STA130 final revision activity: Sample solutions

Correlation between academic self-efficacy and burnout originating from distance learning among nursing students in Indonesia during the coronavirus disease 2019 pandemic

Note: You don't need to open or read any of the following links to do these practice questions

Data source: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JHZDLR

Article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5847840/

Description from Harvard Dataverse (see data source link above):

> This study was approved by the University of Alberta Research Ethics Board (Pro00066510). Participation in the study was voluntary. Informed consent was implied by the overt action of completing the survey after reading the information letter. Students could choose not to respond to a question with no negative consequences to them.

You can find a table with all the survey items here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5847840/table/t2-jeehp-15-02/?report=objectonly#tfn1-jeehp-15-02

The original dataset contains 200 observations of 65 variables. Answers to some of the questions above were already "reverse coded" by the investigators.

For example, consider the following two questions under **engagement**:

4. I feel more and more engaged in my studies.
5. It happens more and more often that I think about my studies in a negative way.

If you were to choose "strongly agree" for Q4, that is a good thing, but to strongly agree with Q5 would generally be considered not a good thing. Q5 has been reverse coded, which basically flips the numbers. Now, in this example, all '4s' indicate positive responses to engagement questions. It isn't the same for each aspect though. All higher values under exhaustion, for example, indicate agreement or strong agreement with being exhausted.

Some additional data wrangling was done for the purposes of this STA130 investigation. Numerical scores for each aspect were calculated by summing scores of the questions within them. E.g., autonomy questions were summed within each student to create the variable `auto`.

Question 1

Consider how this data was collected. Which study design correctly describes this?

A. Prospective cohort.

B. Experiment.

C. Cross-sectional study. [CORRECT]

D. Case-control study.

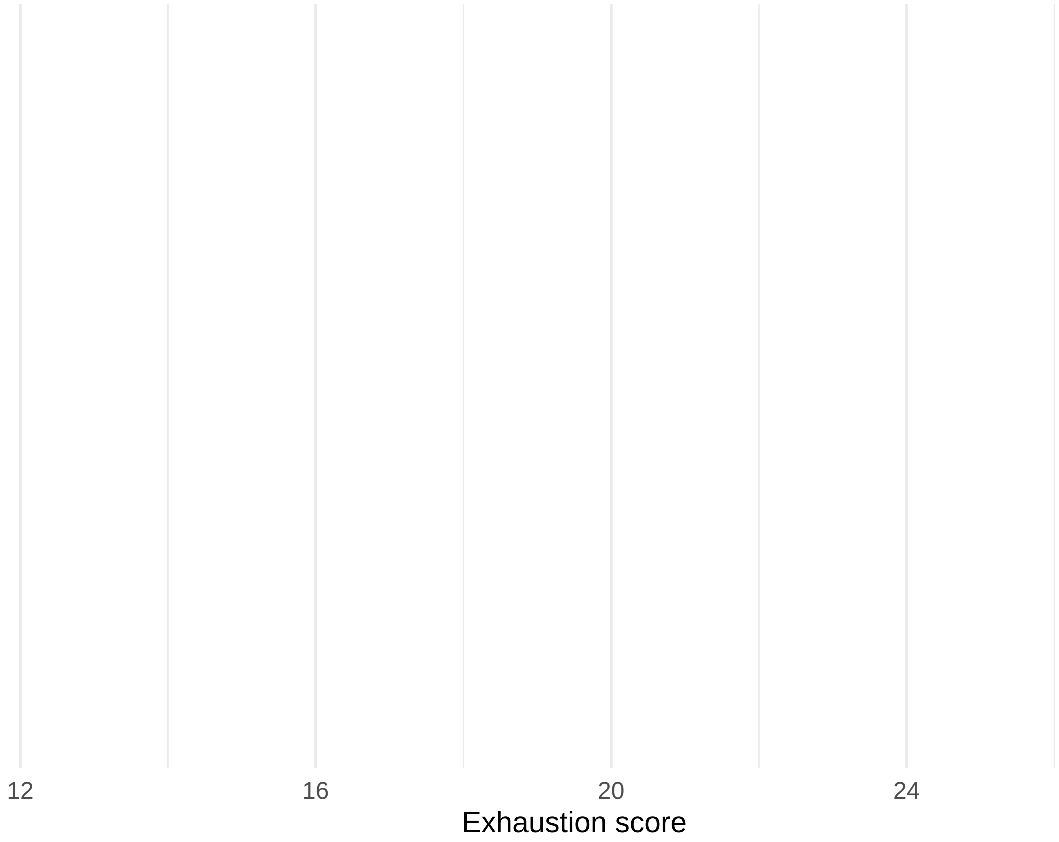
Extra info

Suppose the researchers wanted to understand if there was a difference in average exhaustion scores (`exh`) between 1st and 4th-year medical students.

