STA1510 (BASIC STATISTICS) AND STA1610 (INTRODUCTION TO STATISTICS) NOTES PART 1

Dear student,

I pray that this information finds you in good health. These notes are written an integral part of Unisa's student support programme – a programme that seeks to bridge the distance between the student, the study material and the lecturers. In this work we discuss chapter 1 to chapter 7 of the prescribed text book. Understanding this work will help you cover all the content assessed in assignment 1.

Please note that parts of these notes are extracts from the prescribed text book, the study guide, other statistics sources and large proportion is based from the lecturer's synthesis. This means that many real life examples used are based on the lecturer's understanding and subject to change. In a situation where you do not understand or you disagree with the author please share your view with us. I trust that this information will be helpful and rewarding.

Rajab Ssekuma
Lecturer
Department of Statistics
Tel: +27 12 429 6634
email: ssekur@unisa.ac.za

University of South Africa
Preller Street, Muckleneuk Ridge, Pretoria
PO Box 392, UNISA, 0003, South Africa
Call centre 0861 670 411 / +27 11 670-9000
www.unisa.ac.za

STUDY UNIT 1

Key questions for this unit

What is Statistics?
What is the difference between Population and a Sample?
What is the difference between a parameter and a Statistic?
Distinguish between Qualitative and Quantitative variables.
Distinguish between Nominal and Ordinal variables.
Distinguish between Discrete and Continuous variables.
Distinguish between Scale and Ratio variables.

DEFINITIONS

Statistics is a way to get information from data. In other words, statistics is a tool ''like a toolbox'' used to extract information form collected data. Statistics has two main branches; Descriptive and Inferential statistics.

Descriptive statistics: This deals with methods of organising, summarizing and presenting data in a convenient and informative way. In descriptive statistics, we use graphs, tables, numerical measures like mean, range, median mode etc to summarise data.

Inferential statistics: This is a body of methods used to draw conclusions or inferences about characteristics of population based on sample data.

A population: This is the group of all items of interest to a statistics practitioner. It could be people, cars, house etc. It is frequently very large and may, in fact, be infinitely large.

A sample: This is a set of data drawn from the studied population. In other words, a sample is part of a population.

A parameter: Any descriptive measure of a population is a parameter. Examples of parameters include; population size (N), population variance (sigma-squared σ^2), population standard deviation (sigma σ). In other words, any numerical summary from a population is a parameter.

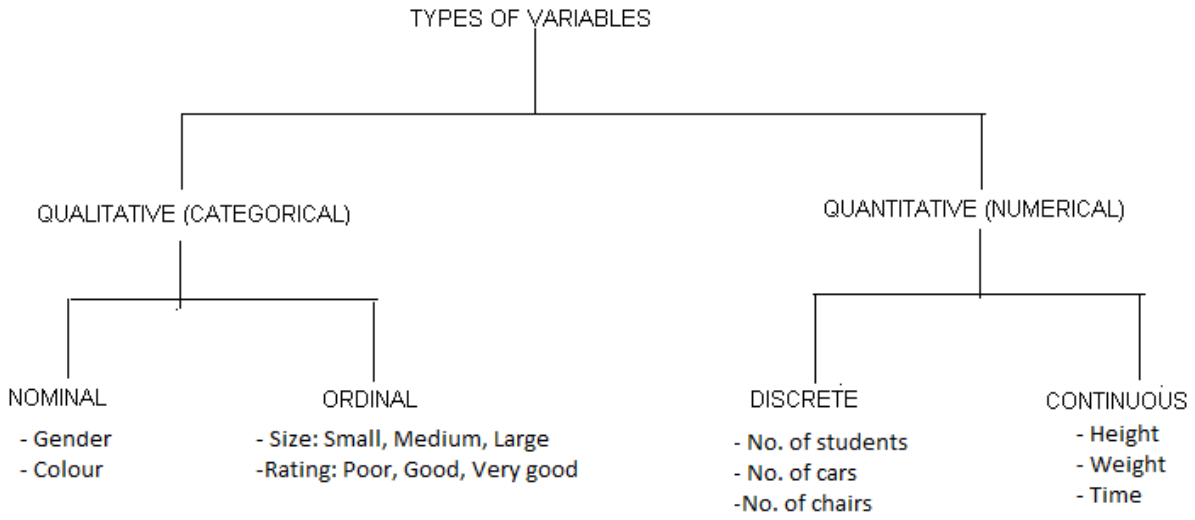
A statistic: Any descriptive measure of a sample is a statistic. Examples include; sample size (n), sample variance (s^2), sample standard deviation (s). In other words, any numerical summary from a sample is a statistic.

TYPES OF VARIABLES

1.1 Introduction to this study unit

This unit introduces the concepts of types of variables. There are basically two types of variables in statistics; Qualitative (*think in terms of quality of life*) and Quantitative (*if you quantify something you could count it*). Qualitative variables are then classified into nominal and ordinal variables. Quantitative variable can be classified into discrete and continuous variables. Once you know your variable is quantitative, it helps to ask yourself if you have actually counted (*then discrete*) or measured (*then continuous*), when you gather the values.

The diagram below is a mind map of what we shall focus on in this section. Please note that though we have to know how to differentiate between variables, questions in this section are set in application form as we shall see when we get to examples and exercises.



1.2 Qualitative Vs Quantitative variables

1.2.1 Qualitative Variables (Categorical Variable)

Also known as categorical variables, qualitative variables are variables with no natural sense of ordering. They are therefore measured on a nominal scale. For instance, hair colour (Black, Brown, Gray, Red, Yellow) is a qualitative variable, as is name (Adam, Becky,

Christina, Dave . . .). Qualitative variables can be coded to appear numeric but their numbers are meaningless, as in male=1, female=2. Variables that are not qualitative are known as quantitative variables.

1.2.2 Quantitative Variables

Quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as colour, religion, city of birth, sport in which there is no ordering or measuring involved.

1.3 Nominal Vs Ordinal variables

1.3.1 Nominal Variables

A nominal variable has values which have no numerical value. As a result the order or sequence of nominal variables is not prescribed. Examples of nominal variables are gender, occupation.

1.3.2 Ordinal variables

An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, economic status, with three categories (low, medium and high). In addition to being able to classify people into these three categories, you can order the categories as low, medium and high.

Please note that the major difference between ordinal and nominal is that order is considered to be important in ordinal variables than in nominal variables.

1.4 Discrete Vs Continuous variables

1.4.1 Discrete variables

Variables that can only take on a finite number of values are called "discrete variables." Or A variable that takes values from a finite or countable set, such as the number of legs of an animal. All qualitative variables are discrete. Some quantitative variables are discrete, such as performance rated as 1,2,3,4, or 5, or temperature rounded to the nearest degree.

1.4.2 Continuous variables

A continuous variable is one for which, within the limits the variable ranges, any value is possible. For example, the variable "Time to solve a mathematical problem" is continuous since it could take 2 minutes, 2.13 minutes etc. to finish a problem.

I like telling my students to look at discrete variables as countable variables with gaps in between say the number of students in a discussion class, and to look at continuous

*variables as countable with decimal point like money R5.13, time, height e.t.c. Please note that this is **not** a standard difference between the two but a personal option.*

1.5 Interval Vs Ratio variables

1.5.1 Interval variables

An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced. For example, suppose you have a variable such as annual income that is measured in Rand, and we have three people who make R10,000, R15,000 and R20,000. The second person makes R5,000 more than the first person and R5,000 less than the third person, and the size of these intervals is the same. If there were two other people who make R90,000 and R95,000, the size of that interval between these two people is also the same (R5,000).

1.5.2 Ratio variables

A variable with the features of interval variable and, additionally, whose any two values have meaningful ratio, making the operations of multiplication and division meaningful.

Now that we are familiar with the definitions, we can take example on how this unit is examined. Please remember that we examine their applications to real life situations in most cases.

Example 1

Which one of the following statements is *incorrect*?

(1) The number of students who attended both discussion classes in 2010 is a discrete variable.

(2) Your marital status is a discrete variable.

(3) Whether one does poor, fair or good in an assignment is an ordinal variable.

(4) The amount of your student loan is a continuous variable.

(5) Your status as a full-time student is a nominal variable.

Solution

The number of students who attended both discussion classes in 2010 a discrete variable (correct).

1.Maritial status (married, not married, single or divorce) is a nominal variable. (Incorrect).

2.Correct.

3.Correct.

4.Correct.

Example 2

The owner of fancy foods chooses a random sample of six people who are at his shop. He asks them a few questions that are summarised as follows:

Sex 1= Male 2= Female	**Age** 1= under 20 2 = 20 to 40 3= 41 to 60 4= over 60	**Method of payment** 1= cash 2= credit card 3= private account	**Satisfaction of service rating** 1= bad 2= average 3= good 4= very good
2	2	1	3
1	2	1	4
1	1	2	1
2	4	3	3
1	3	3	2
1	3	2	3

Consider the following statements:
A: Method of payment is a quantitative variable.
B: The youngest person is male, paid with a credit card and found the service bad.
C: 50% of the people said the service was good.
D: 50% of the males were under 20.
E: The oldest person interviewed said the service was very good.
The correct statements(s) is/are

(1) Only B
(2) C and D
(3) B and C
(4) C,D and E
(5) A and C

Option (1). The youngest person is male, paid with a credit card and found the service bad.

Sex *1= Male*	*Age* *1= under 20*	*Method of payment* *2= credit card*	Satisfaction of service rating 1= bad

SELF ASSESSMENT EXERCISE – TEST YOUR KNOWLEDGE

Question 1

Which one of the following statements is *incorrect*?

(1) Measures for a sample are called statistics while measures for a population are called parameters.
(2) Your marital status is an ordinal variable.
(3) Whether one does poor, fair or good in an assignment is an ordinal variable.
(4) The amount of your student loan is a continuous variable.
(5) The starting salary of MBA graduates is a quantitative variable.

Question 2

Which of the following variables is a qualitative variable?

(1) The most frequent use of your microwave oven (reheating, defrosting, warming, others).
(2) The number of consumers who refuse to answer a telephone survey.
(3) The number of mice used in a maize experiment.
(4) The winning time for a horse running in a Derby.
(5) Weight of a new-born baby.

Question 3

Which one of the following is a discrete variable?

(1) Writing skills of new employees, classified as bad, fair, good and excellent.
(2) A student's yes/no response to a question in a campus newspaper.
(3) The combined weight of parcels sent from a certain post office during a week.
(4) The starting salary of a medical doctor.
(5) The number of students who attended a discussion class.

Question 4
Which of the following statements is *incorrect*?
(1) The number of registered arms dealers in a certain province is a discrete variable.
(2) Your choice of car brand is a nominal variable.
(3) The average mark of statistics students in the exam is a qualitative variable.
(4) The number of building permits for new single-family housing units is a discrete variable.
(5) The opinion of TV viewers on a new program (bad, indifferent, good) is an ordinal variable.

SOLUTIONS TO SELF ASSESSMENT EXERCISES

Question 1
Alternative 2. Your marital status is a nominal variable.

Question 2
Alternative 1. The most frequent use of your microwave oven (reheating, defrosting, warming, others) is a qualitative variable.

Question 3
Alternative 5. The number of students who attended a discussion class is a discrete random variable.

Question 4
Alternative 3. The average mark of statistics students in the exam is a quantitative variable.

STUDY UNIT 2

2 DESCRIPTION OF DATA

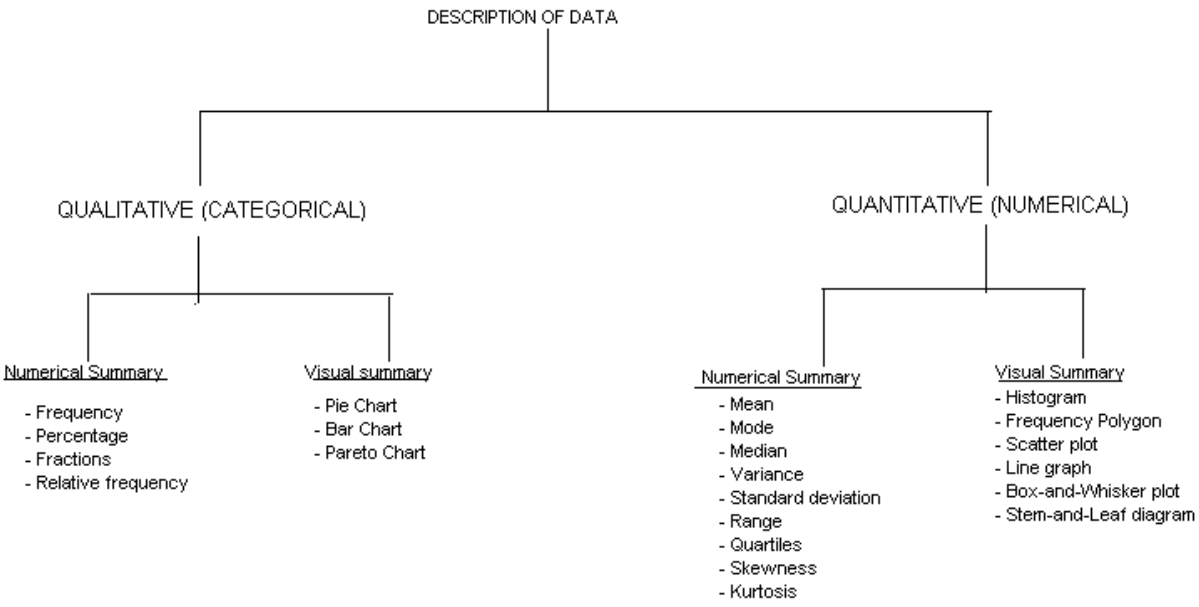
Key questions for this unit

Distinguish between Qualitative and Quantitative data.
How would you represent qualitative data both numerically and visually?
How would you represent qualitative data both numerically and visually?
Interpretation of a frequency distribution and the stem-and-leaf diagram

2.1 Introduction to this study unit

Now that we know that in data can be classified in two ways, that is, qualitative and quantitative. We pose a question, how would we describe data? Description of data can be done in two ways: numerically and visually as shown in the following flow diagram

The diagram below is a mind map of what we shall focus on in this section. Please note that questions in this section are most theoretical. In the past mostly examiners have focused on the stem-and –leaf diagram.



