5. Descriptive Statistics

5.1 Introduction

5.1.1 What is statistics?
5.1.2 Descriptive statistics
5.1.3 Population and sample
5.1.4 Parameter and statistic

5.2 Table Representation

5.2.1 Types of data
5.2.2 Raw data
5.2.3 Frequency distributions
5.2.4 Relative frequency and percentage distributions
5.2.5 Grouped frequency distributions
5.2.6 Constructing grouped frequency distribution tables

5.3 Graphical Representation

5.3.1 Bar graphs
5.3.2 Pie charts
5.3.3 Histograms
5.3.4 Polygons

5.4 Measures of Central Tendency

5.4.1 Mean
5.4.2 Median
5.4.3 Mode

5.5 Measure of Variability

5.5.1 Range
5.5.2 Variance and standard deviation

5.6 Mean, Variance and Standard Deviation for Grouped Data

5.6.1 Mean for grouped data
5.6.2 Variance and standard deviation for grouped data

5.7 Reference

5. Descriptive statistics

5.1 Introduction

5.1.1 What is Statistics?

The collection, processing, interpretation, and presentation of numerical data all belong to the domain of statistics. The word "statistics" is used in several ways. It can refer not only to the mere tabulation of numeric information, but also to the body of techniques used in processing or analysing data.

5.1.2 Descriptive Statistics

A data set in its original form is usually very large. Consequently, such a data set is not very helpful in drawing conclusions or making decisions. It is easier to draw conclusions from summary tables and diagrams than from the original version of a data set. Therefore, it is usually reduce data to a manageable size by constructing tables, drawing graphs, or calculating summary measures such as averages. Methods that helps us to do this type of statistical analysis is called descriptive statistics.

5.1.3 Population and Sample

A population is a collection of all the elements we are studying and about which we are trying to draw conclusions.

A sample is a collection of some, but not all, of the elements of the population

5.1.4 Parameter and Statistic

A parameter is a characteristic of a population.
A statistic is a characteristic of a sample.

5.2 Table Representation

5.2.1 Types of Data

1. Quantitative (numerical) data
   a) Discrete data
      e.g. number of students in a class; number of variables in a program.
   b) Continuous data
      e.g. height of a person; the time to taken to complete an examination.
2. Qualitative (categorical) data
      e.g. sex of a student, presence or absence in a class.

5.2.2 Raw Data

Data recorded in the sequence in which they are collected and before they are processed or ranked are called raw data.

Example 5.2-1 Suppose that the numbers of hours that students spend in working with computer per week are recorded in the following table.

Table 5.1 Hours spend in working with computer per week

21	19	24	25	22	19
22	19	19	25	22	25
23	19	23	26	22	28
21	25	23	18	27	23
18	19	22	21	19	17

5.2.3 Frequency Distributions

A much more informative presentation of the data in table5.1 is an arrangement called a simple frequency distribution. Table 5.2 is a simple frequency distribution. It is an arrangement that shows the frequency of each hour.

Table 5.2 Frequency distribution of numbers of hours working with computer

Hour (*x*)	**Tally marks**	**Frequency (*f*)**
17	/	1
18	//	2
19	//// //	7
20		0
21	///	3
22	////	5
23	////	4
24	/	1
25	////	4
26	/	1
27	/	1
28	/	1

5.2.4 Relative Frequency and Percentage Distributions

A relative frequency distribution lists the relative frequencies for all categories. The relative frequency of a category is obtained by using the following formula.

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequency}}$$

A percentage distribution lists the percentages for all categories. The percentage for a category is obtained by multiplying the relative frequency of that category by 100. i.e.

$$\text{Percentage} = \text{Relative frequency} \times 100$$

Example 5.2-2 Table 5.3 shows the frequency, relative frequency and percentage distribution of a particular piece of information.

Table 5.3 Frequency, relative frequency and percentage distribution

Department	**Frequency**	**Relative Frequency**	**Percentage**
Business	6	6/36 = 0.17	0.17(100) = 17
Computing	8	0.22	22
Engineering	6	0.17	17
Mathematics	4	0.11	11
Others	12	0.33	33
Total =	36	1.00	100

5.2.5 Grouped Frequency Distributions

When the size of raw data becomes large, it would be appropriate to group the data into classes. Data presented in the form of a frequency distribution are called grouped data.

Example 5.2-3 Table 5.4 shows the number of computer keyboards assembled of a company for a sample of 25 days.

