Statistics &Data Analysis Concepts for Data Science and ML 6

6

Continuous Probability Distributions

Learning Objectives

The primary objective of this chapter is to help you understand continuous probability distributions that are critical to decision making in data analysis. After completing this chapter, you should be able to:

- **Understand random variables and continuous probability distributions**
- **Understand normal distribution and solve problems involving normal distribution using the normal formula, normal table, and computer packages**
- **Solve problems involving exponential distribution using the exponential distribution function, exponential table, and computer packages**
- **Understand and solve problems involving uniform distribution using the uniform distribution function and computer packages**

Learning Objectives

- **Gain an in depth understanding of normal, exponential, and uniform distributions through computer simulation and computer generated experiments**

- **Understand the important and widely used distributions related to the normal distribution including the t-distribution, chi-square, and F-distribution**

Background

- **The frequency of occurrence and the probability of occurrence plays an important role in statistical analysis.**

- **Combining these two concepts (frequency of occurrence and the probability theory) provides useful insight that lays the foundation for statistical analysis of samples and making decisions under uncertainty.**

- **Since most statistical methods are based on random sampling, it is important to understand and characterize the behavior of random variables.**

Random Variable

- ***The numerical value is a variable and the value achieved is subject to chance and therefore, it is determined randomly.***

- ***A random variable is a numerical quantity whose value is determined by chance. Note that a random variable must be a numerical quantity.***

- ***The relationship between the values of a random variable and their probabilities is summarized by a probability distribution.***

- ***A probability distribution of a random variable may be described by the set of possible random variable's values and their probabilities. The probability distribution provides a probability for each possible value or outcome of a random variable.***

Discrete and Continuous

The two basic types of random variables are:

***Discrete** and **Continuous**.*

Discrete Random Variable:

A random variable that can assume only integer value or whole number is known as discrete. Chapter 5 dealt with discrete random variables and important discrete probability distributions.

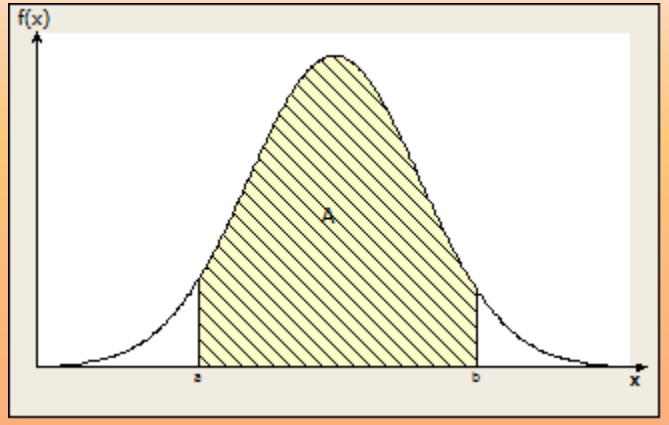
Examples of a discrete random variable: is the number of customers arriving at a bank. Another example of a discrete random variable would be rolling a single die in which the outcomes are 1, 2, 3, 4, 5, or 6.

Continuous random variable:

The random variable may assume any value over a continuous range of possibilities . Since it can take any value within a given range, we are not able to list all possible values of variable.

Continuous Random Variable

The graphical display of the probability distribution x of a continuous random variable is a smooth curve that may assume a shape shown in Figure below.



Using this smooth curve, the entire range of probability can be calculated. The total area under the curve is 1. The curve is a function of x which is denoted by $f(x)$ and is called a ***probability density function*** or a ***probability distribution***.

To determine the probability of a continuous probability distribution such as, the one shown above, one needs to evaluate the area between the points of interest.

Normal Distribution

Background: A continuous random variable x is said to follow a normal distribution with parameters μ and σ if the probability density function of x is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

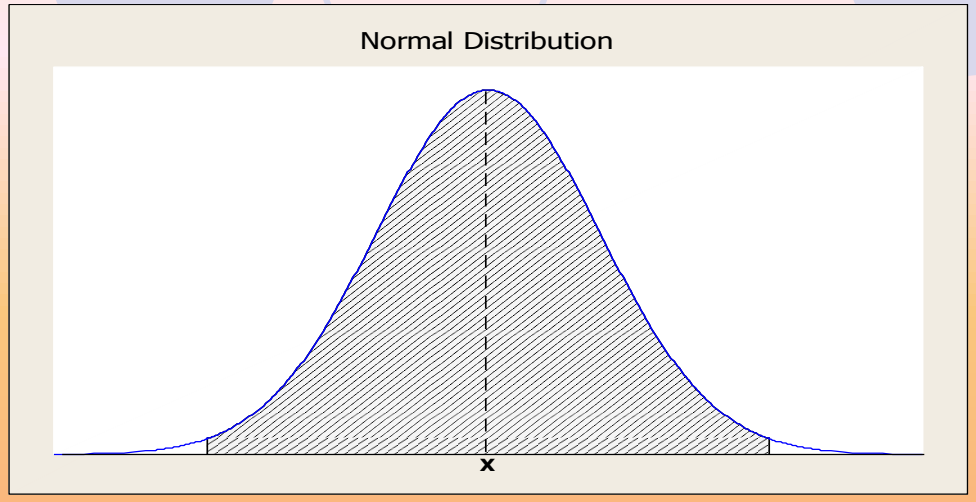
Where f(x) is the probability density function, μ =mean, σ=standard deviation, and e= 2.71828 which denotes the base of the natural logarithm.

Properties of Normal Distribution

The distribution has the following properties:

- The normal curve is a bell shaped curve. It is symmetrical about the line $x = \mu$. The typical shape of the normal curve is shown in Figure on the next slide.

Properties of Normal Distribution…cont.



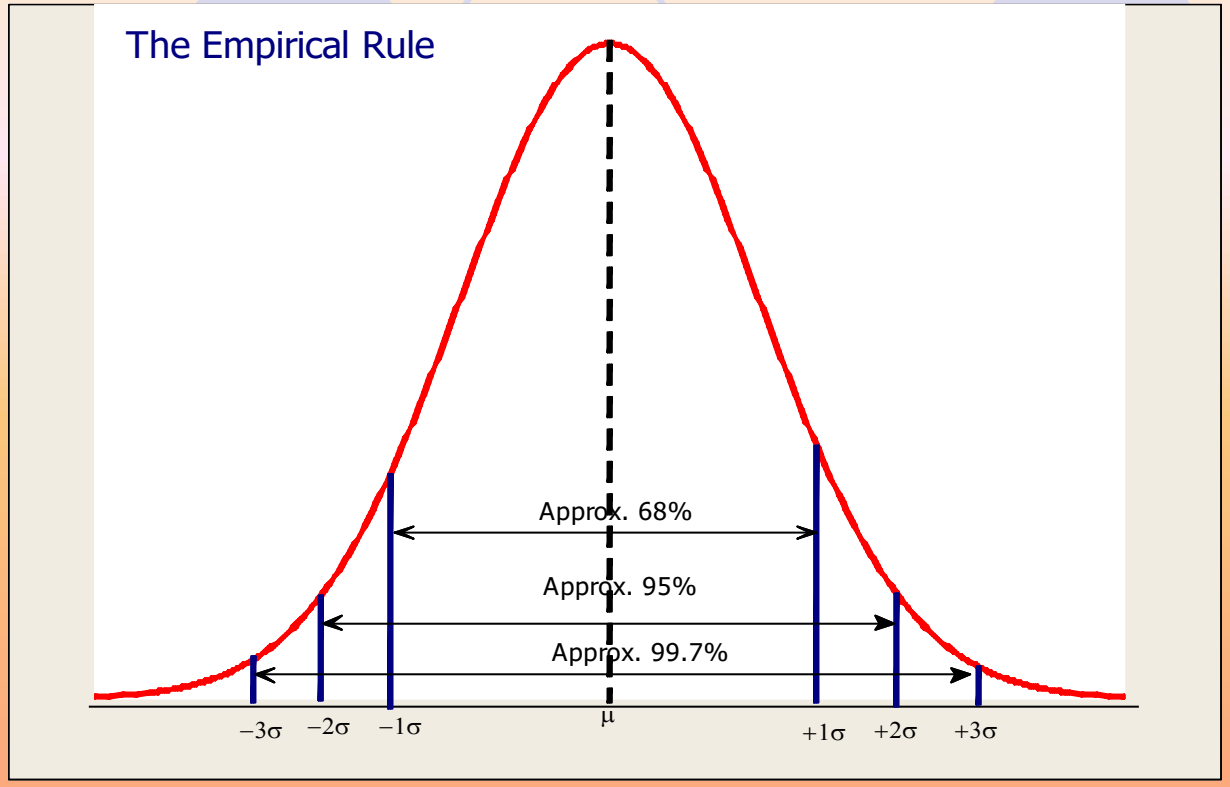
The Normal Curve

- The mean, median, and mode of the distribution have the same value.
- As x increases, f(x) decreases rapidly. The maximum probability occurs at the point where $x = \mu$, and is given by:

$$p(x = \mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

Since the probability f(x) can never be negative, no portion of the curve lies below the x-axis.

Properties of Normal Distribution…cont.



Area Property:

$\mu \pm \sigma = 0.6826$

$\mu \pm 2\,\sigma = 0.9544$

$\mu \pm 3\,\sigma = 0.9973$

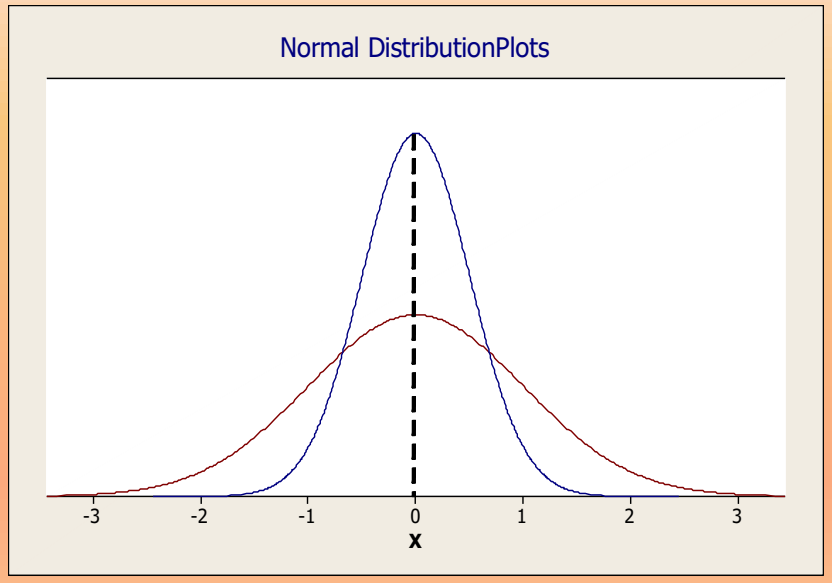
approximately 68% of the observations will fall between the mean and ± 1σ (one standard deviation)

approximately 95% of all observations will fall between the mean and ± 2σ (two standard deviations)

approximately 99.73% of all observations will fall between the mean and ± 3σ (three standard deviations)

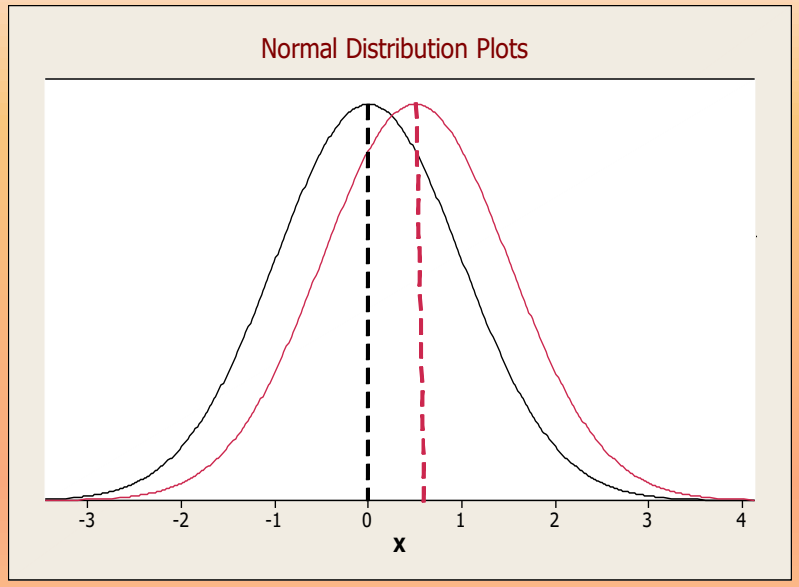
Properties of Normal Distribution…cont.

The shape of the curve depends upon the mean (μ) and standard deviation (σ). The mean μ and the standard deviation σ are the parameters of the normal distribution.



Normal Curves: Same

Mean with different Standard Deviations



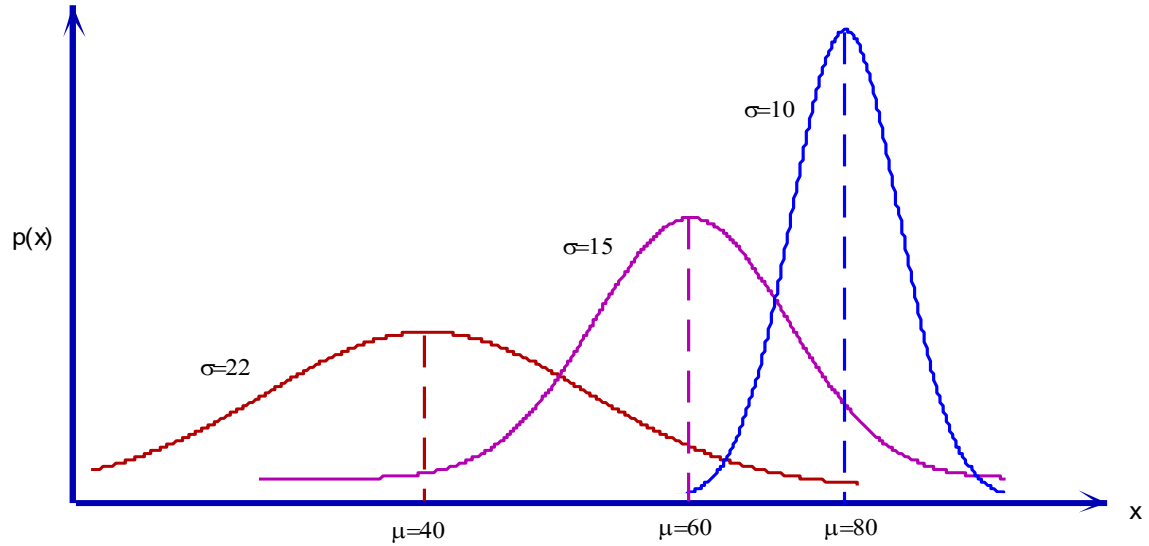
Normal Curves: Same

Standard Deviation with different Means

Properties of Normal Distribution…cont.

The shape of the curve depends upon the mean (μ) and standard deviation (σ). The figure below shows the normal distributions for different values of μ and σ. Note that higher the value of the standard deviation σ, more spread out the curve is.

Normal distribution curves for different values of μ and σ



Properties of Normal Distribution…cont.

The Normal distribution is perhaps the most important distribution in statistics and plays a very important role in statistical theory because of the following characteristics:

- ***Many widely used distributions, such as Binomial, Poisson, and Hypergeometric distributions can be approximated by the normal distribution.***

- ***Many of the sampling distributions, such as t-distribution, F-distribution, and Chi-square distribution, tend to be normal as the sample size increases.***

- ***In some cases, even if a variable is not normally distributed, it sometimes can be brought to normality by a simple transformation of variable. For example: if the distribution of a variable X is skewed and therefore not normal, it may be that the distribution of might be normal.***

Properties of Normal Distribution…cont.

