Sample Questions - Solutions

- This document contains multiple choice questions that have a similar format to the multiple choice questions on the 2018 STA130 final exam.
- The sample questions are published to give students a "feel" for the exam questions.
- These sample questions are based on the course content after the midterm test.
- The final exam will cover all the material before and after the midterm test.

Continues on next page ⟶

From Wikipedia

> "The SAT is a standardized test widely used for college admissions in the United States. Introduced in 1926, its name and scoring have changed several times; originally called the Scholastic Aptitude Test, it was later called the Scholastic Assessment Test, then the SAT I: Reasoning Test, then the SAT Reasoning Test, and now, simply the SAT."
>
> "The SAT has four sections: Reading, Writing and Language, Math (no calculator), and Math (calculator allowed)."

Questions 1 through 4 use the `SAT_2010_1` data set, described below. The questions start on page 5.

A data set `SAT_2010` contains results by state for 2010. A few new variables were defined using the `R` code below.

```
glimpse(SAT_2010)
```

```
## Observations: 50
## Variables: 9
## $ state               <fct> Alabama, Alaska, Arizona, Arkansas, Califo...
## $ expenditure         <int> 10, 17, 9, 10, 10, 10, 16, 13, 9, 10, 13, ...
## $ pupil_teacher_ratio <dbl> 15.3, 16.2, 21.4, 14.1, 24.1, 17.4, 13.1, ...
## $ salary              <int> 49948, 62654, 49298, 49033, 71611, 51660, ...
## $ read                <int> 556, 518, 519, 566, 501, 568, 509, 493, 49...
## $ math                <int> 550, 515, 525, 566, 516, 572, 514, 495, 49...
## $ write               <int> 544, 491, 500, 552, 500, 555, 513, 481, 47...
## $ total               <int> 1650, 1524, 1544, 1684, 1517, 1695, 1536, ...
## $ sat_pct             <int> 8, 52, 28, 5, 53, 19, 87, 74, 64, 80, 64, ...
```

```
SAT_2010_1 <- SAT_2010 %>%
  mutate(total_high = ifelse(total >= quantile(SAT_2010$total, 0.75),
                        "Yes", "No")) %>%
  mutate(pupil_teacher_ratio_high =
           ifelse(pupil_teacher_ratio >= 16, "Yes", "No")) %>%
  mutate(id = row_number())
```


Continues on next page ⟶

Below is an explanation of all the variables.

`state` - a factor with levels for each state.

`expenditure` - average expenditure per student (in each state)

`pupil_teacher_ratio` - pupil to teacher ratio in that state

`salary` - teacher salary (in 2010 US $)