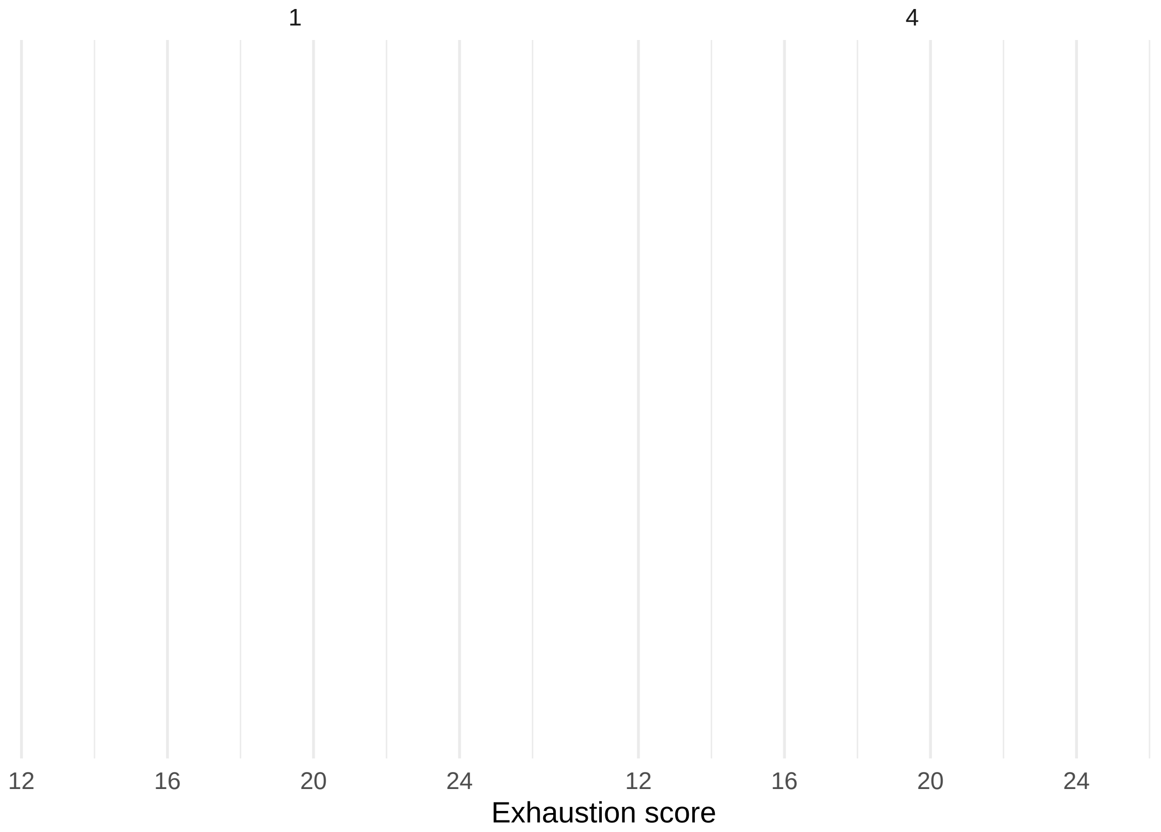
Question 2

Which pair of geometries would be the most sensible to add to the code below to create plots to compare the distributions of exhaustion scores for 1st and 4th-year medical students?

```r
### Plot A
burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  ggplot(aes(x = exh, group = `year in program`)) +
  # geometry here
  theme_minimal() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) +
  xlab("Exhaustion score")
```



```r
### Plot B
burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  ggplot(aes(x = exh)) +
  # geometry here
  theme_minimal() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) +
  xlab("Exhaustion score") +
  facet_wrap(~`year in program`)
```



