Stat 20: Problem Set 1

Solutions

Question 1

Below are the results of a study of conducted on a sample of 200 people in the city of Banjul, The Gambia. They sent an inspector to each sampled resident and asked subjects whether they slept under a bed net and whether or not they had a malarial infection at any point in the past. The results of the study are displayed below.

	Healthy	Infected
No net	40	60
Bed net	50	50

For each claim that follows, classify it as descriptive, predicitive, a generalization, or causal.

part a

"Our study demonstrates an association across Africa between the use of bed nets and the prevalence of malarial infection."

Generalization

part b

"Those in the study who slept under a bed net were 10 percent less likely to have had a malarial infection than those who did not."

Summary

part c

"Using bed nets leads to less likelihood of a malarial infection."

Causal Claim

Question 2

An online auction platform is rolling out a new feature called "Price Genie". When a seller inputs all of the information about the object they wish to sell, the site generates a dollar amount that is their suggested price that the seller start the bidding at. This dollar amount is calculated from data on past sales of similar items.

Identify the type of claim which is generated by the online auction platform.

Prediction

Questions 3-12

These questions will help you get started with coding!

Question 3

Open a .qmd document in RStudio. You can do this by going to the top left of your window and selecting File -> New File -> Quarto Document. A window will then pop-up. Give your document a title, toggle the output to HTML and then click "Create Empty Document" on the bottom left hand side of the window. *Fill in the blank below with your initials to testify that you completed this question.*

JS

Question 4

When you type in a qmd document, it will be output as plain text (not as code). You can organize your plain text into sections by using headers (# sign) and subheaders (## or more pound signs). The more pound signs you add, the smaller the headers become.

Type in, on a new line, "# Header" to make a heading. You need to take the space after the pound sign. Type in "## Subheading" to make the next largest heading (subheading). Play around with the different sized headers. *Fill in the blank below with your initials to testify that you completed this question.*

JS

Question 5

Use Ctrl + Alt + I on a Windows computer or Cmd + Option + I on macOS computer to toggle an R chunk in your qmd file. This will prepare you for the next section.

You will want to get in the habit of making many small chunks rather than one large chunk when working! *Fill in the blank below with your initials to testify that you completed this question.*

JS

Question 6

Create a line of code that adds the numbers 1 and 2 together and run it to get a result. *Write the code you used in the space below.*

```
1 + 2
```

Question 7

Now, write code that multiplies 3 and 4 together and saves the result into an object called `x`. Run it to get a result. *Write the code you used in the space below.*

```
x <- 3*4
```

Do you get an output on the screen like before? If not, where does the output "appear?"

I do not get an output on the screen like before. The variable `x` *and its value, 12, appear in the environment pane.*

Question 8

Create a vector with 4 numbers of your choosing and save it to the object `my_numbers`. *Write the code you used in the space below.*

```
my_numbers <- c(20,21,22,23)
```

Question 9

Create three lines of code:

- one which calculates the mean of my_numbers,
- one which calculates the sum of my_numbers,
- and one which calculates the maximum of my numbers.

If you don't know how to do these, you can look up statements online such as "how to calculate the maximum of a vector in R!" *Write all three lines of code you used in the space below.*

```
mean(my_numbers)
```

```
sum(my_numbers)
```

```
max(my_numbers)
```

Question 10

Create a vector containing the first names of the Stat 20 instructor and tutors in alphabetical order and save it to an object called `my_course_staff`. *Write the code you used in the space below.*

```
my_course_staff <- c("Chris", "Emma", "Evelyn","Jeremy")
```

Question 11

What you will notice throughout the semester is that when it comes to situations where categorical variables are plotted, R chooses to plot them in alphabetical order.

This is not always useful. Use the `factor()` function and the `levels` argument to reorder the levels of `my_course_staff` so that the instructor's name is first. Save this new,re-leveled vector into a new object. *Write the code you used in the space below.*

```
my_course_staff_ordered <-
  factor(my_course_staff, levels = c("Jeremy","Chris", "Emma", "Evelyn"))
```

Question 12

Using the `data.frame()` function, make a data frame with the vectors `my_numbers` and the re-made `my_course_staff` from **Question 11**. *Write the code you used in the space below.*

```
my_data_frame <- data.frame(my_numbers,my_course_staff)
```

Question 13

Below are the first few rows of a music-related data set I compiled. Columns three and four were compiled from Spotify and Twitter, respectively on July 10, 2022.

