UNIT 6

DESCRIPTIVE STATISTICS AND BASIC PROBABILITY

Contents:

Introduction

Statistics is the mathematical science that consists of methods of collecting, organizing and analyzing data to drawn meaningful conclusions. In general, statistics is categorized into two broad parts, namely, descriptive statistics and inferential statistics. Investigations part in statistics are discussed in detail in descriptive statistics while analyses fall into two inferential analyses statistics.

Descriptive statistics deals with the processing of data without attempting to draw any inferences from it. The data are presented in the graphical forms such as tables and graphs. The characteristics of the data are described in simple terms. Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.

Inferential statistics is a scientific subject area that uses mathematical technics to make predictions and estimates by analyzing the given data. This subject area is of practice to people employed in such fields as engineering, economics, biology, the social sciences, business, agriculture and communications.

Session 21

Data Representation using Graphical Views

Introduction

Statistics is concerned with the scientific method by which information is collected, organised, analysed and interpreted for the purpose of description and decision making.

There are two statistical methods, namely Descriptive Statistics and Inferential Statistics.

(a) Descriptive Statistics

Descriptive Statistics deals with the presentation of quantitative data or qualitative data, in either tables or graphs form, and with the methodology of analysing the data.

(b) Inferential Statistics

Inferential Statistics involves techniques for making inferences about the whole population based on the observations obtained from samples.

21.1 Basic Definitions

(a) Population

A population is the group from which data are to be collected.

(b) Sample

A sample is a subset of a population.

(c) Variable

A variable is a feature characteristic of any member of a population differing in quality or quantity from one member to another.

(d) Quantitative variable

A variable differing in quantity is called quantitative variable, for example, the weight of a person, number of people in a car.

(e) Qualitative variable

A variable differing in quality is called a qualitative variable or attribute, for example, colour, the degree of damage of a car in an accident.

(f) Discrete variable

A discrete variable is one which no value may be assumed between two given values, for example, number of children in a family.

(g) Continuous variable

A continuous variable is one which any value may be assumed between two given values, for example, the time for 100-meter run.

21.2 Methods of Collecting Data

Statistics very often involves with a collection of data. There are many ways to obtain data such as facet to face interviews, telephone conversations, providing questioners etc. Nowadays, through the World Wide Web is the most important way of collecting data.

The advantages and disadvantages of common data collecting method are discussed below.

21.2.1 Postal Questionnaires

Advantages

- The apparent low cost compared with other methods although the cost per useful answer may well be high.
- No need for a closely grouped sample as in personal interviews, since the Post Office is acting as a field force.
- There is no interviewer bias.
- A considered reply can be given - the respondent has time to consult any necessary documents.

Disadvantages

- The whole questionnaire can be read before answering (which in some circumstances it is undesirable).
- Spontaneous answers cannot be collected. Only simple questions and instructions can be given.
- The wrong person may complete the form.
- Other persons' opinions may be given e.g. by a wife consulting per husband.
- No control is possible over the speed of the reply.
- A poor "response rate" (a low percentage of replies) will be obtained.

The fact that only simple questions can be asked and the possibility of a poor response rate are the most serious disadvantages and are the reasons why other methods will be considered. Only simple questions can be asked because there is nobody available to help the respondent if they do not understand the question. The respondent may supply the wrong answer or not bother to answer at all. If a poor response rate is obtained only those that are interested in the subject may reply and these may not reflect general opinion

21.2.2 Telephone Interviews

Advantages:

- It is cheaper than personal interviews but tends to be dearer on average than postal questionnaires.
- It can be carried out relatively quick.
- Help can be given if the person does not understand the question as worded.
- The telephone can be used in conjunction with other survey methods, e.g. for encouraging replies to postal surveys or making appointments for personal interviews.
- Spontaneous answers can be obtained.

Disadvantages:

- In some countries, not everybody owns a telephone, therefore, a survey carried out among telephone owners would be biased towards the upper social classes of the community. But the telephone can be used in industrial market research anywhere since businesses are invariably on the telephone.
- It is easy to refuse to be interviewed on the telephone simply by replacing the receiver. The response rate tends to be higher than postal surveys but not as high as when personal interviews are used.
- As in the postal questionnaire, it is not possible to check the characteristics of the person who is replying, particularly about age and social class.
- The questionnaire cannot be too long or too involved.

21.2.3 Personal Interviews

In market research this is by far the most commonly used way of collecting information from the general public.

Advantages:

- A trained person may assess the person being interviewed in terms of age and social class and area of residence, and even sometimes assess the accuracy of the information given.
- Help can be given to those respondents who are unable to understand the questions, although great care must be taken that the interview's own feelings do not enter into the wording of the question and so influence the answers of the respondents.
- A well-trained interviewer can persuade a person to give an interview who might otherwise have refused on a postal or telephone enquiry, so that a higher response rate, giving a more representative cross-section of views, is obtained.
- A great deal more information can be collected than is possible by the previous methods. Interviews of three quarters of an hour are commonplace, and a great deal of information can be gathered in this time.

Disadvantages:

- It is far more expensive than either of the other methods because interviewers must be recruited, trained and paid a suitable salary and expenses.
- The interviewer may consciously or unconsciously bias the answers to the question, despite being trained not to do so.
- Persons may not like to give confidential or embarrassing information at a face-to-face interview.
- In general, people may tend to give information that they feel will impress the interviewer, and show themselves in a better light, e.g. by claiming to read "quality" newspapers and journals.
- There is a possibility that the interviewer will cheat by not carrying out the interview or carrying out only parts of it. All reputable organisations carry out quality control checks to lessen the chances of this happening.
- Some types of people are more difficult to locate and interview than others, e.g. travellers. While this may not be important in some

surveys, it will be on others, such as car surveys. One problem is that of the working housewife who is not at home during the day: hence special arrangements must be made to carry out interviews in the evenings and at weekends.

21.2.4 Observation

This may be carried out by trained observers, cameras, or closed circuit television. Observation may be used in widely different fields; for example, the anthropologist who goes to live in a primitive society, or the social worker who becomes a factory worker, to learn the habits and customs of the community they are observing. Observation may also be used in "before and after" studies, e.g. by observing the "traffic" flow in a supermarket before and after making changes in the store layout. In industry many Work Study techniques are based upon observing individuals or groups of workers to establish the system of movements they employ with a view to eliminating wasteful effort. If insufficient trained observers are available, or the movements are complicated, cameras may be used so that a detailed analysis can be carried out by running the film repeatedly. Quality control checks and the branch of market research known as retail audits may also be regarded as observation techniques.

Advantages:

- The actual actions or habits of persons are observed, not what the persons say they would do when questioned. It is interesting to note that in one study only 40% of families who stated they were going to buy a new car had bought one when called upon a year later.
- Observation may keep the system undisturbed. In some cases, it is undesirable for people to know an experiment or change is to be made or is taking place to maintain high accuracy.

Disadvantages:

- The results of the observations depend on the skill and impartiality of the observer.
- It is often difficult in practice to obtain a truly random sample of persons or events.
- It is difficult to predict future behaviour on pure observation.
- It is not possible to observe actions which took place before the study was contemplated.
- Opinions and attitudes cannot usually be obtained by observation.
- In marketing, the frequency of a person's purchase cannot be obtained by pure observation. Nor can such forms of behaviour as church-going, smoking and crossing roads, except by employing a continuous and lengthy (and hence detectable) period of observation.

21.2.5 Reports and Published Statistics

Information published by international organisations such as the United Nations Organisation gives useful data. Most governments publish statistics of population, trade, production etc. Reports on specialised topics including scientific research are published by governments, trade organisations, trade unions, universities, professional and scientific organisations and local authorities. The World Wide Web is also an efficient source of obtaining data.

