431 Quiz 1 Answer Sketch

Thomas E. Love

2017-10-10

Setup

There are a total of 41 questions on the quiz.

Please select or type in your best response (or responses, as indicated) for each question. The questions are not arranged in any particular order, and your score is based on the number of correct responses, so you should answer all questions. Questions 1-40 are each worth 2, 2.5 or 3 points, and the maximum possible score on the quiz is 100 points.

The deadline for completing the quiz is **Noon on Monday October 9**. If you wish to work on some of the quiz and then return to work on the rest of the quiz or edit your responses, you can do this by [1] completing the final question which asks you to type in your full name, and then [2] submit the quiz. You will then receive a link which allows you to return to the quiz without losing your progress.

You are welcome to consult the materials provided on the course website, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the other helpful folks at `431-help` at `case` dot `edu`.

Grading

You must answer Question 41 for Dr. Love to grade your quiz, but it doesn't count otherwise.

For the other 40 questions, 18 are worth 2 points, 4 are worth 2.5 and the remaining 18 are worth 2 points, as indicated below.

Value	Questions
2 pts	1, 4, 7, 10, 13, 15, 16, 17, 22, 24-29, 31, 34-35
2.5 pts	2, 8, 20-21
3 pts	3, 5-6, 9, 11-12, 14, 18-19, 23, 30, 32-33, 36-40

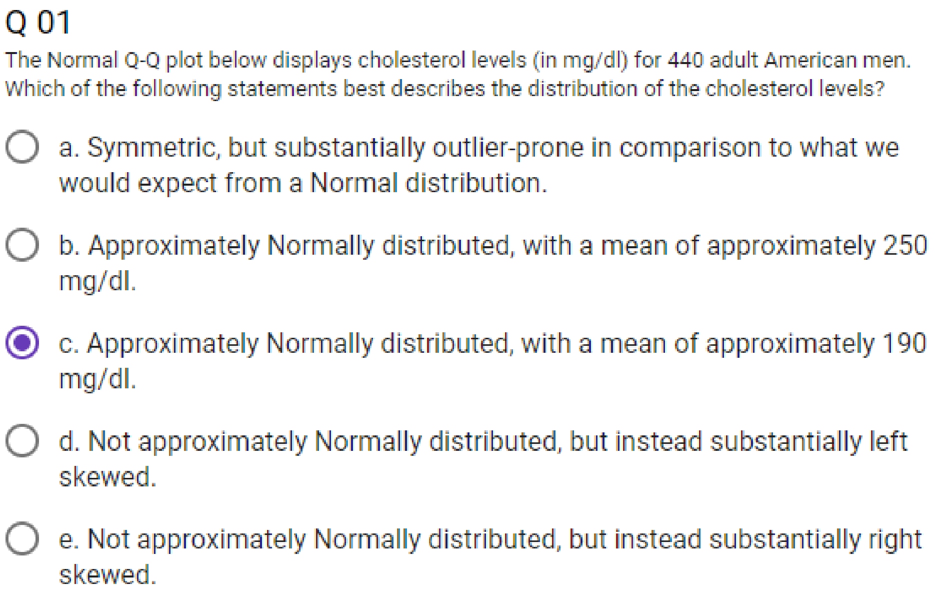
The maximum possible score is thus: `18 * 2 + 4 * 2.5 + 18 * 3` or **100** points. In advance, I expect plenty of students to score in the A range, but a score of 100 is unlikely.

Partial Credit?

Generally, items are scored as either correct (full points) or incorrect (0 points).

- Questions 2, 5, 6, 8, 9, 19, 24, 30, 36 and 39 are checkbox items or items where multiple responses are required. In each case, a series of individual responses are marked, so that students can earn less than full credit for partially correct responses.
- On Questions 3, 13, 16 and 18, partially correct responses on questions requiring a calculation and rounding are possible.
- On Question 31, I decided to award partial credit to one incorrect response.
- On Questions 14, 23, 32 and 33 (items requiring R code), I reserve the right to award partial credit. The plan for 33 is outlined below, and a similar approach will be taken on the other items as necessary.

Answer 01 is c, for 2 points.



Q 01

The Normal Q-Q plot below displays cholesterol levels (in mg/dl) for 440 adult American men. Which of the following statements best describes the distribution of the cholesterol levels?

○ a. Symmetric, but substantially outlier-prone in comparison to what we would expect from a Normal distribution.

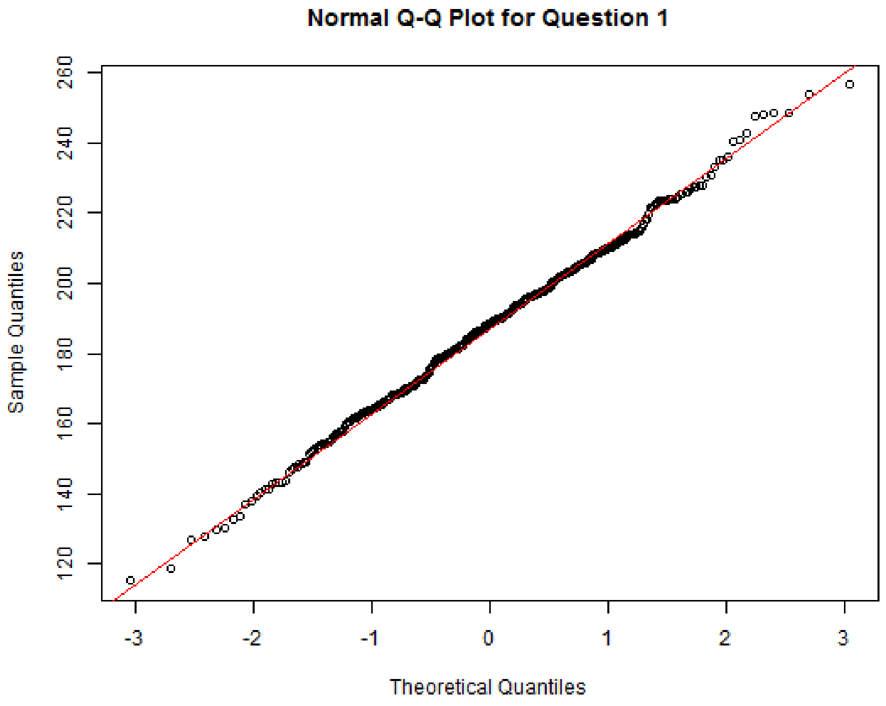
○ b. Approximately Normally distributed, with a mean of approximately 250 mg/dl.

● c. Approximately Normally distributed, with a mean of approximately 190 mg/dl.

○ d. Not approximately Normally distributed, but instead substantially left skewed.

○ e. Not approximately Normally distributed, but instead substantially right skewed.

Figure for Question 1



The plot clearly mirrors what we would expect from a Normal distribution (a straight diagonal line in the Normal Q-Q plot), and the mean should thus be at the value on the Sample Quantiles scale corresponding to a Theoretical Quantiles value of 0 (a Z score of 0 will be the mean.) The center of the distribution corresponds to a sample value (as shown on the Y axis) between 180 and 200.

Answer 02 is worth 2.5 points

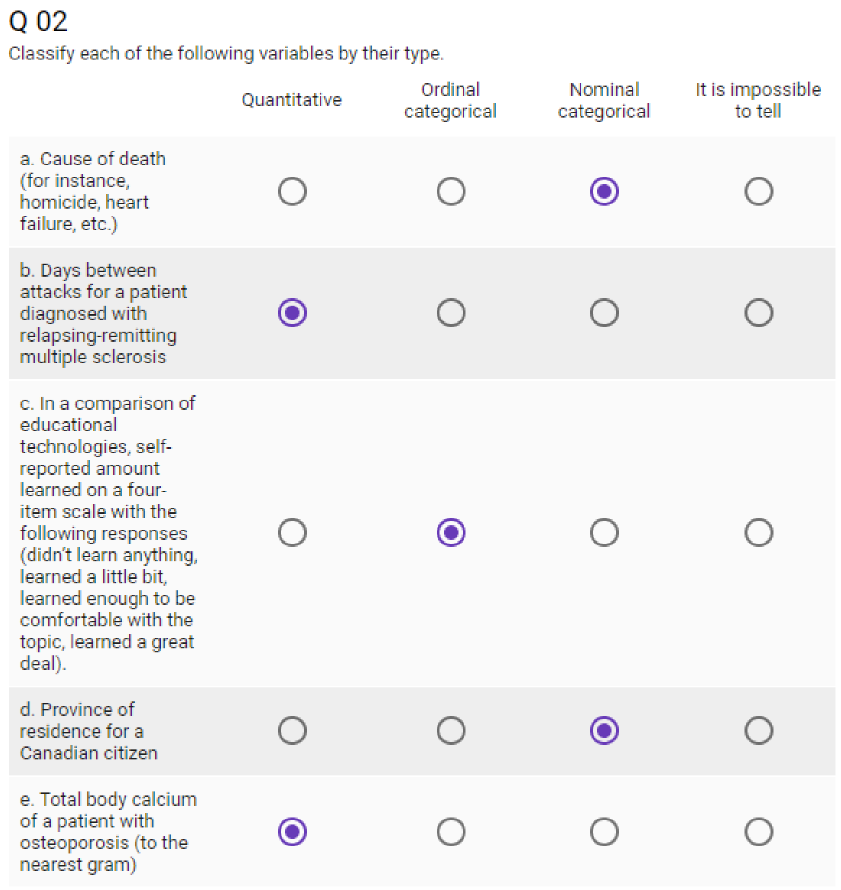
Answer 02a is Nominal categorical

Answer 02b is Quantitative

Answer 02c is Ordinal categorical

Answer 02d is Nominal categorical

Answer 02e is Quantitative



Q 02

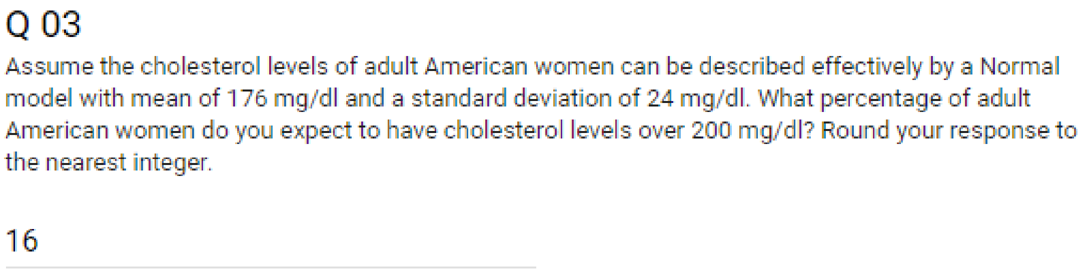
Classify each of the following variables by their type.

	Quantitative	Ordinal categorical	Nominal categorical	It is impossible to tell
a. Cause of death (for instance, homicide, heart failure, etc.)	○	○	●	○
b. Days between attacks for a patient diagnosed with relapsing-remitting multiple sclerosis	●	○	○	○
c. In a comparison of educational technologies, self-reported amount learned on a four-item scale with the following responses (didn't learn anything, learned a little bit, learned enough to be comfortable with the topic, learned a great deal).	○	●	○	○
d. Province of residence for a Canadian citizen	○	○	●	○
e. Total body calcium of a patient with osteoporosis (to the nearest gram)	●	○	○	○

Partial Credit

For each correct response in a-e, you received 0.5 point.

Answer 03 is 16 percent, for 3 points.



Q 03

Assume the cholesterol levels of adult American women can be described effectively by a Normal model with mean of 176 mg/dl and a standard deviation of 24 mg/dl. What percentage of adult American women do you expect to have cholesterol levels over 200 mg/dl? Round your response to the nearest integer.

16

I had hoped you would figure this out by realizing that this is equivalent to finding the probability that a random draw from a Normal distribution will be one standard deviation or more above the mean. Since 68% of a Normal distribution is within one standard deviation of the mean, 32% must be outside that range, and since the distribution is symmetric, half of that 32% must be one standard deviation or more above the mean, so the answer is 16%.

But, you could also have run the following R calculation:

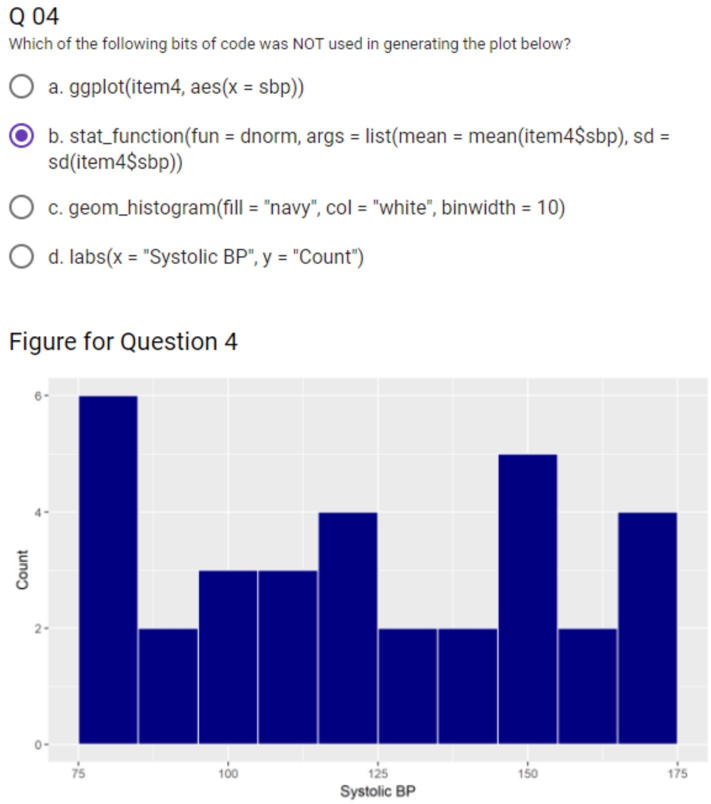
```
[1] 0.1586553
```

which, converted to a percentage then rounded to the nearest integer, also yields 16%.

Partial Credit

- I gave full credit to 16 or 16%.
- I gave 2 points to responses of 15.9 or 15.9%, where the student didn't round.
- I gave 1 point to responses of 0.16 or 0.159, where the student didn't convert to a percentage or round.

Answer 04 is b, for 2 points.



Q 04

Which of the following bits of code was NOT used in generating the plot below?

○ a. ggplot(item4, aes(x = sbp))

● b. stat_function(fun = dnorm, args = list(mean = mean(item4$sbp), sd = sd(item4$sbp))

○ c. geom_histogram(fill = "navy", col = "white", binwidth = 10)

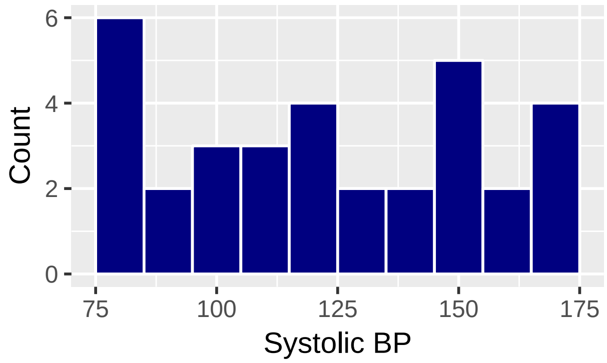
○ d. labs(x = "Systolic BP", y = "Count")

Figure for Question 4

The plot was actually formed with the other three bits of code.

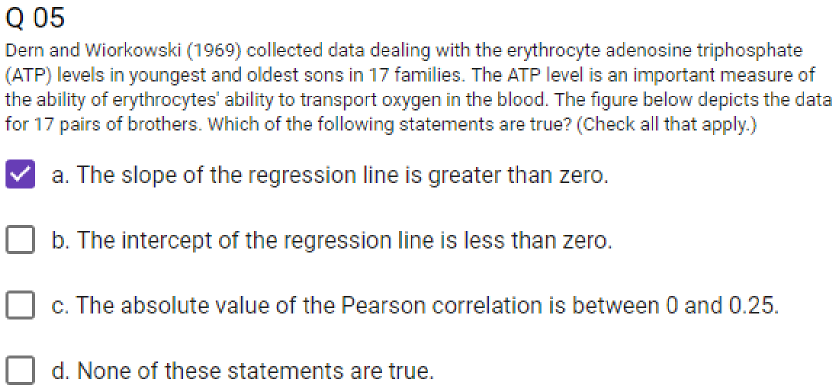
```
item4 <- read.csv("item04.csv") %>% tbl_df

ggplot(item4, aes(x = sbp)) +
  geom_histogram(fill = "navy", col = "white",
                 binwidth = 10) +
  labs(x = "Systolic BP", y = "Count")
```



If we'd added the `stat_function` statement shown in part b, I suppose that would have been moving towards adding a Normal density estimate on top of the plot, although it wouldn't have worked properly because it's short one right parenthesis) at the end, and because we didn't plot the histogram on a density scale.

Answer 05 is a, for 3 points.



Q 05

Dern and Wiorkowski (1969) collected data dealing with the erythrocyte adenosine triphosphate (ATP) levels in youngest and oldest sons in 17 families. The ATP level is an important measure of the ability of erythrocytes' ability to transport oxygen in the blood. The figure below depicts the data for 17 pairs of brothers. Which of the following statements are true? (Check all that apply.)

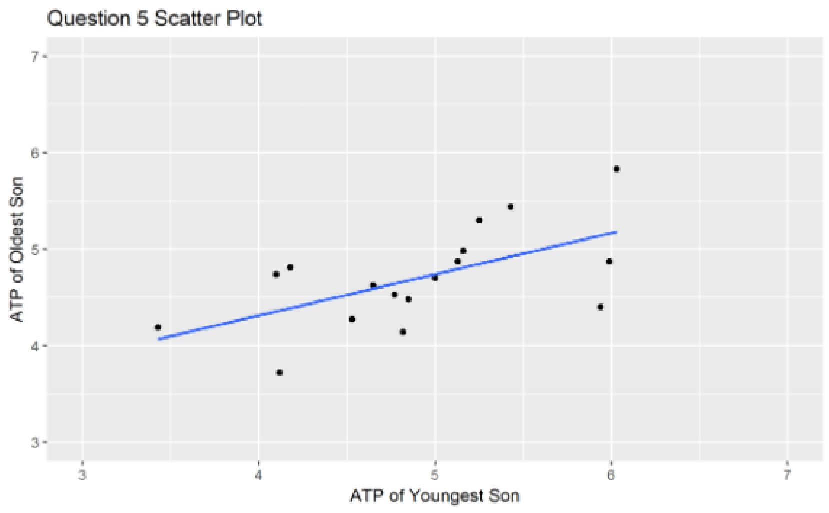
☑ a. The slope of the regression line is greater than zero.

□ b. The intercept of the regression line is less than zero.

□ c. The absolute value of the Pearson correlation is between 0 and 0.25.

□ d. None of these statements are true.