Artist	Genre	Listeners (in millions)	Followers (in millions)
Kendrick Lamar	Hip Hop	40	12
Drake	Hip Hop	66	39.5
Doja Cat	Pop	60	5.5
Harry Styles	Pop	73	38
Taylor Swift	Pop	57	90.5
The Weeknd	Pop	75	16
Luke Combs	Country	13	1

part a

What is the unit of observation in this data set?

A single musical artist.

part b

Identify each variable in this data set according to the Taxonomy of Data.

- Artist: *categorical nominal*
- Genre: *categorical nominal*
- Listeners: *numerical discrete*
- Followers: *numerical discrete*

Remember that discrete values take jumps. Even though listeners and followers are being measures in millions, which allows the use of a decimal point, there is some point at which you cannot add any more decimal digits because you would be dividing a person. At this point, the values start taking jumps. This is the essence of a numerical, discrete variable.

Questions 14-16

Here is a contingency table of college students with their **Favorite Color** (Red or Blue) down the columns and their **School** (Berkeley or Stanford) across the rows.

	Red	Blue
Berkeley	10	90
Stanford	60	40

Question 14

Find the proportion of all students who attend Berkeley. What type of proportion is this?

The proportion in question is $\frac{100}{200} = \frac{1}{2}$. This is a *marginal proportion.*

Question 15

Find the proportion of all students who attend Berkeley that like red best. What type of proportion is this?

The proportion in question is $\frac{10}{200} = \frac{1}{20}$. This is a *joint proportion.*

Question 16

Of the students who attend Berkeley, find the proportion that like red best. What type of proportion is this?

The proportion in question is $\frac{10}{100} = \frac{1}{10}$. This is a *conditional proportion.*

Questions 17-19

The following questions are based on a dataset containing information on Google Reviews for a number of Berkeley restaurants. Here are the first few rows of this dataset.

Some other notes:

- `neighborhood` can be one of `"Downtown"`, `"North Berkeley"` or `"Southside"`.
- `rating` is on a scale from 1 to 5.

`name`	`neighborhood`	`cuisine`	`reviews`	`photos`	`rating`
Berkeley Social Club	Downtown	Korean	920	1319	4.4
Cheese Board	North Berkeley	Pizza	2941	1891	4.8
Top Dog	Southside	Hot Dog	1056	381	4.6

Question 17

Say I want to create a two-variable visualization with:

- the `reviews` variable
- one other variable in the dataset.

Which of the following plots cannot be used for this purpose? *Circle the correct choice(s).*

Correct answer(s) in bold.

a. **Stacked, normalized bar chart**

b. Side-by-side Box Plots

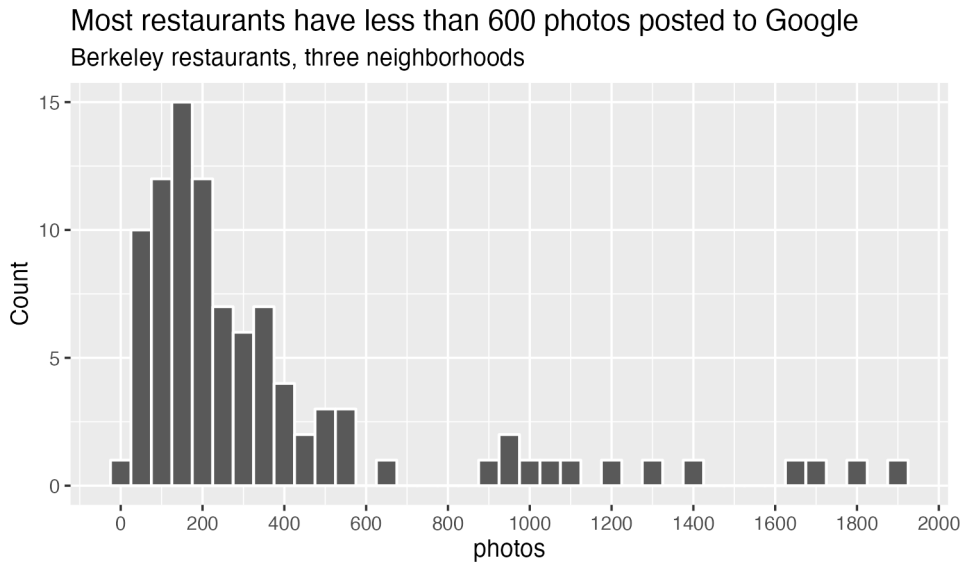
c. Side-by-side Violin Plots

d. Scatter Plot

Technically, we will not introduce a scatter plot until later in the course. A scatter plot is a type of visualization that works with two numerical variables. You will not be tested on this particular plot for your first quiz.

