Unit 3

Data Presentation: Tables and Charts

Overview

Having studied the data collection procedures in Unit 2, we shall now address the question of data presentation. In this unit, we will study a few presentation formats categorised under tables and charts. These formats help us examine and interpret the data without much difficulty. However, we should exercise caution in selecting the formats. We will study in this unit that the choice of a particular presentation format depends, to a considerable extent, on the nature of the data at hand. Further, the kind of analysis intended and the level of statistical sophistication expected determine the choice of a format. Irrespective of the format chosen, it is essential that we follow a few general principles in presenting the data. This, also, we will learn in this unit. You should note that there are now a wide variety of computer packages available for constructing the tables and charts. All that you have to provide is data. However, there may be situations where you will have to construct them on your own manually. This unit should help you present the data manually or by using a computer package or both.

Unit 3 Learning Objectives

After completing this unit, you should be able to:

- list the general principles required to present data in the form of tables and charts, and suggest the factors to be considered in selecting presentation formats;
- use pie charts and bar charts to present data and explain the limitations of their use;
- compare the use of bar charts and histograms in presenting data;
- discuss the problems of creating frequency tables with quantitative data only and explain the associated concepts of class, class limits, class marks and class width;
- list the different types of summary tables;
- discuss the use of frequency charts including the *ogive*;
- create pie charts, bar charts, histograms, frequency charts, stem and leaf displays, boxplots and scattergraphs using MINITAB.

The sessions in this unit are

Session 1: General Principles.
Session 2: Presenting Data: Tables.
Session 3: Presenting Data: Charts.

Note to students

This unit contains several activities and one practice assignment at the end of the unit. You are to work on the activities on your own. If you have any questions or concerns please post a message in the unit discussion forum so that your E-tutor can provide assistance to you. The assignment is to be uploaded in the practice assignment area.

Session 1

General Principles

Introduction

There are certain guidelines that one should follow in the construction of data presentation formats, i.e., tables and charts. Some formats appear so frequently in the literature (Statistics) that we accept them as standard. In this session we will provide guidelines for presenting data as well as discuss the nature of data.

Objectives

On completing this session students should be able to:

- discuss the quantitative and qualitative methods of data collection;
- state general principles for data presentation.

Guidelines for Presenting Data

The choice of presentation formats is dependent on the data available, the analysis required and the statistical sophistication expected. Regardless of the format adopted, however, we should follow certain general principles such as the ones given below.

(i) A table or a chart must be fully labelled. That is, it must contain clear, concise and self-explanatory headings, sub-headings and legends.

(ii) All items should be adequately numbered. That is, each table or chart or figure being used should be appropriately numbered, each type being numbered separately from the others.

(iii) The measurement units used (for example, TT$, EC$, %, metric tons, etc.) must be clearly stated.

(iv) The use of unwieldy numbers must be avoided. For example, numbers such as 2.348562, 24.39785 may be rounded off as, respectively, 2.35 and 24.40, and a large number like 175 321 429.514 may be approximated in millions as 175.3 million. Remember to use the same number of significant figures throughout. If you are using three significant figures, the three numbers above should be written as 2.35, 24.4 and 175 million, respectively.

(v) If no data are available for any part of the exercise, these missing values in the data should be clearly identified.

(vi) Where relevant, and especially when secondary data are used, the source of the data should be clearly stated. This practice is more than just common courtesy to those who had produced the data. It is particularly useful for future reference in cases where another researcher may wish to corroborate or challenge the results, etc.

(vii) The data should be presented as clearly as possible – the presentation should not look cluttered with too many details.

Before we go any further, let us spend some time on the nature of data.

Qualitative and quantitative data

Our choice of a format for presenting the data, i.e., chart or table, depends on whether the data are qualitative or quantitative. We can view qualitative data as categorical or attributive in nature. That is to say, these data are not classified according to numerical size. For instance, if we were to classify our Statistics students (referred to in Unit 2) according to their birth month, this would be categorical data. On the other hand, quantitative data, as we have mentioned earlier, are numerical. To extend the example given above, if the students were classified in terms of age, we would get numerical data. We can subdivide each of qualitative and quantitative into two sub-types.

Qualitative data can be either of the following:

- **ordinal**, where the data are shown simply in order of magnitude or rank (i.e. "greater than" or "less than" but the differences between data objects cannot be measured. For example, the ICC ratings (as released on June 16, 2008) list the top 10 test cricket batsmen in the world as follows:

1. Kumar Sangakara (SL)
2. Shivnarine Chanderpaul (WI)
3. Michael Hussey (AUS)
4. Mohammed Yousef (PAK)
5. Ricky Ponting (AUS)
6. Jacques Kallis (SA)
7. Matthew Hayden (AUS)
8. Mahela Jayawardena (SL)
9. Younis Khan (PAK)

10 . Kevin Pietersen (ENG).

- **nominal**, where the subjects are allocated to distinct named categories and can neither be measured nor ordered. For example, the local Water Authority might list three categories of customers namely, residential, commercial and industrial. In its computer database it might use a "1" to identify the residential customers, a "2" for commercial customers and a "3" for industrial customers. However, the numbers used have no unit of measurement.

Similarly, quantitative data can be either of the following:

- **discrete** (often **integer**), usually obtained by counting, and have jumps or gaps in the observable values. For instance, if we counted the number of students obtaining specific grades in the Statistics class, we would get discrete values. (Discrete values are distinct values on the real line. Suppose, for instance, that employees of a firm are allowed to take any of the following casual leave days at a time: 1/2, 1, 1 1/2, 2. An employee will have to choose one of these as nothing else is available. These are discrete numbers. (Note that these are not integer numbers although integer numbers are also discrete.)