2.1 Qualitative Data:

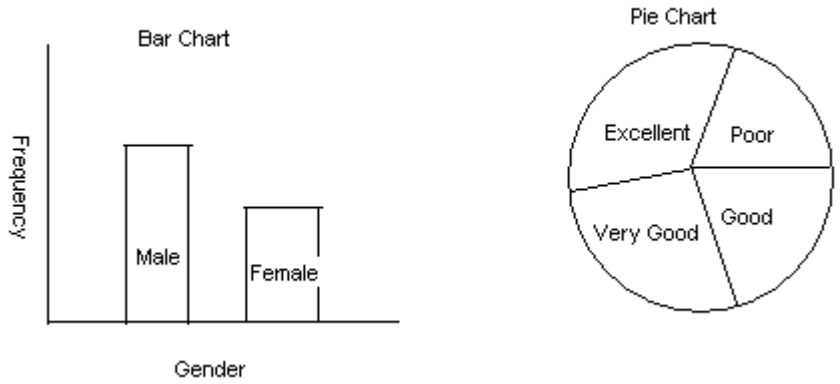
Remember in study unit 1 we classified qualitative as categorical data. Think in terms of gender, say, you have a class of female and male students.

2.1.1 Numerical Summary

To summarise this data numerically, you would perhaps first think of how many are female or male (frequency), what percentage are male or female, what is the fraction (Ratio) of male to female which is the relative frequency. There is not too much we can do in terms of summarising qualitative data numerically.

2.1.2 Visual Summary

Visually if data is qualitative, in most cases we use the bar chart or the pie chart to represent it. The figures below are examples of Bar chart and Pie charts respectively.



2.2 Quantitative data

From study unit 1, we classified quantitative data as countable or measurable on a numeric scale. In this case think in terms of salaries.

2.2.1 Numerical Summary

If you are to access employees salaries, you would first look at the average (mean) salary, the middle(median) salary, the most occurring (mode) salary, the variance (see study unit 3), the standard deviation (see study unit 3), the range , kurtosis, correlation, skewness. In brief most of the statistical analysis is done on quantitative data.

2.2.2 Visual Summary

Visually if data is quantitative, we use the histogram, the frequency polygon, the stem-and-leaf diagram, scatter plot, line graph and the box-and-whisker plot to represent it. With the exception of the stem-and-leaf diagram, the reset are examinable theoretically. If the examiner wants to examine the features of any diagram, it will be drawn for you.

Example 3

Which one of the following statements is *incorrect*?

(1) A bar graph cannot be used for two categorical variables.
(2) Adjacent rectangles in a histogram share a common side.
(3) A stem-and –leaf plot provides sufficient information to determine whether a dataset contains an outlier.
(4) Box plots display the centre, spread and outliers of a distribution.
(5) A histogram is better than a box plot for evaluating the shape of a dataset.

Solution

Option 1:
A bar graph can be used for two categorical variables

Example 4

The following table gives the cumulative relative frequency of the mass of 100 youngsters:

Class interval	Cumulative relative frequency
$19.5 - 29.5$	0.04
$29.5 - 39.5$	0.18
$39.5 - 49.5$	0.35
$49.5 - 59.5$	0.60
$59.5 - 69.5$	0.80
$69.5 - 79.5$	0.94
$79.5 - 89.5$	1.00

Which of the following statements is incorrect?
(1) The interval 49.5-59.5 has the largest number of observations.
(2) There are 35 youngsters having a mass of more than 49.5 kg.
(3) The interval 39.5-59.5 has 42 observations.
(4) 94% of the youngsters have a mass of less than 79.5 kg.
(5) The interval 19.5-29.5 has 4 observations.

Solution

Class interval	Cumulative relative frequency	Relative frequency	Frequency
$19.5 – 29.5$	0.04	0.04	4
$29.5 – 39.5$	0.18	0.14	14
$39.5 – 49.5$	0.35	0.17	17
$49.5 – 59.5$	0.60	0.25	25
$59.5 – 69.5$	0.80	0.20	20
$69.5 – 79.5$	0.94	0.14	14
$79.5 – 89.5$	1.00	0.06	6

Option (1) Correct
There are 25 youngsters in the interval $49.5 – 59.5$

Option (2) Incorrect
There are (25 + 20 + 14 + 6) = 65 youngsters having a mass of more than 49.5 kg

Option (3) Correct
The interval $39.5 – 59.5$ 39 has (17 + 25) = 42 observations.

Option (4) Correct
The number of youngsters with less than 79.5 kg is (4 + 14 + 17 + 25 + 20 + 14) = 94.
The percentage is therefore $\frac{94}{100} \times 100 = 94\%$

Option (5) Correct
The interval $19.5 – 29.5$ has 4 observations

STUDY UNIT 3

In this study unit we discuss the following

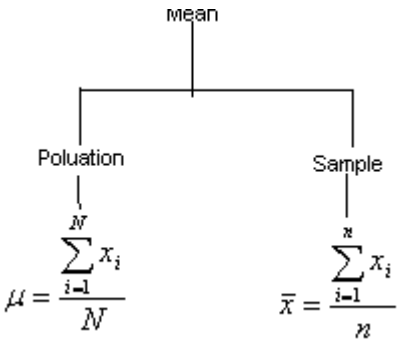
1. *Measures of location / measures of central tendency.*
2. *Measures of spread / measures of dispersion*
3. *Quartiles, Box plots and Percentiles*
4. *Measures of linear relationships*

3.1 Measures of central tendency / Measures of location

These include the mean, the median and the mode.

3.1.1 The mean / Average

The mean (averages) is calculated by summing all the observations and dividing by their number. Calculation of the mean depends on the source of the data. This can either be the population or the sample



Example:

Calculate the mean following sample data: 29, 39, 43, 52, 39

The sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{29+39+43+52+39}{5}$$

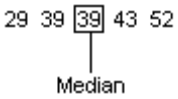
$$= \frac{202}{5}$$

$$= 40.4$$

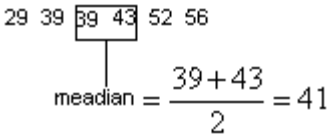
3.1.2 The median

The median of the data set is the middle value of an ordered data set. Before calculating the median, the data set has to be arranged in order (either ascending or descending).
Please note that:

(i) If the data set is odd in number, its quite easy to identify the middle value which, is the median.
For example: consider the following data set: 29, 39, 52, 43, 39



(ii) If the data set is even in number, the median is the average of the two middle value.
For example: Consider the following data set; 29, 43, 39, 39, 56, 52



3.1.3 The mode

The mode is the most occurring observation in a data set. Or we can say the observation with the highest frequency. For example in the following data set: 29, 39, 39, 43, 52, 56, the mode is 39.
Please note that:

(i) It's possible for a data set not to have a mode. E.g: there is no model in the following data set 29, 39, 43. However, this does not mean that the mode is zero. If you say that mode is zero, it implies that the value (0) occurs most, which is not true in this case.

(ii) It's also possible for the data set to have two modes. Such a data set is called a bimodal data set. Plotting such a data set will lead to two peaks as shown below

Example 5

The following stem-and-leaf display is for a set of values where the stem is formed by the units and the leaf represents the decimal digits:

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

Which of the following statements is *incorrect*?

(1) The number of values larger than 4.0 is 12

(2) The median of this data set is 3.7

(3) 20% of the values lie between the values 2 and 3

(4) The mode of the data set is 3.6

(5) The sixth smallest value in the data set is 2.8

Solution
Option (2)

$$\text{Median} = \frac{3.7 + 3.8}{2} = 3.75$$

3.2 Measures of Dispersion / Measures of Spread

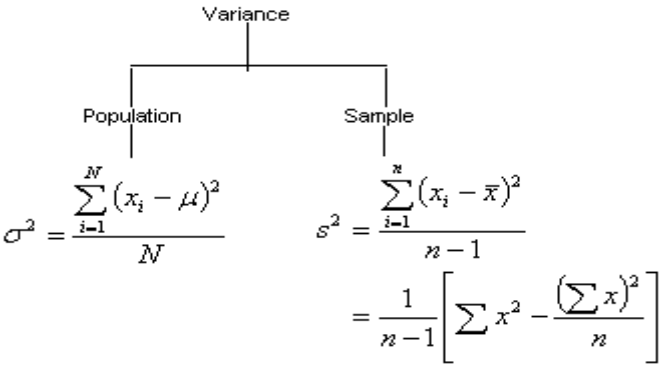
3.2.1 Range

This is perhaps the most easiest to calculate. The range is the difference between the largest and the smallest observation of a data set.

Range = Largest - Smallest observation

3.2.2 Variance

Calculation of the variance depends on the source of the data, which is either from a population of the sample.

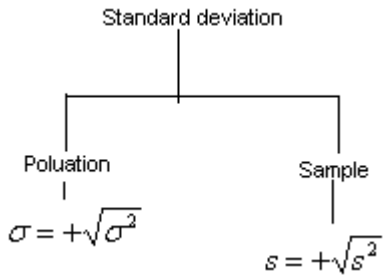


3.2.3 Standard deviation

The standard deviation is the positive square root of the variance.

$$\text{Std deviation} = +\sqrt{\text{var}\,iance}$$

Calculation of the standard deviation also depends on the source of the data, which is either from a population of the sample.



Please note that the mean, the standard deviation and the variance can also be excuted directly from any scientific calculator. If you are using the SHARP EL531WH advanced D.A.L like mine, you follow the following steps.

1. Set you calculator in Stat 0 mode as follows; Press mode, press 1, press 0. You will have

   Stat 0 on your screen
2. Enter the data set as follows. E.g. Consider the data set as follows: 42, 45, 48, 79

   42 Press m+ a button next to STO
   45 Press m+ a button next to STO
   48 Press m+ a button next to STO
   79 Press m+ a button next to STO

The m+ button next to the STO button stores the observations in the memory of your calculator. Your will have

DATA SET=
4

This means that you have 4 observations stored in the memory.

3. To get the mean, press RCL and 4. On the top of number 4, there is a small $\bar{x}$, which standard for the mean. It's green in colour and to use green keys we either use RCL (recall) or use ALPHA. You will have

   $\bar{x}$ =
   53.5

   This is equivalent to calculating the mean manually as;

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$= \frac{42+45+48+79}{4}$$
$$= \frac{214}{4}$$
$$= 53.5$$

4. To get the standard deviation, we press RCL(recall) , then press number 5. The standard deviation is the small green s_x on the top of number 5. You will have

   s_x =
   17.17556404

   This will have saved you time spent in using the following formula.

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

$$=\sqrt{\frac{1}{n-1}\left[\sum x^2-\frac{(\sum x)^2}{n}\right]}$$

$$=\sqrt{\frac{1}{4-1}\left[12334-\frac{(214)^2}{4}\right]}$$

$$=\sqrt{\frac{1}{3}[12334-11449]}$$

$$=\sqrt{\frac{(885)}{3}}$$

$$=\sqrt{295}$$

$$=17.17556404$$

Please remember that $\sum x^2=42^2+45^2+48^2+79^2=12334$. We clearly see that working it out manually takes a lot of time and we are likely to make mistakes.

5. To get the variance using our calculator, we just need to square the answer of the standard deviation. After pressing RCL number 5, press x^2, then press equal sign. You will have

$$s_x^{\,2}=$$

$$295$$

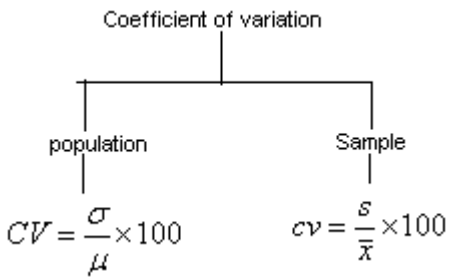
Otherwise we have to square $(17.17556404)=295$. Please remember that since

$$\text{Std deviation}=+\sqrt{\text{var}\,iance}$$

Then $\text{variance}=(\text{Standard deviation})^2$

3.2.4 Coefficient of variation

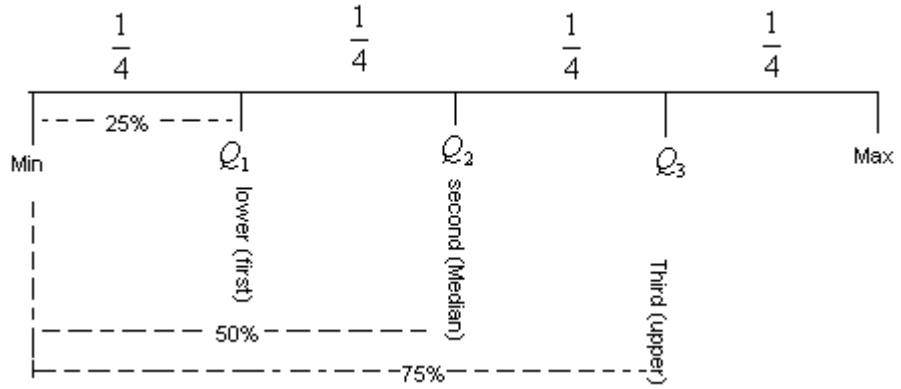
This measures the scatter in the data relative to the mean. In many Statistics book, its expressed as a a percentage. The coefficient of variation also depends on the source of the data.



3.3 Quartiles Box plot and Percentiles

3.3.1 The first (lower), the second (median) and the upper (third) quartiles

For purposes of this module, we shall only discuss the quartiles. The word quartile perhaps comes from quarter $\left(\frac{1}{4}\right)$. This means that quartiles divide a data set into four equal parts as follows:



The calculation of the quartiles requires to first arrange the data set in order, preferably in ascending order. Once the data is arranged in order, we then obtain the position of a particular quartile as follows:

(i) The location/position of first (lower) quartile is given by $\left(\frac{n+1}{4}\right)$ where n is the number of observations in the given data set.

