# 431 Quiz 1

Thomas E. Love

Version: 2021-09-30 11:24:35

## Instructions for Students

There are 25 questions on this Quiz, and this PDF is 20 pages long. Be sure you have all 20 pages.

It is to your advantage to answer all 25 Questions. Your score is based on the number of correct responses, so there's no chance a blank response will be correct, and a guess might be, so you should definitely answer all of the questions.

### The Google Form Answer Sheet

All of your answers should be placed in the Google Form Answer Sheet, located at https://bit.ly/431-2021-quiz1-answer-sheet. All of your answers must be submitted through the Google Form by 9 PM on Monday 2021-10-04, without exception. The form will close at that time, and no extensions will be made available, so do not wait until Monday evening to submit. We will not accept any responses except through the Google Form.

The Google Form will contain places to provide your responses to each question, and a final affirmation where you'll type in your name to tell us that you followed the rules for the Quiz. You must complete that affirmation and then submit your results. When you submit your results (in the same way you submit a Minute Paper) you will receive an email copy of your submission, with a link that will allow you to edit your results.

If you wish to work on some of the quiz and then return later, you can do this by [1] completing the final question (the affirmation) which asks you to type in your full name, and then [2] submitting the quiz. You will then receive a link at your CWRU email which will allow you to return to the quiz as often as you like without losing your progress.

### Getting Help

This is an open book, open notes quiz. You are welcome to consult the materials provided on the course website and that we've been reading in the class, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love and the teaching assistants. You will be required to complete a short affirmation that you have obeyed these rules as part of submitting the Quiz.

If you need clarification on a Quiz question, you have exactly two ways of getting help:

1. You can ask your question in a private post on Piazza to the instructors. (This is the only kind of post you will be able to make on Piazza during the Quiz.)
2. You can ask your question via email to **431-help at case dot edu**.

While the complete Quiz is available, we will not answer questions about the Quiz except through the two approaches listed above. We promise to respond to all questions received before 5 PM on 2021-10-04 at that time, if not sooner.

A few cautions:

- Specific questions are more likely to get helpful answers.
- We will not review your code or your English for you.
- We will not tell you if your answer is correct, or if it is complete.
- We will email all students if we find an error in the Quiz that needs fixing.

### When Should I ask for help?

We recommend the following process.

- If you encounter a tough question, skip it, and build up your confidence by tackling other questions.
- When you return to the tough question, spend no more than 10-15 minutes on it. If you still don't have it, take a break (not just to do other questions) but an actual break.
- When you return to the question, it may be much clearer to you. If so, great. If not, spend 5-10 minutes on it, at most, and if you are still stuck, ask us for help.
- This is not to say that you cannot ask us sooner than this, but you should **never, ever** spend more than 20 minutes on any question without asking for help.

## Scoring and Timing

Most questions are worth 4 points, and the rest are worth either 2, 5 or 6 (as indicated), summing to 100 points. The questions are not in any particular order, and range in difficulty from "things I expect everyone to get right" to "things that are deliberately tricky".

The Quiz is meant to take 4-6 hours. I expect most students will take 3-8 hours, and some will take as little as 2 or as many as 10. It is not a good idea to spend a long time on any one question.

Dr. Love will grade all Quizzes, and you should have your result by class time on Thursday 2021-10-07.

## Writing Code

Occasionally, we ask you to provide a single line of code. If not otherwise specified, a single line of code in response should contain no more than one pipe, although you may or may not need the pipe in any particular setting. Moreover, you need not include the `library` command at any time for any of your code. Assume in all questions that all relevant packages have been loaded in R.

### The Data Sets

We have provided three data sets that are mentioned in the Quiz. You will find them in our Shared Google Drive in the Quiz 1 folder. They may be helpful to you.

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

### Packages used in building this document

This doesn't mean you need to use all of these packages or even that I used them all, but it does mean that I did not add any additional packages to this list in building the quiz or the answer sketch. You will note that the `mosaic` package is not listed here, but I did use the `favstats` function from that package using the `mosaic::favstats()` approach in preparing this material.

```
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())
```