- **continuous**, usually obtained by measurement, and have no jumps or gaps, i.e., in theory, any value is possible within a specific range. An example would be the measured heights of the students in the Statistics class.

Further, since qualitative data always have a limited number of alternative values, we can describe them as discrete as well. For statistical analysis, we can convert qualitative data into discrete numeric data by simply counting the number of different values that appear.

To reinforce what we have discussed as regards these two types of data, qualitative and quantitative, we have given below a few examples.

Qualitative data are obtained when the observations fall into separate distinct categories. Consider the following:

- colour of eyes (blue/green/brown/black);
- examination result (pass/fail);
- socio-economic status (low income group/middle income group/high income group).

The possible categories under which these data can be grouped are finite in number and cannot be obtained by counting or measurement, but arise from the category or group in which the subject is placed.

Quantitative data arise when the observations can be counted or measured. Remember that quantitative data can be discrete (integer) or continuous. If the values are integers (that is whole numbers), these are classified as discrete, and as continuous if the values when measured can take any value within a range. An example of the former is ‘number of cigarettes smoked in a day’ and the latter is ‘weight’.

Consider also the following example, though it is oversimplified. Suppose that we wish to analyse the amount of funds allocated to each faculty at a university. The data on the amount of funds are quantitative in nature while data on the faculty names are qualitative.

So, you can see that both qualitative and quantitative data can be obtained from the same source. Now, be sure that you can distinguish between the data types before you move on to the next section on data presentation formats.

Summary

In this session we listed several guidelines for the presentation of tables and charts. We also discussed the nature of quantitative and qualitative data.

Session 2

Presenting Data

Introduction

In this session, we will introduce two types of tables: summary tables and frequency tables. We will also discuss frequency tables with only quantitative data, and relative frequency tables.

Objectives

After completing this session you should be able to:

- identify the different types of summary tables.

Summary tables

If you take up any statistical publication like the *Quarterly Statistical Digest* of the Central Bank of Trinidad and Tobago or the *Annual Statistical Digest* of the Central Statistical Office, Trinidad and Tobago, it is not uncommon to see tables with lots of data. The purpose of such tables as mentioned is to provide the users with a summary of the data on a certain topic or item. In fact, in Statistics, we refer to tables of this kind as **summary tables**.

A summary table, thus, brings together a mass of connected information for digestion at a glance. Table 3.1, for example, shows the performance of 20 students, distinguished by gender and by degree option followed, in various courses during the 1993/4 academic year. This summary table helps us draw some preliminary conclusions about the relative performances of men and women by degree option. The use of such tables, however, is not always advisable as they can become quickly cluttered and, hence, confusing. (When we have to use them we should take great care in their construction and, in fact, it is generally better to avoid them.)

Table 3.1
Performance of 20 selected students in the faculty of Social Sciences
Academic Year 1993/4

				MARKS OBTAINED BY COURSE (Percent)											
No.	SEX	AGE	OPTION	EC14E	EC16A	EC10D	EC10F	SY13E	SY13F	MS15A	MS15B	GT11C	GT11D	No. of courses Taken	Average per Courses
1	MALE	20	Accounting	65	63	43	45	50	N.A.	82	62	N.A.	37	8	55.9
2	MALE	23	Economiics	45	50	36	40	46	25	78	43	41	N.A.	9	44.9
3	FEMALE	20	Accounting	45	84	54	58	63	N.A.	89	54	N.A.	40	8	60.9
4	MALE	20	Soc. & Man.	31	81	35	N.A.	56	47	N.A.	N.A.	37	60	8	43.4
5	FEMALE	47	Other	41	61	32	N.A.	53	46	N.A.	N.A.	43	N.A.	6	46.0
6	FEMALE	23	Soc. & Man.	48	82	51	N.A.	54	48	N.A.	N.A.	47	N.A.	6	55.0
7	FEMALE	19	Economics	N.A.	73	44	57	58	N.A.	65	56	N.A.	N.A.	6	58.8
8	MALE	28	Economics	71	80	51	58	66	N.A.	68	55	40	66	9	61.7
9	MALE	20	Management	51	81	57	52	70	N.A.	66	60	N.A.	60	8	62.1
10	FEMALE	2	Management	91	76	47	64	60	N.A.	91	92	N.A.	35	8	69.5
11	FEMALE	19	Other	17	N.A.	29	N.A.	N.A.	60	N.A.	N.A.	51	N.A.	4	39.2
12	FEMALE	39	Other	38	35	23	34	45	56	N.A.	N.A.	38	N.A.	7	38.4
13	FEMALE	21	Soc. & Man.	92	90	44	N.A.	70	66	N.A.	N.A.	50	N.A.	6	68.7
14	FEMALE	26	Management	85	79	43	47	57	N.A.	80	71	N.A.	45	8	63.4
15	MALE	24	Economics	70	76	58	55	42	54	72	49	55	N.A.	9	59.0
16	FEMALE	22	Management	N.A.	N.A.	N.A.	N.A.	41	N.A.	57	49	N.A.	53	4	50.0
17	MALE	38	Management	49	61	44	54	18	N.A.	85	44	N.A.	18	8	46.6
18	MALE	23	Economics	88	83	38	64	54	N.A.	89	73	51	51	9	65.7
19	MALE	20	Economics	59	83	55	58	64	48	72	N.A.	40	45	9	58.2
20	MALE	22	Economics	60	51	36	54	60	53	59	41	38	N.A.	9	50.2
No. of Students Enrolled in Course				18	18	19	14	19	10	14	13	12	11		
Average Mark per Course				58.1	71.6	43.2	52.9	54.1	50.3	75.2	57.6	44.3	46.4		
Total				1046	1289	820	740	1027	503	1053	749	531	510		

Source: U.W.I. Registry, St. Augustine

Frequency tables

A second means of summarising data is by a frequency table, an example of which is shown in Table 3.2a. On the left hand side of the table, we have listed the names of the Faculties and on the right hand side the number of students enrolled in each Faculty during the academic year 1994/5. We call tables of this nature frequency tables or frequency distribution tables. In the present case, Table 3.2a, for example, shows the number or frequency of students enrolled in each Faculty.

