2B03 Assignment 1*

Descriptive Statistics (Chapters 1 & 2)

Yusef Elsherif 400543068

2024-09-19

Instructions: *You are to use Quarto Markdown for generating your assignment output file. You begin with the Quarto Markdown script downloaded from A2L, and need to pay attention to information provided via introductory material posted to A2L on working with R, Quarto Markdown, and downloading data from ODESI. Having downloaded all necessary files, placed them in the same folder/directory, and added your answers to the Quarto Markdown script, you then are to generate your output file using the "Render" button in the RStudio IDE and, when complete, upload both your Quarto Markdown file and your PDF file to the appropriate folder on A2L.*

1. Define the following terms in a sentence (or *short* paragraph) and state a formula if appropriate; cite your source, e.g., "2B03 Lecture Slides" (this question is worth 5 marks).

    a. Categorical Data

    Categorical, also known as qualitative, data refers to non numeric data. Source: ECON 2B03 Introduction & Chapter 1 Slides

    b. Frequency Distribution

    Frequency distributions are different ways of summarizing data according to the number of frequencies within a given interval. There are different types of frequency distribution, which follow different rules, including absolute, cumulative, relative, and cumulative relative frequency distributions. This data can be presented in graphical or tabular formats. Source: ECON 2B03 Chapter 1 & 2 Slides

    c. Sturges' Rule

    Sturges' rule is a method used to determine the desirable number of classes for a distribution.

*DESKTOP-GMDSDEO, x86-64, Windows 10

$$K = 1 + 3.3log_{10}n$$

"n" is the sample size, the resulting integer K rounded to the nearest whole number would be the desirable number of classes for that distribution. Source: ECON 2B03 Chapter 1 & 2 Slides

d. Cross Tabulations

Cross tabulations are tabular summaries for two categorical variables, in which one variable is represented by row heads and the other is represented by column heads. Source: ECON 2B03 Chapter 1 & 2 Slides

e. Sample Median

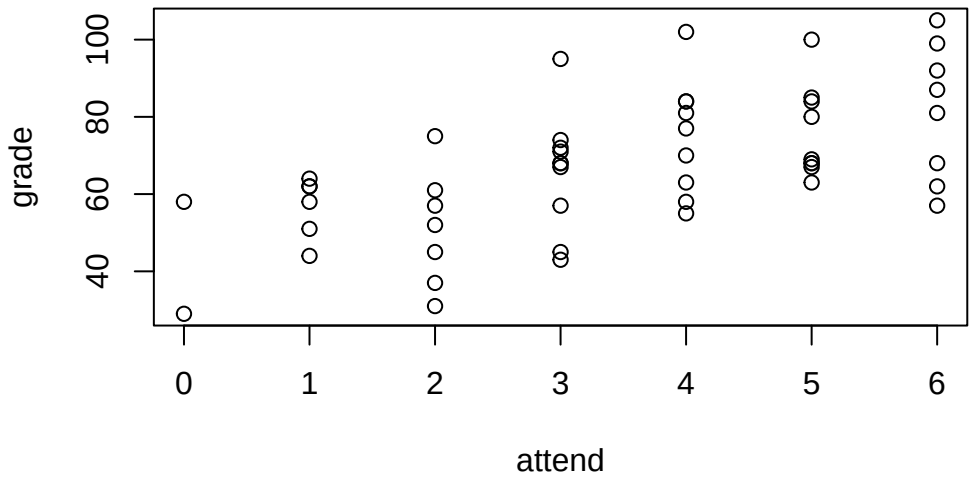
It is a measure of central tendency that divides data in into equal halves. It is the middle observation and the average of two middle observations for odd and even number of observations respectively, when placed in an ordered array. Source: ECON 2B03 Chapter 1 & 2 Slides

2. Consider the following dataset on the final grade received in a particular course (`grade`) and attendance (`attend`, number of times present when work was handed back during the semester out of a maximum of six times). Note that R has the ability to read datafiles directly from a URL, so here (unlike the `odesi` data that you manually retrieve) you do not have to manually download the data *providing you are connected to the internet* (this question is worth 8 marks).

```
course <- read.table("https://jeffreyracine.github.io/2B03/attend.RData")
attach(course)
```

a. Create a scatterplot of the data with `attend` on the horizontal axis and `grades` on the vertical axis via the command

```
plot(attend,grade)
```



