Business Analytics

The Science of Data-Driven Decision Making

U Dinesh Kumar

WILEY

Descriptive Analytics 2

"The Purpose of Visualization is Insight – Not Pictures".

—**Ben Shneiderman**

LEARNING OBJECTIVES

- **LO 2-1** Understand the basic concepts in descriptive analytics and how it is used in data-driven decision making.
- **LO 2-2** Learn different variable types such as qualitative and quantitative along with scales of measurement such as nominal, ordinal, interval and ratio.
- **LO 2-3** Understand data types such as cross-sectional data, time series data and panel data.
- **LO 2-4** Understand the difference between population and sample and gain insights through fundamental concepts in statistics such as measures of central tendency, measures of variability and measures of shape.
- **LO 2-5** Learn data visualization and various types of visual charts.
- **LO 2-6** Understand the application of descriptive analytics in decision making.

ESSENCE OF DESCRIPTIVE ANALYTICS

Descriptive analytics is about finding "what has happened" by summarizing the data using innovative methods and analysing the past data using simple queries. Analysing past data can provide insights that can assist organizations to take appropriate decisions. Consider the Walmart example discussed in Chapter 1, where they found that during the hurricane season the demand for strawberry pop-tart increased seven times the normal season; this is a very good example for application of descriptive statistics. Based on this insight, Walmart ensured that there is enough stock of strawberry pop-tarts in the stores during a hurricane season. John Snow's spot map on Cholera outbreak in London and his final hypothesis that Cholera is water-borne disease is another classic example of application of descriptive analytics through data visualization. There are many such examples where simple analysis of the past data has revealed interesting facts such as difference in shopping behaviour of men and women, relationship freeze, etc. Descriptive analytics involves data summarization – using techniques such as pivot tables, descriptive statistics and data visualization that can be used for analysing past data to gain insights and hidden patterns.

IMPORTANT

Descriptive analytics is the starting point of analytics based solution to problems. It helps to understand the data and provide directions for predictive and prescriptive analytics. Business Intelligence (BI), which largely involves creating reports and business dashboard that lead to actionable insights, is essentially a descriptive analytics exercise.

2.1 | INTRODUCTION TO DESCRIPTIVE ANALYTICS

Descriptive analytics is the science of describing past data and thus capturing "what happened" in a given context. Primary objective of descriptive analytics is simple comprehension of data using data summarization, basic statistical measures and visualization. Various tools and techniques are used in describing the data. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape can provide useful insights. Many different plots such as histogram, bar chart, pie-chart, box-plot, scatter plot and tree diagram can provide insights about past data and subsequently assist with further analysis by generating new hypotheses.

Descriptive analytics is an important part of reporting across several industries which enables top management to monitor key performance indicators and take decisions. Most companies generate reports and dashboards at regular intervals as part of business intelligence (BI) to communicate various aspects of the business to the top management, stakeholders, and the external world. Business reports include descriptive analytics in the form of tables, charts, and innovative diagrams such as Treemap. With the advent of mobile technology, many real-time reports are generated and are accessed by the top management in their mobile handsets enabling them to take quick actions if necessary. For example, a retailer such as Bigbazaar or Reliance retail in India may like to know the top 5 (in terms of revenue generated) products that are sold by region, by city, by store, etc. Such information would assist the management to plan their inventory, shelf space, pricing, etc. They can also monitor trend in revenue generated at regional, city, and store levels over the past several periods. Several companies use dashboards and scorecards to communicate KPIs that are relevant to them; one of the primary applications of descriptive analytics is designing effective dashboards and scorecards.

2.2| DATA TYPES AND SCALES

Data is classified into different categories based on data structure and scale of measurement of the variables.