Table 3.2a
Distribution of Student Enrolment (1994/5)

Faculty (UWI, St. Augustine)	Enrolment
Agriculture	392
Arts & General Studies	787
Education	250
Engineering	1025
Law	39
Medical Sciences	630
Natural Sciences	791
Social Sciences	1301
Total	5215

Source: Office of Planning and Development, UWI, St. Augustine

As you notice here, at the bottom of the table, we have acknowledged the source of the data. Obviously, what is presented in the table is secondary data, i.e., it was not collected for the purpose of writing this unit. That is to say, whenever we use secondary data, it is imperative that we acknowledge the source.

Now, let us revisit Table 3.2a. It shows the relationship between a qualitative category (the different Faculties) and a quantitative category (the number of students enrolled). Generally, at least one of the two categories appearing in any frequency table must be numeric. (We will discuss later the special case where both categories are numeric.)

In our example, where do you think the data, the number of students enrolled, would have originated from? If you have guessed University application or registration forms, you are correct. These forms, as you are aware, contain such categories as name, age, religion, educational qualifications, and so on for the students to supply information. Thus, these forms provide data on various categories.

Now, we may have the data, but how do we present them manually (when a computer package, MINITAB, for example, is not available) so as to arrive at a table like the one at 3.2a? For the present purpose, let us use the same example of student enrolment in Faculties.

The first step is to list the eight faculties vertically. Then go through the pile of application forms and each time the name of a specific faculty is seen, mark a vertical line or a slash against that faculty. After each of the four vertical lines or slashes, mark a horizontal line across them, as indicated in Table 3.2b. (The symbol 𝍸 denotes five entries. In Statistics, entries are presented this way for ease of counting.) Continue doing this until you have exhausted the pile of application forms. At the end, you should get something that looks like Table 3.2b. When these entries are counted and recorded numerically, we get Table 3.2a.

Table 3.2b
Frequency Table: A Generic View

Agriculture	𝍸 𝍸 ... 𝍸 ||	(up to 392)
Arts and General Studies	𝍸 𝍸 ... 𝍸 ||	(up to 787)
Education	𝍸 𝍸 ... 𝍸 𝍸	(up to 250)
Engineering	𝍸 𝍸 ... 𝍸 𝍸	(up to 1025)
Law	𝍸 𝍸 ... 𝍸 ||||	(up to 39)
Medical Sciences	𝍸 𝍸 ... 𝍸 𝍸	(up to 630)
Natural Sciences	𝍸 𝍸 ... 𝍸 |	(up to 791)
Social Sciences	𝍸 𝍸 ... 𝍸 |	(up to 1031)

You should note that the information extracted and displayed in Table 3.2a are among the simplest examples of frequency tables. For example, if we want to get more information about the students, we could use the same application or registration forms (used for Table 3.2a) to extract other useful information such as gender distribution of students among the faculties, distribution by religion, by age, by country of birth, and so on. (And, when we use all these data together, the nature of the table changes. In essence, we will get a summary table.)

ACTIVITY 3.1

Write a brief note on *frequency distributions.*

Note:

a) Write your answer in the space given here.
b) Discuss your answer with your tutor in the online environment.

Now, we shall discuss the following two variants of frequency tables in the given order:

(i) frequency tables with quantitative (numeric) data only;
(ii) relative frequency tables.

Frequency tables with quantitative data only

We said earlier that it is possible to have frequency tables with only quantitative data. When a table contains only quantitative data, we need to take a few special considerations into account. Consider the following table which lists the actual marks obtained in a Mathematics final examination by all female students who took the examination:

Table 3.3
Marks of Female Students in Mathemathics (1994/5)

45	29	59	42	67	52	89	63	77
41	41	84	85	71	52	16	58	65
48	74	64	97	72	78	93	91	
91	28	13	72	77	43	92	35	
17	71	85	85	86	87	82	92	
38	33	79	94	90	89	95	72	
92	94	98	54	74	93	69	40	
85	52	49	58	76	73	62	61	
30	44	44	83	91	95	85	80	
15	79	55	60	97	47	62	50	

Source: DATA.MTV

The categories to be matched here are mark obtained with number of female students obtaining mark. The second category is the frequency variable. But how do we define mark obtained? Do we draw up a two-column table with the first column listing all possible marks representable by the 101 integers ranging from 0 to 100 and, next to each integer, record the number of students scoring this mark? If we do this, we are likely to find that the frequencies are small (0 or 1 in most cases). The table resulting from this exercise will not provide any meaningful analysis.

Although it does not apply in this case, the situation could have been complicated even more, if a student scored fractional marks like 42.5% or even 55.75%. Of course, we could round these off to the nearest whole number but this may not always be reliable especially for scores falling within a small range. Should we then include all possible values (including fractional) between 0 and 100? This will not help matters since we will get a continuum and our data are discrete scores. If we were to modify this somewhat by dividing the range between 0 - 100 into intervals into which we can group the scores, this should solve our problem. This has the effect of grouping data which can then be more easily put into a summary form. We do not need to consider the entire range of possible scores but just the actual range of marks, or something close to it. The actual range here is 13 to 98 and, for convenience, we can use a range that starts at 10 and ends at 100.

