Chapter Eight: Probability Distributions

When you completed this chapter, you will be able to:

- ✓ understand the properties of a probability distribution;
- ✓ identify the difference and similarity of discrete and continuous random variables;
- ✓ understand the concepts of expected value and variance of a discrete variable;
- ✓ show how certain types of discrete data may be represented by Binomial Distribution;
- ✓ show how certain types of discrete data may be represented by Poisson Distribution;
- ✓ approximate a Binomial Distribution by a Poisson Distribution;
- ✓ illustrate the properties of a Normal Distribution;
- ✓ describe the parameters of a Normal Distribution;
- ✓ approximate Binomial and Poisson Distribution by Normal Distribution.

Reference(s): Mason Chapters 6 and 7, Berenson Chapter 7, Newbold Chapter 4 and 5
Exercise(s): Seminars 16, 17, 18, 19 and 20, Mason Chapter6 Exercises 3, 7, 9, 15,19, 33, 35,
Mason Chapter7 Exercises 9, 11, 15, 19.

The outcomes of an experiment are said to be taking place randomly, since they cannot, by definition, occur in any particular order or pattern. Such variables, whose values thus cannot be known in advance by the person conducting the experiment, are called *Random Variables*.

Random Variables is a variable that takes on *numerical value* and is dependent on *chance*.

Random variables are denoted by *capitals* : X, Y, Z, ... etc. And the values it takes are denoted by *small letters* : x, y, z, ... or sometimes : x_1, x_2, x_3,... etc.

e.g. 1 A bond is given a rating.
Outcomes : Discrete rating A, B, C, D or E
Sample space : *Discrete (finite)*
Random variable : Define {Y = 1} if rating A
{Y = 2} if rating B
{Y = 3} if rating C
{Y = 4} if rating D
{Y = 5} if rating E

e.g. 2 Invest $10,000 in a common stock.
Outcomes : value of yield
Sample space : *Continuous (infinite)*
Random variable : Define X = value of yield
$(0 \le x \le \infty)$

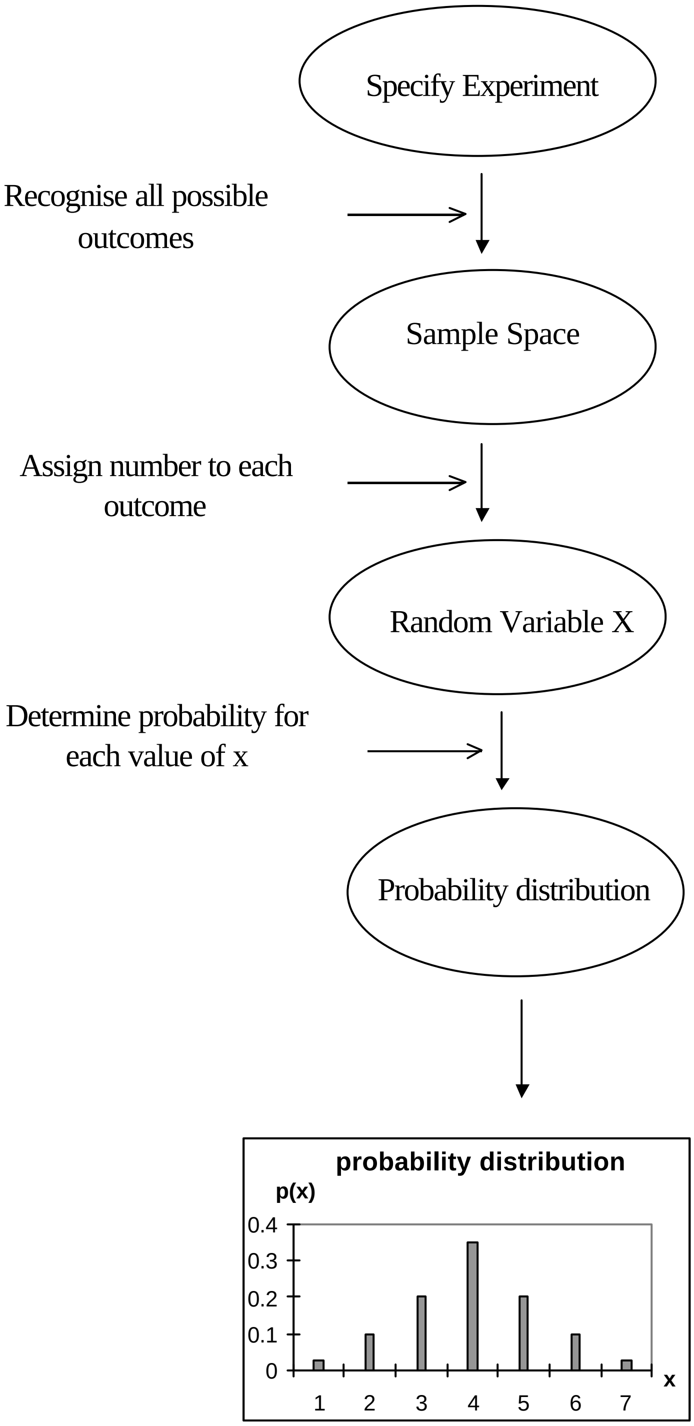
If a sample space contains a ***finite*** *number* of possibilities or an ***unbounded sequence*** with as many elements as there are *whole numbers*, it is called a ***Discrete*** *Sample Space*.

