PSY 611 Homework #1

YOUR NAME HERE

Table of Contents

# Instructions

Please complete this assignment using the RMarkdown file provided [Link to Rmarkdown file here]. Once you download the RMarkdown file please (1) include your name in the preamble, (2) rename the file to include your last name (e.g., “weston-homework-1.Rmd”). When you turn in the assignment, include both the .Rmd and knitted .doc files.

To receive full credit on this homework assignment, you must earn **30 points**. You may notice that the total number of points available to earn on this assignment is 65 – this means you do not have to answer all of the questions. You may choose which questions to answer. You cannot earn more than 30 points, but it may be worth attempting many questions. Here are a couple things to keep in mind:

1. Points are all-or-nothing, meaning you cannot receive partial credit if you correctly answer only some of the bullet points for a question. All must be answered correctly.
2. After the homework has been graded, you may retry questions to earn full credit, but you may only retry the questions you attempted on the first round.
3. The first time you complete this assignment, it must be turned in by 9am on the due date (October 25). Late assignments will receive 50% of the points earned. For example, if you correctly answer questions totalling 28 points, the assignment will receive 14 points. If you resubmit this assignment with corrected answers (a total of 30 points), the assignment will receive 15 points.
4. You may discuss homework assignments with your classmates; however, it is important that you complete each assignment on your own and do not simply copy someone else’s code. If we believe one student has copied another’s work, both students will receive a 0 on the homework assignment and will not be allowed to resubmit the assignment for points.

**Data:** Some of the questions in this homework assignment use the dataset referred to as homework-world. [Link to data goes here.] This dataset is similar to one you’ve seen in class and contains a new variable called World. So-called “first world” countries (coded 1) are those that were aligned with the United States after World War II (e.g., members of NATO) or were considered to be clearly in the U.S. sphere of influence. “Second world” countries (coded 2) are former members of the Soviet Union or countries considered to have been clearly in the Soviet Union’s sphere of influence. “Third world” countries (coded 3) include those considered by the United Nations to be among the least developed countries in the world. The remaining countries are coded 4 for this variable.

# 2-point questions

### Question 1

This question is about order of operations:

4 divided by 2 plus 2 times 5

* Using R, show that this value equals 12.
* Using R, show that this value equals 5.
* Using R, show that this value equals 20.

### Question 2

Calculate descriptive statistics (mean, standard deviation, median, min, max, range, skew, and kurtosis) for all continuous variables in the homework-world dataset.

### Question 3

Calculate descriptive statistics (mean, standard deviation, median, min, max, range, skew, and kurtosis) for all continuous variables in the homework-world dataset, but only for the “second world” countries.

### Question 4

Using the homework-world dataset, create a boxplot graph for the Corruption variable, with separate boxplots for the four groups defined by the new variable, World. What does this graph tell you, in broad strokes, about:

* group differences in corruption central tendency?
* group differences in corruption variability?
* the presence of outliers within each group?

### Question 5

# 5-point questions

### Question 1

In class, you were shown a formula for variance as a function of expected values, .

* Create a variable in **R** called X and give it the values 4, 2, 5, 3, 2, 4, 5, 1, 1, 5.
* Calculate the variance of this variable as if it represents an entire population.
* Calculate the variance as if this variable is a sample.
* Using the formula of expected values, calculate the variance. (Treat E as if it means “take the mean.”)
* Which formula, population parameter or sample statistic, does the expected value formula calculate?

### Question 2

Using the homework-world dataset, create a matrix of scatterplots for the happiness, freedom, and support variables. The scatterplot for each pair should include the linear best fit line (a straight line). (You may need to consult the help page for a particular function.)

* Does a linear relationship seem to describe any of the pairwise relationships?
* Are there any countries that stand out as being unusual by their distance from the remaining countries in these plots? [Identify these countries by name and describe how they differ from the others.]

### Question 3

Using the homework-world dataset:

* Calculate the correlation between the average perception of freedom in a country and the typical generosity of its citizens.
* Create a scatterplot showing the relationship between freedom and generosity.
* Referencing the scatterplot, consider the different threats to the validity of a correlation. Are there any that you are concerned about? Are there any that you don’t think are a problem?
* If you had to guess, do you think the correlation you calculated (a) underestimate the strength of the true relationship between freedom and generosity, (b) overestimates this relationship, or (c) does a good job representing the relationship?

# 10-point questions

### Question 1

Use the following steps to create a composite score for the non-categorical variables (Happiness, GDP, Life, Support, Generosity, Freedom, and Corruption) and answer some questions using that composite:

* The variables are scored in a consistent direction (higher is “better”) with one exception: Corruption. Begin by reversing the direction of this variable.
* The variables are in different metrics, so transform them to Z scores.
* Create a mean of the Z scores. Note that there are missing data for some countries. Average the variables that are available for each country.
* Create a figure (or two?) for this new variable. Is it normally distributed? Any outliers?
* Identify the 10 countries with the highest scores and the 10 countries with the lowest scores.
* By collapsing all of these variables into one composite score we are claiming them to be good proxies for some “thing.” What might you name that thing and how would you justify it quantitatively?

### Question 2

Using the homework-world dataset, create a histogram of the life expectancy variable (Life) that includes the following:

* a vertical line at the mean
* a vertical line at the median
* the normal distribution density curve (not the density of the data in hand)

Is this distribution skewed? What are two ways you can tell?

Apply a transformation to the Life variable to try and make it more normal ([see here for some common ones.](https://bookdown.org/mike/data_analysis/variable-transformation.html#variable-transformation)) Plot the transformed variable with the same vertical lines and density curve. Is your transformed variable skewed? What are two ways you can tell?

# 20-point questions

### Question 1

We evaluate statistics on the degree that they are biased and efficient. We discussed bias in class. Now, you’ll have to evaluate efficiency. While bias is a comparison of the statistic to the population parameter (e.g., how close is the expected value of the statistic to the parameter), efficiency is the comparison of two different statistics. More specifically, efficiency refers to the ratio of precision of the estimates of two different parameters at a given sample size.

* Use simulation to estimate the value of the mean and median from a known distribution 100K times at many (at least 10) different sample sizes, between and . At each sample size, estimate the variance of the estimates of the mean and median. Use those estimates to calculate the relative efficiency of the mean compared to the median at each sample size, defined as:
* What does the an efficiency value of 1 mean? What does an efficiency value of less than one mean?
* Create a figure to visually display these results; include a description of this figure, including any key takeaway points.