431 Quiz 1 and Answer Sketch

Thomas E. Love

due 2021-10-04, draft: 2021-10-05 00:23:13

Packages for building the Quiz (and sketch)

```
library(broom)
library(Epi)
library(equatiomatic)
library(glue)
library(ggrepel)
library(janitor)
library(knitr)
library(magrittr)
library(naniar)
library(NHANES)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

Data Sets provided to students

- `oscar.csv`, first mentioned in Question 07
- `zips.csv`, first mentioned in Question 14
- `fastfood.csv`, first mentioned in Question 18

1 Question 01

Suppose that the height (in cm) of adult women living in the state of Ohio follows a Normal model with mean 162 and standard deviation 6. If this is the case, then what percentage of adult women living in the state of Ohio would have a height of 168 cm or larger? Please round your response to the nearest integer.

1.1 Answer 01 is 16 (percent).

I'd hoped you'd simply realize that 16% of data falls more than one standard deviation above the mean in a Normal distribution, and thus save yourself a calculation, since the remaining 84% must fall above that level, but if not, you certainly could use `pnorm` in R to do the calculation.

```
pnorm(168, mean = 162, sd = 6, lower.tail = FALSE)
```

```
[1] 0.1586553
```

1.2 Results 01

- No partial credit was available on Question 01.

Question 01	%
Correct response	94
Available points earned	94

- No incorrect response was selected by more than one person.

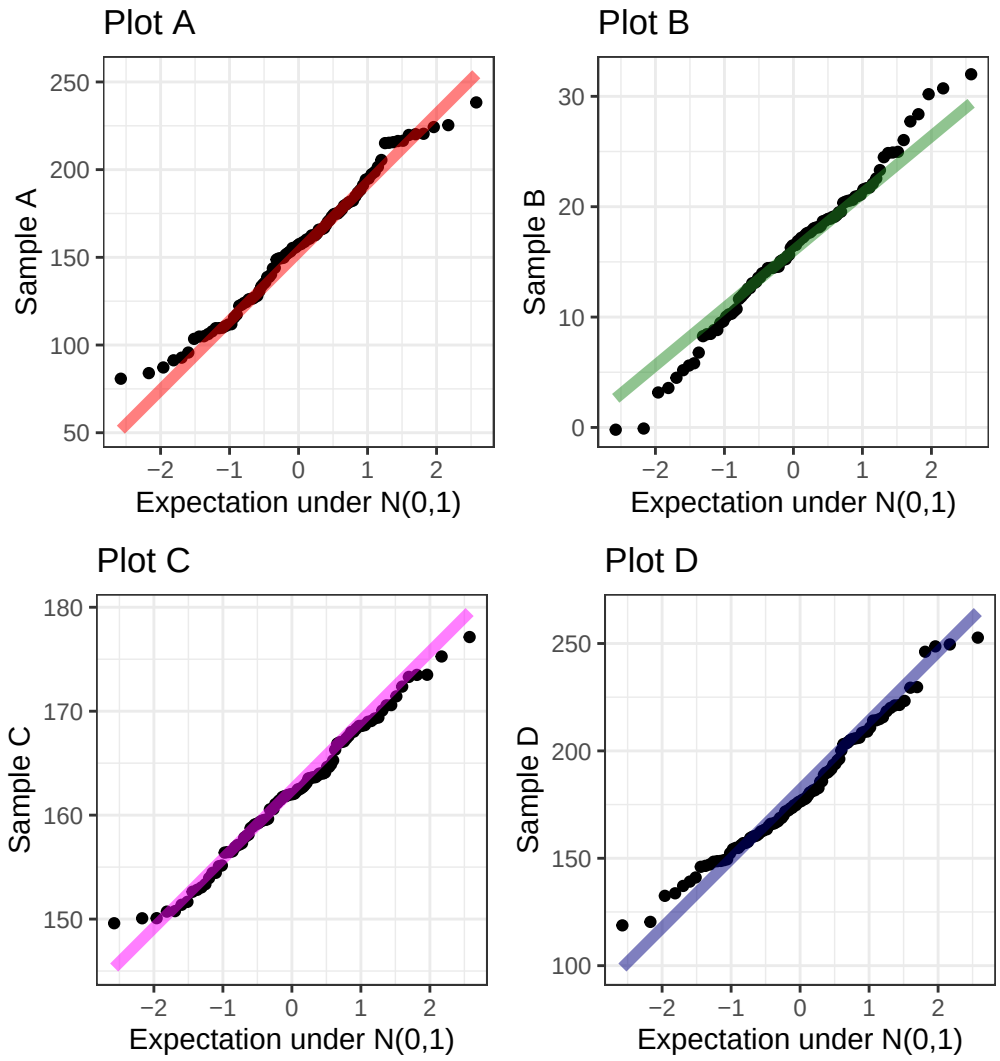
2 Question 02

There are four plots shown in the Figure for Question 02. Each shows a Normal Q-Q plot describing a different set of 200 heights. Which of the plots in the Figure for Question 02 shows data that could plausibly come from the Normal model specified in Question 01?

(Check all responses that are appropriate.)

a. Plot A
b. Plot B
c. Plot C
d. Plot D
e. None of the above.

Figure for Question 02



2.1 Answer 02 is c.

The Normal model of interest has mean 162 and standard deviation 6. Each of the four plots is reasonably close to a Normal distribution in terms of shape. So, let's think about the implications of a Normal model with mean 162 and standard deviation 6. For example, almost all of the data should fall within 3 standard deviations of the mean (from 162-18 = 144 to 163+18 = 181, roughly).

Looking at the Y axis in each plot shows us the range of each sample. Plot A displays a much larger standard deviation than 6, as does Plot D, and Plot D also seems to have a larger mean than 162. Plot B displays a standard deviation that's not far from 6, but a much smaller mean than 162. Only Plot C has an appropriate mean near 162 and standard deviation near 6, and in fact Plot C is the only data set derived from the Normal model specified in Question 01.

2.2 Results 02

- If you selected c and one other option, you received 2 points.
- If you selected c but two other options1, you received 1 point.

Question 02	%
Correct response	65
Available points earned	75

- The most common incorrect response was selecting both a and c.

3 Question 03

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. In a follow-up study, a new set of researchers measured a new set of 30 children, gathering each child's weight percentile and the number of cigarettes smoked per day in the child's home. For the 30 children in the new study, the Pearson correlation coefficient between weight percentile and cigarettes smoked was reported as r = -0.6. In interpreting the results in the responses below, the slope refers to the slope of a regression model predicting weight percentile using cigarettes smoked in the home for the 30 children. Which of the following interpretations of this result is most correct?

Select the best response.

a. The slope will be negative, and the model will account for less than one-quarter of the variation in weight percentiles.
b. The slope will be positive, and the model will account for less than one-quarter of the variation in weight percentiles.
c. The slope will be negative, and the model will account for between 25% and 49% of the variation in weight percentiles.
d. The slope will be positive, and the model will account for between 25% and 49% of the variation in weight percentiles.
e. The slope will be negative, and the model will account for at least half of the variation in weight percentiles.
f. The slope will be positive, and the model will account for at least half of the variation in weight percentiles.
g. None of these interpretations are correct.

3.1 Answer 03 is c.

If r = -0.6, then r^2 will be 36% (so the model accounts for 36% of variation), and the slope will be negative, because the least squares regression line's slope and the Pearson correlation coefficient are defined so that they must always have the same sign.

3.2 Results 03

- We gave 1 point for choice `d` which correctly identified the R^2 range.
- We gave 1 point for choice `e` which correctly identified the slope's sign.

Question 03	%
Correct response	79
Available points earned	83

- The most common incorrect response was e.

4 Question 04 (6 points)

The process of inductive inference, as described in *The Art of Statistics*, requires us to think hard about how we move from looking at the raw data to making general claims about the target population. Consider the following principles of effective measurement in this context.

```
I. We want to actually measure what we really want to measure
   without introducing systematic bias.

II. We want to sample at random whenever possible from the
    available subjects we are trying to make inferences about.