Question 18

The following is a histogram of the distribution of photos posted to Google among the restaurants in the dataset.



Based off of the plot, which measure of center would be **least** representative of a typical observation of the data? *Circle the correct choice.*

Correct answer in bold.

a. Mode (pick the bin with the most observations)

b. **Mean**

c. Median

Question 19

Based off of the same plot, Which measures of spread would be **most** representative of a typical observation of the data? *Circle the correct choice(s).*

Correct answer(s) in bold.

a. Standard Deviation

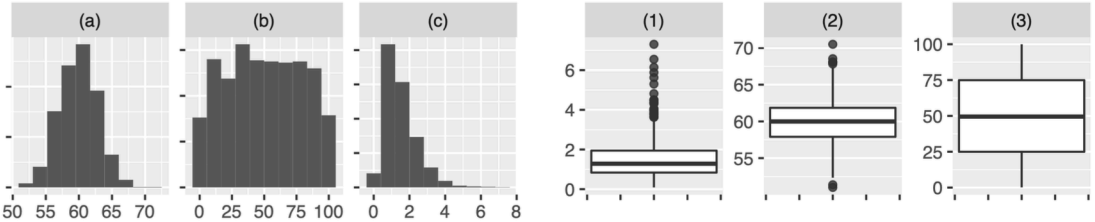
b. **Median Absolute Deviation**

c. **IQR**

d. Range

Question 20

Circle the correct matching between each of the three distributions represented as a histogram and as a boxplot.



Correct matching in bold.

a. a-1, b-2, c-3
b. a-2, b-1, c-3
c. a-3, b-2, c-1
d. a-1, b-3, c-2

e. **a-2, b-3, c-1**

f. a-3, b-1, c-2

Question 21

As discussed in the lecture notes, this is not a data frame:

Handed-ness / Sex	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

It does, however, depict a data set, just in a different format. Sketch the data that is summarized here but structured as a data frame. Think through: what was the unit of observation? What were the variables? How many rows are there? How many columns? (you need not fill out the entire data frame; just a schematic)

You should sketch a data frame that contains 100 rows and 2 columns.

- The unit of observation is a single person.
- The two variables are *handedness* (left or right) and *sex* (Male or Female).

Questions 22-23

RMS Titanic was a British passenger liner with 2,224 people aboard (including both passengers and crew). The Titanic sank in the North Atlantic Ocean on 15 April 1912 after striking an iceberg on her way to New York City. It was the deadliest sinking of a single ship at the time with almost 70% of the 2,224 passengers and crew dying.

Below are the first five rows of 2,224 for a data frame called `titanic`.

Name	Sex	Age	Fare	Class	Survived
Braund, Mr. Owen Harris	male	22	7.25	3	FALSE
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	71.28	1	TRUE
Heikkinen, Miss. Laina	female	26	7.92	3	TRUE
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	53.10	1	TRUE
Allen, Mr. William Henry	male	35	8.05	3	FALSE

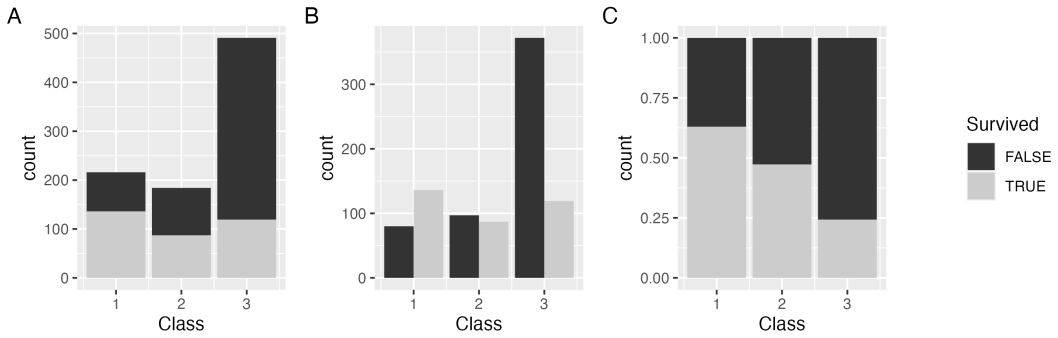
Question 22

What is the unit of observation for `titanic`?