Table 5.4 Number of keyboards assembled

Classes	**Frequency**
41 – 50	5
51 – 60	8
61 – 70	8
71 – 80	4

Lower and Upper Limit

The smallest value in a class is called the lower limit of the class. e.g. 41, 51, 61, 71 and 81. The largest value in a class is called the upper limit of the class. e.g. 50, 60, 70 and 80.

Class Midpoint

The midpoint of a class is obtained by using the following formula.

$$\text{Class midpoint} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

e.g. 45.5, 55.5, 65.5 and 75.5.

Class Boundary

The class boundary is given by the midpoint of the upper limit of one class and the lower limit of the next class. e.g. 50.5, 60.5 and 70.5.

Class Width

The class width is obtained by using the following formula.

$$\text{Class width} = \text{Upper boundary} - \text{Lower boundary}$$

e.g. 60.5 – 50.5 = 10

5.2.6 Constructing Grouped Frequency Distribution Tables

Number of Classes

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set. It is preferable to have more classes as the size of a data set increase. E.g. there are four classes in Table 5.4.

Class Width

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are of the same size, the approximate width of a class is obtained by using the following formula

$$\text{Class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

Usually this approximate class width is rounded to a convenient number. E.g. the class width of the data in Table 5.4 is (50 – 41 + 1) = 10.

Starting Point

Any convenient number which is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example 5.2-4 Construct a grouped frequency distribution table for the following hourly output rate data.

Table 5.5 Hourly output rate

81	76	78	84	76	78
79	80	79	76	82	84
73	78	73	74	72	86
77	80	83	82	83	79
75	80	83	81	77	79

Approximate width of a class $= \frac{86-72}{5} = 2.8$, $\therefore$ class width = 3

Starting value = 72

Table 5.6 Frequency and percentage distributions table

Output Rate	Tally	*f*	Boundary	Relative Frequency	Percentage
72 – 74	////	4	71.5 to < 74.5	0.133	13.3
75 – 77	𝍸 /	6	74.5 to < 77.5	0.200	20.0
78 – 80	𝍸 𝍸	10	77.5 to < 80.5	0.333	33.3
81 – 83	𝍸 //	7	80.5 to < 83.5	0.233	23.3
84 – 86	///	3	83.5 to < 86.5	0.100	10.0
Total :		30		0.999	99.9%

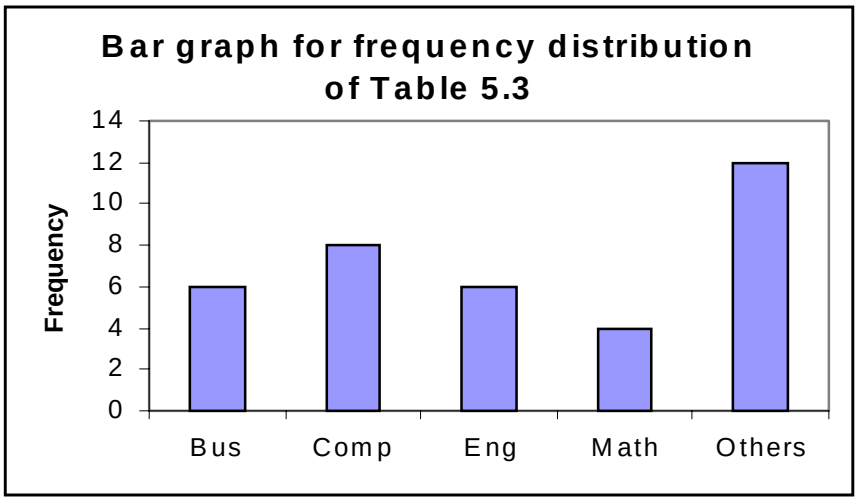
5.3 Graphical Representation

All of us have heard the saying "a picture is worth a thousand words". A graphic display can reveal at a glance the main characteristics of a data set.

5.3.1 Bar Graphs

Bar graphs or bar charts are used to display qualitative data. To construct a bar chart, we mark the various categories on the horizontal axis. All categories are represented by intervals of the same width. We mark the frequencies on the vertical axis, then draw one bar for each category such that the height of the bar represents the frequency of the corresponding category. We leave a small gap between adjacent bars.

