CHAPTER 1

DESCRIPTION OF DATA

Objectives

After the completion of this chapter, the student should be able to calculate and interpret the following from a given sample of data:

(1) Measures of centrality: Arithmetic Mean, Mode, and Median. Relation between these three.
(2) Summary Measures: Quartiles, Deciles and Percentiles.
(3) Measures of variability: Range, Interquartile Range, Variance and Standard Deviation.
(4) Composite Summary Measures: Coefficient of Variation and the Coefficient of Skewness.
(5) Graphical representation of data: Box Plot, Histogram, Polygon and Ogive.

1.1 Introduction

Statistics is one of the most important branches of applied mathematics. It is used in fields like agriculture, physical sciences and human sciences. Descriptive statistics is mainly concerned with summary calculations and graphical displays of data.

In this module we will study the description of data obtained from samples.

1.2 Definitions

1.2.1 Population / Sample

A population is the total set of elements of interest for a given problem. A sample is a subset of a population. In later studies we use statistics obtained from samples to make predictions for the population the sample came from. This is called **Statistical Inference.**

1.2.2 The statistic

A statistic is a function of the sample data. Examples of a statistic are the arithmetic mean, mode, median and standard deviation. These will be defined in the following sections.

1.2.3 Arithmetic mean

The arithmetic mean is the sum of all the data divided by the number of observations. Thus if X_1, X_2, ..., X_n are n observations, then we write the arithmetic mean as follows:

$$\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$$

$$\bar{X}=\frac{\sum X_i}{n}$$

The Greek capital letter Σ (Sigma) is used to indicate the sum of a number of elements. The subscript i in formula (1.1) is a discrete variable which assumes the values 1, then 2, 3,... up to the last observation n. We will often use this notation.

1.2.4 The mode

The mode is the most frequently occurring value. If it is important to know which observation occurred most often, the mode is determined.

1.2.5 The median

The median is the middle value when the data are arranged in numerical sequence. Fifty percent of the observations are greater than the median and 50% are less.

1.3 Measures of deviation

All the previous statistics, the mean, mode and median are measures of central tendency, i.e. a number which tries to indicate a "central" value in a set of data. Statistics which describe the spread or variation of data are called measures of deviation or dispersion.

1.3.1 The range

The range is the simplest statistic which gives a indication of the spread of a set of numbers. It is simply equal to the difference between the lowest and highest value in a sample. The range has the serious disadvantage in that it only uses the extreme two values of the sample and ignores the information contained in the rest of the sample. The following statistics overcome this problem.

1.3.2 The variance

If we add all the squared differences between each sample value and the sample mean and divide this quantity by (n -1), we obtain the variance:

$$S^2=\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}$$

And if we use the Σ notation:

$$S^2=\frac{\sum(X_i-\bar{X})^2}{n-1}$$

The **standard deviation** S is the square root of the variance:

$$S=\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$$

1.3.3 **Example 1**

Consider the following data set:

X_i : 24, 13, 18, 35, 5, 28, 31, 24, 29

The mean:

$$\bar{X} = \frac{1}{n}\sum X_i$$
$$= \frac{24+13+...+29}{9}$$
$$= 23$$

The mode: the number 24 occurs twice, hence this is the mode.

The median: first arrange the data set in ascending order:

X_i : 5, 13, 18, 24, 24, 28, 29, 31, 35

↑

Hence the median is 24.

Note: If the set of data has an even number of elements we obtain the median by computing the mean of the two middle numbers.

The standard deviation:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$
$$= \frac{(5-23)^2+(13-23)^2+...+(35-23)^2}{9-1}$$
$$= 9{,}49$$

Remarks:

(1) The mode and median are equal in this case, but this is in general not true.

(2) Another data set Y_i might have a equal mean but a smaller or greater standard deviation, depending on the variation in the data set.

1.4 Grouped data

When summarizing large masses of raw data, it is often useful to distribute the data into classes or categories. The number of individuals belonging to each class is determined, and this is called the class frequency. A tabular arrangement of data by classes together with the corresponding class frequency is called a frequency distribution or frequency table. Table 1.1 is a sample of 60 determinations of the copper content of a standard solution.