Do you see any pattern present in the data? If so describe it in your own words.

b. Construct the average grades for persons attending 0 times, and then repeat for those attending 1 time, 2 times, and so on through 6 times using something like

```
mean(grade[attend==0])
```

```
[1] 43.5
```

```
mean(grade[attend==1])
```

```
[1] 56.83333
```

```
mean(grade[attend==2])
```

```
[1] 51.14286
```

```
mean(grade[attend==3])
```

```
[1] 66
```

```
mean(grade[attend==4])
```

```
[1] 74.88889
```

```
mean(grade[attend==5])
```

```
[1] 76
```

```
mean(grade[attend==6])
```

```
[1] 81.375
```

Do you see any pattern present in the means?

Students who attend classes more often are the ones who achieve higher grades.

3. This question requires you to download data obtained from Statistics Canada. If you are working on campus go to ODESI (off campus users must first sign into the McMaster library via link McMaster libaccess, search for odesi via the library search facilities then select odesi from these search results). Next, select the "Find data" field in odesi and search for "Labour Force Survey June, 2024", then scroll down and select the *Labour Force Survey, June 2024 [Canada]*. Next scroll down to the "DOWNLOAD FILES" section and click on the link "View in Data Explorer Tool", then click on the download tab in the upper right and select "Download RData Format File" which, after a brief pause, will download the data to your hard drive. Finally, make sure that you place this file in the same directory/folder as your R code file (this file ought to have the name *"LFS_June_2024.RData"*, and in RStudio select the menu item Session -> Set Working Directory -> To Source File Location). Note that it would be prudent to retain this file as we will use it in future assignments (this question is worth 8 marks).

   Next, open RStudio, make sure this RData file and your Quarto Markdown script are in the same directory (in RStudio open the Files tab (lower right pane by default) and refresh the file listing if necessary). Then read the file as follows:

```
load("LFS_June_2024.RData")
```

   This data set contains some interesting variables on the labour force status of a random subset of Canadians. We will focus on the variable `HRLYEARN` (hourly earnings) summarized in the link on odesi "View DDI Codebook" (this appears directly below the link for "View in Data Explorer Tool". We will also consider other variables so that we can condition our analysis on these variables by restricting attention to subsets of the data, e.g., for full-time workers only (`FTPTMAIN==1`) reporting positive earnings. We also look at the highest educational attainment for people in the survey and consider both high school graduates (`EDUC==2`) and those holding a bachelors degree (`EDUC==5`). To construct these subsets we can use the R command `subset` as follows (the ampersand is the logical operator *and* - see `?subset` for details on the `subset` command):

```
hs <- subset(table, FTPTMAIN==1 & EDUC==2 & HRLYEARN > 0)$HRLYEARN
ba <- subset(table, FTPTMAIN==1 & EDUC==5 & HRLYEARN > 0)$HRLYEARN
```

These commands simply tell R to take a subset of the data frame `table` for full-time workers having either a high school diploma or university bachelors degree for those reporting positive earnings, and then retain only the variable `HRLYEARN` and store these in the variables named `hs` (hourly earnings for high-school graduates) or `ba` (hourly earnings for university graduates). The following questions ask you to compute various descriptive statistics and other graphical summaries of these two variables.

Note that nothing will be printed out by running the two lines above - they simply create subsets of the data for subsequent use.

a. Report the five number summary for each subset (hint: `fivenum(hs)` etc.). Indicate what each number tells us (hint: see help by typing `?fivenum` in the console pane).

```
fivenum(hs)
```

```
[1]   6.60  20.00  25.00  33.00 197.12
```

Table 1: Five Number Summary HS

Minimum	Lower-Hinge	Median	Upper-Hinge	Maximum
6.60	20.00	25.00	33.00	197.12

```
fivenum(ba)
```

```
[1]   8.65  27.00  38.46  51.28 196.68
```

Table 2: Five Number Summary BA

Minimum	Lower-Hinge	Median	Upper-Hinge	Maximum
8.65	27.00	38.46	51.28	196.68

Comparions: As we can see from the five number summaries, the minimum, lower-hinge, median, and upper-hinge, show higher average salaries for university graduates when compared to high school graduates. The maximum wage for both groups is similiar.

b. What can you say about relative wages of high school and university graduates?