(ii) The position of the second (median) is given by $2\left(\frac{n+1}{4}\right)$

(iii) The position of the third/ upper quartile is $3\left(\frac{n+1}{4}\right)$

Please note that according to some books, like Business Statistics by Levene if;

(i) $\left(\frac{n+1}{4}\right)=\frac{10+1}{4}=2.75$, we then take Q_1 to be the 3th observation.

(ii) $\left(\frac{n+1}{4}\right)=2.35$, we then take Q_1 to be the 2th observation.

This means that if the decimal point is 5 and above, you round it off to the nearest whole number.

Example 6

Consider the following data set : 240, 260, 350, 350, 420, 510, 530, 550.

The position of lower/ first quartile is $\left(\frac{n+1}{4}\right)=\frac{8+1}{4}=\frac{9}{4}=2.25$. Hence, the values of $Q_1 = 2^{nd}$ observation, which is 260.

The position of upper/ third quartile is $3\left(\frac{n+1}{4}\right)=\frac{3(8+1)}{4}=\frac{27}{4}=6.75$. Hence, the values of $Q_1 = 7^{th}$ observation, which is 530.

3.3.2 The interquartile range (IQR)

The Interquartile range is the difference between the third and the first quartiles.

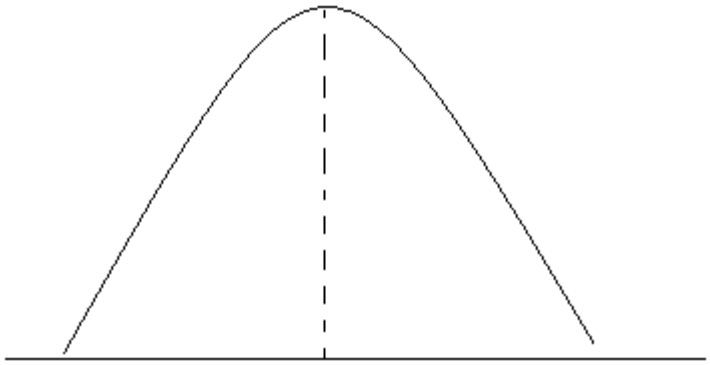
$IQR = Q_3 - Q_1$

Considering the above data set $IQR = 530 - 260 = 270$

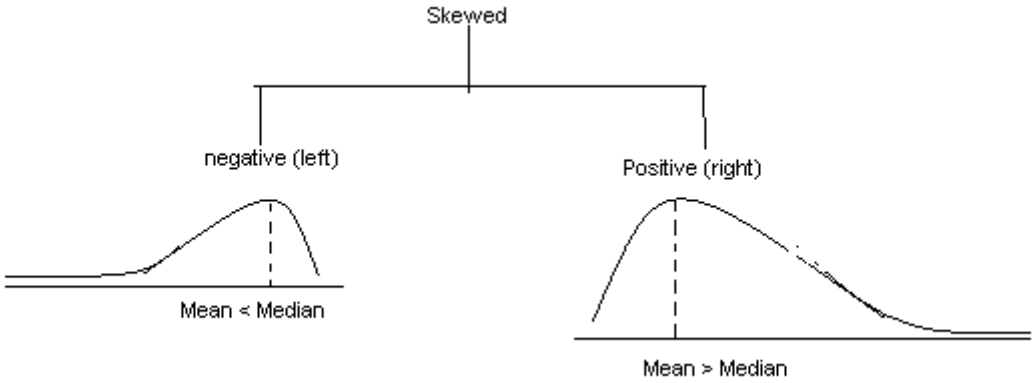
3.33 The distribution of data

Data can be symmetrical (normally) distributed or can be skewed.

In symmetrical distribution the values below the mean are distributed exactly as the values above the mean. This can be demonstrated using the following graph.



In skewed distribution, the values are not symmetrical. Skewness can either be negative (left-skewed) or positive (right-skewed). What determines the skewness the position of the longer tail. If the long tail of the distribution on the left, we have negative (left) skewed and if it's on the right we have positive (right) skewed.



Generally, skewness is caused by presence of extreme values.

In left skewness, the extreme values pull the mean downwards so that the mean is less than the median. This is comparable to the examination session. When we write exams, most students tend to finish towards the end of allocated time, although there a few who walk out of the examination center shortly after the start, especially those who work so fast. These few students are the one responsible for the long tail on the left of the distribution.

On other hand, in right skewness, most values are in the lower portion of the distribution. A long tail on the right is caused by the presence extremely large values that pull the mean upwards so that it's greater than the median. This is comparable to salary allocations in most workplace (UNISA inclusive). Most people (including me) get low salaries.

However, there is a category of people (managers, directors, professors etc.) who get huge amount of salaries. These few employees are the one responsible for the long tail on the right of the distribution.

3.4 The measures of linear relationship

We shall discuss much about the calculation of measures of linear relationship when we discuss a chapter on Simple linear regression. In this section, we shall put our emphasis on the interpretation and the understanding of these measures. These include;

3.4.1 Covariance.

The covariance measures the strength of the linear relationship between two numerical variables (X and Y). Say for example the strength of the relationship between income and expenditure. It's is believed that the more you earn, the more you spend. We generally expect this relationship to be positive and increasing. In some economic variables, indication of the relationship is not straight forward. For example, the relationship between interest rates and the oil price. In this we have to calculate the covariance between the two variables.

Calculation of the covariance also depends on the source of data. For this module we concentrate on sample data where the covariance is given by;

$$COV(x;y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

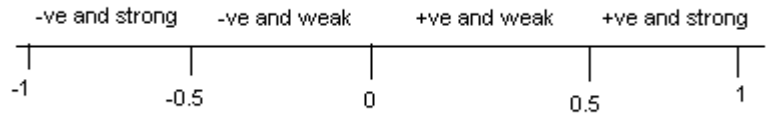
$$=\frac{1}{n-1}\left[\sum x_i y_i-\frac{(\sum x_i)(\sum y_i)}{n}\right]$$

The breakdown of this formulae will be covered under a chapter on simple linear regression.

3.4.2 The coefficient of correlation (r)

This measures the strength and the direction of the relationship between two numerical variables. It lies between -1 and 1. i.e. $-1\le r\le 1$.

The representation of the coefficient of correlation also depends on the original source of data, which is either population or sample. Again, for purposes of this module, we shall stick on the sample coefficient of correlation. Its interpretation can be summarised as follows;



In summary, we can say;

(i) If $r=\pm 1$ we have a perfect positive or a perfect negative relationship. This however, very difficult to meet. If you are in love or you have ever been in love, you perhaps understand what this statement means!

(ii) If $0.5\le r\le 1$ we have a positive strong in magnitude relationship. This can be compared to love at first sight or when you are beginning a love relationship (dating).

(iii) If $-1\le r\le -0.5$ we have strong negative in magnitude relationship. This is comparable to a situation of divorce or in the process of terminating a love relationship.

(iv) If $\pm 0.5 \le r \le 0$ we have generally a weak in magnitude positive or weak negative relationship depending on the sign of coefficient of correlation (r).

Please remember that the calculation of the coefficient of correlation shall be covered in a chapter that deals with Simple linear regression.

SELF ASSESSMENT EXERCISE – TEST YOUR KNOWLEDGE

QUESTION 1

The following is a set of data from a sample of eight students.

12 15 4 9 1 10 6 3

Which of the following statements is *incorrect*?

(1) The minimum value is 1

(2) The median is 7.5

(3) The distribution is symmetrical

(4) The maximum value is 15

(5) The range is 15

QUESTION 2

A study was conducted on the 12-month earnings per share (in rand) of six large airline companies.

4.36 6.19 − 0.42 3.73 0.26 6.27

Based on the above data, which one of the following statements is *incorrect*?

(1) The mean earnings per share is 3.3983.
(2) The sample standard deviation is 2.8811.
(3) The sample variance is 8.300
(4) Only one airline did not make a profit
(5) The coefficient of variation is 1.1795.

QUESTION 3

The following is a set of data from a sample of eight students.

12 15 4 9 1 10 6 3

Which of the following statements is *incorrect*?

(1) The mean is 7.5

(2) The median is 7.5

(3) The interquartile range is 12

(4) The position of the first quartile is 2.25

(5) The third quartile is 12

QUESTION 4

Which one of the following statements is correct?

(1) In a symmetrical distribution the mean, median and mode are not the same.

(2) If the mean is greater than the median this is a negative skew distribution.

(3) If the mean is less than the median this is a positive skewed distribution.

(4) The value of the quartile Q_2 is always equal to the median.

(5) There cannot be more than one mode in the distribution of data.

QUESTION 5

The following data represent the number of children in a sample of 11 families from a certain community:

2 0 4 1 1 5 1 1 4 0 2

Which one of the following statement is incorrect?

(1) The mean is 1.909

(2) The median is 5

(3) The mode is 1

(4) The standard deviation is 1.700

(5) The range is 5

SOLUTIONS TO SELF ASSESSMENT EXERCISE

QUESTION 1

We begin by arranging the data set in ascending order as follows:

1 3 4 6 9 10 12 15

Option (1) Correct

Option (2) Correct

Median $= \frac{6+9}{2} = 7.5$

Option (3) Correct

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+3+4+....+15}{8} = \frac{60}{8} = 7.5$$

Option (4) Correct

Option (5) Incorrect

Range = largest – smallest observation which 15- 1=14

QUESTION 2

Using the calculator as explained in section 3.2, option 1, option 2 option 3 and option 4 are all correct.

The incorrect option should be option (5). This should be

$$cv = \frac{s}{\bar{x}} \times 100$$
$$= \frac{2.881}{3.398} \times 100$$
$$= 84.78\%$$

QUESTION 3

We begin by arranging the data set in ascending order as follows:

1 3 4 6 9 10 12 15

Option (1) Correct

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+3+4+....+15}{8} = \frac{60}{8} = 7.5$$

Option (2) Correct

Median $= \frac{6+9}{2} = 7.5$

Option (3) Incorrect

The position of first / lower quartile is $\frac{(n+1)}{4} = \frac{(8+1)}{4} = 2.25$. Thus the values of $Q_1 = 2^{nd}$ observation which is 3

The position of third / upper quartile is $\frac{3(n+1)}{4} = \frac{3(8+1)}{4} = \frac{27}{4} = 6.75$. Thus the values of $Q_3 = 7^{th}$ observation which is 12

Hence the $IQR = Q_3 - Q_1 = 12 - 3 = 9$

Option (4) Correct

Option (5) Correct

QUESTION 4

Option (4)

The value of the quartile Q_2 is always equal to the median.

QUESTION 5

You can now answer this question on your own.

4 BASIC PROBABILITY

STUDY UNIT 4

Key units to this chapter

Define probability. What is meant by an event?
Understand what is meant with the following concepts: Joint event, Union event, Independent event, Marginal probability, Complement of an event, Mutually exclusive events and Sample space.
Understand conditions under which P(A/B) = P(A)
Probability rules such as Addition rule, Multiplication rule and Complement rule.
Constructing and interpreting a probability tree and the basic concepts of the bayes' law

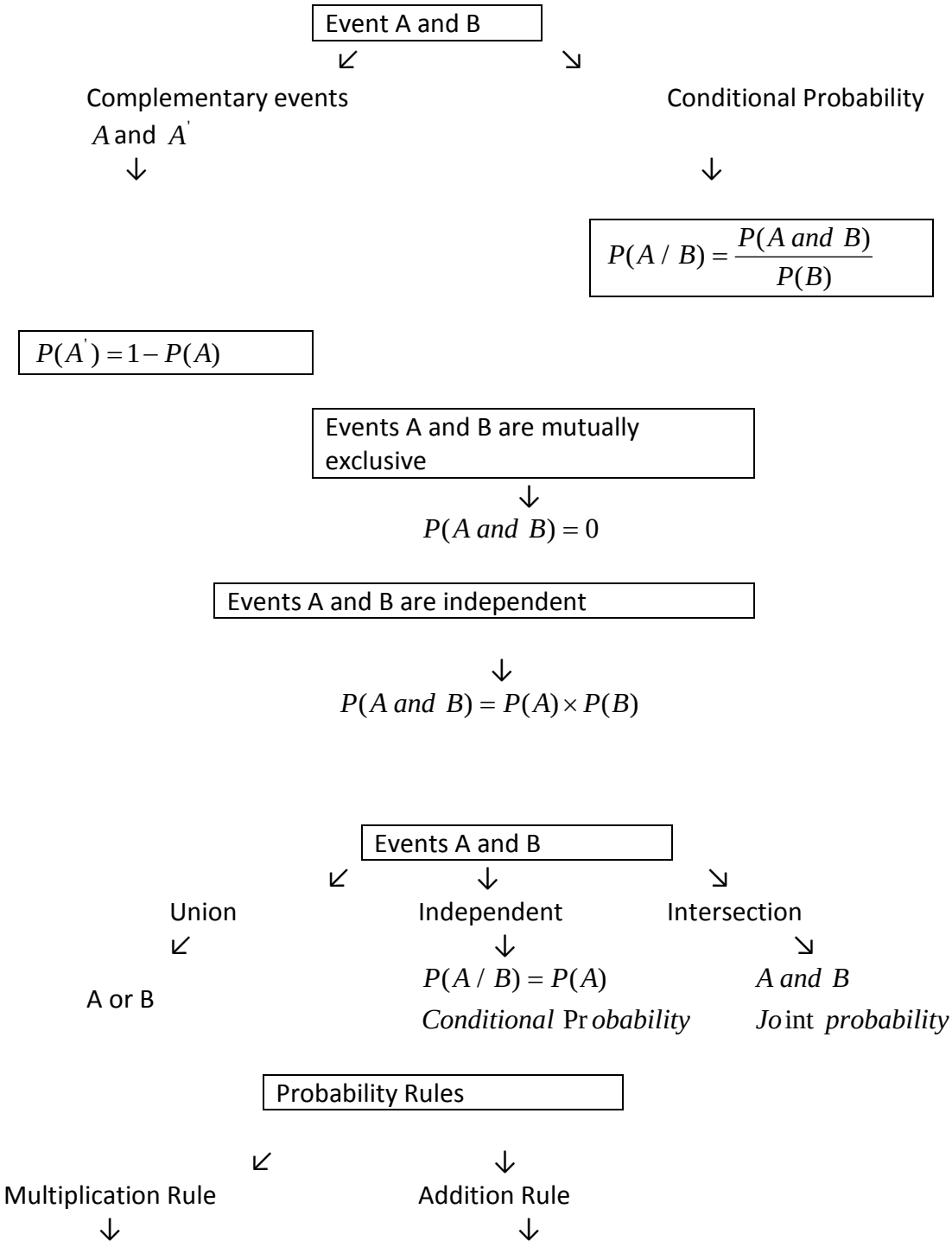
4.1 Introduction to this study unit

This unit introduces the basic concepts of probability. It outlines rules and techniques for assigning probabilities to events. Probability plays a critical role in statistics. All of us form simple probability conclusions in our daily lives. Sometimes these determinations are based on facts, while others are subjective. If the probability of an event is high, one would expect that it would occur rather than it would not occur. If the probability of rain is 95%, it is more likely that it would rain than not rain.