III. We want to use measures that give us a good chance of getting
     a similar result in a new study using the same measures.
```

Each of the principles listed above is associated primarily with a particular step in the process of building inductive inference. Identify the step (a, b or c) in the process associated with each of the statements (I, II or III) above.

a. Moving from the raw data to the sample
b. Moving from the sample to the study population
c. Moving from the study population to the target population

4.1 Answer 04 is that I is a, II is b, and III is a

See Spiegelhalter, Chapter 3.

Statement I describes the principle of validity, in the sense (as Spiegelhalter puts it) of measuring what you really want to measure and not having a systematic bias. This is part of moving from raw data to the sample, so I is primarily associated with a.

Statement II is about the importance of random sampling in order to generate a representation of the study population from the data, so II is primarily associated with b. This is related to the notion of internal validity - does the sample we observe accurately represent what is going on in the group we are studying?

Statement III describes the principle of reliability, in the sense (as Spiegelhalter puts it) of having low variability from occasion to occasion, and so being a precise or repeatable number. Reliability is also part of moving from raw data to the sample, so III is also primarily associated with a. Some people are likely to confuse what we're talking about here (reliability) with external validity, but that's not correct. We're talking about building a perfect measurement here, not about whether or not we're sampling the people we are particularly interested in sampling.

4.2 Results 04

- Each of the three parts of this question were worth 2 points.
- As expected, Part III was especially difficult. You had to read the text closely, and you had to not be afraid to use the same response twice.

Question 04	Part I %	Part II %	Part III %	As a Whole
Correct response	84	81	27	23
Available points earned	84	81	27	64

- The most common incorrect responses were (Part I): b, (Part II): a and (Part III): c
- Lots of people, I think, assumed that each response had to be selected once.

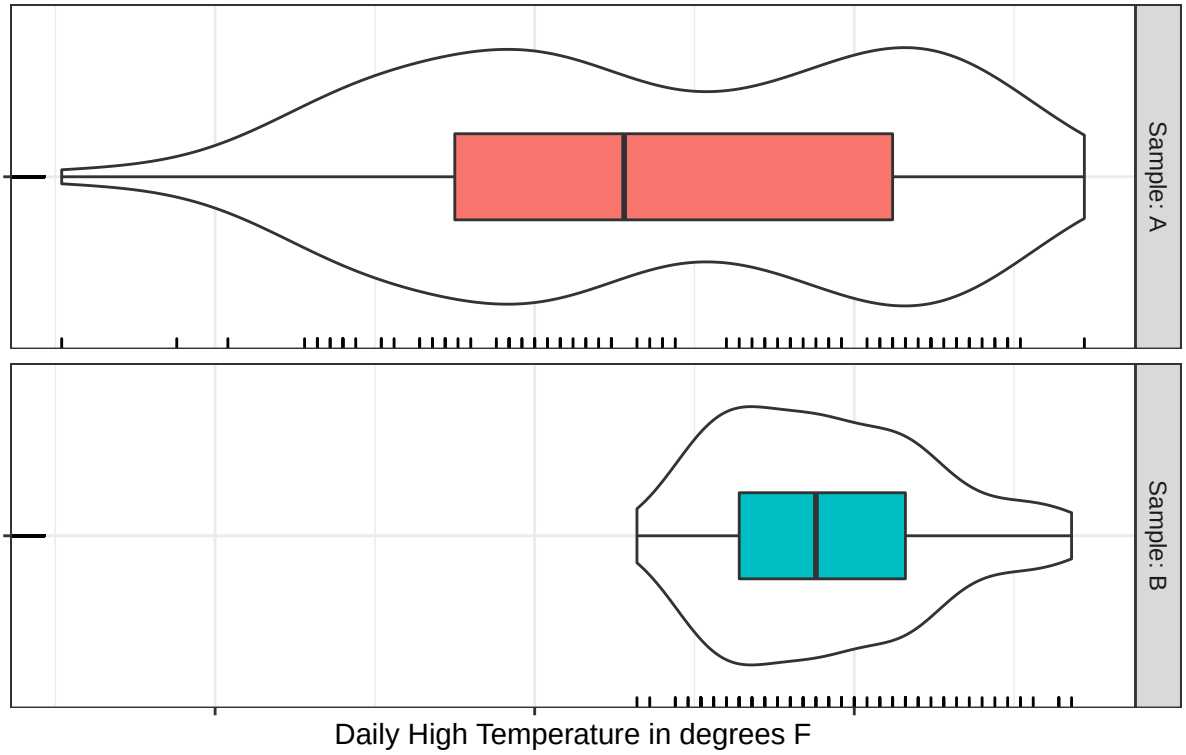
5 Question 05

The Figure for Question 05 shows the high daily temperatures (in degrees Fahrenheit) measured at Burke Lakefront Airport in Cleveland, Ohio in two groups of dates, drawn from the past few years.

- One of the samples was formed from a random selection of 100 dates in the month of September.
- The other sample includes a random selection of 100 dates from the entire year. Unfortunately, the x-axis (which was the same for each subplot) was left unlabeled, **but the missing x-axis labels are the same** for each of the two samples of data. The plot below provides some evidence regarding the distributions of the two samples.

Figure for Question 05. Comparing Samples A and B

Daily High Temperatures (in degrees F) at Burke Lakefront Airport in Cleveland, OH, USA



Consider the following statements:

```
I. Sample A describes the data gathered only in September.
II. The interquartile range in Sample A is wider than that of Sample B.
III. Sample A would be less accurately modeled using a Normal distribution than Sample B.
```

Which of the statements listed above are true?

a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

5.1 Answer 05 is f.

Statement I is false. September is a relatively warm month in Cleveland (much warmer, for instance, than the winter months.) Of these two samples, Sample B is clearly centered at a much higher temperature.

Statement II is true. We can see from the boxplots that the width of the box (which is the IQR) in Sample A is clearly greater than that shown for Sample B.

Statement III is also true. The Sample A data do not appear to describe a Normal distribution. There is, for instance, no central peak to that distribution as we can see from the violin plot. The Sample A data appear more uniform than a Normal or perhaps might be interpreted as multi-modal. Sample B, on the other hand, is reasonably symmetric, has a central peak, and no obvious signs of unusual tail behavior.

5.2 Results 05

- We gave 1 point for choice `f` which included both II and III but also included I.
- We gave 1 point for choice `b` which included II but left out III.

Question 05	%
Correct response	89
Available points earned	91

- The most common incorrect response was b.

6 Question 06

Here we consider data describing the age at onset (in years) for women with a diagnosis of multiple sclerosis. The oldest age at onset was 44 years. The stem-and-leaf display shows the data for the first 19 subjects.

```
The decimal point is 1 digit(s) to the right of the |

1 | 46788889
2 | 033667
3 | 239
4 | 24
```

If the next subject added to the data is 28 years of age, which of the following summary values will increase, as a result?

```
I. The mean
II. The standard deviation
III. The median
```

a. I only

b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

6.1 Answer 06 is a.

Only the mean would increase, in this case. The median would stay the same, and the standard deviation would decrease. Here's the demonstration.

```
msage <- tibble(age_onset = c(14, 16, 17, 18, 18, 18, 18,
                              19, 20, 23, 23, 26, 26, 27,
                              32, 33, 39, 42, 44))

mosaic::favstats(~ age_onset, data = msage)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
 min Q1 median   Q3 max     mean       sd  n missing
  14 18     23 29.5  44 24.89474 9.115888 19       0
```

```
msage_2 <- msage %>% add_row(age_onset = 28)

mosaic::favstats(~ age_onset, data = msage_2)
```

```
 min Q1 median Q3 max  mean       sd  n missing
  14 18     23 29  44 25.05 8.899882 20       0
```

6.2 Results 06

- We gave 1 point for choice d or choice e because you got I right but missed II or III.

Question 06	%
Correct response	81
Available points earned	84

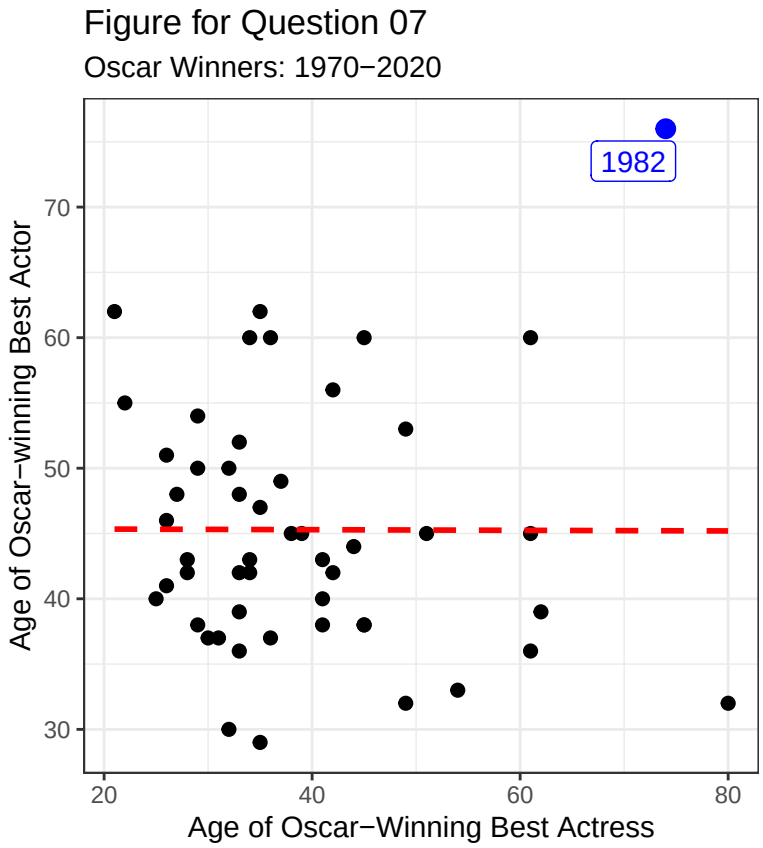
- The most common incorrect response was e.

7 Question 07

The data in the `oscar.csv` file I have provided to you describe the winners of the Academy Awards (also called the "Oscars") for Best Actor and Best Actress from 1970 to 2020.

The Figure for Question 07 is a scatterplot of 51 points, in each case displaying the age of the Best Actor (on the vertical, or y, axis) and the age of the Best Actress (on the horizontal, or x, axis) from the Academy Awards. Note that the Pearson correlation coefficient associated with these data is -0.003.