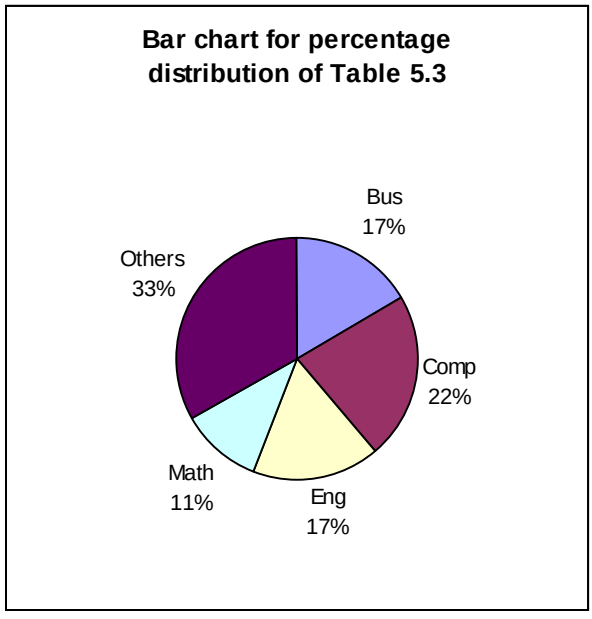
Example 5.3-1 Represent the frequency distribution of Table 5.3 using a bar graph. i.e. Bus. = 6, Comp. = 8, Eng. = 8, Math. = 4, Others = 12



5.3.2 Pie Charts

A pie chart is more commonly used to display data in percentage form. The whole pie or circle represents the total sample or population. The pie is divided into different portions that represent the percentages of the population or sample belonging to different categories.

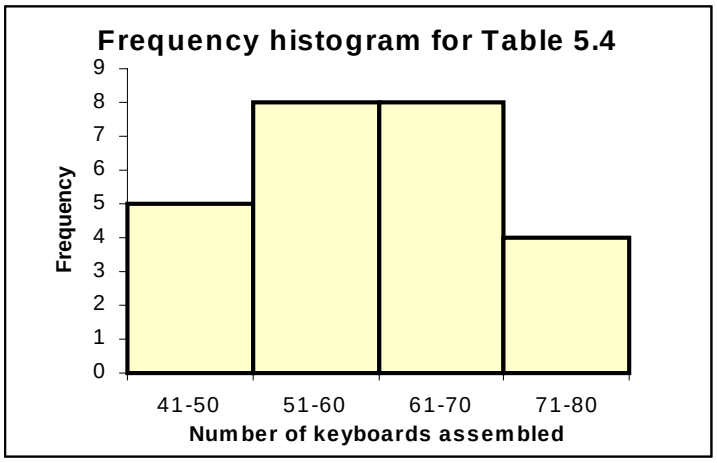
Example 5.3-2 Represent the percentage distribution of Table 5.3 using a pie chart. i.e. Bus.=17%, Comp=22%, Eng.=17%, Math.=11%, Others=33%.



5.3.3 Histograms

A histogram is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies or percentages are marked on the vertical axis. The frequencies, relative frequencies or percentages are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other and without leaving any gap between them.

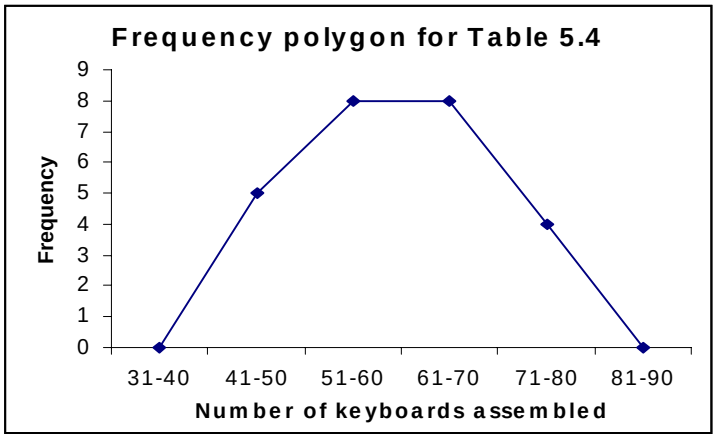
Example 5.3-3 Use a histogram to represent the frequency distribution of Table 5.4. i.e. 41 – 50 = 5, 51 – 60 = 8, 61 – 70 = 8, 71 – 80 = 4.



5.3.4 Polygons

A polygon is a graph that can be used to represent quantitative data. To draw a frequency polygon, we first mark a dot above the midpoint of each class at a height equal to the frequency of that class. Next, we mark two more classes, one at each end of the existing classes, with zero frequencies and mark their midpoints. In the last step, we join the adjacent dots with straight lines.

