6 Chapter 2 - Describing Data with Numerical Measures

> I do hate sums. There is no greater mistake than to call arithmetic an exact science. There are ... hidden laws of Number which requires a mind like mine to perceive. For instance, if you add a sum from the bottom up, and then again from the top down, the result is always different.
>
> M.P. LaTouche (1878)

1. Previously introduced - the notion of populations and samples - more often than not, researchers are working with a sample of the total population. It is understood that a correctly specified sample represents a good approximation of the total population.
   (a) Often we cannot obtain complete information regarding the population. If we cannot obtain information regarding the population and our sampling techniques are sound, numerical information regarding the population can be inferred from the numerical information from the sample.
2. Motivation for using numerical measures of populations and samples.
3. Restrictions on what can be measured numerically.
   (a) Some measures cannot be used with some variables. The best examples, categorical & ordinal variables - you cannot find the average value of a categorical & ordinal variables.
4. Measure of Center (measuring the center of the distribution)
   (a) Arithmetic Mean (Interval Variables) [8]
      i. Population mean (μ) (mu)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

      ii. Sample mean $\bar{x}$ (x bar)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[8] N refers to the population count, n is the sample count.

(b) Median (Interval Variables)

i. Order values from lowest to highest.

ii. Position of the median $\rightarrow$ 0.5*(N+1)

iii. If the position of the median is a number that ends in 0.5, you need to average the two adjacent points.

(c) Mode (All variables)

i. The category (or categories) with the highest frequency of values. A distribution that has two modes is called bimodal.

(d) Sensitivity of the mean value to outliers.

Valid numeric measures for the three main variable types.

Variable Type	**Mean**	**Median**	**Mode**
Categorical	No	No	Yes
Ordinal	No	No	Yes
Interval	Yes	Yes	Yes

5. Measure of variability (or measure of the variation or dispersion around the center of the distribution) (Interval Variables)

(a) Range - the difference between the largest and smallest values.

(b) Deviation - difference between a given value in the sample or population distribution (x_i) and its mean ($\bar{x}$ or μ) value.

(c) Variance Average of the squared deviation from the mean.

i. Population variance (σ^2) (sigma squared)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

ii. Sample variance (s^2)

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

(d) Standard Deviation Square root of the variance

i. Population standard deviation (σ) (sigma)

ii. Sample variance (s)

(e) Points regarding Variance and Standard Deviations

i. s and σ are always greater than or equal to zero.
ii. The larger the value of the variance or standard deviation (population or sample), the greater the dispersion.
iii. If the variance is equal to zero, all values are equal to the mean.
iv. To measure or describe the variation or dispersion using the units of analysis, use the standard deviation.

(f) Tchebysheff's and the Empirical Rule - knowledge of these two tricks allows you to describe the distribution more in depth.

i. Tchebysheff's [9] Theorem - Given a number k greater than or equal to one, and a set of n measurements, at least $(1 - (\frac{1}{k^2}))$ of the distribution lies in the interval within k*(standard deviation) from the mean.
A. Works for all interval distributions.
ii. Empirical Rule - Given a distribution of measurements that is approximately symmetrical ...
A. The interval $(\mu \pm \sigma)$ or $(\bar{x} \pm s)$ contain approximately 68% of the measurement.
B. The interval $(\mu \pm 2\sigma)$ or $(\bar{x} \pm 2s)$ contain approximately 95% of the measurement.
C. The interval $(\mu \pm 3\sigma)$ or $(\bar{x} \pm 3s)$ contain nearly all of the measurement.

(g) Applications of Tchebysheff's and the Empirical Rule

6. Measures of Relative Standing (the position of one value compared to the center and variation of the distribution) (Interval variables)

(a) Z score - measurement of the value as a weighted value of the standard deviation.

$$Zscore = \frac{x - \bar{x}}{s}$$

i. Application of Tchebysheff's and the Empirical Rule to the Z-score.

(b) Percentile - Values in the distribution are sorted or ordered in ascending order. Percentile of x is its percent position within this order.

i. Median is at the 50th percentile.

(c) Quartiles. Quintiles, inner and outer fences
Quartiles and qunitiles are percentile measures of a distribution of data. Suppose you can show the relative frequencies of data either in terms of a table or a graph.

[9] Pafnuty Livovich Tchebyshev (1821-1894) Russian mathematician noted for his foundation in the mid 19th century of the Petersburg mathematical school

Percentiles measure the percent of values from the minimum to some percent (such as 10%). Quartiles are points at 25%, 50%, and 75% of the distribution. Quintiles are points at 20%, 40%, 60% and 80% of the distribution.

Quartiles are used to assess the existence of outliers. Before getting into this, let's look at how quartiles are computed.

i. First quartile (Q1) is the point in position 0.25*(n+1).

ii. Second quartile (Q2) (also the median) is the point in position 0.50*(n+1).

iii. Third quartile (Q3) is the point in position 0.75*(n+1).
When the measure is not an integer, the quartile is found by taking the two integers between the measure and mutiplying the difference by decimal value of the solution. Here is an example - find Q1 for the following

$$4, 8, 9, 11, 11, 13, 16, 18, 20, 25$$

Position of Q1 = 0.25*(n+1) = 0.25*(11)=2.75

Given that answer is not an integer, we need to use the method above. An answer of 2.75 means that the answer lies between the second and third value or lies between 8 and 9. Take the difference of these values, which is one. Use the decimal value of your answer (the .75 of 2.75) and multiply it by the difference (0.75*1=0.75). This means that the correct value of Q1 lies 0.75 beyond the value in the second position or 8.75

(d) The interquartile range (IQR) is the middle 50% range of values. It is the difference between the Q1 and Q3 or

$$IQR = Q3 - Q1$$

(e) Suspect outliers lie in the inner fence. The formulas for these regions of the distribution are:

$$\text{Left inner fence} = Q1 - 1.5 * IQR$$

$$\text{Right inner fence} = Q3 + 1.5 * IQR$$

(f) Extreme outliers lie in the outer fence. The formulas for these regions of the distribution are"

$$\text{Left outer fence} = Q1 - 3 * IQR$$

$$\text{Right outer fence} = Q3 + 3 * IQR$$