# 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.

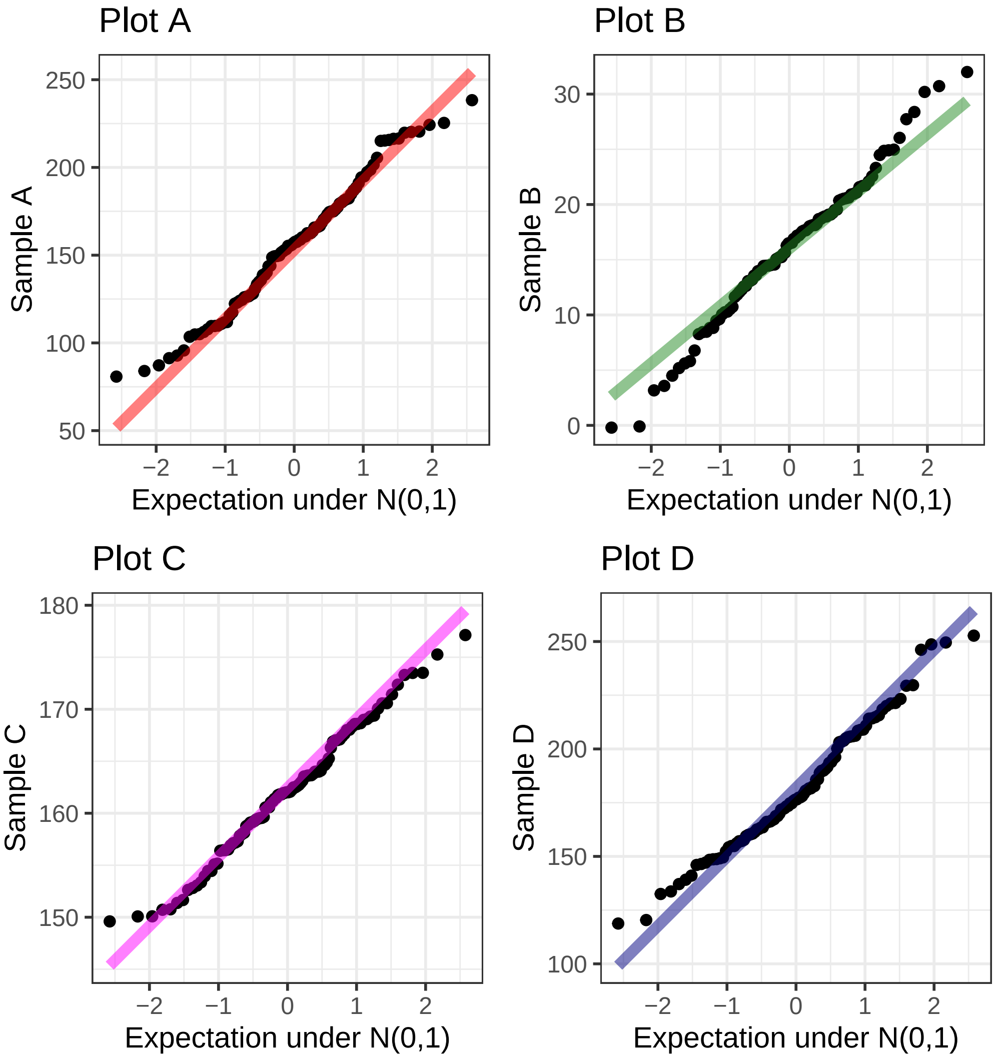
## 2 Question 02

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



# 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.

# 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.
```

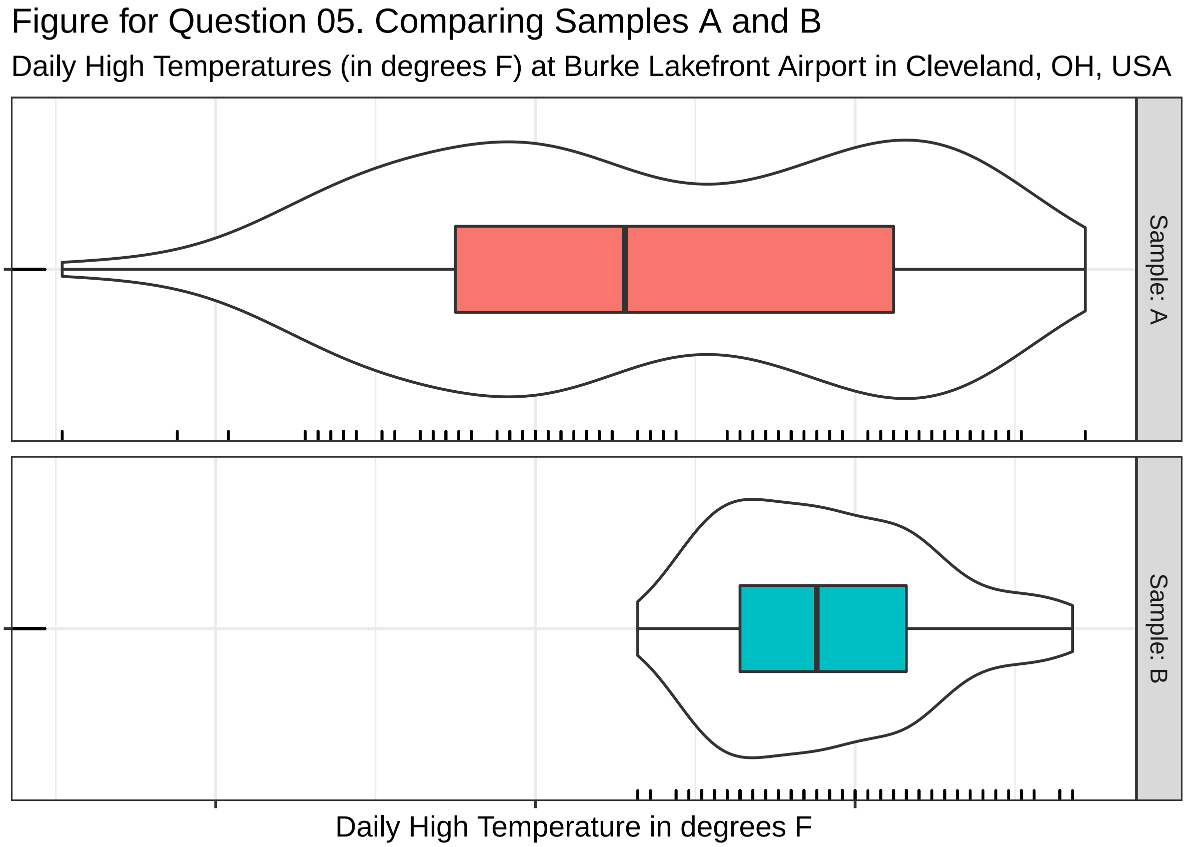
Each of the principles listed above is associated primarily with a particular step in the process of building inductive inference. Identify the step (a, b or c) in the process associated with each of the statements (I, II or III) above.

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

## 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.



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

# 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
```

