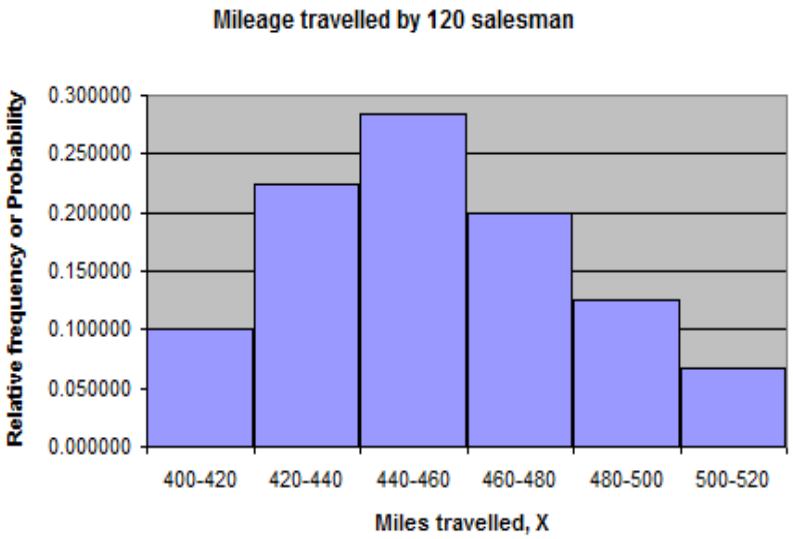
Probability Distributions



Probability?

Continuous probability distributions?

Discrete probability distributions?

The concept of probability is an important aspect of the study of statistics and within this presentation we shall introduce the reader to some of the concepts that are relevant to probability distributions.

However, the main emphasis of the chapter is to focus on the concepts of discrete and continuous probability distributions and not on the fundamentals of probability theory.

Learning Objectives

On completing this unit you should be able to do the following.

- Understand terms: experiment, outcome, sample space, relative frequency, sample probability, mutually exclusive and independent.
- Use the basic probability laws to solve simple problems.
- Understand the concept of a probability distribution and calculate a measure of average and dispersion.
- Understand when to apply the Binomial distribution.
- Understand when to apply the Poisson distribution.
- Use the normal distribution to calculate the values of a variable that correspond to a particular probability.
- Use the normal distribution to calculate the probability that a variable has a value between specific limits.
- Understand when to apply approximations to simplify the solution process.
- Solve problems using the Microsoft Excel spreadsheet.

Introduction to Probability

The concept of probability is an important aspect of the study of statistics and we shall introduce the reader to some of the concepts that are relevant to probability distributions. However, the main emphasis of the chapter is to focus on the concepts of discrete and continuous probability distributions and not on the fundamentals of probability theory.

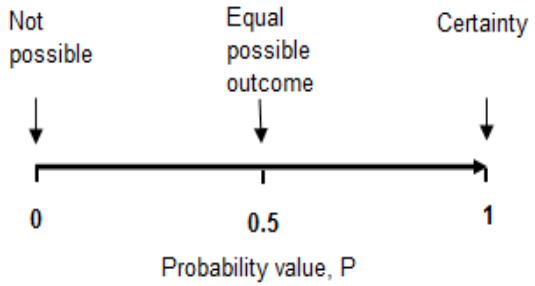
Table 5.1 summarizes the probability distributions that are applicable to whether the data variables are discrete/continuous and whether the distributions are symmetric/skewed.

Table 5.1

Measured characteristic	Variable type			
	Discrete		Continuous	
Shape	Symmetric	Skewed	Symmetric	Skewed
Distribution	Binomial	Poisson	Normal	Exponential

Basic Ideas

There are a number of words and phrases that encapsulate the basic concept of probability: chance, probable, odds. In all cases we are faced with a degree of uncertainty and concerned with the likelihood of a particular event happening. Statistically these words and phrases are too vague; we need some measure of likelihood of an event occurring. This measure is termed probability and is measured on a scale ranging between 0 and 1.



