Summary of Descriptive Statistics

Dimitrios D. Thomakos

University of Peloponnese

Quantitative Methods

Descriptive Statistics are used to:

- Allow us to "know" our data.
- To summarize the sample information in a compact and meaningful way.
- To check the possible underlying probabilistic structure of our data.
- To examine the presence of outliers and decide what to do with them.

There are 7 types of descriptive statistics

- The order statistics (empirical distribution of the data) and percentiles.
- Measures of central tendency.
- Measures of dispersion (around the central tendency).
- Measures of asymmetry (around the central tendency).
- Measures of 'tail-heaviness' or kurtosis.
- Measures of normality (of the underlying theoretical distribution).
- Measures of dependence and correlation.

Order Statistics and Percentiles

Let $\{X_i\}_{i=1}^{N}$ denote our sample of N observations. We have that:

- The order statistics are the sample observations ordered from the smallest to the largest, that is $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$, where and $X_{(1)} \doteq \min_i X_i$ similarly $X_{(N)} \doteq \max_i X_i$.
- The p-% percentiles (also known as quantiles), denoted by Q_p are these observations (or interpolated values between observations) such that p-% of the observations are less than or equal to Q_p. That is, in the sequence of the order statistics the percentiles

  $$\underbrace{X_{(1)}, X_{(2)}, \ldots, Q_p}_{p\text{-\% of observations}}, \ldots, X_{(N)}.$$

- The 25%, 50% and 75% percentiles are also called the quartiles of the empirical distribution.

The Sample Mean and the Sample Median

- The sample mean $\bar{X}_N$ is the equally weighted sum of all the sample observations, that is $\bar{X}_N \doteq (1/N)\sum_{i=1}^{N} X_i$
- The sample median M is the central percentile of the empirical distribution of our data, that is $M \doteq Q_{0.50}$.
- Both the above measures are solutions to dispersion optimization (minimization) problems and are, therefore: (a) 'optimal' in measuring central tendency and (b) they define their 'own' measures of dispersion. We have that:

$$\bar{X}_N \doteq \min_{\alpha}(1/N)\sum_{i=1}^{N}(X_i-\alpha)^2 \qquad M \doteq \min_{\alpha}(1/N)\sum_{i=1}^{N}|X_i-\alpha|$$

  where α is a constant. Graphical interpretation of these problems?
- For data corresponding to discrete measurements (example?) we also have the mode of the empirical distribution.
- The sample median is a more 'robust' measure of central tendency (interpretation? examples?).

The Sample Variance/Std. Deviation and the Mean Absolute Deviation

- The optimization problems for the measures of central tendency provide us with the corresponding measures of dispersion.
- The sample mean, which minimizes the sum of squared deviations from a constant, gives us the sample variance (or mean squared deviation) $s_N^2 \doteq (1/N)\sum_{i=1}^{N}(X_i - \bar{X}_N)^2$. The standard deviation s_N is used since it is measured in the same units as the original variable.
- The sample median, which minimizes the sum of the absolute deviations from a constant, gives us the sample mean absolute deviation $MAD \doteq (1/N)\sum_{i=1}^{N}|X_i - M|$.
- There are a couple of alternative measures of dispersion that are useful in the empirical analysis. These are:
  - The mean absolute deviation *from the mean* $(1/N)\sum_{i=1}^{N}|X_i - \bar{X}_N|$.
  - The scaled range $(1/N)(X_{(N)} - X_{(1)})$.
  - The scaled interquartile range $IQR \doteq (Q_{0.75} - Q_{0.25})/1.34$.

- There is also a measure of relative dispersion (variation) called the coefficient of variation given as the ratio of the standard deviation over the sample mean $CV \doteq s_N/\bar{X}_N$.
- The coefficient of variation shows us the relationship between the dispersion of the values in the sample relative to the value of the mean. A large value indicates large dispersion relative to the mean value while a small value indicates small dispersion relative to the mean. This can be used to judge the predictive capability of the sample mean for future observations (more when we talk about inference using the sample mean).

The Sample Skewness and the Sample Kurtosis

When analyzing data were are frequently interested not only in their central tendency but also in the 'tails' of the empirical distribution. Tail-based information can be used to (a) improve our understanding of the data and (b) improve our inference later on. There are two types of tail-based information that we most frequently use in descriptive statistics:

- The sample skewness is a measure of the degree of asymmetry of the empirical distribution of the data around their mean. It is useful since it tells us whether we have a few large values either to the right (positive skewness) or to the lest (negative skewness) of the sample mean. The sample skewness is defined as:

$$\hat{\beta}_3 \doteq \frac{(1/N)\sum_{i=1}^{N}(X_i - \bar{X}_n)^3}{s_N^3}$$

The Sample Skewness and Sample Kurtosis cont.

- The sample kurtosis is a measure of the degree of 'tail-heaviness' of the empirical distribution of the data. It is useful since distributions with 'fat' tails indicate that there is increased chance of encountering not just few but many extreme values. Note that this extreme values are not outliers but rather an integral part of the data we are analyzing. The sample kurtosis is defined as:

$$\hat{\beta}_4 \doteq \frac{(1/N)\sum_{i=1}^{N}(X_i - \bar{X}_n)^4}{s_N^4}$$

Describing more than one variables

When our data include more than one variables we would like to summarize the potential relationship that exists amongst them. Formally this is part of statistical modeling but there are a couple of descriptive devices that can be used to summarize multivariate information.

- The simplest such measure is the sample correlation coefficient. This is a unit-free measure ranging between -1 and 1 that measures the degree of linear association between two variables. Note the term 'linear' here. Correlation cannot capture properly non-linear relationships and is not synonymous to causation! It is defined as:

$$r_{XY} \doteq \frac{\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)}{s_X \cdot s_Y}$$

Describing more than one variables cont.

- Another simple to use measure is a contingency table. In this table we split the data of both variables into categories (if there is no such natural split) and count the number of joint occurrences in each table cell. If we convert these counts into proportions we can then compute the total squared deviation of each cell from the equal proportions ratio to see whether there is some dependence in the two variables. Specifically, let the number of categories for the X variable be C_X and the number of categories for the Y variable be C_Y. For each table cell we have a proportion π_k for $k = 1, 2, \ldots, K$, with $K \doteq C_X \cdot C_Y$ the total number of cells. The total squared deviation can be computed as:

$$\chi^2_{total} \doteq K \cdot \sum_{k=1}^{K} (\pi_k - 1/K)^2$$