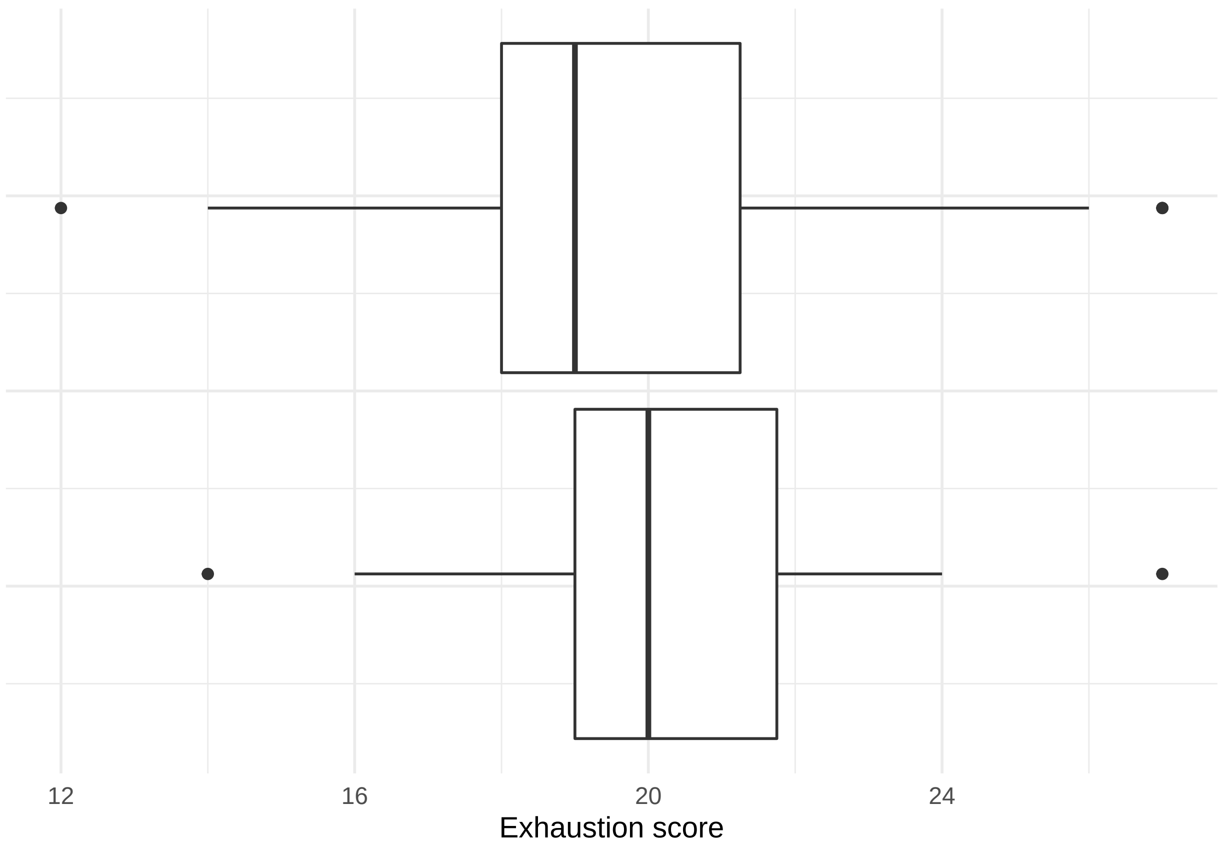
A. Plot A: `geom_bar()+` , Plot B: `geom_histogram(bins = 200, fill = "grey", colour = "black") +`

