**Terminology**

Q1) Match the number of the definitions below to following terms *(Two are not used.)*: (0.1 each)

\_\_**7**\_\_ Ratio \_\_**8**\_\_ Percentage \_\_**9**\_\_ Standard Deviation

\_\_**14**\_ Rate \_\_**10**\_ Frequency distribution \_\_**13**\_ Mean, μ (mu)

\_\_**12**\_ Proportion \_\_**4**\_\_ Σ (uppercase Sigma) \_\_**11**\_ Mean, (X-bar)

\_\_**2**\_\_ Percent change \_\_**1**\_\_ Percentile \_\_**15**\_ Control group

\_\_**6**\_\_ Right (pos.) skew \_\_**3**\_\_ Left (neg.) skew \_\_**17**\_ Associated

1. A point below which a specific percentage of the cases fall.
2. A statistic that expresses the magnitude of change in a variable from time 1 to time 2.
3. The extent to which a distribution of scores has a few scores that are extremely low.
4. The summation of
5. Statistics that indicate the amount of variety, or heterogeneity, in a distribution.
6. The extent to which a distribution of scores has a few scores that are extremely high.
7. The number of cases in one category divided by the number of cases in some other category.
8. The number of cases in a category divided by the number of cases in all categories of a variable, the entire quantity multiplied by 100.
9. Square root of the squared deviations of scores around the mean, divided by N (for population) or n-1 (for sample).
10. A table the displays the number of cases in each category of a variable.
11. The arithmetic average of the scores of the sample.
12. The number of cases in one category of a variable divided by the number of cases in all of the categories of the variable.
13. The arithmetic average of the scores of the population.
14. The number of actual occurrences of some phenomenon or trait divided by the number of possible occurrences per some unit of time.
15. Experiment or study participants / subjects that do not receive treatment or receive a standard treatment.
16. Experiment or study participants / subjects that receive a treatment
17. Two or more variables that are related in some way.

**Theory**

Q2) What is the best central tendency measure to use when the data points are skewed and why? (0.50 points)

**The best measure of central tendency to use when data points are skewed is the median because it’s not affected by outlier values.**

Lecture 1.5

Q3) Why do we use n-1 instead of using n in the denominator of the standard deviation formula. Why not just use n? (0.50 points)

**We use n-1 in the denominator of the standard deviation formula because it allows us to be cautious. Subtracting 1 from n makes n smaller so that when we divide by n-1, it makes our estimates larger, which provides more wiggle room and reduces bias. Dividing by n makes us more likely to be biased.**

**Application**

Q4) Download the Airbag and other influences on accident fatalities dataset called nassCDS.csv and the corresponding codebook from ELMS. Review the codebook and import this dataset into R.

4a) Describe the shape of the dataset. (0.25 points)  
How many columns of data (number of variables)?   
How many rows of data observations?

**There are 16 total columns, but only 15 variables. There are 26,217 rows.**

**Text

Description automatically generated**

4b) Examine the injured severity variable. (0.25 points)

Are there any missing observations? If yes, how many?

What percentage of observations were either incapacitated or killed?

**Yes, there are 153 missing observations and 36.9% of the observations were either incapacitated or killed.**

**Text

Description automatically generated**

4c) Create bivariate tables using the seatbelt and dead variables—one with frequencies (0 decimals) and one with percentages (round to 1 decimal). If seatbelts is in the columns, use column %, otherwise use row %. Report your findings below. You may put them together in 1 table or keep them separate in two tables—but the tables must be presentation quality (PQ) formatted. (0.25 points)

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

Q4d) Report the following summary statistics for the year of the vehicle variable in a presentation quality (PQ) table: sample size (n), minimum, maximum, 1st & 3rd quartiles, median, mean, and standard deviation. Round your answers to the nearest 1 decimal if needed. (0.25 points)

![Table

Description automatically generated]()

**Sample Size (n): 26,217**

**Minimum: 1953**

**Maximum: 2003**

**Q1: 1989**

**Q3: 1997**

**Median: 1994**

**Mean: 1993**

**Standard Deviation: 5.6**

Text

Description automatically generated

Q4e) Create a histogram for the year of the vehicle variable. Describe the type and shape of the distribution. Insert a copy of the graph below and be sure to change the title and x-axis labels from the default to be more presentation quality. (0.25 points)

**Chart, histogram

Description automatically generated**

**This histogram has a left/negative skew. It also somewhat resembles a unimodal left skew. The data’s left skew indicates that the mean is less than the median.**

**Text

Description automatically generated**

Q5) Fix a messy variable: Load the 2016 General Social Survey (GSS). It is an .Rdata file, so load it with this command:  
load(file.choose()) and do not try to name it anything like you would a .csv file.

Examine the “hours worked in the past week” variable: HRS1.

5a) Run this command: table(gss$HRS1)

You should see something odd with the output. You’ll get a similar result if you run summary(gss$HRS1).

What is wrong with this output?

What is the data type in R currently?

What data type should it be?What function can you run in R to confirm the data type? (There are a few options). If the function you answer comes from a library, list that library too.

(0.25 points)

**When I ran the table(gss$HRS1), I noticed that the second to last row in the fifth column returned “89+ hrs.” The current data type is factor, but it should be numeric. The function I ran in R to determine the data type was class(gss$HRS1).**

**A picture containing calendar

Description automatically generated**

5b) Provide code in your R script to recode HRS1 so that it coded properly, meaning that it looks how we should expected it to look for data analysis. *Hint: you will likely have to recode with at least two function lines.* (0.50 points)

*Calendar

Description automatically generated*

5c) Condense (Recode) your newly ‘fixed’ hours worked variable into 3 categories, where anything less than 40 hours per week = “part-time”, 40 hours per week = “full time”, and anything over 40 hours per week = “More than full-time”. Use whatever recoding strategy you want. Provide a PQ table below showing the 3 work hour categories. (0.50 points)

**\*\* NOTE \*\***

**My code is pictured below, but I unfortunately continued to run into an error (also pictured below). I’m not sure how to fix it, so any tips or guidance would be greatly appreciated. I also included code for how I would have designed my PQ table. I’m very sorry I couldn’t figure it out, but I tried my best to debug.**

Text

Description automatically generated