`read` - state average Reading SAT score

`math` - state average Math SAT score

`write` - state average Writing SAT score

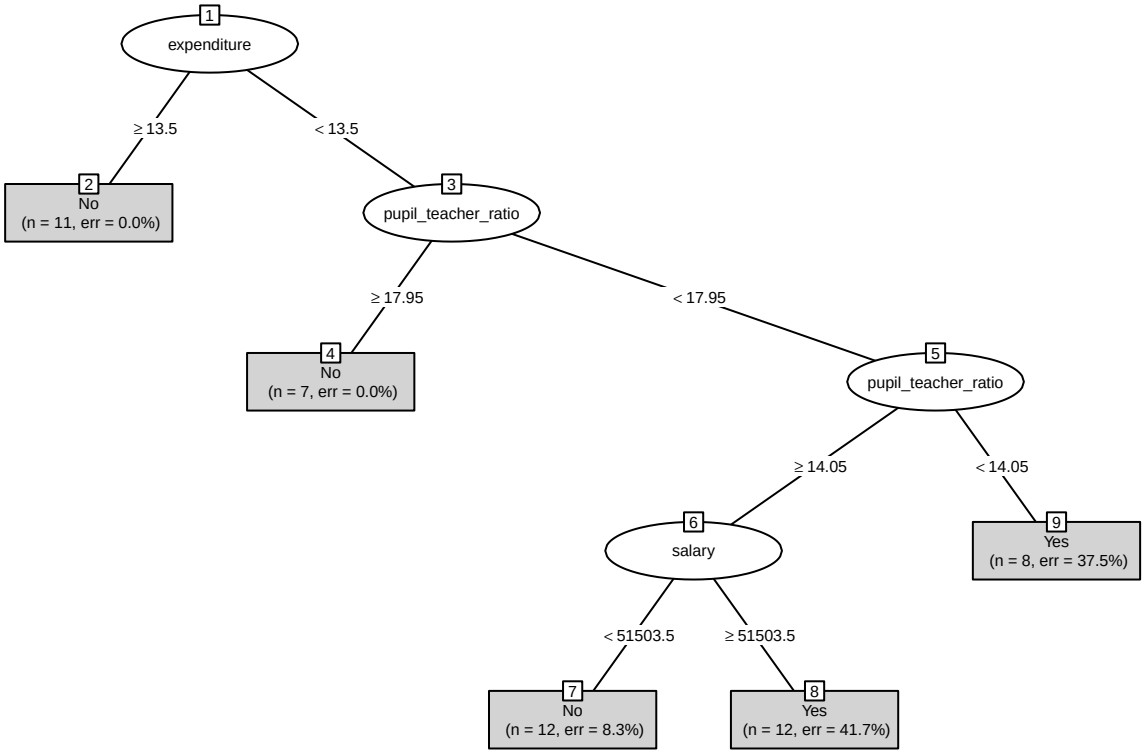
`total` - state average Total SAT score

`sat_pct` - percent of students taking SAT in that state

Does spending more on education improve SAT scores? Consider the classification tree to model to predict whether a state has an average total score at or above the 75^{th} percentile.

```
set.seed(25)
# Select training set
SAT_2010_1_train <- SAT_2010_1 %>% sample_frac(size = 0.8)
# Select test set
SAT_2010_1_test <- SAT_2010_1 %>% anti_join(SAT_2010_1_train, by = "id")

treemod <- rpart(total_high ~ expenditure + pupil_teacher_ratio + salary,
                 data = SAT_2010_1)
plot(as.party(treemod), type = "simple", gp = gpar(cex = 0.5))
```




```
predicted_tree <-
  predict(object = treemod, newdata = SAT_2010_1, type = "class")
m <- table(predicted_tree,SAT_2010_1$total_high)
m
```

```
##
## predicted_tree No Yes
##            No  29   1
##            Yes  8  12
```


Continues on next page ⟶

Answer the following questions based on the output for the classification tree on page 4.

1. Which of the following statements are TRUE?

   I) If expenditure is less than 13.5 and pupil teacher ratio is less then 14.05 then eight students had an average total SAT scores in the top 25% of SAT scores.

      False. The observations are for states not individual students.

   II) If expenditure is less than 13.5 and pupil teacher ratio is less then 14.05 then five states had an average total SAT scores in the top 25% of SAT scores.

      True. See terminal node 9. The predicted class for terminal node 9 is "Yes" and the error is 37.5% (i.e., 37.5% of the 8 states in this node are "No"). So, $(1 - 0.375) \times 8 = 5$ states are "Yes" (i.e., have average SAT scores in the top 25% of states).

   III) The dependent variable in the classification tree is average total SAT score.

      False. The dependent variable is a categorical variable based on average total SAT score.

   IV) The dependent variable in the classification tree is equal to "Yes" if a state's average SAT score is greater than the 25^{th} percentile, and "No" otherwise.

      False. The `R` code `mutate(total_high = ifelse(total >= quantile(SAT_2010$total, 0.75), "Yes", "No"))` defines the dependent variable as an average SAT score at least as large as the 75^{th} percentile.

   V) The dependent variable in the classification tree is equal to "Yes" if a state's average total SAT score is greater than or equal to the 75^{th} percentile, and "No" otherwise.

      True. See answer above.

Which of the above statements are TRUE?

A) All the statements are TRUE.

B) Only the statements I) and II) are TRUE.

C) Only the statements II) and V) are TRUE. **(Correct)**

D) Only the statements II) and III) are TRUE.

E) Only the statements III) and V) are TRUE.

Answer the following question based on the output for the classification tree on page 4.

2. Which of the following statements are FALSE?

   I) The overall accuracy of the classification using the test set is $\frac{29+12}{(29+12+8+1)} = 0.82$.

   False. A test and validation set were not used, but they are in the code.

   II) The sensitivity of the classification using `SAT_2010_1` is $\frac{12}{13} = 0.92$.

   True. See confusion matrix.

   III) The specificity of the classification using `SAT_2010_1` is $\frac{8}{8+29} = 0.22$.

   False. This is the false-positive rate.

   IV) Nine states were misclassified using the classification tree.

   True. See confusion matrix.

Which of the above statements are FALSE?

A) All the statements are FALSE.