The principles of probability help bridge the words of descriptive statistics and inferential statistics. Studying this unit will help you learn different types of probabilities, how to compute probability, and how to revise probabilities in light of new information. Probability principles are the foundation for the probability distribution, the concept of mathematical expectation, and the Binomial and Poisson distributions, topics that are discussed in study unit 5.

Challenges in understanding statistics usually start from this chapter. There are basic concepts we have to master to understand probability in general. The concepts are summarised in the following mind map and are explained as follows:

Mind-map on the concepts of probability in general.



$P(A\ and\ B) = P(A\ /\ B) \times P(B)$

↓

If A and B are INDEPENDENT then

$P(A\ and\ B) = P(A) \times P(B)$

$P(A\ or\ B) = P(A\ B) + P(B) - P(A\ and\ B)$

↓

If A and B are MUTULLAY EXCLUSIVE then,

$P(A\ or\ B) = P(A) + P(B)$

4.1.1 Definition

The Probability of an event can be defines as follows:

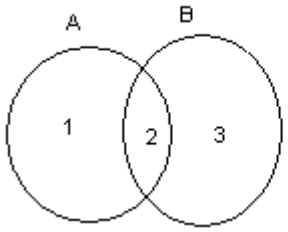
$$\Pr ob = \frac{number\ of\ successes}{number\ of\ possible\ outcomes} \text{ or } \Pr ob = \frac{number\ of\ successes}{samples\ space}$$

4.2 Event

An event is defined as a set of possible outcomes of a variable. A simple event is described by a single character.

Example 7

Consider the following Venn diagram contain set A and B



The simple events from the above diagram include; $\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}$. We can further break down the classification of events as follows;

4.2.1 Joint (Intersection) event

These are simple events common in both sets. In this modules we use the word ''*and*'' to represent joint events. Form the above Venn diagram, the joint event, $A\ and\ B = \{2\}$.

4.2.2 Union (Combination) event

This represents a combination of one or more simple events in a sample space. In this modules we use the word ''*or*'' to represent union events. Form the above Venn diagram , the union of events in set A and B is $A\ or\ B = \{1, 2, 3\}$

4.2.3 Independent events.

These are events in which the occurrence of one does not affect or depend on the occurrence of another. Like in real life, the word independent (common used by ladies) to mean that she is looking after herself, that is, she does not depend on her boy friend or parents. We carry the same meaning when we use the same word in statistics.

However, when it comes to probability, if event A and B are independent, we interpret this it as $P(A\ and\ B) = P(A) \times P(B)$

In other words, the joint probability of two independent events is equal to the product of two marginal probabilities. In the previous statement we introduced a new term "Marginal probability"

Marginal probability

This term is used to indicate the sum of two joint probabilities.

Example 8

Consider the following contingency table;

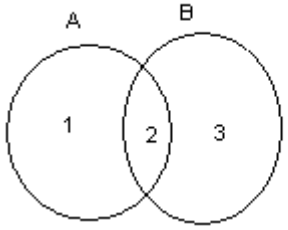
	B	$B^{'}$	Total
A	$P(A\ and\ B)$	$P(A\ and\ B^{'})$	$P(A)$
$A^{'}$	$P(A^{'}\ and\ B)$	$P(A^{'}\ and\ B^{'})$	$P(A^{'})$
Total	$P(B)$	$P(B^{'})$	

The marginal probability of A is $P(A) = P(A\ and\ B) + P(A\ and\ B^{'})$. Likewise the marginal probability of B is $P(B) = P(A\ and\ B) + P(A^{'}\ and\ B)$

4.2.4 Complement event.

A complement of set A is the event that will occur if event set A does not occur. "*It sounds confusing! Not so?*" For example, if it does not shine, it will rain. So the complement of raining is shining and vice versa. Or if a pregnant woman does not give birth to a baby boy, she will give birth to a baby girl. The complement of giving birth to a baby boy is giving birth to a baby girl.

If we use the Venn diagram



The complement of set A, represent as $A^{'} = \{3\}$, and the complement of set B, represented as $B^{'} = \{1\}$.

4.2.5 Mutually exclusive events

The word mutually exclusive is used to indicate that the two event do not have an intersection. For example, gender is mutually exclusive. You are either a male or a female. Now each time it comes up in probability, it will mean that; $P(A\ and\ B) = 0$

4.2.6 Conditional event

A conditional event is the event that will occur ***given that*** another event occurred. For example, sometimes when it rains the number of accidents on the road tends to increase. So the increase in the number of accidents is conditioned on rain. Or when can say that because it rained, there was an increase in the number of accidents on the road.

In probability concepts, a conditional probability is represent as

$$P(A/B) = \frac{P(A\ and\ B)}{P(B)} \text{ or } P(B/A) = \frac{P(A\ and\ B)}{P(A)}$$

The key term to look for here, that is, if you want to know that this is a conditional probability is the term "***given that***".

Please note that if events are independent, that is, $P(A\ and\ B) = P(A) \times P(B)$, then the conditional probability changes to $P(A/B) = P(A)$ or $P(B/A) = P(B)$

4.3 Probability Rules

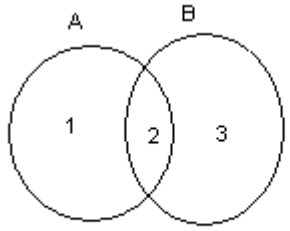
4.3.1 Addition Rule

When two events A and B occur simultaneously, the general addition rule is applied for finding P(A or B) = probability that event A occurs or event B occurs or both occur. In probability concepts, this rule is expressed as follows:

$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$.

This is the sum of two marginal probabilities less the joint probability. It makes sense to subtract off the joint probability because we have to remove an overlap.

For example: Consider the Venn diagram below



A or B = {1, 2, 3} = A={1, 2} + B={2, 3} - A and B = {2}

Please note that is events are mutually exclusive, that is, $P(A\ and\ B) = 0$, the addition rule changes to $P(A\ or\ B) = P(A) + P(B)$

4.3.2 Multiplication Rule

The multiplication rule defines the probability that events A and B both occur. In probability concepts, the multiplication rule is expressed as;

$P(A\ and\ B) = P(A/B) \times P(B)$

Or

$P(A\ and\ B) = P(B/A) \times P(A)$

This rule can be derived from the conditional probability by cross multiplying. We already know that the conditional probability that of A given B is $P(A/B) = \dfrac{P(A\ and\ B)}{P(B)}$.

Now cross multiplying this expression and making the joint probability the subject of the formula yields $P(A\ and\ B) = P(A/B) \times P(B)$

4.3.3 Complement rule

We have already defined a complement event in section 4.2. Now, according to this rule

$P(A') = 1 - P(A)$

Likewise

$P(B') = 1 - P(B)$

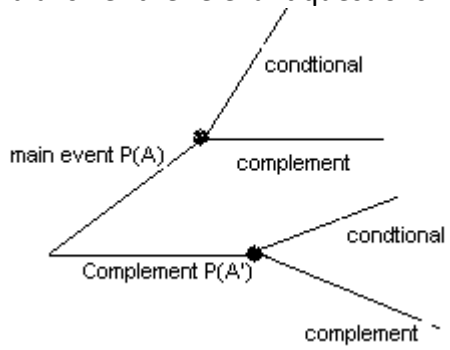
4.4 Probability Tree diagrams

A probability tree is built on two concepts, namely;

(i) Complement rule

(ii) Conditional probability

Once you must how to play around with these concepts, you can build all probabilities trees and answer the relevant questions. In summary, the probability tree is built as follows:



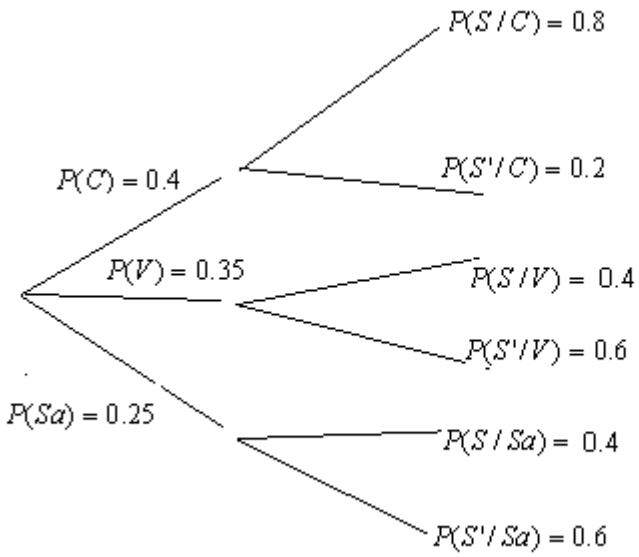
Let's take an example demonstrating how this works.

Example 9

A sidewalk ice-cream seller sells three flavours: chocolate, vanilla and strawberry. Of his sales 40% is chocolate, 35% vanilla and 25% strawberry. Sales are by cone or cup. The percentages of cone sales for chocolate, vanilla and strawberry are 80%, 60% and 40% respectively. Use a tree diagram to determine the relevant probabilities of a randomly selected sale of one ice cream. Which one of the following statements is *incorrect?*

1) $P(strawberry) = 0.25$
2) $P(vanilla\ in\ a\ cup) = 0.14$
3) $P(chocolate\ in\ a\ cone) = 0.32$
4) $P(chocolate\ or\ vanilla) = 0.75$
5) $P(vanilla\ /\ in\ a\ cone) = 0.3889$

Let Sa = event strawberry flavour
C = chocolate flavour
V =vanilla flavour
S = percentage of cone sales



Option (1) Correct

$P(strawberry) = 0.25$

Option (2) Correct

$P(vanilla\ in\ a\ cup) = P(V) \times P(S'/V) = 0.35 \times 0.4 = 0.14$

Option (3) Correct

$P(chocolate\ in\ a\ cone) = P(C) \times P(S/C) = 0.4 \times 0.8 = 0.32$

Option (4) Correct

6) $P(chocolate\ or\ vanilla) = P(C) + P(V) = 0.4 + 0.35 = 0.75$. Remember that these are mutually exclusive events.

Option (5) Incorrect

$$P(vanilla\ /\ in\ a\ cone) = \frac{P(vanilla\ in\ a\ cone)}{P(cone)}$$

$$= \frac{P(V) \times P(S/V)}{P(C)P(S/C) + P(V)P(S/V) + P(Sa)P(S/sa)}$$

$$= \frac{0.35 \times 0.6}{(0.4 \times 0.8) + (0.35 \times 0.6) + (0.25 \times 0.4)}$$

$$= \frac{0.21}{0.63}$$

$$= 0.333$$

SELF ASSESSMENT EXERCISE – TEST YOUR KNOWLEDGE

Question 1

Which statement is *correct*?

(1) Probability takes on a value from 0 to 1

(2) Probability refers to a number which express the chance that an event will occur.

(3) Probability is zero if the event A of interest is impossible.

(4) The sample space refers to all possible outcomes of an experiment

(5) All the above statements are correct.

Question 2

Assume that X and Y are two independent events with $P(X) = 0.5$ and $P(Y) = 0.25$.

Which of the following statements is *incorrect*?

(1) $P(X') = 0.75$

(2) $P(X \text{ and } Y) = 0.125$

(3) $P(X \text{ or } Y) = 0.625$

(4) X and Y are not mutually exclusive

(5) $P(X/Y) = 0.5$

Question 3

Refer to the following contingency table:

Event	C_1	C_2	C_3	C_4	Total
D_1	75	125	65	35	300
D_2	90	105	60	45	300
D_3	135	120	75	70	400
Total	300	325	200	150	1000

Which one of the following statements is *incorrect*?

(1) $P(C_1 \text{ and } D_1) = 0.075$

(2) $P(D_1) = 0.3$

(3) $P(C_1 \text{ or } D_1) = 0.6$

(4) $P(D_3/C_4) = 0.4667$

(5) $P(C_4/D_3) = 0.175$

Question 4

Numbers $1, 2, 3, 4, 5, 6, 7, 8, 9$ are written on separate cards. The cards are shuffled and the top one turned over. Let $A = \text{an even number}$, $B = a \text{ number greater than } 6$.

Which one of the following statements is *incorrect*?

(1) The sample space is $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

(2) $P(A) = \frac{4}{9}$

(3) $P(B) = \frac{1}{9}$

(4) $P(A \text{ and } B) = \frac{1}{9}$

(5) $P(A \text{ or } B) = \frac{7}{9}$

Question 5

If A and B are independent events with $P(A) = 0.25$ and $P(B) = 0.60$, then $P(A/B)$ is equal to

(1) 0.25
(2) 0.60
(3) 0.35
(4) 0.85
(5) 0.15

Question 6

Given that $P(A) = 0.7$, $P(B) = 0.60$ and $P(A \text{ and } B) = 0.35$, which one of the following statements is *incorrect*?