Let us summarize what we have said so far about frequency tables with quantitative data only. The first step in constructing a frequency distribution from a table which contains only quantitative data is to divide the range of (actual) marks scored (the interval 10–100) into a series of equal sub-intervals or classes. There are no set rules by which this must be done—the choice is based purely on the characteristics of the data. Suppose we choose nine classes as follows: 10–20, 20–30, ... , 90–100. We would get the frequency column of Table 3.4, i.e., column 2.

Let us pause briefly to familiarise ourselves with a few more terms that we need to know in this context. The terms are

- **class limits:** these are the lower and upper values attached to each class. The first class (10–20) in Table 3.4 has a lower limit of 10 and an upper limit of 20;
- **class marks**: these are the midpoints of each class, obtained from the average of the class limits. Our first two class marks would then be 15 and 25, obtained as follows:
  (10 + 20)/2 = 15 and (20 + 30)/2 = 25;
- **class width**: this is the difference between consecutive class marks (in our case here, it is 25 - 15 which is equal to10).

Now, can you identify these values for each class in Table 3.4?

Table 3.4

Frequency Distribution and Cumulative Frequency Distribution: Marks of Female Students in Mathematics (1994/5)

1	2	3
Class Interval	Frequency	Cumulative Frequency
10–20	4	4
20–30	2	6
30–40	4	10
40–50	12	22
50–60	9	31
60–70	10	41
70–80	14	55
80–90	11	66
90–100	16	82

Source: DATA.MTW

As shown in Table 3.4, the first column defines the class intervals for each of the nine classes. The second column shows the number of scores that fall within each interval, called the frequency, and the last column shows the cumulative frequency. Although we have not mentioned this quantity before, you are probably already familiar with the word cumulative. This means increasing the quantity by successive additions, and if you look at columns 2 and 3 in the table again, you will see that this is exactly what we have done. Columns 1 and 2 in Table 3.4 constitute the frequency distribution while columns 1 and 3 constitute the cumulative frequency distribution.

There is one other peculiarity in the table that we need to bring to your attention. Did you notice that the class intervals defined above overlap: that is, the upper limit of one interval is the same as the lower limit of the next one? For instance, the mark 40 is the upper limit of the third class and the lower limit of the succeeding one. So, where do we put a score of 40? There is, in fact, no hard and fast rule. We just need to ensure that we are consistent. In this example, the score is placed in the (40-50) interval. Whatever is done, you must ensure that the score is only counted once. Had we chosen to include this mark in the lower class (30 - 40), then, while the frequency of this class will now rise to 5, the frequency of the second class (40 – 50) will fall to 11. It is entirely your decision. However, you should maintain uniformity in presentation.

Before we move on to the next section, let us reinforce what we said about cumulative frequency. The third column shows the addition of the frequency counts or what we call the cumulative frequency of the data. These figures tell us the number of students falling into the current interval plus all preceding ones. For instance, the cumulative frequency count of 22 associated with the class 40–50 (obtained from adding 4 + 2 + 4 and 12) tells us that 22 students obtained marks in the range 10–50. The cumulative frequency of the last class, 90–100, is quite naturally the total number of female students (82).

Relative frequency tables

While a frequency distribution shows the number of data points of the sample that falls within each class, we may be sometimes interested, not in this exact figure, but rather in the proportion or percentage of the total that each figure represents. For instance, the first class in our example contains four students out of the total of 82. We may represent this, not as the absolute figure of 4, but as the relative figure of 4/82 (or 4.88). Similarly, the second class contains 2/82 (or 2.44% of the total). Repeating this procedure for all of the defined classes and presenting the results in a table gives us a relative frequency distribution (see Table 3.5 below):

Table 3.5
Relative Frequency Distribution and Cumulative Relative Frequency: Marks of Female Students in Mathematics (1994/5)

Class Interval	Relative Frequency (%)	Cumulative Relative Frequency (%)
10–20	4.88	4.88
20–30	2.44	7.32
30–40	4.88	12.2
40–50	14.6	26.8
50–60	11.0	37.8
60–70	12.2	50.0
70–80	17.1	67.1
80–90	13.4	80.5
90–100	19.5	100

Notice that we have added a third column here as well, showing the relative cumulative frequencies. These are obtained in the same fashion as are the relative frequencies, that is, by finding the proportion of the total that each cumulative frequency represents. Let us say a brief word here about significant figures as related to this table. You will have noticed that all the relative frequencies and relative cumulative frequencies are recorded as three digits or three significant figures which include whole numbers and decimals. As we pointed out before, we must use a consistent number of significant figures in any table. Columns 1 and 2 in Table 3.4 constitute the frequency distribution, while columns 1 and 3 constitute the cumulative frequency distribution.

Summary

So far, we have talked about frequency tables and introduced you to those containing both qualitative and quantitative data or quantitative data only. We said that for tables with only quantitative data, you need to select appropriate data classes and divide the data between the classes. It is recommended that you always select between 6 and 15 data classes in order to summarise the data so that they are most easily understood. We will now move on to discuss graphical ways of summarising data.

Session 3

Presenting Data: Charts

Introduction

In this session, we will discuss data presentation using various types of charts. These will include pie charts, bar charts, histograms, frequency charts, stem and leaf display, boxplots and scattergraph. We will introduce these in the order listed.

