UNIT

04 DESCRIPTIVE STATISTICS

Huge amount of data has been collected over the years from different domains. In order for these data to uncover, statistical analysis may be applied. The data must be analyze in order to: relate it to other pieces of data, summarize by a certain categories or to different areas and visualize it or display everything accurately. There are a lot of things that can be done to an analyzed data, and this concept represent descriptive analytics.

This unit covers the discussion about descriptive analytics. It will be followed by outlining the different type of data that will be used in analysis, which is the first task in data mining process. These include nominal, ordinal, internal and ratio type of attributes. Basic statistical descriptions can be used to learn more about each attribute's values. Mean, median and mode are the measures of central tendency, which describe the central position of a data on a given dataset. Measures of dispersion of variation of data or simply measures of variability denote the width of the distribution of the data. Understanding the variability of the distribution may provide information lacking by the central tendency. There are different ways of measuring variability and these are range, interquartile, variance and standard deviation.

LESSON 1:

DESCRIPTIVE STATISTICS

OBJECTIVES:

At the end of this lesson, the student will be able to:

- Describe descriptive analytics as a statistical procedure in analysing data
- Understand the importance of descriptive analytics as a statistical method
- Illustrate an example of descriptive analytics conducted by an analysts or statistician.

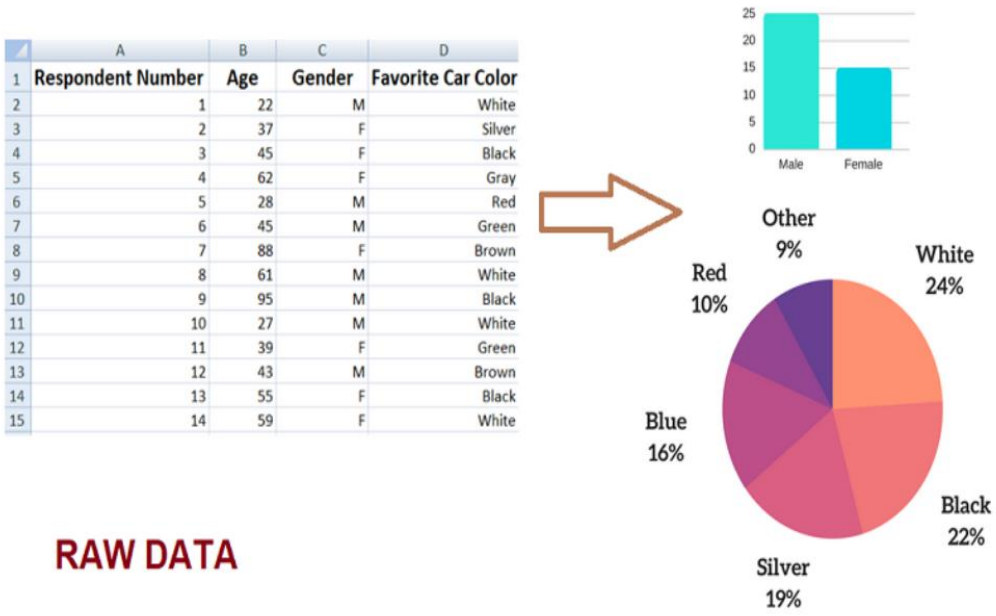
Duration: 2 hours

Defining Descriptive Analytics

Descriptive analytics is a statistical technique used to search and summarize historical data to identify patterns or meaning. It is an initial stage in data processing which creates a summary of historical data to obtain useful information and possibly prepare the data for further analysis.

Data mining and data aggregation methods are used to organize data and make it possible to identify patterns and relationships in it. Cleaning, relating, summarizing, reporting and data visualization may be applied to yield more insight in descriptive analytics.

Descriptive analytics is a conventional form of Business Intelligence (BI) and data analysis. It aims to provide a picture or summary view of facts and figures in an understandable format, to either inform or prepare data for further analysis. It uses two primary techniques, namely data aggregation and data mining to report past events. It presents past data in an easily digestible format for the benefit of a wide business audience. Refer to Figure 1 as an example.



