Chapter 3: Descriptive Measures

November 27, 2019

Section 3.1

Mean of a Data Set

The mean of a data set is the sum of the observations divided by the number of observations, also known as the arithmetic mean.

Median of a Data Set

The median of a data set is the middle value in its ordered list. Arrange the data in increasing order.

- If the number of observations is odd, then the median is the observations exactly in the middle of the ordered list.
- If the number of observations is even, then the median is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position $(n + 1)/2$ in the ordered list.

Mode of a Data Set

The mode of a data set is the most frequently occuring value. Find the frequency of each value in the data set.

- If no value occurs more than once, then the data set has no mode, or all values are modes.
- Otherwise, any value that occurs with the greatest frequency is a mode of the data set

Sample Mean

For a variable x,the mean of the observations for a sample is called a sample mean and is denoted $\bar{x}$. Symbolically,

$$\bar{x} = \frac{\sum x_i}{n}, \tag{1}$$

where n is the sample size.

Section 3.2

Range of a Data Set

The range of a data set is the difference between its largest and smallest values. The range of a data set is given by the formula

$$Range = Max - Min, \tag{2}$$

where Max and Min denote the maximum and minimum observations, respectively.

Sample Standard Deviation

For a variable x, the standard deviation of the observations for a sample is called a sample standard deviation. It is denoted s_x, or, when no confusion will arise, simply s. We have

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, \tag{3}$$

where n is the sample size and $barx$ is the sample mean.

Roughly speaking, the sample standard deviation indicates how far, on average, the observations in the sample are from the mean of the sample.

Variation and the Standard Deviation

The more variation that there is in a data set, the larger is its standard deviation.

Computing Formula for a SampLe Standard Deviation

A sample standard deviation can be computed using the formula

$$s = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}} \tag{4}$$

where n is the sample size.

Three-Standard Deviation Rule

Almost all the observations in any data set lie within three standard deviation to the either side of the mean.

Section 3.3

Chebyshev's Rule

For any the quantitative data set and any real number k greater than or equal to 1, at least $1 - 1/k^2$ of the observations lie within k standard deviations to either side of the mean, that is, between $\bar{x} - k * s$ and $\bar{x} + k * s$.

Empirical Rule

For any quantitative data set with roughly a bell-shaped distribution, the following properties hold.

- Property 1: Approximately 68% of the observations lie within one standard deviation to either side of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$.
- Property 2: Approximately 95% of the observations lie within two standard deviations to either side of the mean, that is, between $\bar{x} - 2s$ and $\bar{x} + 2s$.
- Property 3: Approximately 99.7% of th observations lie withing three standard deviations to either side of the mean, that is, between $\bar{x} - 3s$ and $barx + 3s$.

Section 3.4

Quartiles

Quartiles divide a data set into quarters (four equal parts).

- Q1: The first quartile is the median of the bottom half of the data set.
- Q2: The second quartile is the median of the entire data set.
- Q3: The third is the median of the top half of the data set.

Procedure for Quartiles

- Step 1: Arrange the data in increasing order.
- Step 2: Find the median of the entire data set. This value is the second quartile, Q_2
- Step 3: Divide the ordered data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.
- Step 4: Find the median of the bottom half of the data set. This value is the first quartile, Q_1.
- Step 5: Find the median of the top half of the data set. This is the third quartile, Q_3.
- Step 6: Summarize the results.

Interquartile Range

The interquartile range, or IQR, is the difference between the first and third quartiles, that is, $IQR = Q_3 - Q_1$. Roughly speaking, the IQR gives the range of the middle 50% of the observations.

Five-Number Summary

The five-number summary of a data set consists of the minimum, maximum, and the quartiles, in increasing order.

$$Min, Q_1, Q_2, Q_3, Max.$$

Lower and Upper Limits

The lower limit is the number that lies 1.5 IQRs below the first quartile; the upper limit is the number that lies 1.5 IQRs above the third quartile.

$$Lowerlimit = Q_1 - 1.5 * IQR;$$
$$Upperlimit = Q_3 + 1.5 * IQR.$$

Procedure for Boxplots

- Step 1; Determine the quartiles.
- Step 2: Determine the outliers and adjacent values.
- Step 3: Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and adjacent values with vertical lines.
- Step 4: Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.
- Step 5: Plot each potential outlier with an asterisk.

Note:

- In a boxplot, the two lines emanating from the box are called whiskers.
- Boxplots are frequently drawn vertically instead of horizontally.
- Symbols other than an asterisk are often used to plot potential outliers.

Section 3.5

Population Mean (Mean of a Variable)

A population mean is the arithmetic mean of a population data (variable). For a variable x, the mean of all possible observations for the entire population is called the population mean or mean of the variable x. It is denoted μ_x or, when no confusion will arise, simply μ. For a finite population,

$$\mu = \frac{\sum x_i}{N},$$

where N is the population size.

Population Standard Deviation (Standard Deviation of a Variable)

For a variable x, the standard deviation of all possible observations for the entire population is called the population standard deviation or standard deviation of the variable x. It is denoted σ_x or, when no confusion will arise, simply σ. For a finite population, the defining formula is

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

where N is the population.

The population standard deviation can also be found from the computing formula

$$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \mu}$$

Parameters and Statistics

- Parameter: A descriptive measure for a population.
- Statistic: A decriptive measure for a sample.

Standardized Variable

For a variable x, the variable

$$z = \frac{x - \mu}{\sigma},$$

is called the standardized version of x or the standardized variable corresponding to the variable x.

z=Score

For an observed value of a variable x, the corresponding value of the standardized variable z is called the z-score of the observation. The term standard score is often used instead of z-score. The z-score of an observation tells us the number of standard deviations that the observation is from the mean, that is, how far the observation is from the mean in units of standard deviations.