21.3 Primary and Secondary Data

Before considering whether to instigate a data collection exercise at all it is wise to ascertain whether data which could serve the purpose of the current enquiry is already available, either within the organisation or in a readily accessible form elsewhere.

When data is used for the purpose for which it was originally collected it is known as primary data; when it is used for any other purpose subsequently, it is termed secondary data. For example, if a company Buyer obtains quotations for the price, delivery date and performance of a new piece of equipment from several suppliers with a view to purchase, then the data as used by the Buyer is primary data. Should this data later be used by the Budgetary Control department to estimate price increases of machinery over the past year, then the data is secondary.

Secondary data may be faced with the following difficulties:

- The coverage of the original enquiry may not have been the same as that required, e.g. a survey of house building may have excluded council built dwellings.
- The information may be out of date, or may relate to different period of the year to that required. Intervening changes in price, taxation, advertising or season can and do change people's opinions and buying habits.
- The exact definitions used may not be known, or may simply be different from those desired, e.g. a company which wishes to estimate its share of the "fertilizer" market will find that the government statistics included lime under "fertilizers".
- The sample size may have been too small for reliable results, or the method of selecting the sample a poor one.
- The wording of the questions may have been poor, possibly biasing the results.
- No control is possible over the quality of the collecting procedure, e.g. by seeing that measurements were accurate, questions were properly asked and calculations accurate.

However, the advantage of secondary data, when available and appropriate, is that a great deal of time and money may be saved by not having to collect the data oneself. Indeed, in many cases, for example with import-export statistics,

it may be impossible for a private individual or company to collect the data which can only be obtained by the government.

21.4 Graphical Descriptions of Data

21.4.1 Graphical Presentation

A graph is a method of presenting statistical data in visual form. The main purpose of any chart is to give a quick, easy-to-read-and-interpret pictorial representation of data which is more difficult to obtain from a table or a complete listing of the data. The type of chart or graphical presentation used and the format of its construction is incidental to its main purpose. A well-designed graphical presentation can effectively communicate the data's message in a language readily understood by almost everyone. You will see that graphical methods for describing data are intuitively appealing descriptive techniques and that they can be used to describe either a sample or a population; quantitative or qualitative data sets.

Some basic rules for the construction of a statistical chart are listed below:

(a) Every graph must have a clear and concise title which gives enough identification of the graph.

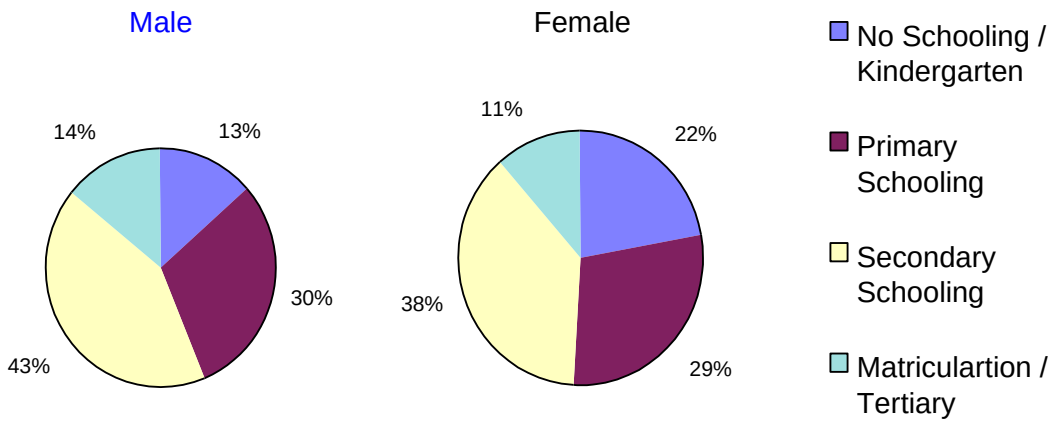
(b) Each scale must have a scale caption indicating the units used.

(c) The zero point should be indicated on the co-ordinate scale. If, however, lack of space makes it inconvenient to use the zero-point line, a scale break may be inserted to indicate its omission.

(d) Each item presented in the graph must be clearly labelled and legible even in black and white reprint. There are many varieties of graphs. The most commonly used graphs are described as below.

(a) Pie chart - Pie charts are widely used to show the component parts of a total. They are popular because of their simplicity. In constructing a pie chart, the angles of a slice from the centre must

be in proportion with the percentage of the total. The following example of pie charts gives the percentage of education attainment in Hong Kong.



(b) Simple bar chart - The horizontal bar chart is also a simple and popular chart. Like the pie chart, the simple horizontal bar chart is a one-scale chart. In constructing a bar chart, it is noted that the width of the bar is not important, but the height of the bar must be in proportion with the data. The following bar chart gives the monthly household income of Hong Kong.

(c) Two-directional bar chart - A bar chart can use either horizontal or vertical bars. A two-

directional bar chart indicates both the positive and negative values. The following

example gives the top 5 cities which have the highest/ lowest recorded temperature.

(d) Multiple bar chart - A multiple bar chart is particularly useful if one desires to make quick comparison between different sets of data. In the following example, the marital status of male and female in Hong Kong are compared using multiple bar char.

(e) Component bar chart - A component bar chart subdivides the bars in different sections. It is useful when the total of the components is of interest. The following example gives the nutritive values of food.

(f) Other type of graphs - Graphic presentations can be made more attractive using careful layout and appropriate symbols. Sometimes information pertaining to different geographical area can even be presented using so-called statistical map.

A pictograph illustrates statistical data by means of a pictorial symbol. It can add greatly to the interest of what might otherwise be a dull subject. The chosen symbol must have a close association with the subject matter, so that the reader can comprehend the subject under discussion at a glance.

21.5 Frequency Distribution

Statistical data obtained by means of census, sample surveys or experiments usually consist of raw, unorganized sets of numerical values. Before these data can be used as a basis for inferences about the phenomenon under investigation or as a basis for decision, they must be summarized, and the pertinent information must be extracted.

Example 1

A traffic inspector has counted the number of vehicles passing a certain point in 100 successive 20-minute time periods. The observations are listed below.

23	20	16	18	30	22	26	15	5	18
14	17	11	37	21	6	10	20	22	25
19	19	19	20	12	23	24	17	18	16
27	16	28	26	15	29	19	35	20	17
12	30	21	22	20	15	18	16	23	24
15	24	28	19	24	22	17	19	8	18
17	18	23	21	25	19	20	22	21	21
16	20	19	11	23	17	23	13	17	26
26	14	15	16	27	18	21	24	33	20
21	27	18	22	17	20	14	21	22	19

A useful method for summarizing a set of data is the construction of a frequency table, or a frequency distribution. That is, we divide the overall range of values into several classes and count the number of observations that fall into each of these classes or intervals.

The general rules for constructing a frequency distribution are

i) There should not be too few or too many classes.

ii) Insofar as possible, equal class intervals are preferred. But the first and last classes can be open-ended to cater for extreme values.

iii) Each class should have a class mark to represent the classes. It is also named as the class midpoint of the ith class. It can be found by taking simple average of the class boundaries or the class limits of the same class.

1. Setting up the classes

Choose a class width of 5 for each class, then we have seven classes going from 5 to 9, from 10 to 14, …, and from 35 to 39.