Figure for Question 5

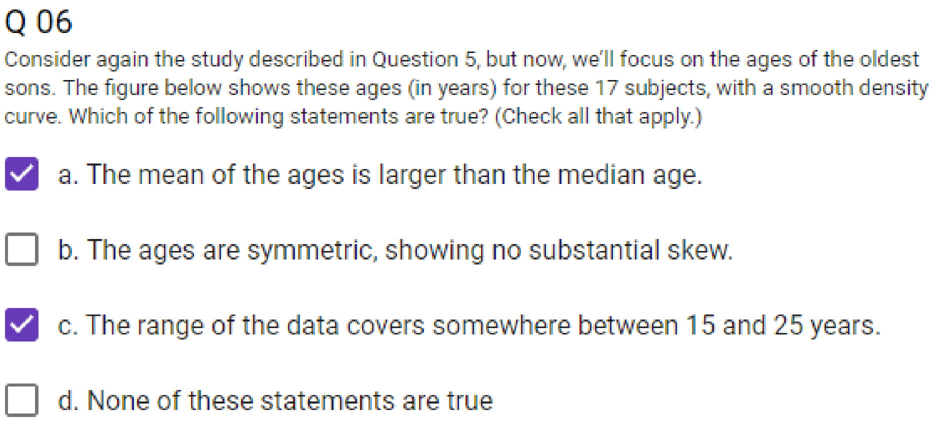


- Statement a is clearly true - the slope of the regression line is definitely positive. Higher levels of ATP of the youngest son are associated with higher ATP in the older son.
- Statement b requires a little thought, but extrapolating the line to see where it would cross the y-axis when the ATP of the youngest son is 0 suggests that the intercept is going to be somewhere between 2 and 3, in any case not negative, so Statement b is false. The actual regression line is `old.atp` $= 2.6 + 0.43$ `young.atp`
- As for Statement c, The correlation is pretty strong here, in fact it turns out to be 0.6, at any rate much higher than 0.25. A correlation as low as 0.25 would indicate a very weak relationship, with points scattered far away from the straight line, so c is false.

Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

Answer 06 is a and c, for 3 points.



Q 06

Consider again the study described in Question 5, but now, we'll focus on the ages of the oldest sons. The figure below shows these ages (in years) for these 17 subjects, with a smooth density curve. Which of the following statements are true? (Check all that apply.)

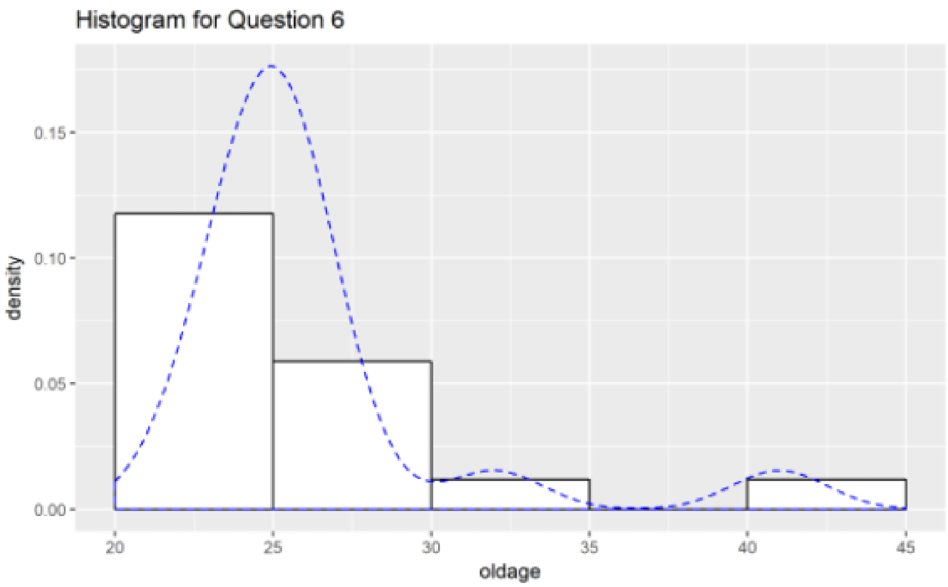
☑ a. The mean of the ages is larger than the median age.

□ b. The ages are symmetric, showing no substantial skew.

☑ c. The range of the data covers somewhere between 15 and 25 years.

□ d. None of these statements are true

Figure for Question 6



- These are right-skewed data, according to the histogram, so statement a is true, and statement b is false.
- The data range from a bin marked 20-25 to a bin marked 40-45, so the range could be as small as 15 (40-25) and as large as 25 (45-20), so statement c is true, too.

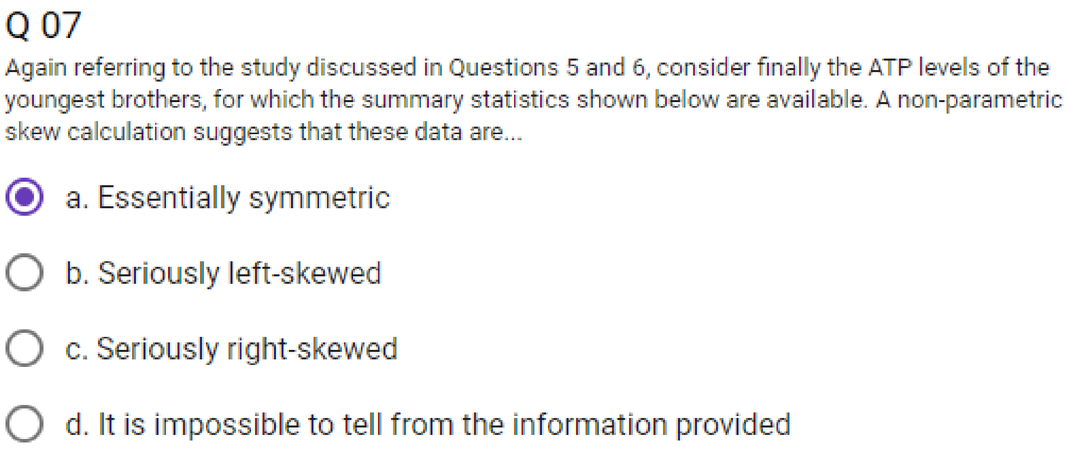
Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

This item didn't go well.

Most people who got it wrong got either `a` or `c`, but not both.

Answer 07 is a, for 2 points.



Q 07

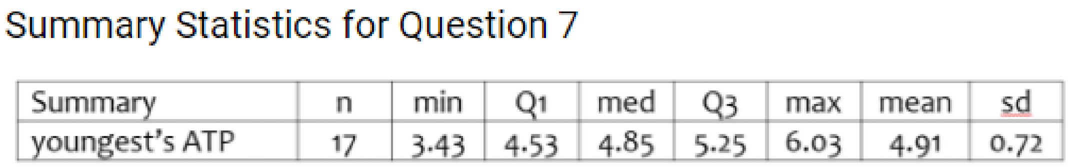
Again referring to the study discussed in Questions 5 and 6, consider finally the ATP levels of the youngest brothers, for which the summary statistics shown below are available. A non-parametric skew calculation suggests that these data are...

● a. Essentially symmetric

○ b. Seriously left-skewed

○ c. Seriously right-skewed

○ d. It is impossible to tell from the information provided

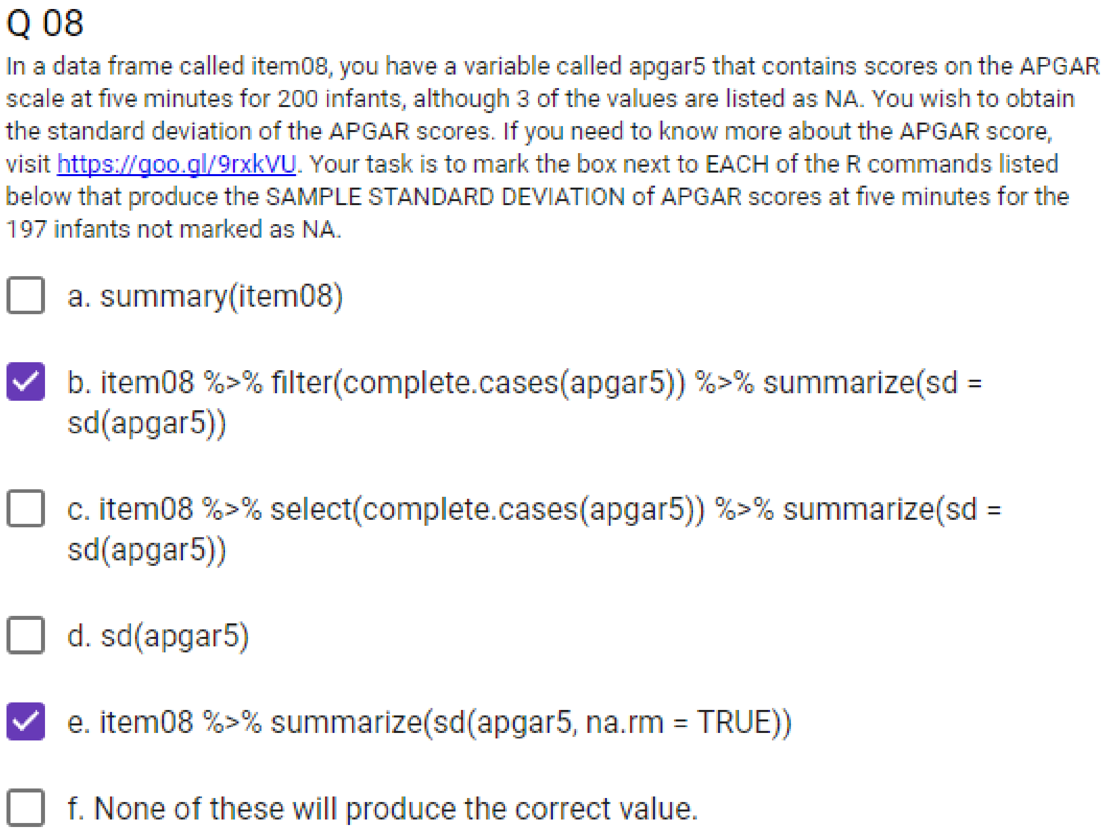


Summary Statistics for Question 7

Summary	n	min	Q1	med	Q3	max	mean	sd
youngest's ATP	17	3.43	4.53	4.85	5.25	6.03	4.91	0.72

The skew1 = (4.91 - 4.85)/0.72 = 0.08, which indicates no substantial skew. To show substantial skew, the nonparametric skew would have to be greater than 0.2 in absolute value. So, by this measure, the data are essentially symmetric.

Answer 08 is b and e, for 2.5 points.



Q 08

In a data frame called item08, you have a variable called apgar5 that contains scores on the APGAR scale at five minutes for 200 infants, although 3 of the values are listed as NA. You wish to obtain the standard deviation of the APGAR scores. If you need to know more about the APGAR score, visit https://goo.gl/9rxkVU. Your task is to mark the box next to EACH of the R commands listed below that produce the SAMPLE STANDARD DEVIATION of APGAR scores at five minutes for the 197 infants not marked as NA.

□ a. summary(item08)

☑ b. item08 %>% filter(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))

□ c. item08 %>% select(complete.cases(apgar5)) %>% summarize(sd = sd(apgar5))

□ d. sd(apgar5)

☑ e. item08 %>% summarize(sd(apgar5, na.rm = TRUE))

□ f. None of these will produce the correct value.

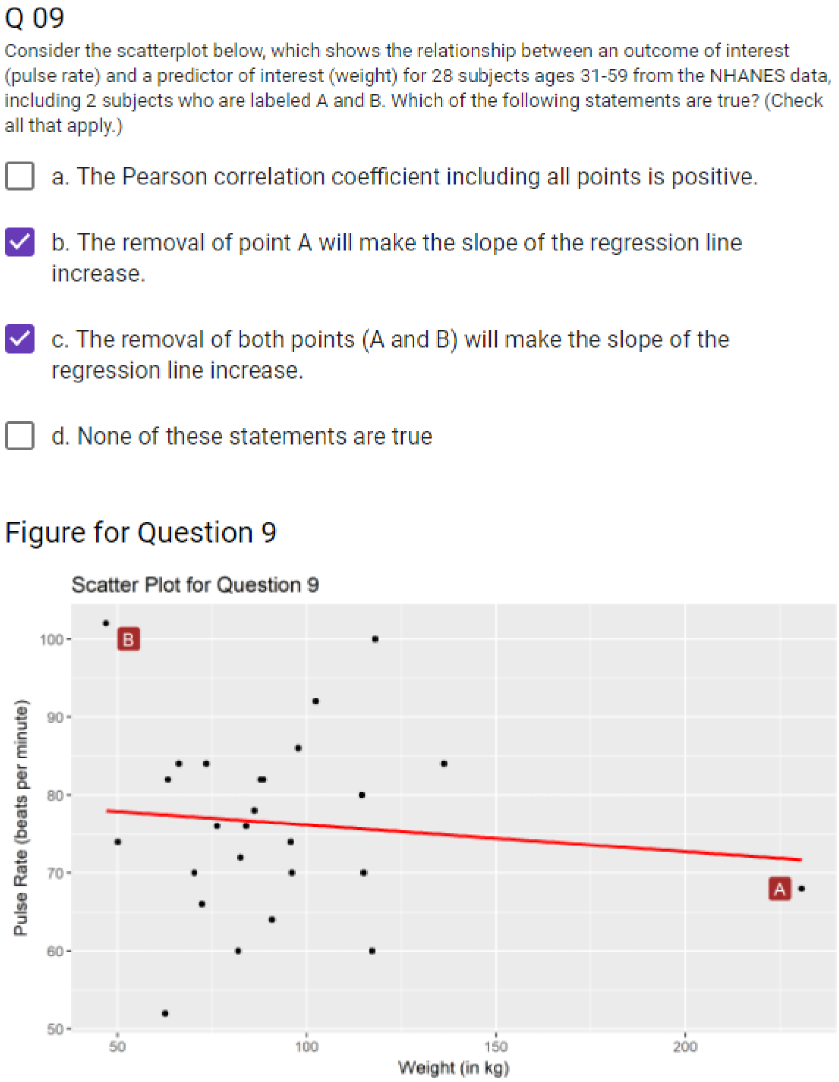
Statements b and e will produce the appropriate standard deviation.

- Statement a doesn't work because the `summary` function doesn't present the standard deviation.
- Statement c doesn't work because `select` is for picking columns (variables) rather than rows (observations) and you need `filter` to pick rows when using `complete.cases`.
- Statement d doesn't work because there are missing values in apgar5, so the result this gives is NA. You'd have to include na.rm=TRUE to make it work.

Partial Credit

For each correct check or no check in boxes a-e, you received 0.5 point. If you checked f, regardless of what else you checked, you received no points.

Answer 09 is b and c, for 3 points.



Q 09

Consider the scatterplot below, which shows the relationship between an outcome of interest (pulse rate) and a predictor of interest (weight) for 28 subjects ages 31-59 from the NHANES data, including 2 subjects who are labeled A and B. Which of the following statements are true? (Check all that apply.)

☐ a. The Pearson correlation coefficient including all points is positive.

☑ b. The removal of point A will make the slope of the regression line increase.

☑ c. The removal of both points (A and B) will make the slope of the regression line increase.

☐ d. None of these statements are true

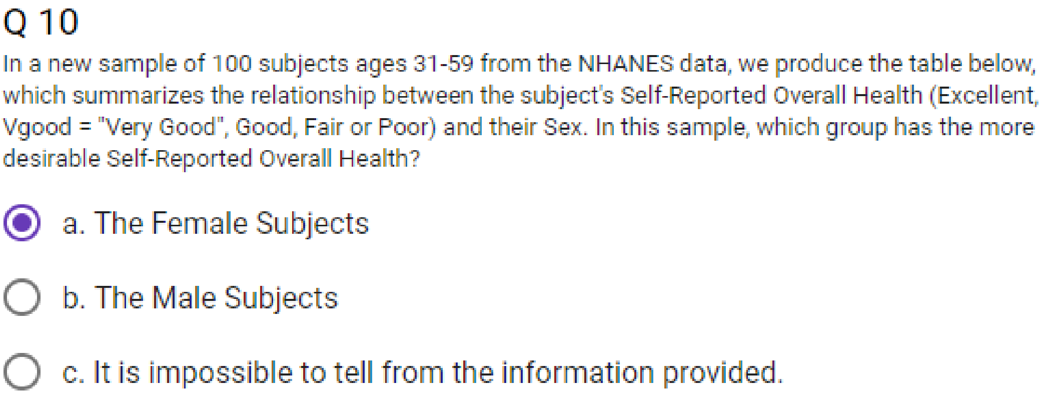
Figure for Question 9

- The slope of the regression line is negative when all points are included. Therefore, the Pearson correlation will also be negative. So statement a is false.
- Point A is dragging the line down on the right side. Removing it will therefore increase the slope (making it closer to 0 or even positive, it appears.) So statement b is true.
- Point B is pulling the line up on the left side, so removing it will also increase the slope. Removing *both* A and B will certainly increase the slope. So statement c is true.

Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

Answer 10 is a, for 2 points.



Q 10

In a new sample of 100 subjects ages 31-59 from the NHANES data, we produce the table below, which summarizes the relationship between the subject's Self-Reported Overall Health (Excellent, Vgood = "Very Good", Good, Fair or Poor) and their Sex. In this sample, which group has the more desirable Self-Reported Overall Health?

● a. The Female Subjects

○ b. The Male Subjects

○ c. It is impossible to tell from the information provided.

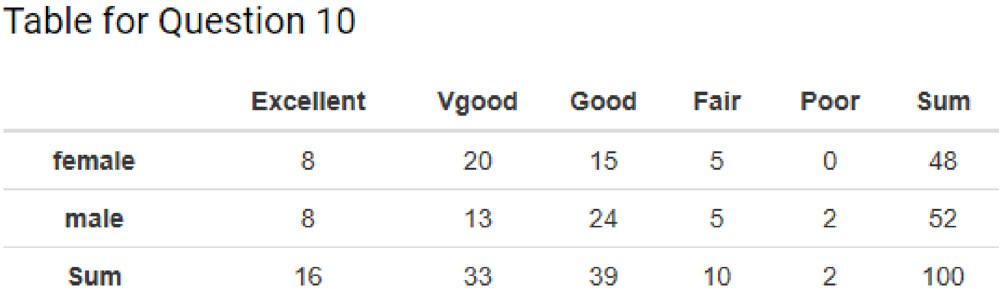


Table for Question 10

	Excellent	Vgood	Good	Fair	Poor	Sum
female	8	20	15	5	0	48
male	8	13	24	5	2	52
Sum	16	33	39	10	2	100

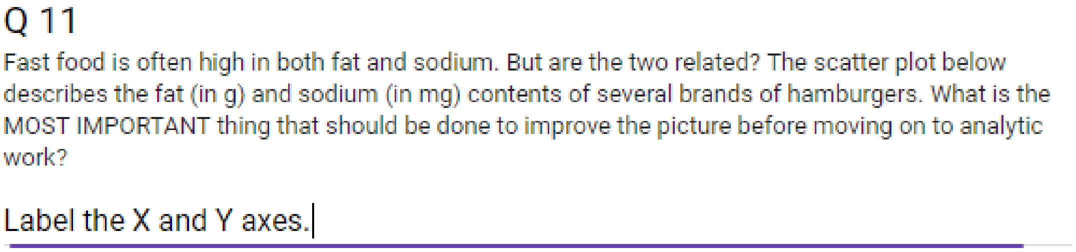
The female group has 48 subjects, while the male group has 52. If we look at the difference between the groups, it is entirely in the Vgood, Good and Poor groups. The female group has more Vgood and the male group has more Good and Poor results. So the **female** distribution is more desirable (better health).