	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

Figure 4.1 Sample of Descriptive Analytics
Source: http://www.intellspot.com/descriptive-statistics-examples/

Different types of descriptive analytics will be discuss later on. The most common measures that is being utilized by an analysts or a statistician are: measures of central tendency, measures of dispersion or variability and measures of position. Measures of central tendency include the mean, median, and mode, while measures of dispersion or variability include the standard deviation, variance, range, and interquartile range.

People use descriptive statistics to repurpose hard-to-understand quantitative insights across a large data set into bite-sized descriptions. A student's grade point average (GPA), for example, provides a good understanding of descriptive statistics. The idea of a GPA is that it takes data points from a wide range of exams, classes,

and grades, and averages them together to provide a general understanding of a student's overall academic abilities. A student's personal GPA reflects his mean academic performance (Kenton, 2020).

Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question “What happened?” (or What is happening?), characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives (Gartner_Inc, 2020).

Even without knowing it, many organizations use descriptive analytics extensively in their everyday operations. For most businesses, descriptive analytics form the core of their everyday reporting. This includes simpler reports such as inventory, workflow, warehousing, and sales, which can be aggregated easily and provide a clear picture of a company’s operations. One straightforward example of how descriptive analytics are used in operations revolves around annual revenue reports. (What is Descriptive Analytics? 2020).

LESSON 2:

CENTRAL TENDENCY

OBJECTIVES:

At the end of this lesson, the student will be able to:

- Differentiate the different measures of central tendency.
- Compute the mean, median and mode given a particular data sets.
- Understand the importance of the measures of central tendency

Duration : 3 hours

Measures of Central Tendency

The measure of central tendency can be defined as a descriptive statistical method which describes or shows the center value in a dataset. It can be referred to as the measure of central location where most values in a distribution fall. The mean, median and mode, these are the common measures of central tendency. Each measures has different method of calculating the location of the central point. In choosing what measure of central tendency to use would really depends on the type of data being given.

Measures of central tendency is also one of the most common measures used in statistics. It can provide a comprehensive summary of the dataset, but it does not deliver information about the individual values in the dataset.

Mean (Average)

Mean is a representation of the sum of all values in a dataset divided by the total number of the values. The mean is the most common measure of central tendency. It has a formula of:

$$\overline{X} = \frac{\sum X}{N}$$

Where:

X – represents the mean,

∑x – represents the summation of all scores or values, and

N – represents the number of cases.

Example of calculating mean:

These are the scores of first year students in a History test,

10 5 9 8 6 5 9 8 7 6 5 6

1. Add all scores or values ; 10+5+9+8++6+5+9+8+7+6+5+6= 84
2. Divide the result by the number of cases (or the number of scores): 12
3. Compute the mean using the formula:
   X= 84/12= 7

The average or mean is 7.

Example of calculating the mean using a frequency table.

Given a table of frequencies of the scores that is obtained in a History test. The first column is the test scores, and the other column is the frequency the number of students obtained that score.

X (score)	Frequency
10	1
5	3
8	2
2	5
4	5

1. Multiply each test score by its frequency, it will calculate the sum of all scores: 10X1+5x3+8x2+2x5+4X5= 71
2. Divide the sum of all scores by the sum of all the frequencies: 1+3+2+5+5= 16
3. Applying the formula: 71/16= 4.43

Calculating the mean incorporate all values in the dataset. If there is change to any value, the mean result also changes.

Median

The median is simply the middle value in a dataset. In the case where the dataset has even number of values, the median of that dataset is the average or mean of the two middle values. It is the location in which half of the values are above, and half of the values are below. Median is a preferred measure of central tendency when a distribution has extreme values. In order to look for the median values should be sorted first from lowest to highest.