(1) $P(B') = 0.40$ P (B') = 0.4
(2) A and B are not mutually exclusive
(3) A and B are dependent
(4) $P(B/A) = 0.60$
(5) $P(A \text{ or } B) = 0.95$

Question 7

The Burger Queen Company has 4755 locations along the west coast. The general manager is concerned with the profitability of the locations compared with major menu items sold. The information below shows the number of each menu item selected by profitability of store.

	Baby Burger M_1	Mother Burger M_2	Father Burger M_3	Nachos M_4	Tacos M_5	Total
High profit R_1	250	424	669	342	284	1969
Medium Profit R_2	312	369	428	271	200	1580
Low profit R_3	289	242	216	221	238	1206
Total	851	1035	1313	834	722	4755

If a menu order is selected at random, which statement is *incorrect*?

(1) $P(M_5) = 0.1518$
(2) $P(R_3) = 0.0501$
(3) $P(R_2 \text{ and } M_3) = 0.0900$
(4) $P(M_2 / R_2) = 0.2335$
(5) $P(R_1 / M_4) = 0.4101$

Question 8

In a particular country, airport A handles 50% of all airline traffic, and airports B and C handle 30% and 20% respectively. The detection rates for weapons at the three airports are 0.9, 0.5 and 0.4 respectively.

If a passenger at one of the airports is found to be carrying a weapon through the boarding gate, what is the probability that the passenger is using airport C?

(1) 0.2206
(2) 0.6618
(3) 0.5000
(4) 0.2941
(5) 01176

SOLUTIONS TO SELF ASSESSMENT EXERCISES

Question 1

Alternative 5.
All options are correct.

Question 2

Option (1) Incorrect

$$P(X) = 1 - P(X)$$
$$= 1 - 0.5$$
$$= 0.5$$

Option (2) Correct
Since event X and Y are independent,

$$P(X \text{ and } Y) = P(X) \times P(Y)$$
$$= 0.5 \times 0.25$$
$$= 0.125$$

Option (3) Correct
Using the addition rule

$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)$$
$$= 0.5 + 0.25 - 0.125$$

Option (4) Correct
Remember that mutually exclusive events are defined by $P(X \text{ and } Y) = 0$. Since $P(X \text{ and } Y) \neq 0$, event X and Y are not mutually exclusive.

Option (5) Correct
Using the conditional probability

$$P(X/Y) = \frac{P(X \text{ and } Y)}{P(Y)}$$
$$= \frac{0.125}{0.25}$$
$$= 0.5$$

Question 3

Option (1) Correct

$$P(C_1 \text{ and } D_1) = \frac{75}{1000}$$
$$= 0.075$$

Option (2) Correct

$$P(D_1) = \frac{300}{1000}$$
$$= 0.3$$

Option (3) Incorrect

$$P(C_1 \text{ or } D_1) = P(C_1) + P(D_1) - P(C_1 \text{ and } D_1)$$
$$= \frac{300}{1000} + \frac{300}{1000} - \frac{75}{1000}$$
$$= 0.30 + 0.30 - 0.075$$
$$= 0.525$$

Option (4) Correct

$$P(D_3 / C_4) = \frac{P(D_3 \text{ and } C_4)}{P(C_4)}$$
$$= \frac{70/100}{150/100}$$
$$= \frac{70}{150}$$
$$= 0.4667$$

Option (5) Correct

$$P(C_4 / D_3) = \frac{P(D_3 \text{ and } C_4)}{P(D_3)}$$
$$= \frac{70/100}{400/100}$$
$$= \frac{70}{400}$$
$$= 0.175$$

Question 4

In this question we first define the events with respective probabilities as follows;

$A = \{2, 4, 6, 8\}$, $B = \{7, 8, 9\}$, $A \text{ and } B = \{8\}$. This means that $P(A) = \frac{4}{9}$, $P(B) = \frac{3}{9}$ and $P(A \text{ and } B) = \frac{1}{9}$.

So the $P(A \text{ or } B) = \frac{4}{9} + \frac{3}{9} - \frac{1}{9} = \frac{2}{3}$

Question 5

$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A) = 0.25$$

Alternative 1

Question 6

Option (1) Correct

$$P(B') = 1 - P(B)$$
$$= 1 - 0.6$$
$$= 0.4$$

Option (2) Correct

Remember that mutually exclusive events implies that $P(A \text{ and } B) = 0$. Since $P(X \text{ and } Y) \neq 0$, event X and Y are not mutually exclusive.

Option (3) Correct

Recall that event A and B are independent only and only if $P(A \text{ and } B) = P(A) \times P(B)$. Since

$$P(A \text{ and } B) \neq P(A) \times P(B)$$
$$0.35 \neq 0.6 \times 0.7$$

We can say that event A and B are not independent. Hence they are assumed to be dependent.

Option (4) Incorrect

$$P(B/A) = \frac{P(B \text{ and } A)}{P(A)}$$
$$= \frac{0.35}{0.70}$$
$$= 0.5$$

Option (5) Correct

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.95$$

Question 7

Option (1) Correct

$$P(M_5) = \frac{722}{4755}$$
$$= 0.1518$$

Option (2) Incorrect

$$P(R_3) = \frac{1206}{4755}$$
$$= 0.2536$$

Option (3) Correct

$$P(R_2 \text{ and } M_3) = \frac{428}{4755}$$
$$= 0.0900$$

Option (4) Correct

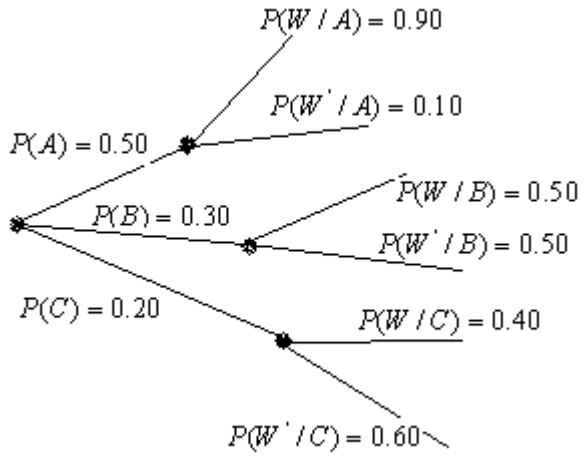
$$P(M_2 / R_2) = \frac{P(M_2 \text{ and } R_2)}{P(R_2)}$$
$$= \frac{369/4755}{1580/4755}$$
$$= \frac{369}{1580}$$
$$= 0.2335$$

Option (5) Correct

$$P(R_1 / M_4) = \frac{P(R_1 \text{ and } M_4)}{P(M_4)}$$
$$= \frac{342/4755}{834/4755}$$
$$= \frac{342}{834}$$
$$= 0.4101$$

Question 8

Let W = event person is carrying a weapon. We first construct a probability tree to represent the situation. This is done as follows:



Now we know that what is required is $P(C/W) = \frac{P(C \text{ and } W)}{P(W)}$. Please remember that, from the multiplication rule $P(C \text{ and } W) = P(W/C) \times P(C)$, and the $P(W) = P(W \text{ and } A) + P(W \text{ and } B) + P(W \text{ and } C)$

Hence.

$$P(C/W) = \frac{P(C \text{ and } W)}{P(W)}$$

$$= \frac{P(C \text{ and } W)}{P(W \text{ and } A) + P(W \text{ and } B) + P(W \text{ and } C)}$$

$$= \frac{P(W/C) \times P(C)}{P(W/A) \times P(A) + P(W/B) \times P(B) + P(W/C) \times P(C)}$$

$$= \frac{0.2 \times 0.4}{(0.2 \times 0.4) + (0.3 \times 0.5) + (0.2 \times 0.4)}$$

$$= \frac{0.08}{0.68}$$

$$= 0.1176$$

Alternative (5).

5 DISCRETE PROBABILITY DISTRIBUTION

STUDY UNIT 5

Key concepts in this unit are:

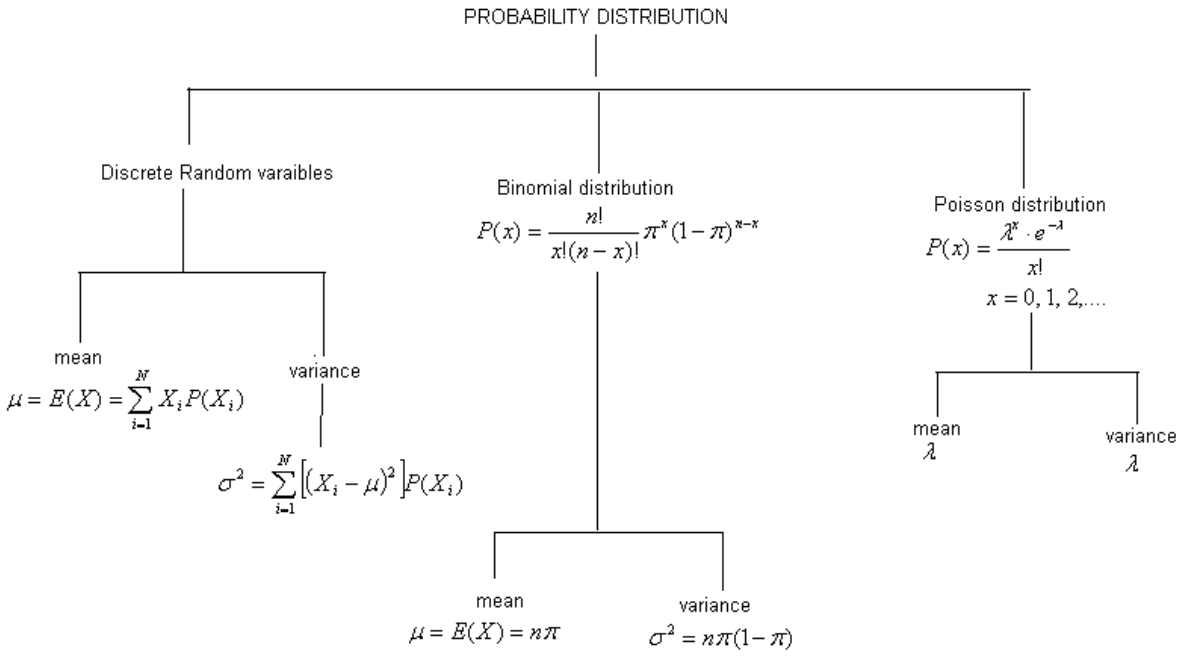
Define a discrete probability distribution.
How would you construct a probability distribution for a discrete random variable?
Distinguish between discrete and continuous random variables.
How would you compute the expected value and the variance of a discrete random variable?
How would you compute the expected value and the variance of a Binomial distribution?
Using the Binomial formula and the tables in general.
The concept of the Poisson distribution in general
How would you compute the expected value and the variance of a Poisson distribution?

5.1 Introduction

In study unit 4 learnt much about probability in general. In this study unit we discuss discrete random variables and their probability distributions. Probability distributions are classified as either discrete or continuous, depending on the random variable. Please revisit study unit 1 to remember the difference between discrete and continuous variables.

A random variable is a variable that can take on different values according to the outcome of an experiment . It is described as *random* because we don't know ahead of time exactly what value it will have following the experiment.

For example, when we toss a coin, we don't know for sure whether it will land heads or tails. Likewise, when we measure the diameter of a roller bearing, we don't know in advance what the exact measurement will be.
In this study unit the emphasis is on discrete random variables and their probability distributions. In the next unit we will cover random variables of continuous type.
The mind map to this study unit is as follows:



5.1.1 Definition

A **probability function**, denoted $p(x)$, specifies the probability that a random variable is equal to a specific value. More formally, $p(x)$ is the probability that the random variable X takes on the value x, or $p(x) = P(X = x)$.

5.1.2 Properties of probability density functions.

The two key properties of a probability function are:

- For any value of x, $0 \le p(x) \le 1$.
- $\sum p(x) = 1$, the sum of the probabilities for all possible outcomes, x, for a random variable, X, equals one.

5.2 Probability distribution for discrete random variables

The probability distribution for discrete random variable as a mutually exclusive list of all possible numerical outcomes along with the probability of occurrence of each outcome. That is, if X is a discrete random variable associated with a particular chance experiment, a list of all possible values X together with their associated probabilities is called a discrete probability distribution. The total probability of all outcomes is 1.

5.2.1 Expected value of a Discrete Random Variable

The mean of a discrete probability distribution for a discrete random variable is called expected value, repersented as $E(x), or\ \mu$. It is calculated as the sum of the product of the random variable X by its corresponding probability, $P(x)$, as follows

$$\mu = E(x) = \sum_{i=1}^{N} X_i P(x_i)$$

Where

X_i = the i[th] outcome of the discrete random variable X

$P(x_i)$ = the probability of occurrence of the i[th] outcome of X

Example 10

Based on her experience, a professor knows that the probability distribution for X = number of students who come to her office on Wednesdays is given below.

x	0	1	2	3	4
$P(X = x)$	0.01	0.20	0.50	0.15	0.05

What is the expected number of students who visit her on Wednesdays?

(1) 0.50
(2) 0.70
(3) 1.85
(4) 0.90
(5) 0.30

Solution: The expected number (the mean) is calculated as the sum of the product of the random variable X by its corresponding probability, $P(X)$, as follows:

$$\mu = E(x) = \sum_{i=1}^{N} X_i P(x_i)$$
$$= (0 \times 0.01) + (1 \times 0.20) + (2 \times 0.5) + (3 \times 0.15) + (4 \times 0.05)$$
$$= 1.85$$

Alternative 3

5.2.2 Variance of a discrete random variable

The variance of a probability distribution is computed by multiplying each possible squared difference $\left[(X_i - \mu)^2\right]$ by its corresponding probability, $P(x_i)$, and then summing the resulting products as follows:

$$\sigma^2 = \sum_{i=1}^{N} \left[(X_i - \mu)^2\right] P(x_i)$$

Where

X_i = the i[th] outcome of the discrete random variable X

$P(x_i)$ = the probability of occurrence of the i[th] outcome of X

Please note that we have to compute the mean first before we think of calculating the variance of a discrete random variable.