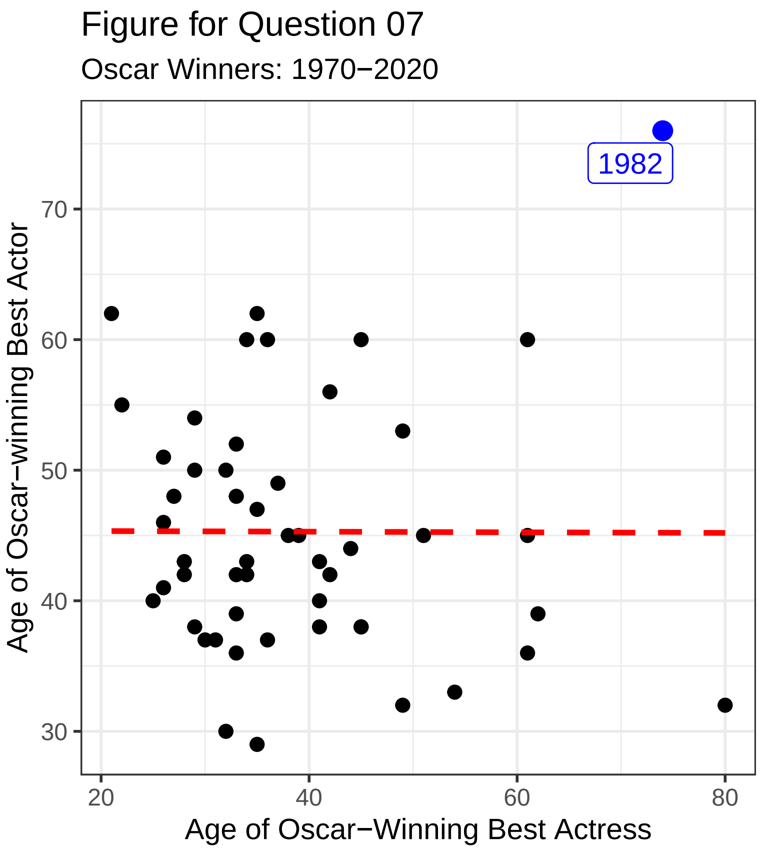
a. I only
b. II only
c. III only
d. I and II
e. I and III
f. II and III
g. All three statements
h. None of the three statements

