HSCI 50 LAB 1: Exploratory Data Analysis ANSWER KEY

INSTRUCTIONS

The data set that we will be using for the entire course is resampled data from the MIT COVID-19 Beliefs, Behaviors & Norms Survey (https://covidsurvey.mit.edu/api.html). This is a multi-country, online survey that examined different COVID-19 perceptions across time, from July 6, 2020 to March 28, 2021. We will be using data from the Philippines aged 20 to 60.

In this lab, you will learn how to draw by hand and interpret some common exploratory data analyses.

You have many options to submit this worksheet. Either you work on this by hand and scan/take a clear photo of your submission and save as PDF, or type your responses in a Word processor or PowerPoint presentation. You do not have to copy the questions again, but please number them accordingly.

1. Identify the level of measurement of each of the variables in this data set. Your choices are: nominal, ordinal, discrete, continuous

a. *gender*: Takes the values "male" or "female" (This survey was not powered beyond heteronormative gender norms)
b. *age*: Age in years, expressed as an integer
c. *age_grp*: Age group, expressed in 10 year intervals (20-30, 31-40, 41-50, 51-60)
d. *response*: Response to the question, "If a vaccine for COVID-19 becomes available, would you choose to get vaccinated?" The responses are: Yes, No, Don't Know, Already vaccinated

The answers are as follows:

a. *gender*: This is nominal since there are two categories that are not ordered
b. *age*: This is discrete because the values are expressed as integers. If the age was expressed as a decimal value, then this would have been continuous.
c. *age_grp*: This is ordinal because there are four categories that are ordered by age. This could also be a discrete variable if it was analyzed as discrete, given that this is an equal interval ordinal variable. The variable could be something like "decade of age" instead.
d. *response* This is nominal since there are four categories that are not ordered

2. The following data is a listing of 20 randomly drawn ages from the full data set.

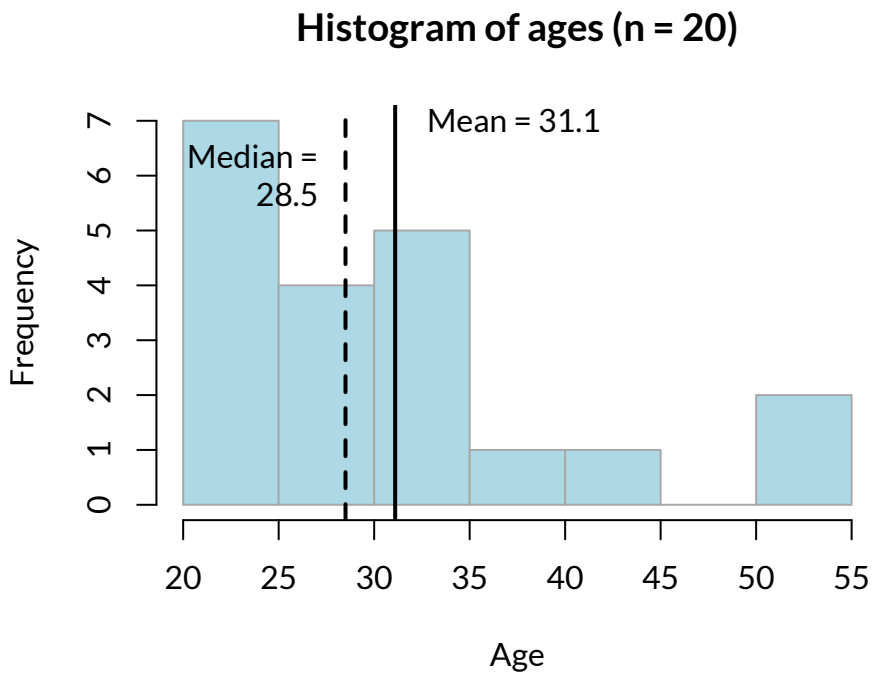
```
## 30 44 25 27 20 27 24 51 31 25 33 32 21 33 53 21 40 24 27 34
```

By hand, draw a stem-and-leaf plot where the stem is the tens digit and the leaf is the ones digit.

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 | 0114455777
##   3 | 012334
##   4 | 04
##   5 | 13
```

3. Describe what the skewness will look like if the same set of data was visualized as a histogram. Where is the mean relative to the median?

The histogram would be positively- (or right-) skewed, meaning that there is a long tail of higher values. We expect the mean to be higher (or on the right of the histogram) relative to the median.



4. From that list of data, calculate by hand (you may use a calculator) the mean.