- ***The distributions of many sample statistics, such as the distribution of a sample mean, sample variance, sample proportion, etc., follows a normal distribution as the sample size gets larger. These can be studied with the help of the normal distribution.***

- ***The theory of small sample tests, (e.g., t, F, χ^2) is based on the fundamental assumption that the population from which samples are drawn is normal or approximately normal.***

- ***The normal distribution has wide application in statistical quality control/process control and in process capability analysis.***

- ***The normal distribution represents the distribution of random errors in many kinds of measurements. Many experimental data often turn out to follow the normal distribution.***

The Standard Normal Distribution

To calculate the normal probability, $p(x_1 \leq X \leq x_2)$ where X is a normal variate with parameters μ and σ, we need to evaluate:

$$\int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

To evaluate the above expression, none of the standard integration techniques can be used. However, the expression can be numerically evaluated for $\mu = 0$ and $\sigma = 1$.

When the values of the mean $\mu = 0$ and standard deviation $\sigma = 1$, the normal distribution is known as the standard normal distribution and is usually denoted by Z.

Calculating Normal Probabilities

When the random variable x is normally distributed with mean μ and variance σ^2 that is; $x \sim N(\mu, \sigma^2)$ we can calculate the probabilities between the points of interests by standardizing the normal curve.

The standardized value is known as the standard or standardized normal distribution and is given by:

$$z = \frac{x - \mu}{\sigma}$$

Z= distance from the mean to the point of interest (x) in terms of standard deviation units

x = point of interest

μ = the mean of the distribution, and

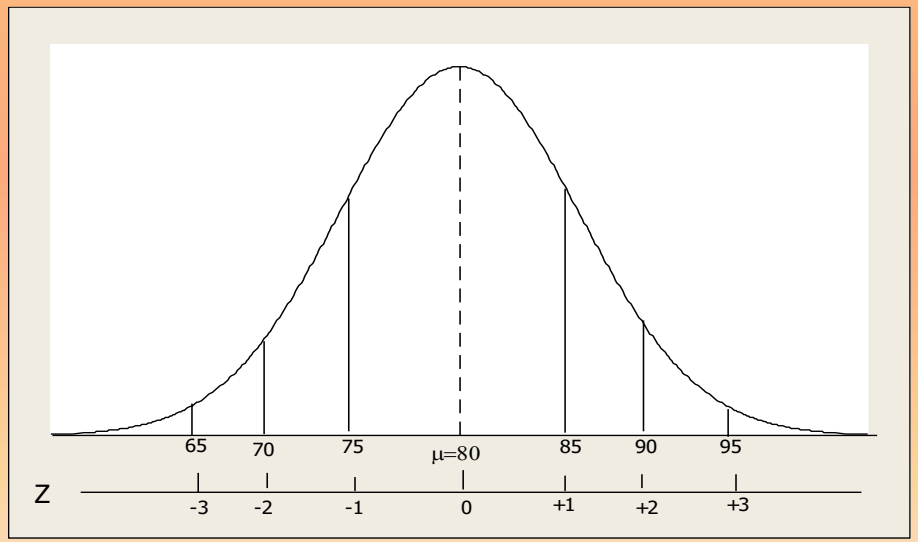
σ = the standard deviation of the distribution.

The idea behind standardizing is that any probability involving X can be expressed in terms of a standard normal random variable Z so that a single table can be used to calculate the areas or the probabilities.

Example 1

Consider a data set that has a normal distribution with a mean, $\mu = 80$ and standard deviation, $\sigma = 5$. Determine the percentage of observations within each of the following range of values: [a] 75 and 85 [b] 70 and 90 [c] 65 and 95.

Solution: ***To determine the percent of observations between the points of interest, first calculate the z values (using the Z-score formula) corresponding to the points of interest. The Z-value tells us how far the point is from the mean in standard deviations. The area or the probability corresponding to the Z values can be obtained from a Standard Normal Table (explained below).***

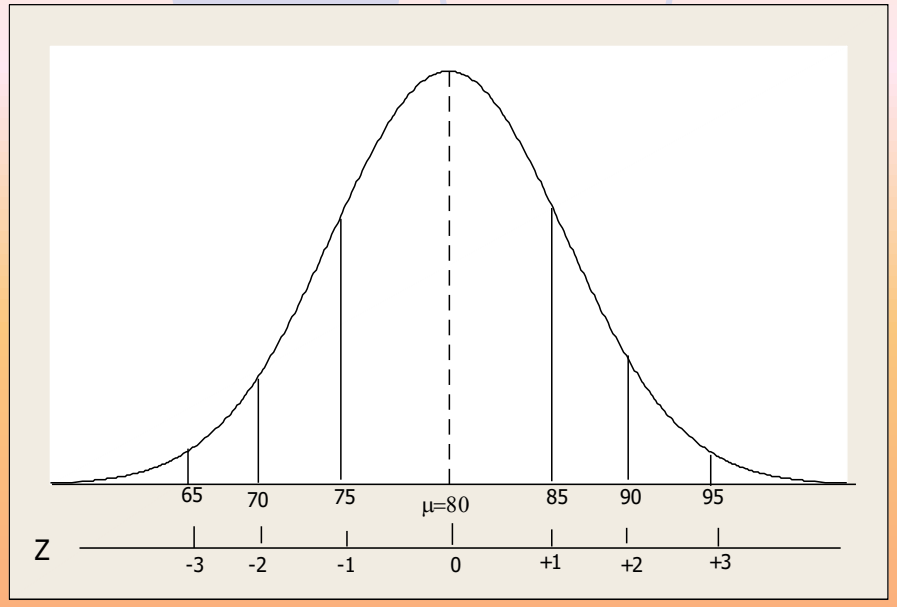


The Z-score formula)

$$z = \frac{x - \mu}{\sigma}$$

[a] to find the percent between 75 and 85 first, find the z values for these points

Example 1…cont.



z-score for 75 or, $x = 75$

Given $\mu = 80$ ***and*** $\sigma = 5$

$$z = \frac{x - \mu}{\sigma} = \frac{75 - 80}{5} = -1$$

z-score for 80 or, $x = 80$

$$z = \frac{x - \mu}{\sigma} = \frac{80 - 80}{5} = 0$$

z-score for 85 or, $x = 85$

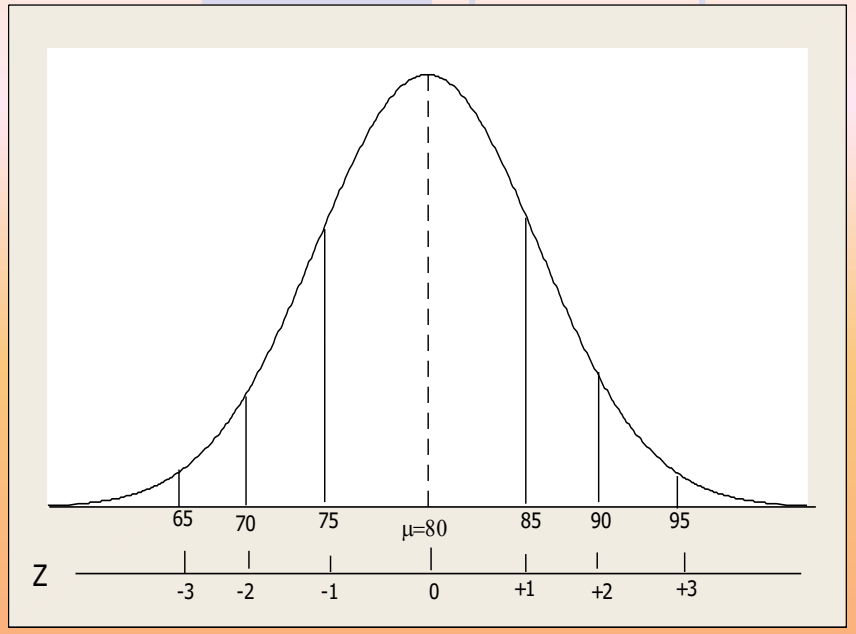
$$z = \frac{x - \mu}{\sigma} = \frac{85 - 80}{5} = 1.0$$

The z-values for the other points are calculated in a similar way using the z-score formula .

Note: Z =1 means that the point of interest x is 1 standard deviation from the mean.

The z-value is always zero at the center of the distribution because at this point the x and μ are the same for z - formula. The z values on the right hand side of the mean are positive and the values on the left side are negative.

Example 1…cont.



In this example, the range of values: [a] 75 and 85 [b] 70 and 90 [c] 65 and 95 fall between ±1, ±2, and ±3 standard deviations from the mean. See the z-scores corresponding to these values in the figure

From the empirical rule we know that the mean and ±1, ±2, and ±3 standard deviations contain approximately 68%, 95% and 99.7% of the observations.

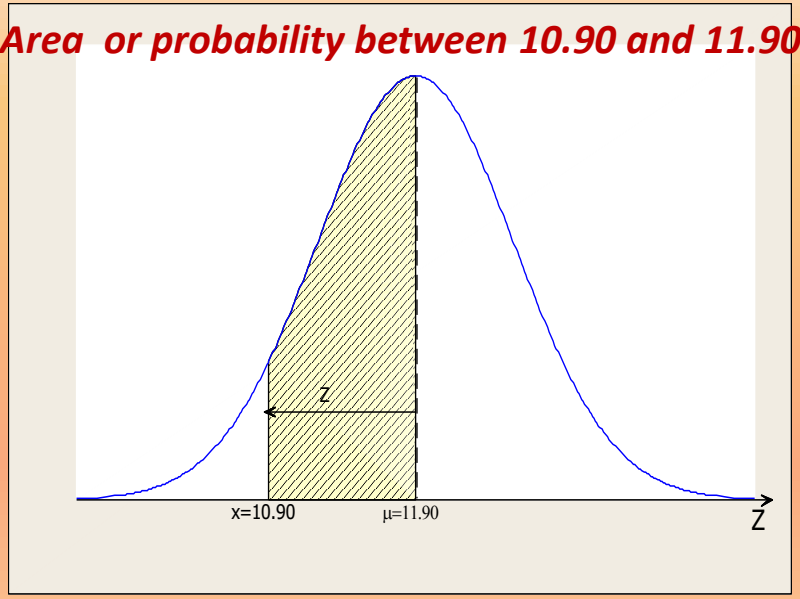
Therefore, the percentage of observations within each of the following range of values: [a] 75 and 85 [b] 70 and 90 [c] 65 and 95 are approximately 68%, approximately 95%, and approximately 99.7%

If the Z-values are other than ±1, ±2, and ±3 then we need to use the standard normal table to find the percentages or the probabilities. This is demonstrated in the next example.

Example 2

In a company, the hourly wages are normally distributed with a mean μ = \$11.90 and a standard deviation σ = \$0.40.

[a] What percentage of the workers earn between \$10.90 and \$11.90?



To determine the required percentage, we need to determine the area between \$10.90 and \$11.90. The area is calculated by calculating the z-score. The figure shows the required area is shaded. The value is calculated as

$$z = \frac{x - \mu}{\sigma} = \frac{10.90 - 11.90}{0.40} = -2.5$$

Note that = -2.5 means that the value 10.90 is 2.5 standard deviation away from the mean. The negative sign indicates that the value is on the left side of the mean.

To determine the percentage for = -2.5, go to the standard normal table (z-table- next slide) and locate 2.5 under the z column. Read the value for z =2.5 and the column 0.00. This value is 0.4938. (Note that z =-2.5 and z = +2.5 will have the same area because the normal distribution is symmetrical).

This means that = -2.5 is equivalent to an area of 0.4938 or 49.38%. So the required percentage is

$$p(10.90 \le x \le 11.90) = 0.4938 or 49.38\%$$

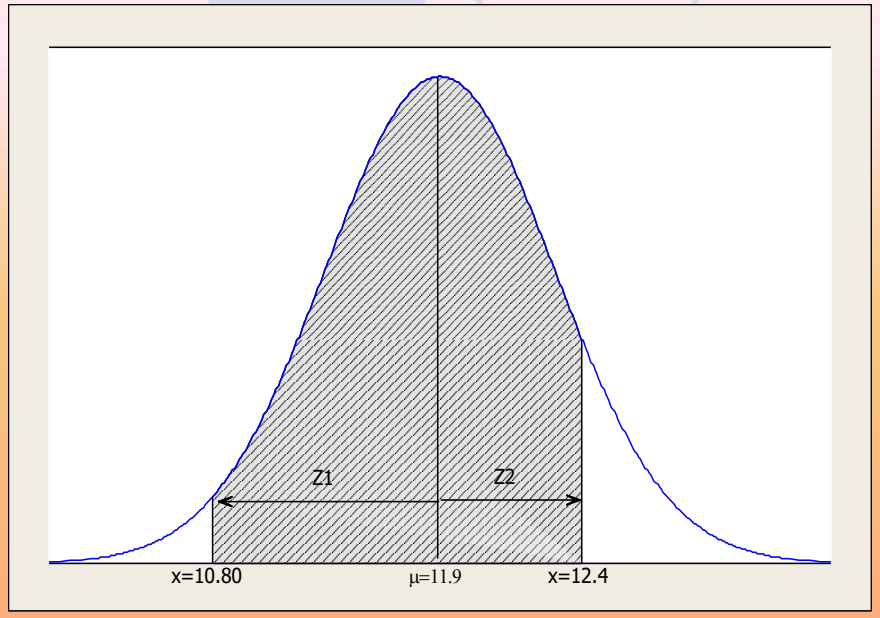
Note: In normal distribution, finding the area or the percentage between the points of interest is same as finding the probability.

We will refer to the standard normal table on page 22 to evaluate the percentages for All the examples.

Standard Normal Distribution Table

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

[b] What percentage of the workers earn between $10.80 and $12.40?



The required area or percentage is shown in Figure. In this case, we find the areas between $10.80 and $11.90 and between $11.90 and $12.40 and ***add*** them. The calculations are shown below:

$$z_1 = \frac{x-\mu}{\sigma} = \frac{10.80-11.90}{0.40} = -2.75 \Rightarrow 0.4970$$

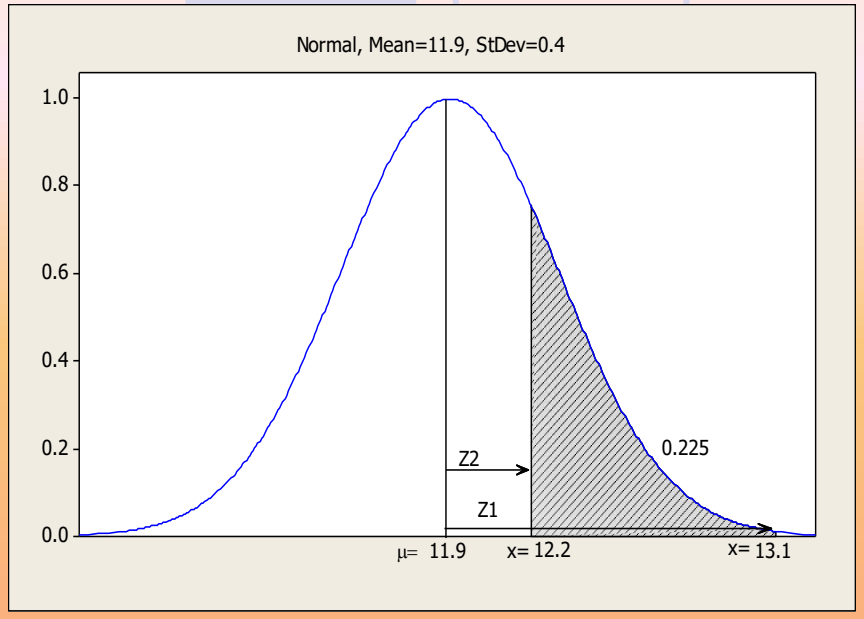
$$z_2 = \frac{x-\mu}{\sigma} = \frac{12.40-11.90}{0.40} = +1.25 \Rightarrow 0.3944$$

[note that 0.4970 is the area corresponding to z=-2.75 and 0.3944 is the area corresponding to z = +1.25 from the normal table-page 22]

The required percentage is $p(10.80 \le x \le 12.40) = 0.4970 + 0.3944 = 0.8914$

or, 89.14% of the workers earn between $10.80 and $12.40.