B. Plot A: `geom_histogram(bins = 200, fill = "grey", colour = "black") +` , Plot B: `geom_point() +`

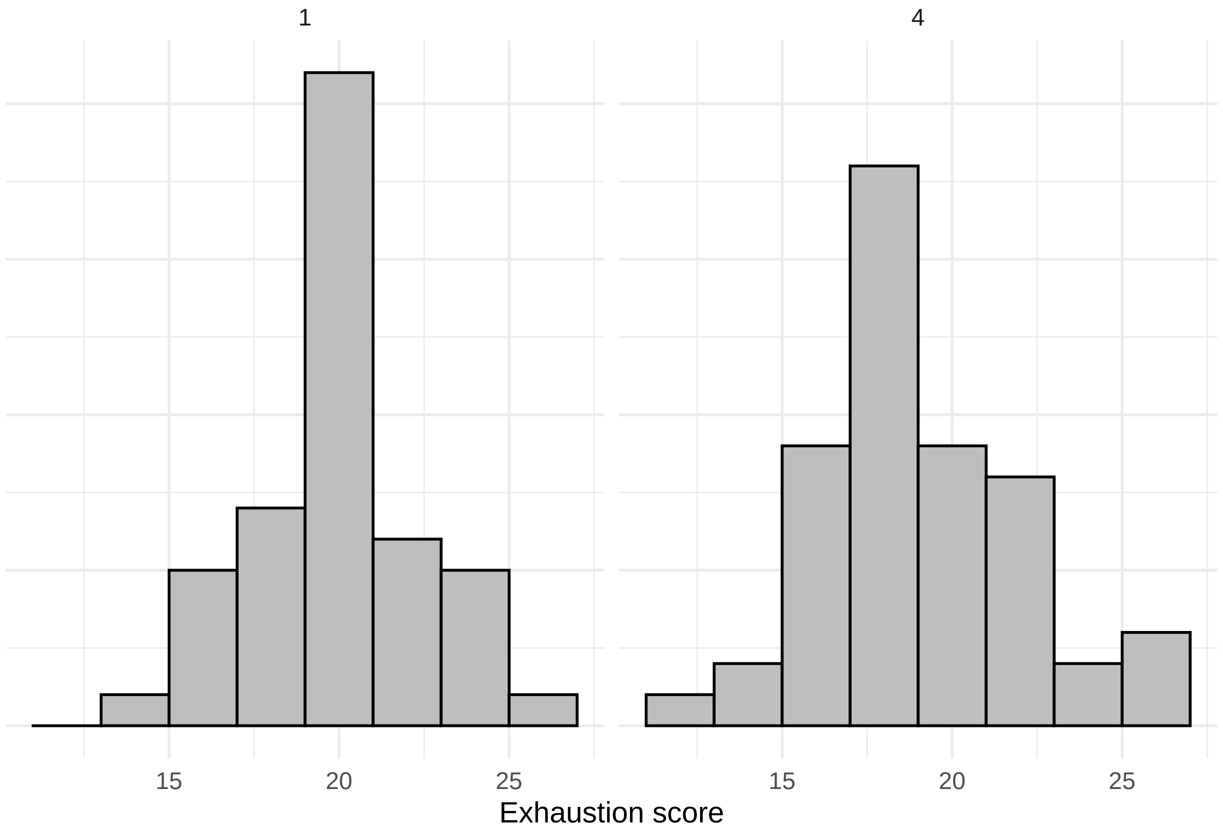
C. Plot A: `geom_point()`, Plot B: `geom_distribution() +`

D. Plot A: `geom_boxplot()`, Plot B: `geom_histogram(binwidth = 2, fill = "grey", colour = "black") +` [CORRECT]

```
### ANSWERS ###

### Plot A
burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  ggplot(aes(x = exh, group = `year in program`)) +
  geom_boxplot()+
  theme_minimal() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) +
  xlab("Exhaustion score")
```



```
### Plot B
burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  ggplot(aes(x = exh)) +
  geom_histogram(binwidth = 2, fill = "grey", colour = "black") +
  theme_minimal() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) +
  xlab("Exhaustion score") +
  facet_wrap(~`year in program`)
```



Question 3

What is the goal in this situation?

A. Estimation.

B. Testing. [CORRECT]

C. Prediction

Question 4

What parameter(s) or variable(s) is/are of primary interest in this situation?

A. The average exhaustion score for medical students.

B. The median exhaustion score for 1st year medical students.

C. The difference in average exhaustion scores between 1st and 4th year medical students. [CORRECT]

D. The median difference in exhaustion scores between each possible pair of 1st and 4th year medical students.

Question 5

We have learned about a number of statistical procedures for testing, estimation and prediction to address questions with data.

Which statistical procedure (among those covered in STA130) is most appropriate in this situation? (Bonus practice: Briefly justify your answer.)