A random variable is called a *Discrete Random Variable* if its set of possible outcomes is *countable*.

If a sample space contains an ***infinite number*** of possibilities equal to the number of points on a line segment, it is called a *Continuous Sample Space*.

A random variable is called a *Continuous Random Variable* if it takes on values on a continuous scale.

A *probability distribution* is a specification (in a form of graph, a table or a function) of the probability associated with each value of the random variable.



Discrete Random Variable

A probability distribution involving only discrete value of x is usually called a ***Probability Mass Function*** *(p.m.f.).*

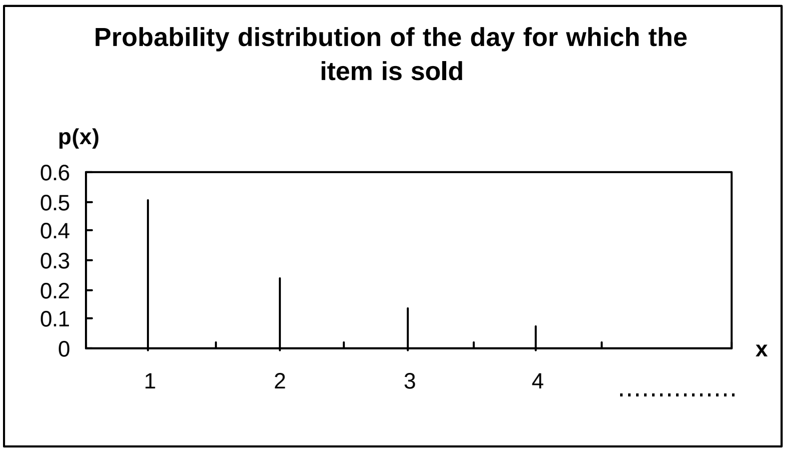
e.g.3 For a relatively low-selling item, the estimate is that there is a 50 -50 chance this item is sold on any given day.

Sol] Let X be the day on which the item is sold

P(X=1) = 1/2
P(X=2) = (1/2)(1/2) = 1/4
P(X=3) = (1/2)(1/2)(1/2) = 1/8
and so on

Although an infinite number of x values is possible, X is ***discrete random variable***

$$P(X=x) = \begin{cases} (1/2)^x & \text{for } x = 1, 2, 3, 4, \ldots, \infty \\ 0 & \text{otherwise} \end{cases}$$



Properties of Discrete Distribution

1. **P(X=x) = p(x)**

2. **p(x) ³ 0** for all x

3. å **p(x) = 1**

Cumulative Mass Function (c.m.f.) is the sum of values of the probability mass function for all values of the random variable x that are less than or equal to x. The value of the cumulative mass function at any given point x is usually denoted by the symbol *F(x)*.

$$\mathbf{F(x) = P(X \le x) = \sum_{i \le x} P(i)}$$

e.g.4 If 50% of the automobiles sold by an agency for a certain foreign car are equipped with diesel engine, find a formula for the probability distribution of the number of diesel models among the next 4 cars sold by this agency.

Sol] By counting rule, the total possible outcomes is 2 x 2 x 2 x 2 = 2^4 = 16 ways

In general, the event of selling x diesel models can occur in ${}_4C_x$ ways.

$p(x) = P(X=x) = {}_4C_x / 16$

$p(0) = {}_4C_0 / 16 = 1/16$
$p(1) = {}_4C_1 / 16 = 4/16$
$p(2) = {}_4C_2 / 16 = 6/16$
$p(3) = {}_4C_3 / 16 = 4/16$
$p(4) = {}_4C_4 / 16 = 1/16$

find the cumulative distribution of the random variable.

