5 - BASIC STATISTICS

5.1 Random Variables

In this section, we define random variables and study probability distributions of discrete and continuous random variables, expected value and variance, and special distributions such as Binomial, Poisson, Uniform, Normal and Exponential. Further, we study the properties and applications of these distributions as well.

What is a random variable?

Definition:
Let E be an experiment and S a sample space associated with the experiment. A function X assigning to every element s in S, a real number (say X(s)), is called a random variable.

Note that the values of a random variable will always be numbers and are determined by chance. Therefore, the values of the random variable cannot be predicted in advance. However, it is possible to tell beforehand the possible values that the random variable could take, and the chance (probability) of getting those values.

By convention, we use CAPITAL LETTERS for random variables (e.g. X), and lower case letters to represent the values taken by the random variable (e.g. x). Note that the standard abbreviation for 'random variable' is r.v.

Example 1:
Consider selecting a student in a class and recording his or her average mark for all the subjects. Here, the sample space is the set of students; the 'average mark obtained by a student in the class' is the random variable X because it is a function from the set of students to the real number, $X(s)$ which is the average mark of student s.

Example 2:
Toss a coin 3 times. The sample space is

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Suppose we define X as the number of heads.

Now, let us consider the value taken by each element of the sample space on X.
So, X(HTH) = 2, X(THT) = 1, etc.

Here, notice that the value taken by X on each element of the sample space is a real number. Therefore, X is a random variable according to the above definition.

Suppose we define Y such that $Y = \begin{cases} 1 & \text{if } 2^{nd} \text{ toss gives a head} \\ 0 & \text{otherwise} \end{cases}$

Then Y(HTH) = 0, Y(THH) = 1, Y(HHH) =1, etc.

So, Y is also a random variable according to the definition.

Any function is a random variable as long as it is defined on all elements of S, and takes only real values.

There are two types of random variables, discrete and continuous.

5.1.1 Discrete Random Variables

A **discrete random variable** is a random variable that has either a finite number of possible values or a countable number of possible values.

Usually, discrete random variables result from counting, such as 0, 1, 2, 3 and so on. For example, the number of members in a family is a discrete random variable.

5.1.2 Continuous Random Variables

A **continuous random variable** is a random variable that has either an infinite number of possible values that is not countable.

Continuous random variables are variables that result from measurements. For example, air pressure in a tyre of a motor vehicle represents a continuous random variable, because air pressure could in theory take on any value from 0 lb/in^2 (psi) to the burning pressure of the tyre.

The distinction between discrete and continuous random variables is important because the statistical techniques associated with the two types of random variables are different.

Probability Distribution of Random Variables

As the value of a random variable cannot be predicted in advance, it may be useful to find the probabilities that correspond to the possible values of the random variable.

Definition:

The **probability distribution** of a random variable provides the possible values of the random variable and their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

5.2 Probability Distribution of a Discrete Random Variable

Let X be a discrete random variable. The most basic question we can ask is: what is the probability that X takes the value x? In other words, what is $P(X=x)$?

5.2.1 Definition

Let X be a discrete random variable. Suppose X takes countable number of values x_1, x_2, x_3, … . With each possible value x_i, we associate a number $p(x_i) = P(X = x_i)$, then we call $p(x_i)$, i=1,2,3, … the probability of x_i if they satisfy the following conditions.

1. $\sum P(X=x_i)=1$
2. $0 \le P(X=x_i) \le 1$ for all i

Example 3:
Toss a fair coin 3 times. Define X as the number of heads obtained. Find the probability distribution of X.

Solution:
All the 8 outcomes HHH, HHT, HTH, THH, HTT, THT, TTH, TTT are equally likely:

So, we can use our knowledge of probability to find out that

P(HHH) = P(HHT) = ….. = P(TTT) = 1/8

Then
P(X = 0) = P({TTT}) = 1/8
P(X = 1) = P({HTT, THT, TTH}) =3/8
P(X = 2) = P({HHT, HTH, THH}) =3/8
P(X = 3) = P({HHH,}) =1/8

These probabilities constitute a probability distribution because they provide the corresponding probabilities (> 0) of the possible values of the random variable X, and that P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1 (i.e. the property 1 of the distribution).

Definition:
The **Cumulative Distribution Function (c.d.f.)** of a discrete r.v. X is given by

$F_X(x) = P(X \le x) = \sum^{x} P(X = x_i)$, which is the cumulative probability up to the value x.

Example 4:
Consider the previous example 3. Let X = No. of heads. Find the cumulative probability up to each of the values *0, 1, 2, 3.*

Solution:

$F_X(0) = P(X \le 0) = \frac{1}{8}$, $F_X(1) = P(X \le 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$

$F_X(2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$, $F_X(3) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.$

Verify also that $F_x(-1) = 0$, $F_x(0.5) = 1/8$, $F_x(4) = 1$.