Objectives

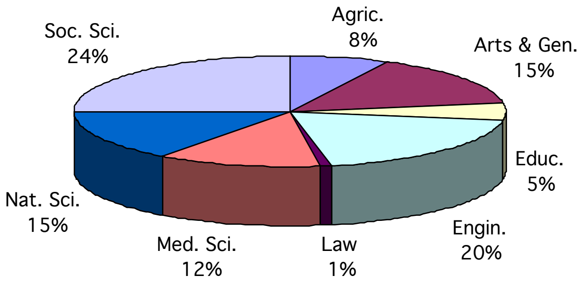
On completing this session you should be able to:

- use pie charts, bar graphs and histograms to present data;
- explain the relevance of the different types of charts to data collection.

Pie charts

If you look at Figure 3.1 below, you will see why it is called a pie chart. It consists of a circle or 'pie' which is divided into sectors or slices. The pie chart here contains the data that we have presented in the frequency table earlier (see Table 3.2a), i.e., student enrollment in Faculties during the 1994/5 academic year.

Figure 3.1
Student Enrollment in Faculty (1994/5)
UWI, St. Augustine Campus



The relative size of the slices represents the proportion of the total enrollment in that Faculty (the qualitative category). Pie charts, therefore, represent the relative size of the component parts of a whole. A complete circle represents the total, and this circle is divided into appropriate segments, the size of which shows the relative proportion of each constituent part of the whole. Thus, in our case, what we have done is to distribute the total student enrollment (5215) across the Faculties. Since, for example, 392 students enrolled in the Faculty of Agriculture, we represent this, approximately, in the pie chart.

Let us see how this is done. Since a circle, by definition, has 360°, we must divide these 360°, proportionally, between the faculties. To find the number of degrees each Faculty represents, we do this simple calculation:

$$\frac{\textit{No. of students enrolled}}{\textit{Total no. of students enrolled}} \; \textit{x 360 degrees}$$

So, the Faculty of Agriculture should be represented by:

392 x 360 / 5215 degrees at the centre

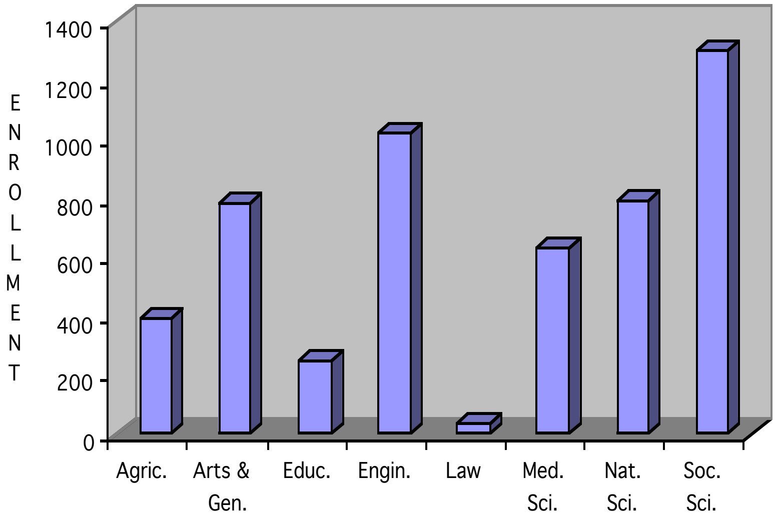
This could be a tedious exercise, especially in cases where there are a large number of categories. Fortunately, we no longer need to plot these charts manually. Computer programmes are now available to convert data into pie charts, and the other charts we will introduce later. Since you are using the MINITAB statistical package with this course, try to produce a pie chart for this data.

Let us now move on to another format for presenting the data, the **bar chart.**

Bar charts

Figure 3.2 below plots the student enrollment data Table 3.2a as a bar chart. The bar chart is one of the most common formats for presenting data. As you can see in the figure, the length of the bar is proportional to the size of the variable (in this case the number of students enrolled in a Faculty).

Figure 3.2
Student Enrollment in Faculties (1994/5)
UWI, St. Augustine Campus



Notice that the bar chart shows the actual enrollment values, whereas the pie chart showed them as percentages. The bar chart has qualitative data (the Faculties) on the horizontal axis with the quantitative data (student enrollments) on the vertical axis. (We will see later that there are some similarities between a bar chart and a histogram, which is another device used to show frequency distributions.)

The bar chart depicted in Figure 3.2 is really a **simple** bar chart. There are other types such as the **multiple** bar chart, the **percentage** bar chart, and the **component** bar chart. However, in this course we are only interested in the simple bar chart.

Now, do Learning Activity 3.2 and then go to the next section on **histograms.**

ACTIVITY 3.2

Describe a **pie chart** and a **bar chart** and explain the relevance of each in data presentation.

Note:
a) Write your answer in the space given here.