$F(0) = p(0) = 1/16$
$F(1) = p(0) + p(1) = 5/16$
$F(2) = p(0) + p(1) + p(2) = 11/16$
$F(3) = p(0) + p(1) + p(2) + p(3) = 15/16$
$F(4) = p(0) + p(1) + p(2) + p(3) + p(4) = 1$

Hence,

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1/16 & \text{for } 0 \le x < 1 \\ 5/16 & \text{for } 1 \le x < 2 \\ 11/16 & \text{for } 2 \le x < 3 \\ 15/16 & \text{for } 3 \le x < 4 \\ 1 & \text{for } x \ge 4 \end{cases}$$

e.g.5 A fair dice is tossed twice, find the probability distribution of ***sum*** of the outcomes of the experiment, and find the cumulative probability distribution of the ***sum*** of outcomes.

Sol] List all possible ***sum*** of outcomes:

1	2	3	4	5	6	7
8	9	10	11	12	13	14

determine the probabilities of each possible sum of outcomes:

x						
p(x)						
x						
p(x)						

cumulative probabilities distribution :

x						
F(x)						
x						
F(x)						

Expected Value of Discrete Random Variable:

The expected value of a random variable is the *center of gravity* or the balancing point of the value x "weighted" by their probability.

The expected value of a discrete random variable, x, is found by multiplying each x-value by its probability and then summing over all values of the random variable.

Expected value of X:

$$= E[X] = \sum_{\text{all } x} x\, p(x)$$

e.g.6 Inland Container company sells custom-made boxes. suppose these boxes must be ordered in units of 1000, 2000, 4000, 6000, or 12000. Based on past records, the p.m.f. are described as follows:

x (in ' 000)	1	2	4	6	12
P(X=x)	0.08	0.27	0.10	0.33	0.22

What is the expected value of X? What is the total number of boxes, in 500 orders, expected to be sold?

Sol] $E[X] = \sum x\, p(x)$

$= 1 \times 0.08 + 2 \times 0.27 + 4 \times 0.10 + 6 \times 0.33 + 12 \times 0.22$

$= 5.64$ (in ' 000 boxes)

$= 5640$ (boxes)

The total number of boxes sold :

$= 500 \times 5640$

$= 2820000$ (boxes)

Variance of Discrete Random Variable:

The variance of a random variable X is defined as the *expected squared deviation* of the values of the random variable about the mean $\{E[X] = \mu\}$.

The variable of a random variable is often denoted by the symbol **V[X]**.

$$V[X] = \sigma^2 = E[(x-\mu)^2]$$

The variance of a discrete random variable

$$\mathbf{V[X]} = \sigma^2 = \sum_{\text{all } x} (x-\mu)^2\, p(x)$$

the standard deviation is

$$\sigma = \sqrt{\mathbf{V[X]}}$$

e.g.7 Find the variance and the standard deviation of the box order example.

Sol] the variance is:

$$V[X] = \sigma^2 = \sum (x - 5.64)^2 p(x)$$
$$= (1 - 5.64)^2 \times 0.08 + (2 - 5.64)^2 \times 0.27 + (4 - 5.64)^2 \times 0.10 + (6 - 5.64)^2 \times 0.33 + (12 - 5.64)^2 \times 0.22$$
$$= 14.51 \text{ (boxes squared)}$$

the standard deviation is:

$$\sigma = \sqrt{V[X]}$$
$$= \sqrt{14.51}$$
$$= 3.81 \text{ (in ‘ 000 boxes)}$$

e.g.8 The director of a publication company is trying to decide how many copies to print everyday. Each copy costs $2.0 to print and sell for $5.0. Any copies unsold at the end of each day must be discarded. The director has estimated the following probability distribution for the copy sales, using data from past program sales:

copies sold	25,000	40,000	70,000
probability	0.10	0.60	0.3

The director has decided to print either 25-, 40-, or 70-thousands copies. Which number of copies will maximize the company’ s expected profit?

