FTDS // DESCRIPTIVE STATISTICS

Hacktiv8 DS Curriculum Team

Phase 0 Learning Materials

Hacktiv8 DS
Curriculum
Team

Contents

Objectives

- able to grasp the concept of Descriptive Statistics
- able to understand various type of measurements
- able to understand concept of Skewness and Kurtosis
- able to grasp the concept of Five Number Summary
- able to understand Box and Whisker Plot
- able to understand Outliers
- able to understand Covariance & Correlation

Definition

- **Descriptive Statistics** is methods that deals with collecting, organizing, representing, and presenting data using tables, graphs, and other numerical parameters to provide useful information.

- **Inferential Statistics** is methods related to drawing conclusions or estimating the population from the sample through hypothesis testing and statistical testing by utilizing information from samples generated from Descriptive Statistics.

Measurement Types

- Mean : Sum of the values in the dataset, divided by the number of observations in the dataset.
- Median : The middle-most value after the values are sorted ascendingly.
- Mode : The most frequently occurring value.
- Skewness : measure of a dataset’s symmetry – or lack of symmetry.
- Kurtosis : measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

Skewness

- Right-skewed :
  - ✓ Data is bunched together on the left
  - ✓ Creating a long tail on the right
  - ✓ Pulling the mean > median to the right
  - ✓ Skew value is positive

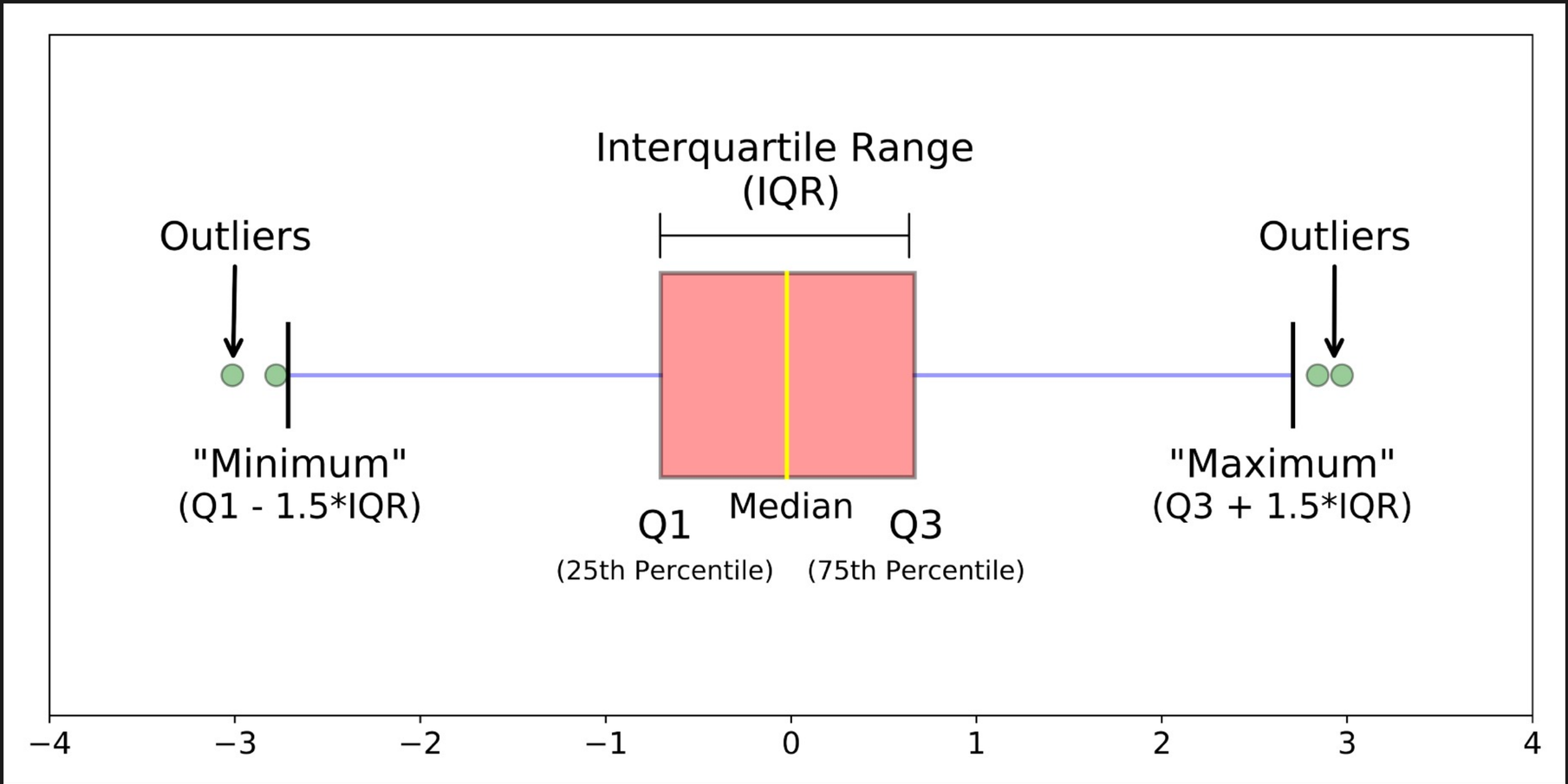
- Left-skewed :
  - ✓ Data is bunched together on the right
  - ✓ Creating a long tail on the left
  - ✓ Pulling the mean < median to the left
  - ✓ Skew value is negative

Skewness

- A perfectly symmetrical data set will have a skewness of 0. The normal distribution has a skewness of 0.