____

____

____

____

____

____

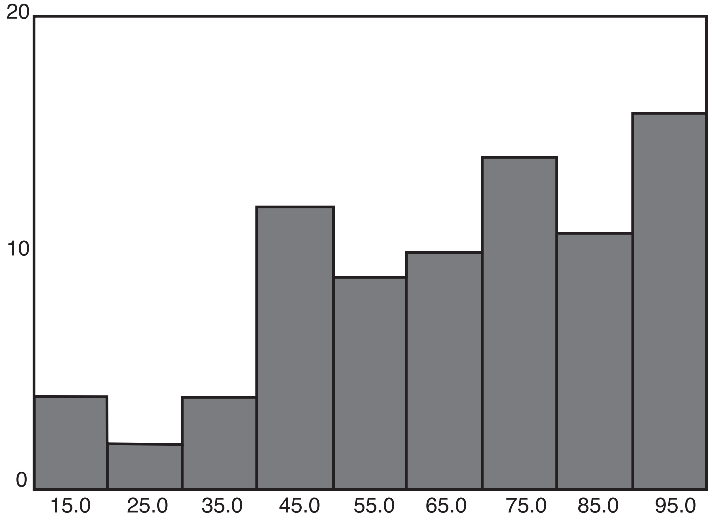
____

____

Histograms

The frequency distributions of Table 3.3 are plotted as a histogram in Figure 3.3. Like the bar chart (Figure 3.2) the histogram is made up of a set of vertical bars whose heights represent the frequencies of the items being considered. Unlike the bar chart, however, the width of the bars in the histogram is significant and represents the class interval associated with each particular frequency represented by the height of the bar. The ends of the bar (on the horizontal axis) represent the lower and upper limits of the class interval and the mid-points (labelled in Figure 3.3) represent the class marks.

Figure 3.3
Histogram: Marks of Female Students in Mathematics (1994/5)



You should note that the histogram (Figure 3.3) was constructed using a statistical package called SPSS for Windows. MINITAB will also plot a histogram but it looks a little different.

When you compare the histograms of the SPSS and MINITAB (two computer packages) you will notice that the class marks and associated frequencies differ slightly. This is because MINITAB uses different class intervals (5–15), (15–25), … , (95–105) and so the number of data points in each interval changes. If you look at the raw data in Table 3.3 again, you will see that only the number 13 falls in the first interval (hence the frequency of 1) and 17, 15, and 16 fall within the second (hence the frequency of 3).

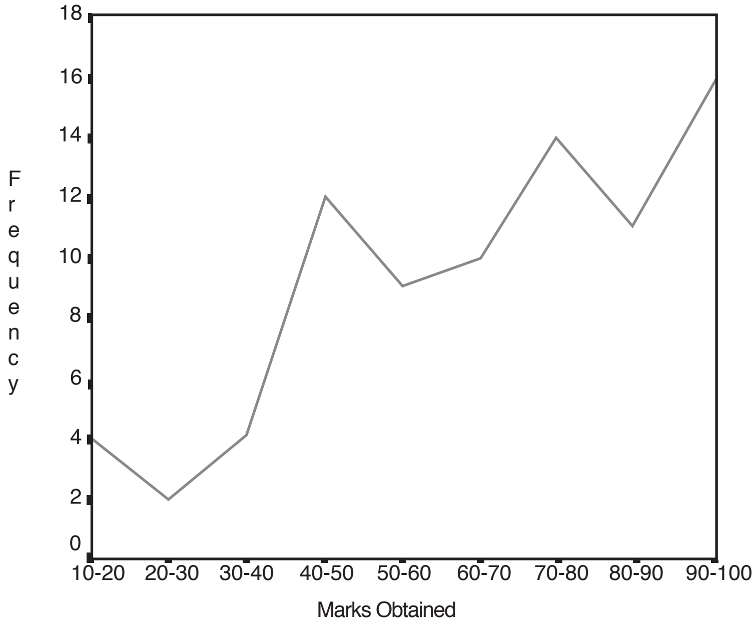
So far, we have seen that we can construct a histogram by plotting class intervals against frequencies. However, if we plotted the class intervals against the relative frequencies (as in Table 3.5), we would obtain a relative frequency histogram. Try this now using the same data in Table 3.3 or if you prefer, use any other set of data.

Let us now move on to yet another data presentation format, **frequency charts**.

Frequency charts

In addition to the bar chart and the histogram, we can also represent frequency distributions by line graphs called frequency charts. Figure 3.4 below shows the frequency chart for the data. Here, we have represented the class intervals on the horizontal axis, against the frequencies on the vertical axis. We have actually plotted the class mark for each class interval against the corresponding frequency to obtain the line graph in the figure.

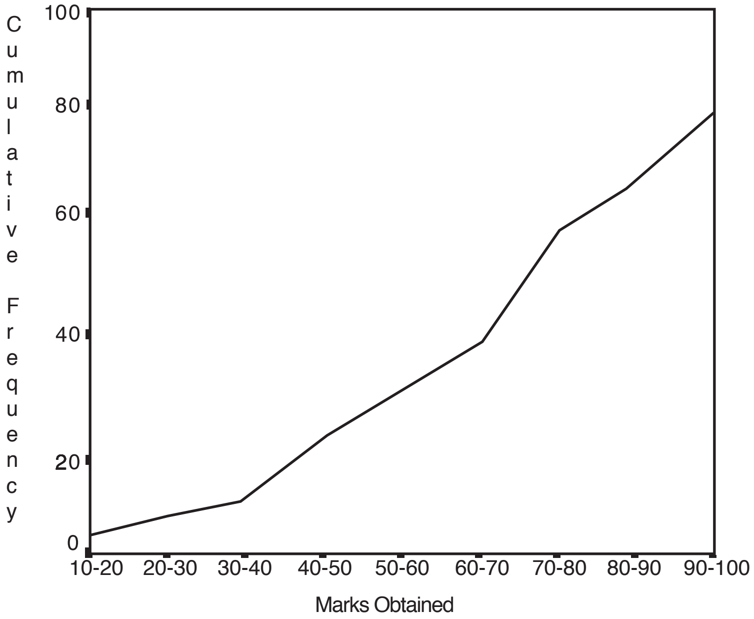
Figure 3.4
Frequency Chart: Marks of Female Students in Mathematics (1994/5)



If you could find some way to fit the line graph of Figure 3.4 over the histogram in Figure 3.3, you would find that each kink on the line graph exactly matches the centre of the top of each bar on the histogram. You can check this yourself by marking the midpoint of the top of each bar on the histogram and joining each of those points together.