A single passenger on the Titanic.

Question 23

The three graphs below were generated using `titanic`.



Which graph is **best** used to answer the following questions? *Write the letter associated with the graph to the right of each question*

How many first class passengers survived the Titanic's sinking? _____**B**_____

Is there an association between a passenger's class and their survival? ***C***_

What is the mode of the `Class` variable? ***A***_

Questions 24-25

Open up RStudio and write down the code you used to complete each of the following questions in the space below. *Make sure you load in any libraries where necessary!*

Question 24

Consider the vector `q6` which was made as follows:

```
q6 <- c(1,2,3,4,5,6,NA)
```

Load `q6` as is into your session. Then, write R code to calculate the mean of `q6` so that the result is 3.5 (this is the mean of the numbers 1 through 6). *Hint: how can you learn more about the function `mean()`?*

```
mean(q6, na.rm = TRUE)
```

Question 25

The `promote` dataset can be found in the `stat20data` package. Using this data, write `ggplot2` code to make a stacked, normalized bar chart having identified gender on the x axis, with the bars being filled in by promotion decision. Write the code you used below. Then, make a claim about the association between the two variables (we will revisit this study in more detail later in the course)!

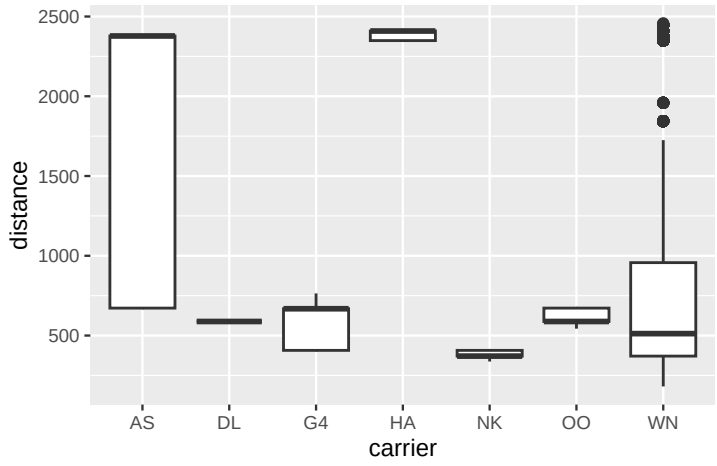
Make sure you have loaded the `stat20data` and `tidyverse` libraries using the `library()` function

```
ggplot(data = promote,
       mapping = aes(x = gender,
                     fill = decision)) +
  geom_bar(position = "fill")
```

Claim: *Resumes in the study that had traditionally male names were associated with more positive promotion outcomes.*

Question 26

Below is a smaller version of the data from a future lab called `flights_mini`. It contains all flights out of Oakland (OAK) from December 2020. This data frame is used to create the plot that follows. `distance` refers to the distance a given plane travels on its flight, measured in miles. `carrier` refers to the carrier code for a specific airline.



Which of the following interpretations of the plot above are true? *Circle all that apply.*

Correct choice(s) in bold.

A. The carrier with the most heavily skewed distance distribution is HA.
B. The median distance of the flights operated by DL, G4, and OO are roughly equivalent.
C. The minimum distance traveled in this data set is roughly 200.
D. There is no clear association between the carrier and the distance of their flights.
E. The carrier with the greatest variability in distance, as measured by the IQR, is AS.

Questions 27-28

The `mpg` dataset is available as a part of the `tidyverse` library. It contains information on fuel consumption for 38 models of car between 1999 and 2008. *Datasets can have help files, too!* You do not need to include code for loading in libraries or accessing help files in your answers to the below questions.

Question 27

Write `dplyr` code to calculate the median and IQR city miles per gallon for the vehicles in the dataset and copy it below. The result of your code should be one data structure.

```
summarise(mpg,
          cty_median = median(cty),
          cty_IQR = IQR(cty))
```

Question 28

Write `dplyr` code to calculate the mean and standard deviation city miles per gallon for the vehicles in the dataset *for each class of car* and copy it below. The result of your code should be one data structure.

```
grouped_mpg <- group_by(mpg, class)
summarise(grouped_mpg,
          cty_mean = mean(cty),
          cty_sd = sd(cty))
```