# 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")
```



This question continues on the next page.

### Question 07 (continued)

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.

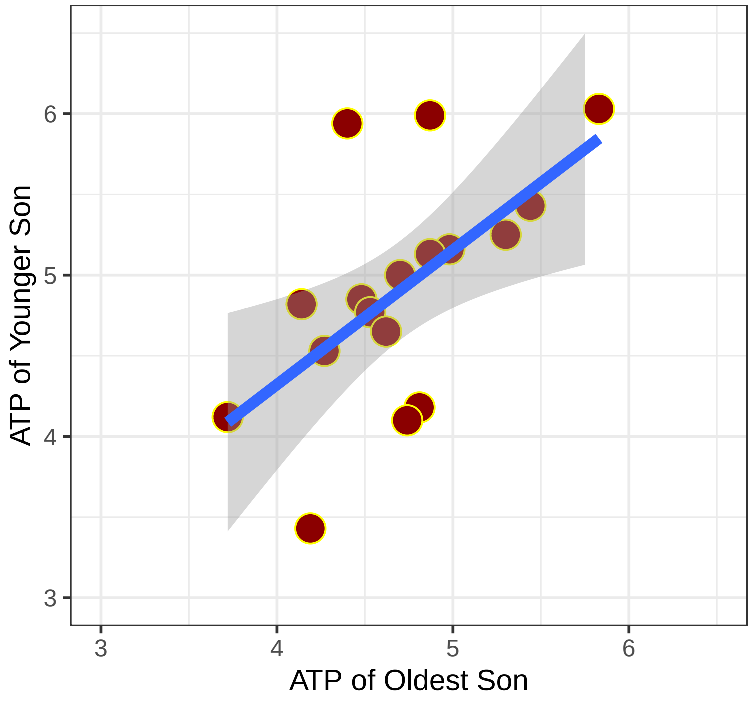
## 8 Question 08

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.

Figure for Question 08

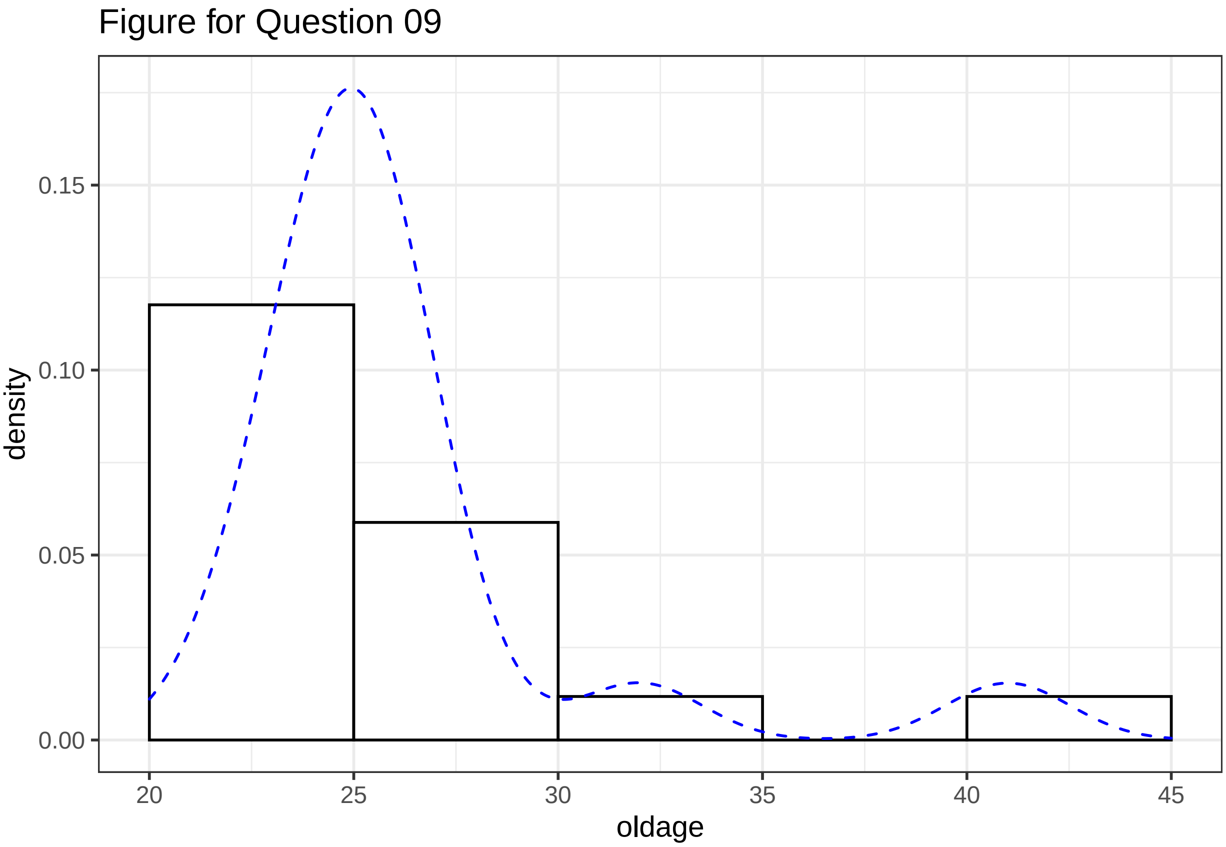


# 9 Question 09

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



# 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.

# 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 |

## 12 Question 12 (5 points)

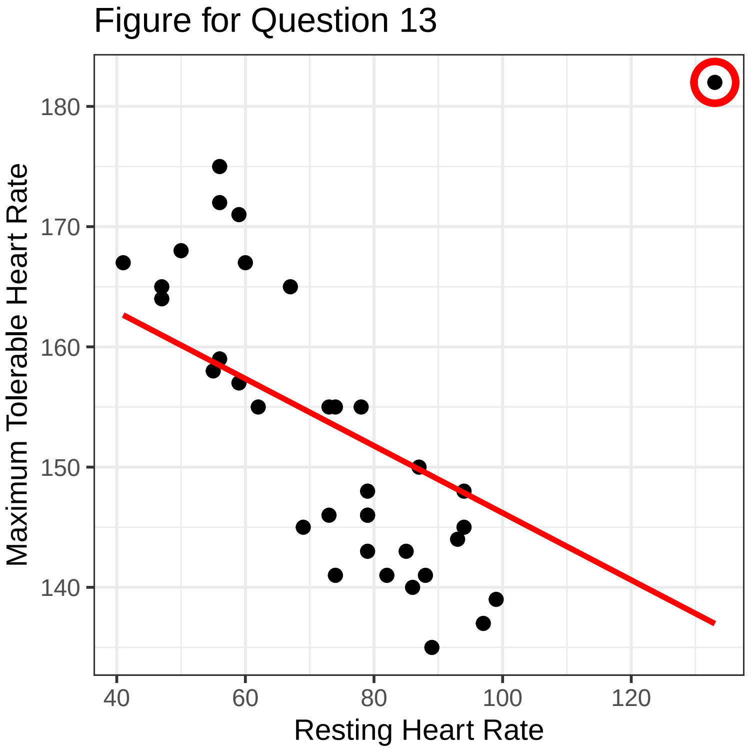