2. Tallying and counting

Classes	Count
5 – 9	3
10 – 14	9
15 – 19	36
20 – 24	35
25 – 29	12
30 – 34	3
35 – 39	2

3. Illustrating the data in tabular form

Frequency Distribution for the Traffic Data

Number of vehicles period	Number of periods
5 – 9	3
10 – 14	9
15 – 19	36
20 – 24	35
25 – 29	12
30 – 34	3
35 – 39	2
Total	100

In this example, the class marks of the traffic-count distribution are 7, 12, 17, …, 32 and 37.

21.5.1 Histogram

A histogram is usually used to present frequency distributions graphically. This is constructed by drawing rectangles over each class. The height of each rectangle should be proportional to its frequency.

Notes :

1. The markings on the horizontal scale of a histogram can be class limits, class boundaries, class marks or arbitrary key values.
2. The range of the random variable should constitute the major portion of the graphs of frequency distributions. If the smallest observation is far away from zero, then a 'break' sign ( ) should be introduced in the horizontal axis.

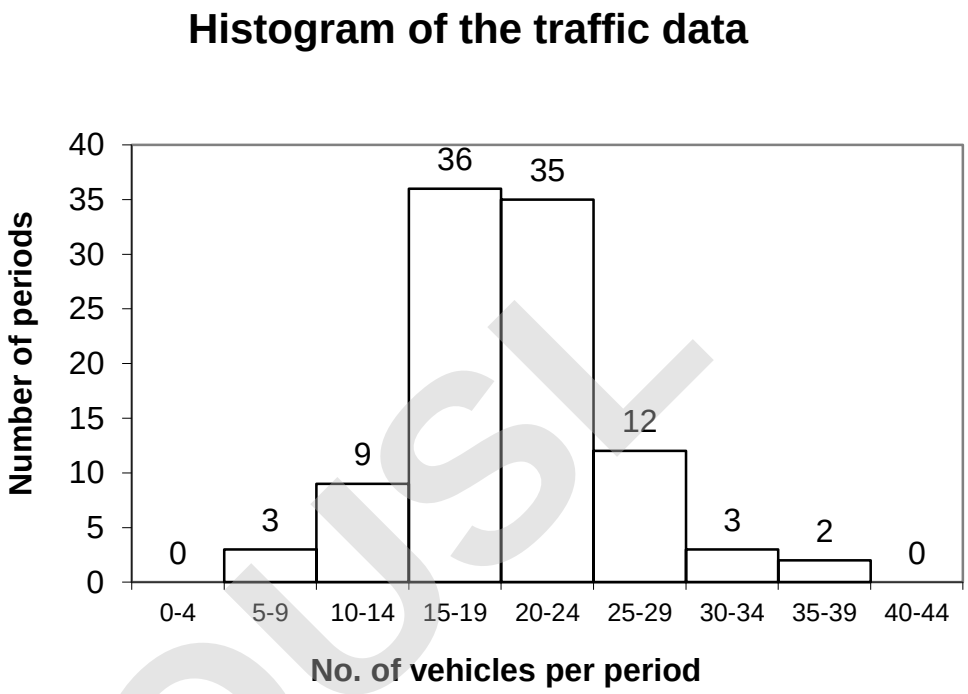
21.5.2 Frequency Polygon

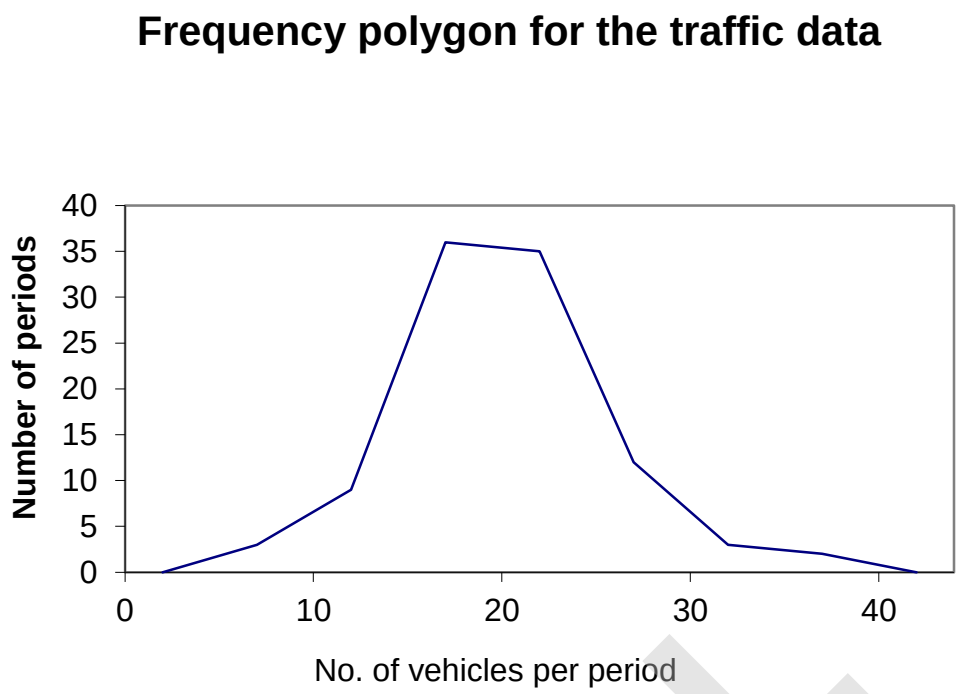
Another method to represent frequency distribution graphically is by a frequency polygon. As in the histogram, the base line is divided into sections corresponding to the class-interval, but instead of the rectangles,

the points of successive class marks are being connected. The frequency polygon is particularly useful when two or more distributions are to be presented for comparison on the same graph.

Example 2

Construct a histogram and a frequency polygon for the traffic data in Example 1.





21.5.3 Frequency Curve

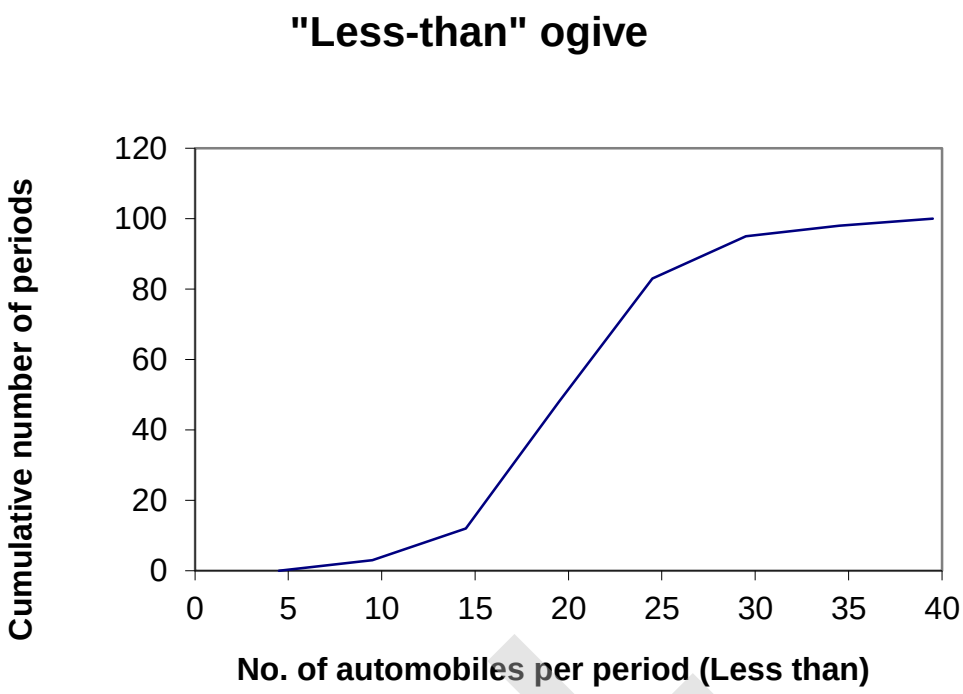
A frequency curve can be obtained by smoothing the frequency polygon.

