Chapter 1

Descriptive Statistics

1.1. Representation of data

Consider a generic sample of n observations for a given variable x, i.e. $x_1, x_2, ...x_n$. There are at least three possible representation of these observations

1. *Raw representation*: You write down the data in the order you observe them
2. *Ordered representation*: You put the observations on the real line and renumber them in an ascending order
3. *Frequency representation*: For each observation, you construct absolute frequency

$$n_i = \#\text{observation}_i$$

and relative frequency

$$f_i = \frac{n_i}{\sum_{i=1}^{n} n_i}$$

Notice that

$$\sum_{i=1}^{n} n_i = n \implies \sum_{i=1}^{n} f_i = 1$$

1.2. Measures of centrality

Consider a generic sample of n observations for a given variable x, $x_1, x_2, ...x_n$.

Mean The sample mean of x is the usual arithmetic mean, defined as follows

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

Median The sample median of x is the observation having the property that at least 50% percent of the data are less than or equal to it and at least 50% percent of the data values are greater than or equal to it. If two data values satisfy this condition, then the sample median percentile is the arithmetic average of these values.

Mode The mode of x is the most frequent observation. Note that it is possible for a frequency distribution to have more than one most frequent observation.

Geometric Mean The geometric mean is used to measure the rate of change of a variable over time. For a sample of n observations, it is equal to the n^{th} root of a product of n values:

$$\bar{x}_g = (x_1 x_2 ... x_n)^{\frac{1}{n}}$$

The geometric mean is used to calculate the average rate of return/growth over a series of time periods. In a sample of n of one-period return/growth $x_i = 1 + r_i$, the average rate of return/growth over n period is equal to

$$\bar{r}_g = \bar{x}_g - 1$$

1.3. Measures of dispersion

Range Denoted by x_{min} the smallest observation in the sample, and by x_{max} the largest observation. Then the range is defined as the difference of these two values, i.e.

$$\text{range} = x_{max} - x_{min}$$

Percentiles The idea behind percentiles is the same as with the median: certain intervals should contain certain fractions of the total number of observations. To determine the sample $100p$ percentile in a sample of size n, we must determine the data value such that

1. At least np of the data values are less than or equal to it.
2. At least $n(1-p)$ of the data values are greater than or equal to it.

Therefore, to find the sample $100p$ percentile of a data set

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
3. If np is an integer, then the average of the values in positions np and is the sample $100p$ percentile.

Interquartile range The sample $25th$ percentile is called the first quartile. The sample $50th$ percentile is called the median or the second quartile. The sample $75th$ percentile is called the third quartile. The interquartile range is difference between the third and the first quartile of a sample.

Variance The sample variance of x is defined by

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

where $x_i - \bar{x}$ are called deviations from the mean. The quantity

$$s_x = \sqrt[2]{s_x^2}$$

is called sample standard deviation.

1.4. Measures of relationship between variables

Consider a generic sample of n observations for two given variables x and y, i.e. $x_1, x_2, ...x_n$, and $y_1, y_2, ...y_n$.

Covariance The sample covariance between x and y is defined by

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where (x_i, y_i) are observations on a pair of variables

Correlation coefficient A sample correlation coefficient between x and y is defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} is the sample covariance and s_x and s_y are sample standard deviations. We say that the variables x and y are positively correlated if $r_{xy} > 0$ and negatively correlated if $r_{xy} < 0$. In case $r_{xy} = 0$ the variables are called uncorrelated. In the extreme cases where $r_{xy} = 1$ (-1) we talk about perfect positive (negative) correlation.

1.5. Exercises

Exercise 1 The time (in seconds) that a sample of $n = 10$ employees took to complete a task is: $\{14, 28, 40, 13, 25, 27, 20, 29, 49, 66\}$. Find the following for this data:

- Sample mean
- Sample median

- Sample variance
- Sample standard deviation
- Sample coefficient of variation

Exercise 2 Consider the following sample of prices of certain goods and quantities sold:

$$\text{Price} = \{10, 15, 20, 25, 30\}$$
$$\text{Quantity} = \{100, 90, 75, 50, 0\}$$

- Sketch a scatter plot of price (along the x-axis) and quantity (along the y-axis). Do you think these two variables will have a negative covariance, a positive covariance, or zero covariance?
- Compute and interpret the covariance and the correlation coefficient.

Exercise 3 Consider the following sample of observation for monthly income (in 1000GBP):

$$X = \{1.0, 1.0, 1.1, 1.1, 1.1, 1.2, 1.2, 1.4, 1.5, 1.6, 1.6, 2.2, 3.4, 7.6, 10.6, 26.4\}$$

Find the following for this data:

- Sample mean
- Sample median
- Sample mode
- Sample range
- Sample variance
- Sample standard deviation.

Plot the relative frequency distribution and the cumulative relative frequency distribution.