5.2.3 Standard deviation of a discrete random variable

The standard deviation is the positive square root of the variance of a discrete random variable

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} (X_i - \mu)^2 P(x_i)}$$

Example 11

Let the probability distribution for X = number of jobs held during the past year for students at a college be as follows:

x	1	2	3	4	5
$P(X = x)$	0.25	0.33	0.17	0.15	0.10

The standard deviation of the number of jobs held is

(1) 8.000

(2) 1.3682

(3) 2.5200

(4) 1.2844

(5) 1.6496

Solution:

We first calculate the mean

$$\mu = E(X) = \sum_{i=1}^{N} X_i P(x_i)$$
$$= (1 \times 0.25) + (2 \times 0.33) + (3 \times 0.17) + (4 \times 0.15) + (5 \times 0.10)$$
$$= 2.52$$

Then we use the mean to calculate the variance

$$\sigma^2 = \sum_{i=1}^{N} \left[(X_i - \mu)^2 \right] P(x_i)$$
$$= (1 - 2.52)^2 \times 0.25 + (2 - 2.52)^2 \times 0.33 + (3 - 2.52)^2 \times 0.17 + (4 - 2.52)^2 \times 0.15 + (5 - 2.52)^2 \times 0.10$$
$$= 1.6496$$

The standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.6496} = 1.2844$$

Alternative 4

5.2.4 The language

Please note that though this is not part of Statistics, sometimes the language used in this section tend to confuse students, especially if you are not a mathematics student or you did not take mathematics prior to this module.
The key terms usually used and their interpretations are;

(i) Exactly: This is used to indicate equals to $(=)$. For example the probability of obtaining exactly two is interpreted as $P(X = 2) = ..$

(ii) At least: This is used to indicate greater than or equal to . For example the probability of obtaining at least two is interpreted as $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 3) + ...$

(iii) At most: This is used to indicate less than or equal to. For example the probability of obtaining at most two is interpreted as $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

Exercise

Question 1

The number of telephone calls coming into a switchboard and their respective probabilities for a 3-minute interval are as follows:

x	0	1	2	3	4	5
$P(X = x)$	0.60	0.20	0.10	0.04	0.03	0.03

How many calls might be expected over a 3-minute interval?

(1) 0.04
(2) 3
(3) 0.2
(4) 0.79
(5) 3.75

Question 2

The probability distribution of a discrete random variable x is shown below.

x	0	1	2	3
$P(X = x)$	0.25	0.40	0.20	0.15

Find the *incorrect* statement:

(1) This is an example of a discrete probability distribution.
(2) The expected value of x is 1.25
(3) The variance of x is 2.55
(4) If $x = 0$, after multiplication by $P(x)$, the answer 0, which means that the probability associated with the value $x = 0$ has no influence on the answers of the mean and the variance.
(5) The standard deviation of x is 0.9937

Question 3

Use the data set given in question 2 and find the incorrect statement.

(1) $P(x > 1) = 0.35$

(2) $P(x \le 2) = 0.65$
(3) $P(1 < x \le 2) = 0.20$
(4) $P(0 < x < 1) = 0.00$
(5) $P(1 \le x < 3) = 0.60$

Solutions

Question 1
Recall, the expected number is also the mean of a discrete random variable, calculate as:

$$\mu = E(X) = \sum_{i=1}^{N} X_i P(X_i)$$
$$= (0 \times 0.60) + (1 \times 0.20) + (2 \times 0.10) + (3 \times 0.04) + (4 \times 0.03) + (5 \times 0.03)$$
$$= 0.79$$

Alternative 4
Question 2

1. Correct. The variable takes on discrete values, therefore the statement is correct. Remember in section 5.2 of this unit we defined the probability distribution for discrete random variable as a mutually exclusive listing of all possible numerical outcomes along with the probability of occurrence of each outcome which is exactly the case in this option.
2. Correct.
$$\mu = E(X) = \sum_{i=1}^{N} X_i P(X_i)$$
$$= (0 \times 0.25) + (1 \times 0.40) + (2 \times 0.20) + (3 \times 0.15)$$
$$= 1.25$$
3. Incorrect. This figure was incorrect computed. It should be
$$\sigma^2 = \sum_{i=1}^{N} \left[(X_i - \mu)^2 \right] P(X_i)$$
$$= (0 - 1.25)^2 \times 0.25 + (1 - 1.25)^2 \times 0.40 + (2 - 1.25)^2 \times 0.20 + (3 - 1.25)^2 \times 0.15$$
$$= 0.9875$$

4. Correct. You can see it if you study the calculation of the mean and the variance**.**

5. Correct. $\sigma = \sqrt{\sigma^2} = \sqrt{0.9875} = 0.9937$

Question 3

1. Correct. We add from two (greater than one) up to three as follows;
$$P(x > 1) = P(x = 2) + P(x = 3) = 0.20 + 0.15 = 0.35$$
2. Incorrect. Here we take values from zero to two. One could also consider this question as at most two as discussed in study unit 4.
$$P(x \le 2) = P(x = 0) + P(x = 1) + P(x = 2) = 0.25 + 0.40 + 0.20 = 0.85$$
3. Correct. In this case one is not included but two is. $P(1 < x \le 2) = P(x = 2) = 0.20$
4. Correct. $P(1 < x < 1) = 0.00$ because between 0 and 1 there is no discrete value for x.

5. Correct. Here one is included but three is not.
   $P(1 \le x < 3) = P(x = 1) + P(x = 2) = 0.40 + 0.20 = 0.60$

Having understood discrete random variable, we can now discuss their probability distributions. This is very small but important section in Statistics.
There are quite a number of discrete probability distributions. However, in this module we only study two of the many, namely;

- the Binomial distribution and
- the Poission distribution.

5.3 The Binomial Distribution

The binomial distribution describes the probability distribution resulting from the outcome of a binomial experiment. A binomial experiment usually involves several repetitions (trials) of the basic experiment. The binomial probability distribution gives us the probability that a success will occur x time in n trials, for $x = 0, 1, 2,.....n$.

5.3.1 Characteristic of a Binomial experiment

- The experiment must consist of n identical trials.
- Each trial has 1 of 2 possible mutually exclusive outcomes: success or failure (success refers to the occurrence of the event of interest).
- The probability (π) that the trial results in a success remains the same from trial to trial.
- The trials are independent of each other (the outcome of a trial does not affect the outcome of any other trial).

5.3.2 The Binomial formula

The probability distribution of number of successes x of the random variable X in n trials of a binomial experiment is:

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

$n =$ number of trials or sample size
$\pi =$ probability of success on each trial
$x =$ the binomial variance 0, 1, 2,....e.t.c.

Please note that, the mathematical sign(!) is called the factorial sign of a positive integer n. It is interpreted as the product of all positive integers less than or equal to n. For example $5! = 5 \times 4 \times 3 \times 2 \times 1$, $4! = 4 \times 3 \times 2 \times 1$ and $0! = 1$ *"Interesting! Not so?"*

5.3.3 The use of table

Instead of using the formula, students are advised to use Binomial table. We shall learn how this is done when we come to examples.

5.3.4 The mean of the binomial distribution

The mean , μ, of the binomial distribution is equal to the sample size, n, multiplied by the probability of an event of interest π.

$$\mu = E(X) = n\pi$$

5.3.5 The variance of the binomial distribution

$$\sigma^2 = n\pi(1-\pi)$$

5.3.6 The standard deviation of the binomial distribution

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} = \sqrt{n\pi(1-\pi)}$$

Example 12

A textile firm has found from experience that only 20% of the people applying for certain stitching-machine job are qualified for the work. If 5 people are interviewed, what is the probability of finding at least three qualified persons?

$n = 5, \quad \pi = 0.20, \quad P(x \geq 3)?$

Please do not forget that at least three means add from three, four and so on. Using the formula:

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

We then have

$$P(x \geq 3) = P(x=3) + P(x=4) + P(x=5)$$

$$= \frac{5!}{3!(5-3)!}0.20^3(1-0.20)^{5-3} + \frac{5!}{4!(5-4)!}0.20^4(1-0.20)^{5-4} + \frac{5!}{5!(5-5)!}0.2^5(1-0.20)^{5-5}$$

$$= \frac{120}{6\times 2}\times 0.008\times 0.064 + \frac{120}{24\times 1}\times 0.0016\times 0.8 + \frac{120}{120\times 1}\times 0.00032\times 1$$

$$= 0.0512 + 0.0064 + 0.0003$$

$$= 0.0579$$

Please note that the same result can be obtained if we use Binomial tables.

Exercise:

Question 1

A new car salesperson knows that he sells cars to one customer out of 10 who enters the showroom. The probability that he will sell a car to exactly two of the next three customers is

(1) 0.027
(2) 0.973
(3) 0.000
(4) 0.090
(5) 0.901

Question 2

Use the information given in question 1. Let X be number of cars the salesperson sells to the next three
customers. Which one of the following statements is *incorrect*?

(1) X has a binomial distribution

(2) The expected number of cars sold if n = 3 is 0.3
(3) The variance of this distribution is 0.27
(4) P(X ≤ 1) = 0.9720
(5) P(X >2) = 0.0280

Question 3
Suppose that 62% of new cars sold in a country are made by one small car manufacturer. A random sample of 7 purchases of new cars is selected. The probability that 4 of those selected purchases are made by this car manufacturer is

(1) 0.5800
(2) 0.5714
(3) 0.2838
(4) 0.4200
(5) 0.7162

Solutions
Question 1

$$n = 3, \quad \pi = \frac{1}{10} = 0.1, \quad P(x = 2)?$$

$$P(x = 2) = \frac{3!}{2!(3-2)!} 0.1^2 (1-0.1)^{3-2}$$
$$= 0.027$$

Question 2

1 Correct.
2 Correct. $E(x) = n\pi = 3 \times 0.1$
3 Correct. $\sigma^2 = n\pi(1-\pi) = 3 \times 0.1(1-0.1) = 0.27$
4 Correct.

$$P(x \le 1) = P(x = 0) + P(x = 1)$$
$$= \frac{3!}{0!(3-0)!} 0.1^3 (1-0.1)^{3-0} + \frac{3!}{1!(3-1)!} 0.1^1 (1-0.1)^{3-1}$$
$$= 0.7290 + 0.2430$$
$$= 0.9720$$

5 Incorrect

$$P(x > 2) = P(x = 3)$$
$$= \frac{3!}{3!(3-3)!} 0.1^3 (1-0.1)^{3-3}$$
$$= 0.001$$

Question 3

$$n = 7, \quad \pi = 0.62, \quad P(x = 4)?$$

$$P(x=4)=\frac{7!}{4!(7-4)!}0.62^{4}(1-0.62)^{7-4}$$
$$=0.2838$$

Alternative 3

5.4 Poisson Distribution

5.4.1 Introduction

The Poisson distribution is a discrete distribution for which the probability of occurrence over the given span of time, space, or distance is extremely small. There is no specific upper limit to the count (n is unknown), although a finite count is expected. The Poisson distribution tends to describe the phenomena like:

- Customers arrival at a service point during a given period of time, such as the number of motorist approaching a toll booth, the number of hungry persons entering a McDonald's restaurant, or the number of calls received by a company call center. In this context it is also useful in a management science technique called queuing (waiting-line) theory.
- Defects in manufacturer materials, such as the number of flaws in wire or pipe products over a given number of feet, or the number of knots in wooden panels for a given area.
- The number of work-related deaths, accidents, divorces, suicides, and homicides over a given period of time.

Although it is closely related to the Binomial distribution, the Poisson distribution has a number of characteristics that makes it unique.

5.4.2 Characteristics of Poisson distribution

- The number of successes that occur in a specified interval is independent of the number of occurrence in any other interval.
- The probability that success will occur in an interval is the same for all intervals of equal size, and is proportional to the size of the interval.
- x is the count of the number of successes that occur in a given interval, and may take on any value from 0 to infinity.

5.4.3 The formula

If X is a Poisson random variable, the probability distribution of the number of successes of X is

$$P(x)=\frac{\lambda^{x}\cdot e^{-\lambda}}{x!}$$
$$x=0,1,2,....$$

$\lambda=$ the average number of successes occurring in the given time or measurement.

$e=2.71828$ (the base of natural logarithms)

5.4.4 Use of tables

Instead of using the formula, students are advised to use Binomial table. We shall study how this is done when we come to examples

Example 13

The average number of a certain radio sold per day by a firm is approximately Poisson, with mean of 1.5. The probability that the firm will sell at least two radios over a three-day period is equal to

(1) 0.5578
(2) 0.1255
(3) 0.9344
(4) 0.0447
(5) 0.4422

Solution:

Recall that this distribution has no upper bound. Therefore we have to express at least in another equivalent way such as

$$\begin{aligned} P(x \geq 2) &= 1 - P(x \leq 1) \\ &= 1 - \{P(x=0) + P(x=1)\} \\ &= 1 - \left\{ \frac{1.5^0 \cdot e^{-1.5}}{0!} + \frac{1.5^1 \cdot e^{-1.5}}{1!} \right\} \\ &= 1 - \{0.2231 + 0.3347\} \\ &= 0.4422 \end{aligned}$$

Alternative 5

Example 14

A bank receives on average 6 bad cheques per day. The probability that it will receive exactly 4 bad cheques on a given day is

(1) 0.0892
(2) 0.1393
(3) 0.2851
(4) 0.1339
(5) 0.6667

Solution

Given that $\lambda = 6, \;\; P(x=4)?$

$$P(x=4) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{6^4 \cdot e^{-6}}{4!} = 0.1339$$

Alternative 4

SELF ASSESSMENT EXERCISE – TEST YOUR KNOWLEDGE

In this section we have selected questions based on whole study unit. Please attempt them before referring to the solutions

Question 1

Bank robbers brandish firearms to threaten their victims in 80 percent of the incidents. An announcement that six bank robberies are taking place is being broadcast. The probability that a firearm is being used in at least one of the robberies is

(1) 0.0015
(2) 0.7379
(3) 0.0001
(4) 0.9999
(5) 0.0016

Question 2

In an urban country, health official anticipate that the number of births this year will be the same as last year, when 438 children were born – an average of 438/356, or 1.2 births per day. Daily births have been distributed according to a Poisson distribution.