Properties of the Cumulative Distribution Function (when X is Discrete)

1. $F(-\infty) = 0$, $F(+\infty) = 1$
2. $F_X(x)$ is a non-decreasing function of x: that is
   if $x_1 < x_2$, then $F_X(x_1) \le F_X(x_2)$.
3. If $P(a \le X \le b) = F_X(b) - F_X(a)$

5.2.2 Mean and Variance of a Discrete Random Variable

Having identified the distribution of a discrete random variable, it may now be important to introduce the centre and spread of a random variable. We usually use the mean to describe the centre of a random variable. The mean of a random variable is often called the *expected value* of the random variable. The variance and standard deviation are used to describe the spread of a random variable. Variance is in fact a measure of how spread-out the values are around their mean.

Definition:
The mean and variance of a discrete random variable are given by the following formulas. Note that the mean (or expected value) is denoted by μ (or $E(X)$), and the variance is denoted by σ^2 or $V(X)$.

$$\mu = E(X) = \sum [x.P(X = x)]$$

where x is the value of the random variable and $P(X = x)$ is the probability that X takes the value x. Note that μ is the population mean because the sum ∑ is taken over all values of the r.v.

$\sigma^2 = V(X) = E(X - E(X))^2$. Here σ is the population standard deviation.

Note:

- The expected value of X always lies between the smallest and largest values of X.
- The variance of X is never negative.

To find the standard deviation of the random variable, take the square root of the variance.

When computing the V(X), it may be easy to use $V(X) = E(X^2) - [E(X)]^2$ which can be shown as follows.

By definition,

$$V(X) = E(X - E(X))^2$$

$$= \mathrm{E}(\mathrm{X}^2 - 2X.E(X) + [E(X)]^2)$$
$$= \mathrm{E}(\mathrm{X}^2) - 2E(X).E(X) + [E(X)]^2$$
$$= \mathrm{E}(\mathrm{X}^2) - [E(X)]^2$$

Example 5:
Consider the previous example 3, and consider X as the number of heads. What are the expected value and variance of X?

Solution:
We have seen that the probability distribution of X was:

$X = x$	0	1	2	3
$P(X=x)$	1/8	3/8	3/8	1/8

So, $E(X) = 0.(1/8) + 1.(3/8) + 2.(3/8) + 3.(1/8) = 3/2$.

$V(X) = 0^2.(1/8) + 1^2.(3/8) + 2^2.(3/8) + 3^2.(1/8) - [3/2]^2 = 3/4$,
since $V(X) = E(X^2) - [E(X)]^2$.

Now, consider the following theorems with regard to the expected value and the variance. The proofs of the theorems are not given here.

Theorem 1: If 'a' is a constant, then E(a) = a

Theorem 2: If 'a' is a constant and X is any random variable, then E(aX) = a E(X)

Theorem 3: If 'a' and 'b' are constants, then E(aX + b) = a E(X) + b

Theorem 4: If 'a' is a constant, then V(a) = 0

Theorem 5: If 'a' is a constant and X is any random variable, then
$V(aX) = a^2 V(X)$

Theorem 6: If 'a' and 'b' are constants, then $V(aX + b) = a^2 V(X)$

Note that these theorems are valid whether X is a discrete or a continuous random variable.

5.3 The Binomial Probability Distribution

Binomial Experiment

Suppose that we have a biased coin for which the probability of obtaining a head is 2/3. We toss the coin 100 times and count the number of heads obtained. This problem is typical of an entire class of problems that are characterized by the feature that there are exactly two possible outcomes (for each trial) of interest. These problems are called binomial experiments.

Features of a Binomial experiment

1. There are a fixed number of trials. We denote this number by n.
2. The n trials are independent (result of one trial does not depend on any other trial), and are repeated under identical conditions.
3. Each trial has only two outcomes; success denoted by *S,* and failure denoted by *F*.
4. For each trial, the probability of success is the same. We denote the probability of success by p and that of failure by q. Since each trial results in either success or failure, $p + q = 1$ and $q = 1 - p$.

In the above experiment of tossing a biased coin, let us now see how it meets the above criteria of a binomial experiment. Consider the above features one at a time.

1. The coin is tossed 100 times, so there are $n = 100$ trials (fixed) in this case.

2. The trials can be considered as independent, as the outcome of one trial has no effect on the outcome of another trial.

3. There are only two outcomes, head or tail. As we are interested in getting a head, it can be considered as a success, and getting a tail can be considered as a failure.

4. On each trial, the probability p of success is 2/3 (same for all trials).

In this type of binomial experiments, our interest is to find the probability of a certain number of successes (say r) out of n trials.

Here, if X is defined as the number of getting r successes, then we say X is distributed as Binomial with parameters n and p. That is denoted by:

$$X \sim \text{Bin}(n, p)$$

Example 6:
Suppose a student is taking a multiple-choice question paper, and he has only three more multiple-choice questions left to do. Each question has 4 suggested answers, and only one of the answers is correct. He has only few seconds left to do these three questions, so he decides to randomly select (guess) the answers. The interest here is to know the probability that he gets zero, one, two, or all three questions correct.

Solution:
This is a binomial experiment. Each question can be considered as a trial, so the number of trials (n) is 3.

There are only two outcomes for each trial – success (S) indicating a correct answer, and failure (F) indicating a wrong answer.