[c] What percentage of the workers earn between $12.20 and $13.10?



In this case we need to find the areas between $11.90 and $13.10 and between $11.90 and $12.20, and ***subtract*** them. The calculations are shown below.

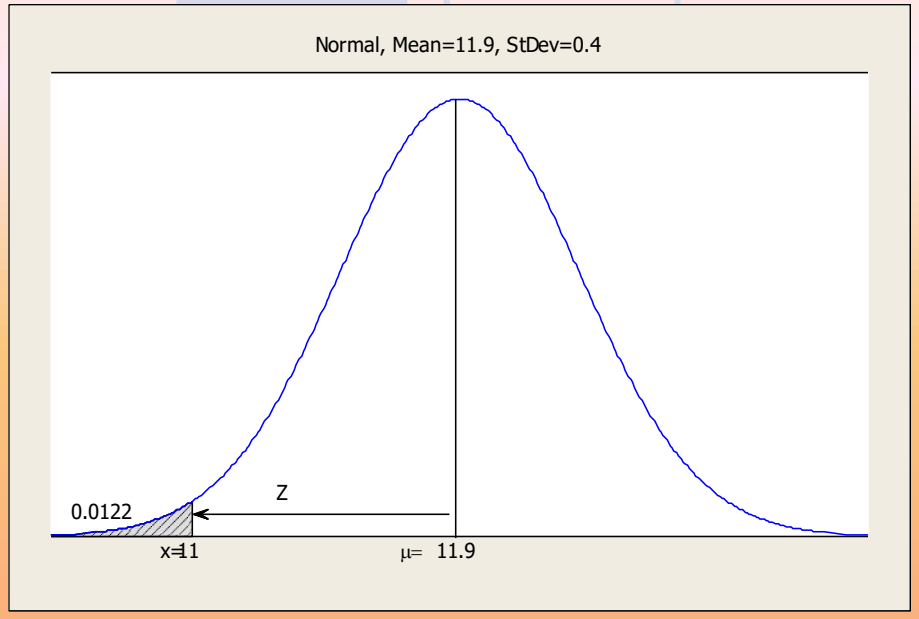
$$z_1 = \frac{x-\mu}{\sigma} = \frac{13.10-11.90}{0.40} = 3.0 \Rightarrow 0.4987$$

$$z_2 = \frac{x-\mu}{\sigma} = \frac{12.20-11.90}{0.40} = +0.75 \Rightarrow 0.2734$$

The required percentage is $p(12.20 \le x \le 13.10) = 0.4987 - 0.2734 = 0.2253$

which means 22.53 % of the workers earn between $12.20 and $13.10.

[d] What percentage of the workers earn less than $11.00?



The required area or percentage is shown in Figure . In this case, we need to find the areas between $11.90 and $11.00 and ***subtract these area from 0.5***. The calculations are shown below

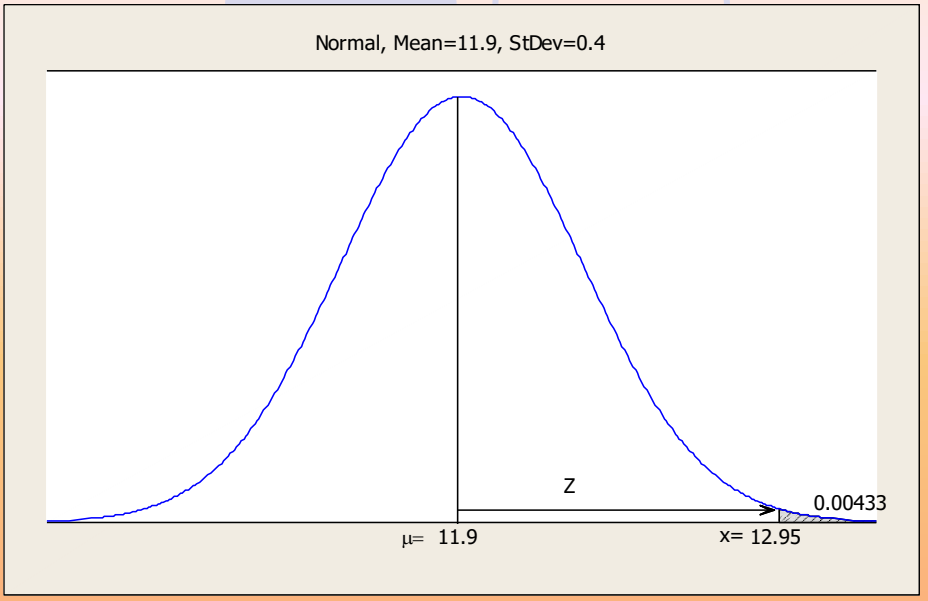
$$z = \frac{x-\mu}{\sigma} = \frac{11.0-11.90}{0.40} = -2.25 \Rightarrow 0.4878$$

The required percentage is

$$p(x<11.0) = 0.5 - 0.4878 = 0.0122$$

or, 1.22% of the workers earn less than $11.0.

[e] What percentage of the workers earns more than $12.95?



Find the area between \$11.90 and \$12.95 and ***subtract the area from 0.5***. The calculations are shown below.

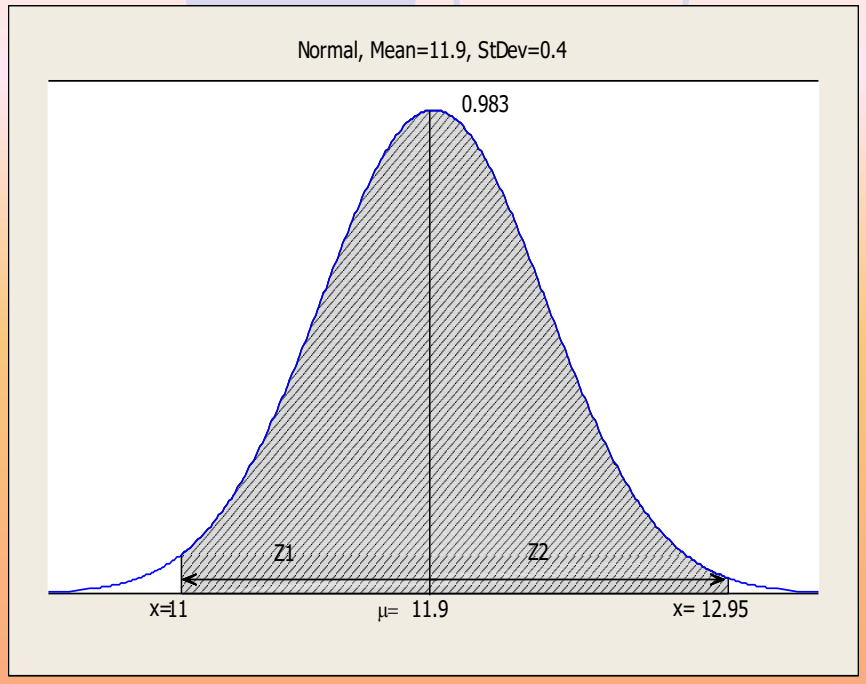
$$z = \frac{x-\mu}{\sigma} = \frac{12.95-11.90}{0.40} = 2.63 \Rightarrow 0.4957$$

The required percentage is

$$p(x > 12.95) = 0.5 - 0.4957 = 0.0043$$

or, 0.43% of the workers earn more than \$12.95.

[f] What percentage of the workers earns less than $11.0 or more than $12.95?



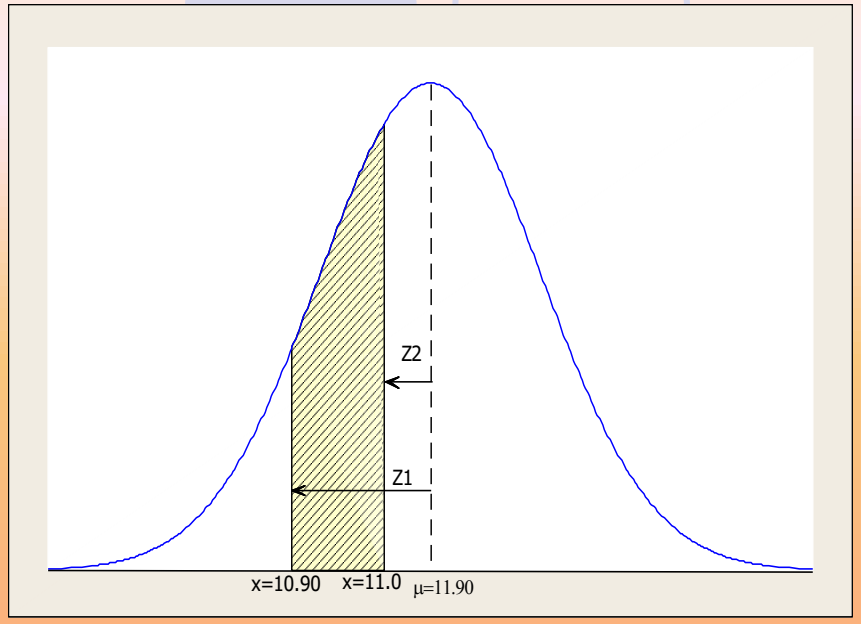
we need to find the areas for less than $11.0 and more than $12.95 and add them. We have determined these areas in part (d) and (e) . The required areas are shown in the figure .

From part (d) and (e) the percentage of earning less than $11.0 and more than $12.95 is 0.0122 and 0.0043. Therefore, the percentage of workers earning less than $11.0 and more than $12.0 is

$$p(x < 11 or x > 12.95) = 0.0122 + 0.0043 = 0.0165$$

or, 1.65% of the workers earn less than $11.0 or more than $12.95.

[g] What percentage of the workers earn between $10.90 and $11.00?



Find the areas between $11.90 and $10.90 and between $11.00 and $11.90, and ***subtract*** them.

The required percentage shown as the shaded area in the figure and can be evaluated as

$$z_1 = \frac{x-\mu}{\sigma} = \frac{10.90-11.90}{0.40} = -2.5 \Rightarrow 0.4938$$

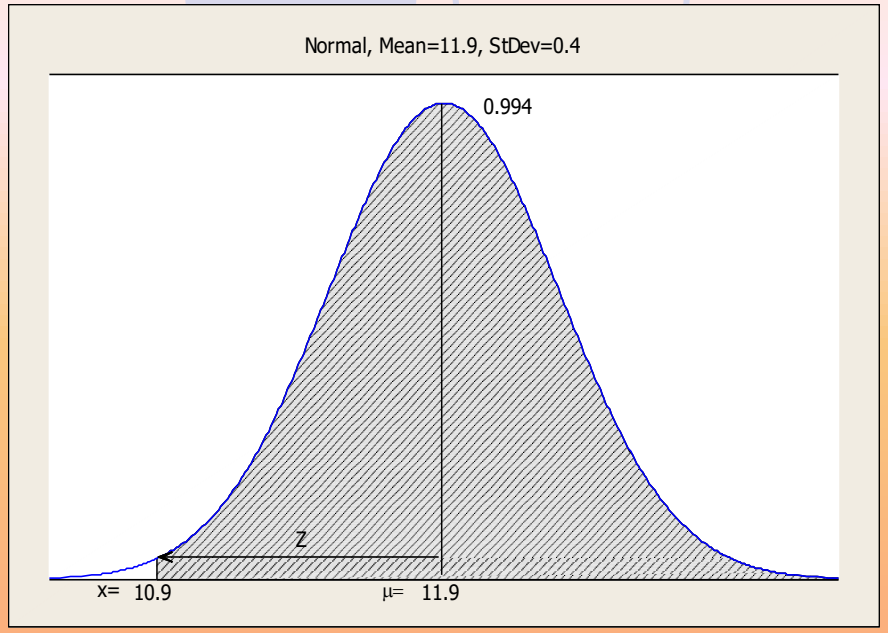
$$z_2 = \frac{x-\mu}{\sigma} = \frac{11.0-11.90}{0.40} = -2.25 \Rightarrow 0.4878$$

The required percentage is

$$p(10.90 \le x \le 11.00) = 0.4938 - 0.4878 = 0.006$$

or, 0.6 % of the workers earn between $10.90 and $11.00.

[h] What percentage of the workers make $10.90 or more?



Find the area between $11.90 and between $10.90 and ***add 0.5*** to this area.

The required percentage shown as the shaded area in the figure can be evaluated as:

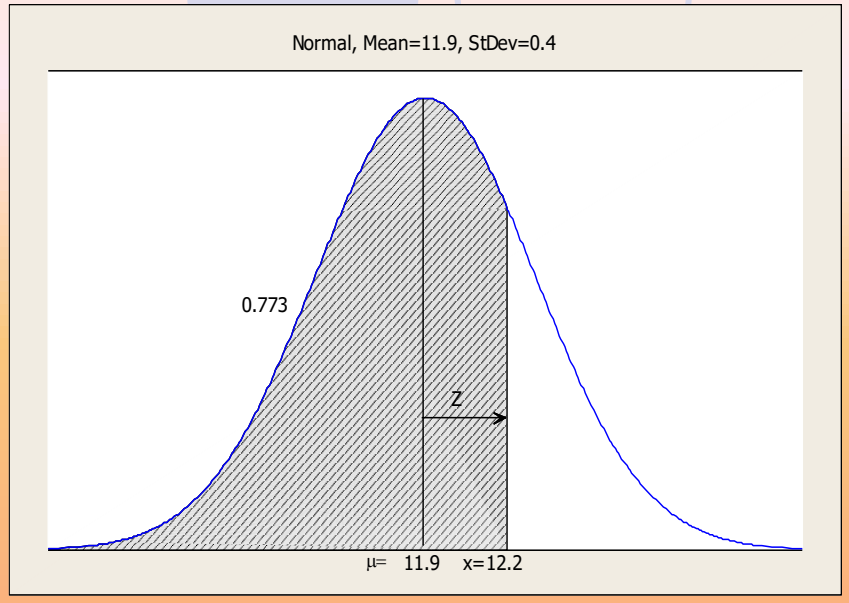
$$z = \frac{x-\mu}{\sigma} = \frac{10.90-11.90}{0.40} = -2.50 \Rightarrow 0.4938$$

The required percentage is

$$p(x \geq 10.90) = 0.4938 + 0.5 = 0.9938$$

Therefore, 99.38 % of the workers make $10.90 or more.

[h] What percentage of the workers earn $12.20 or less?



