Homework 1 SOLUTIONS

P8130 Fall 2022

Due: September 23, 2022 at midnight EST

P8130 Guidelines for Submitting Homework

- Your homework must be submitted through Courseworks. No email submissions!
- Only one PDF file should be submitted, including all derivations, graphs, output, and interpretations. When handwriting is allowed (this will be specified), scan the derivations and merge ALL PDF files (http: //www.pdfmerge.com/).
- You are encouraged to use R for calculations, but you must show all mathematical formulas and derivations. Please include the important parts of your R code in the PDF file but also submit your full, commented code as a separate R/RMD file.
- To best follow these guidelines, we suggest using Word (built in equation editor), R Markdown, Latex, or embedding a screenshot or scanned picture to compile your work.

DO NOT FORGET: You are encouraged to collaborate on homeworks, explain things to each other, and test each other's knowledge. But Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

Problem 1 (5 points)

Please classify each of the following variables as qualitative (specify if binary, nominal, or ordinal) or quantitative (specify if discrete or continuous):

a) homework feedback, labeled as "poor", "fair", "good", "very good"

Qualitative, ordinal

b) homework feedback, labeled as "fail", "pass"

Qualitative, binary

c) country of birth

Qualitative, nominal

d) the quantity of grapes (in lbs) to make 3 liters of wine

Quantitative, continuous

e) number of TAs in the P8130 course

Quantitative, discrete

Problem 2 (15 points)

In a study of 133 individuals with a recent bike crash history, depression scores were measured using a standardized test. The depression scores for 14 of these individuals are as follows:

45, 39, 25, 47, 49, 5, 70, 99, 74, 37, 99, 35, 8, 59

Note: this can be done by hand or using R. R code is shown in the solutions.

a) Compute the following descriptive summaries of these data: mean, median, range, SD.

```
# import the data as a vector
bike_dep <- c(45, 39, 25, 47, 49, 5, 70, 99, 74, 37, 99, 35, 8, 59)
```

```
# mean depression score
mean(bike_dep)
```

```
[1] 49.35714
```

```r
# median depression score
median(bike_dep)
```

```
[1] 46
```

```r
# range of depression scores
max(bike_dep) - min(bike_dep)
```

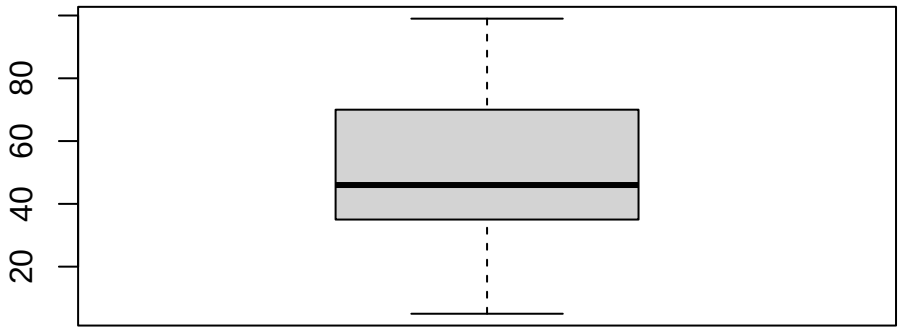
```
[1] 94
```

```r
# standard deviation of depression scores
sd(bike_dep)
```

```
[1] 28.84603
```

b) Describe the box plot and the underlying distribution of the data. Use some of the following terms: left-skewed, right-skewed, symmetric, bimodal, unimodal distribution.

```r
# boxplot of depression scores
boxplot(bike_dep)
```



This boxplot does not indicate any skewness - it is mostly symmetric. There are no outliers.

