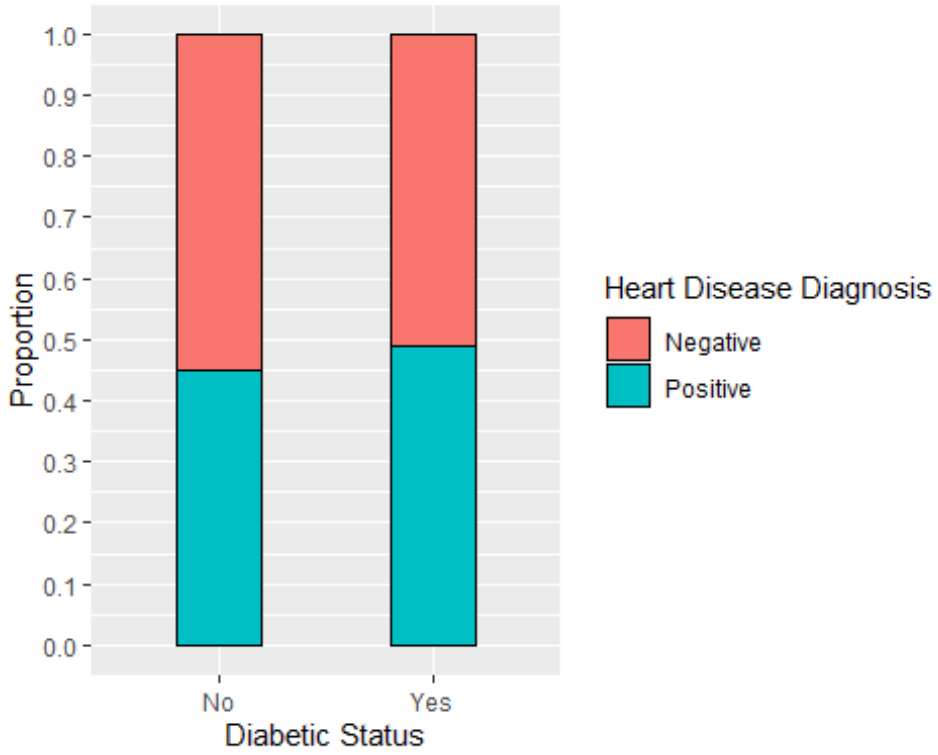
Include an image of your barplot in the report and Include your R code

- 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
- 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!

```
cat("\014")
```

```
shell("cls")
ggplot(data = Heart, aes(x = fbs, fill = target)) +
  geom_bar(position = "fill", color = "black", width = 0.4) +
  labs(x = "Diabetic Status", y = "Proportion", fill = "Heart Disease
Diagnosis") +
#  theme_hc() +
  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.

```
cat("\014")
```

```r
shell("cls")
#how many people do have/don't have HD based on whether they are diabetic or
not
#help(prop.test)
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
```

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

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

```r
Rr_2 = (22/45)/(116/258)

print("X-squared = 0.10627, df = 1, p-value = 0.3722")
```

```
## [1] "X-squared = 0.10627, df = 1, p-value = 0.3722"
```

```r
print("22 diabetics out of 45 have angiographic disease status, and 116 out
of 258 non-diabetics have angiographic disease status.")
```

```
## [1] "22 diabetics out of 45 have angiographic disease status, and 116 out
of 258 non-diabetics have angiographic disease status."
```

```r
#sprintf("The relative risk of diabetic's with heart disease versus non-
diabetics with heart disease is %f", Rr_2)
print("Based on a p-value of 0.3722, we have little to no evidence of a
```

```
difference in diabetic rates between these two groups based on the type of
screening condition they have. This difference could be explained as random
chance here.")
```