We could calculate the row percentages to see this more clearly, if you like. . .

	Excellent	Vgood	Good	Fair	Poor
female	16.7	41.7	31.2	10.4	0
male	15.4	25.0	46.2	9.6	3.8

If we describe Excellent as 5, VGood as 4, etc. down to Poor as 1, we have the female mean at 3.65 and the male mean at 3.39.

Answer 11 is Label the axes for 3 points.

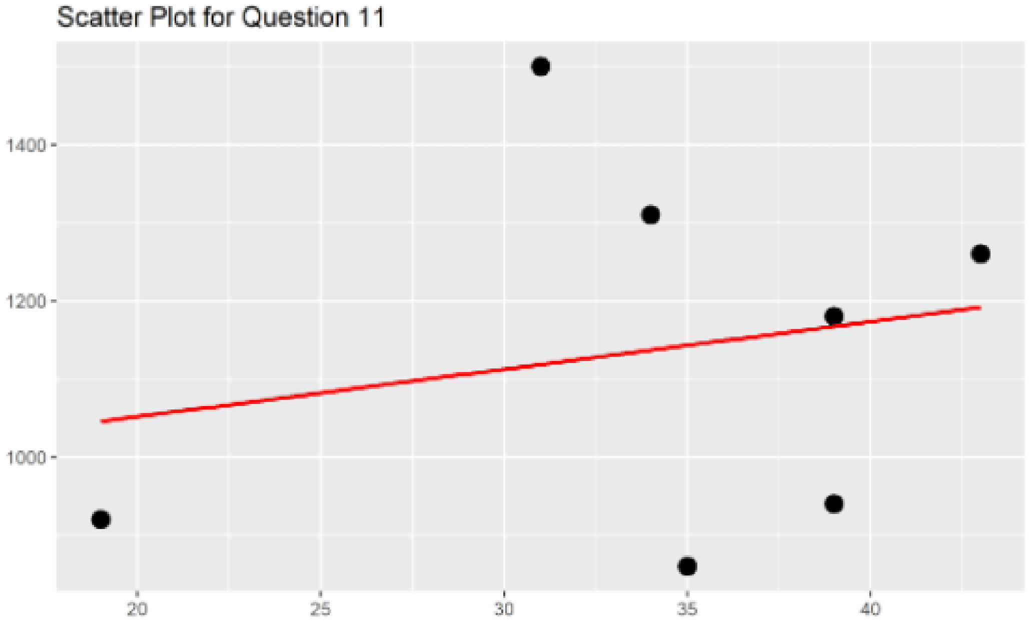


Q 11

Fast food is often high in both fat and sodium. But are the two related? The scatter plot below describes the fat (in g) and sodium (in mg) contents of several brands of hamburgers. What is the MOST IMPORTANT thing that should be done to improve the picture before moving on to analytic work?

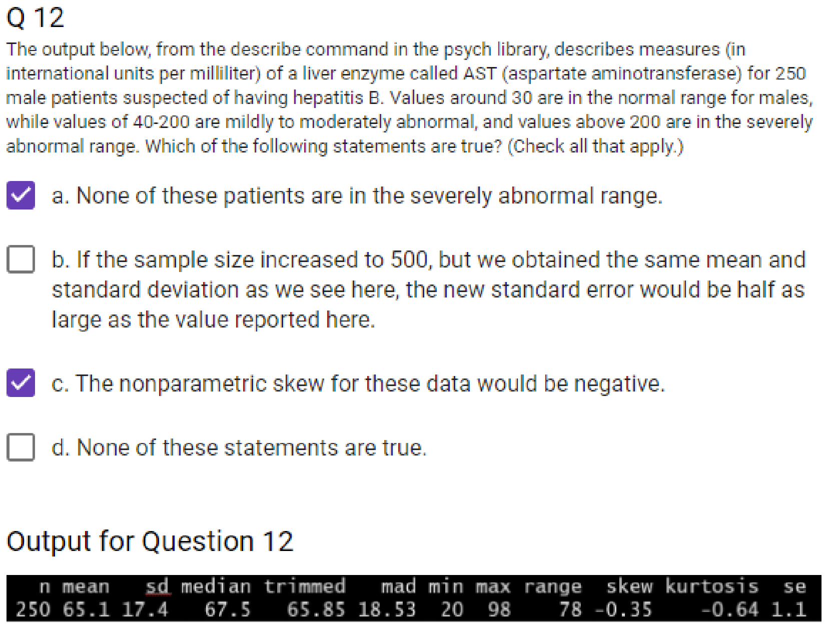
Label the X and Y axes.

Figure for Question 11



The correct response is to label the axes. Those who didn't come up with this response usually focused on making some sort of transformation.

Answer 12 is a and c for 3 points.



Q 12

The output below, from the describe command in the psych library, describes measures (in international units per milliliter) of a liver enzyme called AST (aspartate aminotransferase) for 250 male patients suspected of having hepatitis B. Values around 30 are in the normal range for males, while values of 40-200 are mildly to moderately abnormal, and values above 200 are in the severely abnormal range. Which of the following statements are true? (Check all that apply.)

☑ a. None of these patients are in the severely abnormal range.

□ b. If the sample size increased to 500, but we obtained the same mean and standard deviation as we see here, the new standard error would be half as large as the value reported here.

☑ c. The nonparametric skew for these data would be negative.

□ d. None of these statements are true.

Output for Question 12

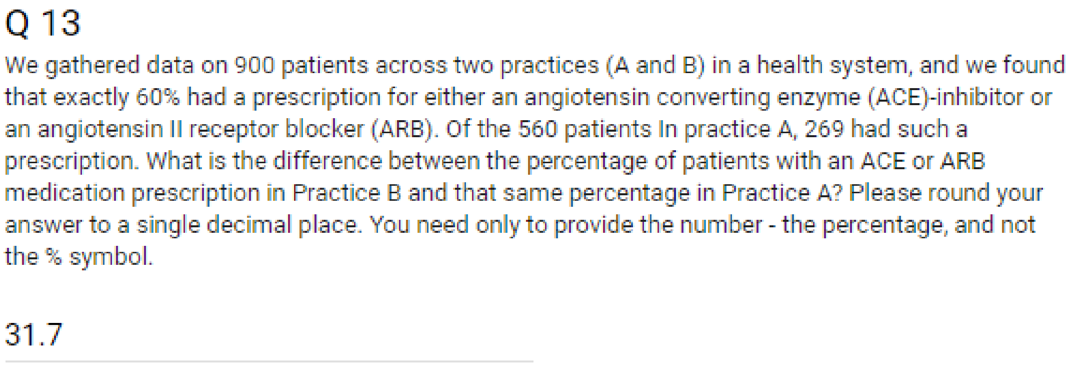
```
  n mean   sd median trimmed   mad min max range  skew kurtosis  se
250 65.1 17.4   67.5   65.85 18.53  20  98    78 -0.35    -0.64 1.1
```

- These data range from 20 to 98, so none of these patients are in the severely abnormal (defined as above 200) range, so a is true.
- The standard error is proportional to the square root of n, not to n, so statement b is false. The standard error is the standard deviation divided by the square root of the sample size. Right now, the standard error is $17.4/\sqrt{250} = 1.1$. Under these new circumstances, the standard error would be $17.4/\sqrt{500} = 0.78$, which isn't half of the original standard error, but instead about 3/4 of the size. So b is not true.
- The mean is less than the median, so the nonparametric skew would be negative. Specifically, skew1 = (65.1 - 67.5)/17.4 = -0.14. So c is true.

Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

Answer 13 is 31.7 for 2 points.



Q 13

We gathered data on 900 patients across two practices (A and B) in a health system, and we found that exactly 60% had a prescription for either an angiotensin converting enzyme (ACE)-inhibitor or an angiotensin II receptor blocker (ARB). Of the 560 patients In practice A, 269 had such a prescription. What is the difference between the percentage of patients with an ACE or ARB medication prescription in Practice B and that same percentage in Practice A? Please round your answer to a single decimal place. You need only to provide the number - the percentage, and not the % symbol.

31.7

The question helps us fill out the relevant 2x2 table. We start with. . .

Practice	ACE or ARB	No Prescription	Total
A	269		560
B			
Total	*900(0.6)* = 540		900

The complete 2x2 table, with margins, after some arithmetic, is:

Practice	ACE or ARB	No Prescription	Total
A	269	*560-269* = 291	560
B	*540-269* = 271	*360-291* = 69	*271 + 69* = 340
Total	540	*900-540* = 360	900

- The percentage in practice B with a prescription is 100(271/340) = 79.7.
- The percentage in practice A with a prescription is 100(269/560) = 48.0.
- So the difference is 79.7 - 48.0 = 31.7 percentage points.

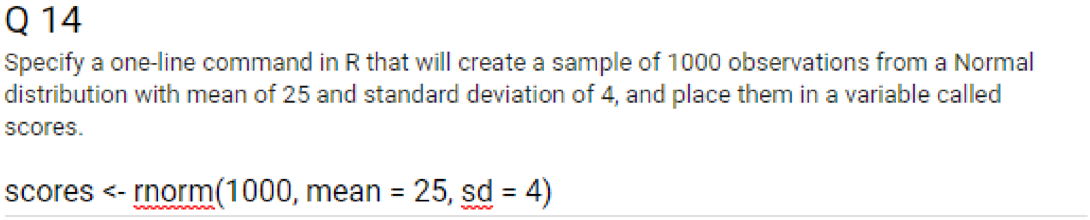
Partial Credit

- I gave 1 point for 31.6, 31.8 or 32.

This item didn't go well.

It required several pieces of a calculation. I didn't see obvious patterns in the wrong answers.

Answer 14 is a piece of R code, for 3 points.



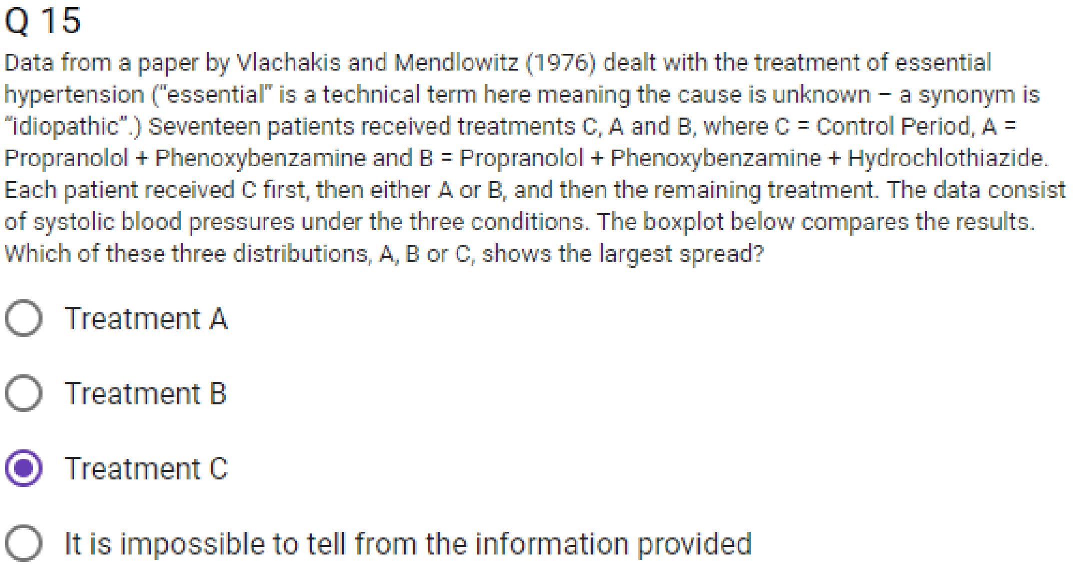
Q 14

Specify a one-line command in R that will create a sample of 1000 observations from a Normal distribution with mean of 25 and standard deviation of 4, and place them in a variable called scores.

```
scores <- rnorm(1000, mean = 25, sd = 4)
```

The correct response is anything equivalent to `scores <- rnorm(1000, mean = 25, sd = 4)`. Anything that would have accomplished all of the goals was acceptable. For instance, you might have embedded your `scores` variable inside a data frame, even though I didn't ask you to do that. If so, OK.

Answer 15 is c, for 2 points.



Q 15

Data from a paper by Vlachakis and Mendlowitz (1976) dealt with the treatment of essential hypertension ("essential" is a technical term here meaning the cause is unknown – a synonym is "idiopathic".) Seventeen patients received treatments C, A and B, where C = Control Period, A = Propranolol + Phenoxybenzamine and B = Propranolol + Phenoxybenzamine + Hydrochlothiazide. Each patient received C first, then either A or B, and then the remaining treatment. The data consist of systolic blood pressures under the three conditions. The boxplot below compares the results. Which of these three distributions, A, B or C, shows the largest spread?

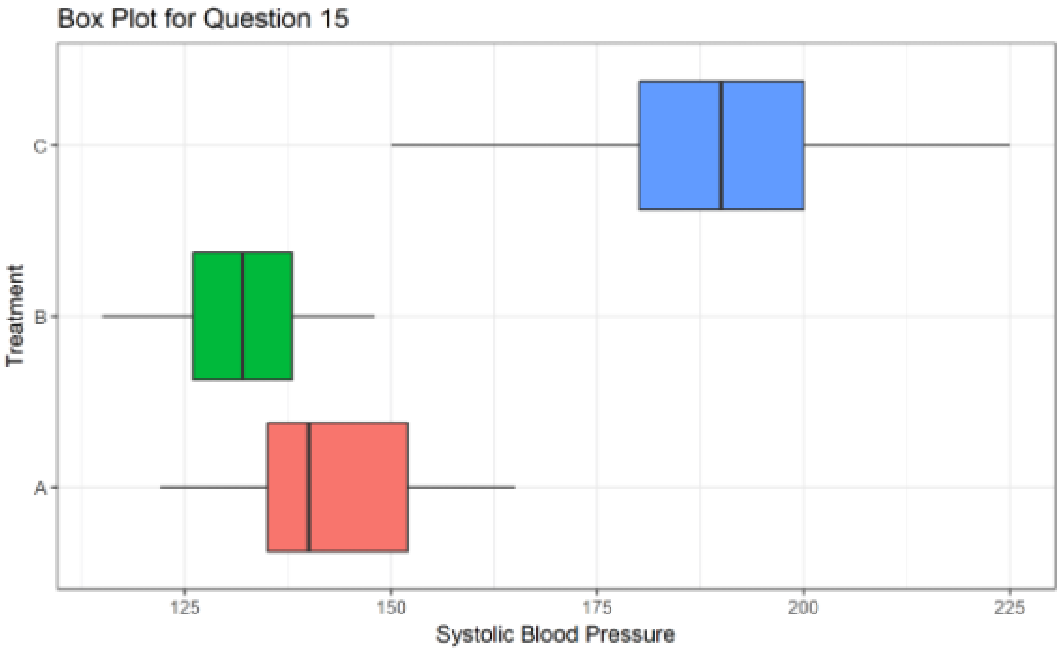
○ Treatment A

○ Treatment B

● Treatment C

○ It is impossible to tell from the information provided

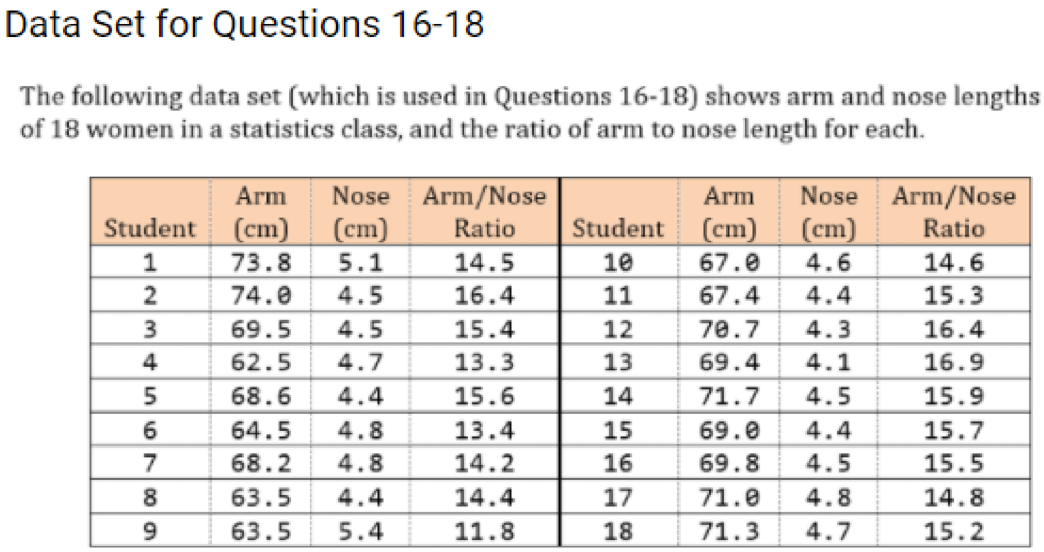
Figure for Question 15



This is evident from the whiskers, which show a much larger range in Treatment C results than in either of the other treatment groups.

Incidentally, there was a typo here. I was missing `ro` in hydrochlorothiazide, which is usually abbreviated HCTZ to avoid this sort of problem.

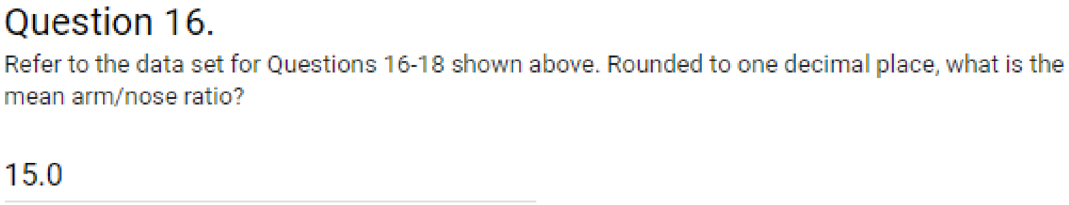
Answer 16 is 15.0, for 2 points.



Data Set for Questions 16-18

The following data set (which is used in Questions 16-18) shows arm and nose lengths of 18 women in a statistics class, and the ratio of arm to nose length for each.