2.2.1 | Structured and Unstructured Data

Data at a macro-level can be classified as structured and unstructured data. Structured data means that the data is described in a matrix form with labelled rows and columns. Any data that is not originally in the matrix form with rows and columns is an unstructured data. For example, e-mails, click streams, textual data, images (photos and images generated by medical devices), log data, and videos. Machine-generated data such as images generated by satellite, magnetic resonance imaging (MRI), electrocardiogram (ECG) and thermography are few examples of unstructured data. There is an increasing trend in

the generation of unstructured data due to social media platforms such as Facebook and YouTube and analysis of unstructured data is important for effective management. Internet of things (IoT) is another source unstructured data.

The importance of unstructured data in decision making has increased many folds in the recent past due to its applications to different sectors of the industry. For example, analysing social media data is important for companies to understand the sentiments expressed by the customers about their products/services and take necessary remedial measures. Significant proportion of social media data is natural language (text) apart from images and videos. Apart from social media, machine-generated data are usually unstructured (e.g. data generated from medical devices such as ECG, MRI, etc.). High percentage of Big Data problems constitute unstructured data. One of the main challenges in analysing unstructured data is in the conversion of unstructured data to structured data, which then enables model development. Examples of structured and unstructured data are shown in Tables 2.1 and 2.2.

The data in Table 2.2 is a clickstream data (search behaviour of an internet user that captures the websites visited by the user). Clickstream data is useful for understanding the behaviour of internet users. Based on their surfing (internet browsing) behaviour, individuals are targeted with advertisement for products and services. The unstructured data as shown in Table 2.2 does not have matrix structure as in the case of structured data in Table 2.1. Before any analytics model can be built, unstructured data has to be converted into a structured data.

TABLE 2.1 Structured data consisting of nominal and ratio scales

No.	Gender	Age	Percentage SSC	Board SSC	Percentage HSC	Percentage Degree	Salary
1	M	23	62	Others	88	52	270000
2	M	21	76.33	ICSE	75.33	75.48	220000
3	M	22	72	Others	78	66.63	240000
4	M	22	60	CBSE	63	58	250000
5	M	22	61	CBSE	55	54	180000
6	M	23	55	ICSE	64	50	300000
7	F	24	70	Others	54	65	240000
8	M	22	68	ICSE	77	72.5	235000
9	M	24	82.8	CBSE	70.6	69.3	425000
10	F	23	59	CBSE	74	59	240000

TABLE 2.2 Unstructured data (sample clickstream data)

https://en.wikipedia.org/wiki/Clickstream
http://hortonworks.com/hadoop-tutorial/how-to-visualize-website-clickstream-data/
http://searchcrm.techtarget.com/definition/clickstream-analysis
https://www.qubole.com/blog/big-data/clickstream-data-analysis/

2.2.2 | Cross-sectional, Time Series, and Panel Data

Another important classification of data is based on the type of data collected. Based on the type of data collected, the data is grouped into the following three classes:

1. **Cross-Sectional Data:** A data collected on many variables of interest at the same time or duration of time is called cross-sectional data. For example, consider data on movies such as budget, box-office collection, actors, directors, genre of the movie during year 2017.
2. **Time Series Data:** A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.
3. **Panel Data:** Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data). Example of a panel data is data collected on variables such as gross domestic product (GDP), Gini index, and unemployment rate for several countries over several years.

2.3 | TYPES OF DATA MEASUREMENT SCALES

Structured data can be either numeric or alpha numeric and may follow different scales of measurement (level of measurement). It is important to understand the type of variables within the data with respect to the measurement scale since the model specification while building analytics models such as regression may depend on the scale of measurement.

2.3.1 | Nominal Scale (Qualitative Data)

Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables. For example, variables such as marital status (single, married, divorced) and industry type (manufacturing, healthcare, banking and finance) fall under nominal scale. During data collection, it is usual to assign a numerical code to represent a nominal variable. For example, the data collector may have used number 1 to represent singles, 2 for married, and 3 for divorced category for categorical variable marital status. The codes 1, 2, and 3 used here do not have any value attached to them. That is, basic mathematical operations are meaningless in a nominal scale (e.g., subtraction: married – unmarried or ratio: married/unmarried are meaningless). While developing statistical models, nominal scale data are usually transformed before building the model. For example, when developing a regression model, categorical variables are converted using dummy variables before building the regression model (is discussed in Chapter 10).