Example 5.3-4 Use a polygon to represent the frequency distribution of Table 5.4. i.e. 41-50 = 5, 51-60 = 8, 61-70 = 8, 71-80 = 4.



5.4 Measures of Central Tendency

We often represent a data set by numerical summary measures, usually called the typical values. A measure of central tendency gives the centre of a histogram or a frequency distribution curve.

5.4.1 Mean

The mean or average is the most frequently used measure of central tendency. It is obtained by dividing the sum of all values by the number of values in the data set. Thus,

Mean for population data: $\mu = \dfrac{\sum x}{N}$

Mean for sample data: $\bar{x} = \dfrac{\sum x}{n}$

where $\sum x$ is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean.

Example 5.4-1 The following data give the prices of five telephone handsets sold in a shop yesterday.

158 189 265 127 191

Find the mean sale price for these telephone handsets.

$$\bar{x} = \frac{\sum x}{n} = \frac{158+189+265+127+191}{5} = \frac{930}{5} = 186$$

A major shortcoming of the mean as a measure of central tendency is that it is very sensitive to outliers. (outliers or extreme values are very small or very large relative to the majority of the values in a data set.)

5.4.2 Median

Another important measure of central tendency is the median. The median is the value of the middle term in a data set that has been ranked in increasing order. The calculation of the median consists of the following two steps.

1. Rank the given data set in increasing order.
2. Find the middle term. The value of this term is the median.

The position of the middle term in a data set with n values is obtained as follows.

$$\text{Position of the middle term} = \frac{n+1}{2}$$

If the number of observations in a data set is odd, then the median is given by the value of the middle term in the ranked data. If the number of observation is even, then the median is given by the average of the values of the two middle terms.

Example 5.4-2 Find the median for the following data.

37.1 42.2 53.1 53.2 70.0 71.9
74.9 79.6 93.6 109.6 137.1 168.8

Position of the middle term = $\frac{n+1}{2} = \frac{12+1}{2} = 6.5$

Therefore, the median is given by the mean of the 6th and 7th values in the ranked data. i.e.

$$\text{Median} = \frac{71.9 + 74.9}{2} = 73.40$$

The advantage of using the median as a measure of central tendency is that it is not influenced by outliers.

5.4.3 Mode

The mode is the value that occurs with the highest frequency in a data set.

Example 5.4-3 The following data give the speeds (in miles per hour) of eight cars that were stopped on a highway or speeding violations.

77 69 74 81 71 68 74 73

Find the mode.

Mode = 74 miles per hour

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median.

5.5 Measures of Variability

The measure of central tendency, such as mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among values of observations for one data set may be much larger or smaller than for the other data set.

Note that the words dispersion, spread, and variation have the same meaning.

5.5.1 Range

The range is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set. For a set of n values $x_1, x_2, \ldots\ldots, x_n$, the

$$\text{Range} = \max\{x_i\} - \min\{x_i\}.$$

Example 5.5-1 Consider the following two data sets on the ages of all workers for each of the two small company.

Company 1:	47	38	35	40	36	45	39
Company 2:	70	33	18	52	27		

Find the mean and range for these data sets.

Company 1: Mean = 40 Range = 47 – 35 = 12
Company 2: Mean = 40 Range = 70 – 18 = 52

The range, like the mean, has the disadvantage of being influenced by outliers. Thus, it is not a good measure of dispersion to use for data set that contains outliers. Another disadvantage is that its calculation is based on two values only. All other values in a data set are ignored while calculating the range. Thus, it is not very satisfactory measure of dispersion.

5.5.2 Variance and Standard Deviation

The standard deviation is the most frequently used measure of dispersion. The value of the standard deviation tells how closely the values of a data are spread around the mean. In general, a lower value of the standard deviation of a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. On the other hand, a large value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The standard deviation is obtained by taking the positive square root of the variance. The following formulas are used to calculate the variance.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \qquad \text{and} \qquad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

where δ^2 is the population variance and s^2 is the sample variance. The quantity $x-\mu$ or $x-\bar{x}$ in the above formulas is called the deviation of x value from the mean. The sum of the deviations of x values from the mean is always zero.

However, for large number of data, it would be easier and more efficient to calculate the variance and hence the standard deviation by using the following computational formulas.

$$\sigma^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} \qquad \text{and} \qquad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

Example 5.5-2 A building subcontractor pays his eight employees the daily wages (in dollar) is given in the following table. Find the variance and standard deviation for these data.