Student	Arm (cm)	Nose (cm)	Arm/Nose Ratio	Student	Arm (cm)	Nose (cm)	Arm/Nose Ratio
1	73.8	5.1	14.5	10	67.0	4.6	14.6
2	74.0	4.5	16.4	11	67.4	4.4	15.3
3	69.5	4.5	15.4	12	70.7	4.3	16.4
4	62.5	4.7	13.3	13	69.4	4.1	16.9
5	68.6	4.4	15.6	14	71.7	4.5	15.9
6	64.5	4.8	13.4	15	69.0	4.4	15.7
7	68.2	4.8	14.2	16	69.8	4.5	15.5
8	63.5	4.4	14.4	17	71.0	4.8	14.8
9	63.5	5.4	11.8	18	71.3	4.7	15.2



Question 16.

Refer to the data set for Questions 16-18 shown above. Rounded to one decimal place, what is the mean arm/nose ratio?

15.0

```
data16 <- data_frame(Student = 1:18,
                     arm_nose = c(14.5, 16.4, 15.4, 13.3, 15.6, 13.4, 14.2, 14.4, 11.8,
                                  14.6, 15.3, 16.4, 16.9, 15.9, 15.7, 15.5, 14.8, 15.2))
summary(data16)
```

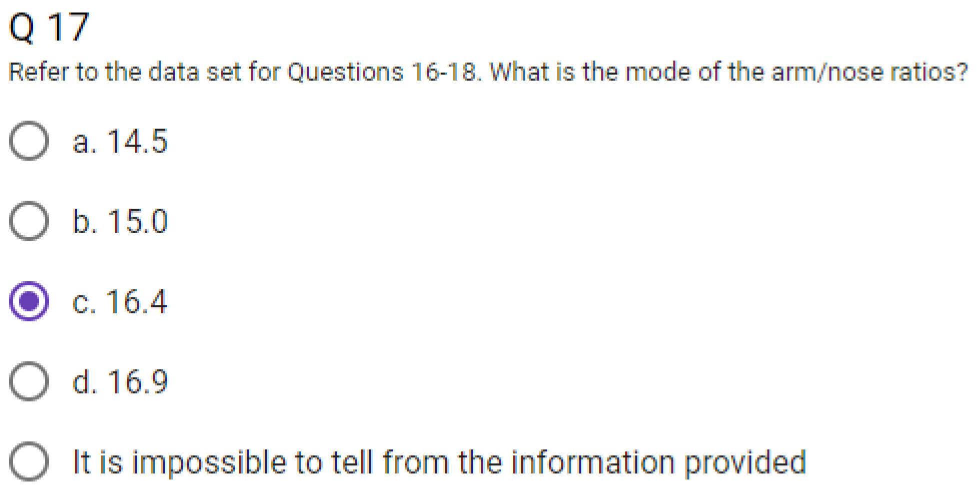
```
    Student         arm_nose    
 Min.   : 1.00   Min.   :11.80  
 1st Qu.: 5.25   1st Qu.:14.43  
 Median : 9.50   Median :15.25  
 Mean   : 9.50   Mean   :14.96  
 3rd Qu.:13.75   3rd Qu.:15.68  
 Max.   :18.00   Max.   :16.90  
```

So the mean is in fact 14.96, which rounds to 15.0. Or, you could just sum the 18 values, get 269.3, and divide for 14.96111, which still rounds to 15.0

Partial Credit

- I gave 1 point for 14.9 or for 14.96

Answer 17 is c, for 2 points.



Q 17

Refer to the data set for Questions 16-18. What is the mode of the arm/nose ratios?

○ a. 14.5

○ b. 15.0

● c. 16.4

○ d. 16.9

○ It is impossible to tell from the information provided

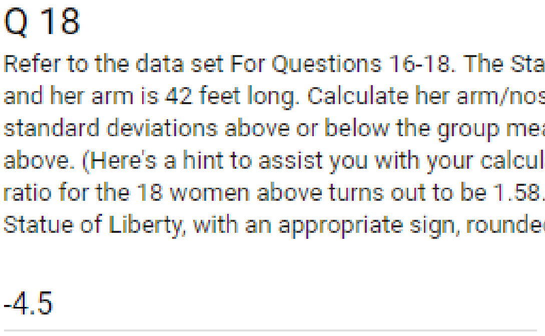
Here are the sorted data…

```
data16 %>% arrange(arm_nose) %>%
  knitr::kable()
```

Student	arm_nose
9	11.8
4	13.3
6	13.4
7	14.2
8	14.4
1	14.5
10	14.6
17	14.8
18	15.2
11	15.3
3	15.4
16	15.5
5	15.6
15	15.7
14	15.9
2	16.4
12	16.4
13	16.9

There are two values of 16.4 (students 2 and 12), and those are the only repeated values. Hence, 16.4 is the mode.

Answer 18 is -4.5, for 3 points.



Q 18

Refer to the data set For Questions 16-18. The Statue of Liberty's nose measures 4 feet, 6 inches, and her arm is 42 feet long. Calculate her arm/nose ratio, and use it to specify her Z score (# of standard deviations above or below the group mean) as compared to the 18 women described above. (Here's a hint to assist you with your calculations: The sample VARIANCE of the arm/nose ratio for the 18 women above turns out to be 1.58.) Your response should be the Z score for the Statue of Liberty, with an appropriate sign, rounded to one decimal place.

-4.5

To calculate the arm/nose ratio of the Statue of Liberty, we need to get her arm and nose lengths on the same scale. (Note that it doesn't have to be the same scale as was used for the women in the class, mathematically.) So, her arm length is 42 feet, and her nose length is 4.5 feet. Thus, the Statue of Liberty has arm/nose ratio of $42/(4.5) = 9.33$.

The mean of the 18 women is 14.96, and the standard deviation is the square root of the specified variance (1.58) and $\sqrt{1.58} = 1.26$.

Thus, the Z score for the statue is

$$(9.33 - 14.96)/1.26 = -4.47$$

And so Z = -4.5 after rounding to one decimal place. Note that if you used 15 for the mean instead of 14.96, you'd get Z = -4.5, also.

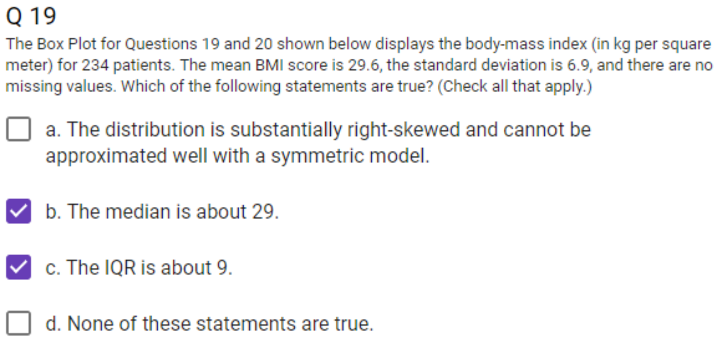
Partial Credit

- I gave 2 points for -4.47 (failure to round)
- I gave 1 point for -4.4 or -5 (incorrect or too much rounding)

This item didn't go well.

It required several pieces of a calculation. I didn't see obvious patterns in the wrong answers.

Answer 19 is b and c, for 3 points.



Q 19

The Box Plot for Questions 19 and 20 shown below displays the body-mass index (in kg per square meter) for 234 patients. The mean BMI score is 29.6, the standard deviation is 6.9, and there are no missing values. Which of the following statements are true? (Check all that apply.)

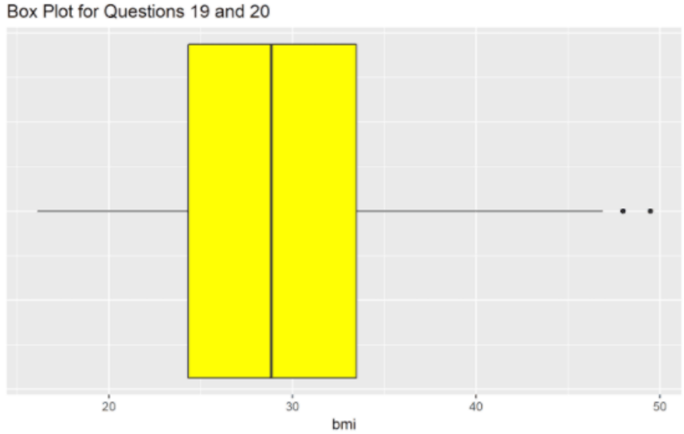
□ a. The distribution is substantially right-skewed and cannot be approximated well with a symmetric model.

☑ b. The median is about 29.

☑ c. The IQR is about 9.

□ d. None of these statements are true.

Figure for Questions 19 and 20



- The median, as indicated by the center line in the box, is clearly near 29, so b is true.
- The data cannot be severely right skewed, if the mean (29.64) is close to the median (29), and that difference in this case is less than a tenth of a standard deviation, so a is false.
- The IQR is the difference between the 25th and 75th percentiles, as shown by the edges of the box. Clearly the difference between them is near 9. In fact, the IQR is actually 33.48 - 24.32 = 9.16. So c is true.

Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

Answer 20 is b, for 2.5 points.

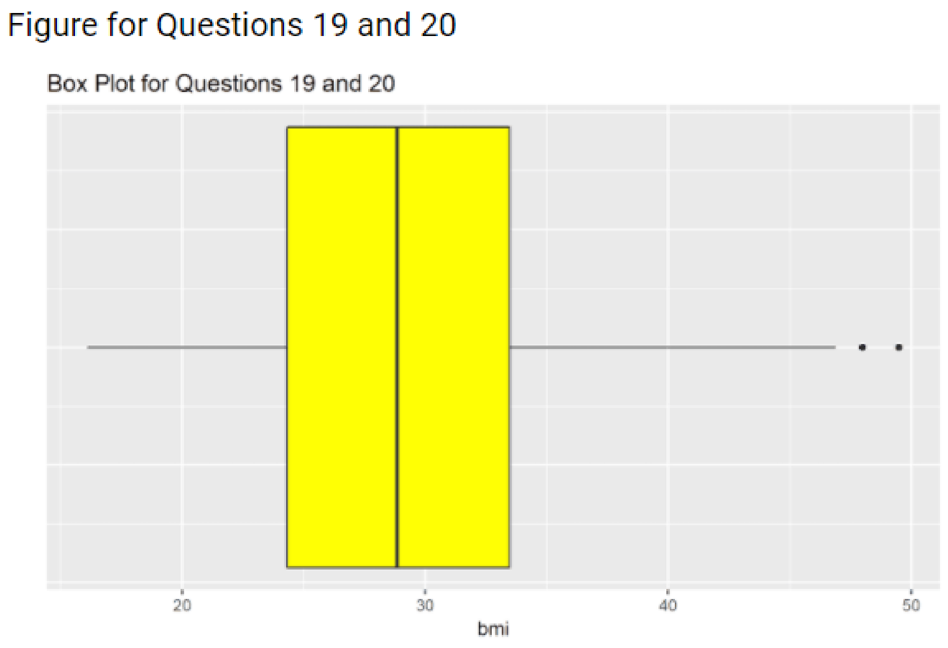
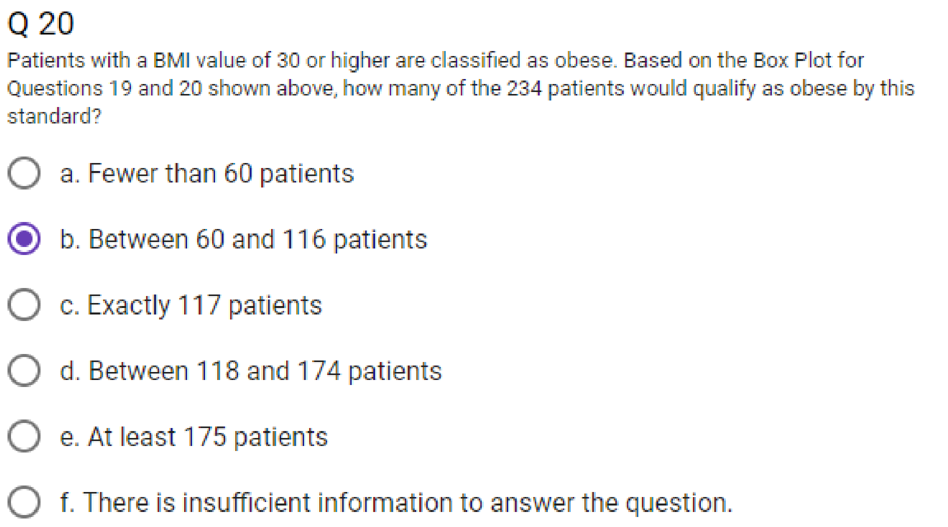


Figure for Questions 19 and 20

Box Plot for Questions 19 and 20



Q 20

Patients with a BMI value of 30 or higher are classified as obese. Based on the Box Plot for Questions 19 and 20 shown above, how many of the 234 patients would qualify as obese by this standard?

○ a. Fewer than 60 patients

● b. Between 60 and 116 patients

○ c. Exactly 117 patients

○ d. Between 118 and 174 patients

○ e. At least 175 patients

○ f. There is insufficient information to answer the question.

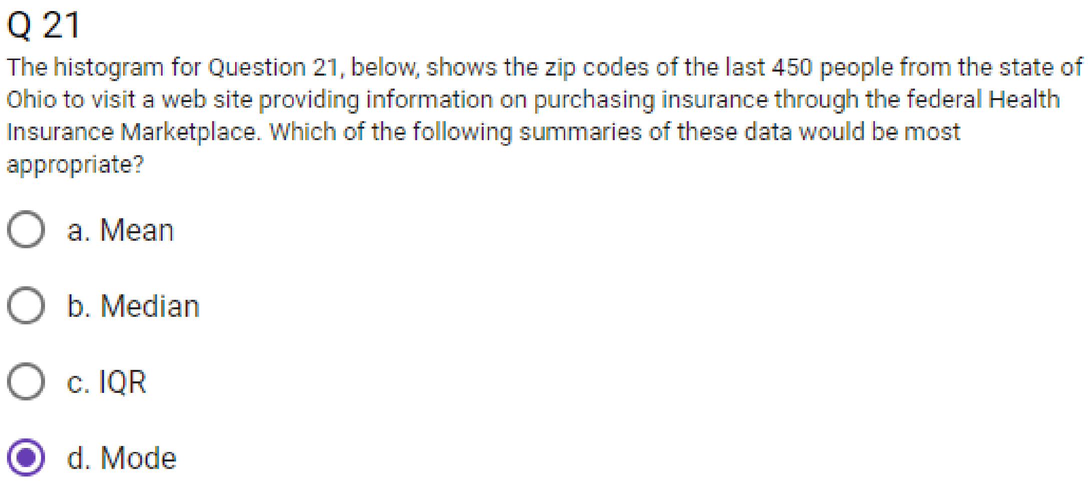
Clearly 30 is above the median (so less than 50% of the 234 patients can have BMI 30 or more) but a good deal less than the 75th percentile (so at least 25% of the 234 patients have BMI 30 or more).

- Since 50% of 234 is 117 and 25% of 234 is 58.5, it looks like the right answer is between those values. That's response b.

This item didn't go well.

`d` and `f` were the more common incorrect responses.

Answer 21 is d, for 2.5 points.



Q 21

The histogram for Question 21, below, shows the zip codes of the last 450 people from the state of Ohio to visit a web site providing information on purchasing insurance through the federal Health Insurance Marketplace. Which of the following summaries of these data would be most appropriate?

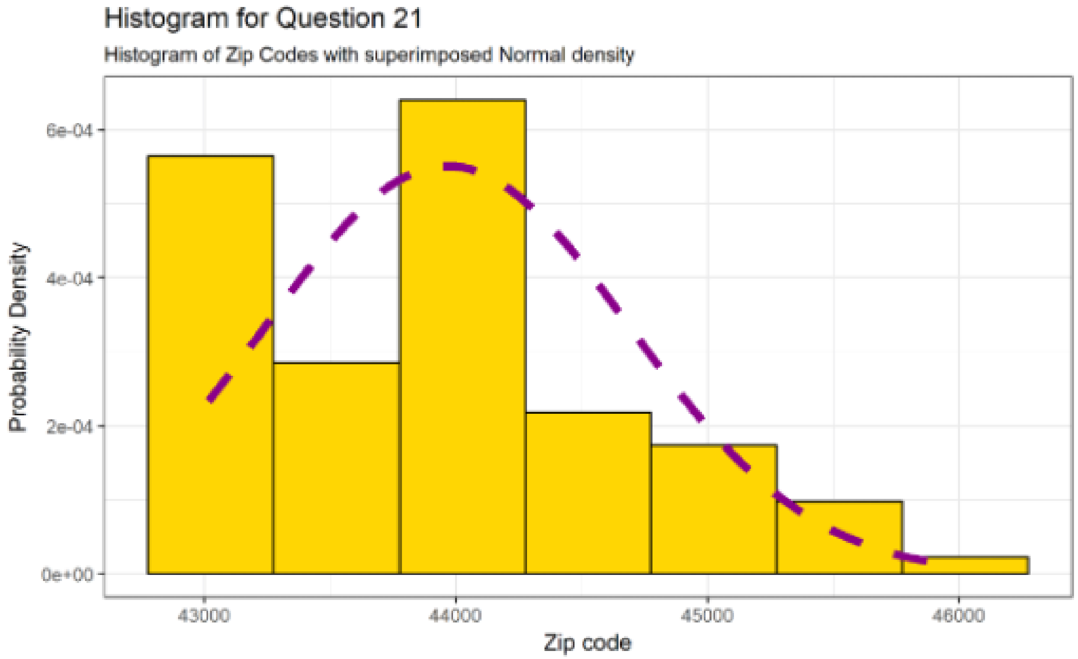
○ a. Mean

○ b. Median

○ c. IQR

● d. Mode

Figure for Question 21

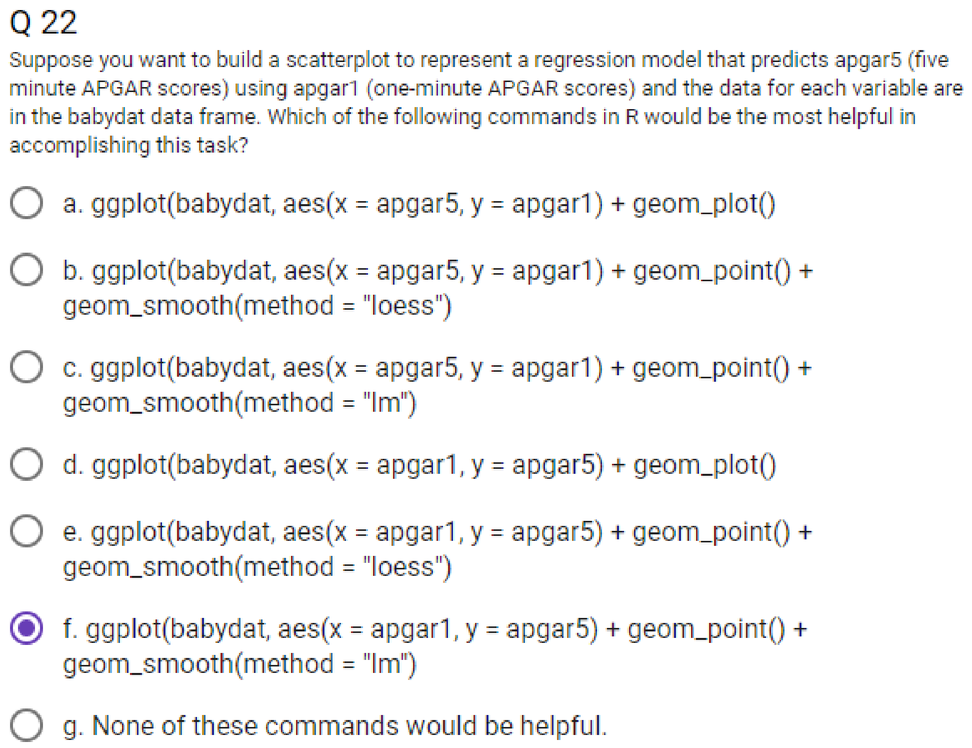


Zip codes are numbers, but they're not quantitative. Instead, they are nominal categorical data. Of these four choices, only a mode could possibly be relevant.