A. Hypothesis test for one proportion.

B. Randomization test on difference in parameters. [CORRECT]

C. The bootstrap confidence interval for a parameter.

D. Classification trees.

E. Linear regression.

Question 6

Below is a hypothesis test and two linear regressions.

```
burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  group_by(`year in program`) %>%
  count() # count is the same as summarize(n = n())
```

```
## # A tibble: 2 x 2
## # Groups:   year in program [2]
##   `year in program`     n
##               <dbl> <int>
## 1                 1    46
## 2                 4    52
```

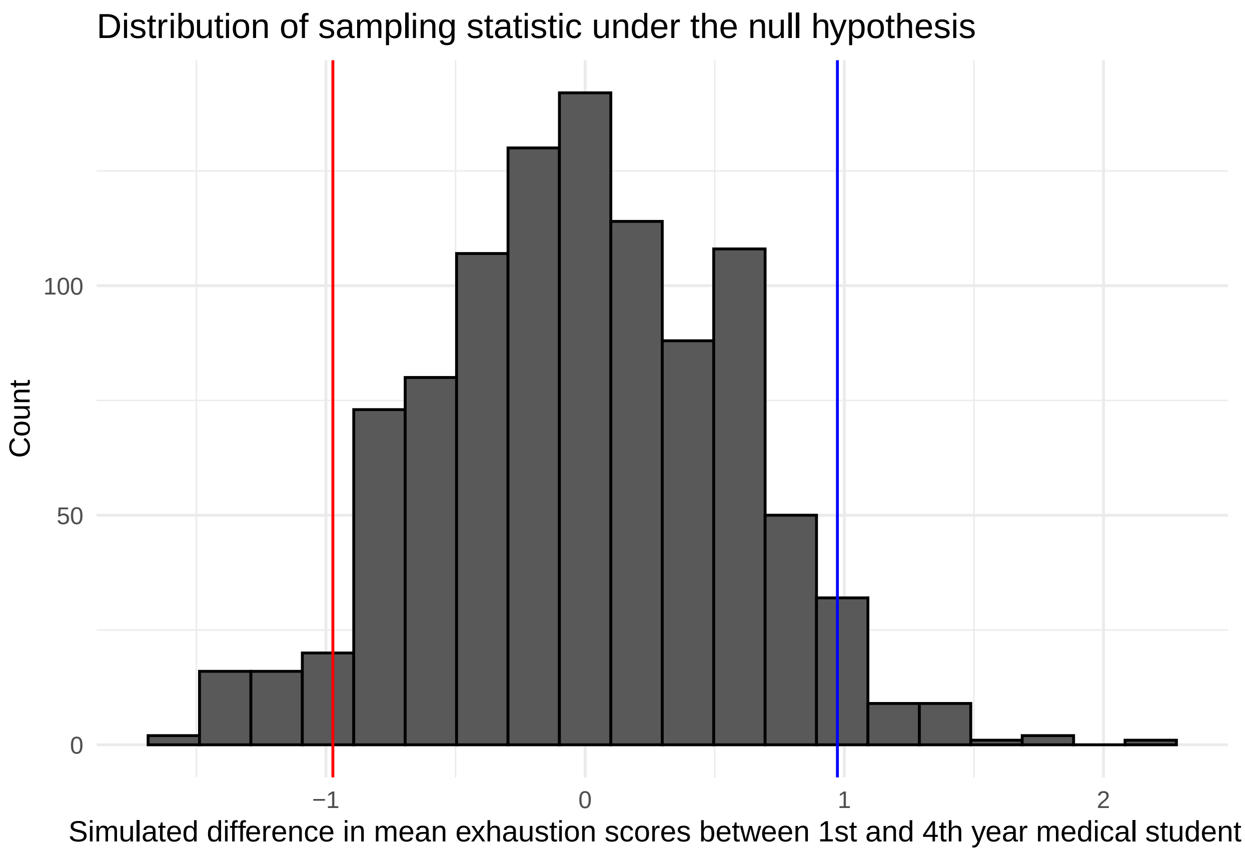
```
test_stat_setup <- burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  group_by(`year in program`) %>%
  summarise(mean = mean(exh))

test_stat_setup
```

```
## # A tibble: 2 x 2
##   `year in program`  mean
##               <dbl> <dbl>
## 1                 1  20.4
## 2                 4  19.5
```

```
test_stat <- test_stat_setup %>%
  ungroup() %>%
  summarise(test_stat = diff(mean)) %>%
  as.numeric()

test_stat
```

```
## [1] -0.9732441
```

```
set.seed(88)

repetitions <- 1000

simulated_stats <- rep(NA, repetitions)

for (i in 1:repetitions){
  # this is were you create one simulation
  new_sim <- burnout %>%
  filter(`year in program` %in% c(1, 4)) %>%
  mutate(`year in program` = sample(`year in program`))

  # you'll basically use whatever code you used for test_stat above,
  # but on the new_sim data/vector
  sim_val <- new_sim %>%
  group_by(`year in program`) %>%
  summarise(mean = mean(exh)) %>%
  ungroup() %>%
  summarise(test_stat = diff(mean)) %>%
  as.numeric()

  simulated_stats[i] <- sim_val
}

sim_tibble <- tibble(simulated_statistics = simulated_stats)

hypothesized_value <- 0

ggplot(sim_tibble, aes(x = simulated_statistics)) +
  geom_histogram(bins = 20, color = "black") +
  labs(x="Simulated difference in mean exhaustion scores between 1st and 4th year medical students",
       y="Count",