2.3.2 | Ordinal Scale

Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude. For example, in many survey data, Likert scale is used. Likert scale is finite (usually a 5 point scale) and the data collector would have defined the order of preference. For example, assume that a feedback is collected on a training program using 5-point Likert scale in which 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent. In this case, we know that

5 is better than 4 and 4 is better than 3; however, the difference 5 – 4 (Excellent – Very Good) is meaningless.

2.3.3 | Interval Scale

Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade (°C) or intelligence quotient (IQ) score are examples of interval scale. In interval scale, the ratios do not make sense. For example, 40°C is not twice hot as 20°C. Similarly, a person with an IQ score of 160 is not twice smarter than a person with an IQ score of 80. However, 40°C is 20°C more than 20°C, IQ score of 160 is 80 more than an IQ score of 80. In an interval scale, the reference is fixed arbitrarily, for example 0°C is fixed based on the freezing point of water.

2.3.4 | Ratio Scale

Any variable for which the ratios can be computed and are meaningful is called ratio scale. Most variables come under this type; for example: demand for a product, market share of a brand, sales, salary, and so on. If Ms Hawai Sundari's salary is 40,000 per month and Ms Dawai Sundari's salary is 90,000 per month then we can interpret that Dawai Sundari earns 2.25 times the salary of Hawai Sundari.

2.4 | POPULATION AND SAMPLE

Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem. The size of the population can be very large in many cases. For example, in 2014, close to 834.08 million people were eligible to vote in the Indian general elections (Source: Election Commission of India). Thus, the population size of the voters in 2014 was 834.08 million which included all eligible voters. During every election, media and other organizations collect data to predict likely winner of election through opinion polls (and they rarely get it right due to complexities associated with collecting right sample). It is very difficult (also practically impossible) to collect data from all 834.08 million eligible voters about their choice of candidate, so the opinion polls are based on opinion expressed by a subset of voters called **sample**.

Population (also known as universal set) is the set of all possible data for a given context whereas sample is the subset taken from a population. In many analytical problems, we make inference about the population based on the sample data. There are many challenges in sampling (process of selecting an observation from the population). An incorrect sample may result in bias and incorrect inference about the population. Sampling is discussed in detail in Chapter 4.

2.5 | MEASURES OF CENTRAL TENDENCY

Measures of central tendency are the measures that are used for describing the data using a single value. **Mean, median** and **mode** are the three measures of central tendency and are frequently used to compare different data sets. Measures of central tendency help users to summarize and comprehend the data.

2.5.1 | Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency. Assume that the data has n observations in a sample, and let X_i be the value of the i^{th} observation. Then the mean is given by

$$\text{Mean} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{2.1}$$

Symbol $\bar{X}$ is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we get the population mean which is denoted by μ. Among the measures of central tendency, mean is the most frequently used measure since it uses all the observations (all X_i values) in the data set (either sample or population) to calculate the mean value. Table 2.1 has the salary of graduating students from a business school; the average salary is given by

$$\bar{X} = \frac{(270 + 220 + 240 + 250 + 180 + 300 + 240 + 235 + 425 + 240) \times 1000}{10} = 260000$$

The average (or mean) salary is 260000. Note that the average value need not be a part of the data set, that is, none of the graduating student's salary is 260000. In Microsoft Excel, function 'Average(array)' can be used for calculating the mean value of the data. Mean can be interpreted as the centre of gravity of the distribution of the data. An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

Associated with the mean is a phenomenon often called "wisdom of crowd", according to which the collective wisdom of people is better than any individual person's knowledge. For example, in 1906, Francis Galton attended a contest in Plymouth, UK in which the villagers were asked to guess the weight of an Ox, the one who guessed the closest won the prize. Around 800 villagers participated in the contest. Francis Galton found that the average of all the weights entered was very close to the actual weight. In fact, the difference was less than a pound. Also, the average turned to be better than the guess by the winner of the contest (Surowiecki, 2004).