This item didn't go well.

And I wasn't surprised. It's a bit tricky.

Answer 22 was meant to be f but was really g, for 2 points.



Q 22

Suppose you want to build a scatterplot to represent a regression model that predicts apgar5 (five minute APGAR scores) using apgar1 (one-minute APGAR scores) and the data for each variable are in the babydat data frame. Which of the following commands in R would be the most helpful in accomplishing this task?

○ a. ggplot(babydat, aes(x = apgar5, y = apgar1) + geom_plot()

○ b. ggplot(babydat, aes(x = apgar5, y = apgar1) + geom_point() + geom_smooth(method = "loess")

○ c. ggplot(babydat, aes(x = apgar5, y = apgar1) + geom_point() + geom_smooth(method = "lm")

○ d. ggplot(babydat, aes(x = apgar1, y = apgar5) + geom_plot()

○ e. ggplot(babydat, aes(x = apgar1, y = apgar5) + geom_point() + geom_smooth(method = "loess")

● f. ggplot(babydat, aes(x = apgar1, y = apgar5) + geom_point() + geom_smooth(method = "lm")

○ g. None of these commands would be helpful.

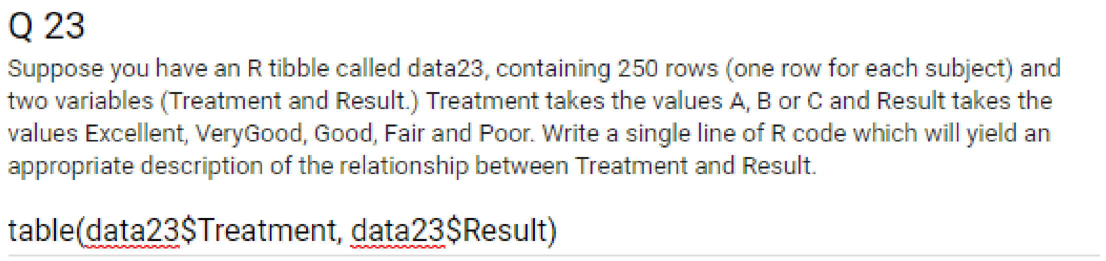
- To obtain a scatterplot, you use `geom_point`, not `geom_plot` (`geom_plot` isn't a thing) so we can eliminate a and d as options.
- The five-minute APGAR score needs to go on the vertical (y) axis because it is the outcome, and the one-minute APGAR score needs to go on the horizontal (x) axis because it is the predictor. So we need x = apgar1, and y = apgar5, so that eliminates b and c as options.
- Choice e shows a loess smooth, rather than the linear model we are trying to represent, so that eliminates e.
- That leaves f, which shows the linear model we are trying to represent. So that's the most helpful choice.

BUT a very observant student noted that I'd somehow left off the closing) in the `ggplot` part of each option. Thus, the truly correct answer is **g**. As a result, I accepted **either f or g** for full credit.

This item was right on the borderline of not going well.

`c` and `e` were the most common wrong answers.

Answer 23 is a piece of R code, for 3 points.



Q 23

Suppose you have an R tibble called data23, containing 250 rows (one row for each subject) and two variables (Treatment and Result.) Treatment takes the values A, B or C and Result takes the values Excellent, VeryGood, Good, Fair and Poor. Write a single line of R code which will yield an appropriate description of the relationship between Treatment and Result.

```
table(data23$Treatment, data23$Result)
```

We're looking at the association of two categorical variables. The Treatment is nominal and the Result is ordinal. We need a table to compare the results by treatment group.

I'd accept any response that generates a cross-tabulation of frequencies or a table of percentages or proportions comparing Results by Treatment. That includes things like:

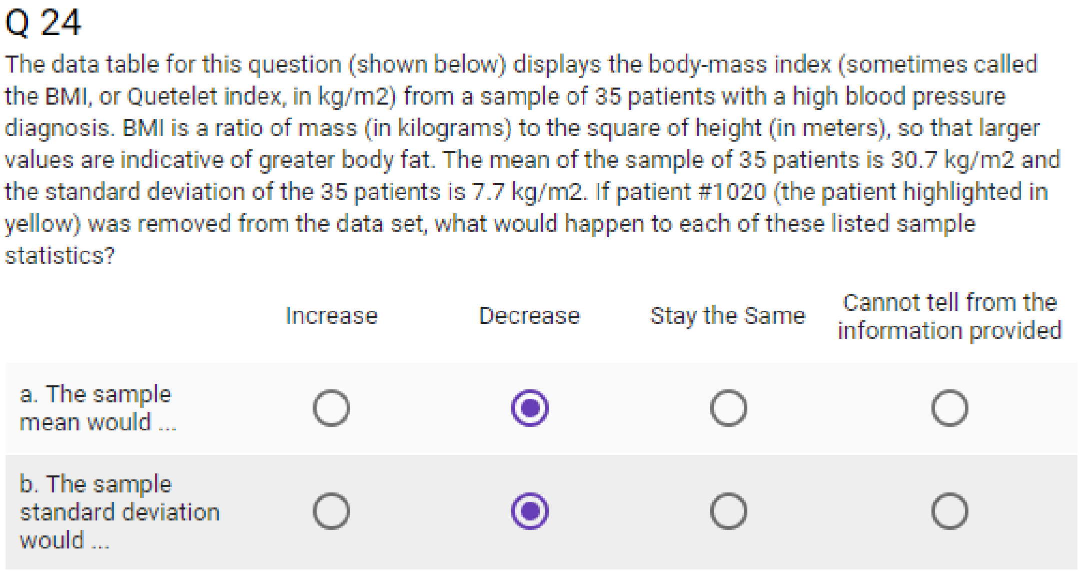
- `data23 %>% select(Treatment, Result) %>% table()`
- `addmargins(table(data23$Treatment ,  data23$Result))`
- `xtabs(~ Treatment + Result, data = data23)`

This item didn't go well.

Things that won't work and thus earned 0 points:

- A boxplot or other graphical tool isn't appropriate here, because it wouldn't show all of the data, would require assigning numbers to the Results categories, and thus assume that Result can be treated as if it's a linear scale, where Excellent is as far away from Good, for instance, as Good is from Poor.
- Using `favstats` or `summary`, even in a `by` statement, or `group_by` and `summarize` approach will generate means and standard deviations, but not counts or percentages.
- Using a chi-square test on the data without fitting a table won't give us the frequencies, proportions or percentages.
- Using a linear model is no help.
- Using ggplot could possibly be helpful with, for instance, a bar chart, but everyone who tried it made mistakes, so their code would not run, or would not produce anything useful.

Answer 24 is that a and b would decrease, for 2 points.



Q 24

The data table for this question (shown below) displays the body-mass index (sometimes called the BMI, or Quetelet index, in kg/m2) from a sample of 35 patients with a high blood pressure diagnosis. BMI is a ratio of mass (in kilograms) to the square of height (in meters), so that larger values are indicative of greater body fat. The mean of the sample of 35 patients is 30.7 kg/m2 and the standard deviation of the 35 patients is 7.7 kg/m2. If patient #1020 (the patient highlighted in yellow) was removed from the data set, what would happen to each of these listed sample statistics?

	Increase	Decrease	Stay the Same	Cannot tell from the information provided
a. The sample mean would ...	○	●	○	○
b. The sample standard deviation would ...	○	●	○	○

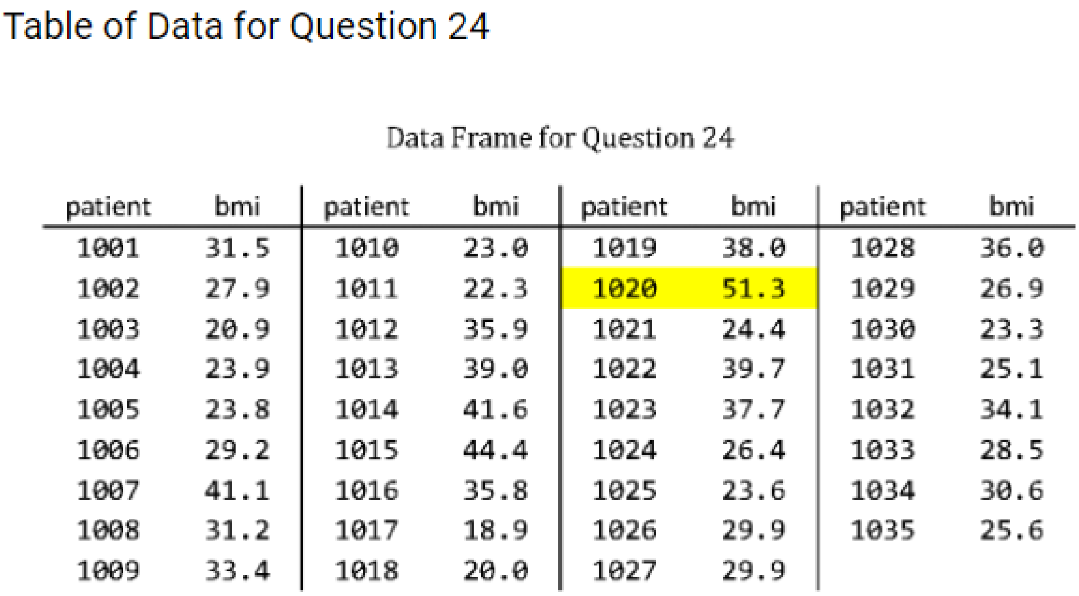


Table of Data for Question 24

Data Frame for Question 24

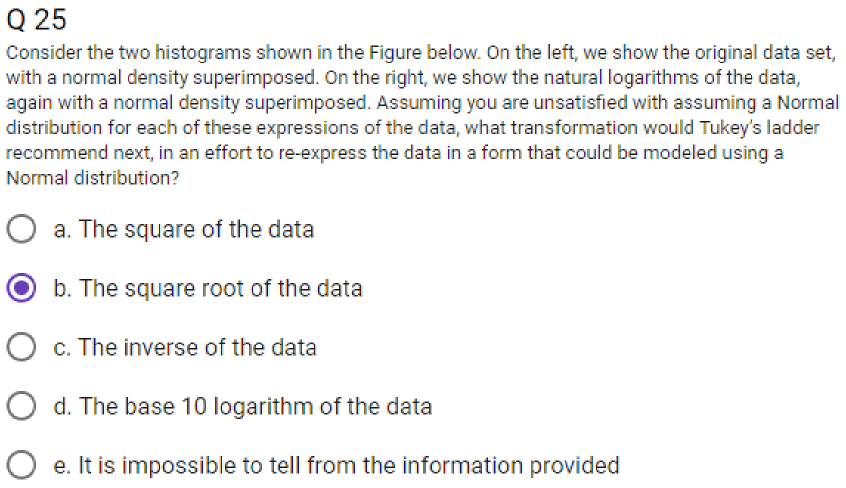
patient	bmi	patient	bmi	patient	bmi	patient	bmi
1001	31.5	1010	23.0	1019	38.0	1028	36.0
1002	27.9	1011	22.3	1020	51.3	1029	26.9
1003	20.9	1012	35.9	1021	24.4	1030	23.3
1004	23.9	1013	39.0	1022	39.7	1031	25.1
1005	23.8	1014	41.6	1023	37.7	1032	34.1
1006	29.2	1015	44.4	1024	26.4	1033	28.5
1007	41.1	1016	35.8	1025	23.6	1034	30.6
1008	31.2	1017	18.9	1026	29.9	1035	25.6
1009	33.4	1018	20.0	1027	29.9		

Patient 1020 has the maximum value in the data set. Dropping this patient will decrease both the mean and the standard deviation. The original mean was 30.7, and the original standard deviation was 7.7. These would drop to 30.1 and 6.9, respectively.

Partial Credit

1 point for a correct response to each part (a and b)

Answer 25 is b, for 2 points.



Q 25

Consider the two histograms shown in the Figure below. On the left, we show the original data set, with a normal density superimposed. On the right, we show the natural logarithms of the data, again with a normal density superimposed. Assuming you are unsatisfied with assuming a Normal distribution for each of these expressions of the data, what transformation would Tukey's ladder recommend next, in an effort to re-express the data in a form that could be modeled using a Normal distribution?

○ a. The square of the data

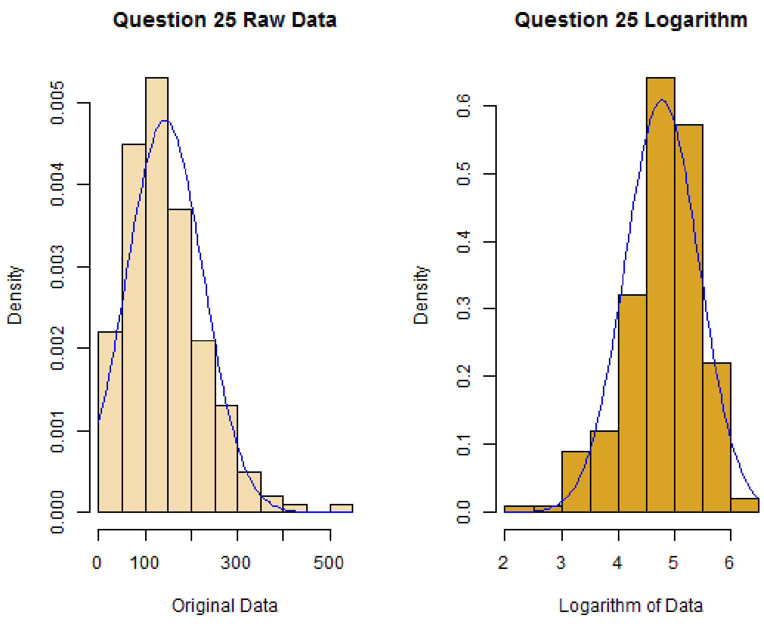
● b. The square root of the data

○ c. The inverse of the data

○ d. The base 10 logarithm of the data

○ e. It is impossible to tell from the information provided

Figure for Question 25

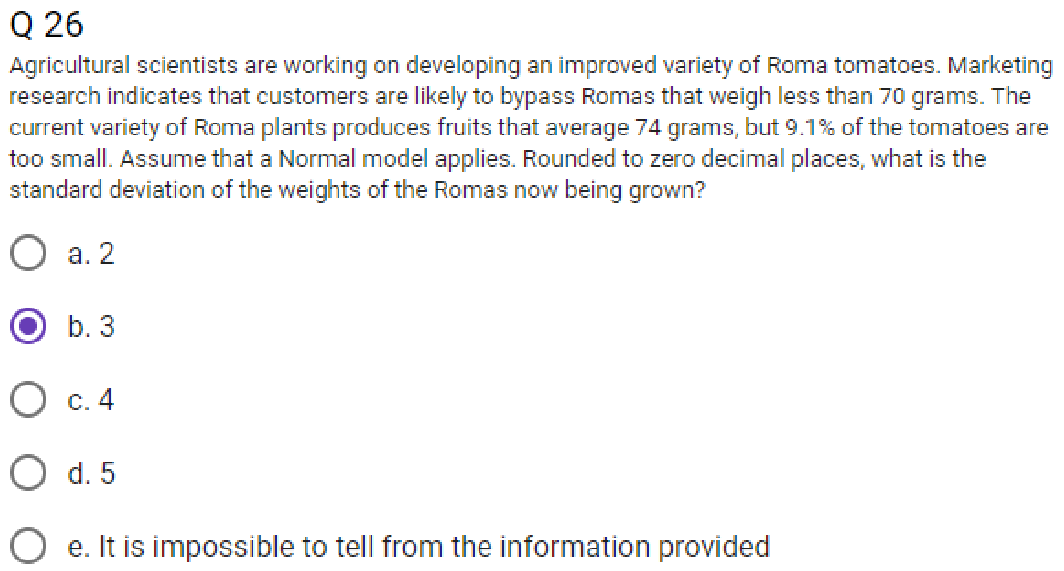


Since the raw data are right skewed, and the logged data are left skewed, something in between seems the best choice. On the ladder of power transformations, the square root (transformation using power p = 0.5) falls between the raw data (p = 1) and the log (p = 0).

This item didn't go well.

c was the most common incorrect response.

Answer 26 is b, for 2 points.



Q 26

Agricultural scientists are working on developing an improved variety of Roma tomatoes. Marketing research indicates that customers are likely to bypass Romas that weigh less than 70 grams. The current variety of Roma plants produces fruits that average 74 grams, but 9.1% of the tomatoes are too small. Assume that a Normal model applies. Rounded to zero decimal places, what is the standard deviation of the weights of the Romas now being grown?

○ a. 2

● b. 3

○ c. 4

○ d. 5

○ e. It is impossible to tell from the information provided

The easiest way to figure this out, I suppose, is to try the calculation with the available choices of standard deviation. We can use a Normal model with each of the available standard deviation choices in turn to see which one matches the probabilities as stated. The mean is known to be 74, we want to know what the standard deviation needs to be to reject 9.1% of tomatoes as being less than 70 grams.

From the output below, we see that with a mean of 74 and standard deviation 2, a rejection rule of "70 grams or less" will happen to 2.3% of tomatoes.