Suppose you have collected data into a tibble in R called `dat12`. The `dat12` data come from a cohort study to look at the impact of exposure to an industrial solvent (stored in the `solvent` variable: a factor taking on the values "none", "moderate" or "profound") on the probability of a bladder cancer 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.

## 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?

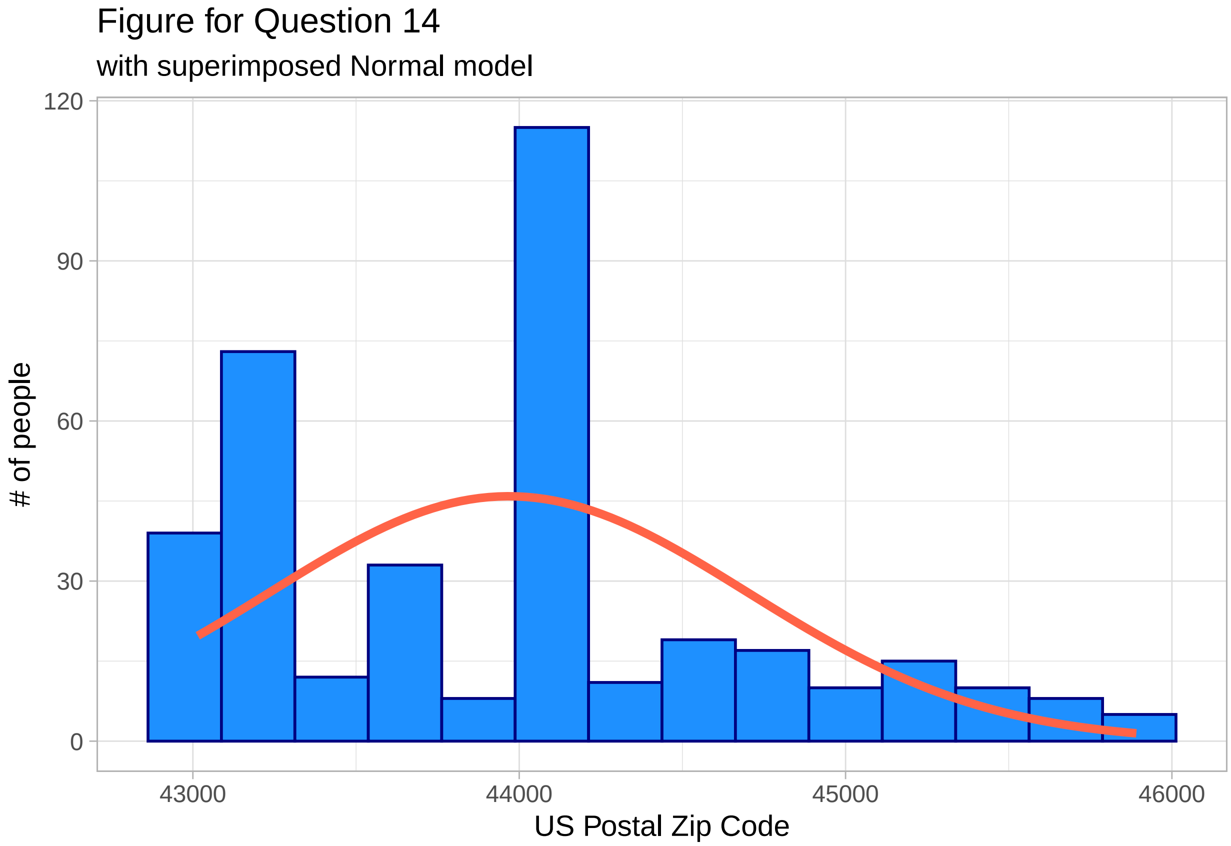
a. decrease
b. increase
c. remain unchanged
d. It is impossible to tell from the information provided.



# 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?

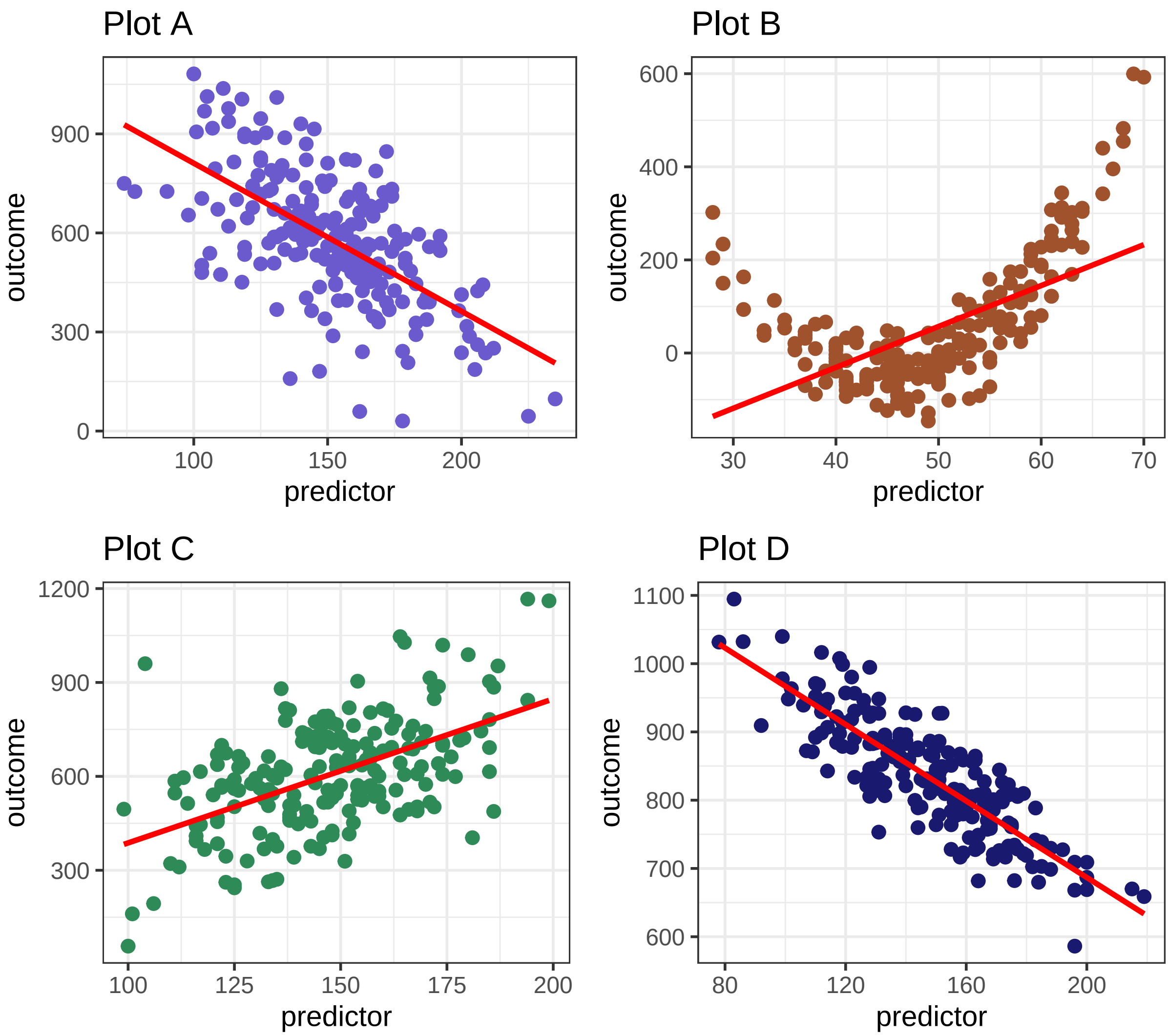
a. Mean
b. Median
c. Mode
d. IQR
e. It is impossible to tell from the information provided