One should be careful about taking decisions based on the mean value of the data. There is a famous joke in statistics which says that, "*if someone's head is in freezer and leg is in the oven, the average body temperature would be fine, but the person may not be alive*". Making decisions solely based on mean value is not advisable. In capital asset procurement such as procurement of fighter aircraft and weapons, defence services across the world use mean time between failures (MTBF) as one of the measures of system reliability (performance). However, MTBF (which is the mean value of the time between failure data) in itself is not a useful measure to assess the reliability of the asset and not very useful in taking operational decisions. It has to be used along with other measures and measures of variability for better understanding of the data. Another issue with mean is, it is affected significantly by presence of

outliers. That is, presence of an outlier can change the mean value significantly. If the data is captured in frequencies, then Eq. (2.2) can be used for calculating the average:

$$\bar{X} = \sum_{i=1}^{n} \frac{f_i X_i}{f_i} \tag{2.2}$$

The frequency of age of students in Table 2.1 is given below:

Age	21	22	23	24
Frequency	1	4	3	2

The average age of students using Eq. (2.2) is given by

$$\bar{X} = \frac{1\times 21 + 4\times 22 + 3\times 23 + 2\times 24}{1+4+3+2} = 22.6$$

2.5.2 | Median (or Mid) Value

Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%. Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position $(n + 1)/2$ when n is odd. When n is even, the median is the average value of $(n/2)^{th}$ and $(n + 2)/2^{th}$ observation after arranging the data in the increasing order.

Consider the example of a bank. The number of deposits in a branch of a bank in a week is shown in Table 2.3.

TABLE 2.3 Number of daily deposits in a Bank

Day	1	2	3	4	5	6	7
Number of Deposits	245	326	180	226	445	319	260

The ascending order of the data in Table 2.3 is given by

180, 226, 245, 260, 319, 326 and 445

Now $(n + 1)/2 = (8/2) = 4$. Thus the median is the 4^{th} value in the data after arranging them in the increasing order; in this case it is 260. There are equal numbers of observation below and above 260. In Microsoft Excel, the function 'Median(array)' can be used for calculating the median of a data set.

Another example is the salary in Table 2.1 that can be arranged as follows:

180000, 220000, 235000, 240000, 240000, 240000, 250000, 270000, 300000, 425000

The 5^{th} and 6^{th} observations are 240000 and 240000 and the average is 240000. Thus, the median salary for the data in Table 2.1 is 240000. Median is much more stable than the mean value, that is adding a new observation may not change the median significantly. However, the drawback of median is that it is not calculated using the entire data like in the case of mean. We are simply looking for the midpoint instead of using the actual values of the data.

2.5.3 | Mode

Mode is the most frequently occurring value in the data set. For example, in the data 'salary' in Table 2.1, the value 240000 is appearing three times and is the mode since all other values are observed only once. In Microsoft Excel, the function 'Mode(array)' can be used for calculating mode. Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless. For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database. In the bar chart (and histogram), mode is the tallest column. It is possible that a data set may not have any mode at all. For example, if each value in the data set appears only once, then there is no mode in the data set.

2.6 | PERCENTILE, DECILE, AND QUARTILE

Percentile, decile and quartile are frequently used to identify the position of the observation in the data set. Percentile score is frequently used in education to identify the position of a student in the group. Another frequent application of percentile is the percentile life used in asset management. Percentile, denoted as P_x, is the value of the data at which x percentage of the data lie below that value. For example, P_{10} denotes the value below which 10 percentage of the data lies. To find P_x, we have to arrange the data in the increasing order and the value of P_x is the position in the data calculated using Eq. (2.3):

$$\text{Position corresponding to } P_x \approx \frac{x(n+1)}{100} \tag{2.3}$$

where n is the number of observations in the data. Note that the value obtained from Eq. (2.3) can be non-integer, in which case we can either round it to the nearest integer or use an approximation which will be explained in Example 2.1. **Decile** corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on. Similarly, **Quartile** divides the data into 4 equal parts. The first quartile (Q_1) contains first 25% of the data, Q_2 contains 50% of the data and is also the median. Quartile 3 (Q_3) accounts for 75% of the data. In Microsoft Excel, the function 'Percentile(array, k)' provides P_x value. That is, Percentile(array, 0.1) will give 10th percentile.