University students earn higher wages on average compared to high school graduates.

c. Using Sturges' rule, how many classes would you construct for the `hs` and `ba` wage data (hint - `length()` gives you the length of the vector, `log10()` may also be useful, so something like `round(1+3.3*log10(length(hs)))` might do the trick for the hs data at least)?

```
round(1+3.3*log10(length(hs)))
```

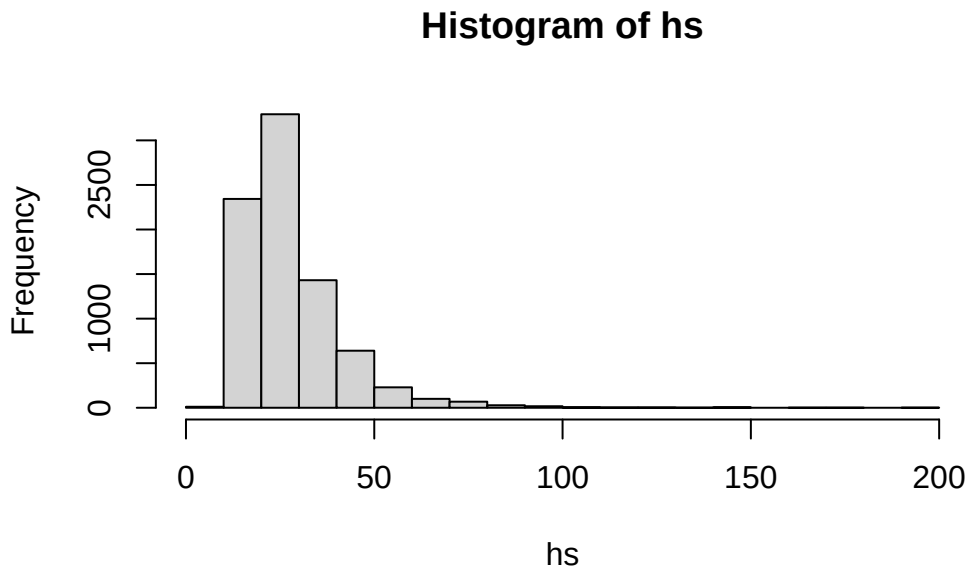
```
[1] 14
```

```
round(1+3.3*log10(length(ba)))
```