21.5.4 Cumulative Frequency Distribution and Cumulative Polygon

Sometimes it is preferable to present data in a cumulative frequency distribution, which shows directly how many of the items are less than, or greater then, various values.

Example 3

Construct a "Less-than" ogive of the distribution of traffic data.



21.5.5 Cumulative Frequency Curve

A cumulative frequency curve can similarly be drawn.

21.5.6 Relative Frequency

Relative frequency of a class is defined as:

$$\frac{\text{Frequency of the Class}}{\text{Total Frequency}}$$

If the frequencies are changed to relative frequencies, then a relative frequency histogram, a relative frequency polygon and a relative frequency curve can similarly be constructed.

Relative frequency curve can be considered as probability curve if the total area under the curve be set to 1. Hence the area under the relative frequency curve between a and b is the probability between interval a and b.

Example 4

Construct a relative frequency distribution and a percentage distribution from the traffic data in Example 1.

Summary

In this session, we have presented the basic definitions regarding the collection of data. We have discussed the methods of collecting data. In each method, the advantages and disadvantages have been discussed in detail. we have introduced Pie charts, frequency distribution and bar charts etc. to represent data.

Learning Outcomes:

After studying this session, you should be able to:

- get familiar with some of the statistical terminology;
- represent data graphically;
- describe basic frequency distribution;

Session 22
Measures of Central Tendency & Dispersion

Contents

Introduction

When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency. For a given set of data, the measure of location we use depends on what we mean by middle; different definitions give rise to different measures. We shall consider some more commonly used measures, namely arithmetic mean, median and mode. The formulas in finding these values depends on whether they are ungrouped data or grouped data.

22.1 Central Tendency

A measure of central tendency is a single value that represents the way in which a group of data around a central value. In other words, it is a way to

describe the centre of a data set. Basically, there are three measures of central tendency, namely, arithmetic mean, media and mode.

22.1.2 Arithmetic Mean

The arithmetic mean, μ, or simply called mean, is obtained by adding together all the measurements and dividing by the total number of measurements taken. Mathematically it is given as

$$\mu = \frac{\sum f_i \cdot x_i}{\sum f_i}$$

Where - for grouped data: f_i - is the frequency in the ith class,

x_i - is the class mark in the ith class;

for ungrouped data: f_i - is the frequency in the ith datum,

x_i - is the value in the ith datum.

Arithmetic mean can be used to calculate any numerical data and it is always unique. It is obvious that extreme values affect the mean. Also, arithmetic mean ignores the degree of importance in different categories of data.

Example 1

Given the following set of ungrouped data:

$$20,\ 18,\ 15,\ 15,\ 14,\ 12,\ 11,\ 9,\ 7,\ 6,\ 4, 1$$

Find the mean of the ungrouped data.

Solution:

$$mean = \frac{20+18+2\cdot 15+14+12+11+9+7+6+4+1}{12}$$

$$= \frac{132}{12}$$

$$= 11$$

22.1.2 Weighted Arithmetic Mean

In order to consider the importance of some data, different weighting factors, w_i , can be assigned to individual datum. Hence the weighted arithmetic mean, μ, is given as:

$$\mu = \frac{\sum f_i \cdot w_i \cdot x_i}{\sum f_i \cdot w_i}$$

Where w_i is the weight for the i^{th} datum.

f_i and x_i are defined same as those in the arithmetic mean for ungrouped and grouped data.

22.1.3 Median

Median is defined as the middle item of all given observations arranged in order. For ungrouped data, the median is obvious. In case of the number of measurements is even, the median is obtained by taking the average of the middle.

Example 2

The median of the ungrouped data: $20, 18, 15, 15, 14, 12, 11, 9, 7, 6, 4, 1$ is

Write these numbers in ascending order,

$$1, 4, 6, 7, 9\ , 11, 12, 14, 15, 15, 18, 20$$

$$\text{Median} = \frac{6+7}{2} \text{ th observation}$$

$$= \frac{11+12}{2}$$

$$= 11.5$$

For grouped data, the median can be found by first identify the class containing the median, then apply the following formula:

$$median = l_1 + \frac{\frac{n}{2} - C}{f_m}(l_2 - l_1)$$

where: l_1 is the lower-class boundary of the median class;

n is the total frequency;

C is the cumulative frequency just before the median class;

f_m is the frequency of the median;

l_2 is the upper class boundary containing the median.

It is obvious that the median is affected by the total number of data but is independent of extreme values. However, if the data is ungrouped and numerous, finding the median is tedious. Note that median may be applied in qualitative data if they can be ranked.

22.1.4 Mode

Mode is the value which occurs most frequency. The mode may not exist, and even if it does, it may not be unique.

For ungrouped data, we simply count the largest frequency of the given value. If all are of the same frequency, no mode exits. If more than one values have the same largest frequency, then the mode is not unique.

Example 5

The value for the mode of the data in Example 5 is 15 (unimodal)

Example 6

{2, 2, 2, 4, 5, 6, 7, 7, 7}

Mode = 2 or 7 (Bimodal)

For grouped data, the mode can be found by first identify the largest frequency of that class, called modal class, then apply the following formula on the modal class:

$$\text{mode} = l_1 + \frac{f_a}{f_a + f_b}(l_2 - l_1)$$

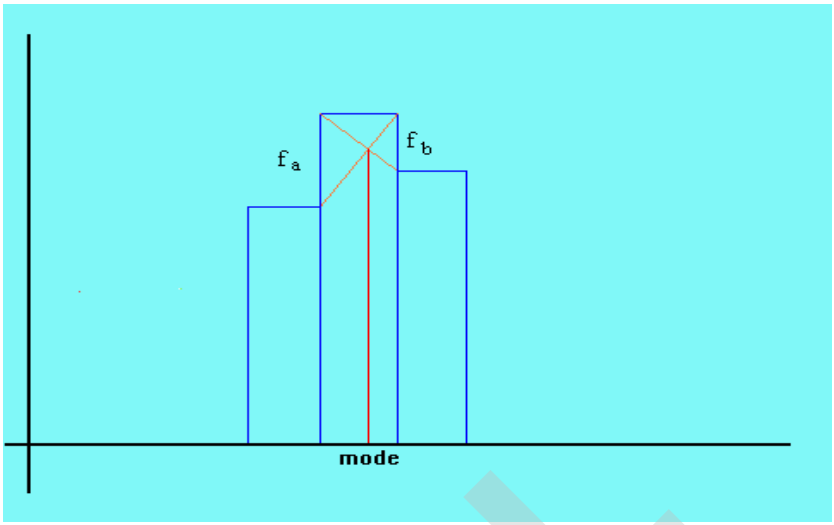
where: l_1 is the lower-class boundary of the modal class;

f_a is the difference of the frequencies of the modal class with the previous class and is always positive;

f_b is the difference of the frequencies of the modal class with the following class and is always positive;

l_2 is the upper class boundary of the modal class.

Geometrically the mode can be represented by the following graph and can be obtained by using similar triangle properties. The formula can be derived by interpolation using second degree polynomial.



Note that the mode is independent of extreme values and it may be applied in qualitative data.

22.1.5 Conclusion

For symmetrically distributed data, the mean, median and mode can be used almost interchangeably.

For moderately skewed distribution data, their relationship can be given by

$$\text{Mean} - \text{Mode} \approx 3 \cdot (\text{Mean} - \text{Median})$$

Physically, mean can be interpreted as the centre of gravity of the distribution. Median divides the area of the distribution into two equal parts and mode is the highest point of the distribution.