```
## [1] "Based on a p-value of 0.3722, we have little to no evidence of a
difference in diabetic rates between these two groups based on the type of
screening condition they have. This difference could be explained as random
chance here."
```

Question 3 (5pts)

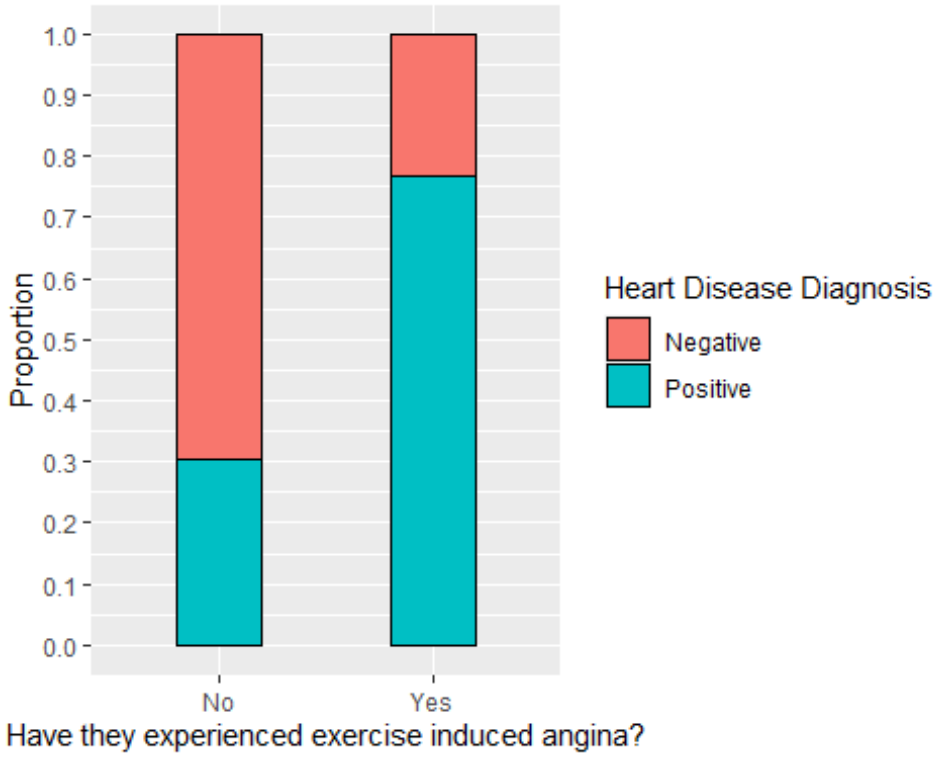
Research Question 2: *Do people who have experienced an exercise induced angina have a* ***higher*** *risk for heart disease?*

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

```
cat("\014")
```

```
shell("cls")
ggplot(data = Heart, aes(x = exang, fill = target)) +
  geom_bar(position = "fill", color = "black", width = 0.4) +
  labs(x = "Have they experienced exercise induced angina?", y =
"Proportion", fill = "Heart Disease Diagnosis") +
#  theme_hc() +
  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).

```
cat("\014")
```

```
shell("cls")
# do those with exang = Yes have a higher risk for target = Negative
#help(prop.test)
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(76, 62),
          n = c(99, 204),
          alternative = "greater")
```

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

```
Rr_4 = (76/99)/(62/204)
```

```
print("X-squared = 55.945, df = 1, p-value = 3.727e-14")
```

```
## [1] "X-squared = 55.945, df = 1, p-value = 3.727e-14"
```

```
print("76 out of 99 people with chest pain have angiographic disease status,
and 62 out of 204 people without chest pain have angiographic disease
status.")
```

```
## [1] "76 out of 99 people with chest pain have angiographic disease status,
and 62 out of 204 people without chest pain have angiographic disease
status."
```

```
#sprintf("The relative risk of people with chest pain with heart disease
versus people without chest pain with heart disease is %f.", Rr_4)
print("Based on a p-value of 3.727e-14, we have sufficient evidence of a
```

```
difference in angiographic disease rates between these two groups based on
the type of screening condition they have. This difference is likely due to
more than random chance here.")
```

```
## [1] "Based on a p-value of 3.727e-14, we have sufficient evidence of a
difference in angiographic disease rates between these two groups based on
the type of screening condition they have. This difference is likely due to
more than random chance here."
```

Question 5 (5pts)