```
ggplot(oscar, aes(x = actress_age, y = actor_age)) +
    geom_point(size = 2) +
    geom_point(data = oscar %>% filter(year == 1982), col = "blue", size = 3) +
    geom_label_repel(data = oscar %>% filter(year == 1982),
                     aes(label = year), col = "blue") +
    geom_smooth(method = "lm", col = "red", lty = "dashed",
                se = FALSE, formula = y ~ x) +
    theme(aspect.ratio = 1) +
    labs(title = "Figure for Question 07",
         subtitle = "Oscar Winners: 1970-2020",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")
```



In 1982, Henry Fonda (age 76) and Katharine Hepburn (74) each won Oscars for the film *On Golden Pond*. This point is marked on the scatterplot by a blue dot, and labeled by its year. If the scatterplot were redrawn eliminating the 1982 awards, and including only the other 50 years, what would happen?

a. The slope of the linear model would INCREASE, and so would the R-squared.
b. The slope of the linear model would INCREASE, and the R-squared would DECREASE.
c. The slope of the linear model would DECREASE, and so would the model's R-squared.
d. The slope of the linear model would DECREASE, and the R-squared would INCREASE.
e. It is impossible to tell from the information provided.

7.1 Answer 07 is d.

Dropping the 1982 outlier will have a substantial effect on the regression line. When applied to the revised data, a regression line will fit the remaining 50 years substantially better (so the R^2, the square of the Pearson correlation, will increase) and the line will rise on the left of the graph and fall on the right, (so that the slope will be substantially decreased).

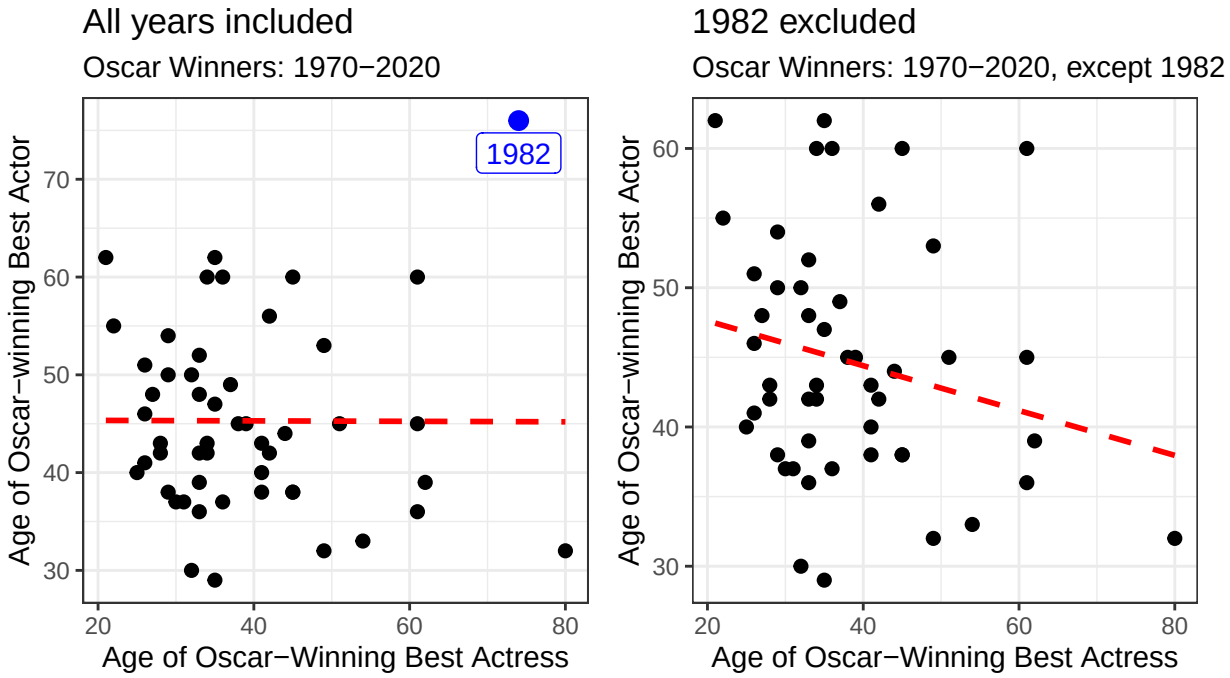
Here's what the plot looks like with and without 1982.

```
p1 <- ggplot(oscar, aes(x = actress_age, y = actor_age)) +
    geom_point(size = 2) +
    geom_point(data = oscar %>% filter(year == 1982), col = "blue", size = 3) +
    geom_label_repel(data = oscar %>% filter(year == 1982),
                     aes(label = year), col = "blue") +
    geom_smooth(method = "lm", col = "red", lty = "dashed",
                se = FALSE, formula = y ~ x) +
    theme(aspect.ratio = 1) +
    labs(title = "All years included",
         subtitle = "Oscar Winners: 1970-2020",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")

oscar_drop82 <- oscar %>% filter(year != 1982)

p2 <- ggplot(oscar_drop82, aes(x = actress_age, y = actor_age)) +
    geom_point(size = 2) +
    geom_smooth(method = "lm", col = "red", lty = "dashed",
                se = FALSE, formula = y ~ x) +
    theme(aspect.ratio = 1) +
    labs(title = "1982 excluded",
         subtitle = "Oscar Winners: 1970-2020, except 1982",
         x = "Age of Oscar-Winning Best Actress",
         y = "Age of Oscar-winning Best Actor")

p1 + p2
```



Here are the models, first including 1982...

```
m1 <- lm(actor_age ~ actress_age, data = oscar)

glance(m1) %>% select(r.squared, nobs)
```

```
# A tibble: 1 x 2
   r.squared  nobs
       <dbl> <int>
1 0.00000984    51
```

```
tidy(m1, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 4)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	45.3865	4.4218	37.9731	52.7998
actress_age	-0.0024	0.1079	-0.1833	0.1786

and now without 1982...

```
m2 <- lm(actor_age ~ actress_age, data = oscar_drop82)

glance(m2) %>% select(r.squared, nobs)
```

```
# A tibble: 1 x 2
  r.squared  nobs
```

```
      <dbl> <int>
1    0.0484    50
```

```
tidy(m2, conf.int = TRUE, conf.level = 0.90) %>%
    select(term, estimate, std.error, conf.low, conf.high) %>%
    kable(digits = 4)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	50.8392	4.1235	43.9233	57.7552
actress_age	-0.1610	0.1030	-0.3337	0.0118

Summarizing in a table, we have:

Dates Included	Model for Actor Age	R^2 (as a %)
1970-2020 (with 1982)	45.4 - 0.002 Actress Age	< 0.001
1970-1981, 1983-2020	50.8 - 0.161 Actress Age	4.8

7.2 Results 07

- No partial credit was available on Question 07.

Question 07	%
Correct response	90
Available points earned	90

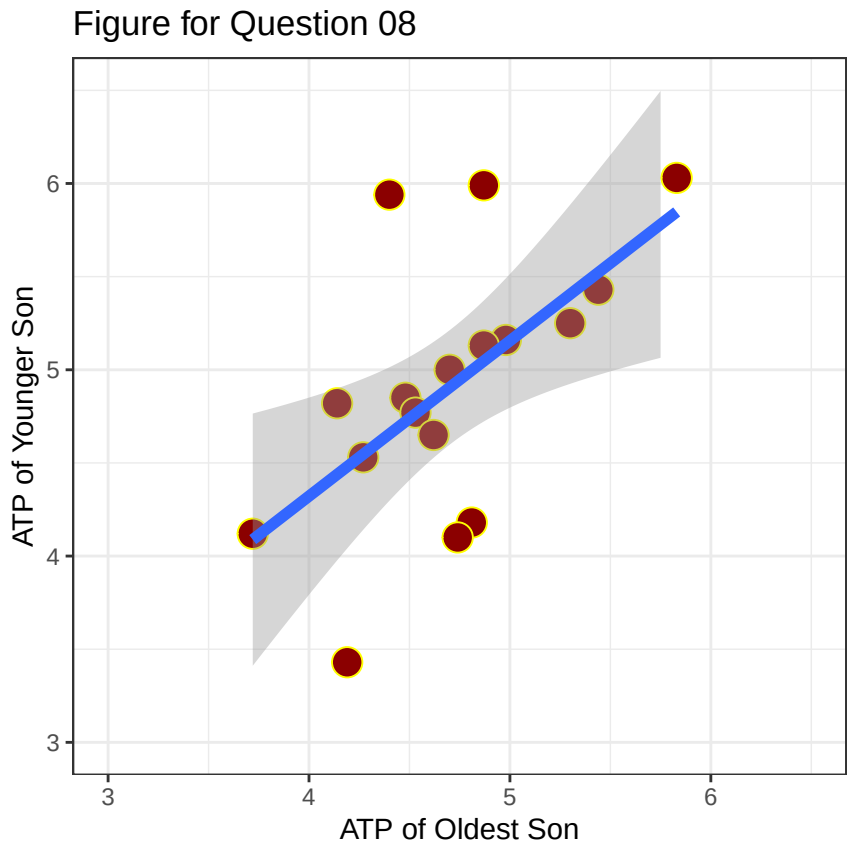
- The most common incorrect response was b.

8 Question 08

Dern and Wiorkowski (1469) 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 to transport oxygen in the blood. The Figure for Question 08 depicts the data for 17 pairs of brothers. Suppose we are also interested in an 18th family, where Kevin is the oldest son and Brian is the youngest son. Which of the following statements are true?

(Check all responses that are appropriate.)

a. If Kevin's ATP is 5, the linear model's point estimate for Brian's ATP exceeds 5.
b. The absolute value of the Pearson correlation is between 0 and 0.25.
c. The intercept of the regression line is less than zero.
d. The slope of the regression line is greater than zero.
e. None of these statements are true.



8.1 Answer 08 is both a and d.

- Statement a is true, from the regression line. The predicted ATP of youngest son if the ATP of oldest son is 5 is clearly less than 5, based on the regression line.
- As for Statement b, The correlation is pretty strong here, in fact it turns out to be 0.6, as we see in the output below, and at any rate is 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.
- Statement c 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 c is false. The actual regression line, as we see in the output below, is `young.atp` = 0.99 + 0.83 `old.atp`
- Statement d is also 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.

```
lm(young.atp ~ old.atp, data = atp)
```

```
Call:
lm(formula = young.atp ~ old.atp, data = atp)