With a sd of 3 we hit the magic number of 9.1% for this probability.

```r
pnorm(70, mean=74, sd=2, lower.tail=TRUE)
```

```
[1] 0.02275013
```

```r
pnorm(70, mean=74, sd=3, lower.tail=TRUE)
```

```
[1] 0.09121122
```

```r
pnorm(70, mean=74, sd=4, lower.tail=TRUE)
```

```
[1] 0.1586553
```

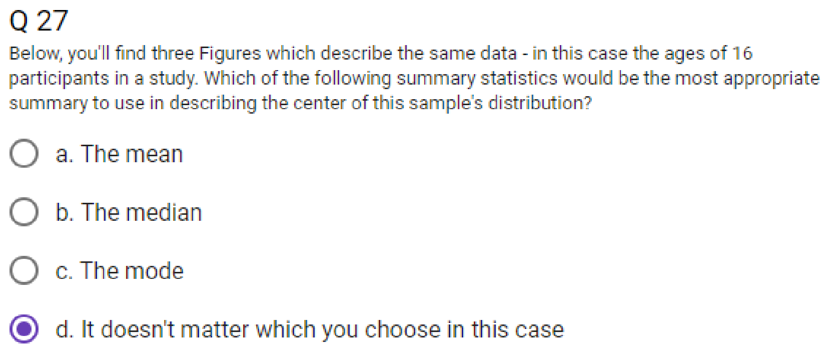
```r
pnorm(70, mean=74, sd=5, lower.tail=TRUE)
```

```
[1] 0.2118554
```

This item didn't go well.

`a` and `e` were the most common incorrect responses.

Answer 27 is d, for 2 points.



Q 27

Below, you'll find three Figures which describe the same data - in this case the ages of 16 participants in a study. Which of the following summary statistics would be the most appropriate summary to use in describing the center of this sample's distribution?

○ a. The mean

○ b. The median

○ c. The mode

● d. It doesn't matter which you choose in this case

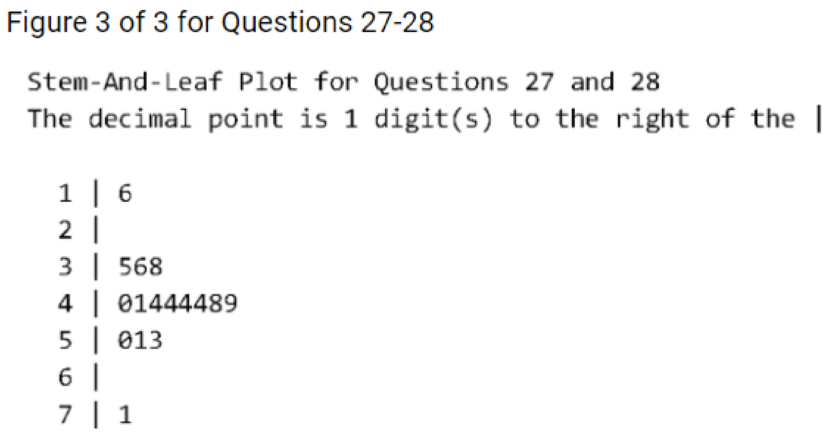


Figure 3 of 3 for Questions 27-28

```
Stem-And-Leaf Plot for Questions 27 and 28
The decimal point is 1 digit(s) to the right of the |

  1 | 6
  2 |
  3 | 568
  4 | 01444489
  5 | 013
  6 |
  7 | 1
```

I haven't shown Figures 1 and 2 here because the stem-and-leaf is sufficient to answer the question.

It doesn't matter which you choose here, because the mean, the median and the mode are all the same value (44).

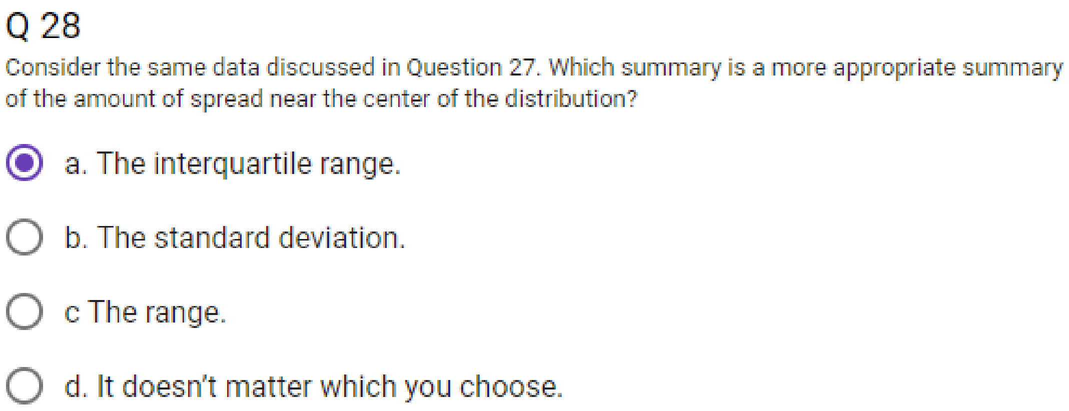
```
tempdat <- c(16, 35, 36, 38, 40, 41, 44, 44, 44, 44, 48, 49, 50, 51, 53, 71)
summary(tempdat)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  16.00   39.50   44.00   44.00   49.25   71.00 
```

This item didn't go well.

And I wasn't surprised. It was tricky. `a` and `b` were the most common incorrect responses.

Answer 28 is a, for 2 points.



Q 28

Consider the same data discussed in Question 27. Which summary is a more appropriate summary of the amount of spread near the center of the distribution?

● a. The interquartile range.

○ b. The standard deviation.

○ c The range.

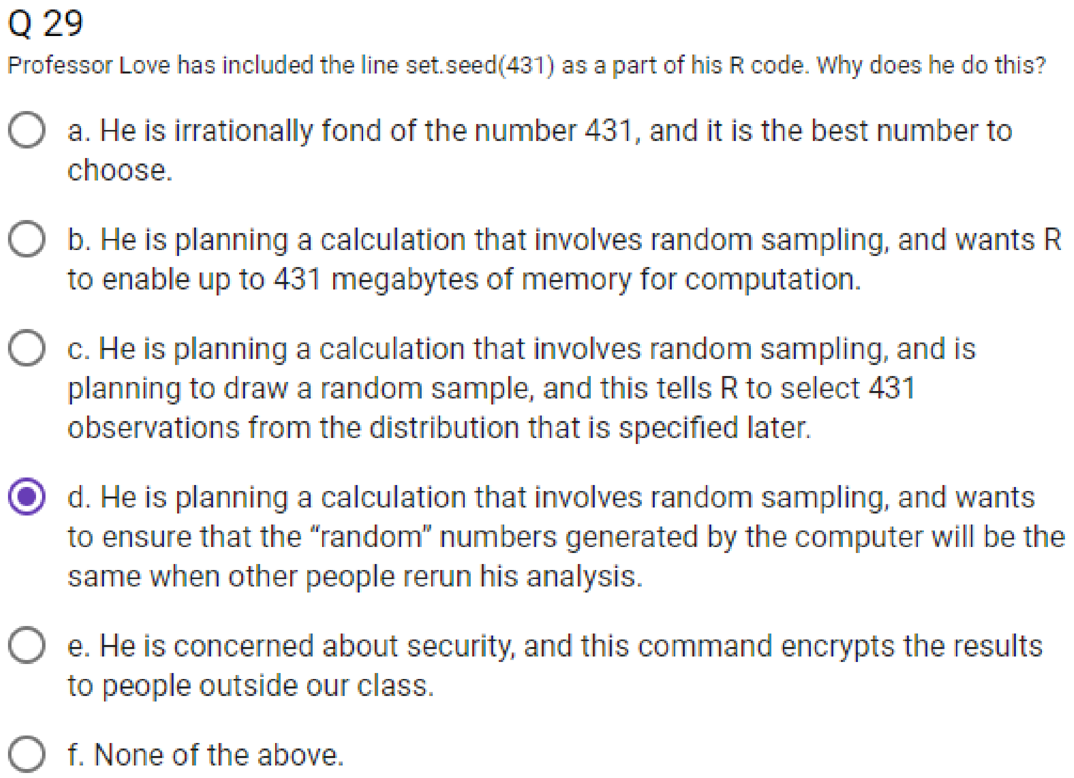
○ d. It doesn't matter which you choose.

The data set is affected by the outlier values at 16 and 71. So an interquartile range (which is essentially unaffected by these outliers) is a better choice than either the standard deviation or the range.

This item didn't go well.

`b` was the most common incorrect response.

Answer 29 is d, for 2 points.



Q 29

Professor Love has included the line set.seed(431) as a part of his R code. Why does he do this?

○ a. He is irrationally fond of the number 431, and it is the best number to choose.

○ b. He is planning a calculation that involves random sampling, and wants R to enable up to 431 megabytes of memory for computation.

○ c. He is planning a calculation that involves random sampling, and is planning to draw a random sample, and this tells R to select 431 observations from the distribution that is specified later.

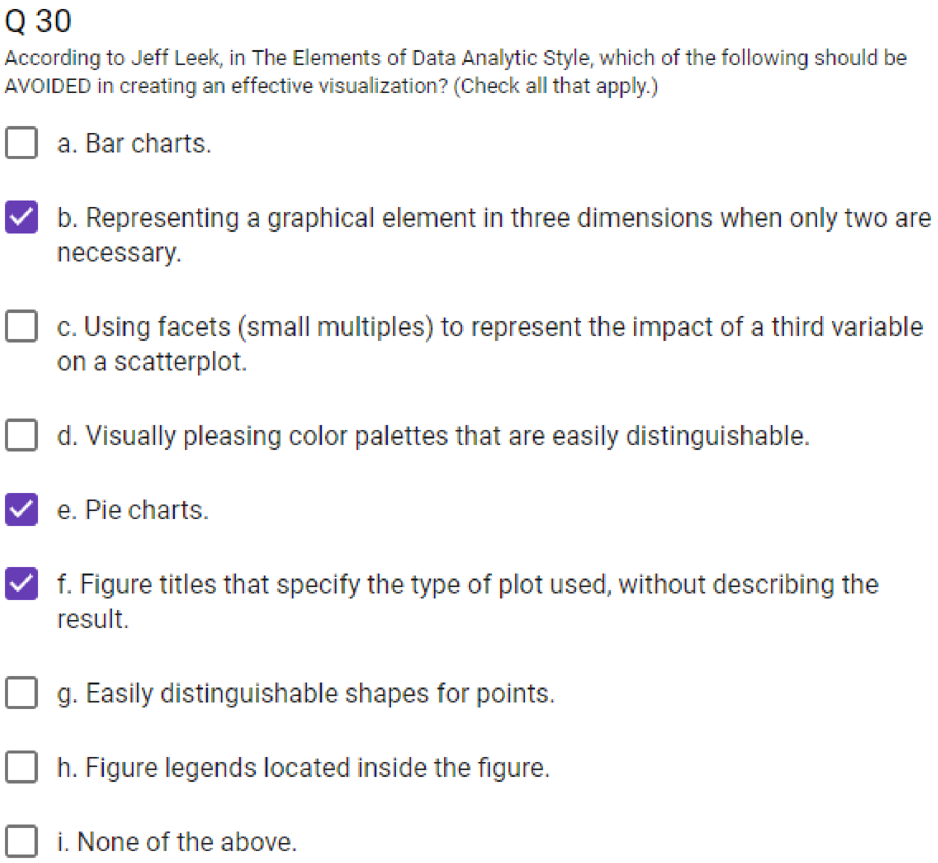
● d. He is planning a calculation that involves random sampling, and wants to ensure that the “random” numbers generated by the computer will be the same when other people rerun his analysis.

○ e. He is concerned about security, and this command encrypts the results to people outside our class.

○ f. None of the above.

Choice d describes what setting a random seed does. For more on this, take a look at Leek’s *The Elements of Data Analytic Style*, section 12.5. Dr. Love has no particular attachment to the number 431. His favorite number is actually 41, but that’s because he is a nearly lifelong fan of the New York Mets.

Answer 30 is b, e and f, for 3 points.



Q 30

According to Jeff Leek, in The Elements of Data Analytic Style, which of the following should be AVOIDED in creating an effective visualization? (Check all that apply.)

☐ a. Bar charts.

☑ b. Representing a graphical element in three dimensions when only two are necessary.

☐ c. Using facets (small multiples) to represent the impact of a third variable on a scatterplot.

☐ d. Visually pleasing color palettes that are easily distinguishable.

☑ e. Pie charts.

☑ f. Figure titles that specify the type of plot used, without describing the result.

☐ g. Easily distinguishable shapes for points.

☐ h. Figure legends located inside the figure.

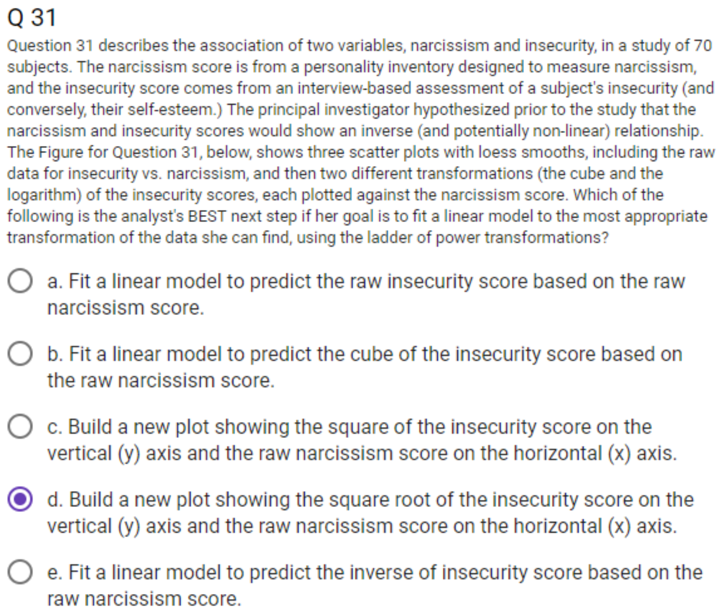
☐ i. None of the above.

See Chapter 11 of Leek's *The Elements of Data Analytic Style*.

Partial Credit

- gain 1 point for selecting each of the three correct responses
- lose 1 point for each other option selected
- minimum score on the item = 0

Answer 31 is d, for 2 points.



Q 31

Question 31 describes the association of two variables, narcissism and insecurity, in a study of 70 subjects. The narcissism score is from a personality inventory designed to measure narcissism, and the insecurity score comes from an interview-based assessment of a subject's insecurity (and conversely, their self-esteem.) The principal investigator hypothesized prior to the study that the narcissism and insecurity scores would show an inverse (and potentially non-linear) relationship. The Figure for Question 31, below, shows three scatter plots with loess smooths, including the raw data for insecurity vs. narcissism, and then two different transformations (the cube and the logarithm) of the insecurity scores, each plotted against the narcissism score. Which of the following is the analyst's BEST next step if her goal is to fit a linear model to the most appropriate transformation of the data she can find, using the ladder of power transformations?

○ a. Fit a linear model to predict the raw insecurity score based on the raw narcissism score.

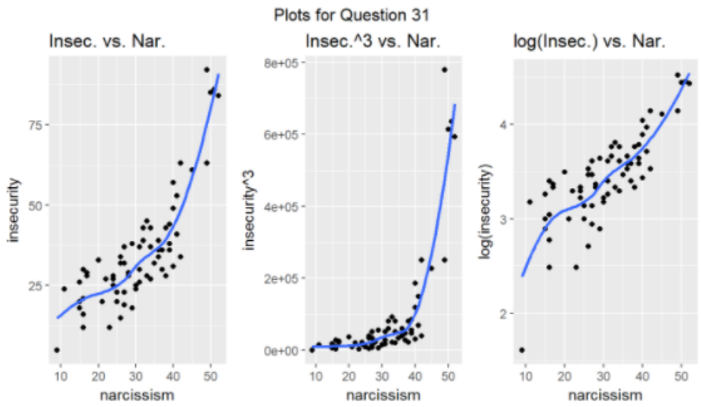
○ b. Fit a linear model to predict the cube of the insecurity score based on the raw narcissism score.

○ c. Build a new plot showing the square of the insecurity score on the vertical (y) axis and the raw narcissism score on the horizontal (x) axis.

● d. Build a new plot showing the square root of the insecurity score on the vertical (y) axis and the raw narcissism score on the horizontal (x) axis.

○ e. Fit a linear model to predict the inverse of insecurity score based on the raw narcissism score.

Figures for Question 31

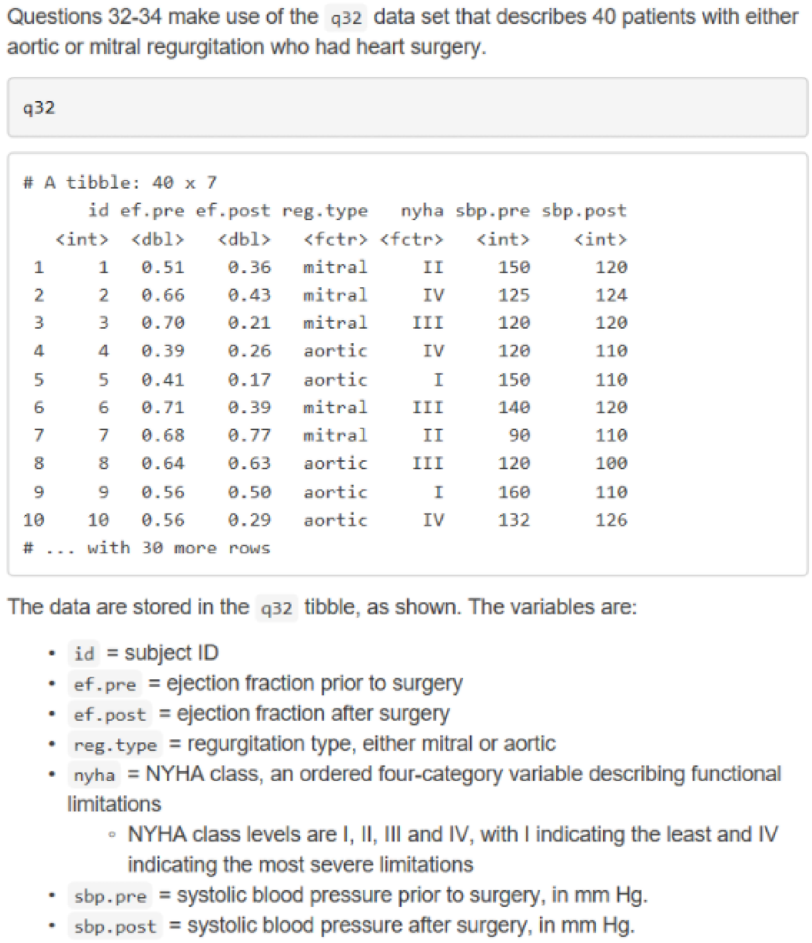


- The raw data shows a rather pronounced curve that will not be picked up well by a straight line model, so choice a isn't the way to go.
- Taking the cube of insecurity is clearly making the curve more pronounced, not less, so choice b isn't going to work, and also since this direction on the ladder moved us toward a more pronounced curve, choice c (fitting the square) is also not productive.
- The logarithm is a more promising transformation, so any analyst trying to get the best possible transformation would like to consider the options around the logarithm as well, which include the square root (and, perhaps, the inverse, too.) You wouldn't want to fit the log model until you looked at the square root and the inverse to decide if the log was the best choice.
- Choice e is inferior to choice d, but is a better choice than a. b or c. So I gave half credit.