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

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

Report the relative risk (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. "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.*

```
cat("\014")
```

```
shell("cls")

#Heart %>%
#  group_by(fbs, target) %>%
#  summarise(count = n())
#prop.test(x = c(22, 116),
#          n = c(45, 258),
#          alternative = "greater")
#Heart %>%
#  group_by(exang, target) %>%
#  summarise(count = n())
#prop.test(x = c(76, 62),
#          n = c(99, 204),
#          alternative = "greater")

RR1=1.08736
CI_L_1RR=1.08736
CI_U_1RR=1.08736
AR1=0.03928
CI_L_1AR=0.03928
CI_U_1AR=0.03928
ARP1=8.03383
CI_L_1ARP=8.03376
CI_U_1ARP=8.03389

PAR_1=0.00583
PE_1=14.85149
PARP_1=1.28075
sprintf("The Point Estimate of Relative Risk for heart disease when patient has had an angina as compared to those who have not is %f with Confidence Internal [Lower C.I., Upper C.I.] between [%f, %f].", RR1, CI_L_1RR, CI_U_1RR)

## [1] "The Point Estimate of Relative Risk for heart disease when patient has had an angina as compared to those who have not is 1.087360 with Confidence Internal [Lower C.I., Upper C.I.] between [1.087360, 1.087360]."

sprintf("The Point Estimate of Attributable Risk is %f with Confidence Internal [Lower C.I., Upper C.I.] between [%f, %f].", AR1, CI_L_1AR, CI_U_1AR)

## [1] "The Point Estimate of Attributable Risk is 0.039280 with Confidence Internal [Lower C.I., Upper C.I.] between [0.039280, 0.039280]."

sprintf("The Point Estimate of Attributable Risk Percent is %f Percent with Confidence Internal [Lower C.I., Upper C.I.] between [%f, %f].", ARP1, CI_L_1ARP, CI_U_1ARP)

## [1] "The Point Estimate of Attributable Risk Percent is 8.033830 Percent with Confidence Internal [Lower C.I., Upper C.I.] between [8.033760, 8.033890]."
```

```
sprintf("The resulting Population Attributable Risk is: %f", PAR_1)
```

```
## [1] "The resulting Population Attributable Risk is: 0.005830"
```

```
sprintf("The resulting Population Exposure is: %f Percent", PE_1)
```

```
## [1] "The resulting Population Exposure is: 14.851490 Percent"
```

```
sprintf("The resulting Population Attributable Risk Percent is: %f Percent",
PARP_1)
```

```
## [1] "The resulting Population Attributable Risk Percent is: 1.280750
Percent"
```

```
RR2=1.08736
CI_L_2RR=1.08736
CI_U_2RR=1.08736
AR2=0.03928
CI_L_2AR=0.03928
CI_U_2AR=0.03928
ARP2=8.03383
CI_L_2ARP=8.03376
CI_U_2ARP=8.03389

PAR_2=0.00583
PE_2=14.85149
PARP_2=1.28075
print("If “Exposed” numbers represent patients who experienced an angina with
heart disease, then non-exposed are the non-anginas with heart disease.")
```

```
## [1] "If “Exposed” numbers represent patients who experienced an angina
with heart disease, then non-exposed are the non-anginas with heart disease."
```

```
sprintf("The Point Estimate of Relative Risk for heart disease when patient
has had an angina as compared to those who have not is %f with Confidence
Internal [Lower C.I., Upper C.I.] between [%f, %f].", RR2, CI_L_2RR,
CI_U_2RR)
```

```
## [1] "The Point Estimate of Relative Risk for heart disease when patient
has had an angina as compared to those who have not is 1.087360 with
Confidence Internal [Lower C.I., Upper C.I.] between [1.087360, 1.087360]."
```

```
sprintf("The Point Estimate of Attributable Risk is %f with Confidence
Internal [Lower C.I., Upper C.I.] between [%f, %f].", AR2, CI_L_2AR,
CI_U_2AR)
```

```
## [1] "The Point Estimate of Attributable Risk is 0.039280 with Confidence
Internal [Lower C.I., Upper C.I.] between [0.039280, 0.039280]."
```

```
sprintf("The Point Estimate of Attributable Risk Percent is %f Percent with
Confidence Internal [Lower C.I., Upper C.I.] between [%f, %f].", ARP2,
CI_L_2ARP, CI_U_2ARP)
```

```
## [1] "The Point Estimate of Attributable Risk Percent is 8.033830 Percent
with Confidence Internal [Lower C.I., Upper C.I.] between [8.033760,
8.033890]."
```

```
sprintf("The resulting Population Attributable Risk is: %f", PAR_2)
```

```
## [1] "The resulting Population Attributable Risk is: 0.005830"
```

```
sprintf("The resulting Population Exposure is: %f Percent", PE_2)
```

```
## [1] "The resulting Population Exposure is: 14.851490 Percent"
```

```
sprintf("The resulting Population Attributable Risk Percent is: %f Percent",
PARP_2)
```

```
## [1] "The resulting Population Attributable Risk Percent is: 1.280750
Percent"
```

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)
- Scale the numeric axis in increments of 40
- Color each group of points differently
- 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