In order to determine a probability of an event occurring, data has to be obtained. This can be achieved through, for example, experience (subjective) or observation (empirical, relative frequency) or theoretical (a priori, deductive reasoning) methods.

The procedure or situation that produces a definite result (or outcome) is termed a random experiment. For example tossing a coin, rolling a die, recording the income of a factory worker, determining defective items on an assembly line, are all examples of 'experiments'.

Relative Frequency

Suppose we perform the experiment of throwing a die and note the score obtained. We repeat the experiment a large number of times, say 1000, and note the number of times each score was obtained. For each number we could derive the ratio of occurrence that an event A will happen (*m*) to the total number of experiments (*n* = 1000). This ratio is called the relative frequency.

In general, if event A occurs *m* times, then your estimate of the probability that A will occur is as follows:

$$P(A) = \frac{m}{n}$$

The result of the die experiment is shown in Table 5.2 below:

Score	1	2	3	4	5	6
Frequency	173	168	167	161	172	159
Relative Frequency	0.173	0.168	0.167	0.161	0.172	0.159

Rules
1.The probability of each event lies between 0 and 1.
2.The sum of the probabilities of these events will equal 1.
3.If we know the probability of an event, then the probability of it not occurring is P(Event not occurring) = 1 - P(Event occurs).

The Probability Laws

In case we are measuring probabilities for multiple events, very often we would like to be able to calculate what is the probability that either one or the other event will happen, or the probability that both events will happen simultaneously.

Addition Law:

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

Mutually Exclusive Events:

$P(A \text{ and } B) = 0$

Independent Events:

$P(A/B) = P(A)$

implies $P(A \text{ and } B) \neq 0$

Multiplication Law for Two Events:

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

Multiplication Law for Two Independent Events:

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

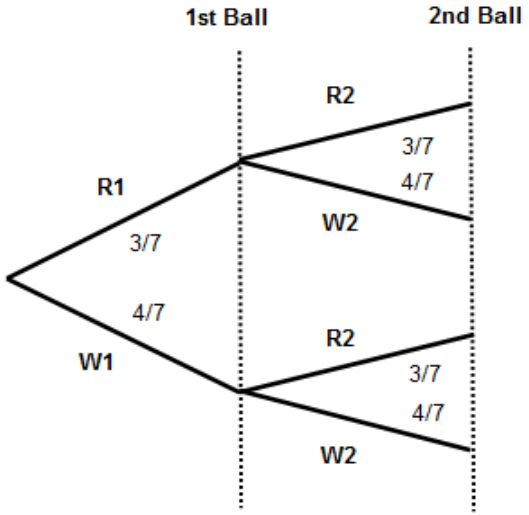
Probability Tree Diagram

Probability tree diagrams provide a visual aid to help you solve complicated probability problems.

Example 5.4

A bag contains three red and four white balls. If one ball is taken at random and then replaced and another ball is taken, calculate the following probabilities: (a) P(R, R), (b) P(just one Red), (c) P(2nd Ball White)?

Figure 5.4 displays the experiment in a tree diagram. Each branch of the tree indicates the possible result of a draw and associated probabilities.



We can now use this diagram to calculate the required probabilities, e.g.

P(R, R) = P(R1) x P(R2) = 9/49

Introduction to Probability Distributions

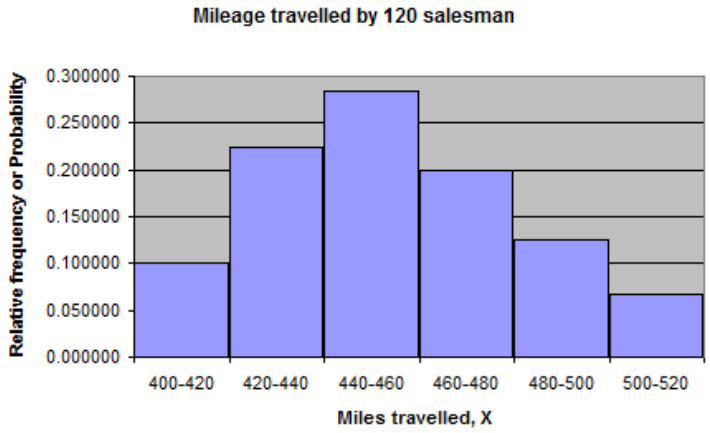
We have already stated that the concept of relative frequency is one way to interpret probability.