Find the area between $11.90 and $12.20 and ***add 0.5*** to this area. The calculations are shown below.

$$z = \frac{x-\mu}{\sigma} = \frac{12.20-11.90}{0.40} = 0.75 \Rightarrow 0.2734$$

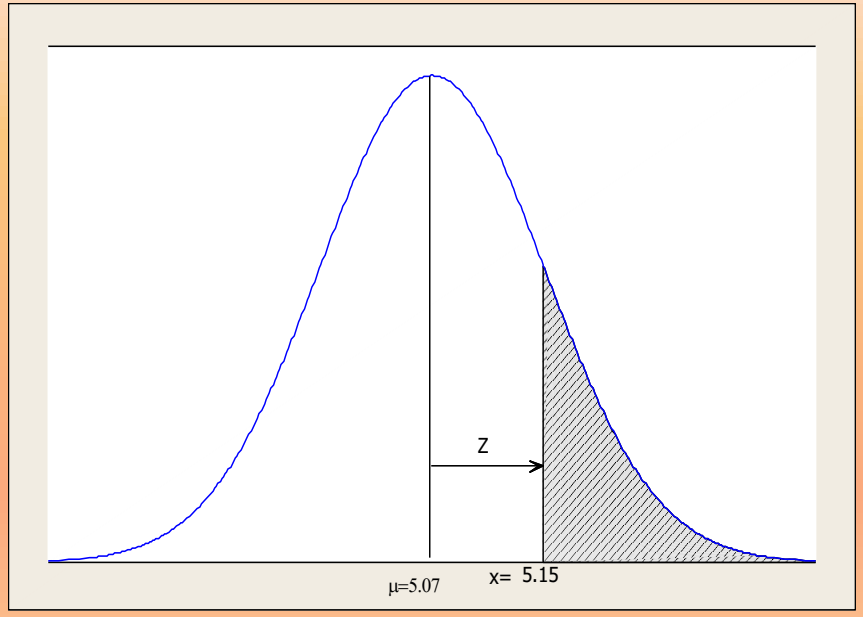
The required percentage is

$$p(x \le 12.20) = 0.5 + 0.2734 = 0.7734$$

This means that 77.34 % of the workers make $12.20 or less.

Example 3

The inside diameter of a piston ring is normally distributed with a mean of 5.07 cm and a standard deviation of 0.07 cm. What is the probability of obtaining a diameter exceeding 5.15 cm?



The required probability is the shaded area shown in the figure. To determine the shaded area, we first find the area between 5.07 and 5.15 using the z-score formula and then subtract the area from 0.5. See the calculations below.

$$z = \frac{x - \mu}{\sigma}$$

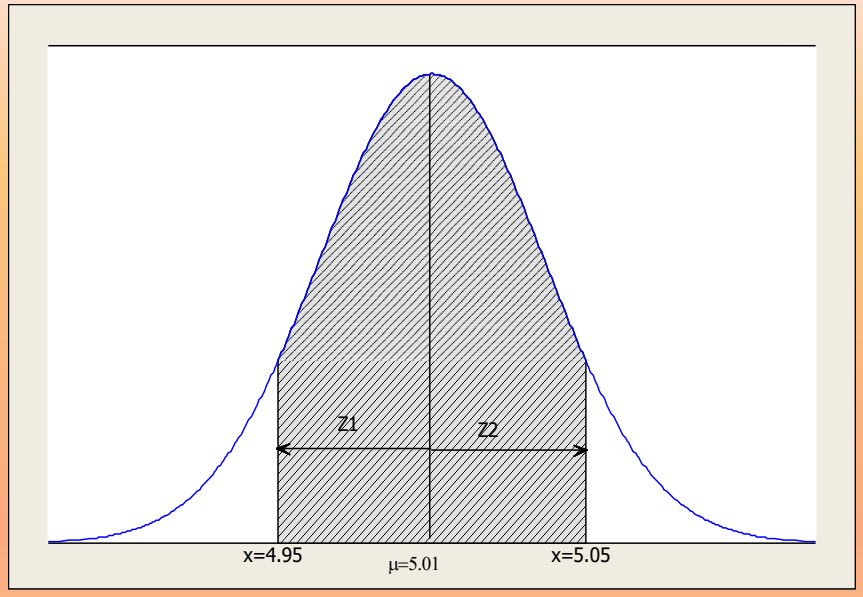
$$z = \frac{5.15 - 5.07}{0.07} = 1.14 \rightarrow 0.3729$$

The required probability is $p(x \geq 5.15) = 0.5 - 0.3729 = 0.1271$

or, there is 12.71% chance that piston ring diameter will exceed 5.15 cm.

Example 4

The measurements on certain type of PVC pipes are normally distributed with a mean of 5.01cm and a standard deviation of 0.03 cm. The specification limits on the pipes are 5.0±0.05 cm. What percentage of the pipes is not acceptable?



The percentage of acceptable pipes is the shaded area shown in the figure.

The required area or the percentage of acceptable pipes is explained below

$$z_1 = \frac{x-\mu}{\sigma} = \frac{4.95-5.01}{0.03} = -2.0 \Rightarrow 0.4772$$

$$z_2 = \frac{x-\mu}{\sigma} = \frac{5.05-5.01}{0.03} = -1.33 \Rightarrow 0.4082$$

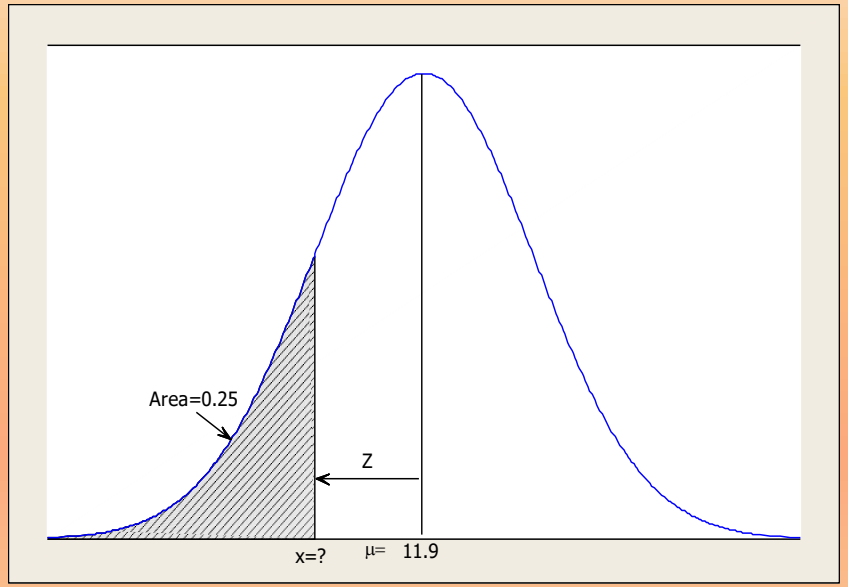
The area 0.4772 is the area between the mean 5.01 and 4.95 (see the figure).
The area left of 4.95 is 0.5 - 0.4772 = 0.0228.
The area 0.4082 is the area between the mean 5.01 and 5.05 (see the figure). The area right of 5.05 is 0.5 - 0.4082 = 0.0918.

Therefore, the percentage of pipes not acceptable = 0.0228+0.0918 =0.1146 or 11.46%.

Using The NORMAL or Z-table In Reverse

Example 5

Consider example 2 in which the hourly wages were normally distributed with a mean μ = \$11.90 and standard deviation σ = \$0.40. Suppose we want to determine the first quartile or the 25th percentile value of the hourly wage. We can determine this as shown below.



The first quartile is the 25th percentile value of the wages. If we want to determine this value, we want to determine x in the z-formula below when mean μ = \$11.90 and standard deviation σ = \$0.40 (see the Figure).

$$z = \frac{x - \mu}{\sigma}$$

Note: Z is the distance from the mean to the point of interest. If we know the area from the mean to x (which is 0.25); we can find the value for this area of 0.25 from the normal table. In the normal table, locate the area closest to 0.25. The closest value to 0.25 is 0.2486 (not exceeding 0.25). See part of the Table in the next slide.

Finding Z for an area of 0.25

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
:									
0.6								0.2486	
:									
:									

the z value for the area of 0.25 is 0.67. Note that the z value will be negative because x lies to the left of the mean. Now, we have all the values to determine x . We can solve for as shown below.

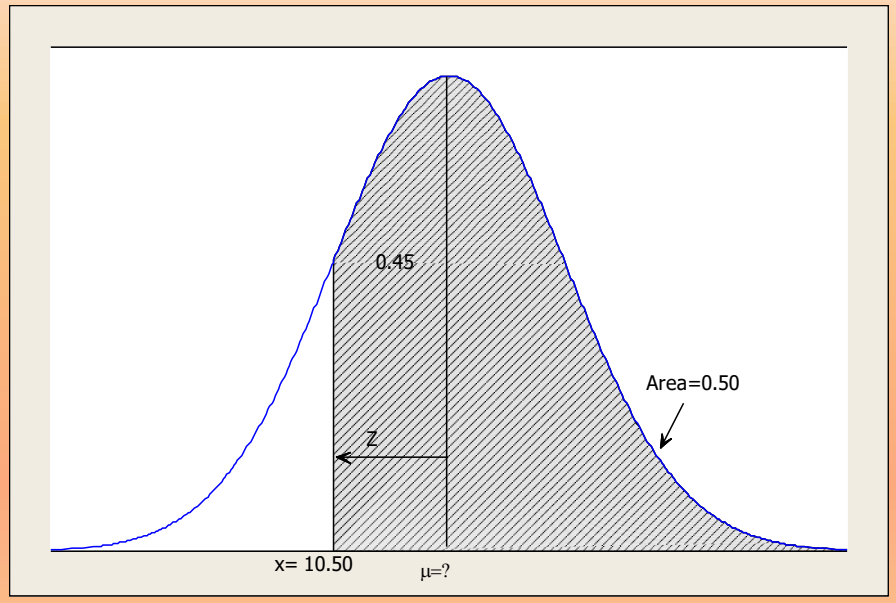
$$-0.67 = \frac{x - 11.90}{0.40}$$

Solving this equation for , we get = 11.632.

Therefore, the first quartile value for the wages is $11.63.

Example 6

Suppose that in example 2, the mean wage is unknown but we know that 95% of the workers make $10.50 or more. Determine the mean wage if the standard deviation for the wage (σ) is known to be $0.40 from the past experience.



The problem is explained in the figure.

To determine the mean (μ) using the z - score formula, we first determine the Z-value corresponding to the area of 0.45. From the normal table, this value is

z = -1.64 or, -1.65

Using this Z-value, we can solve for the mean μ from the formula below

$$z = \frac{x - \mu}{\sigma}$$

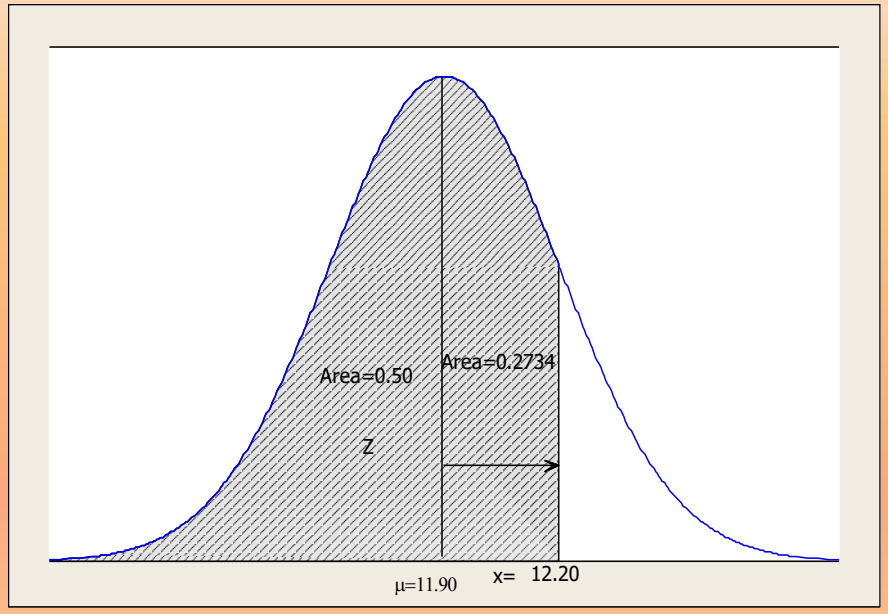
$$-1.65 = \frac{10.50 - \mu}{0.40} \; or,$$

$$\mu = 11.16$$

Therefore, the mean wage m is $11.16

Example 7

Suppose that in example 2, the standard deviation (σ) is unknown but we know that 77.34% of the workers make \$12.20 or less. Determine the standard deviation (σ) of the wages if the mean wage μ is known to be \$11.90. The problem is described in the figure below.



To determine σ using the Z-score formula, we must determine the z-value corresponding to the area of 0.2734. From the normal table, this value is z= 0.75. Using this -value, we can solve for σ as shown below.

$$z = \frac{x - \mu}{\sigma}$$

$$0.75 = \frac{12.20 - 11.90}{\sigma}$$

$$\sigma = 0.4$$

Checking for Normal Data

1. Construct a histogram or stem-and-leaf of the data. The shape of histogram and stem-and-leaf plots will resemble a normal curve if the data are normal or approximately normal.

2. Calculate the mean, median, and mode (if appropriate) of the data. If these measures are approximately equal then the data are symmetrical or approximately normal.

3. Calculate the mean and standard deviation and the intervals $\bar{x} \pm 1s, \bar{x} \pm 2s$ and $\bar{x} \pm 3s$.

   If the data are normal, the percentages of observations for these intervals would be approximately 68%, 95%, and 99.7% respectively.

Checking for Normal Data…cont.

4. Construct a box plot of the data using the five measure summary: minimum, Q1, Q2, Q3, and the maximum. If the data are normal, the distance from minimum data value to Q1 and Q3 to maximum data value will be approximately equal. In addition, the median (Q2) will divide the box in approximately two equal halves.

5. Calculate the interquartile range (IQR) and the standard deviation, s of the sample data. If the ratio

$$IQR / s \approx 1.3$$

then the data are normal or approximately normal.

6. Construct a normal probability plot of the data. If the data are normal or approximately normal, the points on the probability plot will fall on a straight line.