The distribution can be represented as

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062	0.0012	0.0002

What is the probability that at least two births will occur on a given day?

(1) 0.3374
(2) 0.8795
(3) 0.3795
(4) 0.7831
(5) 0.6626

Question 3

Given the following probability distribution for an infinite population with the discrete random variables, x

x	0	1	2	3
$P(x)$	0.2	0.1	0.3	0.4

Which statement is *incorrect*?

(1) The mean of x is 1.9
(2) The probability that x is at most one equals to 0.3
(3) The variance of x is 1.29
(4) The standard deviation of x is 1.14
(5) The probability that x is at least zero equals to 0.2

Question 4
A drug is known to be 80% effective in curing a certain disease. If four people with the disease are to be given the drug, the probability that more than two are cured is:
(1) 0.8464
(2) 0.1536
(3) 0.5000
(4) 0.1808
(5) 0.8192

Question 5
Refer to question 4, the expected value of people cured is
(1) 0.80
(2) 0.20
(3) 3.20
(4) 0.64
(5) 1.00

Question 6
Given a Poisson random variable X, where the average number of successes occurring in a specified interval is 1.8, P(X=0) is equal to
(1) 0.1653
(2) 0.2975
(3) 1.0000
(4) 0.0000
(5) 0.4762

Solutions to the general exercises

Question 1

$$\begin{aligned} P(x \geq 1) &= 1 - P(x \leq 0) \\ &= 1 - \{P(x = 0)\} \\ &= 1 - \frac{6!}{0!(6-0)!} 0.80^{0} (1 - 0.80)^{6-0} \\ &= 1 - 000064 \\ &= 0.9999 \end{aligned}$$

Alternative 4

Question 2

$$P(X \geq 2) = 1 - P(X \leq 1)$$
$$= 1 - \{P(X = 0) + P(X = 1)\}$$
$$= 1 - \{0.3012 + 0.3614\}$$
$$= 0.3374$$

Alternative 1

Question 3

1. Correct.

$$\mu = E(X) = \sum_{i=1}^{N} X_i P(X_i)$$
$$= (0 \times 0.20) + (1 \times 0.10) + (2 \times 0.30) + (3 \times 0.4)$$
$$= 1.9$$

2. Correct.

$$P(x \leq 1) = P(x = 0) + P(x = 1) = 0.2 + 0.1 = 0.3$$

3. Correct.

$$\sigma^2 = \sum_{i=1}^{N} \left[(X_i - \mu)^2 \right] P(X_i)$$
$$= (0 - 1.9)^2 \times 0.2 + (1 - 1.9)^2 \times 0.1 + (2 - 1.9)^2 \times 0.3 + (3 - 1.9)^2 \times 0.4$$
$$= 1.29$$

4 Correct

$$\sigma = \sqrt{1.29} = 1.14$$

5. Incorrect

$$P(x \geq 1) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) = 1.0$$

Question 4

$$P(x > 2) = P(x = 3) + P(x = 4)$$
$$= \frac{4!}{3!(4-3)!} 0.8^3 (1 - 0.8)^{4-3} + \frac{4!}{4!(4-4)!} 0.8^4 (1 - 0.8)^{4-4}$$
$$= 0.4096 + 0.4096$$
$$= 0.8192$$

Alternative 5

Question 5

$$E(X) = n\pi = 4 \times 0.80 = 3.2$$

Alternative 3

Question 6

Given that $\lambda = 1.8,\ P(x = 0)$?

$$P(x = 0) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{1.8^0 \cdot e^{-1.8}}{0!} = 0.1653$$

6 THE NORMAL DISTRIBUTION

STUDY UNIT 6

Key questions for this unit

How would you compute probabilities from the normal distribution?

Can you distinguish between discrete and continuous probability distributions?

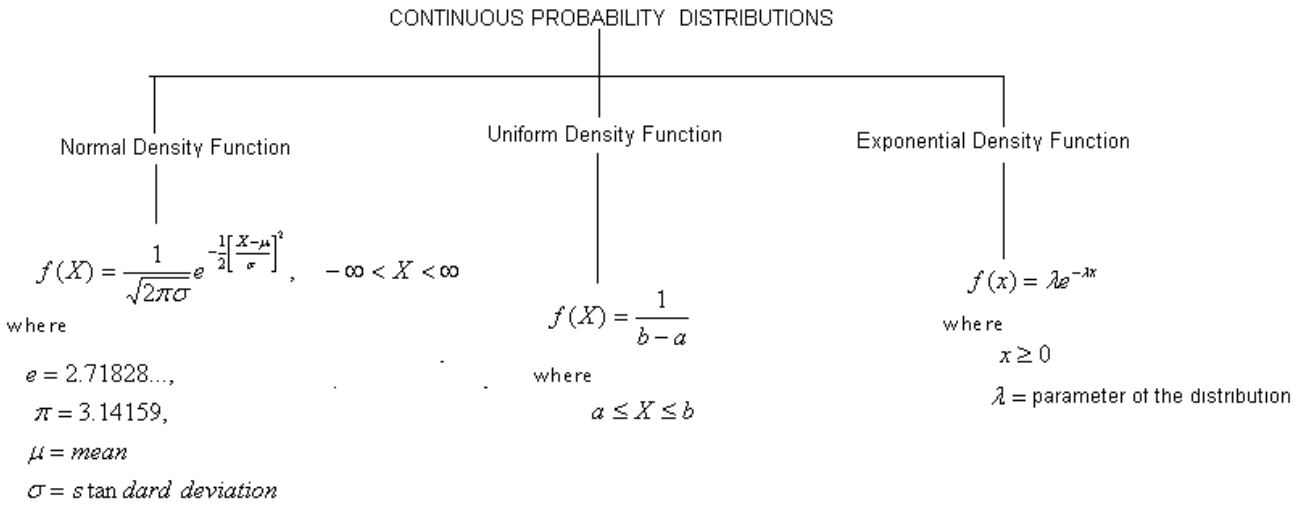
Can you use the normal table to compute probabilities?

Can you determine the Z-variable given the area under the normal curve?

Can you distinguish between the normal, the uniform and the exponential distribution?

6.1 Introduction to this study unit

This chapter describes three continuous distributions, namely the normal, uniform distribution and the exponential distributions. The mind map to this study unit is as follows:



For purposes of this module we shall only concentrate on the normal distributions.

6.2 The normal distribution

The normal is the most important distribution in statistics and the key reasons for this include:

- Numerous continuous variables common in business have distributions that closely resemble the normal distribution.
- The normal distribution can be used to approximate various discrete probability distributions.
- The normal distribution provides the basis for *classical statistical inference* because of its relationship to the *Central Limit Theorem*

You must make sure that you know the characteristics of the normal distribution and how to use the normal table to determine probabilities. For this module it is *not necessary* that you can use the normal distribution to approximate the binomial distribution.

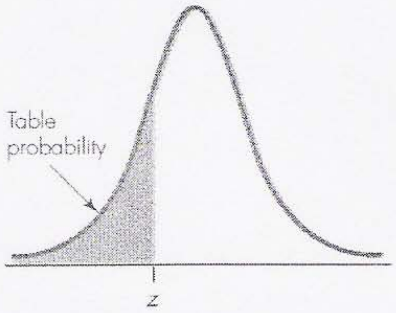
The most important fact to understand is that although a normal distribution is also a probability distribution function, it has its own characteristics with the most obvious one the fact that the variable it describes is continuous. Here are some of the characteristics of the normal distribution.

6.2.1 Characteristics of the normal distribution

- The form of the distribution is described as *bell-shaped*, meaning that it is symmetric (if it was possible to cut out the line forming the bell you would be able to fold it double with the two halves fitting on top of each other).
- The total area equals one, but can be broken up into sections, determined by the values given to the variable X on the horizontal axis.
- The mean and the standard deviation can have any given values and for every pair you have a unique member of the family of normal distributions.
  Please note that, It is not only the placement of the mean μ which determines the distribution of a particular normal distribution, the standard deviation σ (how the values are spread around the mean) also determines the form of the distribution.
- The values in the normal table give the areas for the probabilities of the standard normal distribution, i.e. the one whose mean = 0 and standard deviation =1. This implies that all general normal distributions must first be standardized with the formula $Z = \frac{X - \mu}{\sigma}$ before the normal table may be used.

Here is the table commonly used in this section.

Standard Normal Probabilities (for $z < 0$)



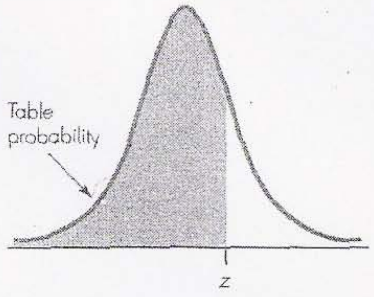
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

In the Extreme (for $z < 0$)

z	−3.09	−3.72	−4.26	−4.75	−5.20	−5.61	−6.00
Probability	.001	.0001	.00001	.000001	.0000001	.00000001	.000000001

S-PLUS was used to determine information for the "In the Extreme" portion of the table.

Standard Normal Probabilities (for $z > 0$)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

In the Extreme (for $z > 0$)

z	3.09	3.72	4.26	4.75	5.20	5.61	6.00
Probability	.999	.9999	.99999	.999999	.9999999	.99999999	.999999999

S-PLUS was used to determine information for the "In the Extreme" portion of the table.

6.2.2 The first thing we learn is how to read the area under the normal curve.

Example 15

Consider the standard normal random variable. Which of the following statements is *incorrect*?

(1) $P(Z \geq 1.63) = 0.0516$
(2) $P(Z \geq 0.50) = 0.3085$
(3) $P(Z < -1.63) = -0.0516$
(4) $P(Z > 1.28) = 0.1003$
(5) $P(-1.00 \leq Z \leq 1.00) = 0.6826$

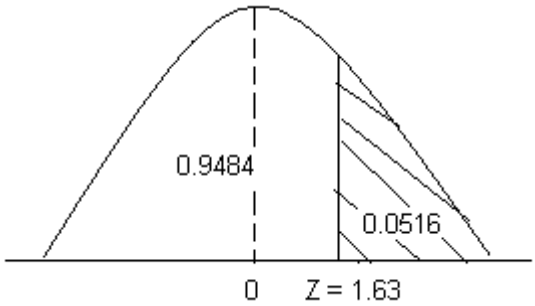
Key points to answer such questions:

- We have to draw the normal curve and shade the required area before attempting to answer the question.
- Negative values are read off from the negative part of the graph, likewise, positives values are read off from the positive part of the graph.
- If you are interested in the area $< Z$ or the area $\leq Z$ we read directly and take the probability given in the table.
- If you are interested in the area $> Z$ or the area $\geq Z$ read the area under the respective graph (negative or positive) but your answer should be "*one minus*" the area under the curve.
- If you are interested in the area between two sides $-z < Z < z$ or $-z \leq Z \leq z$ read off the areas from their respective graphs (positive or negative) and *subtract*. *Please remember that area is never negative*. Therefore, the subtraction must be the larger area minus the smaller area.

Solution

Option 1. Correct

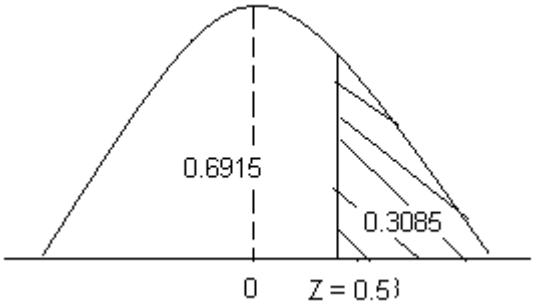
Diagram



$P(Z \geq 1.63) = 1 - 0.9484 = 0.0516$

Option 2. Correct

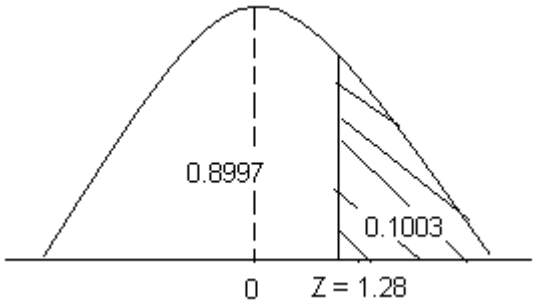
Diagram



$P(Z \geq 0.50) = 1 - 0.6915 = 0.3085$

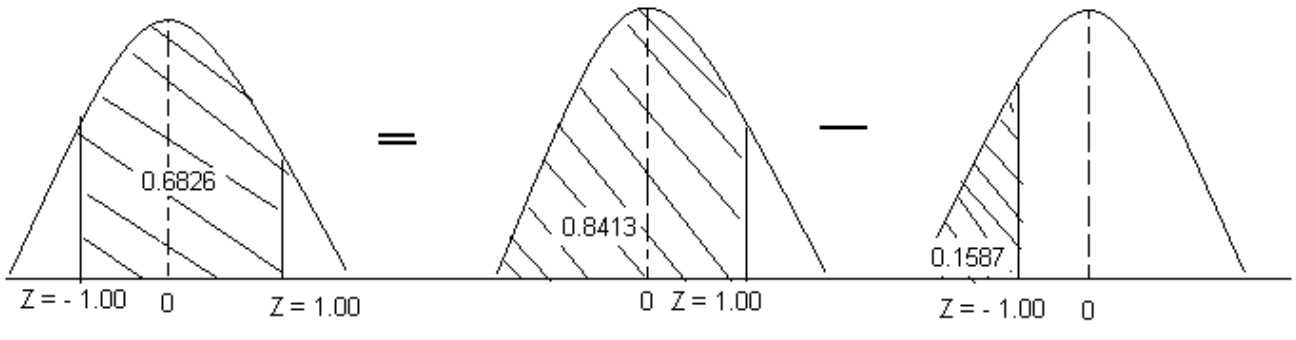
Option 3. Incorrect! Remember the area under the graph cannot be negative.
P(Z < −1.63) = 0.0516

Option 4. Correct
Diagram



$P(Z > 1.28) = 1 - 0.8997 = 0.1003$

Option 5. Correct
Diagram



$P(-1.00 \leq Z \leq 1.00) = 0.8413 - 0.1587 = 0.6826$

Alternative (3)

6.2.3 Give the mean (μ) and the standard deviation (σ).

In this case we first standardize the random variable. This requires us to change a random variable into standard normal using the formula $Z = \dfrac{X-\mu}{\sigma}$ before we can use the tables.