The trials are independent – outcome (correct or incorrect) for any one question does not affect the outcome of the others.

Sine he is guessing and there are 4 answers from which to select, the probability of a correct answer is p=0.25. The probability q of an incorrect answer is then $1 - p = 1 - 0.25 = 0.75$.

So, this is a binomial experiment with $n = 3$, $p = 0.25$. So $X \sim \text{Bin}(3, 0.25)$.

Now what are the possible outcomes in terms of success or failure for these three trials? Here we use the notation SFS to indicate a success on the first question, a failure on the second, and a success on the third. There are 8 possible combinations of S's and F's. They are:

SSS SSF SFS FSS SFF FSF FFS FFF

The probability for each of the above combinations can be computed using the multiplication law as the trials are independent. To illustrate this, let us compute the probability of 'SFS' (success on the first question, failure on the second, and success on the third).

$$P(SFS) = P(S).P(F).P(S) = p.q.p = p^2q = (0.25)^2(0.75) \approx 0.047$$

In a similar way, probability of each of the above eight outcomes can be computed, and they are given in the table below.

Table 1 : Probabilities of outcomes for a Binomial experiment with n=3 & p=0.25

Outcome	**No. of successes (r)**	**Probability**
SSS	3	P(SSS) = P(S)P(S)P(S) = $p.p.p \approx 0.016$
SSF	2	P(SSF) = P(S)P(S)P(F) = $p.p.q \approx 0.047$
SFS	2	P(SFS) = P(S)P(F)P(S) = $p.q.p \approx 0.047$
FSS	2	P(FSS) = P(F)P(S)P(S) = $q.p.p \approx 0.047$
SFF	1	P(SFF) = P(S)P(F)P(F) = $p.q.q \approx 0.141$
FSF	1	P(FSF) = P(F)P(S)P(F) = $q.p.q \approx 0.141$
FFS	1	P(FFS) = P(F)P(F)P(S) = $q.q.p \approx 0.141$
FFF	0	P(FFF) = P(F)P(F)P(F) = $q.q.q \approx 0.422$

Let us now compute the probability that the student gets zero, one, two, or all three questions correct.

P(X = 1) = P[SSF or FSF or FFS]
= P(SFF) + P(FSF) + P(FFS) as SFF, FSF & FFS are mutually exclusive
= 0.423

In the same way, we can find that P(X =2), P(X = 3) and P(X = 0).
Verify that P(X = 3) = 0.016, P(X =2) = 0.141, and P(X = 0) = 0.422. With these results you can see that there is very little chance (0.016) that the student gets all the questions correct.

If X ~ Bin(n, p), then the general formula of computing the probability of getting r successes can be specified as:

$$P(X = r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}, \quad r = 0, 1, 2, ..., n$$

Where n = no. of trials, p = prob. of success
r = no. of successes

The $\frac{n!}{r!(n-r)!} = {}^{n}C_{r}$ is the binomial coefficient which represents the number of combinations with n trials having r successes. As an exercise, compute the probabilities in table 1 using the general formula of the binomial distribution.

Example 7:
Suppose that a computer component has a probability of 0.7 of functioning more than 10,000 hours. If there are 6 such components, what is the probability that at least 4 computer components will function more than 10,000 hours?

Solution:
This is a binomial experiment with n = 6 and p = 0.7. Therefore, if we define X as the number of computer components functioning more than 10,000 hours, then $X \sim Bin(6, 0.7)$.

So, the required probability is:

$$\begin{aligned} P(X \geq 4) &= P(X=4 \text{ or } X=5 \text{ or } X=6) \\ &= P(X=4) + P(X+5) + P(X=6) \\ &= 0.324 + 0.303 + 0.118 \\ &= 0.745 \end{aligned}$$

Mean and Variance of Binomial distribution

Here we can use two formulas to compute the mean (μ) and variance (σ^2) of the binomial distribution.

If $X \sim Bin(n, p)$, then it can be shown that

Mean = $\mu = n.p$ is the expected number of successes and
Variance = $\sigma^2 = n.p.q$ is the variance for the number of successes.

5.4 The Poisson Probability Distribution

This is another discrete probability distribution. The Poisson random variable, unlike the ones we have seen before, is very closely connected with continuous things. Suppose that 'incidents' occur at random times, but at a steady rate overall. The best example is radioactive decay: atomic nuclei decay randomly, but the average number λ which will decay in a given interval is constant.

The Poisson random variable *X* counts the number of independent 'incidents' which occur very often in a given time interval, volume, area and so forth.

So if, on average, there are 2.4 nuclear decays per second, then the number of decays in one second starting now is a Poisson(2.4) random variable. The number of telephone calls in a minute to a busy telephone number, and the number of

customers arriving at a supermarket in an hour, the occurrence of bacteria in air, the number of typographical errors in a page of a book are some more examples.

Although we will not prove it, the probability distribution for a random variable *X,* that is distributed as Poisson is given by the formula:

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

Where λ represents the average number of occurrences of the random event in the interval specified.

We usually write it as $X \sim \text{Poisson}(\lambda)$.