EXAMPLE 2.1

Time between failures (in hours) of a wire cutter used in a cookie manufacturing oven is given in Table 2.4. The function of the wire-cut is to cut the dough into cookies of desired size.

TABLE 2.4 Time between failures of wire-cut (in hours)

2	22	32	39	46	56	76	79	88	93
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

(a) Calculate the mean, median, and mode of time between failures of wire-cuts.
(b) The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
(c) Calculate the values of P_{25} and P_{75}.

Solution:

(a) Mean = 57.64, median = 56, and mode = 46, 89 and 99.
(b) Note that the data in Table 2.4 is arranged in increasing order in columns. The position of $P_{10} = 10 \times (51)/100 = 5.1$. We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table 2.4 have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called P_{10} life.
Instead of rounding the value obtained from Eq. (2.3), we can use the following approximation:

$$\text{Position corresponding to } P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5th position is 21. Value at position 5.1 is approximated as
21 + 0.1 × (value at 6th position – value at 5th position) = 21 + 0.1(1) = 21.1

$$\text{Position corresponding to } P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90 and the value at position 45.9 is
90 + 0.9 (value at 46th position – value at 45th position) = 90 + 0.9 × (3) = 92.7
That is, 90% of the wire-cuts will fail by 92.7 hours.

(c) Position corresponding to P_{25} (1st Quartile or Q_1) = 25 × 51/100 = 12.75
Value at 12th position is 33, so
P_{25} = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75
Position corresponding to P_{75} (3rd Quartile or Q_3) = 75 × 51/100 = 38.25
Value at 38th position is 86, so
P_{75} = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

2.7 | MEASURES OF VARIATION

One of the primary objectives of analytics is to understand the variability in the data. Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X). Variability in the data is measured using the following measures:

1. Range
2. Inter-Quartile Distance (IQD)
3. Variance
4. Standard Deviation

Let us discuss each of them in detail.

2.7.1 | Range

Range is the difference between maximum and minimum value of the data. It captures the data spread. In the data in Table 2.4, the range = 102 – 2 = 100.

2.7.2 | Inter-Quartile Distance (IQD)

Inter-quartile distance (IQD), also called inter-quartile range (IQR), is a measure of the distance between Quartile 1 (Q_1) and Quartile 3 (Q_3). For the data in Table 2.4, we calculated Q_1 as 33.75 and Q_3 as 86. Thus the IQD = 86 – 33.75 = 52.25. IQD is a useful measure for identifying outliers in the data. Outlier is an observation which is far away (on either side) from the mean value of the data. Values of data below Q_1 – 1.5 IQD and above Q_3 + 1.5 IQD are classified as outliers.

For the data in Table 2.4

$$Q_1 - 1.5\ \text{IQD} = 33.75 - 1.5 \times 52.25 = -44.625$$
$$Q_3 + 1.5\ \text{IQD} = 86 + 1.5 \times 52.25 = 164.375$$

In Table 2.4, there are no values either below – 44.625 or above 164.375, thus there are no outliers. Note that IQD is one of the approaches used for identifying outliers; we will discuss other approaches that are used for identifying outliers in Chapters 9 and 10. Also, using IQD for identifying outliers is appropriate only in the case of univariate data (data with one dimension). In the case of multivariate data, we use distance measures such as Mahalanobis distance to identify outliers (discussed in Chapters 9 and 10).