Table 3.4 gives us a cumulative frequency as well as a frequency distribution. Instead of plotting frequencies in the vertical axis, if we plot the associated cumulative frequencies, then we obtain a ***cumulative frequency chart***, also known as an **ogive** as shown in Figure 3.5.

Figure 3.5
Cumulative Frequency Chart (Ogive): Marks of Female Students in Mathematics (1994/5)



The next format we will introduce is **the stem and leaf display.**

Stem and leaf displays

We will discuss stem and leaf displays (as well as boxplots and scattergraphs which follow) briefly to suit our immediate purposes.

The stem and leaf display is a particularly useful technique for presenting data in that it is both a pictorial and a tabular display. It represents summary plots that are similar to histograms, but provide more information. Rather than using the same symbol to represent all cases (filled bars in bar charts and histograms and lines in frequency charts), stem and leaf display uses actual data values. Each observed value is divided into two components: a stem (leading digit) and a leaf (trailing digits). The leaves determine the plot symbols. Consider, for instance, the raw data shown in Table 3.3. A possible stem and leaf chart for the data, obtained using MINITAB, will be as shown below in Figure 3.6:

Figure 3.6
Stem and Leaf Display: Marks of Female Students in Mathematics (1994/5)

```
MTB > Stem and Leaf C6;
SUBC>   By sex.

Stem and leaf of ECON1003    SEX = Female    N  = 82
Leaf Unit = 1.0                 N* =  1


                         1    1  3
                         4    1  567
                         4    2
                         6    2  89
                         8    3  03
                        10    3  58
                        17    4  0112344
                        22    4  57889
                        27    5  02224
                        31    5  5889
                        37    6  012234
                        41    6  5579
                        41    7  11222344
                        33    7  677899
                        27    8  023
                        24    8  55556799
                        16    9  01112223344
                         5    9  55778
```

MINITAB output

From the onset you must pay attention to the following:

- the size of the dataset N
- the size of the Leaf Unit
- the heading (as in all tables & charts).

The Stem-and Leaf Plot can be seen to be both

- a pictorial presentation
  - an implicit bar chart
  - an implicit histogram
- a tabular display
  - an implicit frequcncy table;
  - an implicit cumulative frequency table.

It provides additional information beyond the tables & charts mentioned.

It does not only use actual data; in fact, it is the only table/chart that allows the researcher to directly keep in touch with the actual data.

As the name suggests the Plot is characterized by Stems and Leaves.

Each observation from the dataset is divided into two components:
- a Stem (leading digit);
- a Leaf (trailing digit).

e.g. the observation 58 will be represented by a stem 5 and a leaf 8.

The plot comprises three (3) distinct columns, viz:

- the leftmost column (this provides the cumulative frequencies);
- the second column (this contains the stems);
- the third column (this contains the leaves).

Each row has one and only one stem but may have several leaves. By linking the stem with the leaves in the row we can reconstruct all observations from the dataset that fall in that row.
e.g. The observations imbedded in Row 7 are 40, 41, 41, 42, 43, 44 and 44
The sole observation imbedded in Row 1 is 13.

Thus we can begin to see the Plot as a Frequency Table. Each row is essentially a class of that table. Where then are the frequencies?

The frequencies are located in the leftmost column. e.g. in row #7 we encounter the number 17 in the leftmost column. What does that represent? It is the cumulative frequency of rows #1 thru #7. Check the number of leaves from the top row to be sure.

The frequencies are accumulated in two directions
- from the top row down;
- from the last row up.

One of the entries in the leftmost column may have brackets around it. These brackets are to highlight the row/class in which we can locate the observation that is at the centre of the dataset (i.e. when the dataset is listed in ascending order). Such an observation is called the median and the row/class is called the median class. The number in brackets is the frequency of that median class; it is not a cumulative frequency.

Since we know the observations which fall in each class, we know the class frequency. All that is needed to complete a frequency table are the class limits. These are identified by inspection of the classes as follows:

Class #7 starts at 40
Class #8 starts at 45
Class #9 starts at 50
Class #10 starts at 55
Class #11 starts at 60
Class #12 starts at 65

Working forward we can conclude:
Class #13 will start at 70
Class #14 will start at 75
Class #15 starts at 80
Class #16 starts at 85
Class #17 starts at 90
Class #18 starts at 95

Working backward we can conclude:
Class #6 starts at 35
Class #5 starts at 30
Class #4 will start at 25
Class #3 will start at 20
Class #2 starts at 15
Class #1 will start at 10.

We have located all the lower limits
We can now define the upper limits

Class 1	14
Class 2	19
Class 3	24
Class 4	29
Class 5	34
Class 6	39
Class 7	44
Class 8	49

etc.

Using the lower class limits, the upper class limits and the class frequencies we can construct the Frequency Table imbedded in the Plot

Table 3.6

Frequency Distribution and Cumulative Frequency Distribution from the Stem and Leaf Display for ECON1003 Marks – Class of 1994/95

Class Interval	Data Points in the Interval	Class Frequency	Cumulative Frequency
10 - 14	13	1	1
15 - 19	15,16, 17	3	4
20 - 24		0	4
25 - 29	28, 29	2	6
30 - 34	30, 33	2	8
35 - 39	35, 38	2	10
40 - 44	40, 41,41, 42, 43, 44, 44	7	17
45 - 49	45, 47, 48, 48, 49	5	22
50 - 54	50, 52, 52, 52, 54	5	27
55 - 59	55, 58, 58, 59	4	31
60 - 64	60, 61, 62, 62, 63, 64	6	37
65 - 69	65, 65, 67, 69	4	41
70 - 74	71, 71, 72, 72, 72, 73,74, 74	8	49
75 - 79	76, 77, 77, 78, 79, 79	6	55
80 - 84	80, 82, 83	3	58
85 - 89	85, 85, 85, 85, 86, 87, 89, 89	8	66
90 - 94	90, 91, 91, 91, 92, 92, 92, 93, 93, 94, 94	11	77
95 - 99	95, 95, 97, 97, 98	5	82
		Total = 82	

As shown in Table 3.6 above, we can construct the Cumulative Frequency Table imbedded in the Plot from the Frequency Table,

What about the Charts imbedded in the plot?