Example 8:
A restaurant manager, from his experience knows that vehicles arrive at the drive-through of the restaurant at the rate of 2 vehicles per minute between 5 p.m. and 6 p.m. He wants to find out the following.

(a) exactly 6 vehicles arrive between 5.55 p.m. and 6 p.m.
(b) less than 6 vehicles arrive between 5.55 p.m. and 6p.m.
(c) at least 6 vehicles arrive between 5.55 p.m. and 6p.m.

Solution:
Here we can identify that the random variable X, the number of vehicles that arrive between 5.55 p.m. and 6 p.m., follows a Poisson distribution. It is given that vehicles arrive at the rate of 2 per minute, but the interval of time we are interested in this example is 5 minutes. So, $\lambda = 2 \times 5 = 10$.

(a) The probability that exactly six vehicles arrive between 5.55 p.m. and 6 p.m. is :

$$P(X=6)=\frac{e^{-10}.10^{6}}{6!}\approx 0.0631$$

(b) The probability that fewer than six vehicles arrive between 5.55 p.m. and 6 p.m. is :

$$\begin{aligned} P(X<6) &= P(X\le 5) \\ &= P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5) \\ &= 0.0671 \end{aligned}$$

(c) The probability that at least six vehicles arrive between 5.55 p.m. and 6 p.m. is:

$$\begin{aligned} P(X\ge 6) &= 1 - P(X<6) \\ &= 1 - 0.0671 \\ &= 0.9329 \end{aligned}$$

Mean and Variance of Poisson distribution

If $X \sim \text{Poisson}(\lambda)$, then it can be shown that

Mean = $\mu = \lambda$
Variance = $\sigma^2 = \lambda$

5.5 Probability Distribution of a Continuous Random Variable

A continuous random variable can take any values anywhere in some interval of the real line, e.g. $[0, \infty)$ or $(0, 1)$. Very often quantities such as time, weight, height etc. are commonly considered as continuous random variables.

Recall that, for a discrete random variable X, the probability distribution lists all values that X can take, and give their probabilities. For a continuous random variable X, it is impossible to list all the values that X can take. It is also impossible to think of the probability that X takes any one specific value.

E.g.: Even between the values 0.99999999 and 1.00000001, there are so many values that the probability of each value is negligible. In fact, we write $P(X = x) = 0$ for any x, when X is continuous. Instead, we work with *intervals* for continuous random variables:

E.g. $P(X = 0) = 0$, but $P(0.999 \leq X \leq 1.001)$ can be > 0.

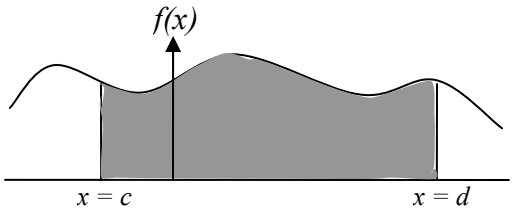
5.5.1 Definition

X is said to be a *continuous random variable* if there exists a function f, called the probability density function (pdf) of X, satisfying the following conditions:

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

(b) $f(x) \geq 0$ for all x

Note:
$P(c \leq x \leq d)$ represents the shaded area under the graph in the following figure of the probability density function between $x = c$ and $x = d$.



So, $P(c \le X \le d) = \int_c^d f(x)dx = F_X(d) - F_X(c)$.

Properties of the Cumulative Distribution Function (when X is Continuous)

1. $F(-\infty) = 0$, $F(+\infty) = 1$

2. $F_X(x)$ is a non-decreasing continuous function of x. *This means that* $F_X(x) < F_X(y)$ *if* $x < y$.

3. $P(a \le X \le b) = F_X(b) - F_X(a)$

It makes no difference whether we say $P(a < X \le b)$ or $P(a \le X \le b)$ because $P(a \le X \le b) = P(X = a) + P(a < X \le b) = P(a < X \le b)$ since $P(X = a) = 0$.

i.e. for a continuous random variable, $P(a < X < b) = P(a \le X \le b)$.

However, this is not true for a discrete random variable.

Definition:
Let X be a continuous random variable with cumulative distribution function $F_X(x)$. The probability density function (pdf) of X is:

$$f(x) = \frac{d\,F_X}{dx}$$

Note : If $f(x)$ is the pdf for a continuous random variable, then $F(x) = \int_{-\infty}^{x} f(y)dy$.

This is true only if X is a continuous r.v.

5.5.2 Mean and Variance of a Continuous Random Variable

Definition:

The mean and variance of a continuous random variable are given by the following formulas.

$$\mu = E(X) = \int_{-\infty}^{+\infty} x.f(x)dx$$

where $f(x)$ is the probability density function.

$$\sigma^2 = V(X) = E(X - E(X))^2 \quad \text{or}$$
$$= E(X^2) - [E(X)]^2 \quad \text{or}$$
$$= \int_{-\infty}^{+\infty} x^2.f(x)dx - \mu^2$$

Note : The properties of variance for continuous r.v.'s are exactly the same as for discrete r.v.'s. See theorems 1- 6 in section 5.2.2.