       title = "Distribution of sampling statistic under the null hypothesis") +
  geom_vline(xintercept = hypothesized_value - abs(test_stat-hypothesized_value),
             colour = "red") +
  geom_vline(xintercept = hypothesized_value + abs(test_stat-hypothesized_value),
             colour = "blue") +
  theme_minimal()
```



```
p_value <- sim_tibble %>%
  filter(simulated_statistics <= hypothesized_value - abs(test_stat-hypothesized_value) |
         simulated_statistics >= hypothesized_value + abs(test_stat-hypothesized_value)) %>%
  summarise(p_value = n() / repetitions) %>%
  as.numeric()
p_value
```

```
## [1] 0.084
```

```
burnout_filtered <- burnout %>%
   filter(`year in program` %in% c(1, 4)) %>%
  mutate(`year in program char` = as.character(`year in program`))

# Linear regression 1
summary(lm(exh ~ `year in program`, data = burnout_filtered))
```

```
##
## Call:
## lm(formula = exh ~ `year in program`, data = burnout_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4615 -1.4615 -0.4348  1.5585  7.5385
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          20.7592     0.5743  36.150   <2e-16 ***
## `year in program`    -0.3244     0.1919  -1.691   0.0941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.844 on 96 degrees of freedom
## Multiple R-squared:  0.02892,    Adjusted R-squared:  0.01881
## F-statistic: 2.859 on 1 and 96 DF,  p-value: 0.09409
```

```r
# Linear regression 2
summary(lm(exh ~ `year in program char`, data = burnout_filtered))
```

```
##
## Call:
## lm(formula = exh ~ `year in program char`, data = burnout_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4615 -1.4615 -0.4348  1.5585  7.5385
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               20.4348     0.4193  48.741   <2e-16 ***
## `year in program char`4  -0.9732     0.5756  -1.691   0.0941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.844 on 96 degrees of freedom
## Multiple R-squared:  0.02892,    Adjusted R-squared:  0.01881
## F-statistic: 2.859 on 1 and 96 DF,  p-value: 0.09409
```