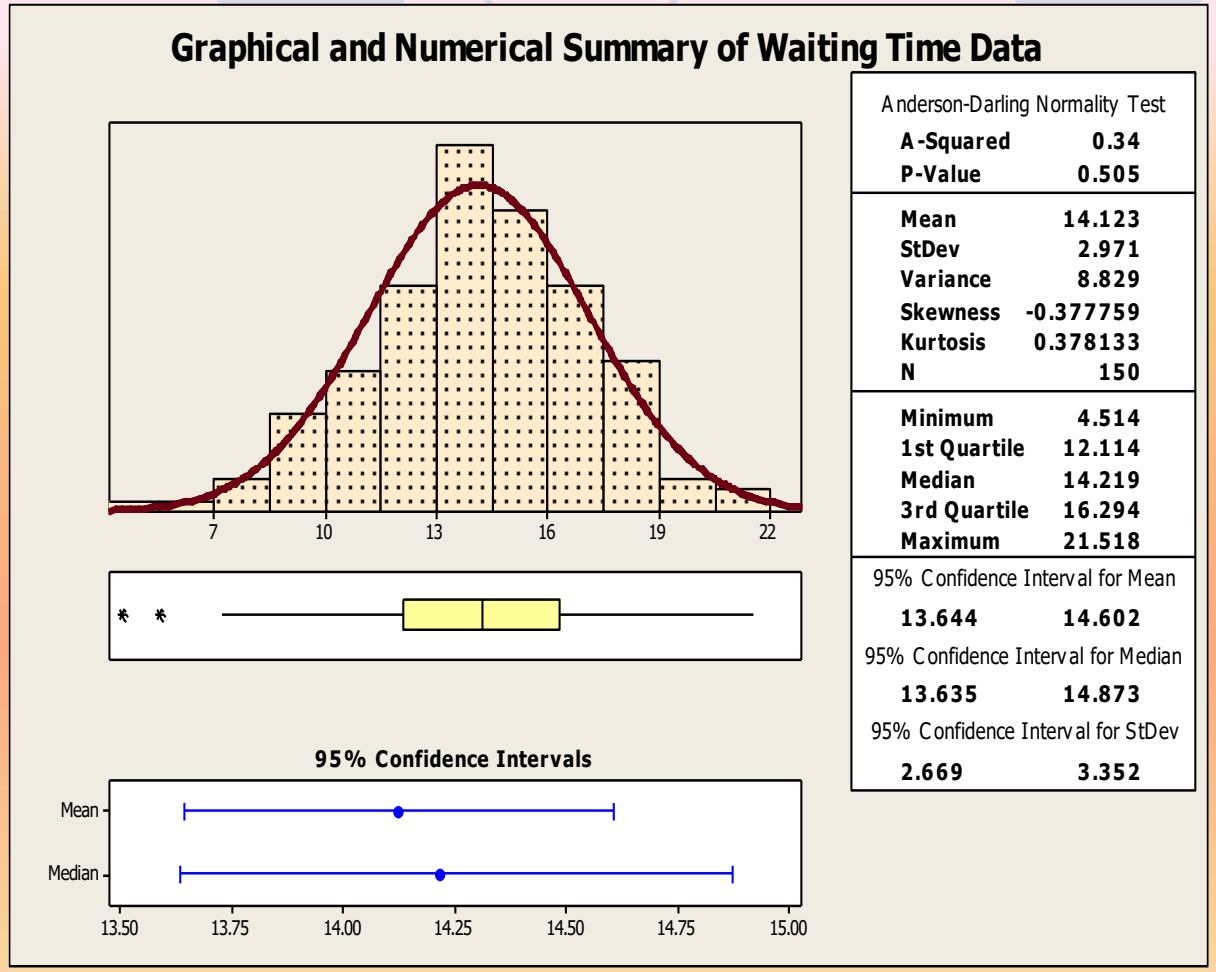
Checking for Normal Data...Example

The waiting time of patients at a hospital emergency service collected the data shown in the table below. Use the six steps outlined in the previous two slides to check if the waiting time data are normal. The distribution of the waiting time data is of interest to draw certain conclusions.

Waiting time (in minutes)

15.9	18.0	16.8	11.7	13.9	19.0	17.6	19.6	11.9	9.8	7.6
14.7	13.8	13.0	15.7	15.6	11.1	7.8	17.2	14.4	14.5	16.4
7.2	15.7	14.3	17.2	17.6	13.6	15.2	13.7	16.5	11.3	10.7
12.9	15.9	12.4	9.7	17.8	14.9	14.8	15.6	13.3	15.2	11.6
13.4	14.2	13.6	18.2	13.1	5.5	15.5	11.9	14.2	9.0	14.0
15.6	15.7	9.3	12.1	13.7	17.2	13.5	16.8	16.3	12.9	18.0
15.0	17.8	11.4	15.0	10.8	17.4	12.7	12.6	21.1	12.3	13.5
9.2	18.3	13.1	16.4	12.0	19.1	16.9	18.8	9.5	12.1	14.4
14.2	13.5	13.5	10.9	11.8	12.4	11.2	14.6	14.4	13.9	14.9
15.0	10.8	18.0	13.6	17.1	15.3	12.1	17.0	11.4	15.3	10.0
18.5	18.6	16.2	16.5	18.3	10.3	13.8	12.8	13.2	11.9	13.7
9.5	16.8	10.1	15.7	15.2	18.2	4.5	13.5	10.5	13.4	10.5
16.7	11.8	15.3	14.8	15.5	15.2	9.0	14.2	13.4	16.0	16.7
14.1	16.7	13.8	15.9	12.8	21.5	16.4				

Graphical and Numerical Summary of Waiting Time using MINITAB



Anderson-Darling Normality Test	
A-Squared	0.34
P-Value	0.505
Mean	14.123
StDev	2.971
Variance	8.829
Skewness	-0.377759
Kurtosis	0.378133
N	150
Minimum	4.514
1st Quartile	12.114
Median	14.219
3rd Quartile	16.294
Maximum	21.518
95% Confidence Interval for Mean	
13.644	14.602
95% Confidence Interval for Median	
13.635	14.873
95% Confidence Interval for StDev	
2.669	3.352

The histogram of the data in indicates that the shape very closely resembles a bell shape or normal distribution.

The normal curve superimposed over the histogram shows that the data are symmetric or normal centered around the mean.

Thus, we can conclude that the data follow a normal distribution.

The stem-and-leaf plot of the data shows that the shape is very close to the normal distribution.

Stem-and-leaf N = 150 Leaf Unit = 0.10

```
 1   4  5
 2   5  5
 2   6
 5   7  167
 6   8  9
14   9  02345679
22   10 13456788
34   11 123345778899
48   12 00012335777889
71   13 012344444455566667778889
(18) 14 012222333446778889
61   15 0012223334555666678889
40   16 022344446667888
25   17 0011235588
15   18 0001123557
5    19 005
2    20
2    21 05
```

Check #2

The values of mean and the median in (Figure – slide 40) are 14.123 and 14.219. If the data are symmetrical or normal, the values of the mean, median are very close. Since the mean and median for the waiting time data are very close, it indicates that the distribution is symmetrical or normal.

Check #3

For the waiting time data

$$\bar{x} \pm 2s \quad \bar{x} = 14.12 \qquad s = 2.97.$$

Table below shows the percentages for the waiting time data between the mean and ± one, two and three standard deviations.

Interval	Percentage in Interval
$\bar{x} \pm s$	69.3
$\bar{x} \pm 2s$	95.3
$\bar{x} \pm 3s$	99.3

The results show that the empirical rule holds. Thus, we can conclude that the waiting time data follows a normal distribution.

Check #4

The box plot of the data in Figure in slide on page 40 shows that the waiting time data very closely follows a normal distribution.

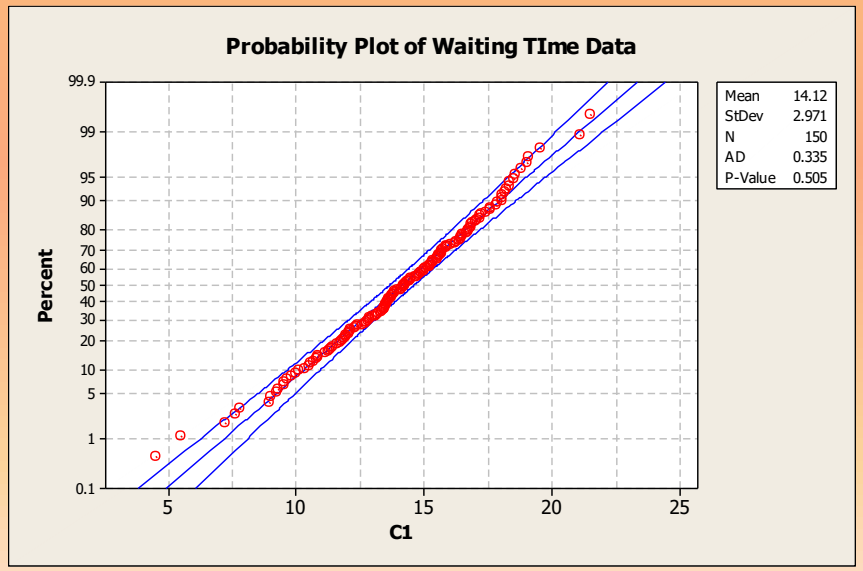
Check #5

The ratio of the interquartile range (IQR) to the standard deviation is calculated below. The values are obtained from the slide on page 40

$$\frac{IQR}{s} = \frac{Q3 - Q1}{s} = \frac{16.294 - 12.114}{2.971} = 1.41$$

The value is close to 1.3 indicating that the data are approximately normal.

Check #6



The normal probability plot of waiting time data shows that the data are very close to a normal distribution.

CALCULATING NORMAL PROBABILITIES USING EXCEL

The normal probabilities can be calculated using NORMDIST function in Excel. The general form of the function is

NORMDIST (x, μ, σ , cumulative)

For the last argument "cumulative," TRUE is specified if a cumulative probability is desired. We will demonstrate the probability calculations using the NORMDIST function. Consider the examples below.

***Example:** The hourly wages for part-time workers with a minimum of two years experience in retail sector are normally distributed with a mean m = $11.90 and a standard deviation s = $0.40.*

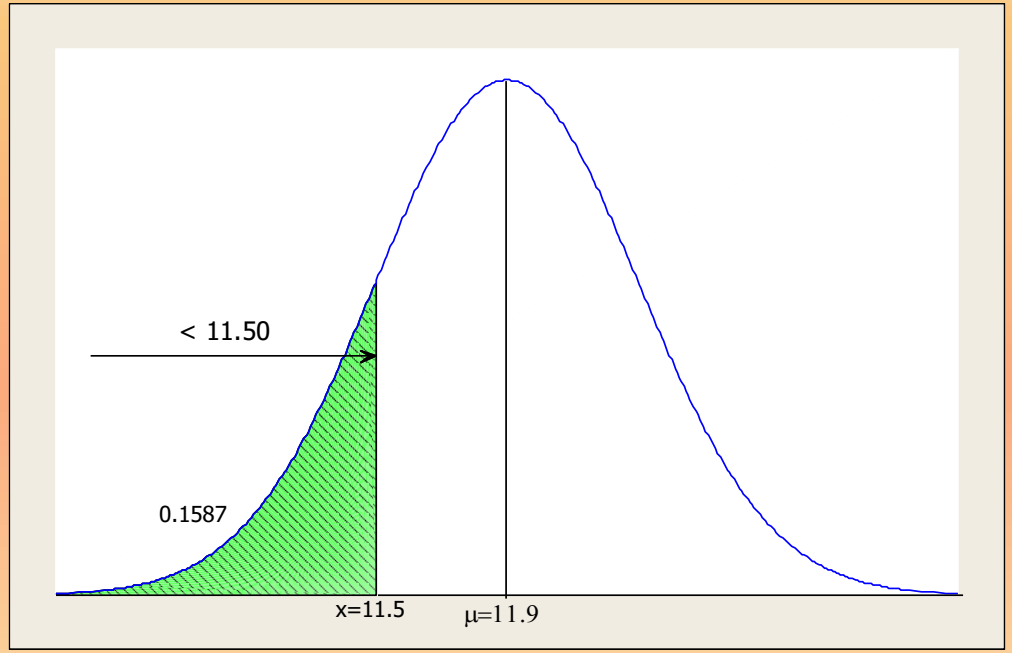
[a] Find the probability that a worker earns less than $11.50?

The required probability that corresponds to the shaded area is shown in the figure below.

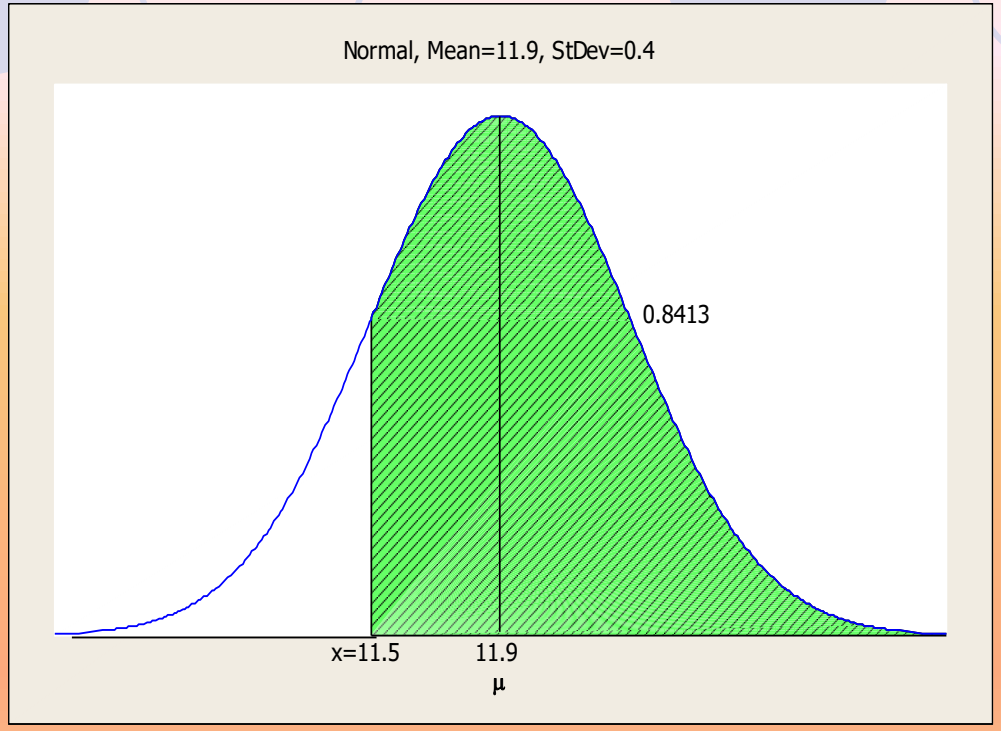
To compute the probability of less than or equal to $11.50, enter the following function in any cell of an Excel worksheet.

=NORMDIST(11.50,11.90,0.40,TRUE)

Hit the enter key after entering the formula and the value 0.158655 will appear in the cell where you entered the formula. This means that the probability that a worker earns $11.50 or less is 0.1587. This is also shown in the figure below.



[b] Find the probability that a worker earns $11.50 or more?



The probability is the shaded area in the figure above. From part [a] we know that the probability of earning less than or equal to $11.50 is 0.1587 or 15.87%. Therefore, the probability of earning $11.50 or more is (1-0.1587) = 0.8413. This is shown in the figure above.

[c]Find the probability that worker earns between $11.50 and $12.40?

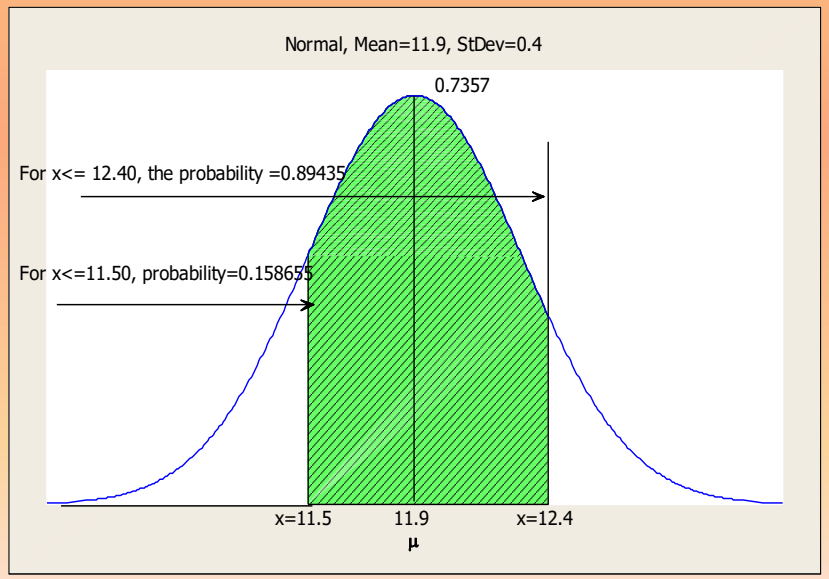
To compute the probability up to $12.40, enter the following function in any cell of an Excel worksheet.

