Vanderbilt University
Political Science Department
Fall 2024

Stats 1
(PSCI 8356)
Professor Jim Bisbee

PROBLEM SET 1: Due Friday, August 30th on Brightspace by 11:59PM.

A reminder: you may work with others in the class on this problem set, and you are in fact encouraged to do so. However, the work you hand in must be your own. Handwritten work is acceptable, but word-processed work (e.g., using LaTeX or `RMarkdown`*) is preferred.*

1. On the class GitHub site may be found a link to a dataset containing information on college undergraduate programs in the United States. In a *very* brief, two-paragraph mini-essay, describe the data therein. In your discussion, you must correctly use each of the following 8 terms. For Martin's sake, please **underline** the terms as you use them in your mini-essay. Don't worry too much about the content or flow of your essay: I just want to see you using these terms correctly. **NB:** You will need to *look* at some of these variables by visualizing them in `R`.

interval-level variable	a distribution skewed to the right
nominal-level variable	mode
units	dichotomous variable
symmetric distribution	median

2. Using the same data, summarize the variables `control`, `adm_rate`, and `md_earn_wne_p6`. What is their "central tendency"? What is their "dispersion"?

3. Consider a standard normal distribution with $\bar{y} = 5$ and $s = 3$. According to Tchebysheff's Theorem, what proportion of the data are greater than zero?

4. Using the `rnorm()` function in `R`, generate a normal distribution of 1,000 observations with $\bar{y} = 5$ and $s = 3$. What proportion of the 1,000 observations are greater than zero. Is this consistent with Tchebysheff's Theorem from your answer above? Why or why not? `set.seed(123)` to ensure you get the same answer every time.