Coefficients:
(Intercept)      old.atp
     0.9867       0.8337
```

```
cor(atp %>% select(old.atp, young.atp))
```

```
            old.atp young.atp
```

```
old.atp   1.0000000 0.5974007
young.atp 0.5974007 1.0000000
```

8.2 Results 08

- We gave one point to those with a, d and exactly one other response.

Question 07	%
Correct response	94
Available points earned	94

- The most common incorrect response was a, c and d.

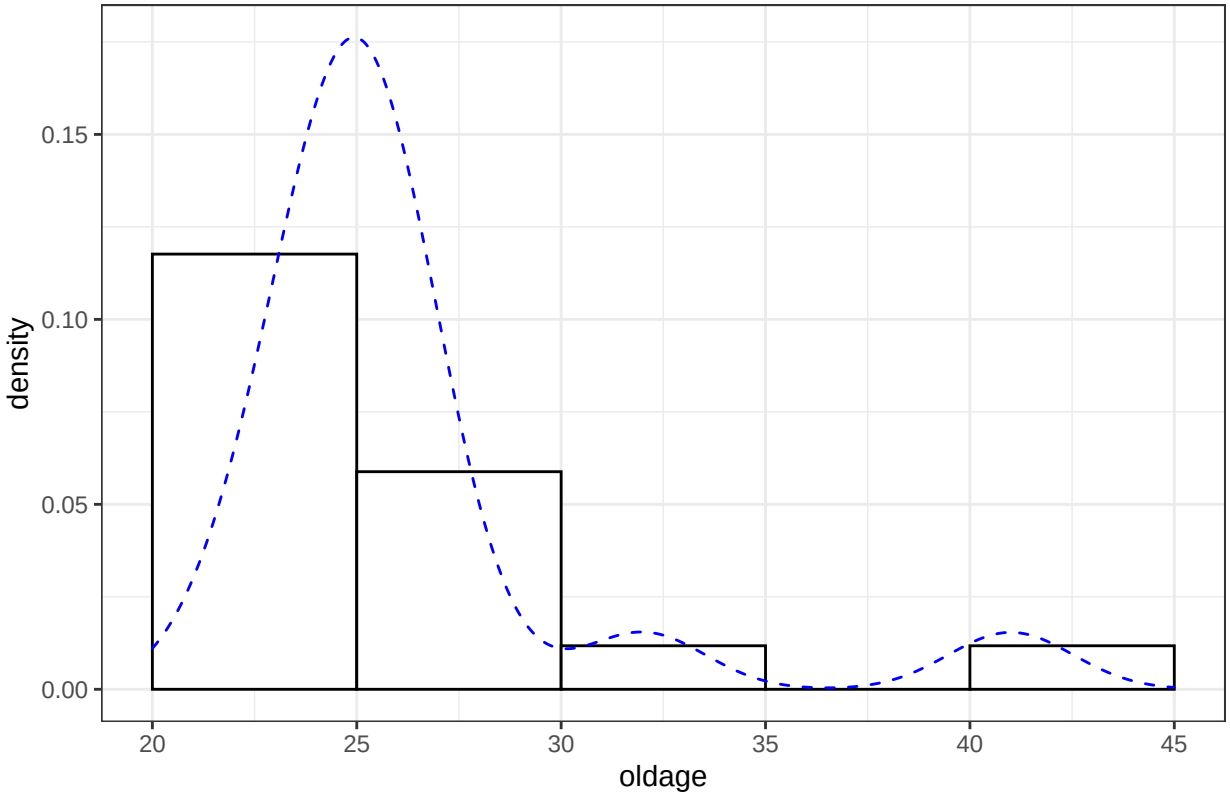
9 Question 09

Consider again the study described in Question 08, but now, we'll focus on the ages of the oldest sons. The Figure for Question 09 shows these ages (in years) for these 17 subjects, with a smooth density curve added. Which of the following statements are true?

(Check all responses that are appropriate.)

a. A Normal Q-Q plot of these data would show an S-shape.
b. The ages are symmetric, showing no substantial skew.
c. The mean of the ages is larger than the median age.
d. The range of the data (max - min) is between 15 and 25 years.
e. None of these statements are true

Figure for Question 09



9.1 Answer 09 is both c and d.

- These are right-skewed data, according to the histogram.
- Statement a would be true only if the data were symmetric but light-tailed. Not the case here.
- Statement b is also false because the data are not symmetric.
- Statement c is true, because the data are right skewed, which pulls the mean above the median.
- Statement d is also true. 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).

9.2 Results 09

- We gave 2 points to students with `c` alone.
- We also gave 2 points to students with `d` alone.
- We also gave 2 points to students with `c` and `d` and one other response.

Question 09	%
Correct response	50
Available points earned	71

- The most common incorrect response was c alone.

10 Question 10

Which of the following will create a sample in R of 500 observations from a Normal distribution with mean of 30 and standard deviation of 8, and place them into a variable called scores. You can assume that the tidyverse package is already loaded, and that an appropriate random seed has been set in a previous command.

(Check all responses that are appropriate.)

a. `scores <- 500*rnorm(n = 1, mean = 30, sd = 8)`
b. `scores <- tibble(rnorm(n = 25, mean = 30, sd = 8))`
c. `scores <- rep(rnorm(mean = 30, sd = 8), 500)`
d. `scores %>% rnorm(n = 500, mean = 30, sd = 8)`
e. `scores <- tibble(y <- rnorm(500, mean = 30, sd = 8))`
f. None of these commands will succeed.

10.1 Answer 10 is f.

None of these commands will do what I asked for. What you need is something like

```
scores <- rnorm(n = 500, mean = 30, sd = 8)
```

or perhaps you could put this within a tibble called `dat10` as:

```
dat10 <- tibble(scores = rnorm(n = 500, mean = 30, sd = 8))
```

- `a` will produce a single value in `scores` that is a random normal variable multiplied by 500.
- `b` produces 25 observations rather than 500, and puts them in a tibble called scores, rather than a variable called scores.
- `c` will produce an error since there's no n value in your rnorm call, among other things.
- `d` uses the pipe %>% rather than the assignment operator <-
- `e` will put the data into a variable called "y <- rnorm(500, mean = 30, sd = 8)" within a tibble called scores, and while I suppose that's a little closer, it's still not right.

10.2 Results 10

- As expected, this was difficult. People are always afraid to select "None of the above."
- We gave 1 point to the many students who selected `e` alone, since it's a little closer to what we're looking for than the other options, as mentioned above.

Question 10	%
Correct response	11
Available points earned	31

- The most common incorrect response was e.

11 Question 11

A new sample of 350 subjects ages 35-59 from the NHANES data generates the Table for Question 11, which summarizes the relationship between the subject's Self-Reported Overall Health (Excellent, Vgood = "Very Good", Good, Fair or Poor) and whether or not they have ever tried marijuana (Yes/No). In this sample, which group is more likely to report their Self-Reported Overall Health in one of the top three categories (Excellent, Very Good or Good)?

a. The "Yes" group, by more than three percentage points.
b. The "Yes" group, by 0.1 to 3 percentage points.
c. Neither group.
d. The "No" group, by 0.1 to 3 percentage points.
e. The "No" group, by more than three percentage points.
f. It is impossible to tell from the information provided.

Table for Question 11

	HealthGen					
Marijuana	Excellent	Vgood	Good	Fair	Poor	Total
No	18	37	60	28	5	148
Yes	21	74	74	29	4	202
Total	39	111	134	57	9	350

11.1 Answer 11 is a.

We could just do the math.

- In the "Yes" group, we have 21 + 74 + 74 = 169 people in the three best health groups, and that's out of a total of 202 people in the "Yes" group, so that's 83.7%.
- In the "No" group, we have 18 + 37 + 60 = 115 people in the three best health groups, and that's out of a total of 148 people in the "No" group, so that's 77.7%.
- So that's a difference of 6 percentage points, favoring the "Yes" group.

Or, we could get some help. Here are the health category percentages, within each marijuana group.

```
dat11 %>% tabyl(Marijuana, HealthGen) %>%
  adorn_percentages(denominator = "row") %>%
  adorn_pct_formatting()
```

```
 Marijuana Excellent Vgood  Good  Fair Poor
        No     12.2% 25.0% 40.5% 18.9% 3.4%
       Yes     10.4% 36.6% 36.6% 14.4% 2.0%
```

- So, in the "No" group, we have 12.2 + 25.0 + 40.5 = 77.7 percent in the three healthiest categories.
- In the "Yes" group, we have 10.4 + 36.6 + 36.6 = 83.6 percent in the three healthiest categories, a little different than what we saw before, thanks to rounding error.

In either case, that is a difference of (essentially) 6 percentage points, favoring the "Yes" group. That's choice **a**.

11.2 Results 11

- No partial credit was available on Question 01.

Question 11	%
Correct response	95
Available points earned	95

- The most common incorrect response was e.

12 Question 12 (5 points)

Suppose you have collected data into a tibble in R called `dat12`. The `dat12` data come from a cohort study to look at the impact of exposure to an industrial solvent (stored in the `solvent` variable: a factor taking on the values "none", "moderate" or "profound") on the probability of a bladder cancer diagnosis (stored as either "yes" or "no" in the `diagnosis` variable.)

Provide a single line of R code to obtain an appropriate summary of the association between these variables. You should not include more than one pipe in your response, but you may not need a pipe at all. Hint: The generation of a p value does not constitute an appropriate summary for this Question.

12.1 Answer 12 is a single line of R code.

The answer I was looking for is `dat12 %>% tabyl(solvent, diagnosis)`, which produces a cross-tabulation with the `solvent` information in the rows and the `diagnosis` information in the columns. Your result needed to yield the cross-tabulation.

Here, I'll make up a data set to show you.

```
dat12 <- tibble(
  id = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10),
  solvent = c("none", "none", "none", "moderate",
              "moderate", "moderate", "profound",
              "profound", "profound", "profound"),
  diagnosis = c("yes", "no", "no", "no", "yes",
                "yes", "no", "yes", "yes", "yes")
)

dat12 %>% tabyl(solvent, diagnosis)
```

```
  solvent no yes
 moderate  1   2
     none  2   1
 profound  1   3
```

Other options that would have worked include:

```
dat12 %$% table(solvent, diagnosis)
```

```
          diagnosis
solvent    no yes
  moderate  1   2
  none      2   1
  profound  1   3
```

or

```
xtabs(~ solvent + diagnosis, data = dat12)
```

```
          diagnosis
solvent    no yes
  moderate  1   2
  none      2   1
  profound  1   3
```

and I would have (somewhat grudgingly) also accepted

```
dat12 %>% count(solvent, diagnosis)
```

```
# A tibble: 6 x 3
  solvent  diagnosis     n
  <chr>    <chr>     <int>
1 moderate no            1
2 moderate yes           2
3 none     no            2
4 none     yes           1
5 profound no            1
6 profound yes           3
```

but I would not have accepted