Example 9:
Suppose that the random variable X has p.d.f. given by

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Verify whether $f(x)$ is a p.d.f.
(b) Find the cumulative distribution function of X
(c) Find the mean and variance of X

Solution:
(a) $f(x) \ge 0$ for all x because $f(x) = 2x \ge 0$ when $0 \le x \le 1$, and
$f(x) = 0$ when x takes all the other values.

Also, $\int_0^1 f(x)dx = \int_0^1 2x.dx = [x^2]_0^1 = 1$

Therefore, $f(x)$ satisfies the two conditions to become a p.d.f.

(b) The Cumulative distribution function

$$F_X(x) = \int_0^1 f(x)dx = \begin{cases} 0 & \text{when } x < 0 \\ x^2 & \text{when } 0 \le x \le 1 \\ 1 & \text{when } x > 1 \end{cases}$$

(c) Mean $= E(X) = \int_0^1 x.2x.dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$

$$E(X^2) = \int_0^1 x^2.2x.dx = 2\left[\frac{x^4}{4}\right]_0^1 = \frac{1}{2}$$

So, $V(X) = E(X^2) - [E(X)]^2 = 1/2 - 4/9 = 1/18$.

5.5.3 The Uniform Probability Distribution

Let a and b be real numbers with $a < b$. A uniform random variable on the interval $[a, b]$ is, roughly speaking, "equally likely to be anywhere in the interval". In other words, its probability density function is constant (say c) on the interval $[a, b]$ (and zero outside the interval). What should the constant value c be? It can be shown that $c = 1/(b - a)$, because the p.d.f. should be $f(x) = c$ when $a \le x \le$ b, and it should satisfy:

$$\int_a^b c.dx = 1 \;\Rightarrow\; c.[x]_a^b = 1 \;\Rightarrow\; c.(b-a) = 1 \;\Rightarrow\; c = \frac{1}{b-a}$$

Therefore, the p.d.f. of the Uniform distribution is :

$$f(x) = \begin{cases} 1/(\text{b - a}) & \text{if a} \le \text{x} \le b \\ 0 & \text{otherwise} \end{cases}$$

The Uniform distribution is usually denoted by $U(a, b)$.

Further calculation (or the symmetry of the p.d.f.) shows that the expected value is given by $(a + b) = 2$ (the midpoint of the interval), and $V(X) = (b - a)^2/12$.

The uniform random variable doesn't really arise in practical situations. However, it is very useful for simulations. Most computer systems include a *random number generator*, which apparently produces independent values of a uniform random variable on the interval [0, 1].

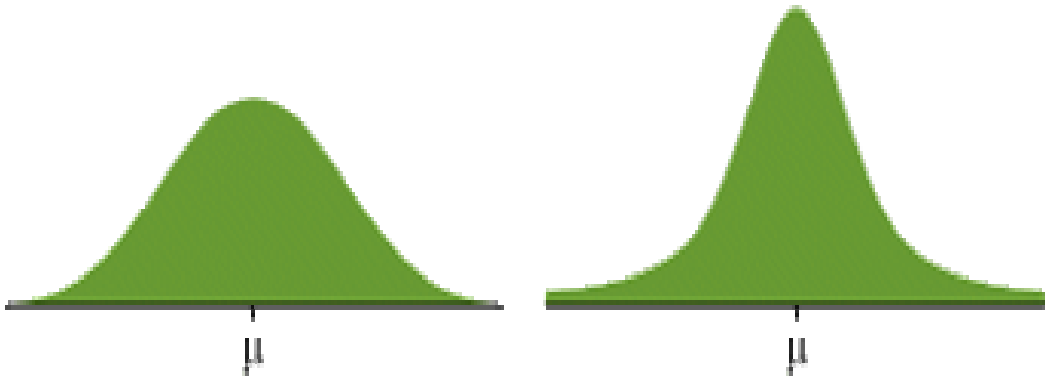
5.5.4 Normal Probability Distribution

Normal distribution is one of the most important examples of a continuous probability distribution. It is probably the most important distribution in statistics since many measurements have (approximate) normal distributions, and hence many statistical methods have been developed for normally distributed data.

The Normal (also known as Gaussian) distribution has two parameters, the mean, μ, and the variance, σ^2. Note that μ and σ^2 satisfy $-\infty < \mu < \infty$, $\sigma^2 > 0$.

If the continuous random variable is distributed as Normal with mean μ and variance σ^2, we write it as $X \sim N(\mu, \sigma^2)$.

The shape of the normal distribution takes the familiar bell-shaped curve which is symmetrical about the vertical line over the mean μ. The parameter σ controls the spread of the curve. If the standard deviation σ is large, the curve is flat and more spread out, and if it is small, the curve is more peaked (see below).



$X \sim N(\mu, \sigma_1^2)$ - Flat **$X \sim N(\mu, \sigma_2^2)$ - Peaked**

Here, $\sigma_1^2 > \sigma_2^2$

The total area under the normal curve is always 1. The graph of the normal distribution is important because the portion of the area under the curve above a given interval represents the probability that a measurement will lie in that interval.