Example 5.6

Consider the frequency distribution representing the mileage travelled by 120 salesmen. From this frequency distribution we can calculate the relative frequency and create a histogram for relative frequency (or probability) against miles travelled.

	A	B	C	D	E
1	**Probability distributions**				
2					
3		Mileage travelled	Frequency, f	Relative frequency	
4		400-420	12	0.100000	=C4/$C$11
5		420-440	27	0.225000	
6		440-460	34	0.283333	
7		460-480	24	0.200000	
8		480-500	15	0.125000	
9		500-520	8	0.066667	=C9/$C$11
10					
11		Total =	120	1.000000	
12			=SUM(C4:C9)	=SUM(D4:D9)	

Probability distribution



Expectation and Variance

For a probability distribution we can apply the definition of probability, using the concept of relative frequency, to create equations that can be used to calculate the mean and standard deviation. For a probability distribution the mean value is called the expected value.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Probability Distributions - Expectation and Variance**											
2												
3												
4	Mileage travelled	Frequency, f	LCB	UCB	Class mid point, X		Relative frequency, P(X = x) = f/N		x * P(X = x)		X^2*P(X = x)	
5												
6	400 - 420	12	400	420	410	=(C6+D6)/2	0.10	=B6/$D$17	41.000	=E6*G6	16810.00	=E6^2*G6
7	420 - 440	27	420	440	430		0.23		96.750		41602.50	
8	440 - 460	34	440	460	450		0.28		127.500		57375.00	
9	460 - 480	24	460	480	470		0.20		94.000		44180.00	
10	480 - 500	15	480	500	490		0.13		61.250		30012.50	
11	500 - 520	8	500	520	510	=(C11+D11)/2	0.07	=B11/$D$17	34.000	=E11*G11	17340.00	=E11^2*G11
12												
13	Summary Statistics											
14												
15		N = Σf =	120.00	=SUM(B6:B11)								
16		ΣXP =	454.50	=SUM(I6:I11)								
17		ΣX^2P =	207320.00	=SUM(K6:K11)								
18		Mean =	454.50	=C16								
19		Variance =	749.75	=C17-C18^2								
20		Standard Deviation =	27.38	=C19^0.5								

Expected value, E(X)

$$E(X) = \sum X \times P(X)$$

Variance value, VAR(X)

$$VAR(X) = \sum X^2 \times P(X) - [E(X)]^2$$

Standard deviation, SD(X)

$$SD(X) = \sqrt{VAR(X)}$$

Continuous Probability Distributions

A random variable is a variable that provides a measure of the possible values obtainable from an experiment.

For example, we may wish to count the number of times that the number 3 appears on the tossing of a fair die, or we may wish to measure the weight of people involved in measuring the success of a new diet programme.

The second example consists of numbers that can take any value with respect to measured accuracy (160.4 lbs, 160.41lbs, 160.414 lbs, etc) and is an example of a continuous random variable.

In this section we shall explore the concept of a continuous probability distribution with the focus on introducing the reader to the concept of a Normal probability distribution.