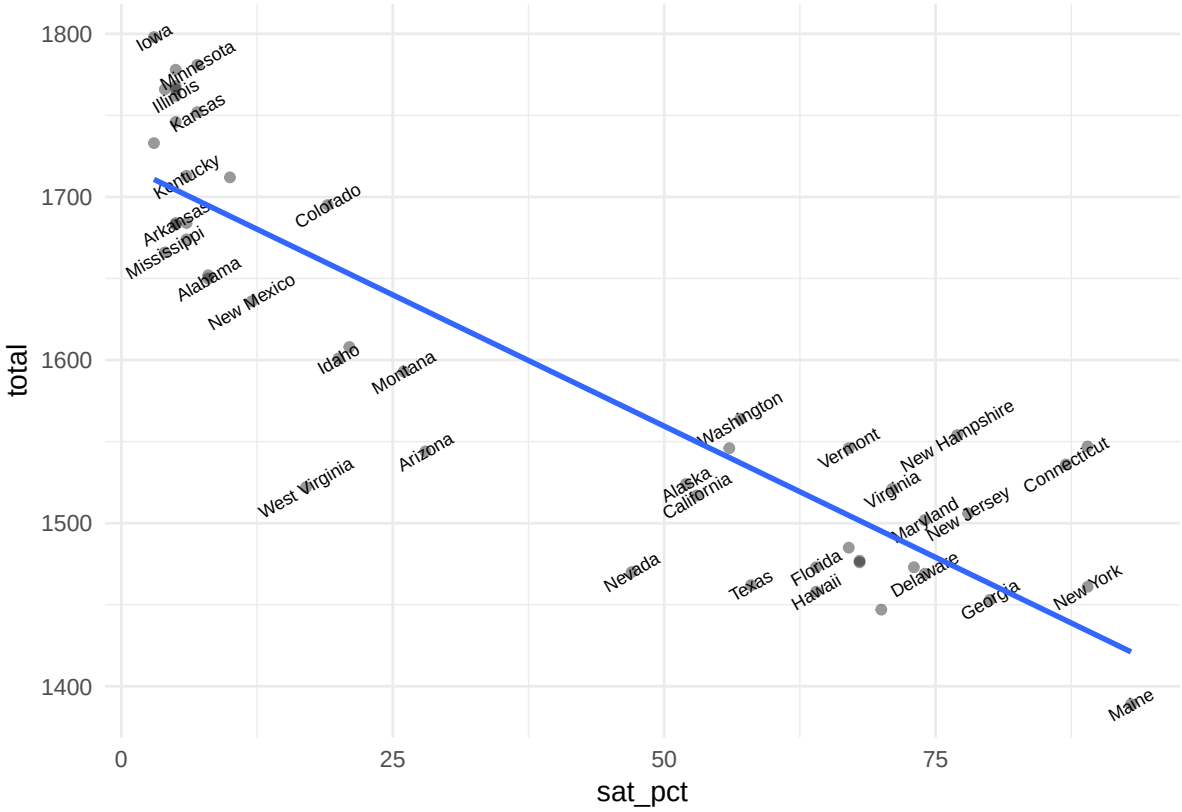
B) Only the statements I) and II) are FALSE.

C) Only the statements II) and V) are FALSE.

D) Only the statements II) and III) are FALSE.

E) Only the statements I) and III) are FALSE. **(Correct)**

A linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where y is total SAT score and x is percent of students taking SAT in that state was fit to the data.




```
reg_mod_simple <- lm(total ~ sat_pct, data = SAT_2010_1)
tidy(reg_mod_simple) %>% select(term, estimate, p.value)
```

```
##          term    estimate      p.value
## 1 (Intercept) 1720.441288 3.056958e-64
## 2     sat_pct   -3.218621 2.512011e-17
```

```
summary(reg_mod_simple)$r.squared
```

```
## [1] 0.7784987
```


Continues on next page ⟶

```
model_dat <- data_frame(state = SAT_2010_1$state,
                        sat_pct = SAT_2010_1$sat_pct,
                        total = SAT_2010_1$total,
                        yhat = predict(reg_mod_simple))
head(model_dat)
```

```
## # A tibble: 6 x 4
##   state      sat_pct total  yhat
##   <fct>        <int> <int> <dbl>
## 1 Alabama          8  1650  1695
## 2 Alaska          52  1524  1553
## 3 Arizona         28  1544  1630
## 4 Arkansas         5  1684  1704
## 5 California      53  1517  1550
## 6 Colorado        19  1695  1659
```


Continues on next page ⟶

Answer the following questions based on the output for the regression model on page 7 and 8.

