Chapter XIV: Fundamentals of Probability and Statistics*

Objectives

- Present fundamental concepts of probability and statistics
- Review measures of central tendency and dispersion
- Analyze methods and applications of descriptive statistics
- Review basic probability distributions.

Fundamentals of Probability and Statistics

An understanding of the fundamentals of probability and statistics is basic to describing and interpreting data. This knowledge is also most valuable in dealing with random events and risk which is inherent in business continuity planning.

Probability is a measure of the likelihood of occurrence of an event. It could refer to the chance that a new machine will work or not; or the chance that a hurricane will occur; or the chance that a neighborhood will be hit by a heavy snow-storm. There are several ways of calculating probability, and we will explore them later in this chapter.

Statistics is the technique used to collect, describe and analyze data. For example, in order to estimate the life expectancy of a machine, one can collect the time-to-failure of a number of similar machines and compute the average time-to failure. This measure, *average*, becomes a statistic. It gives one an idea of how long a machine will last. It is also possible to obtain the *range* of values for the life expectancy of a machine**.** The range will give us an idea of the minimum and maximum values. Such measures as *average***,** *range* and many other measures are known as *statistics***.** We will develop several other statistics in this chapter.

In the study of statistics, there are two basic concepts that are crucial – the concept of *population* and the concept of *sample***.** A *population* is the entity that we are interested in studying. However, it is usually not advisable to collect data on whole populations due to cost and time considerations. Normal practice is to take a *sample* from the *population***.** The data collected from the sample is used to develop sample characteristics known as *statistics***.** This will then be used to make inferences about the population characteristics known as *parameters*.

Data Classification

Discrete variables are variables whose outcomes are counted, for example, the number of requests for emergency response per day.

Continuous variables are variables whose outcomes are measured, for example, the time it takes to respond to an emergency call.

* Chapter prepared by Ore A. Soluade, PhD.

Nominal measurements have no meaningful rank order among values, for example, the classification of a weather event as a tornado, hurricane, or winter storm.
Ordinal measurements have imprecise differences between consecutive values, but have a meaningful order to those values, for example, the classification of a tornado as EF1, EF2, EF3, EF4, or EF5.

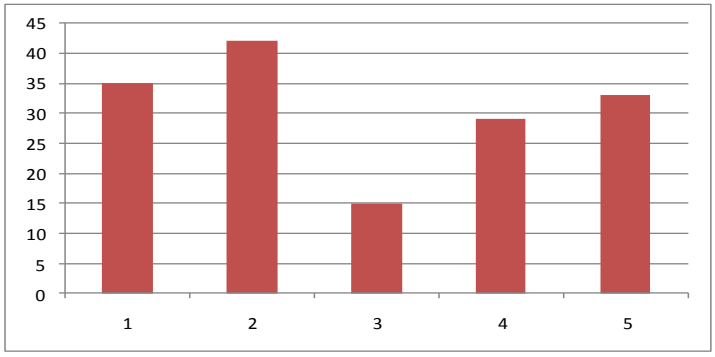
Interval measurements have meaningful distances between measurements defined, but have no meaningful zero value defined, for example, degrees Fahrenheit.

Ratio measurements have both a zero value defined and the distances between different measurements defined, for example, loss measured in dollars.

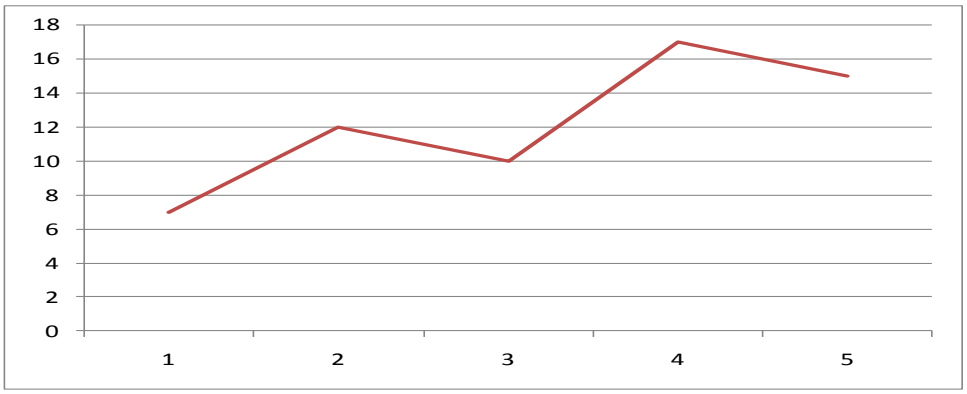
Graphical Presentation of Data

One powerful way of describing data is by displaying it on a chart. With the aid of Microsoft Excel, there are several options for displaying data graphically, depending on what information one is interested in highlighting.

For a set of discrete data, one type of graphical representation is a *bar chart*. An example of a bar chart is shown in the figure below:



A *line graph* displays information as a series of data points connected by straight line segments. Line graphs are particularly revealing if there is trend in the data, as illustrated in the figure below:



Stem and Leaf Plot

A stem and leaf plot is a graphical display of the data that lists them in ascending order and then displays the distribution within a given category. This is used as a preliminary description of the data before more detailed analysis is done. Assume we have the following set of data of the length of time (in minutes) of power outages in a city in the last year, already sorted as follows:

12	14	17	21	26	29	34	45	47	48	50	55	61

A stem-and-leaf plot of this set of data is as shown below:

1	2	4	7
2	1	6	9
3	4		
4	5	7	8
5	0	5	
6	1		

As can be seen, there are 3 data values in the teens, 3 data values in the twenties, 1 in the thirties, 3 in the forties, 2 in the fifties, and 1 in the sixties.

Frequency Distributions

Frequency distribution is a tabular summary of a data set showing the number of occurrences of each value or each class. This is true for both qualitative as well as quantitative data.

Given the following data:

12	14	17	12	26	29	34	45	17	17	50	55	50

The corresponding frequency distribution is as shown below:

x	Frequency
12	2
14	1
17	3
26	1
29	1
34	1
45	1
50	2
55	1

Measures of Central Tendency

Measures of central tendency of a set of data include the mean, the median, and the mode. The mean is the average value, the median is the middle value, and the mode is the value that occurs most frequently. There are two types of means – population mean, and sample mean.

Population Mean, μ, is given by:

$$\mu = \frac{x_1 + x_2 + x_3 \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where x_i = the value of the i^{th} item and N = population size.

Sample mean, $\bar{x}$, is given by:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where n = sample size.

When working with data that is summarized in a frequency distribution comprising classes of data, the computation of the mean of the distribution is calculated using the formula:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 \ldots + f_nx_n}{f_1 + f_2 + f_3 \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

where f_i = frequency count for x_i.

Quartiles and Percentiles

Quartiles are used to split a dataset into four equal parts. One can determine the lowest 25% of the data as values below the first quartile, the lowest 50% of the data as values below the second quartile (also known as the *median*), and the lowest 75% of the data as values below the third quartile.

The p^{th} *percentile* is a value such that p percent of the data are below this value.

… Continued…