```
chisq.test(dat12$solvent, dat12$diagnosis)
```

```
Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be
incorrect

    Pearson's Chi-squared test

data:  dat12$solvent and dat12$diagnosis
X-squared = 1.3194, df = 2, p-value = 0.517
```

because this doesn't provide the actual contingency table.

12.2 Results 12

I gave full credit (5 points) to the following responses:

- `adorn_pct_formatting(adorn_percentages(tabyl(dat12, solvent, diagnosis)))` which is very much unnecessary, but does work to produce a table of row percentages.
- `adorn_percentages(tabyl(dat12, solvent, diagnosis, show_na=F), denominator="row")` which does essentially the same thing.
- `dat12 %>% tabyl(diagnosis, solvent)`, of course, was fine, as was flipping the position of the two variables within the tabyl.
- `dat12 %$% table(diagnosis, solvent)` was fine, too. As was `table(dat12$solvent, dat12$diagnosis)`.
- `dat12 %>% count(solvent, diagnosis)` was less good, but I still gave it full credit.
- I also let someone get away with filtering first to complete cases, which makes no sense here, because they then piped that result into a tabyl appropriately.

I gave no credit to an answer that produced a chi-square test and no table, like:

- `chisq.test(dat12$solvent, dat12$diagnosis, correct=FALSE)`
- or `chisq.test(dat12$solvent, dat12$diagnosis)`
- or `tab <- dat12 %>% tabyl(solvent, diagnosis); chisq.test(tab)`

Nor did I give any credit to a response that wouldn't do anything except throw an error, of course, like trying to generate a correlation or a linear model, or a favstats result (these aren't numeric variables), or leaving out the name of the tibble (dat12) or trying to generate a ggplot without specifying any aesthetics, or trying to add a chi-square test to a table by using a comma or anything like that.

I also gave no credit to answers like the following, which also just produce errors:

- `tabyl(dat12$diagnosis, dat12$solvent) %>% adorn_percentages(denominator = "row")`
- `tabyl(dat12$solvent, dat12$diagnosis) %>% adorn_totals(where = c("row", "col"))`

No credit, too, to an answer that simplified the data in `dat12` to an (incorrect) 2x2 table.

- The only response that achieved partial credit (2/5 points) was if you used an appropriate method but called the tibble `data12` instead of the prescribed `dat12`.

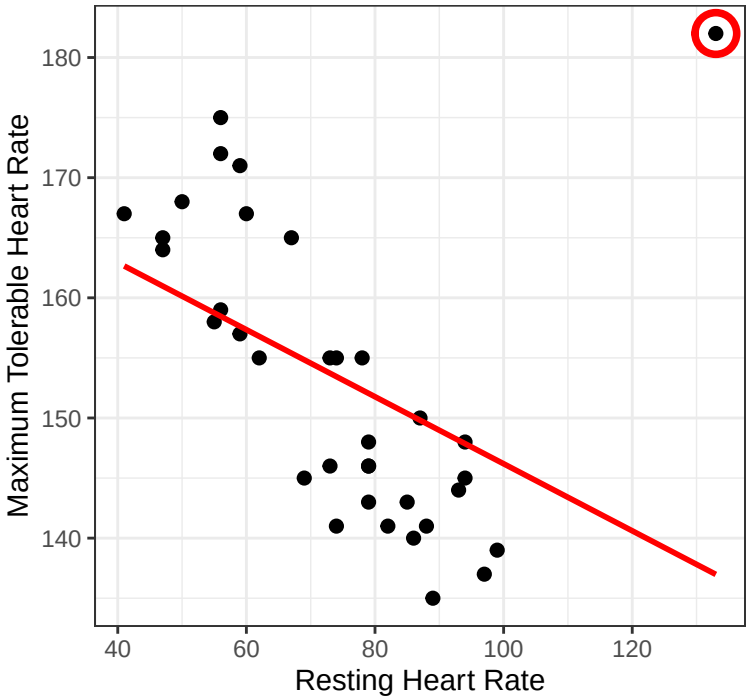
Question 12	%
Correct response	55
Available points earned	55

13 Question 13

The scatterplot shown in the Figure for Question 13 displays data on resting heart rate and maximum tolerable heart rate for 35 subjects in a research study. Subject 11, whose data are circled in red, has a resting heart rate of 133 and a maximum tolerable rate of 182. If the scatterplot was redrawn including only the other 34 subjects, the Pearson correlation coefficient would do what?

a. decrease
b. increase
c. remain unchanged
d. It is impossible to tell from the information provided.

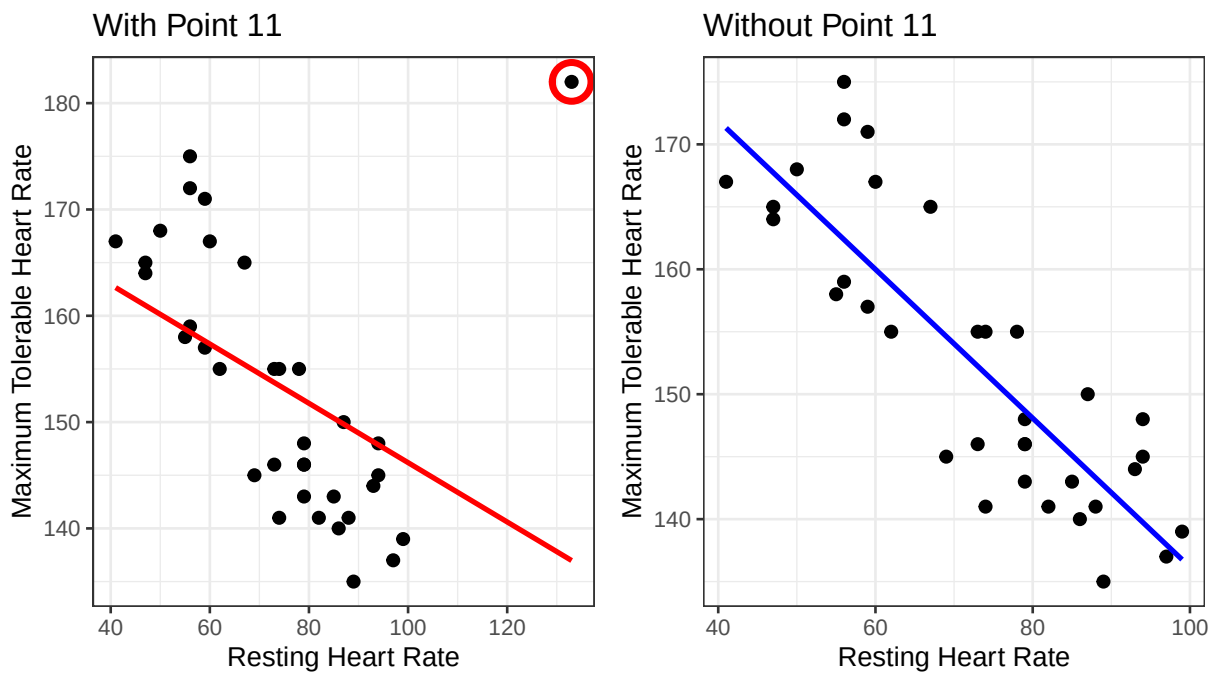
Figure for Question 13



13.1 Answer 13 is a.

Removal of the outlier would cause the association to be stronger, that is to say, it would make the Pearson correlation coefficient (which is already negative) move closer to -1, and thus decrease.

```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using formula 'y ~ x'
```



You didn't have the data, so you couldn't figure this out on your own using R (you had instead to look at the plot and think through what would happen), but in fact the Pearson correlation of the original data set with 35 observations is -0.43, but after we remove point 11 from the data, the Pearson correlation of the remaining 34 observations turns out to be -0.845.