Which ONE of the following statements is TRUE?

A. The hypothesis test and linear regression 2 can answer the same question. [CORRECT]

B. The hypothesis test and linear regression 1 can answer the same question.

C. The the two linear regressions can answer the same question.

D. None of these methods answer the same question.

Question 7

Which, out of models A and B, would fit a linear model to predict exhaustion score from year in program (restricted to 1st and 4th years) and competence score with the MOST coefficients.

```r
burnout_filtered <- burnout %>%
   filter(`year in program` %in% c(1, 4)) %>%
  mutate(`year in program char` = as.character(`year in program`))

## A
summary(lm(exh ~ `year in program char` * compet, data = burnout_filtered))
```

```
## B
summary(lm(exh ~ `year in program char` + compet, data = burnout_filtered))
```

A. A [CORRECT]

B. B

C. Both would have 3 coefficients.

D. Both would have 4 coefficients.

Question 8 (three parts)

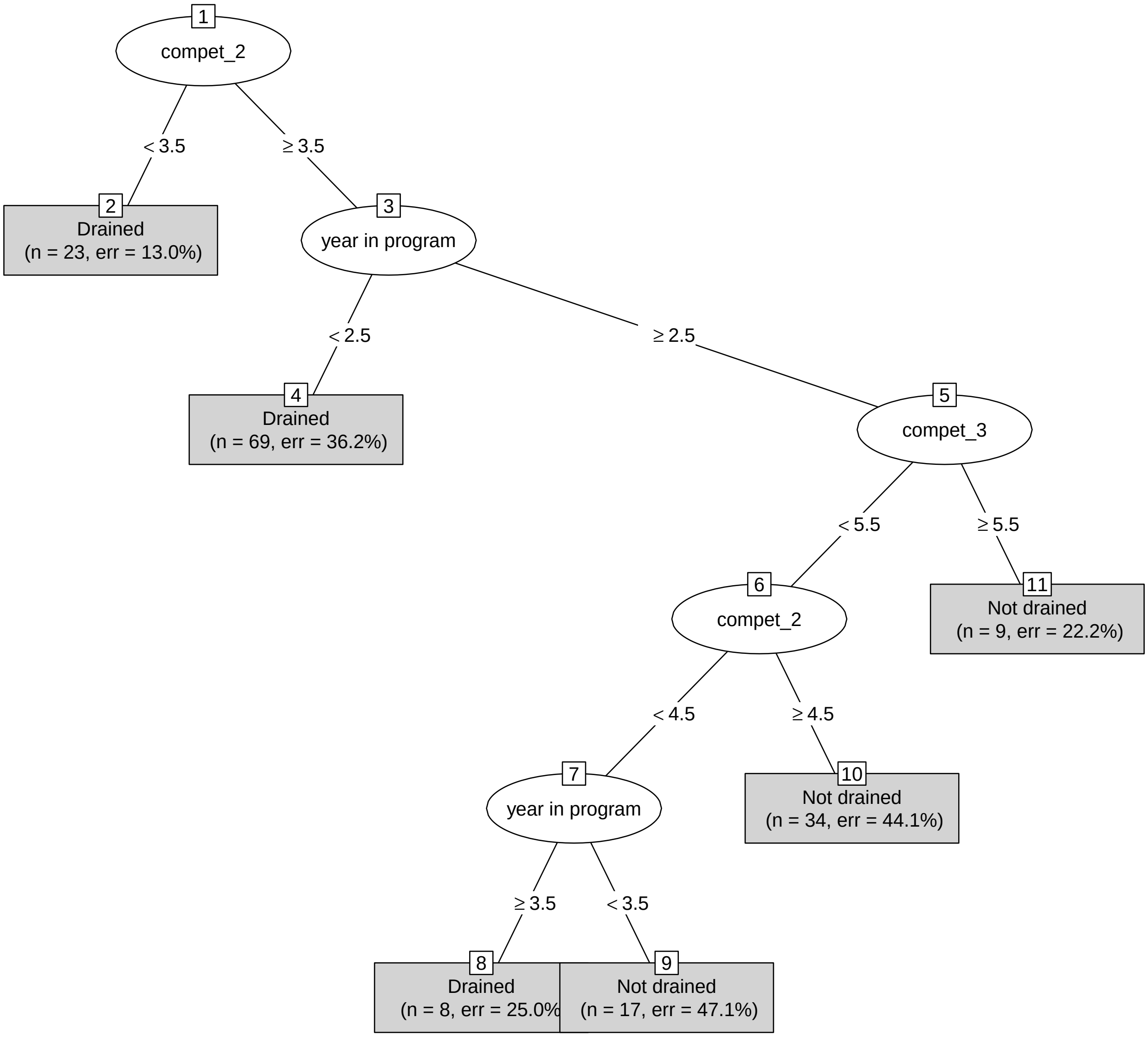
The following code creates a new variable `drained` that takes the value "Drained" if a student chose "Agree" or "Strongly Agree" with the statement "When I am studying or doing school work, I often feel emotionally drained." and "Not drained" otherwise.

The competencies questions were answered on a 1-6 scale, with high values indicating agreement and lower values indicating disagreement.

```
set.seed(42)
n <- nrow(burnout)
n
```

```
## [1] 200
```

```
burnout_b <- burnout %>%
  mutate(drained = case_when(exh_3 >= 3 ~ "Drained",
                             TRUE ~ "Not drained"))

# Random sample of 80% of row indices
training_indices <- sample(1:n, size=round(0.8*n))

train <- burnout_b %>% filter(participant_id %in% training_indices)
test <- burnout_b %>% filter(!(participant_id %in% training_indices))

# Fit the tree based on training data
tree1 <- rpart(drained ~ compet_2 + compet_3 + `year in program`, data=train)
plot(as.party(tree1), type="simple")
```



Based on `tree 1`, what would you predict for the following new students:

Student 1: A 4th year student who rated the statement "In my program, I feel competent." (`compet_2`) at a 2.

