Chapter 1. Looking at Data

(EXAMPLE: Students’ Self-Concept and Grades)

Data and Variables

• Data
  – Consist of some information about individuals.
  – **Individuals** (or Cases):
    the objects (sampling units) described by the data set, typically presented row wise in the data set.
  – **Variables**:
    measurements/characteristics of each individual, typically presented column wise in the data set.

• Distribution of a variable
  – What values the variable takes and how often it takes these values.
  – Types of variables
    i. **Categorical** variables take one of several categories.
    ii. **Quantitative** (numerical) variables take numerical values, and arithmetic operations make sense.

• What do we see from a distribution?
  – Look for the overall pattern and deviations from the pattern.
  – Describe the pattern by
    i. **Shape** (symmetric, skewed to the right or left, unimodal, bimodal and etc.)
    ii. **Center** (approximate midpoint)
    iii. **Spread** (range)
  – Look for outliers if any.

• Data Analysis
  – Extract useful/meaningful information out of data.
  – First, examine the data (exploratory data analysis):
    for a single variable, look at and describe its distribution and
    for multiple variables, examine the relationships among them.
  – Often, the final goal is to answer some specific questions given in the beginning of the study.
Graphical Tools to Describe a Distribution

- For categorical variables
  - **Bar graph** (bar chart): the height of each bar shows the count (or percent) of each category.
  - **Pie chart**: the area of each slice is proportional to the count (or percent) of each category.

- For quantitative variables
  - **Histogram**
    breaks the values into intervals (bins) and shows the frequencies (counts) or relative frequencies (percents) of observations in the intervals. Very common and good for large number of values.
  - **Stem-and-leaf plot**
    gives a quick picture of the shape of the distribution while retaining the actual values in the graph. It separates each observation into a stem (typically all but the final digit) and a leaf (the final digit). Preferable for small to moderate number of observations. **Back-to-back stemplot** is useful for comparing two related distributions.
  - **Box plot**
  - **Dot plot**
  - **Time plot**
    shows values collected over time against the time scale (e.g. daily temperature, monthly unemployment rate). It may reveal seasonal variations, some trends or changes over time.

Example: man’s winning times (in minutes) of the Boston marathon from 1959 to 2001.
Describing Distributions with Numbers

- Numerical summaries can make comparisons of distributions more specific.

  How to measure center?
  How to measure spread or variability?

(EXAMPLE: Babe Ruth Home Runs)

Measuring center

- Mean ($\bar{x}$): the average value of $n$ observations $x_1, x_2, \cdots, x_n$
  
  $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - What is the impact of outliers on the mean?
    
    The sample mean is not a robust (resistant) measure as it is too sensitive to the influence of extreme observations. Generally it gives useful summary for symmetric distributions without outliers.

- Median (M): the middle value
  
  - sort out the observations in increasing order. M is the center observation in the list.
  - if $n$ is odd, M: the $\left(\frac{n+1}{2}\right)$th observation.
  - if $n$ is even, M: the average of the two center observations.
  - resistant measure, good for skewed distributions.

Measuring spread or variability

- Range: Max-Min
  
  - not resistant to outliers.

- Interquartile range (IQR): $Q_3 - Q_1$
  
  - the first quartile $Q_1$ has one quarter of the observations below it.
  - the third quartile $Q_3$ has one quarter of the observations above it.
  - more resistant measure of spread compared to the range.
Boxplot

- Five-number summary: Min, $Q_1$, M, $Q_3$, Max
- Boxplot: graphical representation of the five-number summary
  - a central box spans the quartiles ($Q_1$ and $Q_3$).
  - a line in the box marks the median (M).
  - lines extend from the box out to the smallest (Min) and the largest (Max) observations.

- Less detailed than histograms or stemplots.
- Good for side-by-side comparison of several distributions. (e.g. side-by-side boxplots of GPA split by Gender)

- Identifying suspected outliers
  - A rule of thumb (1.5×IQR criterion):
    Call an observation a suspected outlier if it falls more than 1.5×IQR above $Q_3$ or below $Q_1$.
  - A modified boxplot:
    The lines (or whiskers) stretch to the greatest value less than or equal to ($Q_3 + 1.5\times\text{IQR}$) and the smallest value greater than or equal to ($Q_1 - 1.5\times\text{IQR}$). Any observations beyond the upper or the lower limit are plotted individually.

Measuring spread

- Variance ($s^2$): the average squared deviations of $n$ observations, $x_1, x_2, \cdots, x_n$.
  $x_i - \bar{x}$: the deviation of the $i$th observation from the mean $\bar{x}$.
  - What if we average the deviations?
  - The sample variance is defined as $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.
  - Why divide by $n-1$ rather than $n$?
    It has to do with the degrees of freedom (the number of free terms in the sum of squares). Since the sum of deviations is zero, only $(n-1)$ deviations are free.
• **Standard Deviation** \( (s) \): the square root of the variance

\[
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})^2}
\]

and it has the same unit as the original observations.

- Can \( s \) be zero?
  - Unless all the observations are the same, \( s > 0 \).
- The more spread, the larger \( s \).
- Not resistant measure.
- In general, \( s \) and \( \bar{x} \) are good for symmetric distributions without outliers.

**Changing the unit of measurement**

- Examples:
  - kilometers → miles
    \[
x_{new} \text{ (miles)} = 0.62 \times x_{old} \text{ (kilometers)}
\]
  - degrees Fahrenheit → degrees Celsius
    \[
x_{new}(^\circ C) = \frac{5}{9}(x_{old} - 32) = -\frac{160}{9} + \frac{5}{9}x_{old}(^\circ F)
\]

- Linear transformations: \( x_{new} = a + bx \)
  - \( a \): a shift constant
  - \( b \): a scale constant (typically \( b > 0 \) for unit conversion)

- Example of linear transformations
• Change in distributions:
  How do the descriptions of a distribution change from one unit to another?

  – Linear transformations do not change the shape of a distribution.
  – They do change the numerical measures of center and spread.
    i. Change a measure of center \( m \) (mean or median) to \( a + bm \).
    ii. Multiply a measure of spread (standard deviation, IQR, or range) by \( b \) for \( b > 0 \) and multiply variance by \( b^2 \).