22.2 Dispersion and Skewness

Sometimes mean, median and mode may not be able to reflect the true picture of some data. The following example explains the reason.

Example 7
There were two companies, Company A and Company B. Their salaries profiles given in mean, median and mode were as follow:

	Company A	Company B
Mean	Rs 30,000	Rs 30,000
Median	Rs 30,000	Rs 30,000
Mode	(Nil)	(Nil)

However, their detail salary structures could be completely different as that:

Company A Rs 5,000 Rs 15,000 Rs 25,000 Rs 35,000 Rs 45,000 Rs 55,000

Company B Rs 5,000 Rs 5,000 Rs 5,000 Rs 55,000 Rs55,000 Rs 55,000

Hence it is necessary to have some measures on how data are scattered. That is, we want to know what is the dispersion, or variability in a set of data.

22.2.1 Range

Range is the difference between two extreme values. The range is easy to calculate but cannot be obtained if open ended grouped data are given.

22.2.2 Deciles, Percentile, and Fractile

Decile divides the distribution into ten equal parts while percentile divides the distribution into one hundred equal parts. There are nine deciles such that 10% of the data are $\leq D_1$; 20% of the data are $\leq D_2$; and so on. There are 99 percentiles such that 1% of the data are $\leq P_1$;

2% of the data are ≤ P_2; and so on. Fractile, even more flexible, divides the distribution into a convenience number of parts.

22.2.3 Quartiles

Quartiles are the most commonly used values of position which divides distribution into four equal parts such that 25% of the data are ≤ Q_1; 50% of the data are ≤ Q_2; 75% of the data are ≤ Q_3.

The first quarter is conventionally denoted as Q_1, while the second and third quarters grouped together is Q_2 and the last quarter is Q_3.

Note that Q_2 includes the median, contains half of the frequency and excludes extreme values. It is also denoted the value $(Q_3 - Q_1) / 2$ as the Quartile Deviation, Q_D, or the semi-interquartile range.

22.2.4 Mean Absolute Deviation

Mean absolute deviation is the mean of the absolute values of all deviations from the mean. Therefore, it takes every item into account. Mathematically it is given as:

$$\frac{\sum f_i |x_i - \mu|}{\sum f_i}$$

where: f_i is the frequency of the ith item;

x_i is the value of the ith item or class mark;

μ is the arithmetic mean.

23.2.5 Variance and Standard Deviation

The variance and standard deviation are two very popular measures of variation. Their formulations are categorized into whether to evaluate from a population or from a sample.

The population variance, σ^2, is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$\sigma^2 = \frac{\sum f_i(x_i - \mu)^2}{\sum f_i}$$

where: f_i is the frequency of the ith item;

x_i is the value of the ith item or class mark;

μ is the population arithmetic mean.

The population standard deviation σ is defined as $\sigma = \sqrt{\sigma^2}$.

The sample variance, denoted as s^2 gives:

$$\frac{\sum f_i(x_i - \bar{x})^2}{(\sum f_i) - 1}$$

where: f_i is the frequency of the ith item;

x_i is the value of the i th item or class mark;

$\bar{x}$ is the sample arithmetic mean.

The sample standard deviation, s, is defined as $s = \sqrt{s^2}$.

Note that when calculating the sample variance, we have to subtract 1 from the total frequency which appears in the denominator. Although when the total frequency is large, $s \approx \sigma$, the subtraction of 1 is very important.

Example 8

Measures of Grouped Data (Refers to the followings Data Set)

Gas Consumption	Frequency
10 – 19	1
20 – 29	0
30 – 39	1
40 – 49	4

50 – 59	7
60 – 69	16
70 – 79	19
80 – 89	20
90 – 99	17
100 – 109	11
110 – 119	3
120 – 129	1
	100

1. $\bar{x} = \dfrac{\sum x_i f_i}{n}, n = \sum f_i$

$$= \frac{1\times 14.5 + 0\times 24.5 + \cdots + 1\times 124.5}{100}$$
$$= 79.7$$

2.

$$median = 79.5 + \frac{50-48}{20}\times 10$$
$$= 80.5$$
$$Q_1 = 59.5 + \frac{25-13}{16}\times 10$$
$$\approx 67$$
$$Q_3 = 89.5 + \frac{75-68}{17}\times 10$$
$$\approx 93.6$$

3.

$$mode = 79.5 + \frac{20-19}{(20-19)+(20-17)}\times 10$$
$$= 82$$

4.

$$\textit{sample}\ \text{s.d.},\ \text{s} = \sqrt{\frac{\text{n}(\sum \text{x}^2 f) - (\sum xf)^2}{n(n-1)}}$$

$$= \sqrt{\frac{100(671705) - (7970)^2}{100(100-1)}}$$

$$= 19.2$$

22.2.6 Coefficient of Variation

The coefficient of variation is a measure of relative importance. It does not depend on unit and can be used to make comparison even two samples differ in means or relate to different types of measurements.

The coefficient of variation gives:

$$\frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

Example 9

	$\bar{x}$	S
Salesman salary	Rs 916.76/month	Rs 286.70
Clerical salary	Rs 98.50/week	Rs 20.55

$$Vs = \frac{286.70}{916.76} \times 100\% = 31\%$$

$$Vc = \frac{20.55}{98.50} \times 100\% = 21\%$$

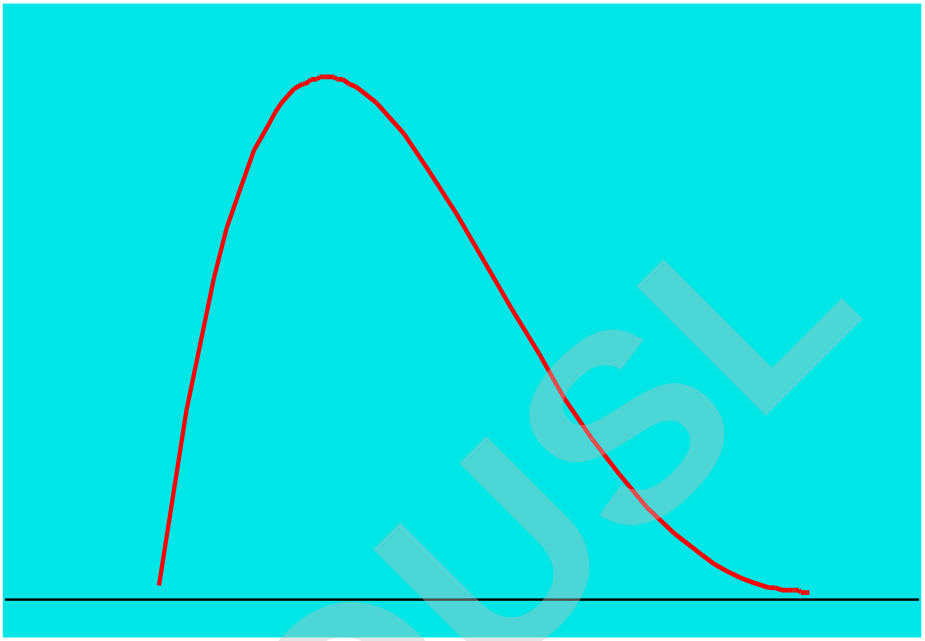
23.2.7 Skewness

The skewness is an abstract quantity which shows how data piled-up. A number of measures have been suggested to determine the skewness of a given distribution. One of the simplest one is known as Pearson's measure of skewness:

