Lesson 01: Introduction to Statistics

References

- Black, Chapter 1 Introduction to Statistics (pp. 1-9), Chapter 2 Charts and Graphs (pp. 15-34)
- Kabakoff, Chapter 2.2.5 Factors (pp. 30), Chapter 5.2.3 Setting the Seed for RNG (pp. 96)
- Davies, Chapter 2.3 Vectors (pp. 23-27), Chapter 8.2 Reading in External Data Files (pp 153-154)
- Stowell, Chapter 3 Selecting a Random Sample from a Dataset (pp. 44)

Data set: home_prices.csv

Description: This data file is derived from a random sample of home resale records maintained by realtors. There are 117 observations and eight variables:

1. PRICE = Selling price ($hundreds)
2. SQFT = Square feet of living space
3. YEAR = Year of construction (year)
4. BATHS = Number of bathrooms
5. FEATS = Number out of 11 features (dishwasher, refrigerator, microwave, disposal, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
6. NBR = Located in northeast sector of city (YES) or not (NO)
7. CORNER = Corner location (YES) or not (NO)
8. TAX = Annual taxes ($)

Exercises:

a) What are the measurement levels of each of the eight variables?

b) Should any variable have its values changed to better reflect its true nature?

c) From the vector "price", select a simple random sample of size 12. Assign the sample to the name "SRS."" Print SRS and determine the mean value.

d) From the vector "price", select a systematic sample of twelve observations. Start with the seventh observation and pick every 10th observation thereafter (i.e. 7, 17, 27,). You should end with the 117th observation. Assign the sample vector to "SS."" Print the values of SS and compute the mean value.

e) Examine the printed values and mean values obtained from the two sampling procedures. Do you see a difference? Try the commands summary(SRS) and summary(SS).

f) Create histograms and stem-and-leaf plots for SRS and SS using hist() and stem(). How do the two samples compare?