Problem Set 1

due Thursday, June 20 at 11:59pm

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."

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."

part c

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

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.

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.*

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.*

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.*

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.*

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.*

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

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.*

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.*

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.*

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.*

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 13**. *Write the code you used in the space below.*

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?

part b

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

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?

Question 15

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

Question 16

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

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).*

a. Stacked, normalized bar chart

b. Side-by-side Box Plots

c. Side-by-side Violin Plots

d. Scatter Plot

Question 18

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

Most restaurants have less than 600 photos posted to Google
Berkeley restaurants, three neighborhoods
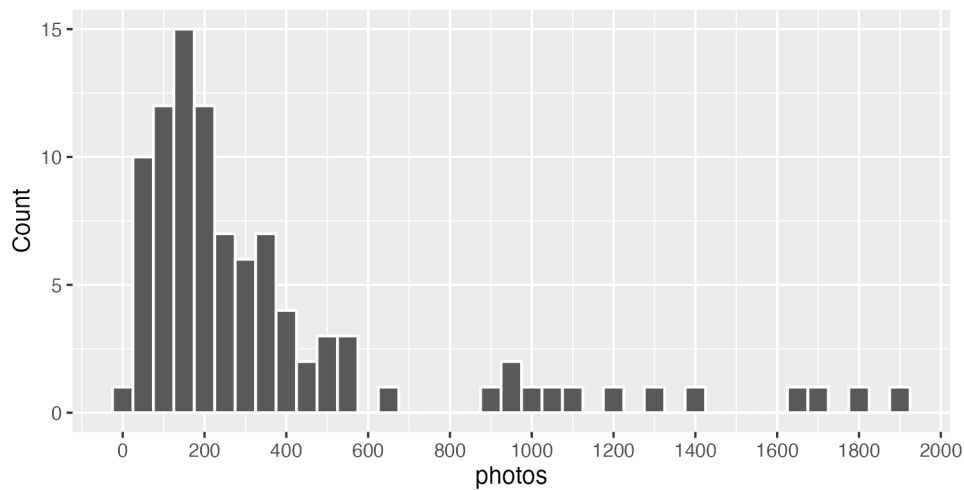
Based off of the plot, which measure of center would be **least** representative of a typical observation of the data? *Circle the correct choice.*

    a. Mode (pick the bin with the most observations)
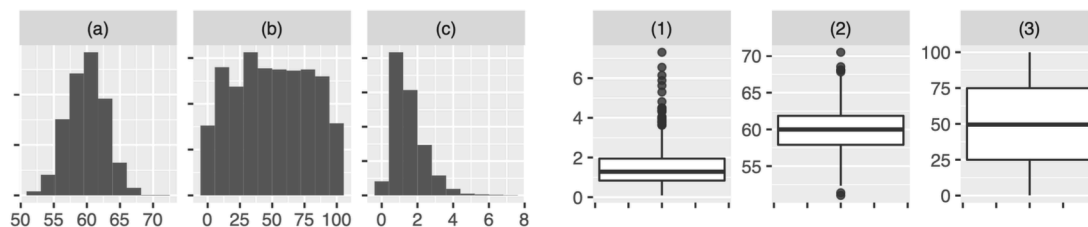   b. Mean
    c. Median

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).*

    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.

    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