Example:

These are the scores of the first year students in a History test:

10 5 9 8 6 5 9 8 7 6 5 6

1. Values should be sorted first from lowest to highest: 5 5 5 6 6 6 7 8 8 9 9 10
2. In this particular case, the number of values or scores are even; so, the median is the mean of the two middle numbers: 6+7/2= 6.5

The median is equal to 6.5

Mode

Mode is defined as the most recurrently occurring value in a dataset. Some dataset may contain multiple modes while in some may not have any mode at all. It is a measure of central tendency with largest frequency in a table.

Example the scores of first year students in a History test:

10 5 9 8 6 5 9 8 7 6 5 6

The mode is 6- which is the most frequent score in the distribution.

LESSON 3:

VARIATION

OBJECTIVES:

At the end of this lesson, the student will be able to:

- Differentiate the different measures of variation or dispersion.
- Compute the variance, range, interquartile and standard deviation given a particular data sets.
- Understand the importance of the measures of variation.

Duration: 3 hours

Measures of Dispersion or Variability

The measure of dispersion or variability refers to how scattered a group of data is. Consider the graph below, these are graph represented by the scores of two quizzes. Quiz1 and Quiz 2 have the same mean result which is 7.0. Even though they have equal means, their distribution is different. In Quiz 1, the scores are more closely distributed but in Quiz 2 they are more scattered or "spread out". The differences between the scores among students were greater in Quiz 2 when compared on Quiz 1.

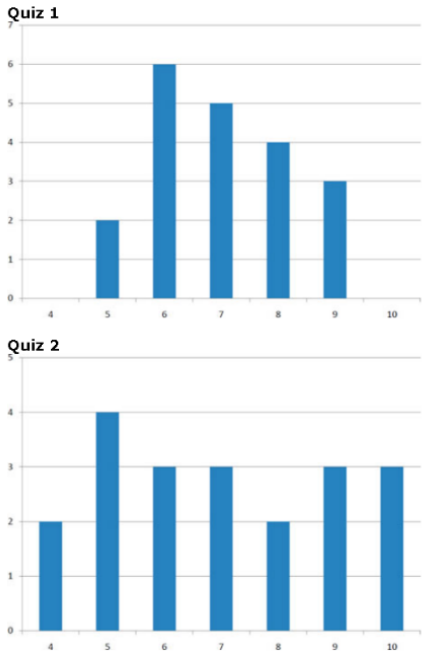


Figure 4.2. Bar Charts of two quizzes.
Source: http://onlinestatbook.com/2/summarizing_distributions/variability.html

Variability and dispersion are some of the terms that describes how spread out a certain distribution is. Measures of variability shows how much the data differs or vary from the average distribution. And the most common measures of variability that is being used are range, interquartile, variance and standard deviation.

Characteristics of Measures of Variability or Dispersion

1. Should be rigidly defined
2. Should be simple and easy to calculate and understand
3. Should not be affected sampling fluctuations and extreme values
4. Should be based on all observations

Categories of Measures of Variability or Dispersion

A. Absolute measure of dispersion:

- A measure which expresses the scattering of observation in terms of distances i.e., range, quartile deviation.
- And a measure which expresses the variations in terms of the average of deviations of observations like mean deviation and standard deviation.

B. Relative measure of dispersion:

Relative measure of dispersion is used for comparing distributions of two or more dataset and for unit free comparison. They are the coefficient of range, the coefficient of mean deviation, the coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.

Range

The simplest measure of variability and dispersion to calculate is the range. It is easy to calculate and easy to understand. It just simply the difference between the highest and the lowest score in a dataset.

Formula of Range:

Range = X max – X min

Where

Xmax – highest score

Xmin – lowest score

Example:

Given the following test score,

10, 2, 5, 6, 7, 3, 4

Using the formula,

Range = 10-2

Range is equal to 8.

In Figure, Quiz 1 has a highest score of 9 and the lowest score is 5, therefore the range is 4. On Quiz 2, the highest score is 10 and the lowest is 6, then range is equal to 6.