Partial Credit

- 2 points for choice d, 1 point for choice e

Answer 32 is a line of R code, for 3 points.



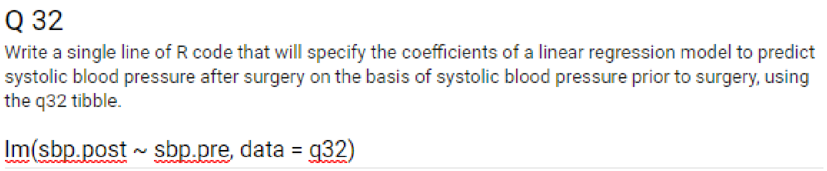
Questions 32-34 make use of the `q32` data set that describes 40 patients with either aortic or mitral regurgitation who had heart surgery.

```
q32
```

```
# A tibble: 40 x 7
      id ef.pre ef.post reg.type   nyha sbp.pre sbp.post
   <int>  <dbl>   <dbl>   <fctr> <fctr>   <int>    <int>
 1     1   0.51    0.36   mitral     II     150      120
 2     2   0.66    0.43   mitral     IV     125      124
 3     3   0.70    0.21   mitral    III     120      120
 4     4   0.39    0.26   aortic     IV     120      110
 5     5   0.41    0.17   aortic      I     150      110
 6     6   0.71    0.39   mitral    III     140      120
 7     7   0.68    0.77   mitral     II      90      110
 8     8   0.64    0.63   aortic    III     120      100
 9     9   0.56    0.50   aortic      I     160      110
10    10   0.56    0.29   aortic     IV     132      126
# ... with 30 more rows
```

The data are stored in the `q32` tibble, as shown. The variables are:

- `id` = subject ID
- `ef.pre` = ejection fraction prior to surgery
- `ef.post` = ejection fraction after surgery
- `reg.type` = regurgitation type, either mitral or aortic
- `nyha` = NYHA class, an ordered four-category variable describing functional limitations
    - NYHA class levels are I, II, III and IV, with I indicating the least and IV indicating the most severe limitations
- `sbp.pre` = systolic blood pressure prior to surgery, in mm Hg.
- `sbp.post` = systolic blood pressure after surgery, in mm Hg.



Q 32

Write a single line of R code that will specify the coefficients of a linear regression model to predict systolic blood pressure after surgery on the basis of systolic blood pressure prior to surgery, using the q32 tibble.

lm(sbp.post ~ sbp.pre, data = q32)

Any code that would produce the estimated slope and intercept coefficients for the correct model is OK. Other options include `lm(q32$sbp.post ~ q32$sbp.pre)`. `coef(lm(q32$sbp.post ~ q32$sbp.pre))`, and `summary(lm(sbp.post ~ sbp.pre, data = q32))`.

This item didn't go well.

Some people obviously read "the coefficients of a linear regression model" to be the Pearson correlation coefficient. No. The coefficients I'm referring to are the estimated slope and intercept.

Others misspelled `sbp.pre` or `sbp.post`.

Answer 33 is a line of R code, for 3 points.

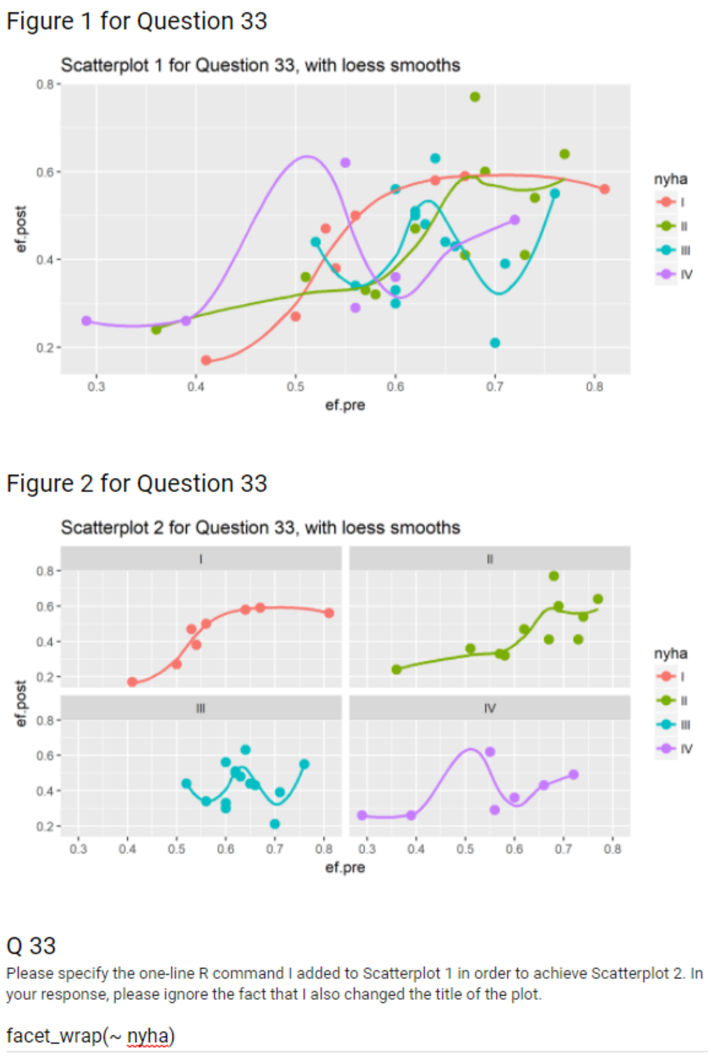


Figure 1 for Question 33

Figure 2 for Question 33

Q 33

Please specify the one-line R command I added to Scatterplot 1 in order to achieve Scatterplot 2. In your response, please ignore the fact that I also changed the title of the plot.

facet_wrap(~ nyha)

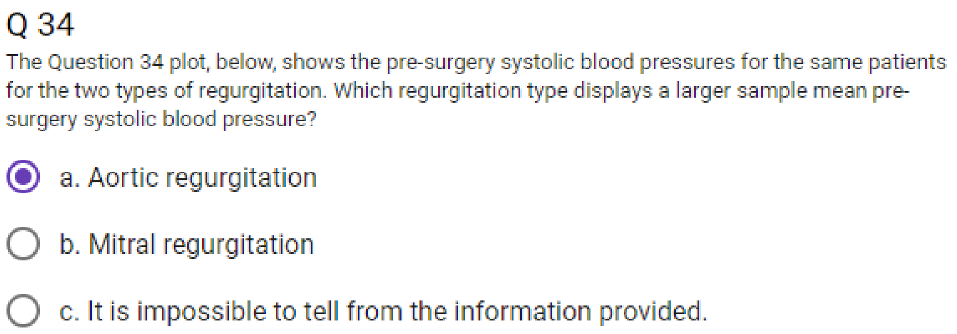
I gave full credit, regardless of whether you included or excluded the + sign, so that either `facet_wrap(~ nyha) +` or `facet_wrap(~ nyha)` would be correct.

Someone used `facet_wrap(~ q32$nyha)` which wouldn't actually work. `ggplot2` specifies the data set elsewhere (in the `aes` section.)

Partial Credit

- It's `nyha` and not `NYHA` and to R, that matters, so if you did that, you lost a point.
- Leaving out the tilde (~) cost you a point.
- Including the entire ggplot command lost you a point.
- You could have also used `facet_grid` to get facets, but that would have produced a different result. Nonetheless, if you used `facet_grid(~ nyha)` or `facet_grid(nyha ~ .)`, I gave full credit.

Answer 34 is a, for 2 points.



Q 34

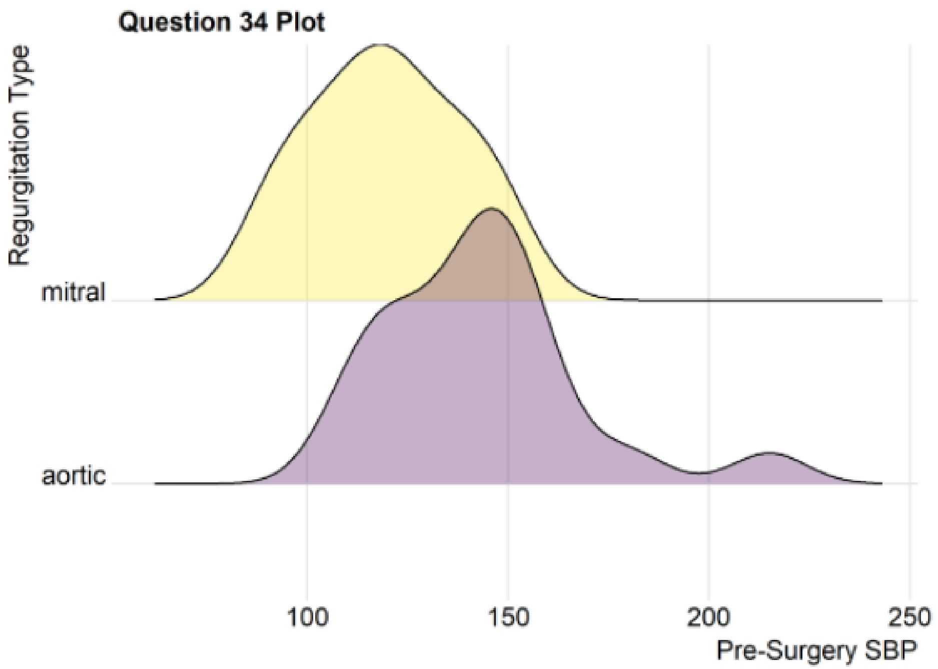
The Question 34 plot, below, shows the pre-surgery systolic blood pressures for the same patients for the two types of regurgitation. Which regurgitation type displays a larger sample mean pre-surgery systolic blood pressure?

● a. Aortic regurgitation

○ b. Mitral regurgitation

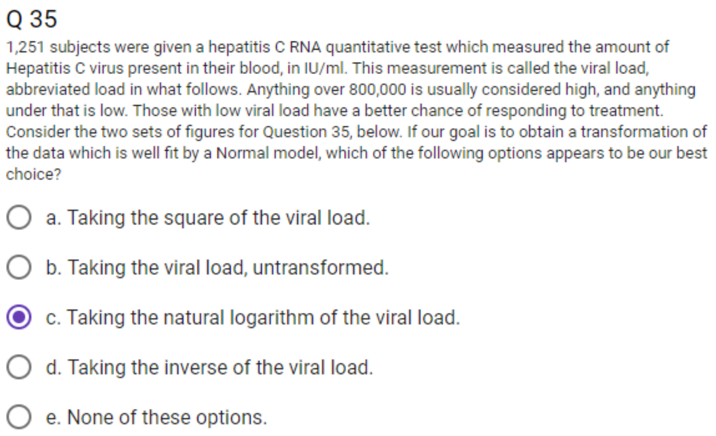
○ c. It is impossible to tell from the information provided.

Figure for Question 34



The mean of the aortic regurgitation data is clearly to the right of the mean in the mitral regurgitation data, according to the ridgeline plot.

Answer 35 is c, for 2 points.



Q 35

1,251 subjects were given a hepatitis C RNA quantitative test which measured the amount of Hepatitis C virus present in their blood, in IU/ml. This measurement is called the viral load, abbreviated load in what follows. Anything over 800,000 is usually considered high, and anything under that is low. Those with low viral load have a better chance of responding to treatment. Consider the two sets of figures for Question 35, below. If our goal is to obtain a transformation of the data which is well fit by a Normal model, which of the following options appears to be our best choice?

○ a. Taking the square of the viral load.

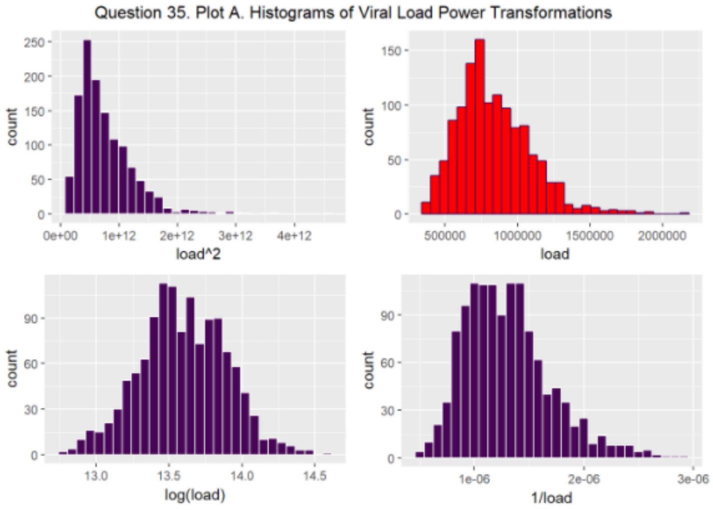
○ b. Taking the viral load, untransformed.

● c. Taking the natural logarithm of the viral load.

○ d. Taking the inverse of the viral load.

○ e. None of these options.

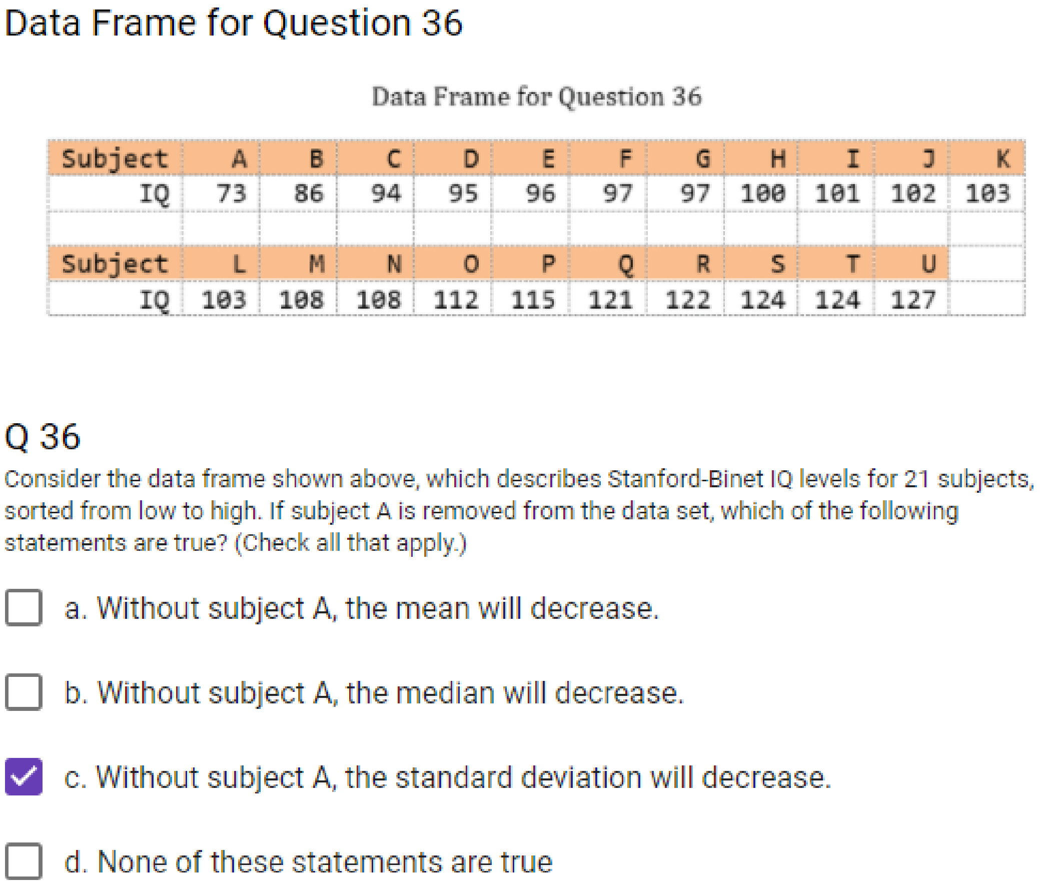
Figure Set A for Question 35



The log transformation is the best choice here. It's the only one that produces a symmetric histogram, or a straight line in the Normal Q-Q plot (which I didn't show here, but was on the quiz.)

- Note that log produces the natural logarithm (base e) in R. To get the base 10 logarithm, you'd use log10 and to get the base 2 logarithm, you'd use log2. Any logarithm will have the same impact on the model, though.

Answer 36 is c, for 3 points.



Data Frame for Question 36

Data Frame for Question 36

Subject	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
IQ	73	86	94	95	96	97	97	100	101	102	103	103	108	108	112	115	121	122	124	124	127

Q 36

Consider the data frame shown above, which describes Stanford-Binet IQ levels for 21 subjects, sorted from low to high. If subject A is removed from the data set, which of the following statements are true? (Check all that apply.)

☐ a. Without subject A, the mean will decrease.

☐ b. Without subject A, the median will decrease.

☑ c. Without subject A, the standard deviation will decrease.

☐ d. None of these statements are true

Without subject A, the mean will increase, since A is the minimum value. The median will stay the same, since we have two values tied at 103, so the middle of the distribution will still yield the same median. The standard deviation is the only one of these statistics that will decrease after A's removal. With subject A, we have mean = 105.1, median = 103 and sd = 13.8 Without subject A, we would have mean = 106.8, median = 103 and sd = 11.9.

Partial Credit

For each correct check or no check in boxes a-c, you received 1 point. If you checked d, regardless of what else you checked, you received no points.

Answer 37 is Sort by Population, for 3 points.