=NORMDIST(12.40,11.90,0.40,TRUE)	(B)

Next, calculate the probability up to $11.40 by entering the following function in any cell of an Excel worksheet.

=NORMDIST(11.50,11.90,0.40,TRUE)	(C)

Formula (B) will return a value of 0.89435 and formula (C) will return the value 0.158655. Therefore,



the probability that worker earns between $11.50 and $12.40=0.89435-0.158655= 0.7357

IMPORTANT NOTES ON PROBABILITY DISTRIBUTIONS

The continuous distributions are defined by their probability density functions; whereas the discrete distributions are defined by their probability mass functions. The density and mass functions depend on one or more parameters.

Continuous distributions can assume different shapes and sizes depending on the values of the parameters. There are three basic types of parameters:

Shape parameter: *controls the basic shape of the distribution. In some distributions, changing the shape parameter will cause major changes in the shape or form of the distribution. In others, changing the shape parameter may not cause major change in the shape and form of the distribution.*

Scale parameter: *controls the unit of measurement within the range of the distribution. A change in the scale parameter either contracts or expands the distribution along the horizontal axis.*

Location parameter: *specifies the location of the distribution relative to zero on the horizontal axis. The location parameter may represent the midpoint or the lower endpoint of the range of the distribution.*

Note that all distributions may not have all three parameters. Some distributions may have more than one shape parameter. It is important to understand the effects of these parameters for successful application and use of the distributions in data analysis.

THE EXPONENTIAL DISTRIBUTION

The exponential distribution has wide applications in modeling.

In the previous chapter we discussed the Poisson distribution which is often used to describe the number of arrivals (or occurrences) over a specified time period.

The exponential distribution is used to describe such phenomenon as the time between failures of components, the time between arrivals of customers or telephone calls, or the lifetime of certain types of components in a machine. This distribution is widely used in reliability engineering to describe the time to failure of certain types of components

RELATIONSHIP BETWEEN THE POISSON AND THE EXPONENTIAL DISTRIBUTION

If X is the random variable that represents the ***number*** of arrivals over a specified period T, then X is said to follow a Poisson distribution, and if Y represents the ***time between successive arrivals***, then Y will follow an exponential distribution. Thus, the Poisson and exponential distributions are closely related.

The exponential distribution is an appropriate model to use when ***the failure rate is constant***.

PROBABILITY DENSITY OF AN EXPONENTIAL DISTRIBUTION

If the random variable x follows an exponential distribution then the probability density function is given by:

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$ where, $x > 0$ and $\mu > 0$ (A)

Cumulative Probabilities for exponential distribution is given by:

$$P(x \le x_0) = 1 - e^{-x/\mu}$$ for $x > 0$ (B)

The ***mean*** and ***standard deviation*** of the exponential distribution are equal and given by:

Mean = μ (c)

Standard deviation, $\sigma = \mu$

The parameter $1/\mu$ in equation (A) is often referred to as the failure rate (time between failures) and is related to the Poisson distribution.

CHARACTERISTICS OF THE EXPONENTIAL DISTRIBUTION

- *The number of arrivals per unit time in the Poisson distribution and the time between arrivals in the exponential distribution can both be used to describe the same thing.*

- *For example, if the number of arrivals per unit time follows a Poisson distribution with mean or average of 10 arrivals per hour, then we can say that the time between arrivals is exponentially distributed with mean time between arrivals being 1/10= 0.1 hour or 6 minutes.*

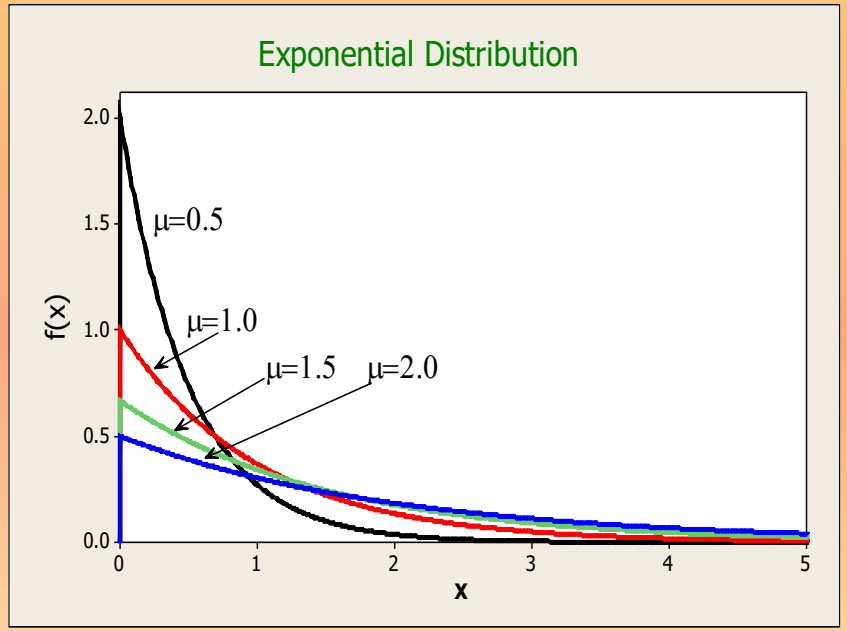
- *Unlike the normal distribution which is described by its location and shape parameters (μ and σ respectively), exponential distribution is described by only one parameter, μ . Each value of μ determines a unique exponential distribution.*

- *The exponential distribution has no shape or location parameter; it is described by a scale parameter which is ($1/\mu$) .*

Investigating the Exponential Distribution

Objective: ***Investigate the general shape of the exponential distribution and observe how the shape of the distribution changes as we change the characteristic scale parameter,(μ) of the distribution.***

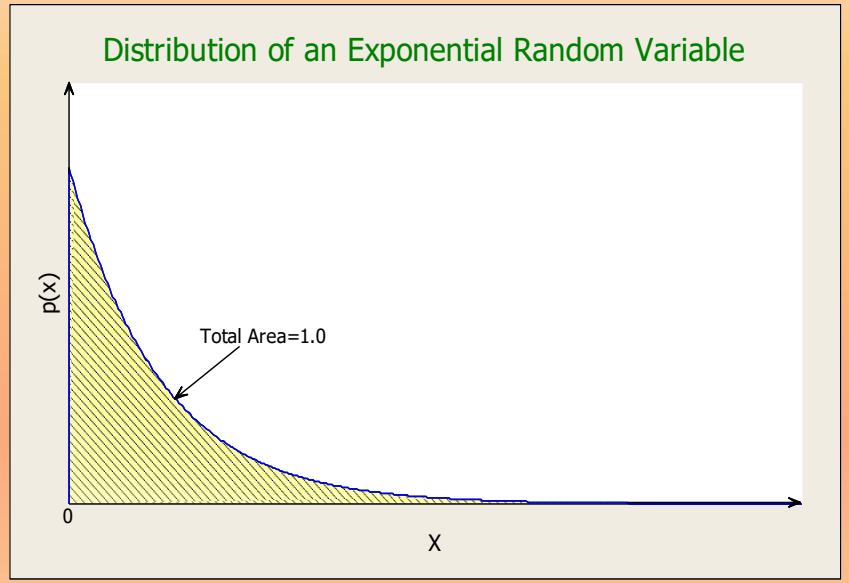
Figure below shows the plot of the density functions of the exponential distribution for different values of $(\mu = 0.5, 0.67, 1.0, 2.0)$



The exponential distribution curve steadily decreases as the value of the random variable x increases. Larger the value of x, the probability of observing a value of at least this large decreases exponentially. Note also that the distribution is not symmetrical and unlike the normal random variable, the exponential random variable is always greater than zero.

Finding Exponential Probabilities

The probabilities for exponentially distributed random variables are found by evaluating the areas between the points of interest of the exponential curve described in figure below.



Suppose X is an exponentially distributed random variable with parameter μ then

$$P(X \geq x) = e^{-x/\mu} \qquad \text{for } x \geq 0$$

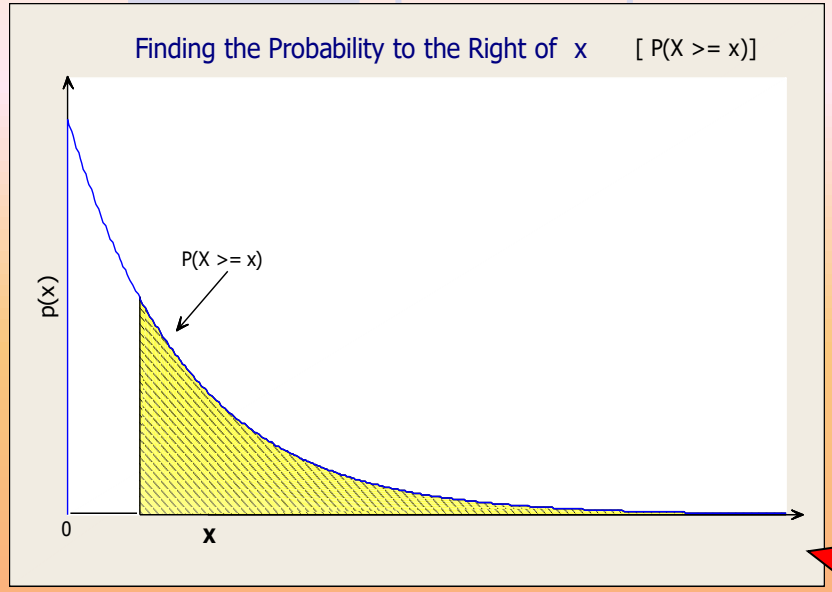
$$P(X \leq x) = 1 - e^{-x/\mu} \qquad \text{for } x > 0$$

$$P(x_1 \leq X \leq x_2) = e^{-x_1/\mu} - e^{-x_2/\mu}$$

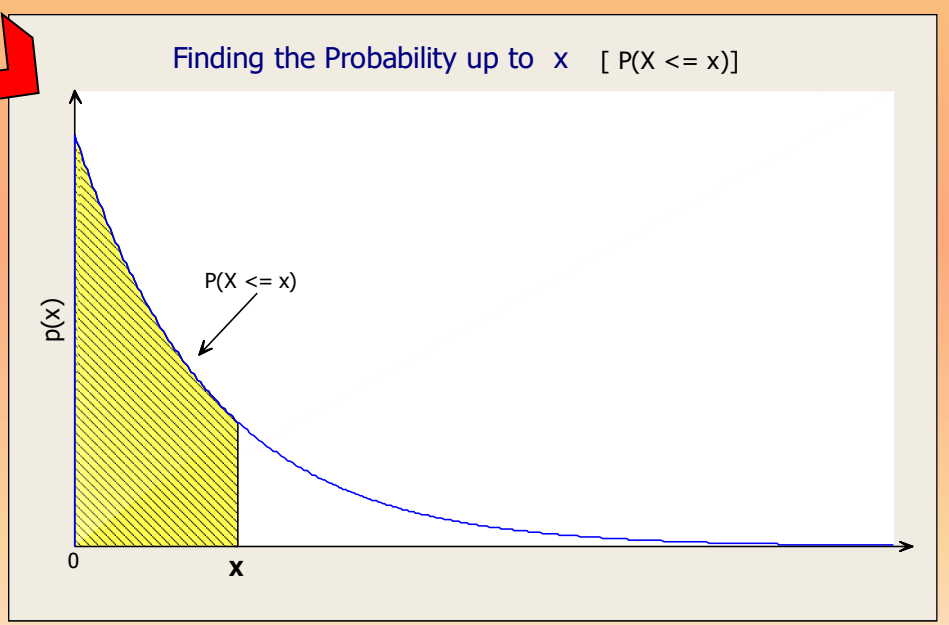
for $x_1, x_2 > 0$

The above equations are used to find the probability between the points of interest in exponential distribution. Figure above shows the total area of the distribution of an exponential random variable.

Figure below demonstrates the probability, P(X ≥ x) for an exponential random variable.



$$P(X \geq x) = e^{-x/\mu} \quad \text{for } x \geq 0$$



The shaded probability is P(X ≤ x) for an exponential random variable

$$P(X \leq x) = 1 - e^{-x/\mu}$$

for $x > 0$

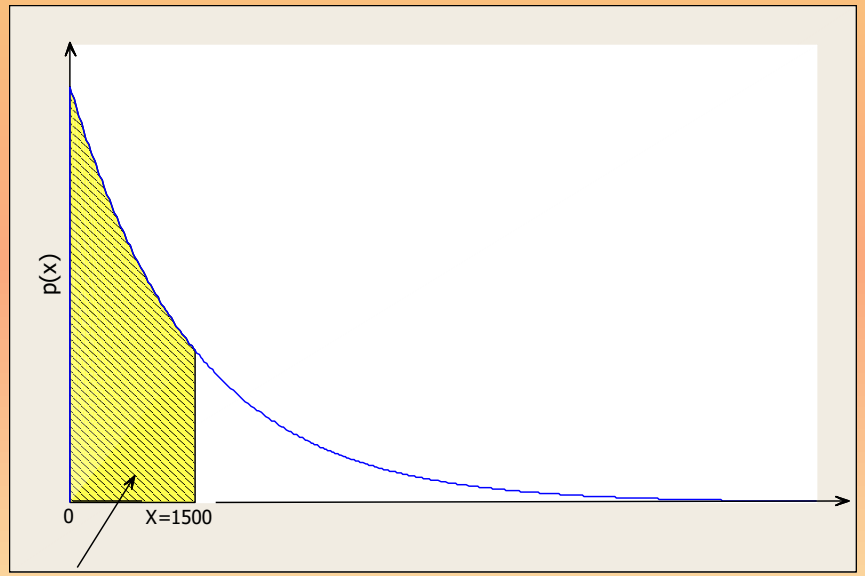
Example 1

The useful life of an electronic component is described by the exponential distribution with a mean life of 1500 days.

[a] Find the probability that the component will fail before its expected life of 1500 days.

Solution:

(a) Note that the life of the component (X) follows an exponential distribution with μ =1500. The probability to be evaluated is shown in the Figure 6.42. From this figure, we can see that



$$p(X \le x) = 1 - e^{-x/\mu}$$

$$p(x < 1500) = 1 - e^{-1500/1500} = 0.632$$

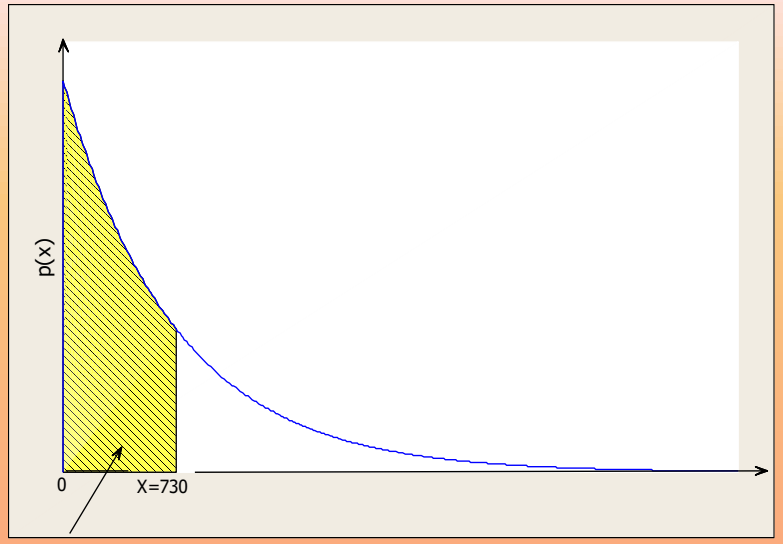
There is a 63.2% chance that the component will fail within 1500 hours.

$$p(x < 1500) = 1 - e^{-1500/1500} = 1 - e^{-1} = 0.632$$

Example 1…cont.

