Descriptive Statistic

Created	@Jan 18, 2021 1:40 PM
Created By	Khanh Vương
Last Edited By	Khanh Vương
Last Edited Time	@Jan 24, 2021 11:54 PM
Module	Introduction to Machine Learning
Status	
Type	Introduction to Machine Learning

- **Descriptive Statistic** is the process that describe the data
- We will consider some **Case**, which is a real-life phenomenon. Then, we will extract all of it's **Variable**
- One thing to notice that, **Variables** are the properties that must vary. Otherwise, it will be a constant
- Each element in the **Case** will be consider as an **Observation**

Describe the data

- Before describing the data, we must first determine the type of the data. There are two main types:
  - **Categorical:**
    - **Nominal Level**: All of the variables are name only, and we cannot consider which one is more important
    - **Ordinal Level**: We can rank all of the variables. These are usually the rankings
  - **Quantitative:** *(Number only)*
    - **Interval**: We have a range show how much a variable is better or worse than other variable
    - **Ratio:** Similar to **Interval**, but have the meaning at 0 point
  - Because **Quantitative** variables are number only, so it has two types of number:
    - Discrete
    - Continuous

- After determine the type of the data, we can convert the data table into frequency table because frequency table can give us an overview of all variables. If we meet continuous values, think of recode it and zip it into **Ordinal Level**, like *from x to y*. Note that this will cause missing information, and this is one way zipping, we cannot re-convert it again.

Plot the Graph

- Next, we will need to describe the data. The best way is show it on a graph:
  - **Categorical Variables**:
    - **Pie Chart** is better if we want to display percentages. However, it can not show the exact number.
    - **Bar Graph** allows us to easily extract the exact number.

    - **Bar Graph** is better than **Pie Chart** if the number of variables increase.
- **Quantitative Variable:**
    - **Dot** method should only be used when we have a very little amount of variable.
    - **Histogram** is nearly the same with **Bar Graph.** However, columns in ****Bar Graph**** have distance between t

Describe the Graph

- After having the graph, look at the shape of the graph to check whether it is:
    - **Bi-Model**
    - **Uni-Model**
    - **Left-Skewed**
    - **Right-Skewed**

- **Determine the center of the graph**:
    - **mode**: Value that occur most frequently. When can have many modes in a same graph. Often used if a variable is measured on a **Nominal** or **Ordinal Level** because we cannot apply calculations on these variables.
    - **median**: The middle value. Note that this is for the value, not the label. If we have an even number (`a % 2 == 0`), then to take the **median**, we simply take the average of the two middle value.
    - **mean**: The average. Note that, the **mean** is very easily to be affected by outlier.

- **Check for the range**:
    - **Range**: The distance between the highest and lowest value in the graph
    - However, the range sometimes cannot show us exactly the density of the graph. Instead, use **Interquartile Range** by split the graph into four parts by three **Quartile**
        - Q_2 is the **median** of the graph, divide the graph into two parts
        - Q_1, Q_3 is the **median** of each parts
        - $IQR = Q_3 - Q_1$
    - With **Interquartile Range**, we can check what the values of outliers are:
        - Lower than $Q_1 - 1.5(IQR)$
        - Greater than $Q_3 + 1.5(IQR)$
    - To show the variability with outliers, we can use **Box Plot**

- **Describe the variability:**
    - **Variance** show us how much the variables are spread out from the **mean**

$$S^2 = \frac{\sum (x - mean)^2}{n - 1}$$

        - Note that, we will use $(x - \mu)^2$ because μ is the balance point of the graph, so the negative and positive value are countered by each other

    - **Standard Derivation** show us the average distance of an observation from the μ

$$S = \sqrt{\frac{\sum (x - avg(x))^2}{n - 1}}$$

- **Standardize**:

- The **Z-Score** show the distance between this value and the μ, divided by the **Standard Derivation** (the number of **Standard Derivation** that removed from the μ)

$$Z = \frac{x - mean}{S}$$