2.7.3 | Variance and Standard Deviation

Variance is a measure of variability in the data from the mean value. Variance for population, σ^2, is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \tag{2.4}$$

Note that, in Eq. (2.4), deviation from mean is squared since sum of deviations from mean will always add up to zero. The variance for the data in Table 2.4 is 818.0304 [using Eq. (2.4)]. In case of a sample, the Sample Variance (S^2) is calculated using

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \tag{2.5}$$

While calculating sample variance S^2, the sum of squared deviation $\sum_{i=1}^{n}(X_i - \bar{X})^2$ is divided by $(n-1)$. This is known as Bessel's correction. For the data in Table 2.4, the sample standard variance is 834.7249. Microsoft Excel functions Var.P(array) and Var.S(array) are used for calculating population variance and sample variance, respectively. The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \tag{2.6}$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.6) is 28.6012.

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} \tag{2.7}$$

For the data in Table 2.4, the standard deviation obtained using the Eq. (2.7) is 28.8916. In Microsoft Excel, functions Stdev.P(array) and Stdev.S(array) are used for calculating population standard deviation and sample standard deviation respectively. There are two arguments for dividing the sum of squared deviations from mean by $(n-1)$ instead of n in Eqs. (2.5) and (2.7). One argument is that, when we take a sample and estimate the mean from the sample $\bar{X}$, we tend to underestimate the sum of squared deviations from the mean. For example, take a sample consisting of first 5 (first column) and last 5 (last column) observations from Table 2.4. The sample is given in Table 2.5.

TABLE 2.5 Sample of 10 observations from Table 2.4

2	3	5	9	21	93	99	99	99	102

The mean $\bar{X}$ for the sample in Table 2.5 is 53.2 and standard deviation [using Eq. (2.7)] is 47.9740. When we estimate the numerator, $(X_i - \mu)^2$, in Eq. (2.4) using $\bar{X}$, instead of μ, we will underestimate $(X_i - \mu)^2$ resulting in underestimation of standard deviation. The calculations of $(X_i - \bar{X})^2$ and $(X_i - \mu)^2$ for the sample in Table 2.5 are shown in Table 2.6.

TABLE 2.6 Underestimation of standard deviation in sample

Data	Standard deviation (using sample mean 53.2)	Standard deviation (using population mean 57.64)
2	2621.44	3095.81
3	2520.04	2985.53
5	2323.24	2770.97
9	1953.64	2365.85
21	1036.84	1342.49
93	1584.04	1250.33
99	2097.64	1710.65
99	2097.64	1710.65
99	2097.64	1710.65
102	2381.44	1967.81
Sample Mean = 53.2	$\sum(X_i - \bar{X})^2 = 20713.60$	$\sum(X_i - \mu)^2 = 20910.74$

In Table 2.6, we can see that the numerator in Eq. (2.4) is underestimated (20713.60) when we use the sample average against population average (20910.74). This will result in underestimation of the standard deviation, a phenomenon called **downward bias**. To overcome this bias, we divide $\sum(X_i - \bar{X})^2$ with $(n - 1)$ instead of n.

Another argument of using Eq. (2.5) is through the concept of **degrees of freedom**. The following two definitions are used for degrees of freedom (Pandey and Bright, 2008):

1. Degrees of freedom is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size n with mean value of $\bar{X}$ by randomly selecting $(n - 1)$ values. We need to fix just one out of n values. Thus the number of independent variables in this case is $(n - 1)$.
2. Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are n observations in the sample and k parameters are estimated from the sample, then the degrees of freedom is $(n - k)$. While using Eq. (2.5) or Eq. (2.7), the value of $\bar{X}$ is estimated from the sample. Thus the degrees of freedom is $(n - 1)$.

Whenever we estimate a parameter from a sample, we lose a degree of freedom. While estimating standard deviation from a sample, we tend to underestimate since mean is also estimated from the sample itself. The downward bias is addressed by dividing the sum of squared deviation from mean with $(n - 1)$ instead of n.