Econometrics HW #1

Ryan Safner

Due: Tuesday, September 25, 2018

For this homework, you may type up or handwrite the answers to each question. When you get to the *R* questions at the end, please attach a copy of your *.R* file (with comments!) of commands for each relevant question. For future homeworks, I will demonstrate how you can complete everything using *R Markdown*, but you will not be obligated to do so.

Theory & Concepts

For the following questions, please answer the questions completely but succinctly (2-3 sentences).

1. Explain the difference between exogeneity and endogeneity.

2. Explain how conducting a randomized controlled experiment helps to identify the causal connection between two variables.

Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use R to verify your answers, but you are expected to reach the answers in this section "manually."

3. A college senior has applied for admission to two medical schools, A and B. She estimates the probability of acceptance at A at 0.7 and the probability of acceptance at B at 0.4 and the probability that she will be admitted to both at 0.2. What is the chance she will *not* be accepted at *either* school? (Consult the probability handout for basic probability rules.)

4. Suppose the probabilities of a visitor to Amazon's website and buying 0, 1, or 2 books are 0.2, 0.4, and 0.4 respectively. What are the *expected number* of books a visitor will purchase and the *standard deviation* of book purchases?

5. Scores on the SAT (out of 1600) are approximately normally distributed with a mean of 500 and standard deviation of 100.

a. What is the probability of getting a score between a 400 and a 700?

b. What is the probability of getting a score between a 300 and a 800?

c. What is the probability of getting *at least* a 700?

d. What is the probability of getting *at most* a 700?

e. What is the probability of getting exactly a 500?

R Problems

For the following problems, please attach/write the answers to each question on the same document as the previous problems, but also include a printed/attached (and commented!) $.R$ script file of your commands to answer the questions.

6. Using the "table" method of finding standard deviation for a random variable discussed in class, use R to find the standard deviation of the following discrete random variable, X, that has the following pdf:

x_i	p_i
0	0.30
5	0.50
10	0.20

To jog your memory, standard deviation is the square root of the sum of the squared deviations from the mean weighted by the probability of the associated value of X.

7. Redo question 5 parts a-d using the `pnorm()` command in `R`.

8. We will use the dataset `mpg`, which is a part of the `ggplot2` package, and describes fuel economy data from the EPA on models of cars released between 1999-2008. Load (or install, if you don't have) the `ggplot2` package in order to use `mpg`.

a. What variables are included in the `mpg` data? (You don't need to explain them, they aren't well-documented, only write down they are).

b. Find summary statistics for `hwy` and `cty` (miles per gallon on the highway and in the city, respectively).

c. How many different manufacturers are in the data, and how many cars from each manufacturer?

d. How many different classes are in the data, and how many cars from each class?

e. Plot two histograms, one of `hwy` and one of `cty`

f. Plot a boxplot of `hwy` mpg by `manufacturer`. Which manufacturer appears to have the highest average highway mpg? Which appears to have the largest variance? Which appears to have outliers? (There are too many manufacturers for `R` to print them on the axis by default, but they are listed in alphabetical order, check using your answers to part c.)

g. Plot a boxplot of `hwy` by `class`. Which car classes appear to have the highest average highway mpg? Highest variance?