5 Introduction and Chapter 1

If you torture the data long enough, Nature will confess.

Ronald Coase

1. Motivation - why economists are interested in statistics and probability
2. What are we measuring $\rightarrow$ unit of analysis [4] The elements are also referred to as observations or rows of data.
   (a) Aggregations of units of analysis - population
      i. Cross-sectional - describing one period of time.
      ii. Time Series - description over time.
   (b) Sample - a special subset of the population
3. The characteristics of interest - variables (elements of variables are values)
4. Types of analyses and variables
   (a) Types of analyses
      i. Univariate
      ii. Bivariate
      iii. Multivariate
   (b) Types of variables
      i. Qualitative
         A. Categorical-multiple categories
         B. Ordinal-Values represent ranking or ordering
      ii. Quantitative (or interval) variable
         A. Discrete - countable set of values.
         B. Continuous
      iii. Why do we worry if the variable is qualitative or quantitative ?
      iv. How to check if a variable is qualitative or quantitative.
   (c) Missing data
      i. Why does it exist?
      ii. Describing missing data.
   (d) Outliers

[4] textbook refers to this as an experimental unit or an element of the sample

Handout - The Summation Operator

The summation operator (Σ) is an efficient way to describe a function. Often, statistical functions are in the form of

$$X_1 + X_2 + X_3 + \ldots + X_n$$

where each X_i is value of the variable X. If a researcher is working with a very large number of values, the long-hand method is cumbersome. A simpler, equivalent representation of the equation above is.

$$\sum_{i=1}^{N} X_i$$

The line above reads, " The sum of X sub i from i equal 1 to N". In English, it reads, "add up all of the values of X from X_1 to X_N."

Rules regarding the Summation Operator

1. The location of variables within parenthesis matters. If the parentheses are located after the summation sign, do the algebraic operation first, then the summation. Here is an example.

$$\sum_{i=1}^{3} (X_i + 1)$$

This line means....

$$(X_1 + 1) + (X_2 + 1) + (X_3 + 1)$$

If the statement was instead this.

$$\sum_{i=1}^{3} X_i + 1$$

The line means

$$X_1 + X_2 + X_3 + 1$$

A similar problem is found between

$$\sum_{i=1}^{3} (X_i^2)$$

and

$$\left(\sum_{i=1}^{3} X_i\right)^2$$

The former works out as

$$X_1^2 + X_2^2 + X_3^2$$

The later is

$$(X_1 + X_2 + X_3)^2$$

2. If the expression being summed contains an "+" or "-" at the highest level, then the summation sign may be taken inside the parentheses. Let's take one of the examples above

$$\sum_{i=1}^{3} (X_i + 1)$$

This can be rewritten as

$$\sum_{i=1}^{3} X_i + \sum_{i=1}^{3} 1 \equiv \sum_{i=1}^{3} X_i + 3$$

So a function written as

$$\sum_{i=1}^{3} (X_i + Y_i)$$

can be rewritten as

$$\sum_{i=1}^{3} X_i + \sum_{i=1}^{3} Y_i$$

3. A sum of a constant times a variable is equivalent to the constant times the sum of the variable. The constant is a value that does not change with the different values of the variable. If every variable is multiplied by the same number and then added up, it would be equal to the sum of the variables times the constant, or

$$\sum_{i=1}^{3} c * X_i$$

is equal to

$$c * \sum_{i=1}^{3} X_i$$

Test these rules out for yourself using the following values.
$X_1 = 10$ $Y_1 = 1$
$X_2 = 57$ $Y_2 = 102$
$X_3 = 87$ $Y_3 = 54$
$c = 2$

1. Major categories of statistical tools
   (a) Descriptive Statistics
   (b) Inferential Statistics
2. Achieving the objectives of Inferential Statistics in Economics
   (a) Specify questions to be answered and identify population of interest
   (b) Decide how to select the sample
   (c) Select the sample and analyze the sample information.
   (d) Use the information from (c) to make inferences about the population.
   (e) Determine the reliability of inference.
3. Graph Concepts

   > Data graphics should draw the viewer's attention to the sense and substance of the data, not to something else. The data graphical form should represent quantitative contents. Occasionally artfulness of design makes a graphic worthy of the Museum of Modern Art, but essentially statistical graphs are instruments to help people reason about quantitative information.
   >
   > Edward R. Tufte
   > *The Visual Display of Quantitative Information*

   (a) Stem and leaf plot - graphical representation of interval and ordinal data using numbers.
      i. Divide data into n main categories - **stem** portion of the graph.
      ii. Within each category, order values (the **leaves**) from lowest to highest.
      iii. List the stem in the first column, leaves in the remaining columns.

      Example

      90 70 70 70 75 70 65 68 60 74 70 95
      75 70 68 65 40 65 70

      4 | 0
      5 |
      6 | 0 5 5 5 8 8
      7 | 0 0 0 0 0 0 0 4 5 5
      8 |
      9 | 0 5

(b) Part of Graphs
   i. Body
   ii. X-axis (horizontal axis)
   iii. Y-Axis (vertical axis)
   iv. Z-Axis (in 3D graphs, the X & Y axes become the horizontal axes, the Z axis is the vertical)
   v. Axis labels
   vi. Data Labels - often stored in a Legend.

(c) Type of graphs
   i. Univariate data
      A. Pie Chart
      B. Bar Chart **(Histogram)**
      C. Stem and Leaf Chart
      D. Box-plot [5]
   ii. Bivariate data [6]
      A. Scatter plot
   iii. Multivariate data
      A. 3D & Contour Plots

4. Tools to prepare univariate data.

(a) For interval variables, group data into K mutually exclusive categories. For other variables, use existing categories.

(b) Count within each category → **Frequency**

(c) Compute Total number of data items → **n** (for sample), **N** (for population)

(d) Compute **Relative Frequency**

$$\text{Relative Frequency(category } j) = \frac{\text{Frequency of category } j}{\text{Total}}$$

   i. Range of values from 0 to 1.00

(e) Compute **Cumulative Relative Frequency**

$$\text{Cumulative Relative Frequency(category } j) = \frac{\sum_{i=1}^{j} \text{Frequency of category } i}{\text{Total}}$$

[5] Presented in Chapter 2, will not be covered in this class.
[6] Preparing bivariate graphs will be covered in Chapter 3

i. Range of values from 0 to 1.00

(f) Compute **Relative Frequency as an angle** - used for PieCharts.

Relative Frequency of category j as an angle = Relative Frequency of category j*360

i. Range of values from 0 to 360

5. Tools to analyze graphical information - univariate data
   (a) Is the distribution symmetric or asymmetric?
   (b) Describing the peak(s) - unimodal or bimodal distribution.
   (c) Describing the shape - degree of skewness.
6. Graphical integrity [7].
   (a) Why do some graphics lie? Why do some publications publish them?
      i. Lack of quantitative skills of the professional artists.
      ii. Doctrine - Statistical data are boring.
      iii. Doctrine - Graphics are only needed by the unsophisticated reader.
   (b) Criteria for graphical excellence - Graphical displays should
      i. show the data.
      ii. induce the viewer to think about the substance rather than the methodology, graphic design, the technology of graphic production.
      iii. avoid distorting what the data have to say.
      iv. present many numbers in a small space.
      v. make large data sets coherent.
      vi. encourage the eye to compare different pieces of data.
      vii. reveal the data at several levels of details, from a broad overview to a fine structure.
      viii. serve a reasonable purpose: description, exploration, tabulation, or decoration.
      ix. be closely integrated with the statistical and verbal descriptions of a data set.

[7] from Tufte