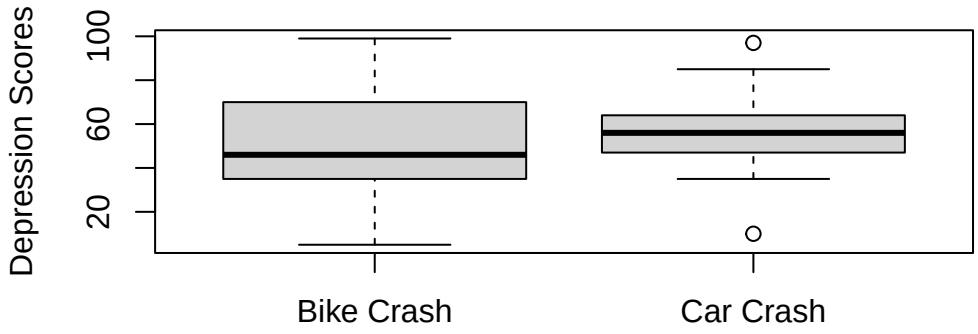
Additionally, 140 individuals with a recent car crash history also participated in the study. The depression scores for 13 of these individuals are given below:

67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50

a) Using R, make a side-by-side box plot of the depression scores stratified by type of accident. Make sure you label your figure appropriately.

```
# import data as a vector
car_dep <- c(67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50)

# make side by side boxplots
boxplot(
  bike_dep,
  car_dep,
  names = c("Bike Crash", "Car Crash"),
  ylab = "Depression Scores"
)
```



b) Describe each of the box plots and the underlying distribution of the data. Use some of the following terms: left-skewed, right-skewed, symmetric, bimodal, unimodal distribution.

- The bike crash boxplot is described above.
- The car crash boxplot is also relatively symmetric, with two outliers (one on either side).
- The range (max - min) for each situation (bike/car) is approximately equal, however, the variance in the car crash group will likely be smaller since the IQR (Q3 - Q1) is much narrower in the car crash group than the bike crash group. (We can check this by calculating the standard deviation of each group (below) – as we suspected, the SD of the car crash group is smaller than that of the bike crash group.)

```
# standard deviation of bike crash group
sd(bike_dep)
```

```
[1] 28.84603
```

```
# standard deviation of car crash group
sd(car_dep)
```

```
[1] 21.58139
```

- The median for the bike crash group is lower than the median of the car crash group.
- We cannot be sure if these distributions are unimodal or bimodal - a histogram would be the better figure to make that judgement.

c) Comparing the 2 box plots, which group appears to have a lower typical depression score?

The median for the bike crash group is lower than the median for the car crash group, so we could conclude that the bike crash group appears to have a lower typical despression score. (Since we decided both of these distributions are relative symmetric, the mean and median will be very close together.)

Problem 3 (10 points)

Suppose we toss one fair 12-sided die:

a) Let's define the event A as "an even number appears". What is the probability of the event A?

$$P(A) = \frac{6}{12} = \frac{1}{2}$$

b) Let's define the event B as "number 10 appears". What is the probability of the event B?

$$P(B) = \frac{1}{12}$$

c) Compute P(B U A).

$$P(B \cup A) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{12} - \frac{1}{12} = \frac{1}{2}$$

d) Are events A and B independent? Why? Prove your answer.

No. The following holds true for independent events: $P(A \cap B) = P(A) \times P(B)$. In this scenario, $P(A \cap B) = \frac{1}{12}$ and $P(A) \times P(B) = \frac{1}{2} \times \frac{1}{12}$ which does not equal $\frac{1}{12}$.

Problem 4 (10 points)

5% of women above age of 75 have dementia. Among women (75+ years old) with dementia, 80% have positive findings on their CT scan. Among women (75+ years old) who don't have dementia, 10% will have a positive CT scan findings. A randomly-selected woman (75+ years old) had a positive CT scan findings.

What is the probability that she actually has dementia? Compute by hand and show the key steps. The answer can be hand written.

Let "woman having dementia" be event D, which means that "woman not having dementia" is event D^c. Let "woman with positive CT findings" be event T, and "woman with negative CT findings" be event T^c.

From the question, we can define the following:

$$P(D) = 0.05 \qquad P(T|D) = 0.80 \qquad P(T|D^c) = 0.10$$

From which follows:

$$P(D^c) = 0.95$$

We want to estimate the probability a woman with positive CT findings has dementia, or $P(D|T)$. We can apply Bayes' Theorem:

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D) \times P(D)}{P(T|D) \times P(D) + P(T|D^c) \times P(D^c)} = \frac{0.80 \times 0.05}{0.80 \times 0.05 + 0.10 \times 0.95} = 0.296$$

The probability of a woman with positive CT findings having dementia is 29.6%.

Note: this problem could also have been solved with a two way table or tree diagram.