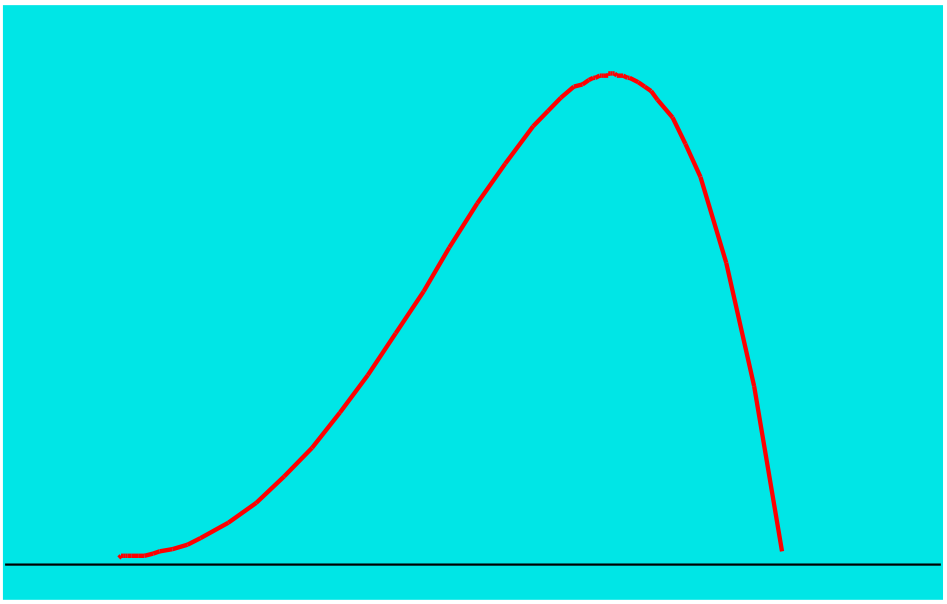
$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\approx \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

If the tail is on the right, we say that it is skewed to the right, and the coefficient of skewness is positive.



If the tail is on the left, we say that is skewed to the left and the coefficient of skewness is negative.



Example 10

We are going to use Example 9 to evaluate the different measurements of variation.

As stated above, the salary scales of the two companies are:

Company A: Rs 5,000 Rs 15,000 Rs 25,000 Rs 35,000 Rs 45,000 Rs 55,000

Company B: Rs 5,000 Rs 5,000 Rs 5,000 Rs 55,000 Rs 55,000 Rs 55,000

Range

Company A: Rs 55,000 - Rs5,000

= Rs 50,000

Company B: Rs 55,000 - Rs 5,000

= Rs 50,000

Mean absolute deviation

Company A: Rs (|5,000 - 30,000| + |15,000 - 30,000| + |25,000 - 30,000| + |35,000 - 30,000| + |45,000 - 30,000| + |55,000 - 30,000|) / 6

= Rs15,000

Company B: Rs (|5,000 - 30,000| + |5,000 - 30,000| + |5,000 - 30,000| + |55,000 - 30,000| + |55,000 - 30,000| + |55,000 - 30,000|) / 6

= Rs 25,000

Variance

Company A: Rs^2 { $(5,000 - 30,000)^2$ + $(15,000 - 30,000)^2$ + $(25,000 - 30,000)^2$ + $(35,000 - 30,000)^2$ + $(45,000 - 30,000)^2$ + $(55,000 - 30,000)^2$ } / 6

= Rs^2 291,666,667

Company B: Rs^2 { $(5,000 - 30,000)^2$ + $(5,000 - 30,000)^2$ + $(5,000 - 30,000)^2$ + $(55,000 - 30,000)^2$ + $(55,000 - 30,000)^2$ + $(55,000 - 30,000)^2$ } / 6

= $Rs^2$625,000,000

Standard deviation

Company A: Rs√ 291,666,667

= Rs 17,078

Company B: Rs √ 625,000,000

= Rs 25,000

Coefficient of variation

Company A: Rs 17,078 / \$30,000 × 100%

= 56.93%

Company B: Rs 25,000 / Rs 30,000 × 100%

= 83.33%

Coefficient of Skewness

Pearson's 1st coefficient of skewness,

$$SK_1 = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Pearson's 2nd coefficient of skewness

$$SK_2 = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Chebyshev's Theorem

For any set of data, the proportion of data that lies between the mean plus and minus k standard deviations is at least $1 - \frac{1}{k^2}$

i.e. $\Pr(\mu - k\sigma \le x \le \mu + k\sigma) \ge 1 - \frac{1}{k^2}$

Summary

In this session, we have introduced the concepts of central tendency of a population. In order to do this, we have discussed the concepts of arithmetic mean, weighted mean, mode and median. We have described the formulae to calculate arithmetic mean, weighted mean, mode and median of a population. We have presented the method how to represent a point in three dimensions using coordinates. To describe the variation of a population, we have given the concepts of Dispersion and skewness. We have defined range, mean deviation, quartiles, variance and standard Deviation to measure the dispersion of a population. We have given formulae to measure the skewness of a population.

Learning outcomes

At the end of this session the student should be able to,

1. Describe the concepts of central tendency, dispersion and skewness of a population.

2. Calculate the arithmetic mean, weighted mean, mode and median to measure the central tendency of a population

3. Calculate range, mean deviation, quartiles, variance and standard deviation to measure the dispersion of a population

4. Describe and measure the skewness of a population.

Session 23

Introduction to Probability

Introduction

The probability of a specified event is the chance or likelihood that it will occur. There are several ways of viewing probability. One would be **experimental** in nature, where we repeatedly conduct an experiment. Suppose we flipped a coin over and repeatedly and it came up heads about half of the time; we would expect that in the future whenever we flipped the coin it would turn up heads about half of the time. When a weather reporter says, "there is a 10% chance of rain tomorrow," she is basing that on prior evidence; that out of all days with similar weather patterns, it has rained on 1 out of 10 of those days.

Another view would be **subjective** in nature, in other words an educated guess. If someone asked you the probability that the Seattle Mariners would win their next baseball game, it would be impossible to conduct an experiment where the same two teams played each other repeatedly, each time with the same starting lineup and starting pitchers, each starting at the same time of day on the same field under the precisely the same conditions. Since there are so many variables to take into account, someone familiar with baseball and with the two teams involved might make an educated guess that there is

a 75% chance they will win the game; that is, if the same two teams were to play each other repeatedly under identical conditions, the Mariners would win about three out of every four games. But this is just a guess, with no way to verify its accuracy, and depending upon how educated the educated guesser is, a subjective probability may not be worth very much.

We will return to the experimental and subjective probabilities from time to time, but in this course we will mostly be concerned with **theoretical** probability, which is defined as follows: Suppose there is a situation with n equally likely possible outcomes and that m of those n outcomes correspond to a particular event; then the **probability** of that event is defined as $\frac{m}{n}$.

23.1 Basic Concepts

If you roll a die, pick a card from deck of playing cards, or randomly select a person and observe their hair color, we are executing an experiment or procedure. In probability, we look at the likelihood of different outcomes. We begin with some terminology.

Events and Outcomes

The result of an experiment is called an **outcome**.

An **event** is any particular outcome or group of outcomes.

A **simple event** is an event that cannot be broken down further

The **sample space** is the set of all possible simple events.

Example 1

If we roll a standard 6-sided die, describe the sample space and some simple events.

Solution

The sample space is the set of all possible simple events: {1, 2, 3, 4, 5, 6}

Some examples of simple events:

We roll a 1

We roll a 5

Some compound events:

We roll a number bigger than 4

We roll an even number

Basic Probability

Given that all outcomes are equally likely, we can compute the probability of an event *E* using this formula:

$$P(E) = \frac{\text{Number of outcomes corresponding to the event E}}{\text{Total number of equally-likely outcomes}}$$

Example 2

If we roll a 6-sided die, calculate

a) *P(rolling a 1).*

b) *P(rolling a number bigger than 4)* Recall that the sample space is {1,2,3,4,5,6}

c) There is one outcome corresponding to "rolling a 1", so the probability is $\frac{1}{6}$

d) There are two outcomes bigger than a 4, so the probability is $\frac{2}{6} = \frac{1}{3}$

Probabilities are essentially fractioning and can be reduced to lower terms like fractions.