The formula of the shape of the normal distribution is the ***normal probability density function***. If $X \sim N(\mu, \sigma^2)$, then the normal probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad -\infty < x < \infty, \ \sigma > 0$$

The following empirical rule gives us information about the percentage of data that lies within one, two, and three deviations of the mean.

Empirical Rule

For a normal distribution,

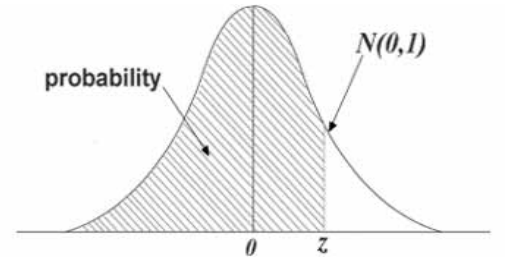
- Approximately 68% of the data values will lie within one standard deviation on each side of the mean.
- Approximately 95% of the data values will lie within two standard deviations on each side of the mean.
- Approximately 99.7% (or almost all) of the data values will lie within three standard deviations on each side of the mean.

Example 10:
The lifetime of a computer component is normally distributed with mean $\mu = 6000$ hours and standard deviation $\sigma = 500$ hours. What is the probability that a computer component selected at random will last from 6000 to 6500 hours?
Solution:
The probability that the lifetime will be between 6000 and 6500 hours can be considered as the corresponding percentage of the area under the curve within that interval. We can identify that this interval is in fact μ and $\mu + \sigma$. As we know from the empirical rule that the area under the normal curve between $\mu - \sigma$ and $\mu + \sigma$ is 68%, the area between μ and $\mu + \sigma$ is half of 68% (or 34%) because the normal distribution is symmetric over μ. This tells us that the probability a computer component will last from 6000 to 6500 hours is 0.34.

Standard Normal Distribution

Normal distributions vary from one another as they depend on two parameters, the mean μ and the standard deviation σ. The computation of the area under the curve in a specified interval of x values (i.e. probability) is not easy due to the complexity of the normal probability density function. It would be a futile task to obtain a table of areas (probabilities) under the normal curve for each different combination of μ and σ.

Mathematicians have found a way to standardize the distributions so that we can use one table of areas for all normal distributions. For this, any normal distribution is converted to the **standard normal distribution**.



Standard Normal Distribution

The standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. We write it as $X \sim N(0, 1)$.

Theorem: If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

The proof of the theorem is not given here.

The above theorem says that any normal random variable X (with mean = μ and variance = σ^2) can be converted to a **standard normal** random variable Z (with mean = 0 and variance = 1). The advantage here is that we can use one table which shows the areas (probabilities) under the standard normal distribution for any interval of *z* values. A table is given at the end to find the probabilities under the standard normal distribution.

Example 11:
Use the table of standard normal distribution to find:

(a) the area (probability) under the standard normal distribution to the left of $z = -1.00$.
(b) the area (probability) between $z = 1.00$ and $z = 2.70$
(c) area (probability) to the right of $z = 0.95$

Solution:

(a) To find the area (probability) to the left of $z = -1.00$, we use the row headed by -1.0 under the column Z of the table, and then move to the corresponding position on the right under the column P. We can see that this value is 0.1587.

(b) Area (probability) between 1.00 and 2.70
= (area left of 2.70) – (area left of 1.00)
= 0.9965 – 0.8413
= 0.1552

(c) Area (probability) to the right of 0.95
= (area under entire curve) – (area to the left of 0.95)
because $P(a \le X \le b) = F_X(b) - F_X(a)$
= 1.0000 – 0.8289
= 0.1711

Alternatively, Area to the right of 0.95 = Area to the left of -0.95
= 0.1711

This is due to the fact that the standard normal r.v. Z is symmetric about zero, and hence, for any positive number c, $P(Z \leq -c) = P(Z \geq c) = 1 - P(Z \leq c)$.

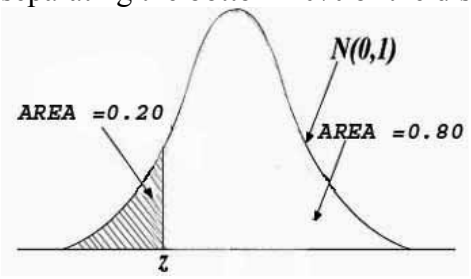
Note:
Any table is limited in the number of entries it contains. Interpolation can be used to extend the range of values tabulated. For example, suppose you need to find the probability (area) under the standard normal distribution to the left of $z = -1.02$. The standard normal table does not give the P value at $z = -1.02$ (see the table), and it only gives the probability values corresponding to $z = -1.00$ and $z = -1.05$. Here we assume that the normal density function is changing at a roughly constant rate between, say, -1.00 and -1.05. So the $z = -1.02$ will be about two fifth of the way between the corresponding P values for $z = -1.00$ (i.e. 0.1587) and $z = -1.05$ (i.e. 0.1469). So, the corresponding P value for $z = -1.02$ is:

$$(2/5)*(0.1587 - 0.1469) + 0.1469 = 0.15162$$

Sometimes, we may need to calculate the value of a normal random variable required for the variable to be at a certain proportion or probability rather than calculating the proportion or probability for a given value of a normal random variable. This is illustrated by an example given below.