Sol]

copies printed	copies demanded	profit/loss	expected profit/loss
25,000	25,000	25,000 x (5 - 2) = 75,000	75,000
	40,000	25,000 x (5 - 2) = 75,000	
	70,000	25,000 x (5 - 2) = 75,000	
40,000	25,000	25,000x(5-2) +15,000(-2) =45,000	112,500
	40,000	40,000 x (5 - 2) = 120,000	
	70,000	40,000 x (5 - 2) = 120,000	
70,000	25,000	25,000x(5-2) + 45,000x(-2) = -15,000	97,500
	40,000	40,000x(5-2) + 30,000x(-2) = 60,000	
	70,000	70,000 x (5-2) = 210,000	

to maximize the company’ s expected profit, 40,000 copies should be printed, and the expected profit is 112,500.

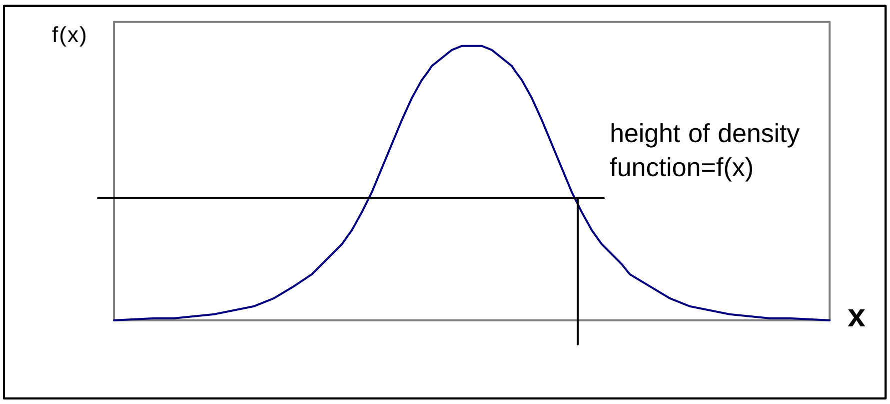
Continuous Random Variable

Probability function defined in terms of a *continuous random variable* are usually referred to as ***Probability Density Function*** *(p.d.f.).*

The symbol **f(x)** will be used to denote a probability density function.

For a continuous function, the value of f(x) must always be given by a *formula*.

The values of f(x) represents the ***height of the density function*** at the point x.

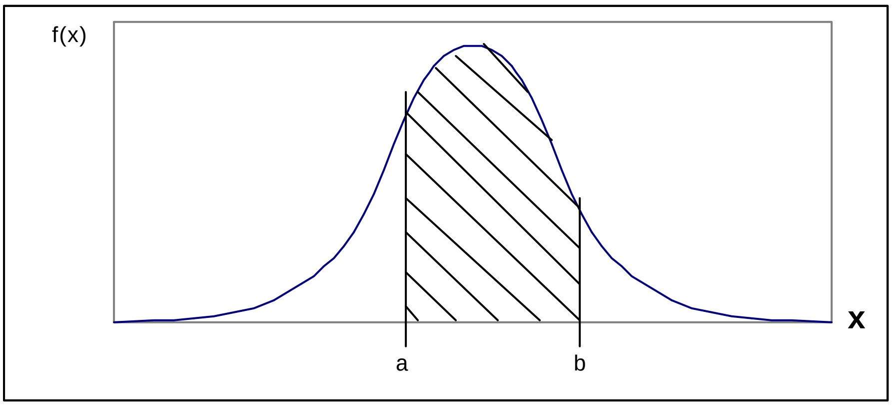


Students are only required to understand the concepts for continuous distribution, *like area enclosed = probability*. Absolutely no **INTEGRATION** is required.
Pages 8 to 11 are for reference only

The probability P(a ≤ X ≤ b) is always given by the *total* ***area*** under the density function f(x) from a to b.

$$P(a \le X \le b) = \text{Area under } f(x) \text{ from a to b}$$

$$= \int_a^b \mathbf{f(x)\,dx}$$



Properties of Continuous Distribution:

1. f(x) ≥ 0 for all x

2. Area under f(x) equals to 1

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

A Cumulative Density Function (c.d.f.) represents the probability that the random variable assumes a *value less than or equal to some specified value*. The symbol *F(x)* denotes a cumulative function, and is defined as

$$\mathbf{F(x) = P(X \le x)} = \text{area up to } (X = x)$$

$$= \int_{-\infty}^{x} \mathbf{f(x)\,dx}$$

$$\mathbf{P(a \le X \le b) = F(b) - F(a)}$$

e.g.9 A continuous random variable X having the probability density function

$$f(x) = \begin{cases} x^2/3 & -1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

a) Find $P(0 < X \leq 1)$.
b) Find $F(x)$, hence use it to evaluate $P(0 < X \leq 1)$.