Example 3

Let's say you have a bag with 20 cherries, 14 sweet and 6 sours. If you pick a cherry at random, what is the probability that it will be sweet?
There are 20 possible cherries that could be picked, so the number of possible outcomes is 20. Of these 20 possible outcomes, 14 are favorable (sweet), so the probability that the cherry will be sweet is $\frac{14}{20} = \frac{7}{10}$

Activity 1

At some random moment, you look at your clock and note the minutes reading.
a. What is probability the minutes reading is 15?
b. What is the probability the minutes reading is 15 or less?

Cards

A standard deck of 52 playing cards consists of four **suits** (hearts, spades, diamonds and clubs). Spades and clubs are black while hearts and diamonds are red. Each suit contains 13 cards, each of a different **rank**: An Ace (which in many games functions as both a low card and a high card), cards numbered 2 through 10, a Jack, a Queen and a King.

Example 4

Compute the probability of randomly drawing one card from a deck and getting an Ace.

There are 52 cards in the deck and 4 Aces so $P(Ace) = \frac{4}{52} = \frac{1}{13} = 0.0769$

We can also think of probabilities as percent: There is a 7.69% chance that a randomly selected card will be an Ace.

Notice that the smallest possible probability is 0 – if there are no outcomes that correspond with the event. The largest possible probability is 1 – if all possible outcomes correspond with the event.

Certain event and Impossible event

An event that has a probability of 0 is called an impossible event.

An event that has a probability of 1 is called a certain event

Notice that the probability of any event must be $0 \le P(E) \le 1$

23.2 Working with Events

23.2.1 Complementary Events

Now let us examine the probability that an event does **not** happen. As in the previous section, consider the situation of rolling a six-sided die and first compute the probability of rolling a six: the answer is $P(\text{six}) = \frac{1}{6}$.

Now consider the probability that we do *not* roll a six: there are 5 outcomes that are not a six, so the answer is $P(\text{not a six}) = \frac{5}{6}$.

Notice that $P(\text{six}) + P(\text{not a six}) = \frac{1}{6} + \frac{5}{6} = \frac{6}{6} = 1$

This is not a coincidence. Consider a generic situation with n possible outcomes and an event E that corresponds to m of these outcomes. Then the remaining $n - m$ outcomes correspond to E not happening, thus

$$P(\text{not } E) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(E)$$

Complement of an Event

The **complement** of an event is the event "E doesn't happen"

The notation $\bar{E}$ is used for the complement of event E.

We can compute the probability of the complement using

$P(\bar{E}) = 1 - P(E)$

Notice also that $P(E) = 1 - P(\bar{E})$

Example 5

If you pull a random card from a deck of playing cards, what is the probability it is not a heart?

There are 13 hearts in the deck, so $P(heart) = \frac{13}{52} = \frac{1}{4}$

The probability of *not* drawing a heart = $P(not\ heart) = 1 - P(heart) = 1 - \frac{1}{4} = \frac{3}{4}$

Example 6

Suppose we flipped a coin and rolled a die and wanted to know the probability of getting a head on the coin and a 6 on the die.

We could list all possible outcomes:

$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$.

Notice there are $2.6 = 12$ total outcomes.

Out of these, only 1 is the desired outcome, so the probability is $\frac{1}{12}$.

The previous example was looking at two independent events.

Independent Events

Events A and B are **independent events** if the probability of Event B occurring is the same whether or not Event A occurs.

Example 7

Are these events independent?

a) A fair coin is tossed two times. The two events are (1) first toss is a head and (2) second toss is a head.

b) The two events (1) "It will rain tomorrow in Kandy" and (2) "It will rain tomorrow in Kegalle".

c) You draw a card from a deck, then draw a second card without replacing the first.

a) The probability that a head comes up on the second toss is 1/2 regardless of whether or not a head came up on the first toss, so these events are independent.
b) These events are not independent because it is more likely that it will rain in Kegalle on days it rains in Kandy than on days it does not.
c) The probability of the second card being red depends on whether the first card is red or not, so these events are not independent.

When two events are independent, the probability of both occurring is the product of the probabilities of the individual events.

$P(A \cap B)$ for independent events

If events *A* and *B* are independent, then the probability of both *A* and *B* occurring is

$$P(A \cap B) = P(A) \cdot P(B)$$

where $P(A \cap B)$ is the probability of events *A* and *B* both occurring, $P(A)$ is the probability of event *A* occurring, and $P(B)$ is the probability of event *B* occurring

If you look back at the coin and die example from earlier, you can see how the number of outcomes of the first event multiplied by the number of outcomes in the second event multiplied to equal the total number of possible outcomes in the combined event.

Example 8

In your drawer you have 10 pairs of socks, 6 of which are white, and 7 tee shirts, 3 of which are white. If you randomly reach in and pull out a pair of socks and a tee shirt, what is the probability both are white?

The probability of choosing a white pair of socks = $\frac{6}{10}$

The probability of choosing a white tee shirt = $\frac{3}{7}$

The probability of both being white = $\frac{6}{10}.\frac{3}{7} = \frac{9}{35}$

Activity 24.2

A card is pulled a deck of cards and noted. The card is then replaced, the deck is shuffled, and a second card is removed and noted. What is the probability that both cards are Aces?

Suppose we flipped a coin and rolled a die and wanted to know the probability of getting a head on the coin *or* a 6 on the die.

Here, there are still 12 possible outcomes:

$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$

By simply counting, we can see that 7 of the outcomes have a head on the coin *or* a 6 on the die *or* both – we use *or* inclusively here (these 7 outcomes are $H1, H2, H3, H4, H5, H6, T6$), so the probability is $\frac{7}{12}$.

How could we have found this from the individual probabilities?

As we would expect $\frac{1}{2}$, of these outcomes have a head, and $\frac{1}{6}$ of these outcomes have a 6 on the die.

If we add these $\frac{1}{2} + \frac{1}{6} = \frac{2}{3}$ which is not the correct probability.

Looking at the outcomes we can see why: the outcome H6 would have been counted twice, since it contains both a head and a 6;

The probability of both a head *and* rolling a 6 = $\frac{1}{12}$

If we subtract out this double count, we have the correct probability = $\frac{8}{12} - \frac{1}{12} = \frac{7}{12}$

***P*(*A* ∪ *B*)**

The probability of either *A* or *B* occurring (or both) is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

23.2.3 Conditional Probability

Often it is required to compute the probability of an event given that another event has occurred

Conditional Probability

The probability the event *B* occurs, given that event *A* has happened, is represented as

$$P(B \mid A)$$

This is read as "the probability of *B* given *A*"

Example 9

Find the probability that a die rolled shows a 6, given that a flipped coin shows a head.

Solution

These are two independent events, so the probability of the die rolling a 6 is $\frac{1}{6}$, regardless of the result of the coin flip.

