HSCI 50 LAB 1: Exploratory Data Analysis

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

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.

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?

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

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)

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. 10.1% for 0.10173.

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

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?

END OF LAB