[b] The component has 2 years or 730 days of warranty. What percentage of the components will fail during the warranty period?

We want to find the probability, P(x ≤ 730). This probability is shown figure.



$$p(x \le 730) = 1 - e^{-730/1500} = 1 - e^{-0.4867} = 1 - 0.6147 = 0.3853$$

The above probability indicates that the manufacturer will have to replace approximately 38.5% of the components during the warranty period.

$$p(x \le 730) = 1 - e^{-730/1500} = 1 - e^{-0.4867} = 1 - 0.6147 = 0.3853$$

Note that the average life of the components is approximately four years. But this high percentage (38.5%) of the components failure is due to the fact that the exponential distribution is positively skewed and there is high concentration of probability on the lower end of the distribution.

Example 2

The time interval between successive failures of air conditioning equipment follows an exponential distribution. If the mean time to failure is 400 hours,

(a) What is the probability that the air conditioning equipment will fail after 300 hours of operation?

(b) What is the probability that the equipment will last 375 hours or less?

Solution:

(a) $p(x > 300) = e^{-x/\mu} = e^{-300/400} = 0.4724$

(b) $p(x \le 375) = 1 - e^{-375/400} = 1 - e^{-0.9375} = 1 - 0.3916 = 0.6084$

The exponential probabilities require evaluating $e^{-\mu}$ ***. A comprehensive table of m for different values of*** μ ***is provided in the appendix. This table can be used to evaluate exponential probabilities).***

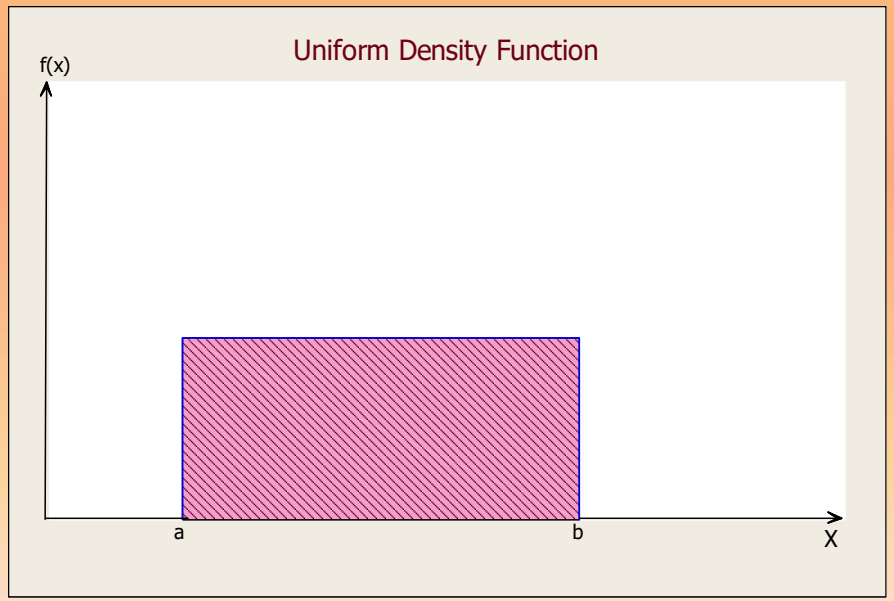
Uniform Distribution

A random variable for which all outcomes between some minimum and maximum values have equal probability of occurrence may be described by a uniform distribution. The density function of the uniform distribution is given by

$$f(x) = \frac{1}{b-a} \qquad a \le x \le b$$

$$= 0 \qquad \text{otherwise}$$

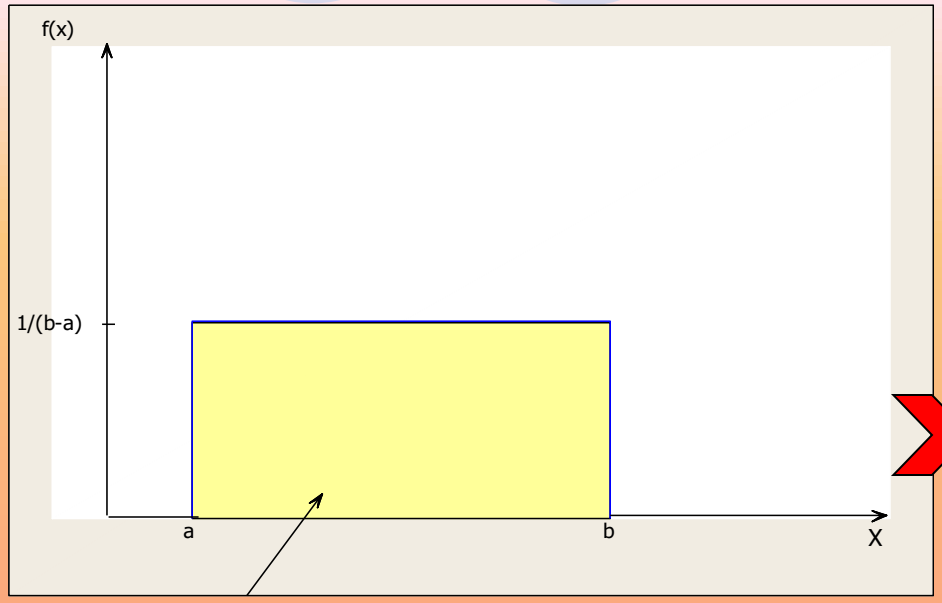
The density function is shown below.



Note that ***a*** and ***b*** are two parameters of the uniform distribution where ***a < b***. Parameter ***a*** is the location parameter and it controls the location of the distribution along the x -axis. The scale parameter is the difference ***(b-a)***. An increase in the difference ***(b-a)*** will elongate the distribution whereas; a decrease in the difference ***(b-a)*** will compress it.

Finding Uniform Probabilities

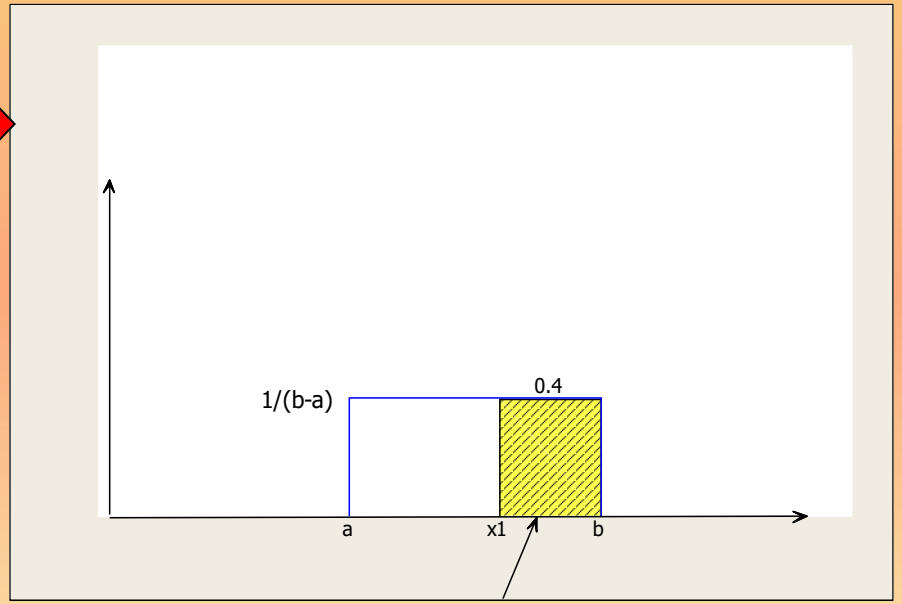
Total probability is unity



Total shaded area= $(b-a)\left[\frac{1}{(b-a)}\right]$

Evaluate the probability when X is greater than or equal to a certain value x_1, that is

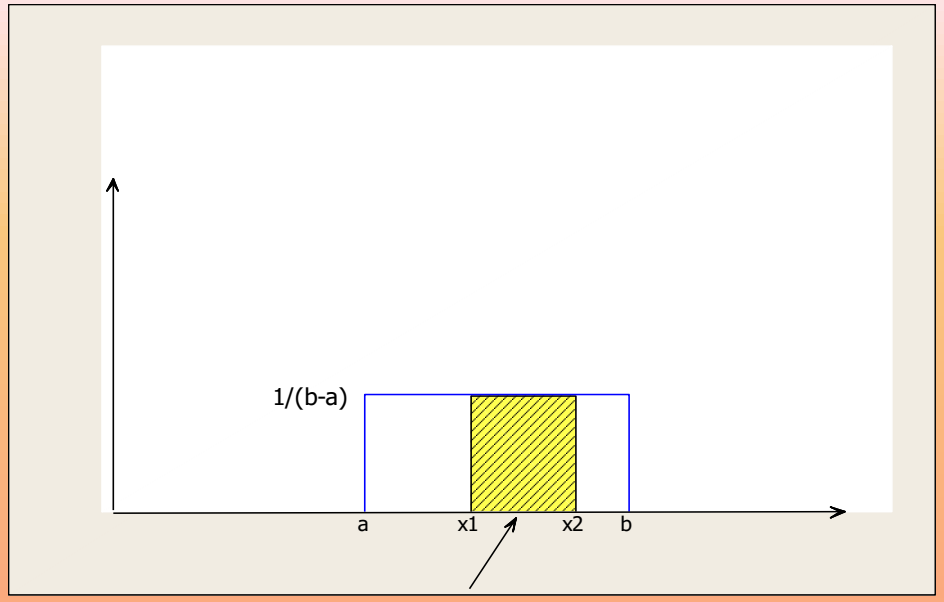
$$P(X \geq x_1)$$



Shaded Area = $(b-x_1)\left[\frac{1}{(b-a)}\right]$

Evaluating the probability when X is between the values x_1 and x_2 or,

$$P(x_1 \le X \le x_2)$$



Area = $(x_2 - x_1)\left[\frac{1}{(b-a)}\right]$

The mean or the expected value and the variance of the uniform distribution are

$$E(x) = \mu = \frac{(a+b)}{2}$$

$$V(x) = \sigma^2 = \frac{(b-a)^2}{12}$$

Example 1

The random variable x is uniformly distributed between 150 and 200.

(i) Calculate the mean and standard deviation of x. Sketch the graph of the distribution and show the mean on the graph.

(i) Evaluate the following probabilities:

(a) $p(x < 170)$

(b) $p(160 \le x \le 190)$

(c) $p(x = 175)$

(d) $p(162 < x < 182)$

Solution:

(i) Mean and standard deviation

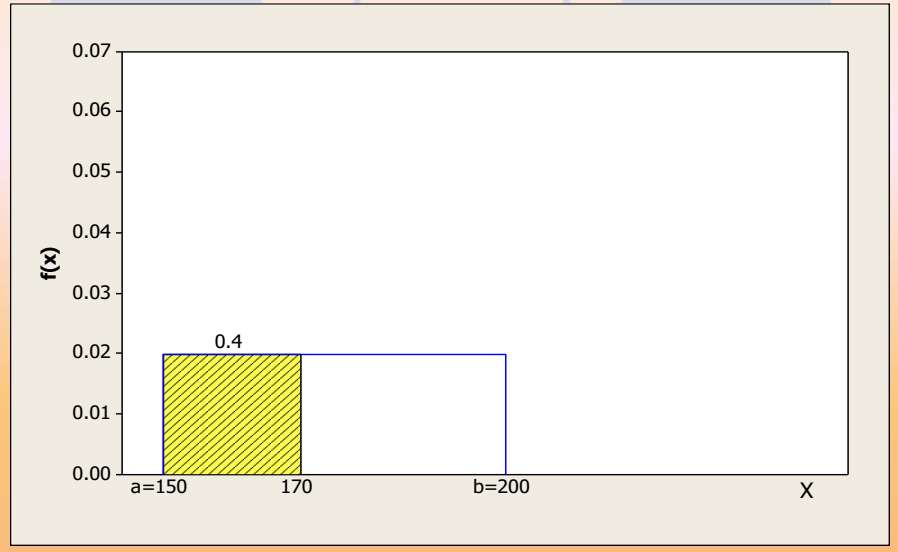
$$E(x) = \mu = \frac{(a+b)}{2} = \frac{150+200}{2} = 175$$

$$V(x) = \sigma^2 = \frac{(b-a)^2}{12} = \frac{(200-150)^2}{12} = 208.33$$

$$\sigma = \sqrt{\sigma^2} = 14.43$$

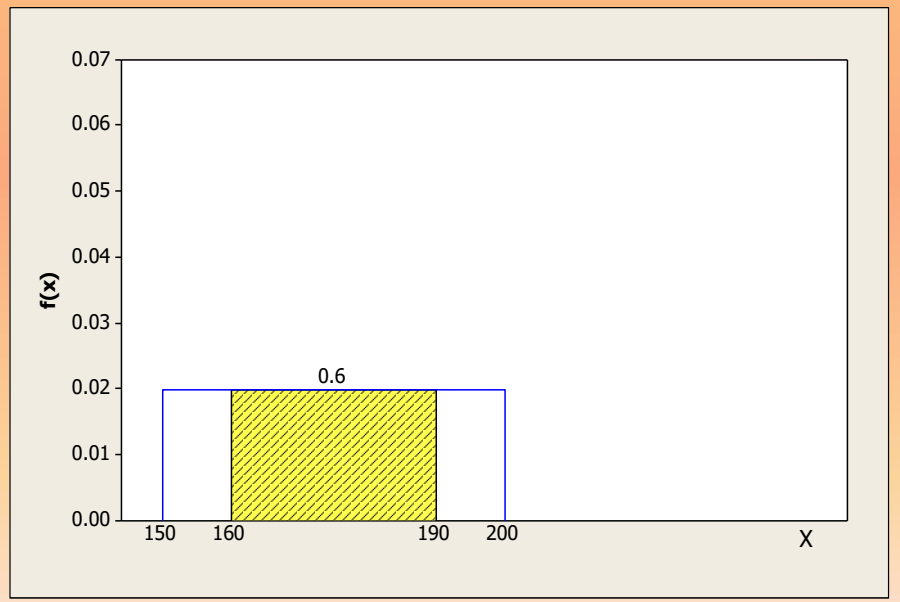
$$(a)\, p(x<170) = p(150<x<170) = \frac{170-150}{200-150} = 0.4$$

The shaded region in the figure below shows the probability.



$$(b)\, p(160 \le x \le 190) = \frac{190-160}{200-150} = 0.6$$

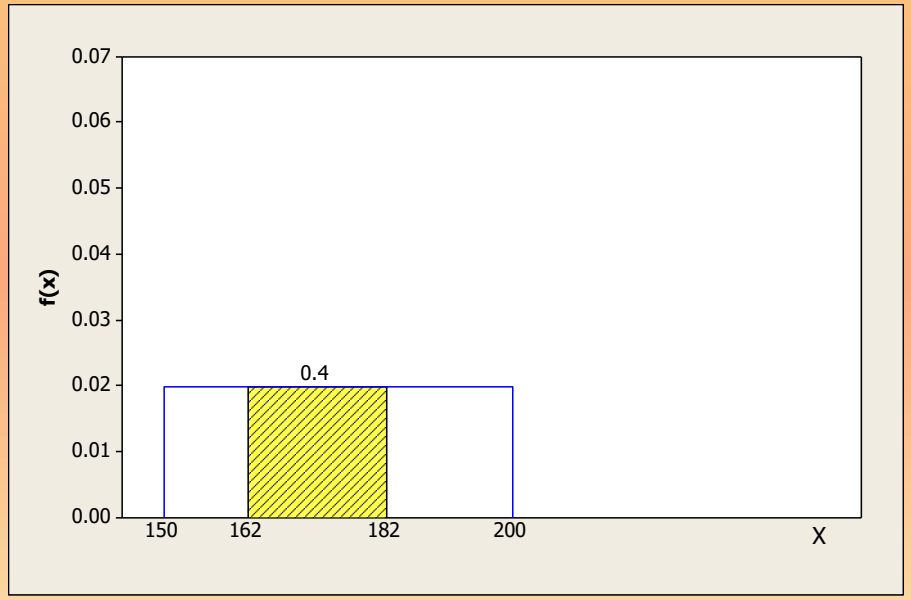
The shaded region below shows the probability.



$$(c)\ p(x = 175) = 0$$

The probability of any single point is zero as the area under the curve above any single point is zero.

$$(d)\ p(162 < x < 182) = \frac{182 - 162}{200 - 150} = 0.40$$ (see the figure below).