The Normal Distribution

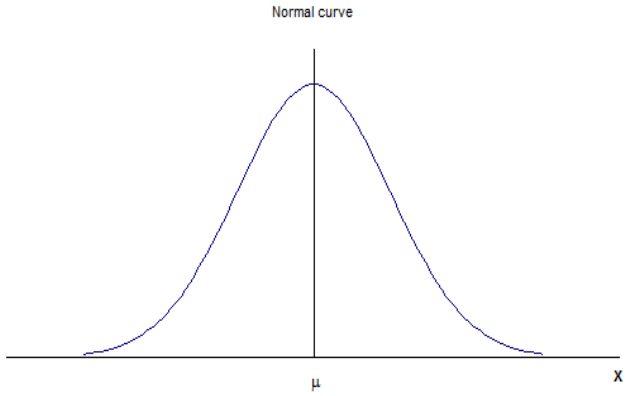
When a variable is continuous, and its value is affected by a large number of chance factors, none of which predominates, then it will frequently appear as a Normal distribution. This distribution does occur frequently and is probably the most widely used statistical distribution. Some of the real-life variables having a Normal distribution can be found, for example, in manufacturing (weights of tin cans), or can be associated with the human population (people's heights).

The Normal distribution is governed by Equation (5.9):

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(X-\mu)^2}{2\sigma^2}\right]$$

This equation can be represented graphically and we note:

1.bell shape,
2.symmetry,
3.mean = median = mode.



Example

Example 5.10

A manufacturing firm quality assures components manufactured and historically the length of a tube is found to be normally distributed with the population mean of 100 cms and a standard deviation of 5 cms.

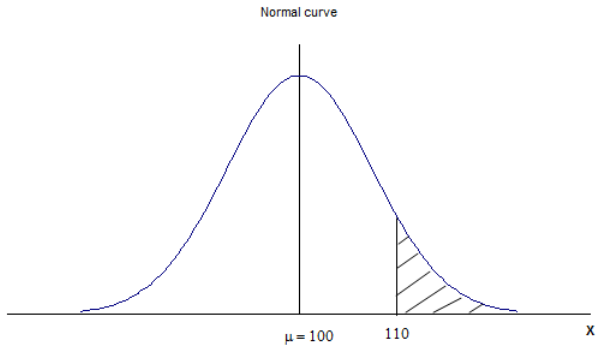
Calculate the probability that a random sample of one tube will have a length of at least 110 cms?

From the information provided we define X has the tube length in cms and population mean μ = 100 and standard deviation = 5. This can be represented using the notation $X \sim N(100, 5^2)$.

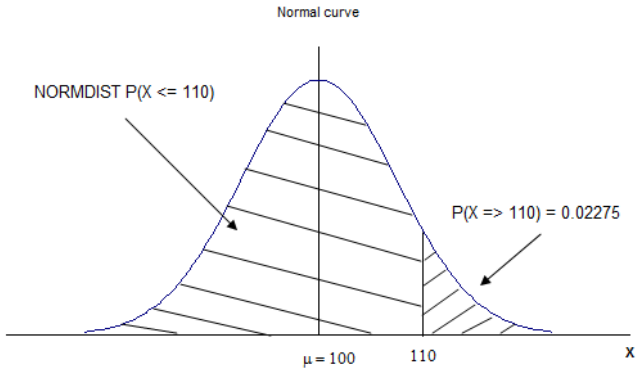
The problem we have to solve is to calculate the probability that one tube will have a length of at least 110 cms. This can be written as $P(X \geq 110)$.

Continued

This can be written as P(X ≥ 110) and is represented by the shaded area.



This problem can be solved by using the Excel function NORMDIST (X, μ, [illegible]^2, TRUE).



From Excel, P(X ≥ 110) = 0.02275 or 2.3%

	A	B	C	D
1	Example 5.10 and 5.12			
2				
3		Normal distribution		
4				
5		Mean μ =	100	
6		Standard deviation σ =	5	
7				
8		X =	110	
9				
10		P(X <= 110) =	0.97725	=NORMDIST(C8,C5,C6,TRUE)
11				
12		P(X => 110) =	0.02275	=1-C10

The Standard Normal Distribution

If we have two different populations, both following normal distribution, it could be difficult to compare them as the units might be different, the means and variances might be different, etc.