## 15 Question 15

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

Figure for Question 15

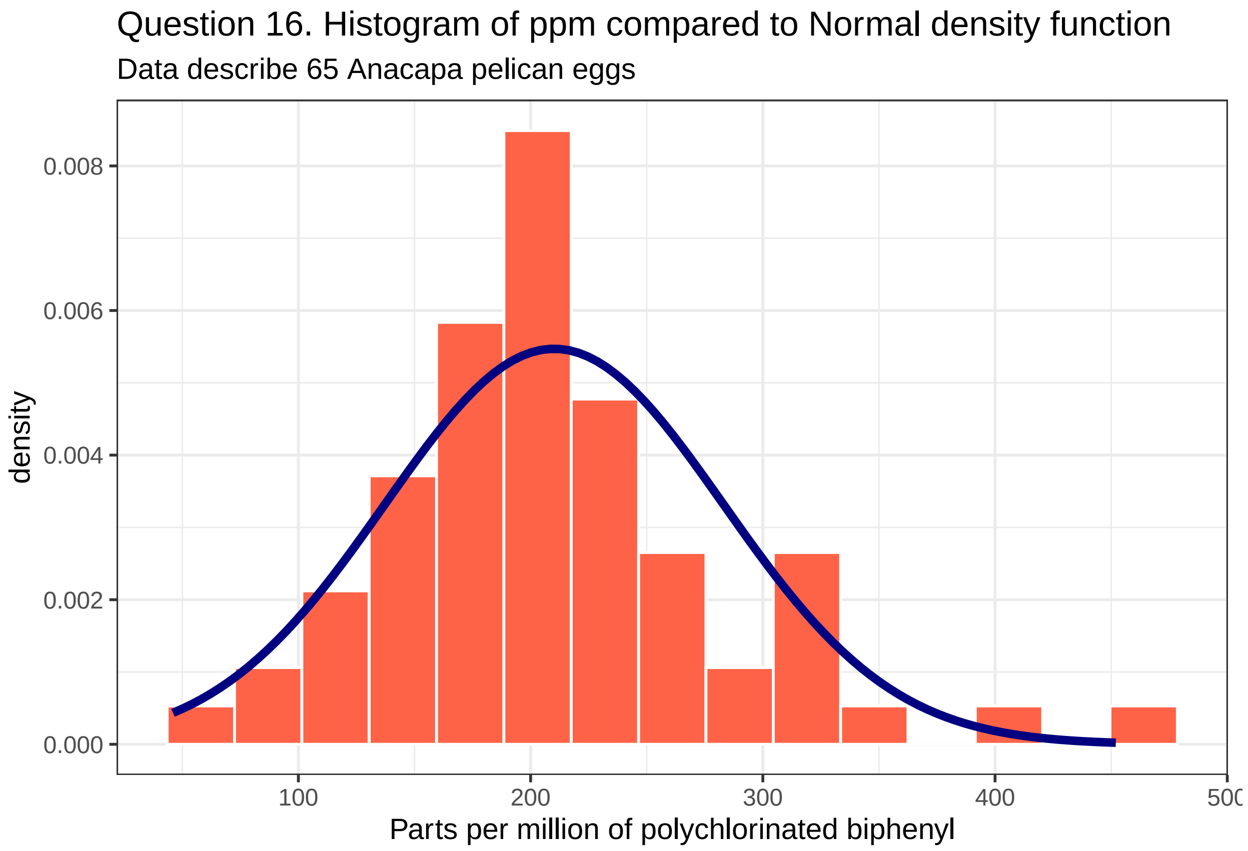


Which of the four scatterplots in the Figure for Question 15 is associated with a linear model for `outcome` using `predictor` that has the largest R-square value?

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

# 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.

### Question 16 (continued)

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

## 17 Question 17

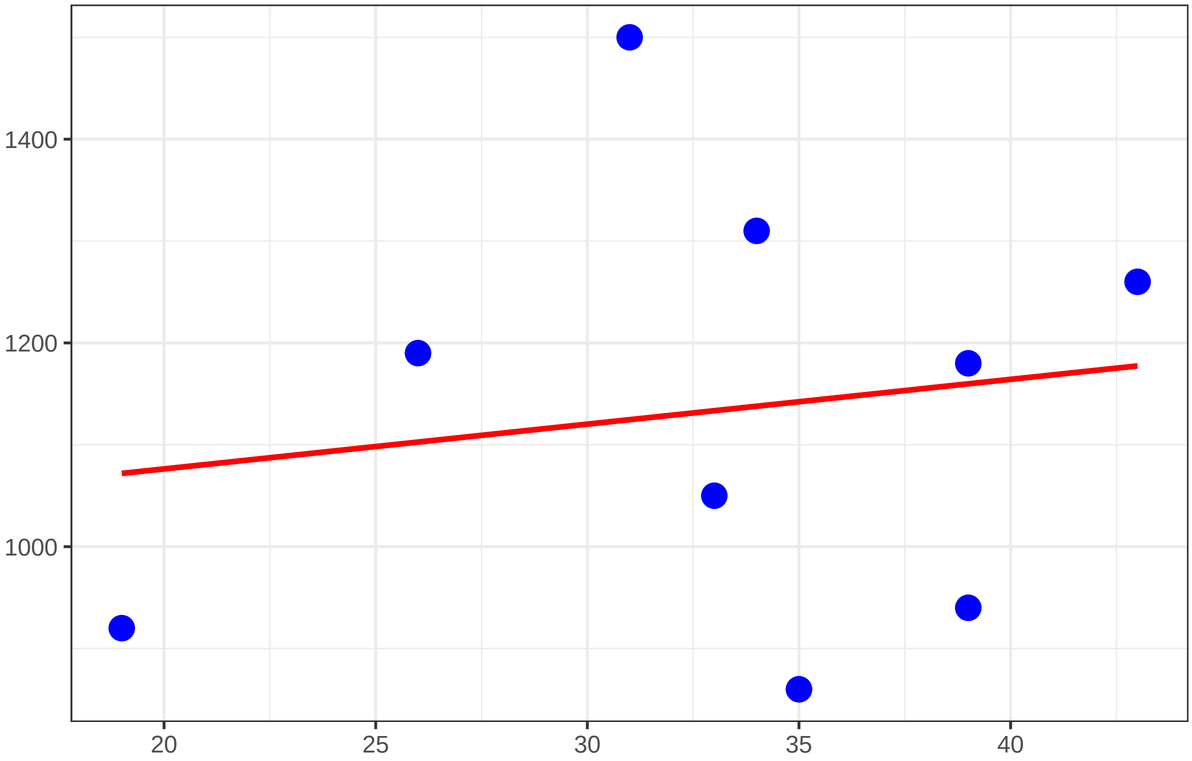
Suppose you are interested in how effectively shell thickness might be used to predict the concentration of