Simply rotate the Plot through 90o and you would recognise that the leaves constitute a Simple Bar Chart

With little additional effort we can convert that bar chart to a Histogram for the dataset. (by replacing the entries in the left column by the corresponding class marks).

If the Leaf Unit is different from 1, each stem in the plot must be multiplied by Leaf Unit before the leading digit(s) can be determined. E.g. if the Leaf Unit = 10, then a Stem of 5 and a Leaf of 7 will represent the data point 507.

The Plot allows us to directly identify the following information:

- The smallest observation in the dataset
- he largest observation in the dataset
- The most frequently recurring observation in the dataset (otherwise called the mode)
- The observation which lies at the middle of the dataset (otherwise called the median)

The Plot allows us to indirectly identify the following information:

- The first quartile of the dataset;
- The third quartile of the dataset;
- Any other percentile of the dataset.

Recall that if we have the smallest observation, the largest observation, the first quartile, the median and the third quartile we can construct the box plot.

Boxplots

A **boxplot**, also known as a box-and-whisker plot, is a graphical display of data using what is called the **five-number summary**.

The five numbers are the

- minimum value;
- maximum value;
- first quartile;
- second quartile;
- third quartile.

We shall discuss **quartiles** in greater detail later. But for the moment, just remember that the second quartile is the median we met earlier.

In a MINITAB constructed boxplot, you will notice a straight line spanning the axis of the diagram. It runs from the minimum value of 13 at the extreme left point or 'whisker', while the right whisker denotes the maximum value of 98. The box, which appears between the two whiskers, displays the cluster of values obtained by eliminating the worst 25% and the best 25% of the scores. The length of the left and the right whiskers, respectively, represents the worst 25% and best 25% of the scores which have been eliminated. Put another way, the box represents the middle 50% of the scores.

The boxplot highlights the first quartile (25%) i.e., the left whisker or the left hand boundary of the box (about 48). That is, 25% of the students would have got less than this score. In the same way, the boxplot highlights the third quartile, which is the score that closes the box on the right hand side (about 85). In this case, 25% of the students would have a higher score than 85. The boxplot also highlights the second quartile or the median, which is the score marked by a cross inside the box (about 70).
We will now discuss scattergraphs.

Scattergraphs

The scattergraph is a plot of two variables on Cartesian co-ordinates (X and Y). The purpose of the plot is to obtain some preliminary idea about the relationship (if any) between two variables. It is quite reasonable, for example, to expect that a student performing well in Mathematics will also perform well in Statistics. That is, we can expect a high grade in one course to be accompanied by a high grade in the other. The hypothesis is, therefore, that the results of the examinations in these two courses are somehow related.

Suppose that we have a pair of scores for each student doing both courses. Each 'observation' therefore consists of the scores obtained by a student in Mathematics and in Statistics. By plotting the pairs of points, we get a scattergraph.
From the scattergraph, we should be able to form an opinion (albeit a very preliminary one) as to whether these two variables are related or not. If they are, then there should

be some discernible pattern to the 'scatter'. If not, the data points, i.e., the pairs of scores, would appear to be scattered fairly randomly over the 'XY' plane or axis. Of course, as you may realise, more analysis is necessary to determine whether these two variables are indeed related.

Now complete Learning Activity 3.3.

ACTIVITY 3.3

What are the advantages of using charts and graphs in statistical investigations?

Note:
a) Write your answer in the space given here.
b) Discuss your answer with your e-tutor

You may also use the scores provided in Table 3.3 to plot a histogram, a boxplot and a stem and leaf display with MINITAB.

Summary

Data can be presented in several ways. In this session we touched on some of the ways data can be presented graphically. This included bar charts and histograms, pie charts and box plots, to name a few.

Wrap Up

In this unit, we have introduced you to different ways of presenting data. We began with a list of principles that one should follow to effectively summarise data in the form of tables and charts, and pointed out that the choice of format depends on whether the data collected are quantitative or qualitative in nature. After discussing frequency distributions, we explained the problems involved in presenting data in a frequency table, if both of its categories are quantitative. We have, however, explained how to overcome such problems by introducing the concepts of class, class marks, class limits, and class width or class intervals. We then explained the statistical relevance and limitations of pie charts and bar charts, and introduced you to other charts such as the stem and leaf display, the boxplot and the scattergraph.

KEY WORDS INTRODUCED IN THIS UNIT

Attributive data

Bar chart

Box plot (box and whisker plot)

Categorical data

Class

Class unit

Continuous value

Cumulative frequency

Discrete value

Five number summary

Frequency

Frequency chart

Frequency distribution

Stem and leaf display

Frequency table

Histogram

Missing value

Nominal data

Numerical data

Ogive

Ordinal date

Class width

Pie chart

Presentation format

Qualitative data

Quantitative data

Relative frequency

Distribution

Scattergraph

Significant figures

Simple bar chart

Summary table