If this is the case, we would like to be able to standardize these distributions so that we can compare them. This is possible by creating the standard normal distribution.

The standard normal distribution is a normal distribution whose mean is always 0 and a standard deviation is always 1.

Normal distributions can be transformed to standard normal distributions by Equation (5.10):

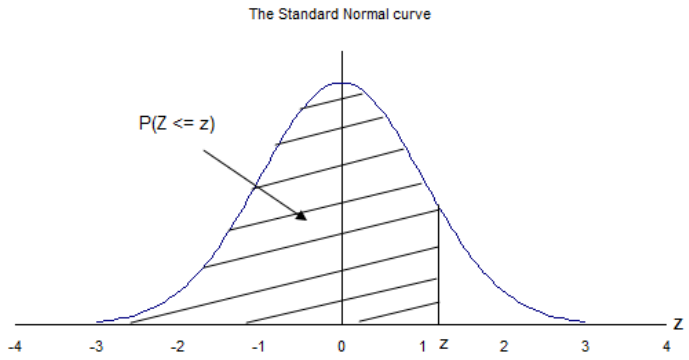
$$Z = \frac{(X - \mu)}{\sigma}$$

Example

Example 5.12.

Example 5.10 consisted of solving the problem, P(X ≥ 110), with μ = 100 and σ = 5. Using Equation (5.12) we can replace X with Z, P(X ≥ 110) = P(Z ≥ + 2).

The value of P(Z ≥ 2) can be calculated using Excel's NORMSDIST () function.



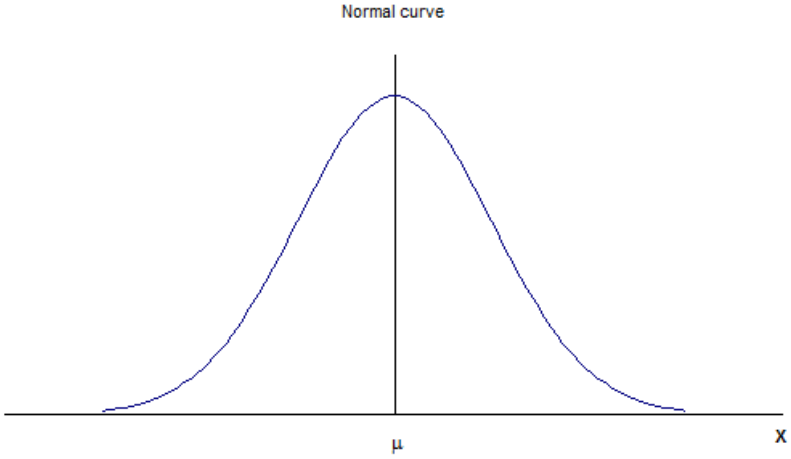
From Excel, P(X ≥ 110) = P(Z ≥ +2) = 0.02275 or 2.3%

	A	B	C	D
1	Example 5.10 and 5.12			
2				
3		Normal distribution		
4				
5		Mean μ =	100	
6		Standard deviation σ =	5	
7				
8		X =	110	
9				
10		P(X <= 110) =	0.97725	=NORMDIST(C8,C5,C6,TRUE)
11				
12		P(X => 110) =	0.02275	=1-C10
13				
14		Z =	2	=(C8-C5)/C6
15		P(Z <= +2) =	0.97725	=NORMSDIST(C14)
16		P(Z => +2) =	0.02275	=1-C15

Empirical Rules

For a normal distribution we can show that a simple relationship exists between the number (or proportion) of data points, the population mean value (μ), and the population standard deviation (σ).

For a normal distribution, the middle (average) value is the population mean, μ, with the data points (or values) symmetrically and evenly spread out either side of this mean value. For one standard deviation either side of this mean value will contain approximately 34% of all the data values.



Empirical rules:

- μ ± σ = 68% of data values
- μ ± 2σ = 95% of data values
- μ ± 3σ = 99.7% of data values.