environmental pollutants, in a setting like the study developed in Question 16. Which variable should go on the vertical (Y) axis of your scatterplot to display and model this association?

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.

## 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?

Figure for Question 18



# 19 Question 19

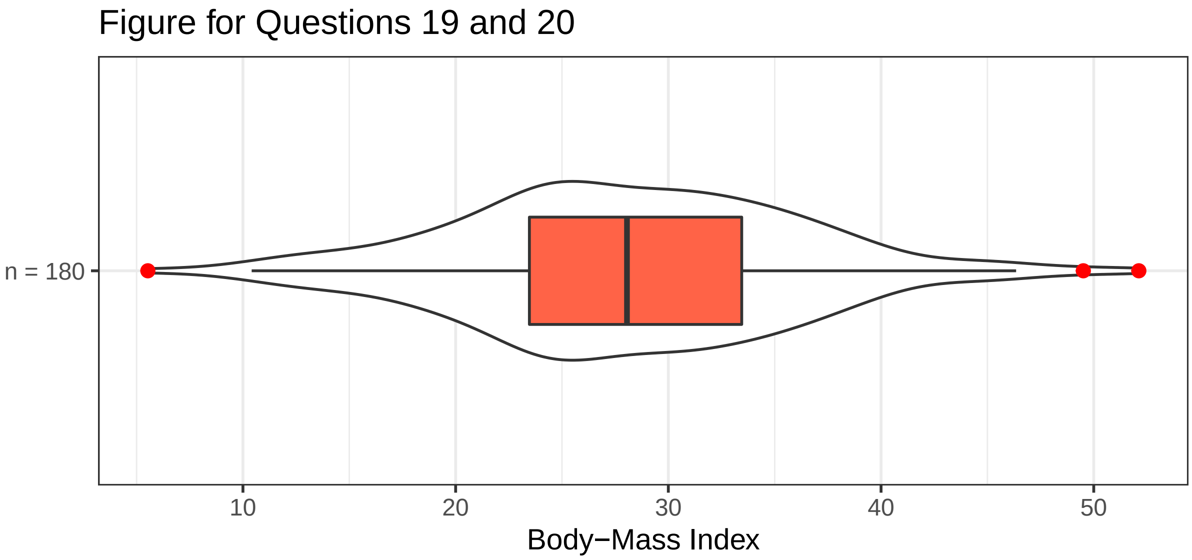
The Figure for Questions 19 and 20 displays the body-mass index (in $\frac{kg}{m^2}$) for 180 adults who suffer from rheumatoid arthritis, who are subjects in a new study. The mean BMI in the sample is 28.1 $\frac{kg}{m^2}$, the standard deviation is 7.9 $\frac{kg}{m^2}$ and there are no missing values.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



# 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.

# 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`?

- Your response should just be the number of characters.
- 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.

# 22 Question 22 (2 points)

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

- Again, your response should just be the number.

# 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`?

- Again, your response should just be the number.

# 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.

# 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
```

# THIS IS THE END OF THE QUIZ.