Employee	1	2	3	4	5	6	7	8
Wage (W) $	1000	600	700	1000	600	1000	1300	800

x	x^2
1000	1000000
600	360000
700	490000
1000	1000000
600	360000
1000	1000000
1300	1690000
800	640000
$\Sigma x = 7000$	$\Sigma x^2 = 6540000$

$$\text{Variance} = s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{6540000 - \frac{(7000)^2}{8}}{8-1} = 59286$$

$$\text{Standard deviation} = s = \sqrt{s^2} = \$243$$

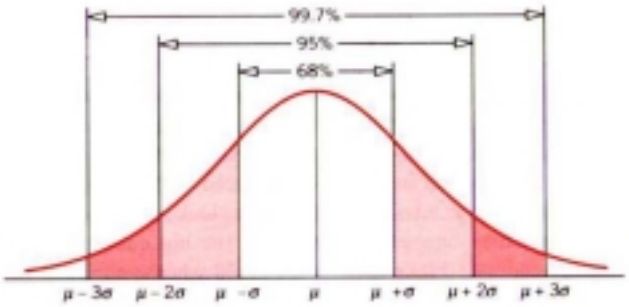
By using the mean and standard deviation, we can find the proportion or percentage of the total observations that fall within a given interval about the mean.

Empirical Rule

For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviation of the mean
3. 99.7% of the observations lie within three standard deviation of the mean

The following figure illustrates the empirical rule



The empirical rule applies to both population and sample data.

5.6 Mean, Variance and Standard Deviation for Grouped Data

5.6.1 Mean for Grouped Data

To calculate the mean for grouped data, first find the midpoint of each class and then multiply the midpoint by the frequencies of the corresponding classes. The sum of these products gives an approximation for the sum of all values. To find the value of mean, divide this sum by the total number of observations in the data. The formulas used to calculate the mean for grouped data are as follows.

Mean for population data: $\mu = \dfrac{\sum mf}{N}$

Mean for sample data: $\bar{x} = \dfrac{\sum mf}{n}$

where m is the midpoint and f is the frequency of a class.

5.6.2 Variance and Standard Deviation for Grouped Data

Following are the basic formulas used to calculate the population and sample variances for grouped data.

$$\sigma^2 = \frac{\sum f\,(m-\mu)^2}{N} \qquad \text{and} \qquad s^2 = \frac{\sum f\,(m-\bar{x})^2}{n-1}$$

where σ^2 is the population variance, s^2 is the sample variance and m is the midpoint of a class. In either cases, the standard deviation is obtained by taking the positive square root of the variance.

Again, the following computational formulas are more efficient for calculating the variance and standard deviation.

$$\sigma^2 = \frac{\sum m^2 f - \dfrac{\left(\sum mf\right)^2}{N}}{N} \quad \text{and} \qquad s^2 = \frac{\sum m^2 f - \dfrac{\left(\sum mf\right)^2}{n}}{n-1}$$

where σ^2 is the population variance, s^2 is the sample variance and m is the midpoint of a class. The standard deviation is obtained by taking the positive square root of the variance.

The population standard deviation: $\sigma = \sqrt{\sigma^2}$

The sample standard deviation: $s = \sqrt{s^2}$

Example 5.6-1 The following gives the frequency distribution of the daily commuting time (in minutes) from home to work for all 25 employees of a company.

Daily commuting time	*Number of employees*
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Calculate the mean, variance and standard deviation of the daily commuting times.

Daily commuting time	f	m	mf	m^2f
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	N = 25		$\sum mf = 535$	$\sum m^2f = 14825$

$$\text{Mean} = \mu = \frac{\sum mf}{N} = \frac{535}{25} = 21.40 \text{ minutes}$$

$$\text{Variance} = \sigma^2 = \frac{\sum m^2 f - \frac{\left(\sum mf\right)^2}{N}}{N} = \frac{14825 - \frac{(525)^2}{25}}{25} = 135.04$$

$$\text{Standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{135.04} = 11.62 \text{ minutes}$$

Thus, the employees of this company spend an average of 21.4 minutes a day commuting from home to work with standard deviation of 11.6 minutes.

5.7 Reference

5.7.1 Modern Elementary Statistics – Freund & Simon, Prentice Hall
5.7.2 Introductory Statistics – Prem S. Mann, Wiley