```
q13 %$% cor(resting, max.tolerance)
```

```
[1] -0.4300022
```

```
q13 %>% slice(-11) %$%
  cor(resting, max.tolerance)
```

```
[1] -0.844986
```

13.2 Results 13

- No partial credit was available on Question 13

Question 13	%
Correct response	74
Available points earned	74

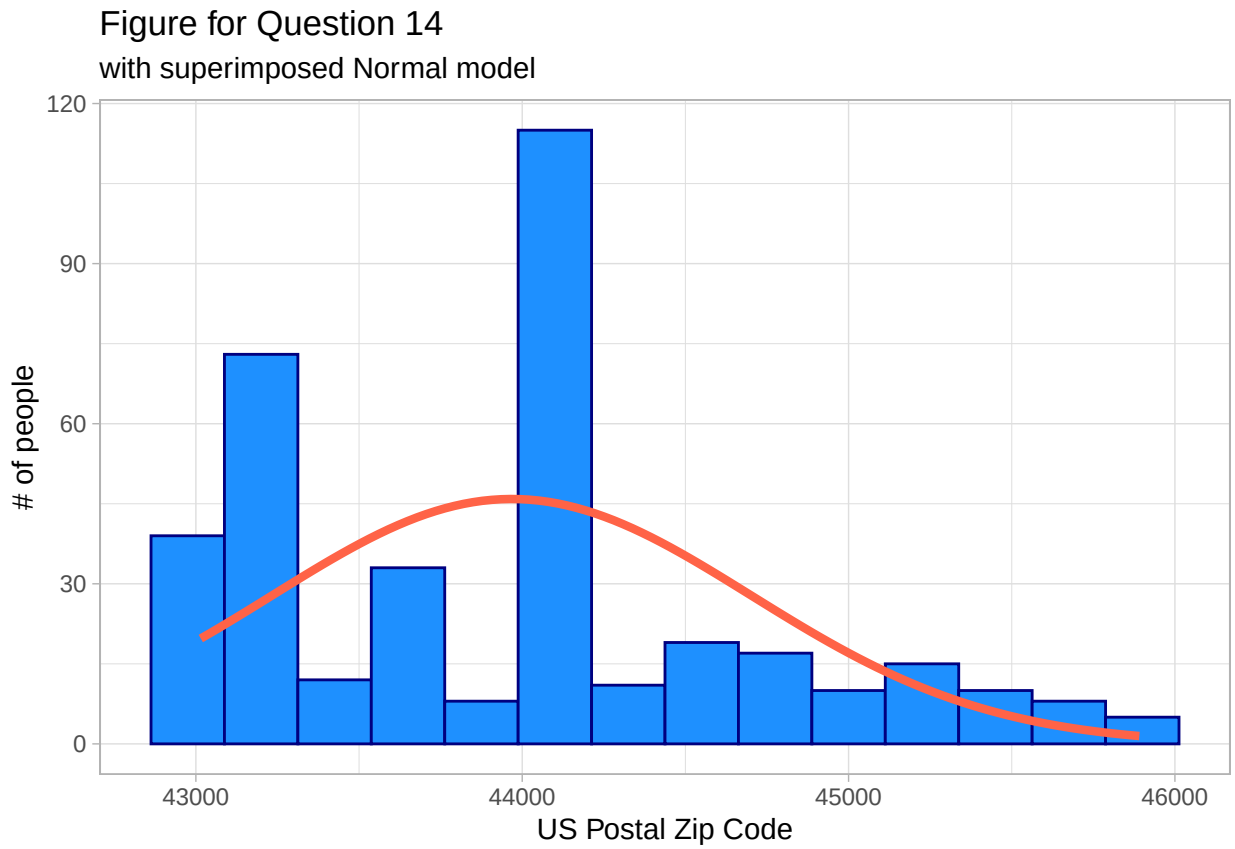
- The most common incorrect response was `b`.

14 Question 14

The Figure for Question 14 shows the US postal zip codes of the last 375 people from the state of Ohio to visit a web site providing information on purchasing insurance through the federal Health Insurance Marketplace. These data are also available to you in the `zips.csv` file provided with the Quiz. Which one of the following

summaries of these data would be most appropriate?

a. Mean
b. Median
c. Mode
d. IQR
e. It is impossible to tell from the information provided



14.1 Answer 14 is c.

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

14.2 Results 14

- No partial credit was available on Question 14

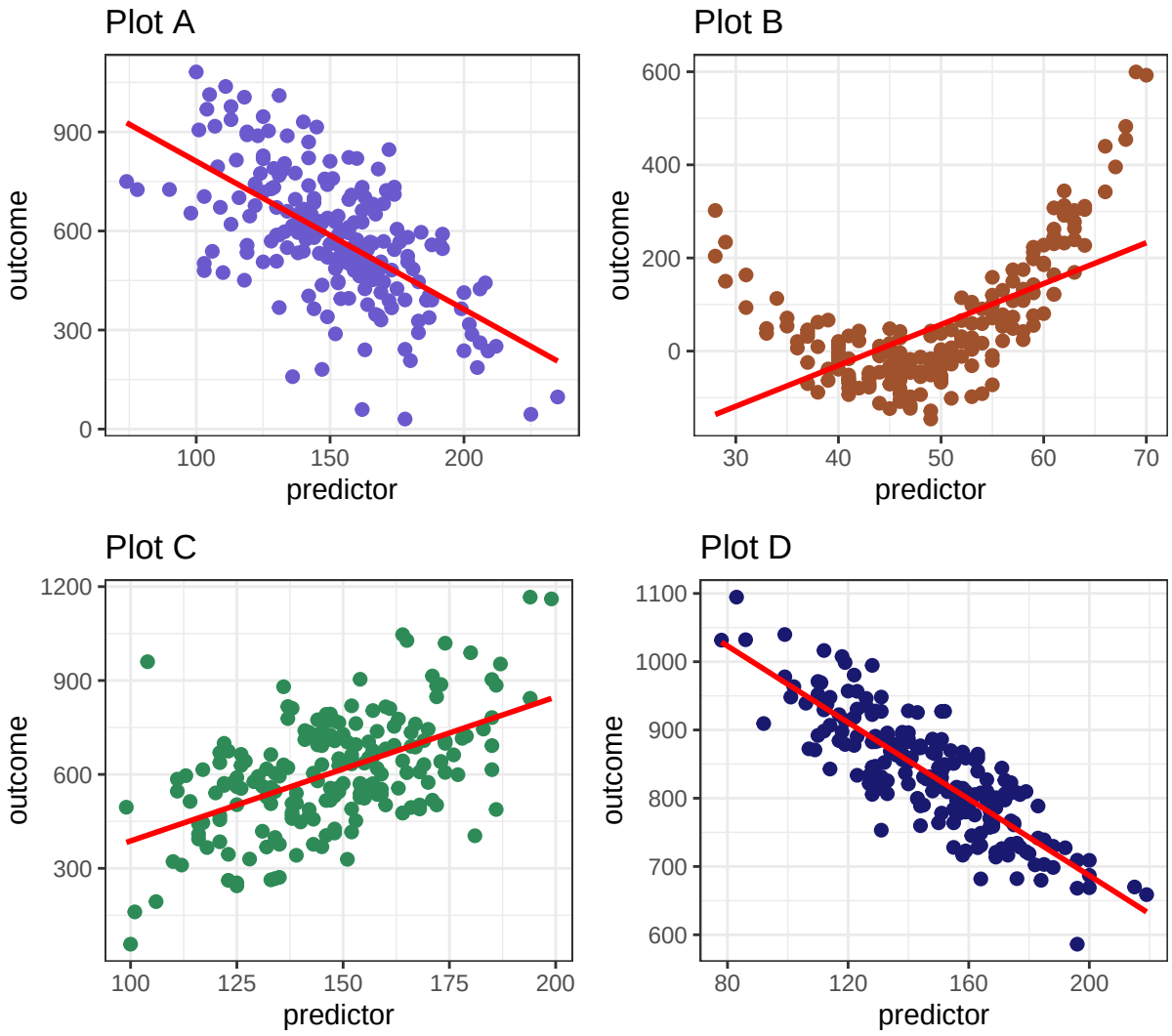
Question 14	%
Correct response	58
Available points earned	58

- The most common incorrect response was `b`.

15 Question 15

Consider the four scatterplots provided in the Figure for Question 15.

Figure for Question 15



Which of the four scatterplots in the Figure for Question 15 is associated with a linear model for `outcome` using `predictor` that has the largest R-square value?

a. Plot A.
b. Plot B.
c. Plot C.
d. Plot D.
e. It is impossible to tell from the information provided.

15.1 Answer 15 is d.

Plots A and C show considerably weaker linear associations than Plot D. Plot B shows a fairly clearly non-linear association. The correlations and resulting R^2 values for the data in each plot are tabulated below.

Plot	Pearson Correlation	R-Square
A	-0.621	0.386
B	0.585	0.342
C	0.533	0.284
D	-0.845	0.713

15.2 Results 15

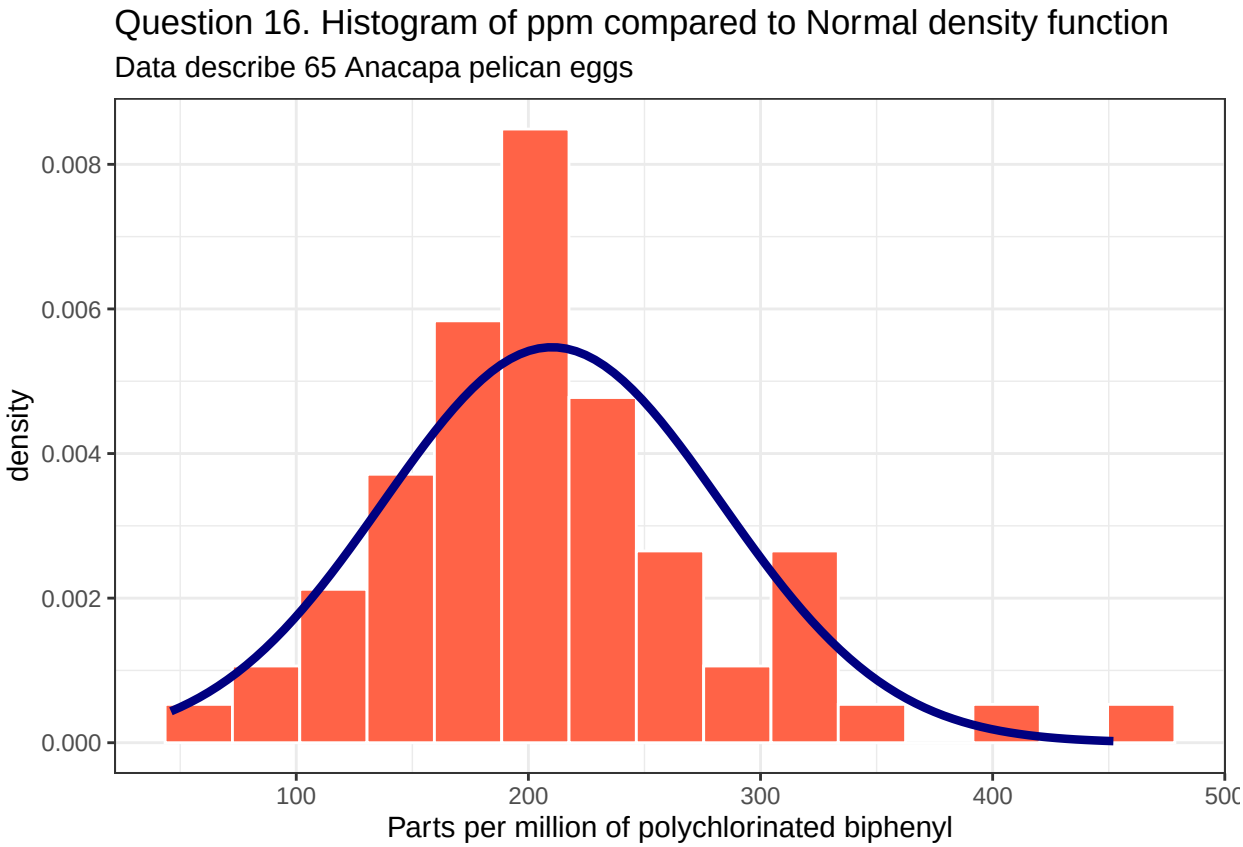
- No partial credit was available on Question 15

Question 15	%
Correct response	95
Available points earned	95

- The most common incorrect response was `e`.

16 Question 16 (5 points)

The data for Question 16 represent the concentration in parts per million of PCB (polychlorinated biphenyl, an industrial pollutant) for 65 Anacapa pelican eggs. The tibble containing the data is called `pelican` and the variable of interest is called `ppm`.



Here are eight lines of code. Note that Dr. Love definitely used lines 1, 2 and 8 in his code. He also used some of the other lines (lines 3-7) but not all of them.

```
1 pelican <- read_csv("data/pelican.csv")

2 ggplot(pelican, aes(x = ppm)) +
3   geom_histogram(aes(y = stat(density)), bins = 15, fill = "tomato", col = "white") +
4   geom_histogram(bins = 15, fill = "tomato", col = "white") +
5   geom_density(col = "navy", lwd = 1.5) +
6   coord_flip() +
7   stat_function(fun = dnorm,
                  args = list(mean = mean(pelican$ppm), sd = sd(pelican$ppm)),
                  col = "navy", lwd = 1.5) +
8   labs(title = "Question 16. Histogram of ppm compared to Normal density function",
        subtitle = "Data describe 65 Anacapa pelican eggs",
        x = "Parts per million of polychlorinated biphenyl")
```

This question continues on the next page. Note that Dr. Love has deliberately not provided the `pelican` data to you.

Please select each of the line numbers that should be REMOVED from the code in order to create the Question 16 plot.

(Check all line numbers that should be removed. More than one line may need to be removed.)

a. Line 3
b. Line 4
c. Line 5
d. Line 6
e. Line 7

16.1 Answer 16 is b, c and d.

Here's the code that generated the plot. As you can see, hashtags comment out what were described as lines 4, 5 and 6. Including any of those lines, or dropping any of the others, changes the plot in clear ways.