- So, when is the skewness too much? The rule of thumb seems to be:
  - ✓ If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
  - ✓ If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
  - ✓ If the skewness is less than -1 or greater than 1, the data are highly skewed.

Box and Whisker Plot



Five Number Summary

The five-number summary involves the calculation of 5 summary statistical quantities, namely :

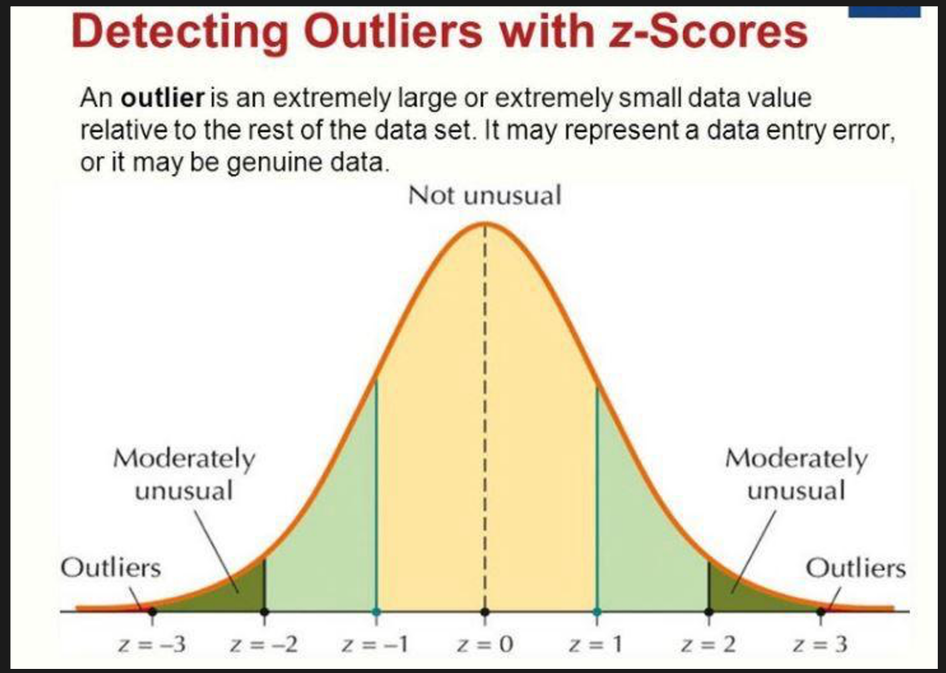
1. Minimum : The smallest observation in the sample.
2. Median : The middle value in the sample, also called the 50th percentile or the 2nd quartile.
3. 1st Quartile : The 25th percentile.
4. 3rd Quartile : The 75th percentile.
5. Maximum : The largest observation in the sample.

How to Detect Outliers?

Using Extreme Value Analysis.

- If the the variable is Normally distributed (Gaussian), then the values that lie outside the mean plus or minus 3 times the standard deviation of the variable are considered outliers.

- *Outliers = mean +/- 3* std.*



How to Detect Outliers?

Using Extreme Value Analysis.

- If the variable is skewed distributed, a general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:

$$IQR = 75th\ quantile - 25th\ quantile$$

- An outlier will sit outside the following upper and lower boundaries:

$$Upper\ boundary = 75th\ quantile + (IQR * 1.5)$$
$$Lower\ boundary = 25th\ quantile - (IQR * 1.5)$$

- Or for extreme cases:

$$Upper\ boundary = 75th\ quantile + (IQR * 3)$$
$$Lower\ boundary = 25th\ quantile - (IQR * 3)$$

How to Handle Outliers?

- Trimming

  Remove the outliers from our dataset.

- Censoring

  Capping the variable distribution at a max and / or minimum value. Censoring is also known as top and bottom coding, winsorization, or capping.

Covariance & Correlation

- Covariance and correlation refers to the measure of how two variables in a data set will change together or related to one another.

- Covariance value : - ∞ to ∞.

- Correlation value : -1 to 1.

Covariance & Correlation

- +1 :
  - ✓ If an increase in one variable results in an increase in the other variable.
  - ✓ Decreases in one variable also cause a decrease in the other.
  - ✓ Both variables move together in the same direction when they change.
- 0 :
  - ✓ No relationship between 2 variables.
  - ✓ If a line is drawn, it will be in a horizontal line.
- -1 :
  - ✓ Decreases in one variable resulting in the opposite change in the other variable.
  - ✓ These variables are inversely related and always move in different directions.

External References

Colab Link

Visit Here