Example 10

The table below shows the number of survey subjects who have received and not received a speeding ticket in the last year, and the color of their car. Find the probability that a randomly chosen person:

a) Has a speeding ticket *given* they have a red car

b) Has a red car *given* they have a speeding ticket

	Speeding ticket	No speeding ticket	Total
Red car	15	135	150
Not red car	45	470	515
Total	60	605	665

Solution

a) Since we know the person has a red car, we are only considering the 150 people in the first row of the table. Of those, 15 have a speeding ticket, so

$$P(\text{ticket} \mid \text{red car}) = \frac{15}{150} = \frac{1}{10} = 0.1$$

b) Since we know the person has a speeding ticket, we are only considering the 60 people in the first column of the table. Of those, 15 have a red car, so

$$P(\text{red car} \mid \text{ticket}) = \frac{15}{60} = \frac{1}{4} = 0.25.$$

These kinds of conditional probabilities are what insurance companies use to determine your insurance rates. They look at the conditional probability of you having accident, given your age, your car, your car color, your driving history, etc., and price your policy based on that likelihood.

Conditional Probability Formula

If Events A and B are not independent, then

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$

Example 11

If you pull 2 cards out of a deck, what is the probability that both are spades?

Solution

The probability that the first card is a spade is $\frac{13}{52}$.

The probability that the second card is a spade, given the first was a spade, is $\frac{12}{51}$, since there is one less spade in the deck, and one less total cards.

The probability that both cards are spades is $\frac{13}{52}\cdot\frac{12}{51}=\frac{156}{2652}\approx 0.0588$

23.3 Bayes Theorem

In this section we concentrate on the more complex conditional probability problems we began looking at in the last section.

Example 12

Suppose a certain disease has an incidence rate of 0.1% (that is, it afflicts 0.1% of the population). A test has been devised to detect this disease. The test does not produce false negatives (that is, anyone who has the disease will test positive for it), but the false positive rate is 5% (that is, about 5% of people who take the test will test positive, even though they do not have the disease). Suppose a randomly selected person takes the test and tests positive. What is the probability that this person actually has the disease?

Bayes' Theorem

$$P(A\mid B)=\frac{P(A)P(B\mid A)}{P(A)P(B\mid A)+P(\bar{A})P(B\mid \bar{A})}$$

In our earlier example, this translates to

$$P(\text{disease}\mid\text{positive})=\frac{P(\text{disease})P(\text{positive}\mid\text{disease})}{P(\text{disease})P(\text{positive}\mid\text{disease})+P(\text{no disease})P(\text{positive}\mid\text{no disease})}$$

Plugging in the numbers gives

$$P(\text{disease}\mid\text{positive})=\frac{(0.001)(1)}{(0.001)(1)+(0.999)(0.05)}\approx 0.0196$$

which is exactly the same answer as our original solution.

Example 13

A certain disease has an incidence rate of 2%. If the false negative rate is 10% and the false positive rate is 1%, compute the probability that a person who tests positive actually has the disease.

Imagine 10,000 people who are tested. Of these 10,000, 200 will have the disease; 10% of them, or 20, will test negative and the remaining 180 will test positive. Of the 9800 who do not have the disease, 98 will test positive. So, of the 278 total people who test positive, 180 will have the disease. Thus

$$P(\text{disease} \mid \text{positive}) = \frac{180}{278}$$

so about 65% of the people who test positive will have the disease.

Using Bayes theorem directly would give the same result:

$$P(\text{disease} \mid \text{positive}) = \frac{(0.02)(0.90)}{(0.02)(0.90) + (0.98)(0.01)} = \frac{0.018}{0.0278} \approx 0.647$$

23.4 Counting

Counting? You already know how to count, or you wouldn't be taking a college-level math class, right? Well yes, but what we'll really be investigating here are ways of counting *efficiently*. When we get to the probability situations a bit later in this chapter, we will need to count some *very* large numbers, like the number of possible winning lottery tickets. One way to do this would be to write down every possible set of numbers that might show up on a lottery ticket but believe me: you don't want to do this.

23.3.1 Basic Counting

We will start, however, with some more reasonable sorts of counting problems in order to develop the ideas that we will soon need.

Basic Counting Rule

If we are asked to choose one item from each of two separate categories where there are *m* items in the first category and *n* items in the second category, then the total number of available choices is $m \cdot n$.

This is sometimes called the multiplication rule for probabilities.

23.3.2 Permutations

In this section we will develop an even faster way to solve some of the problems we have already learned to solve by other means.

Factorial

$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$

If we are choosing *r* items out of *n* possibilities *without replacement* and where the *order of selection is important,* in this situation we write:

Permutations

$$^{n}P_{r} = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)$$

We say that there are $^{n}P_{r}$ **permutations** of size *r* that may be selected from among *n* choices *without replacement* when *order matters*.

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

24.3.3 Combinations

In the previous section we considered the situation where we chose r items out of n possibilities without replacement and where the order of selection was important. We now consider a similar situation in which the order of selection is not important.

We can generalize the situation in this example above to any problem of choosing a collection of items without replacement where the order of selection is **not** important.

If we are choosing *r* items out of *n* possibilities, the number of possible choices will be given by $\frac{{}^{n}P_{r}}{{}^{r}P_{r}}$, and we could use this formula for computation.

If we are choosing *r* items out of *n* possibilities *without replacement* where the *order of selection is **not** important*

Combinations

$$ {}^{n}C_{r} = \frac{{}^{n}P_{r}}{{}^{r}P_{r}} $$

We say that there are ${}_{n}C_{r}$ **combinations** of size *r* that may be selected from among *n* choices *without replacement* where *order doesn't matter*.

We can also write the combinations formula in terms of factorials:

$$ {}^{n}C_{r} = \frac{n!}{(n-r)!r!} $$

Activity 2

1. A ball is drawn randomly from a jar that contains 6 red balls, 2 white balls, and 5 yellow balls. Find the probability of the given event.
    a. A red ball is drawn
    b. A white ball is drawn

2. Suppose you write each letter of the alphabet on a different slip of paper and put the slips into a hat. What is the probability of drawing one slip of paper from the hat at random and getting?
    a. A consonant
    b. A vowel

3. A group of people were asked if they had run a red light in the last year. 150 responded "yes", and 185 responded "no". Find the probability that if a person is chosen at random, they have run a red light in the last year.

4. In a survey, 205 people indicated they prefer cats, 160 indicated they prefer dots, and 40 indicated they don't enjoy either pet. Find the probability that if a person is chosen at random, they prefer cats.

5. Compute the probability of tossing a six-sided die (with sides numbered 1 through 6) and getting a 5.

6. Compute the probability of tossing a six-sided die and getting a 7.

7. Giving a test to a group of students, the grades and gender are summarized below. If one student was chosen at random, find the probability that the student was female.

	A	B	C	Total
Male	8	18	13	39
Female	10	4	12	26
Total	18	22	25	65

8. The table below shows the number of credit cards owned by a group of individuals. If one person was chosen at random, find the probability that the person had no credit cards.

	Zero	One	Two or more	Total
Male	9	5	19	33

Female	18	10	20	48
Total	27	15	39	81

9. Compute the probability of tossing a six-sided die and getting an even number.

10. Compute the probability of tossing a six-sided die and getting a number less than 3.

11. If you pick one card at random from a standard deck of cards, what is the probability it will be a King?

Summary

In this session, we have introduced the concepts of a random experiment, probability, conditional probability, independent events. We have defined the probability of events in a random experiment with equally likely outcomes. We have described Baya's theorem on conditional probability. Using Baya's theorem, we have solved some problems involving conditional probability. We have given the concepts of permutations and combinations. We have construct the formulae to calculate the number of permutations and combinations.

Learning outcomes

At the end of this session the student should be able to

1.Describe the concepts of a random experiment, probability, conditional probability, independent events,

2.Define and calculate the probabilities of events in a given random experiment.

3.Introduce Baya's theorem on probability and calculate probabilities by using Baya's theorem.

4.Define and calculate the permutations and combinations.