```
pelican <- read_csv("data/pelican.csv")

ggplot(pelican, aes(x = ppm)) +
    geom_histogram(aes(y = stat(density)), bins = 15, fill = "tomato", col = "white") +
#    geom_histogram(bins = 15, fill = "tomato", col = "white") +
#    geom_density(col = "navy", lwd = 1.5) +
#    coord_flip() +
    stat_function(fun = dnorm,
                  args = list(mean = mean(pelican$ppm), sd = sd(pelican$ppm)),
                  col = "navy", lwd = 1.5) +
    labs(title = "Question 16. Histogram of ppm compared to Normal density function",
        subtitle = "Data describe 65 Anacapa pelican eggs",
        x = "Parts per million of polychlorinated biphenyl")
```

16.2 Results 16

- I gave 3/5 points to people who selected two of the three lines that had to be removed, and didn't select any of the lines that had to remain. Some of those people chose b and d, while others chose c and d.

Question 16	%
Correct response	73
Available points earned	77

- The most common incorrect response was `a`, `c` and `d`.

17 Question 17

Suppose you are interested in how effectively shell thickness might be used to predict the concentration of environmental pollutants, in a setting like the study developed in Question 16. Which variable should go on the vertical (Y) axis of your scatterplot to display and model this association?

a. the egg identification number (1-65)
b. the concentration in parts per million of PCB
c. the thickness in micrometers of the egg's shell
d. It doesn't matter.
e. It is impossible to tell from the information provided.

17.1 Answer 17 is b.

The outcome goes on the vertical (Y) axis, and the predictor goes on the X axis. Here, we're modeling PCB concentration (our outcome) as a function of the shell thickness (our predictor). If we're just finding a correlation, it wouldn't matter, but if we're fitting a regression or other model, it does matter.

Data Source for Questions 16-17: Data set 165 in Hand DJ et al *A Handbook of Small Data Sets, Volume 1.* From Risebrough RW (1972) Effects of environmental pollutants upon animals other than man. *Proceedings of the 6th Berkeley Symposium on Mathematics and Statistics, VI.* California: University of California Press, 443-463.

17.2 Results 17

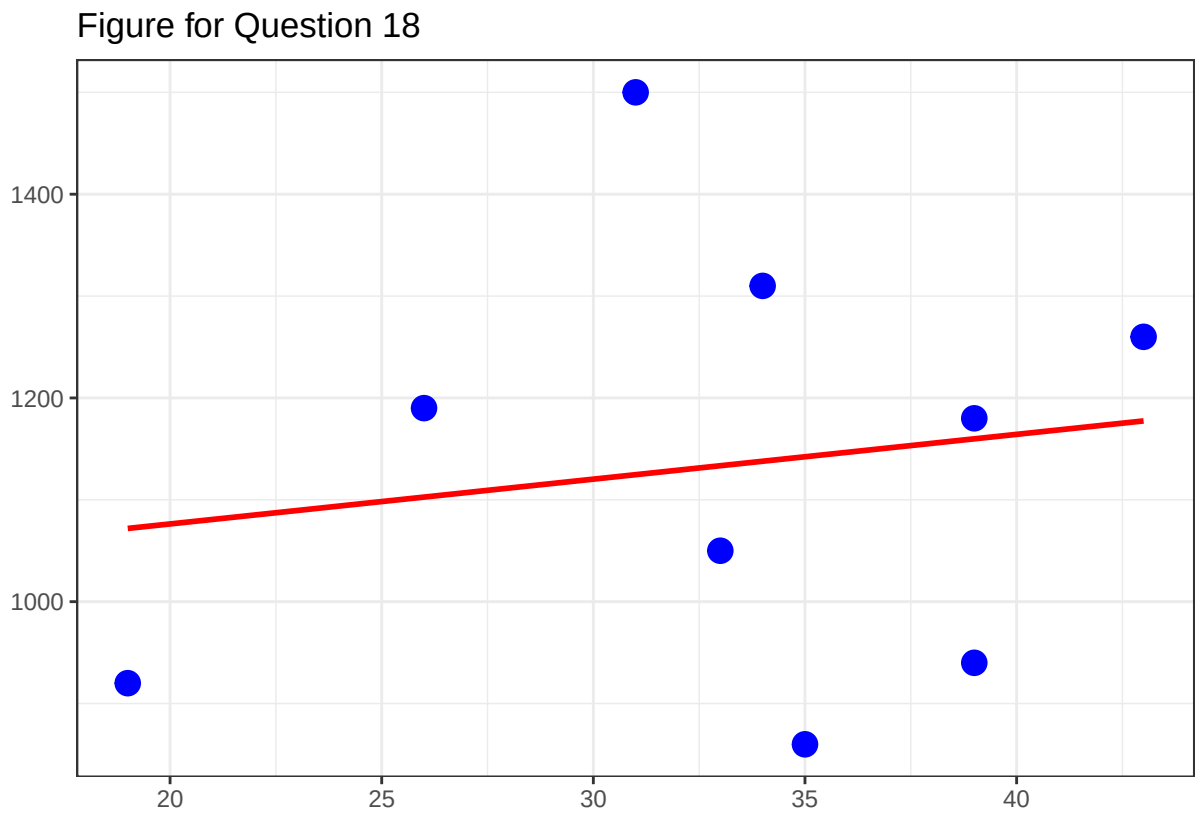
- No partial credit was available on Question 17

Question 17	%
Correct response	82
Available points earned	82

- The most common incorrect response was `c`.

18 Question 18

Fast food is often high in both fat and sodium. But are the two related? The scatter plot shown in the Figure for Question 18 describes the fat (in g) and sodium (in mg) contents of nine brands of hamburgers, and includes a linear model fit with geom_smooth, shown in red. I have provided the data in a file called `fastfood.csv`. In a sentence, what is the MOST IMPORTANT thing that should be done to improve the Figure?



18.1 Answer 18 is Add axis titles.

The correct response is to label the axes.

18.2 Results 18

- I gave full credit (4/4 points) to responses that specifically asked for axis titles.
  - If you also asked for a title for the plot, then that was fine.
  - If you also asked for a linear model to be plotted, that was also fine.
- You received 2/4 points of partial credit for
  - The addition of descriptive labels would improve the Figure for Question 18. [doesn't specifically indicate x and y axis labels.]
  - The most important thing that should be done to improve this figure would be to add x and y axes and a title. [again, no clear indication that axis labels are needed.]
- Some responses that received no credit included:
  - "geom_label which can show the title, the x axis (fat in gram) and y axis (sodium in mg)" [that's not what `geom_label` does]
  - "log of sodium content on Y axis, appropriate title for the figure, x and y axis" [that's not what is on the y axis]
  - "Remove outliers"
  - "Remove the geom_smooth line."
  - "We need to use the cor( ) function to calculate the correlation(R value) and find a line that fit in the relationship between the two variables."

Question 18	%
Correct response	87
Available points earned	89

19 Question 19

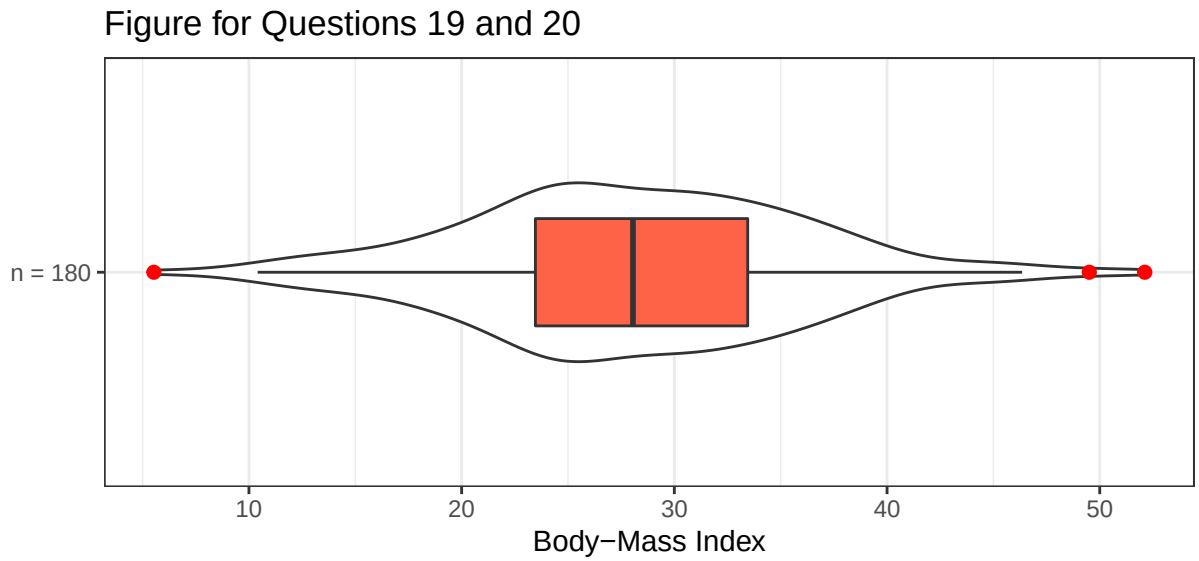
The Figure for Questions 19 and 20 displays the body-mass index (in $\frac{kg}{m^2}$) for 180 adults who suffer from rheumatoid arthritis, who are subjects in a new study. The mean BMI in the sample is 28.1 $\frac{kg}{m^2}$, the standard deviation is 7.9 $\frac{kg}{m^2}$ and there are no missing values. Which of the following statements are true?

(Check all responses that are appropriate.)

a. The distribution is substantially skewed and cannot be approximated well with a symmetric model.
b. The median is about 34 $\frac{kg}{m^2}$.
c. The IQR is about 10 $\frac{kg}{m^2}$.
d. The distribution has too many outliers to be approximated well with a Normal model.
e. None of these statements are true.