Some of the disadvantages of using range is that it is only based on two extreme values that is why it is affected by fluctuations. It is also considered as unreliable measures of variable or dispersion

Interquartile Range

The Interquartile Range or IQR is a measure of dispersion or variation based on distributing a data set into quartiles. Quartiles means dividing the dataset into four equal parts. These values will be separated in parts called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

In the concept of descriptive statistics, the interquartile range (IQR) is thought of as the midspread or the H-spread. It is the range of the middle 50% of the scores in a

distribution. It is equal to the difference between the 75^{th} and 25^{th} percentiles, or the upper and lower quartiles. It has a formula of:

IQR = Q3 − Q1

Referring to the Figure on Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. And for Quiz 2, which is more scattered, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4.

Variance

Variance can be defined as an average of the squared differences of the scores from the computed mean. In descriptive statistics, variance is a measurement of how far each number in the dataset is from the mean and from every other number in the dataset.

The steps of calculating variance:

1. Compute for the mean (average of the numbers)
2. For each number: get the difference between the mean and the number, then square the result (squared difference).
3. Then work out the average of those squared differences.

The data from Quiz 1 in our example are shown on the Table below. The mean is equal 7.0. Thus, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is the previous column squared.

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
Means		
7	0	1.5

Table 1. Calculating the Variance for Quiz 1 scores

Source: http://onlinestatbook.com/2/summarizing_distributions/variability.html

It is important to take note that the mean deviation from the mean is 0. It will always be that way. The mean of the squared deviations is given which is 1.5. Then, the variance is 1.5. Similar calculations with Quiz 2 show that the variance is 6.7. The

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

formula for the variance is:

where

σ2 represents the variance,

μ represents the mean, and

N represents the number of values or numbers.

On Quiz 1, μ = 7 and N = 20.

In the case where variance is used to estimate the variance in a population, then the

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

formula to be used is:

where

s2 represents the estimate of the variance and

M represents the sample mean.

Take note that M is the mean of a sample taken from a population with a mean of μ. In practice, the variance is usually computed in a sample, so this formula is most the most commonly used.

Example

Assume that the scores 1, 2, 4, and 5 were sampled from a bigger population. To estimate the variance in the population, s2 will be needed to compute:

M = (1 + 2 + 4 + 5)/4 = 12/4 = 3.

s2 = [(1-3)2 + (2-3)2 + (4-3)2 + (5-3)2]/(4-1)

= (4 + 1 + 1 + 4)/3 = 10/3 = 3.333

There are alternate formulas that can also be used and is easier to use for with a hand calculator. Remember that these formulas are subject to rounding error if the

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

values are very large or there is an extremely large number of observations.

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

and

For this example,

$$\sum X^2 = 1^2 + 2^2 + 4^2 + 5^2 = 46$$

$$\frac{(\sum X)^2}{N} = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36$$

$$\sigma^2 = \frac{(46-36)}{4} = 2.5$$

$$s^2 = \frac{(46-36)}{3} = 3.333 \text{ as with the other formula}$$

Standard Deviation

The standard deviation is a measure of dispersion or variation that measures the difference between each data point and the mean. When the values in a dataset are closely distributed, the standard deviation is smaller. But when the values in a data set are scattered, the standard deviation is larger for the reason that the distance is greater. The standard deviation is the square root of the variance. The

variance is in squared units. Henceforth, the square root returns the value to the natural units.

From the example in Figure, the standard deviations of the two quiz distributions 1.225 and 2.588 for Quiz 1 and Quiz 2 respectively. The standard deviation is a useful measure of dispersion or variability when the distribution is normal or almost normal because the amount of the distribution within a given number of standard deviations from the mean can be calculated.

Say for example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean.

Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between 50 - 10 = 40 and 50 +10 =60. Similarly, about 95% of the distribution would be between 50 - 2 x 10 = 30 and 50 + 2 x 10 = 70.

The symbol for the population of standard deviation is σ; the symbol for the estimate computed in a sample is s. In Figure , it shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 35 and 45; for the blue distribution, 68% is between 50 and 70.