```
## 30+44+25+27+20+27+24+51+31+25+33+32+21+33+53+21+40+24+27+34 = 31.1
```

5. From that list of data, calculate by hand (you may use a calculator) the parts of a box-and-whisker plot.

a. median
b. lower hinge (first quartile or Q1)
c. upper hinge (third quartile or Q3)
d. interquartile range (IQR)
e. 1.5 * IQR
f. lower fence
g. upper fence
h. lower whiskers
i. upper whiskers
j. values of outliers (if any)

We can use the stem-and-leaf plot to identify the median, lower hinge and upper hinge. Then we can derive the rest. Remember that we have 20 observations.

a. When we order the numbers from lowest to highest, the median should be the ((20+1)/2) = 10.5th observation. Therefore, the median is the value between the 10th and 11th observation. The 10th observation is 27, the 11th is 30. Therefore, the median is (27+30)/2 = **28.5**.

b. The lower hinge or the first quartile is the middle value of the first 10 observations. Therefore, the first quartile is the value between the 5th and 6th observation. The 5th observation is 24, the 6th is 25. Therefore, the first quartile is (24+25)/2 = **24.5**. If you got an answer of 24.75 or 24.25, that's also correct. The sample size here is very low so we expect the values to vary depending on the formula used. Remember this difference of 0.25 isn't really meaningful, considering our age variable increases in increments of 1 at the minimum.

c. The upper hinge or the third quartile is the middle value of the last 10 observations. Therefore, the third quartile is the value between the 15th and 16th observation. The 15th observation is 33, the 16th is 34. Therefore, the third quartile is (33+34)/2 = **33.5**. If you got an answer of 33.25 or 33.75, that's also correct.

   Note that the following values might be different if you had a different answer for (b) and (c). The answers below are based on a lower hinge of 24.5 and upper hinge of 33.5.

d. The IQR can be derived from (b) and (c). IQR = Q3 - Q1 = 33.5 = 24.5 = **9**.

e. 1.5 * IQR = 1.5 * 9 = **13.5**.

f. The lower fence is Q1 - 1.5 x IQR = 24.5 - 13.5 = **11**.

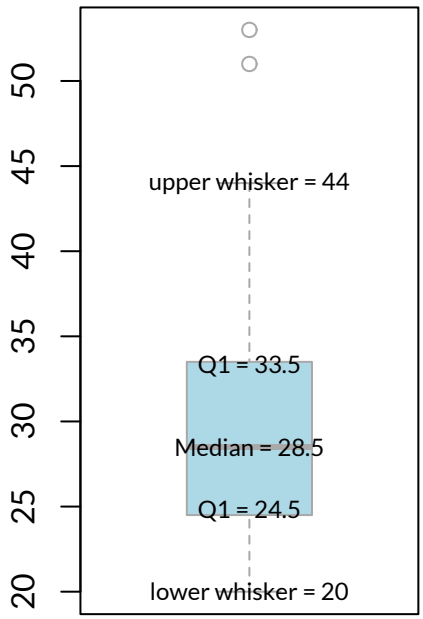
g. The upper fence is Q3 + 1.5 x IQR = 33.5 + 13.5 = **47**.

h. The lower whisker is the lowest value in the data set that is higher than the lower fence. In our data, that value is **20**.

i. The upper whisker is the highest value in the data set that is lower than the upper fence. In our data, that value is **44**.

j. There are no outliers below the lower fence, but there are two outliers above the upper fence. These are **51** and **53**.

The boxplot should look something like this:



6. Now let's start working with the full data set. Given the following frequency counts for each age group, calculate by hand (you may use a calculator) the cumulative frequency, relative frequency, and cumulative relative frequency. Express relative frequencies as percentages to the nearest tenths of a percentage point, e.g. 10.2% for 0.10173. (Note there was a minor typo in the worksheet where the rounding off instruction was incorrect)

It is highly recommended you do not round off values until the very end. So when calculating the relative frequencies, keep two or three extra decimal places, then round them off once you are done populating the table.

Age group	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
20-30	673			
31-40	376			
41-50	234			
51-60	192			

Your table should look something like this:

Age group	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
20-30	673	673	45.6%	45.6%
31-40	376	1049	25.5%	71.1%
41-50	234	1283	15.9%	87.0%
51-60	192	1475	13.0%	100.0%

7. The following table below summarizes the measures of central tendency and dispersion for the age variable, disaggregated by gender. Answer the questions that follow.

Measure	Male	Female
Mean	36.4	33.6
Standard deviation (SD)	11.7	10.7
Variance	137.9	114.2
Median	34.5	30
First Quartile (Q1)	26.8	25.0
Third Quartile (Q3)	45.0	40.0
Interquartile Range (IQR)	18.2	15.0
Mode	20	27

a. Between males and females, which age distribution is more dispersed? Why?
b. What is the skewness of the age distribution for males? What about females?
c. Between males and females, which box-and-whisker plot has wider fences?

The answers are as follows:

a. Looking at either the standard deviation or the interquartile range, male ages are more dispersed than female ages in our data set.
b. For both males and females, mean > median, therefore the data are positively (or right-) skewed.
c. We can calculate the fences for each by hand:

- For males: lower = 26.8 - 1.5 x 18.2 = -0.5 ;upper = 45.0 + 1.5*18.2 = 72.3; width = 72.3 - (-0.5) = 72.8
  - Note that for males, a lower fence of -0.5 might not make a lot of sense (how are negative ages possible?) Take this value as an indication of how spread out the data are.
- For females: lower : 25.0 - 1.5 x 15.0 = 2.5 ;upper = 40.0 + 1.5*15.0 = 62.5; width = 62.5 - 2.5 = 60
- Therefore the male box-and-whisker plot has a wider fence.

END OF LAB