Random Variables and Continuous Probability Distribution

Random Variables and Probability Distribution

Continuous Probability Distributions

Random Variables and Probability Distribution

A random variable is a numerical quantity whose value is determined by chance. A random variable must be a numerical quantity. The relationship between the values of a random variable and their probabilities is summarized by a probability distribution. Probability distributions are characterized by:

The probability density function: the probability density function, f(x), describes the behavior of a random variable and may be viewed as the shape of the distribution. The probability density function represents the entire sample space; therefore, the area under the probability density function must equal one.

$$\int_{-\infty}^{\infty} f(x) = 1$$

The cumulative distribution function: the cumulative distribution function, F(x), denotes the area beneath the probability density function to the left of x.

$$F(x) = \int_{-\infty}^{x} f(r)dr$$

Normal Distribution

To calculate the normal probability, $p(x_1 \le X \le x_2)$ where X is normal with parameters μ and σ, we need to evaluate:

$$\int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

The normal distribution with μ=0 and σ=1 is called a standard normal distribution. Also, a random variable with standard normal distribution is called a standard normal random variable and is usually denoted by Z. If x is normally distributed with mean μ and standard deviation σ, then

$$Z = \frac{x-\mu}{\sigma}$$

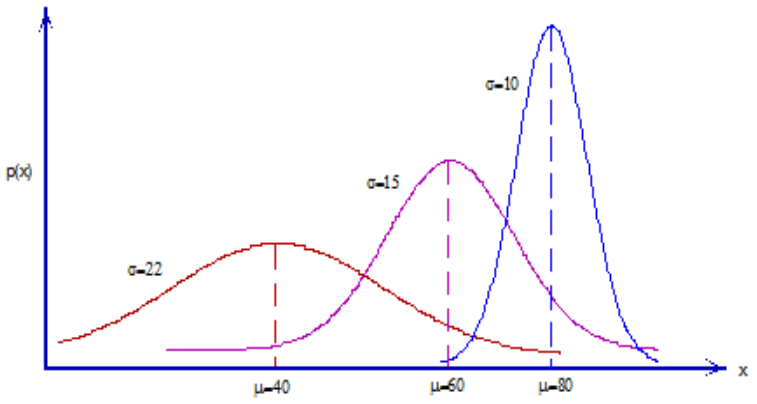
is a standard normal random variable where,
Z = distance from the mean to the point of interest (x) in terms of standard deviation units
x = point of interest
μ = the mean of the distribution, and
σ = the standard deviation of the distribution

Chapter 6: Continuous Probability Distributions - Flow Chart (1)

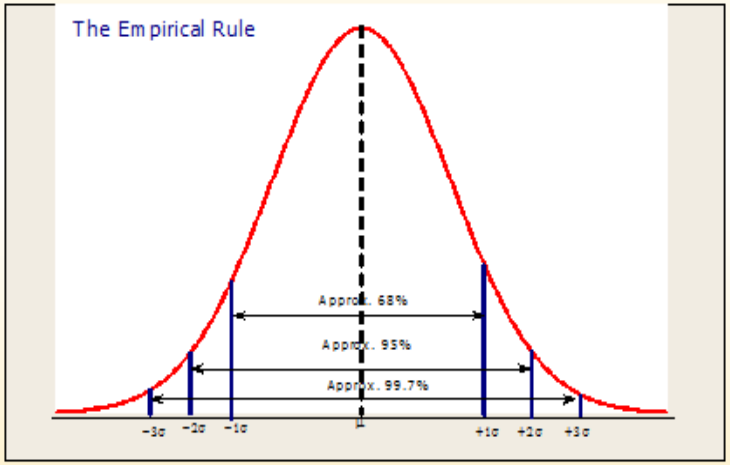
Continuous Probability Distributions...continued

Parameters of Normal Distribution

The shape of the normal curve depends upon the mean (μ) and standard deviation (σ). The mean μ and the standard deviation (σ) are the parameters of the normal distribution. The mean μ determines the location of the distribution whereas; the standard deviation σ determines the spread of the distribution.



Area Property of Normal distribution



Chapter 6: Continuous Probability Distributions - Flow Chart (2)

Exponential Distribution

Probability and Cumulative Density Function of Exponential Distribution

If the random variable X follows an exponential distribution then the probability density function is given by:

$$f(x)=\frac{1}{\mu}e^{-x/\mu} \qquad \text{where,} \quad x>0 \text{ and } \mu>0$$

Cumulative Probabilities for exponential distribution is given by:

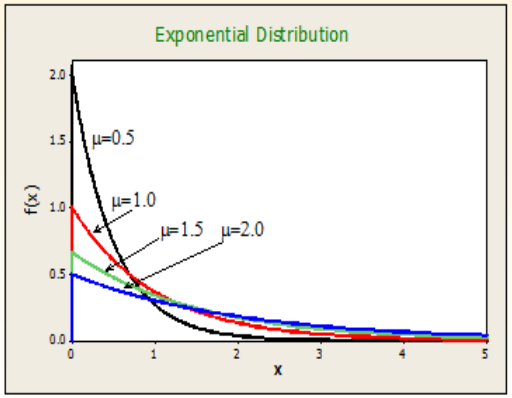
$$P(x\le x_0)=1-e^{-x/\mu} \qquad \text{for} \quad x>0$$

The mean and standard deviation of the exponential distribution are equal and given by:

Mean = μ

Standard deviation, $\sigma = \mu$

Graph of Exponential Distribution for different values of μ



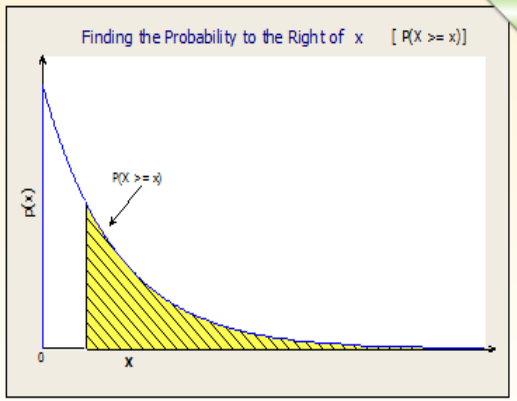
Finding Exponential Probabilities

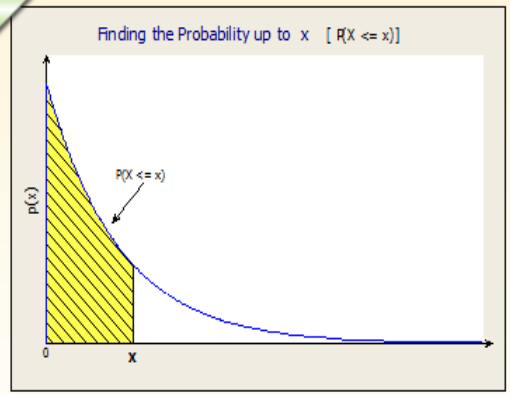
The probabilities for exponentially distributed random variables are found by evaluating the areas between the points of interest of the exponential curve described in Figure 6.39. Suppose X is an exponentially distributed random variable with parameter μ, then

$$P(X \ge x) = e^{-x/\mu} \qquad \text{for } x = 0$$

$$P(X \le x) = 1-e^{-x/\mu} \qquad \text{for } x > 0$$

$$P(x_1 \le X \le x_2) = e^{-x_1/\mu} - e^{-x_2/\mu} \qquad \text{for } x_1, x_2 > 0$$





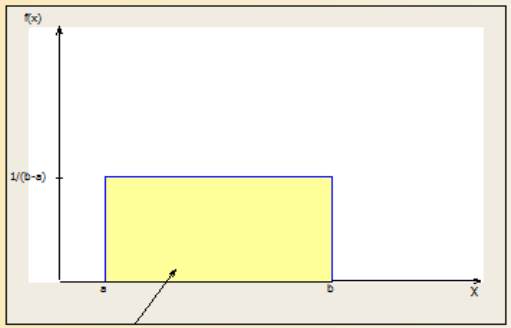
Chapter 6: Continuous Probability distribution - Flow Chart (3)

Uniform Distribution

A random variable for which all outcomes between some minimum and maximum values have equal probability of occurrence may be described by a uniform distribution. The density function of the uniform distribution is given by

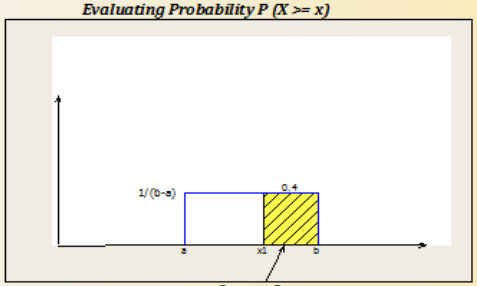
$$f(x) = \frac{1}{b-a} \qquad a = x = b$$
$$= 0 \qquad \text{otherwise}$$

- Note that *a* and *b* are two parameters of the uniform distribution where a < b. Parameter *a* is the location parameter and it controls the location of the distribution along the x-axis. The scale parameter is the difference (b-a). An increase in the difference (b-a) will elongate the distribution, whereas a decrease in the difference (b-a) will compress it.



Total shaded area= $(b-a)\left[\frac{1}{(b-a)}\right]$

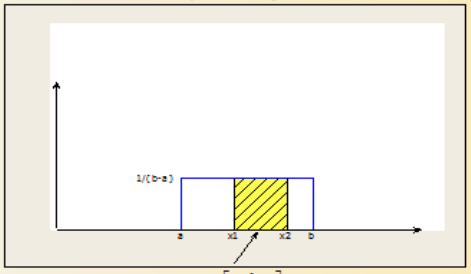
Evaluating Probability P (X >= x)



Shaded Area= $(b-x_1)\left[\frac{1}{(b-a)}\right]$

Evaluating the Probability

$P(x_1 \le X \le x_2)$



Area= $(x_2-x_1)\left[\frac{1}{(b-a)}\right]$

Mean or the Expected Value E(x) and the Variance V(x) of the Uniform Distribution

$$E(x) = \mu = \frac{(a+b)}{2} \qquad V(x) = \sigma^2 = \frac{(b-a)^2}{12}$$

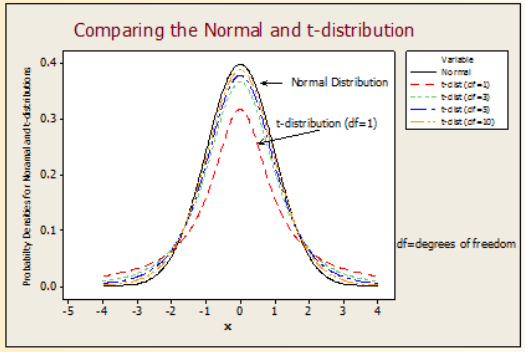
Chapter 6: Continuous Probability Distributions - Flow Chart (4)

Some Important Distributions Related to Normal Distribution

Distributions Related to Normal Distribution — extensively used in Conducting Several Statistical tests

t-distribution

- It was shown by Gosset that the random variable $t=[(\bar{x}-\mu)/(s/\sqrt{n})]$ follows the distribution known as the t-distribution if σ is not known and the sample size ***n*** is small,
- The statistic t has a mean=0 and a variance > 1(unlike the normal distribution whose mean=0 and variance=1).
- Since the variance is greater than 1, this distribution is less peaked at the center compared to the normal distribution and is also higher in the tails compared to the normal distribution.
- As the sample size, n becomes larger; the t-distribution comes closer and closer to the normal distribution.



Chi-square distribution

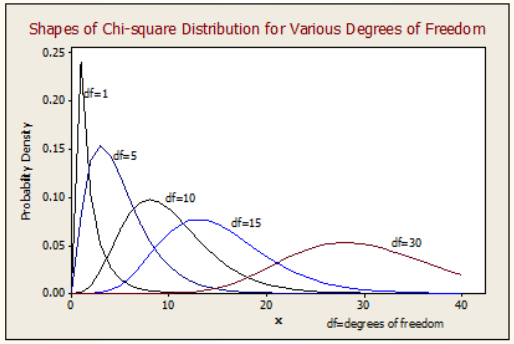
- The chi-square distribution is also defined in terms of the normal distribution. This distribution is always skewed to the right, but as the degrees of freedom increase the distribution tends toward a normal distribution.

- If a set of independent random variables $z_1, z_2, z_3,\ldots,z_k$ are normally distributed with mean zero and variance one then the sum of squares of $z_1, z_2, z_3,\ldots,z_k$ denoted by χ^2 (Chi-square) is also a random variable, and the quantity

$$\chi^2_n = z_1^2 + z_2^2 + z_3^2 + \ldots + z_k^2$$

is distributed as a chi-square distribution with n degrees of freedom. The values of χ^2 are between zero and +8 because χ^2 is the sum of squares.

- The sampling distribution of the sample variance, s^2 ; $(n-1)\,s^2/\sigma^2$ follows a χ^2 distribution with (n-1) degrees of freedom.

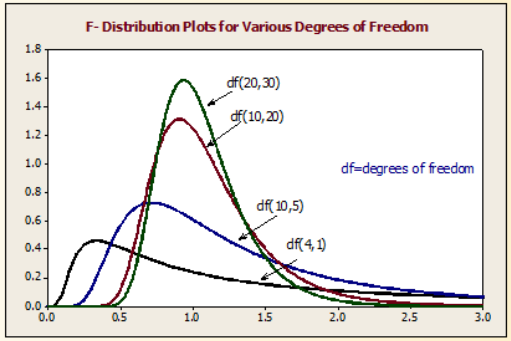


F- distribution

The F-distribution is also related to the normal distribution. Suppose, we have two independent normal variables x_1 and x_2 with the following mean and variances $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ and we draw samples of size n_1 and n_2 from the first and second normal processes. If the sample variances are s_1^2 and s_2^2, then the ratio

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

follows an F-distribution with (n_1-1) and (n_2 -1) degrees of freedom. The shape of the F-distribution depends upon the numerator and denominator degrees of freedom. As the degrees of freedom increase, the distribution approaches the normal distribution.



Chapter 6: Continuous Probability Distributions - Flow Chart (5)