3. Which of the following statements are TRUE?

   I) A linear regression model is appropriate to describe the relationship between `total` and `sat_pct`.

      True. The scatterplot indicates that there is a linear relationship between these two variables.

   II) The residual for Alaska is 1472.

      False. The residual is $total_{\text{Alaska}} - \hat{total}_{\text{Alaska}} = 1524 - 1553 =$ -29.

   III) The least-squares estimate of the intercept is $\hat{\beta}_0 = 1720.441288$, and the least-squares estimate slope is $\hat{\beta}_1 =$ -3.2186212.

      True. See below.

```
tidy(reg_mod_simple) %>% select(term, estimate, p.value)
```

```
##          term    estimate      p.value
## 1 (Intercept) 1720.441288 3.056958e-64
## 2     sat_pct   -3.218621 2.512011e-17
```

   IV) The linear regression model indicates that as the percent of students in a state taking the SAT test increases the average total score increases.

      False. The slope is negative and scatterplot indicates that as the percent of students in a state taking the SAT test *decreases* the average total score *increases*.

Which of the above statements are TRUE?

A) None of the statements are TRUE.

B) Only the statements I) and IV) are TRUE.

C) Only the statements II) and III) are TRUE.

D) Only the statements I) and III) are TRUE. **(Correct)**

E) Only the statements II) and IV) are TRUE.

Is the relationship between SAT scores and percent of students in a state writing the SAT different in states that have smaller classes? A regression analysis was conducted to evaluate this question.

```
reg_mod1 <- lm(total ~ pupil_teacher_ratio_high, data = SAT_2010_1)
tidy(reg_mod1) %>% select(term, estimate, p.value)
```

```
##                          term   estimate      p.value
## 1                 (Intercept) 1602.14706 1.351386e-52
## 2 pupil_teacher_ratio_highYes  -17.77206 6.204414e-01
```


Continues on next page ⟶

Answer the following question based on `R` output on page 10.

4. Which of the following statements are FALSE?

   I) The regression analysis provides strong evidence that the slope of percent of students writing is different for states with high and low pupil teacher ratios.

      False. `sat_pct` is not included in the regression model so there is no estimate of these two slopes.

   II) This is an example of a multiple regression model?

      False. A multiple regression model requires at least two independent variables. There is only one independent variable in this model.

   III) A different regression model is needed to evaluate the primary question: "Is the relationship between SAT scores and percent of students in a state writing the SAT different in states that have smaller classes".

      True.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i,$$

      where, y_i is the i^{th} value of `total`, x_{i1} is the i^{th} value of `sat_pct`, and x_{i2} is the i^{th} value of `pupil_teacher_ratio` is a model that would be appropriate to evaluate this question.

   IV) There is strong evidence that SAT scores are different in states with high versus low pupil teacher ratios since the p-value of the appropriate regression coefficient is very small.

      False. The appropriate regression coefficient to consider is the slope not the intercept. The P-value for the slope is 0.62.

Which of the above statements are FALSE?

A) All the statements are FALSE except for I).

B) All the statements are FALSE except for II).

C) All the statements are FALSE except for III). **(Correct)**

D) All the statements are FALSE except for IV).

E) All the statements are FALSE.

5. A data scientist not affiliated with a university compiled data from several public sources (voter registration, political contributions, tax records) that were used to predict sexual orientation of individuals in a community. Which of the following is an appropriate ethical consideration that should guide the use of such data sets?

   A) The data scientist should ensure that the use of the data sets follows the Nuremberg code's suggestions for the use of data sets.

      This is not an appropriate ethical consideration since the Nuremberg code does not discuss the appropriate use of public data sets.

   B) The analyst should ensure that the individuals cannot be reidentified using the data. The data scientist could, for example, remove certain variables in addition to username that would make it more difficult to reidentify people.

      This is an appropriate ethical consideration. When public data is combined it may result in a data set where individuals are easily identified. **(Correct)**

   C) Ensure that a university ethics board reviews the methods.

      The data scientist is not affiliated with a university so this review is not appropriate.

   D) Obtain informed consent from all users in the data sets.

      This is neither realistic nor necessary.

   E) There are no ethical considerations since the data is public.

      There are almost always ethical considerations when combining public data sets. See explanation to 5B.