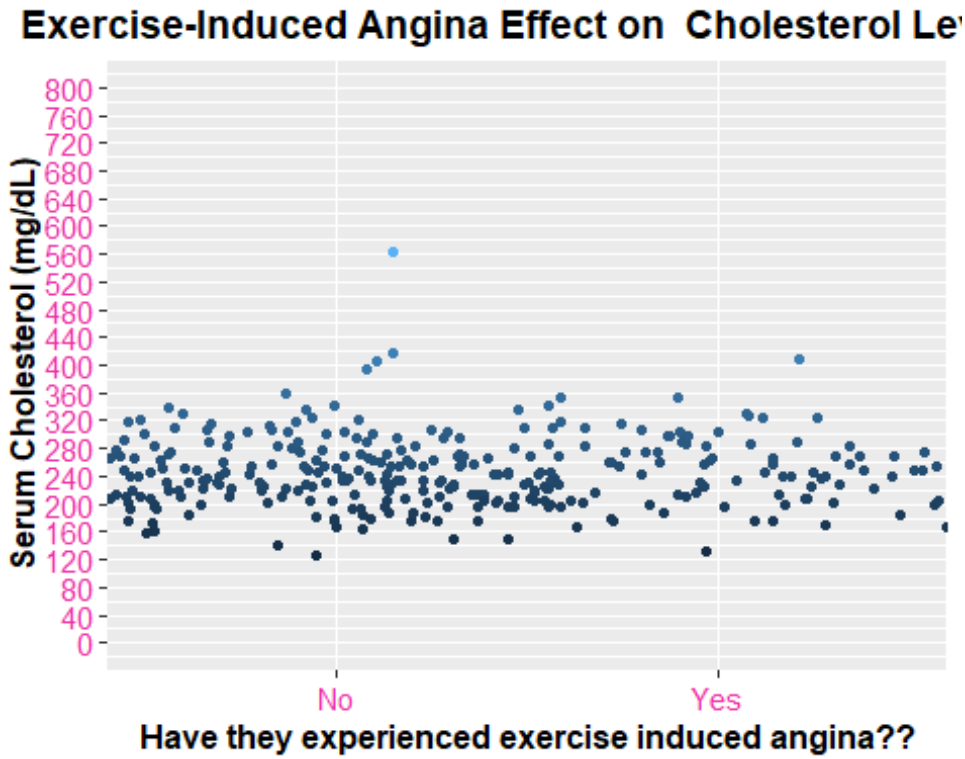
```
cat("\014")
```

```
shell("cls")

#library(scales)

ggplot(data = Heart, aes(x = exang, y = chol), show.legend = NA) +
geom_jitter(aes(colour = chol), width = 0.6, show.legend = NA, show_guide =
FALSE) + labs(title = "Exercise-Induced Angina Effect on  Cholesterol
Levels", x = "Have they experienced exercise induced angina??", y = "Serum
Cholesterol (mg/dL)") + scale_y_continuous(breaks = seq(0, 800, by=40),
limits=c(0,800)) + theme(plot.title = element_text(size = 14, hjust = 0.5,
face = "bold"), axis.title = element_text(size = 12, face = "bold"),
axis.text = element_text(size = 11, color = "maroon2"))

## Warning: The `show_guide` argument of `layer()` is deprecated as of
ggplot2 2.0.0.
## ℹ Please use the `show.legend` argument instead.
```



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.

```
cat("\014")
```

```r
shell("cls")
heart_copy <- Heart
#View(heart_copy)
#select(heart_copy, chol:target) %>%
#  arrange(chol) %>%
#  head()

#t.test(data = heart_copy,
#       chol ~ exang,
#       var.equal = FALSE,
#       alternative = "less")

#cat("\014")
#shell("cls")

t.test(data = heart_copy,
       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

print("A non-directional output yields t = -1.1929, df = 206.28, p-value =
0.2343")

## [1] "A non-directional output yields t = -1.1929, df = 206.28, p-value =
0.2343"

print("The p-value is still greater than the alpha significance level of
0.05, so our results may be different due to random error. There is little to
no evidence that the 'exang' screening condition affected cholesterol
level.")

## [1] "The p-value is still greater than the alpha significance level of
0.05, so our results may be different due to random error. There is little to
no evidence that the 'exang' screening condition affected cholesterol level."

print("Non-angia people have an average serum cholesterol level of 243.8480
mg/dL, while people who have experienced angias have a higher average serum
cholesterol level of 251.2424 mg/dL")
```

```
## [1] "Non-angia people have an average serum cholesterol level of 243.8480
mg/dL, while people who have experienced angias have a higher average serum
cholesterol level of 251.2424 mg/dL"
```

```
print("We are 95% confident that the true difference in mean serum
cholesterol between people who have experienced an Exercise-Induced Angina
and those who have not to be between -19.615616 and 4.826846 mg/dL")
```

```
## [1] "We are 95% confident that the true difference in mean serum
cholesterol between people who have experienced an Exercise-Induced Angina
and those who have not to be between -19.615616 and 4.826846 mg/dL"
```

```
print("A non-directional t-test is used since we are looking for any change
in general, not explicitly expecting a certain directional change in serum
cholesterol in response to the binary screening of experience with an
Exercise-Induced Angina.")
```

```
## [1] "A non-directional t-test is used since we are looking for any change
in general, not explicitly expecting a certain directional change in serum
cholesterol in response to the binary screening of experience with an
Exercise-Induced Angina."
```