Example 12:
Consider the average marks (say X) obtained by 200 students in a particular class. It is assumed that the average marks of students are normally distributed with mean = 58.7 and standard deviation = 15.2. Find the average mark of a student at the 40^{th} percentile. In other words, we need to know the average mark of a student that separates the bottom 20% of students from the top 80% of students.

Solution:
The following figure shows the normal curve with the unknown value of X separating the bottom 20% of the distribution from the top 80% of the distribution.



The area closest to 0.20 in the standard normal table under column P is 0.1977. The corresponding z value is -0.85. As the average marks (X) are distributed as normal with mean 58.7 and standard deviation 15.2, the x value can be found using $z = (x - \mu)/\sigma$. So, $x = \mu + z.\sigma$.

$$= 58.7 + (-0.85)(15.2)$$
$$= 45.78$$

So, the average mark of a student that separates the bottom of the data from the top 80% is 45.78.

5.5.5 Normal Approximation to the Binomial Distribution

In section 5.3, we considered the binomial distribution, and calculated the binomial probabilities using the formula of the binomial probability distribution function. When there are a large number of trials of a binomial experiment, the binomial probability formula can be difficult to use. For example, suppose there are 600 trials of binomial experiment, and we wish to compute the probability of 500 or more successes. For this, it would require that we compute the following probabilities.

$P(X \geq 500) = P(X = 500) + P(X = 501) + \ldots + P(X = 600)$

This would be time consuming to compute by hand. In fact, there are techniques for approximating binomial probabilities, provided that certain conditions are met.

It has been found that the number of trials n in a binomial experiment increases, then the probability distribution of the binomial random variable X becomes more nearly symmetric and bell shaped.

Normal Approximation to Binomial Distribution

Let $X \sim \text{Bin}(n, p)$

Then we say that X is approximately distributed with mean $\mu = n.p$ and variance $\sigma^2 = n.p.(1 - p)$ when n is large.

As a general rule of thumb, we use the condition $n.p.(1 - p) \geq 10$ to make this approximation valid.

Note:
In using this approximation, we are approximating the distribution of a discrete random variable to a distribution of a continuous random variable. Hence, some care must be taken with the endpoints of the intervals involved. For example, for a continuous random variable, $P(X = 10) = 0$, but this probability may be positive for a discrete random variable. The following corrections for continuity have been suggested to improve the above approximation:

(a) $P(a \leq X \leq b) \approx P(a - 0.5 \leq X \leq b + 0.5)$

(b) $P(X \leq b) \approx P(X \leq b + 0.5)$
(c) $P(a \leq X) \approx P(a - 0.5 \leq X)$

Example 13:
The probability that a light bulb will fail in a year is 0.75, and light bulbs fail independently. If 192 bulbs are installed, what is the probability that the number which fail in a year lies between 140 and 150 inclusive?

Solution:
Let X be the number of light bulbs which fail in a year. Then $X \sim \text{Bin}(192, 0.75)$, and so $E(X) = 144$, $V(X) = 36$.

X can be approximated by $X \sim N(144, 36)$ because $n.p.(1 - p) = 36 \geq 10$

The required probability is $P(140 \leq X \leq 150) \approx P(140 - 0.5 \leq X \leq 150 + 0.5)$ by the continuity correction.

$$= P(139.5 \leq X \leq 150.5)$$

Let $Z = (X - 144)/6$. Then $Z \sim N(0, 1)$. So,

$$\begin{aligned} P(139.5 \leq X \leq 150.5) &= P\left(\frac{139.5 - 144}{6} \leq Z \leq \frac{150.5 - 144}{6}\right) \\ &= P(-0.75 \leq Z \leq 1.08) \\ &= 0.8598 - 0.2266 \\ &= 0.633 \end{aligned}$$

5.5.6 The Exponential Probability Distribution

The exponential random variable arises in the same situation as the Poisson: be careful not to confuse them! We have events which occur randomly but at a constant average rate of λ per unit time. The Poisson random variable, which is discrete, counts how many events will occur in the next unit of time. The exponential random variable, which is continuous, measures exactly how long from now it is until the next event occurs (e.g. : inter-arrival time of customers). Note that it takes non-negative real numbers as values.

A continuous random variable X assuming all non-negative values takes the exponential distribution with parameter α (> 0) if the p.d.f of X is given by

$$f(x) = \begin{cases} \alpha.e^{-\alpha x}, & \text{when x >0} \\ 0, & \text{elsewhere} \end{cases}$$

It is usually written as $X \sim \text{EXP}(\alpha)$.

Mean and Variance of the Exponential Distribution

If X ~ EXP(α), then the expected value of X is obtained as follows.

$E(X) = \int_0^\infty x.\alpha e^{-\alpha x} dx$, Integrating by parts and letting $\alpha e^{-\alpha x}\, dx = dv$, $x = u$, we obtain $u = -e^{-\alpha x}$, $du = dx$. Thus

$$E(X) = \left[-x.e^{-\alpha x}\right]_0^\infty + \int_0^\infty e^{-\alpha x} dx = \frac{1}{\alpha}.$$

The variance of X may be obtained by a similar integration. We find that $E(X^2) = 2/\alpha^2$ and therefore $V(X) = E(X^2) - [E(X)]^2 = 1/\alpha^2$.