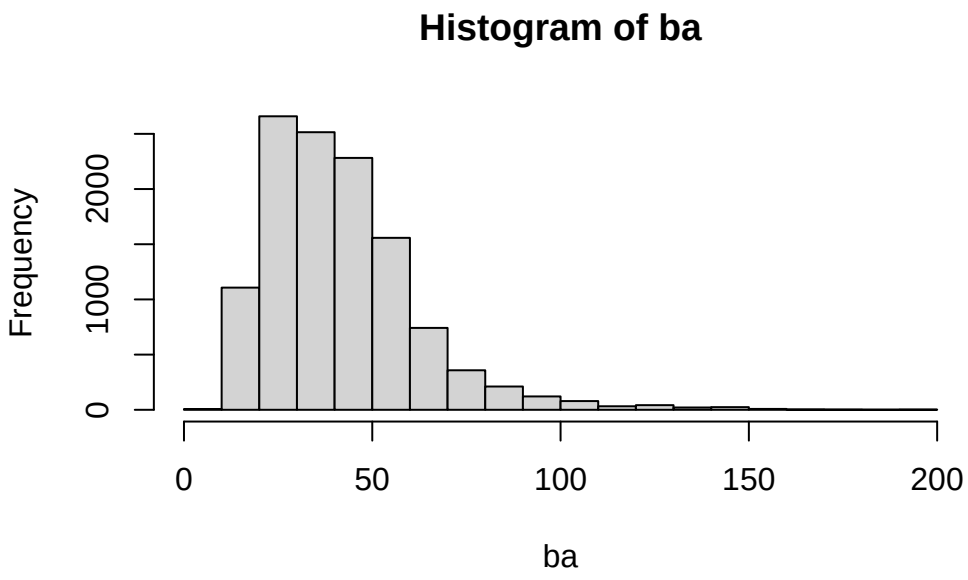
```
[1] 14
```

d. Plot histograms for the `hs` and `ba` data on separate graphs (hint: `hist()`).

```
hist(hs)
```




```
hist(ba)
```

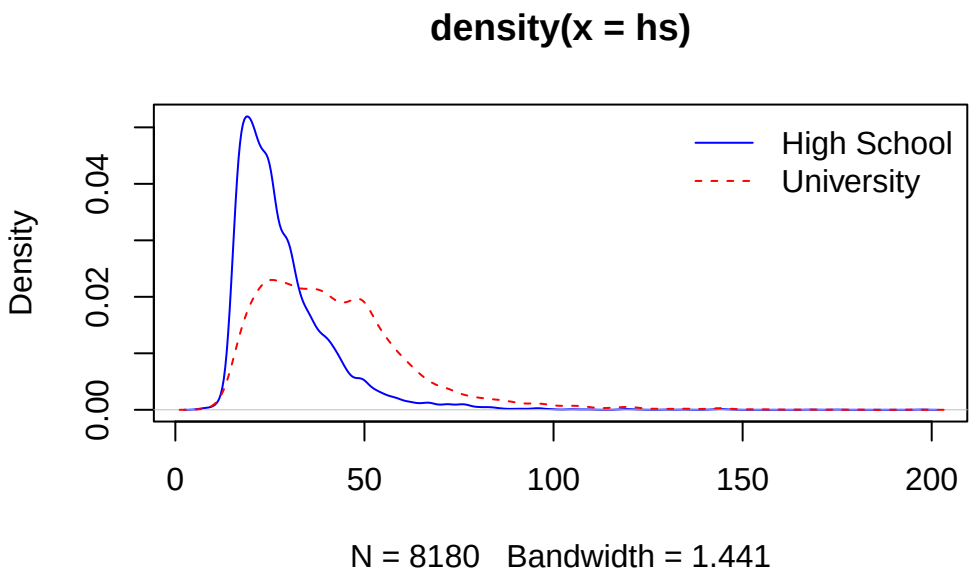


e. Do the number of classes correspond to Sturges' rule?

No, according to Sturges' rules both historgrams should have 14 classes each. The histograms created using the "hist(…)" function have 4 classes.

f. Plot density curves for the `hs` and `ba` data on the same graph and add a legend (hint: first use something like `plot(density(...),col="blue",lty=1)` (you need to fill in `(...)` parts with the name of your data object, e.g., `hs` etc.) then `lines(density(...),col="red",lty=2)`, then see the help page by typing `?legend` in the console pane. Note that you can add a legend after calling `plot()` using something like

```
## This must follow a call to plot() in order to add the
plot(density(hs),col="blue",lty=1)
lines(density(ba),col="red",lty=2)
## legend to an existing plot
legend("topright",c("High School","University"),
       lty=c(1,2),col=c("blue","red"),bty="n")
```



g. What do these density curves tell us about the distribution of hourly wages for high school versus university graduates?

The density curve shows that university graduates on average earn higher wages than high school graduates, since the red curve is more shifted to the right. The flatter red curve also shows that there is a higher variability in wages of university graduates. In comparison high school graduates have a more concentrated peak, at around 25, showing that most high school graduates earn around the same wage.