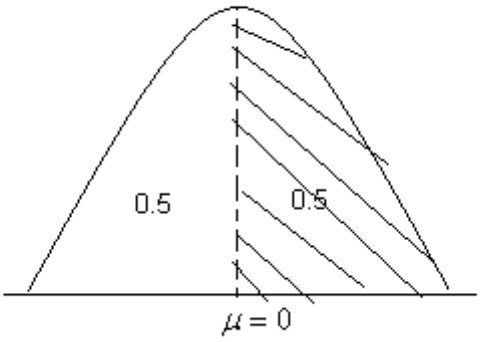
Example 16

Assume X is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 3$, the *incorrect* statement is

(1) $P(X \geq 15) = 0.5$
(2) $P(12 \leq X \leq 18) = 0.955$
(3) $P(X \leq 9) = 0.0228$
(4) $P(X = 20) = 0$
(5) $P(X \geq 12) = 0.8413$

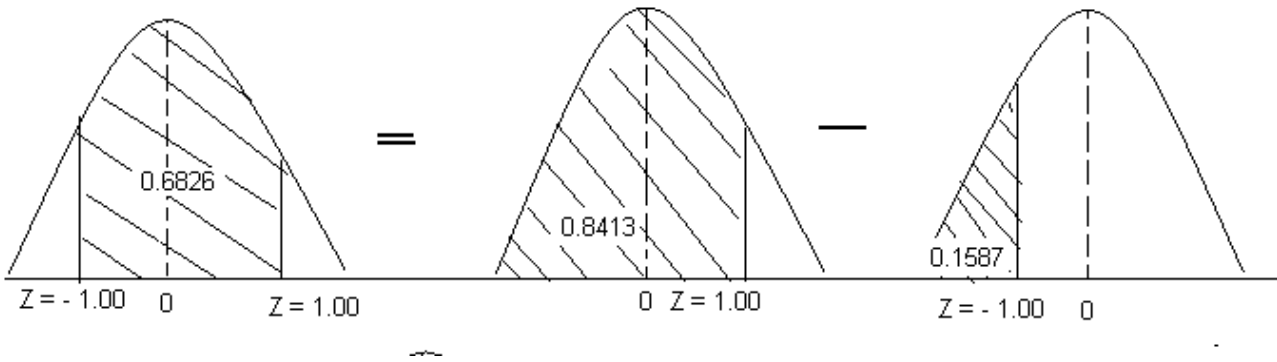
Solution

(1) *Correct.* $P(X \geq 15) = P\left(\dfrac{X-\mu}{\sigma} < \dfrac{15-15}{3}\right) = P\left(Z \leq \dfrac{15-15}{3}\right) = P(Z \leq 0) = 0.5$

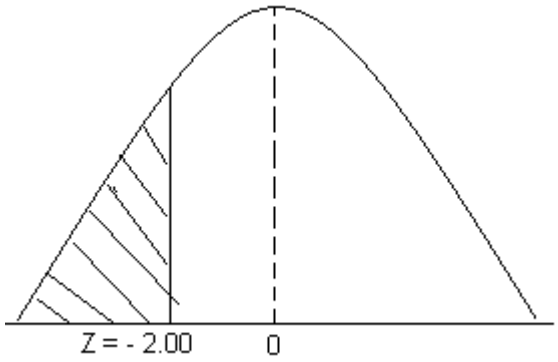


(2) *Incorrect.*

$$P(12 \leq X \leq 18) = P\left(\frac{12-15}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{18-15}{3}\right) = P(-1.00 \leq Z \leq 1.00) = 0.8413 - 0.1587 = 0.6826$$



(3) Correct $\quad P(X \leq 9) = P\left(\dfrac{X-\mu}{\sigma} \leq \dfrac{9-15}{3}\right) = P(Z \leq -2.00) = 0.0228$

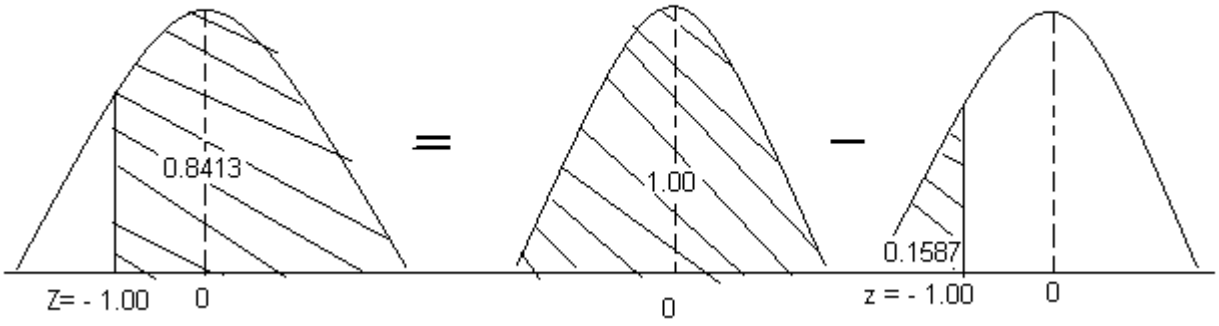


(4) *Correct.* $P(X = 20) = 0$

If the variable is continuous we assume that the probability of it assuming any fixed value is always zero! Remember the continuous variable lies somewhere within a small interval, but we cannot give a fixed value to it.

(5) Correct

$$P(X \geq 12) = P\left(\frac{X-\mu}{\sigma} \geq \frac{12-15}{3}\right) = P(Z \geq -1.00) = 1.00 - 0.1587 = 0.8413$$



Example 17

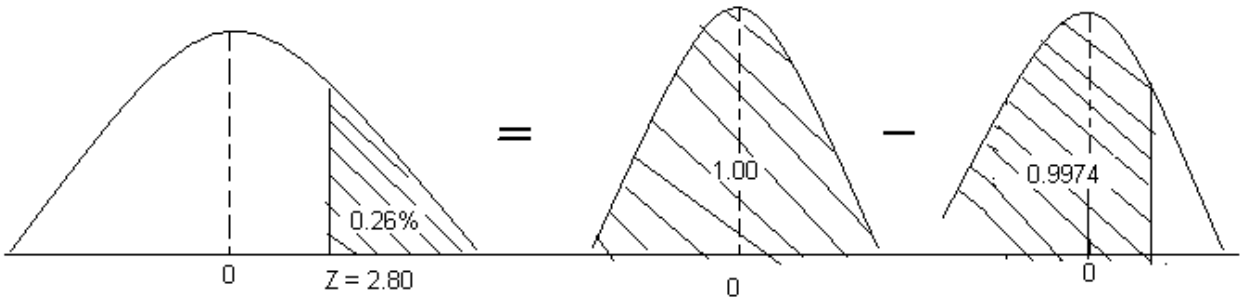
A manufacturer of tow chains finds that the average breaking point is at 3500 kilograms and the standard deviation is 250 kilograms. If you pull weight of 4200 kilograms with this tow chain, the percentage of the time you can expect the chain to break, is

(1) 2.8%
(2) 0.26%
(3) 49.74%
(4) 99.74%
(5) None of the above.

Solution

Option 2

$$P(X > 4200) = P\left(\frac{X-\mu}{\sigma} > \frac{4200-3500}{250}\right) = P(Z > 2.80) = 1.00 - 0.9974 = 0.26\%$$



6.2.4 Given the mean (μ), standard deviation (σ) and the probability

Questions in this section are usually tricky. They usually require you to find a random variable. This means that you have find the area under the curve first before you can read the corresponding Z-value. This is similar to working in the reverse direction.

Example 18

A retailer finds that the demand for a very popular board game averages 100 per week with a standard deviation of 20. If the seller wishes to have adequate stock 95% of the time, how many of the games must she keep on hand?

(1) 132.9
(2) 67.1
(3) 119
(4) 195.0
(5) 109

Solution

This question is about working backwards. To have adequate stock for 95% of the time implies that we are looking for a z-value such that 0.95 of the area lies to the left of it. We use the normal table to look for the value 0.95 inside the normal table because this is an *area*.

The z-value which corresponds to an area of 0.95 is 1.645. This value of 1.645 is the z-value we use to find the w-value.

$$P(X < w) = 0.95$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{w-100}{20}\right) = 0.95$$

$$P(Z < \frac{w-100}{200}) = 0.95$$

$$\frac{w-100}{20} = 1.645$$

$$w = 100 + (20 \times 1.645) = 132.9$$

SELF ASSESSMENT EXERCISE

Question 1

Which one of the following is *not* a characteristic of a normal distribution?

(1) The normal variable can take on only discrete values.
(2) It is a symmetrical distribution.
(3) The mean, median and mode are all equal.
(4) It is a bell-shaped distribution.
(5) The area under the curve is equal to one**.**

Question 2

Given that Z is a standard normal random variable, a negative value of z indicates that
(1) the value Z is to the left of the mean
(2) the value Z is to the right of the median
(3) the standard deviation of Z is negative
(4) the area between zero and Z is negative
(5) the area to the right of Z is equal to 1

Question 3

If Z is a normal variable with $\mu = 0$ and $\sigma = 1$, the area to the left of $Z = 1.6$ is
(1) 0.4452
(2) 0.9452
(3) 0.0548
(4) 0.5548
(5) 0.5000

Question 4

Use the normal table to find the Z- value Z_1 if the area to the *right* of Z_1 is 0.8413. The value of Z_1 is
(1) 1.36
(2) -1.36
(3) 0.00
(4) -1.00
(5) 1.00

Question 5

Let Z be a Z-score that is unknown but identifiable by position and area. If the area to the left of Z is 0.9306, then the value of Z must be
(1)-1.48
(2) 09603
(3) 1.48
(4) 0.4306
(5) -0.0694

Question 6
For a normal curve, if the mean is 20 minutes and the standard deviation is 5 minutes, then the area between 22 and 25 minutes is
(1) 0.1554
(2) 0.3413
(3) 0.4967
(4) 0.1859
(5) 0.0185

Question 7
A bakery firm finds that its average weight of the most popular package of biscuits is 200.5 g with a standard deviation of 10.5 g. What proportion of biscuit packages will weigh less than 180 g?
(1) 0.4744
(2) 0.0256
(3) 0.5226
(4) 0.4713
(5) 0.9744

Question 8
The average labour time to sew a pair of denims is 4.2 hours with a standard deviation of 0.5 hours. If the distribution is normal, then the probability of a worker finishing a pair of jeans in more than 3.5 hours is
(1) 0.0808
(2) 0.4192
(3) 0.5808
(4) 0.9192
(5) 0.9808

Question 9
A retailer finds that the demand for a popular board game averages 50 per week with a standard deviation of 20. If the seller wishes to have adequate stock 99% of the time, how many games must she keep on hand?
(1) 81.0
(2) 89.2
(3) 50.0
(4) 70.0
(5) 96.6

SOLUTIONS TO SELF ASSESSMENT EXERCISE

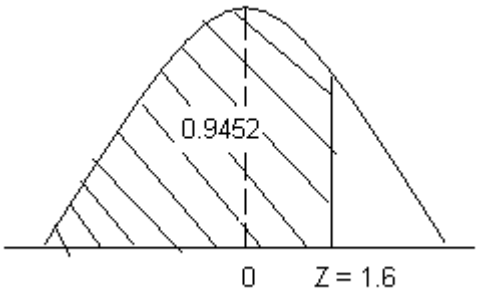
Question 1
The normal variable can only take on continuous values.

Alternative (1).

Question 2

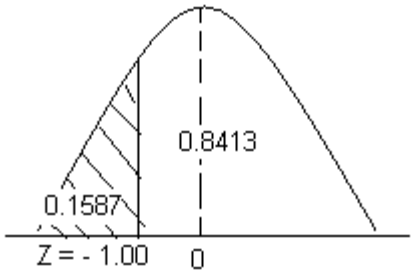
the value Z is to the left of the mean
Alternative (1).

Question 3
$P(Z < 1.6) = 0.9452$



Alternative (2).

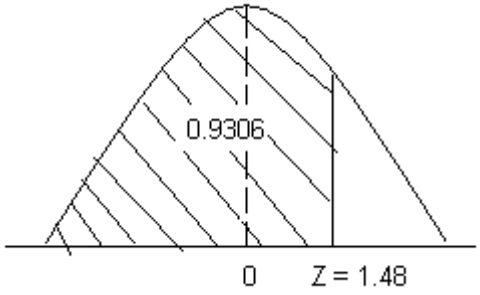
Question 4
$P(Z > Z_1) = 0.8413$
$P(Z < Z_1) = 0.1587 \Longrightarrow Z_1 = -1.00$



Alternative (4).

Question 5

P(Z > z) = 0.9306 ⟹ z = 1.48



Alternative (3).

Question 6

$$P(22 \le X \le 25) = P\left(\frac{22-20}{5} \le Z \le \frac{25-20}{5}\right)$$
$$= P(0.4 \le Z \le 1)$$
$$= 0.8413 - 0.6554$$
$$= 0.1859$$

Alternative (4).

Question 7

$$P(X < 180) = P\left(Z < \frac{180-200.5}{10.5}\right)$$
$$= P(Z < -1.95)$$
$$= 0.0256$$

Alternative (2)

Question 8

$$P(X > 3.5) = P\left(Z < \frac{3.5-4.2}{0.5}\right)$$
$$= P(Z > -1.4) = P(Z < 1.4)$$
$$= 0.9192$$

Alternative (4)

Question 9

This question is about working backwards

$$P(X \leq a) = 0.99$$

$$P\left(Z \leq \frac{a-50}{20}\right) = 0.99$$

$$\frac{a-50}{20} = 2.33$$

$$a = 50 + (20 \times 2.33) = 96.6$$

Alternative (5)