```
set.seed(20214315)
temp <- rnorm(180, mean = 29, sd = 8)
dat19 <- tibble(pt_id = 1:180, bmi = round(temp,2)-0.4)

ggplot(dat19, aes(x = "n = 180", y = bmi)) +
    geom_violin(width = 0.5) +
    geom_boxplot(fill = "tomato", width = 0.3, outlier.size = 2, outlier.color = "red") +
    labs(x = "", y = "Body-Mass Index",
         title = "Figure for Questions 19 and 20") +
    coord_flip()
```



19.1 Answer 19 is is c, and only c.

- The data are not particularly asymmetric.

- The median is about 28, not 34.
- The IQR is about 10.
- There are 3 outliers in 180 observations, which is not in and of itself sufficient to render a Normal model impractical.

Here are the `favstats` results for these data.

```
mosaic::favstats(~ bmi, data = dat19)
```

```
  min      Q1 median    Q3   max    mean       sd   n missing
 5.53 23.4675 28.055 33.45 52.12 28.0795 7.940113 180       0
```

19.2 Results 19

- No partial credit was available on Question 19

Question 19	%
Correct response	94
Available points earned	94

- The most common incorrect response was `e`.

20 Question 20

In the study of subjects with rheumatoid arthritis discussed in Question 19, adults with a BMI value of 25 or higher will be classified as overweight. Based on the Figure for Questions 19 and 20, how many of the 180 subjects would qualify as overweight using this standard?

a. Fewer than 45 subjects
b. Between 45 and 89 subjects
c. Exactly 90 subjects
d. Between 91 and 135 subjects
e. More than 135 subjects
f. There is insufficient information to answer the question.

20.1 Answer 20 is d.

Clearly 25 is below the median (so more than 50% of the 180 patients can have BMI 25 or more) but a good deal more than the 25th percentile (so at least 25% of the 180 patients have BMI below 25). So the correct answer is somewhere above 50% and below 75% of the 180 patients.

- Since 50% of 180 is 90 and 75% of 180 is 135, it looks like the right answer is between those values. That's response d.

20.2 Results 20

- No partial credit was available on Question 20

Question 20	%
Correct response	90
Available points earned	90

- The most common incorrect response was `f`.

21 Question 21 (2 points)

Consider the `starwars` tibble that is part of the `dplyr` package in the tidyverse. How many of the characters listed in that tibble are a good match for Professor Love, in that they are listed in the tibble as being of the Human `species`, having brown `hair_color` and blue `eye_color`?

(Note that we ask for blue `eye_color` and brown `hair_color`, specifically, here, and not other related colors or combinations of these with other colors.)

21.1 Answer 21 is 4.

There are four characters who meet these requirements, as listed in the output below.

```
starwars %>%
  filter(species == "Human" &
           hair_color == "brown" &
           eye_color == "blue") %>%
  select(name, species, hair_color, eye_color, homeworld)
```

```
# A tibble: 4 x 5
  name               species hair_color eye_color homeworld
  <chr>              <chr>   <chr>      <chr>     <chr>
1 Beru Whitesun lars Human   brown      blue      Tatooine
2 Jek Tono Porkins   Human   brown      blue      Bestine IV
3 Qui-Gon Jinn       Human   brown      blue      <NA>
4 Cliegg Lars        Human   brown      blue      Tatooine
```

21.2 Results 21

- No partial credit was available on Question 21. Remember that this question is worth 2 points.

Question 21	%
Correct response	90
Available points earned	90

- The most common incorrect response was 7.

22 Question 22 (2 points)

Of the characters you identified in Question 21, how many are from the `homeworld` of Tatooine?

22.1 Answer 22 is 2.

Two of the four characters identified in Question 21 call Tatooine home, as we can see from the output above.

22.2 Results 22

- No partial credit was available on Question 22. Remember that this question is worth 2 points.

Question 22	%
Correct response	> 95
Available points earned	> 95

- No incorrect response came from more than one person.

23 Question 23

How many of the characters in the entire `starwars` tibble have missing data in at least one of the four variables: `species`, `hair_color`, `eye_color` and `homeworld`?

23.1 Answer 23 is 18.

There are many ways to address this issue. If you wanted to use the `naniar` package, this was my simplest solution.

```
starwars %>%
    select(species, hair_color, eye_color, homeworld) %>%
    miss_case_table()
```

```
# A tibble: 3 x 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
1              0      69     79.3
2              1      17     19.5
3              2       1      1.15
```

and so we have 17 + 1 = 18 observations with at least one missing value across these four variables.

But there are lots of other options, including

```
starwars %>%
  count(is.na(species) | is.na(hair_color) |
        is.na(eye_color) | is.na(homeworld))
```

```
# A tibble: 2 x 2
  `is.na(species) | is.na(hair_color) | is.na(eye_color) | is.na(homeworl~     n
  <lgl>                                                                    <int>
1 FALSE                                                                       69
2 TRUE                                                                        18
```

or perhaps

```
starwars %>%
  filter(!complete.cases(species, hair_color,
                         eye_color, homeworld)) %>%
  nrow()
```

```
[1] 18
```

or maybe

```
starwars %>%
  count(!complete.cases(species, hair_color,
                        eye_color, homeworld))
```

```
# A tibble: 2 x 2
  `!complete.cases(species, hair_color, eye_color, homeworld)`     n
  <lgl>                                                        <int>
1 FALSE                                                           69
2 TRUE                                                            18
```

which gives the answer directly, but there are many, many ways to get it a little less directly, such as by first using

```
nrow(starwars)
```

```
[1] 87
```

to tell you that there are 87 characters in the tibble as a whole, and then something like:

```
starwars %>%
  filter(complete.cases(species, hair_color,
                        eye_color, homeworld)) %>%
  nrow()
```

```
[1] 69
```

to tell you that 69 of them have complete data on these four variables, and thus the other 18 do not.

23.2 Results 23

- No partial credit was available on Question 23.

Question 23	%
Correct response	71
Available points earned	71

- The most common incorrect response was 19.

24 Question 24

Suppose that you built a subset of the `starwars` data called `humanbrown` which consists only of the characters who are Human with brown `eye_color`. Now, you want to obtain the median of their mass in kilograms, among those subjects who have a mass recorded. Which of the following lines of R code would do that?

(Check all responses that are appropriate.)

a. `summary(humanbrown %>% select(mass))`
b. `humanbrown %>% filter(complete.cases(mass)) %>% summarize(quantile(mass, probs = 0.5))`
c. `humanbrown %>% summarize(median(mass, na.rm = TRUE))`
d. `humanbrown %>% filter(complete.cases(mass)) %>% summarize(median(mass))`
e. `mosaic::favstats(~ mass, data = humanbrown)`
f. None of these.

24.1 Answer 24 is that a, b, c, d, and e all work.

First, we'll build the subset.

```
humanbrown <- starwars %>%
    filter(species == "Human", eye_color == "brown")
```

Then we'll try it. The answer turns out to be 79, and all five approaches work.

```
summary(humanbrown %>% select(mass))
```

```
      mass
 Min.   :45.00
 1st Qu.:78.60
 Median :79.00
 Mean   :74.75
 3rd Qu.:82.00
 Max.   :85.00
 NA's   :6
```

```
humanbrown %>% filter(complete.cases(mass)) %>% summarize(quantile(mass, probs = 0.5))
```

```
# A tibble: 1 x 1
  `quantile(mass, probs = 0.5)`
                          <dbl>
1                            79
```

```
humanbrown %>% summarize(median(mass, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `median(mass, na.rm = TRUE)`
                         <dbl>
1                           79
```

```
humanbrown %>% filter(complete.cases(mass)) %>% summarize(median(mass))
```

```
# A tibble: 1 x 1
  `median(mass)`
           <dbl>
1             79
```

```
mosaic::favstats(~ mass, data = humanbrown)
```

```
 min   Q1 median Q3 max     mean       sd  n missing
  45 78.6     79 82  85 74.74545 13.94822 11       6
```

24.2 Results 24

- We gave 3/4 points to all students who selected four of the five correct responses.
- We have 2/4 points to all students who selected three of the five correct responses.

Question 24	%
Correct response	66
Available points earned	77

- The most common incorrect response was b, c and d only.

25 Question 25

I produced the cross-tabulation shown in the Output for Question 25 using the complete `starwars` tibble available in the `tidyverse`. All relevant packages are loaded on my computer. Which one of the following commands did I use?

a. `mosaic::favstats(~ gender + height >= 155, data = starwars)`
b. `starwars %$% table(gender, height >= 155)`
c. `starwars %$% tabyl(gender, height >= 155)`
d. `starwars %$% filter(height >= 155) %>% count(gender)`
e. `table(gender, height >= 155, data = starwars)`
f. None of these would work.

Output for Question 25

```
gender      FALSE TRUE
  feminine      3   13
  masculine     9   53
```

25.1 Answer 25 is b.

I obtained this result with:

```
starwars %$% table(gender, height >= 155)
```

```
gender      FALSE TRUE
  feminine      3   13
  masculine     9   53
```

None of the other codes produce this result. Most would only produce an error message.

25.2 Results 25

- No partial credit was available on Question 25.

Question 25	%
Correct response	90
Available points earned	90

- The most common incorrect response was `f`.

26 Overall Results

Sum up your scores on each item, and then add 2 points (so that the maximum possible score is actually 102, but we'll pretend it's out of 100.)

- Best Score: 100/100 (originally 98/100 + 2 points)
- Median Score: 84/100
- Mean Score: 81.6/100
- Standard Deviation of Scores: 13.6 points

As for how I'd think about those scores:

- Scores in the A range (90 - 100): 19
- Scores between 85 and 89: 11
- Scores in the B range (70 - 84): 24
- There were also some scores below 70.

I encourage all students with scores below 80 should be focusing now on showing meaningful improvement on Project A and Quiz 2. That's probably a good idea for those above 80, too, but I'm most concerned about those in the lower B range and below.