4. Consider the following data on annual profits (in $millions of dollars) for all firms in the textbook publishing industry in Canada (ignore the `[1]`, etc., that appear at the beginning of each line; this is simply the way R displays a vector of numbers):

```
n <- 14
set.seed(123)
profits <- signif(rnorm(n,mean=10,sd=5),3)
profits
```

```
 [1]  7.20  8.85 17.80 10.40 10.60 18.60 12.30  3.67  6.57  7.77 16.10 11.80
[13] 12.00 10.60
```

To set these values in a vector in R, if desired, you can use the command `profits <- c(...)` where `...` are the values above separated by commas, e.g., c(7.2, 8.85, etc.)

```
profits <- c(7.20, 8.85, 17.80, 10.40, 10.60, 18.60, 12.30, 3.67, 6.57, 7.77, 16.10, 11.
```

a. How many observations are there (i.e., what is n, the sample size?)

```
length(profits)
```

```
[1] 14
```

There are 14 observations in the dataset.

b. What is the minimum, maximum, and range?

```
min(profits)
```

```
[1] 3.67
```

```
max(profits)
```

```
[1] 18.6
```

```
max(profits) - min(profits)
```

```
[1] 14.93
```

c. How many classes would you create if you used Sturges' rule?

```
round(1+3.3*log10(length(profits)))
```

```
[1] 5
```

According to Sturges' rules I should create 5 classes.

d. What are the class widths and class boundaries based on your answers to the previous two questions, using Sturges' rule, the sample minimum as the first lower class boundary, and the sample maximum as the last upper class boundary?

```
(max(profits) - min(profits))/5
```

```
[1] 2.986
```

The class width is 2.986 and the class boundaries are as follows: 3.67, 6.656, 9.642, 12.628, 15.614, 18.6.

a. Complete the table below showing the absolute frequency, relative frequency, cumulative frequency, and cumulative relative frequency for the above data. For this question you will need to do some manual data entry in the table skeleton provided below after you have figured out what the counts are based on your answers to the previous set of questions (for help formatting tables, which can be a bit tricky at first, see this link: Quarto Table Online Help or try out the RStudio visual editor table capabilities). In particular, you are to use Sturges' rule (above) to obtain the desired number of classes, and use the range of the data (above) when constructing your class boundaries (note that you need to have a blank line between each new row that you add to the table, and the last class must be closed at the right - this question is worth 8 marks).

Class	Absolute Frequency	Relative Frequency	Cumulative Absolute Frequency	Cumulative Relative Frequency
[3.67,6.656)	2	0.142857	2	0.142857
[6.656,9.642)	3	0.214286	5	0.357143
[9.642,12.628)	6	0.428571	11	0.785714
[12.628,15.614)	0	0.0	11	0.785714
[15.614,18.6)	3	0.214286	14	1.0

5. Since we use the *summation operator* ($\Sigma_{i=1}^{n}$) often in class, let's make sure we understand how to calculate objects that can be expressed succinctly using this operator.

   a. Care must be exercised when expanding certain sums and quantities. Let the sample size be $n = 3$, and let $X_1 = 1$, $X_2 = -1$, and $X_3 = -3$. Demonstrate in R that it is generally not true that $\sum_{i=1}^{n} X_i^2 = \left(\sum_{i=1}^{n} X_i\right)^2$ using the `sum()` function and exponent operator `^2` applied to the data vector `X` created using `X <- c(1,-1,-3)` (this question is worth 2 marks).

   b. Using the same data as in the previous question, compute the sample mean $\bar{X} = \sum_{i=1}^{n} X_k/n$ then compute the sample standard deviation $\hat{\sigma} = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$ in two ways: longhand (you can use R and use longhand notation, e.g., X[1], X[2], and X[3] or 1, -1, and -3, whichever you prefer), then using R functions such as mean() and sd() (this question is worth 2 marks).

   c. Express $\sum_{i=1}^{n} K$, where K is a constant (i.e., a number that does not change hence has no subscript i), in terms of n and K only (Hint - a constant does not have a subscript as it does not change with i, but it is being added/summed, so type out a string of n constants etc.). Then for $K = 3$ and $n = 5$ determine $\sum_{i=1}^{n} K$ using your result purely using n and K (i.e., without a summation sign - this question is

worth 2 bonus marks, and you do not use R, rather use your powerful sense of logic and type out your answer with an explanation).