Example 14:
Let X have an exponential distribution with a mean = 20. Find

(a) the probability that X is less than 18.
(b) the median of X.

Solution:
Since mean of X is $1/\alpha = 20$, the value of $\alpha = 1/20$.
So, $f(x) = (1/20).e^{-x/20}$, $0 < x < \infty$

(a) $P(X < 18) = \int_0^{18} \frac{1}{20} e^{-x/20}\, dx = 1 - e^{-18/20} = 0.593$

(b) The median, m, can be found using the cumulative distribution function,

$$F_X(m) = P(X \le m) = 0.5$$

$$= \int_0^m \frac{1}{20} e^{-x/20} dx = 0.5$$

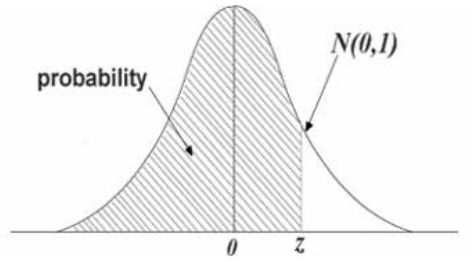
$$= 1 - e^{-m/20} = 0.5$$

$$= e^{-m/20} = 0.5$$

$$\text{so, } m = -(20).\ln(0.5)$$

$$= 13.86$$

The Standard Normal Distribution Table



The distribution tabulated is that of the normal distribution with mean **zero** and standard deviation **1**. For each value of ***Z***, the standardized normal deviate, (the proportion ***P***, of the distribution less than ***Z***) is given. For a normal distribution with mean μ and variance σ^2 the proportion of the distribution less than some particular value ***X*** is obtained by calculating $Z = (X - \mu)/\sigma$ and reading the proportion corresponding to this value of ***Z***.

Z	*P*	*Z*	*P*	*Z*	*P*
-4.00	0.00003	-1.00	0.1587	1.05	0.8531
-3.50	0.00023	-0.95	0.1711	1.10	0.8643
-3.00	0.0014	-0.90	0.1841	1.15	0.8749
-2.95	0.0016	-0.85	0.1977	1.20	0.8849
-2.90	0.0019	-0.80	0.2119	1.25	0.8944
-2.85	0.0022	-0.75	0.2266	1.30	0.9032
-2.80	0.0026	-0.70	0.2420	1.35	0.9115
-2.75	0.0030	-0.65	0.2578	1.40	0.9192
-2.70	0.0035	-0.60	0.2743	1.45	0.9265
-2.65	0.0040	-0.55	0.2912	1.50	0.9332
-2.60	0.0047	-0.50	0.3085	1.55	0.9394
-2.55	0.0054	-0.45	0.3264	1.60	0.9452
-2.50	0.0062	-0.40	0.3446	1.65	0.9505
-2.45	0.0071	-0.35	0.3632	1.70	0.9554
-2.40	0.0082	-0.30	0.3821	1.75	0.9599
-2.35	0.0094	-0.25	0.4013	1.80	0.9641
-2.30	0.0107	-0.20	0.4207	1.85	0.9678
-2.25	0.0122	-0.15	0.4404	1.90	0.9713
-2.20	0.0139	-0.10	0.4602	1.95	0.9744
-2.15	0.0158	-0.05	0.4801	2.00	0.9772
-2.10	0.0179	0.00	0.5000	2.05	0.9798
-2.05	0.0202	0.05	0.5199	2.10	0.9821
-2.00	0.0228	0.10	0.5398	2.15	0.9842
-1.95	0.0256	0.15	0.5596	2.20	0.9861
-1.90	0.0287	0.20	0.5793	2.25	0.9878
-1.85	0.0322	0.25	0.5987	2.30	0.9893
-1.80	0.0359	0.30	0.6179	2.35	0.9906
-1.75	0.0401	0.35	0.6368	2.40	0.9918
-1.70	0.0446	0.40	0.6554	2.45	0.9929

-1.65	0.0495	0.45	0.6736	2.50	0.9938
-1.60	0.0548	0.50	0.6915	2.55	0.9946
-1.55	0.0606	0.55	0.7088	2.60	0.9953
-1.50	0.0668	0.60	0.7257	2.65	0.9960
-1.45	0.0735	0.65	0.7422	2.70	0.9965
-1.40	0.0808	0.70	0.7580	2.75	0.9970
-1.35	0.0885	0.75	0.7734	2.80	0.9974
-1.30	0.0968	0.80	0.7881	2.85	0.9978
-1.25	0.1056	0.85	0.8023	2.90	0.9981
-1.20	0.1151	0.90	0.8159	2.95	0.9984
-1.15	0.1251	0.95	0.8289	3.00	0.9986
-1.10	0.1357	1.00	0.8413	3.50	0.99977
-1.05	0.1469			4.00	0.99997