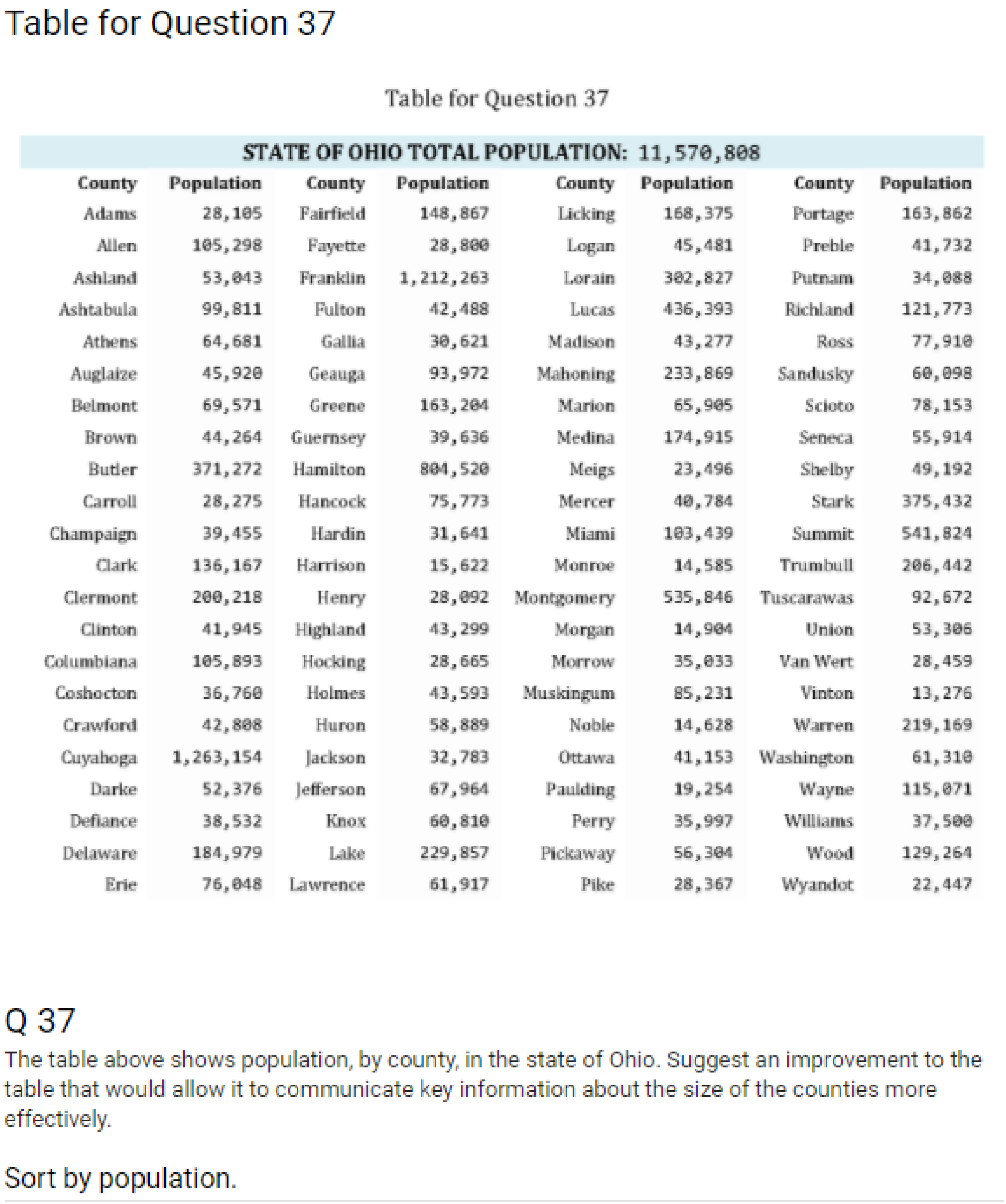


Table for Question 37

Table for Question 37

STATE OF OHIO TOTAL POPULATION: 11,570,808							
County	**Population**	**County**	**Population**	**County**	**Population**	**County**	**Population**
Adams	28,105	Fairfield	148,867	Licking	168,375	Portage	163,862
Allen	105,298	Fayette	28,800	Logan	45,481	Preble	41,732
Ashland	53,043	Franklin	1,212,263	Lorain	302,827	Putnam	34,088
Ashtabula	99,811	Fulton	42,488	Lucas	436,393	Richland	121,773
Athens	64,681	Gallia	30,621	Madison	43,277	Ross	77,910
Auglaize	45,920	Geauga	93,972	Mahoning	233,869	Sandusky	60,098
Belmont	69,571	Greene	163,204	Marion	65,905	Scioto	78,153
Brown	44,264	Guernsey	39,636	Medina	174,915	Seneca	55,914
Butler	371,272	Hamilton	804,520	Meigs	23,496	Shelby	49,192
Carroll	28,275	Hancock	75,773	Mercer	40,784	Stark	375,432
Champaign	39,455	Hardin	31,641	Miami	103,439	Summit	541,824
Clark	136,167	Harrison	15,622	Monroe	14,585	Trumbull	206,442
Clermont	200,218	Henry	28,092	Montgomery	535,846	Tuscarawas	92,672
Clinton	41,945	Highland	43,299	Morgan	14,904	Union	53,306
Columbiana	105,893	Hocking	28,665	Morrow	35,033	Van Wert	28,459
Coshocton	36,760	Holmes	43,593	Muskingum	85,231	Vinton	13,276
Crawford	42,808	Huron	58,889	Noble	14,628	Warren	219,169
Cuyahoga	1,263,154	Jackson	32,783	Ottawa	41,153	Washington	61,310
Darke	52,376	Jefferson	67,964	Paulding	19,254	Wayne	115,071
Defiance	38,532	Knox	60,810	Perry	35,997	Williams	37,500
Delaware	184,979	Lake	229,857	Pickaway	56,304	Wood	129,264
Erie	76,048	Lawrence	61,917	Pike	28,367	Wyandot	22,447

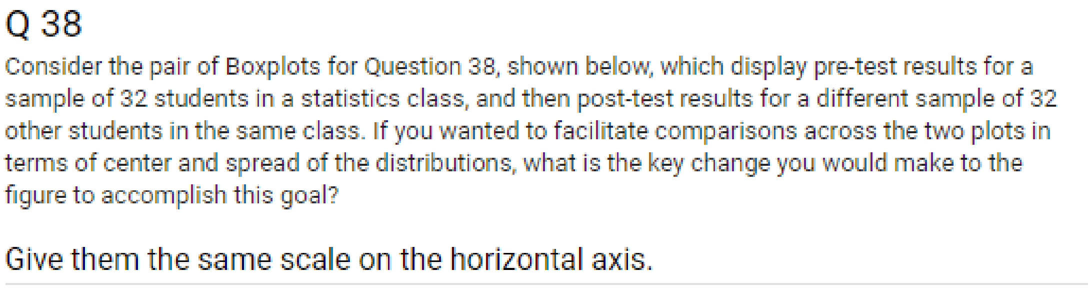
Q 37

The table above shows population, by county, in the state of Ohio. Suggest an improvement to the table that would allow it to communicate key information about the size of the counties more effectively.

Sort by population.

Resort the counties meaningfully, perhaps from highest to lowest population, or perhaps low to high.

Answer 38 is Match the axes, for 3 points.

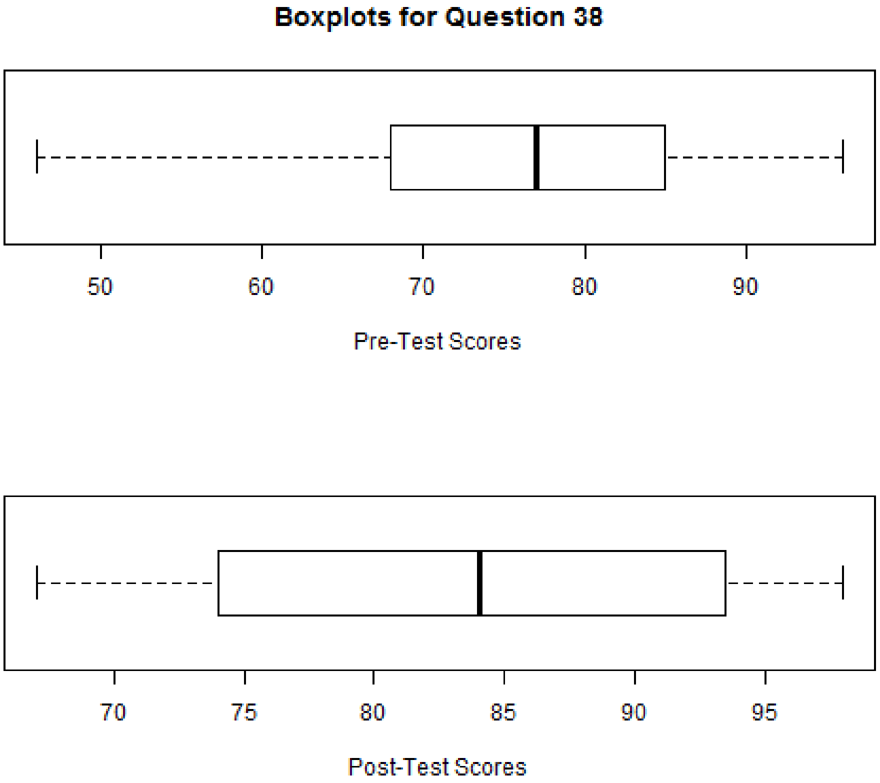


Q 38

Consider the pair of Boxplots for Question 38, shown below, which display pre-test results for a sample of 32 students in a statistics class, and then post-test results for a different sample of 32 other students in the same class. If you wanted to facilitate comparisons across the two plots in terms of center and spread of the distributions, what is the key change you would make to the figure to accomplish this goal?

Give them the same scale on the horizontal axis.

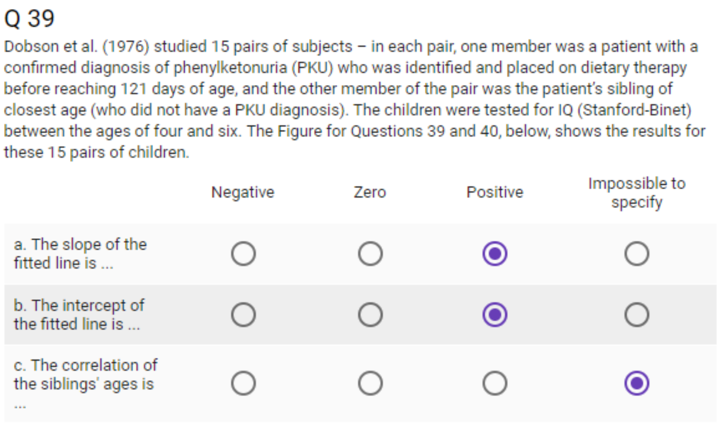
Figure for Question 38



Rescale the plots so they have the same limits on the X-axis. I gave full credit to anything that would accomplish that end.

At least a couple of people tried to add notches to the boxplots, and that alone wasn't the right way to go here.

Answer 39 is positive for a and b, and impossible to specify for c, for 3 points.

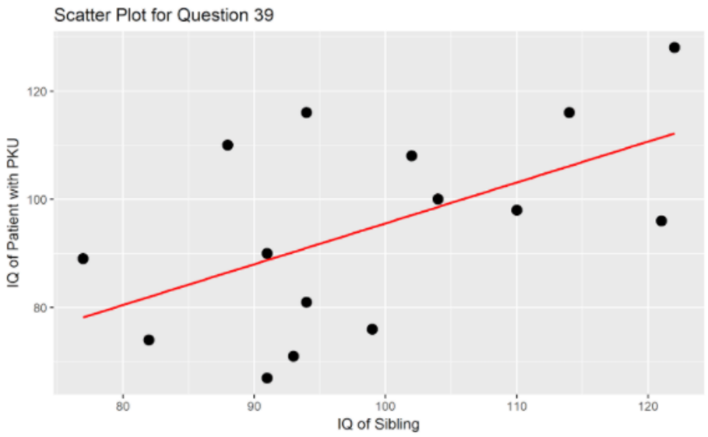


Q 39

Dobson et al. (1976) studied 15 pairs of subjects – in each pair, one member was a patient with a confirmed diagnosis of phenylketonuria (PKU) who was identified and placed on dietary therapy before reaching 121 days of age, and the other member of the pair was the patient's sibling of closest age (who did not have a PKU diagnosis). The children were tested for IQ (Stanford-Binet) between the ages of four and six. The Figure for Questions 39 and 40, below, shows the results for these 15 pairs of children.

	Negative	Zero	Positive	Impossible to specify
a. The slope of the fitted line is ...	○	○	●	○
b. The intercept of the fitted line is ...	○	○	●	○
c. The correlation of the siblings' ages is ...	○	○	○	●

Figure for Question 39



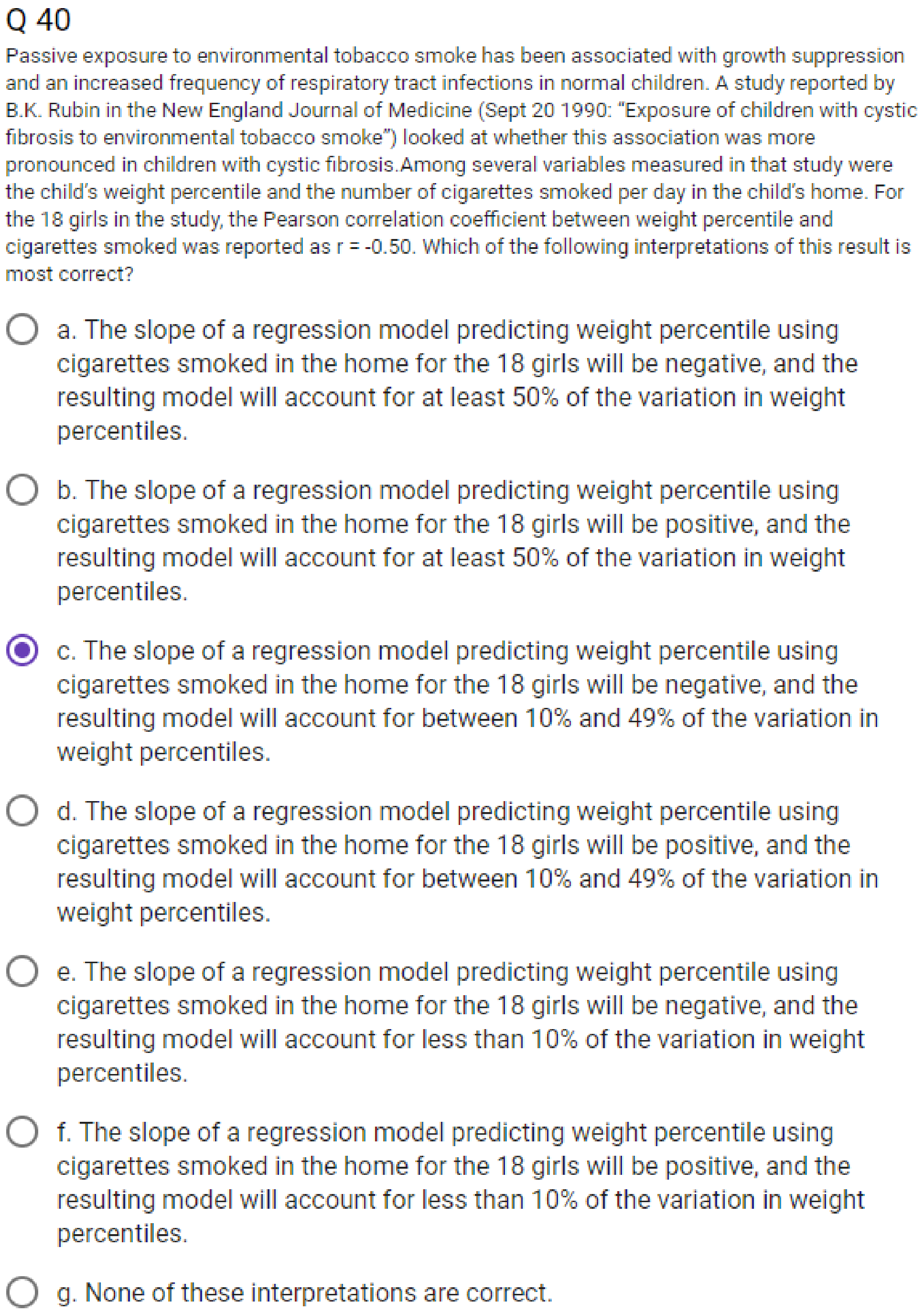
The slope is clearly positive, and the intercept is also positive. One way to see this is to realize that, as we move on the X-axis from IQ of sibling = 120 down to IQ of sibling = 80, we move down about 30 points (from 110 to 80) on the Y axis. If we move down further to 40, we will again move down about 30 points (from 80 to 50) on the Y axis. If we do this again to reach X = 0, then we will again move down about 30 points (from 50 to 20) - so our Y intercept must be near 20. In fact, the line is PKU = 20.0 + 0.76 Sibling.

The correlation question (c) was overly tricky. The correlation of the AGES is what I asked about, so the correct answer is "impossible to specify", but many people clearly assumed I meant the variables that were plotted (which were IQs, not AGEs) and chose "positive" for c as well, because if the slope is positive, then so is the Pearson correlation. So I gave credit for either answer.

Partial Credit

- 1 point for a response of Positive in a
- 1 point for a response of Positive in b
- 1 point for a response of either Positive or Impossible to Specify in part c

Answer 40 is c, for 3 points.



Q 40

Passive exposure to environmental tobacco smoke has been associated with growth suppression and an increased frequency of respiratory tract infections in normal children. A study reported by B.K. Rubin in the New England Journal of Medicine (Sept 20 1990: “Exposure of children with cystic fibrosis to environmental tobacco smoke”) looked at whether this association was more pronounced in children with cystic fibrosis.Among several variables measured in that study were the child’s weight percentile and the number of cigarettes smoked per day in the child’s home. For the 18 girls in the study, the Pearson correlation coefficient between weight percentile and cigarettes smoked was reported as r = -0.50. Which of the following interpretations of this result is most correct?

○ a. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for at least 50% of the variation in weight percentiles.

○ b. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for at least 50% of the variation in weight percentiles.

● c. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for between 10% and 49% of the variation in weight percentiles.

○ d. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for between 10% and 49% of the variation in weight percentiles.

○ e. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be negative, and the resulting model will account for less than 10% of the variation in weight percentiles.

○ f. The slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 18 girls will be positive, and the resulting model will account for less than 10% of the variation in weight percentiles.

○ g. None of these interpretations are correct.

If r = -0.5, then r^2 will be 25% (so the model accounts for 25% of variation), and the slope will be negative.

This item didn’t go well.

a and g were the most common incorrect responses.

Results on the Quiz

Grade Distribution

n	Mean	SD	Q1	Median	Q3	Max
51	81.8	10.3	74	82.5	89.5	100

Range	“Grade”	n
89.5 - 100	A	14
84 - 89	A-/B+	11
73 - 83	B	19
below 73	–	7

Scores by Item

- `# Full Credit` = Number of students (of 51) with full credit on the item
- `% Awarded` = Percentage of total available credit awarded across all students
- **Highlighted items** below are those where less than 80% of points were awarded - these were identified in the preceding pages as items that didn’t go well.

Item	# Full Credit	% Awarded	Item	# Full Credit	% Awarded
1	51	100	**21**	**29**	**57**
2	46+	>95	22	39	76
3	41	83	**23**	**24**	**47**
4	46+	>95	24	51	100
5	36	90	**25**	**33**	**65**
6	**20**	**69**	**26**	**27**	**53**
7	41	80	**27**	**32**	**63**
8	31	91	**28**	**36**	**71**
9	34	81	29	46+	>95
10	46+	>95	30	31	82
11	43	84	31	34	81
12	38	90	**32**	**39**	**77**
13	**31**	**70**	33	46+	95
14	43	84	34	42	82
15	46+	>95	35	46+	>95
16	45	94	36	46+	93
17	46+	94	37	46+	>95
18	**38**	**75**	38	42	82
19	26	80	39	43	94
20	**31**	**61**	**40**	**28**	**55**

Questions about Grades should be emailed to Dr. Love, now.

Dr. Love does all grading for this Quiz. The TAs beta-test the quiz, but do no grading on it. If you have questions about the points you received on specific items, email Dr. Love about your specific concerns. Unlike the homework assignments, he will address quiz grading issues as they appear.