A. Predict drained. [CORRECT]

B. Predict not drained.

C. Not enough information to make a prediction.

Student 2: A 4th year student who rated the statement "In my program, I feel competent." (`compet_2`) at a 5.

A. Predict drained.

B. Predict not drained.

C. Not enough information to make a prediction. [CORRECT]

Student 3: A 1st year student who rated the statement " In my program, I feel competent." (`compet_2`) at a 5.

A. Predict drained. [CORRECT]

B. Predict not drained.

C. Not enough information to make a prediction.

Question 9

```
pred_train_1 <- predict(tree1, newdata=train, type="class")

pred_test_1 <- predict(tree1, newdata=test, type="class")

m1_train <- table(pred_train_1, train$drained)

m <- table(pred_test_1, test$drained)

m
```

```
##
## pred_test_1   Drained Not drained
##   Drained          20          11
##   Not drained       5           4
```

Calculate the sensitivity, specificity, true positive rate, and accuracy for 'tree1'. Suppose we are interested in identifying drained students to provide them with additional support. We will treat the response 'Drained' to be a POSITIVE and 'Not drained' to be a NEGATIVE. Round your answers to 2 decimals.

Sensitivity / True positive rate: 0.80

Specificity / True negative rate: 0.27

False positive rate: 0.73

Accuracy: 0.60

```
### ANSWER CALCULATIONS ###
tp <- m[1,1] / sum(m[,1])
tn <- m[2,2] / sum(m[,2])
accuracy <- sum(diag(m))/sum(m)
fp <- m[1,2]/sum(m[,2])
round(c(tp, tn, fp, accuracy), 2)
```

Question 10

Suppose that the results of this classification tree were going to be used to send support emails to students who were feeling emotionally drained about a 1-1 mentoring opportunity with an alumni member. Suppose the service is actually available to all students, but they wanted to promote it with specific language around coping with feeling emotionally drained. With this in mind, other than the language tailoring, it doesn't matter if students who aren't feeling emotionally drained get this email, but the University really wants to ensure as many students who are feeling drained get it as possible.

Which one of the following is TRUE?

A. The most appropriate prediction model for this purpose would be one where the accuracy is as high as possible.

B. The most appropriate prediction model for this purpose would be one where the sensitivity is as high as possible. [CORRECT]

C. The most appropriate prediction model for this purpose would be one where the specificity is as high as possible.

D. The most appropriate prediction model for this purpose would be one where the sensitivity and specificity are as similar as possible.

Question 11

Suppose that in an original version of this data, the participant IDs were actually ids much like your UTORid. That is, a combination of letters representing your name and then a number. Which one of the following reasons is MOST important reasons why the investigators would need to make sure to remove these before sharing the data?

A. They are not useful predictor variables.

B. They might have a high rate of typos because they are entered by the students.

C. Releasing this variable, especially in conjunction with the gender, age and year in program variables would mean a potentially very high re-identification risk. [CORRECT]

Question 12

In the journal article there is the following statement:

> Participation in the study was voluntary. Informed consent was implied by the overt action of completing the survey after reading the information letter. Students could choose not to respond to a question with no negative consequences to them.

This suggests that this study satisfies the following consent criteria:

A. Information only

B. Comprehension only

C. Voluntariness only

D. Voluntariness and Information only

E. Voluntariness, Information and Comprehension [CORRECT]

Question 13

Suppose another Canadian medical school was interested in using the findings of this research with their own students. For other purposes, they had already run the Basic Psychological Needs Survey with their students and so had scores for Autonomy, Competence, Relatedness and Self-compassion and wanted to see if, without any further surveying, they could get an idea of their own students' exhaustion scores.

Which one of the following is the BIGGEST problem with using this data for this purpose?

A. There is a low response rate for online surveys of students. You know, like course evaluations for example. So there could be non-response bias in this data.

B. Our data was samples from a different population and so it might not be appropriate to make inferences to this new university's population of medical students. [CORRECT]

C. We cannot make predictions without other answers to all the other surveys this data was collected using. I.e., we need answers to the Godin Leisure-Time Exercise Questionnaire (GLTE) and the Achievement Goals Instrument, as well.