Sol] a) $P(0 < X \leq 1) = \int_0^1 x^2/3 \, dx$

$= x^3/9 \Big|_0^1$

$= 1/9$

b) $F(x) = \int_{-\infty}^{x} f(t)\, dt$

$= \int_{-1}^{x} t^2/3 \, dt$

$= t^3/9 \Big|_{-1}^{x}$

$= 1/9\,(x^3 + 1)$

$$F(x) = \begin{cases} 0 & x \leq -1 \\ 1/9\,(x^3 + 1) & -1 \leq x \leq 2 \\ 1 & x \geq 2 \end{cases}$$

$P(0 < X \leq 1) = F(1) - F(0)$
$= 2/9 - 1/9 = 1/9$

Expected Value of Continuous Random Variable

$$\mathbf{E[X]} = \mu = \int_{\text{all } x} \mathbf{x\, f(x)\, dx}$$

Variance of Continuous Random Variable

$$\mathbf{V[X]} = \sigma^2 = \int_{\text{all } x} \mathbf{(x - \mu)^2 f(x)\, dx}$$

The standard deviation is denoted as σ ,

$$\sigma = \sqrt{\mathbf{V[X]}}$$

e.g.10 Find the standard deviation of the following function.

$f(x) = 0.08x$ for $0 \le x \le 5$

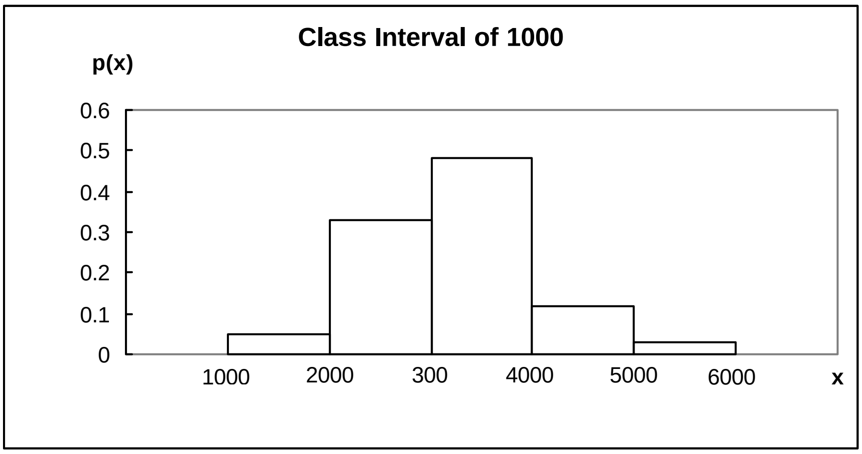
Sol] $E[X] = \mu = \int_{-\infty}^{\infty} x\, f(x)\, dx$

$$= \int_0^5 x\,(0.08x)\,dx = 0.08x^3/3 \Big|_0^5 = 3.33$$

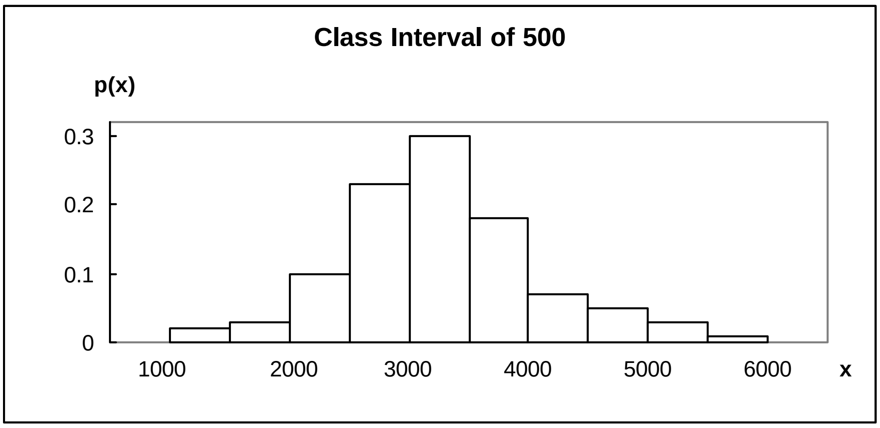
$V[X] = \sigma^2 = \int_0^5 (x - 3.33)^2\,(0.08x)\,dx = 1.41$

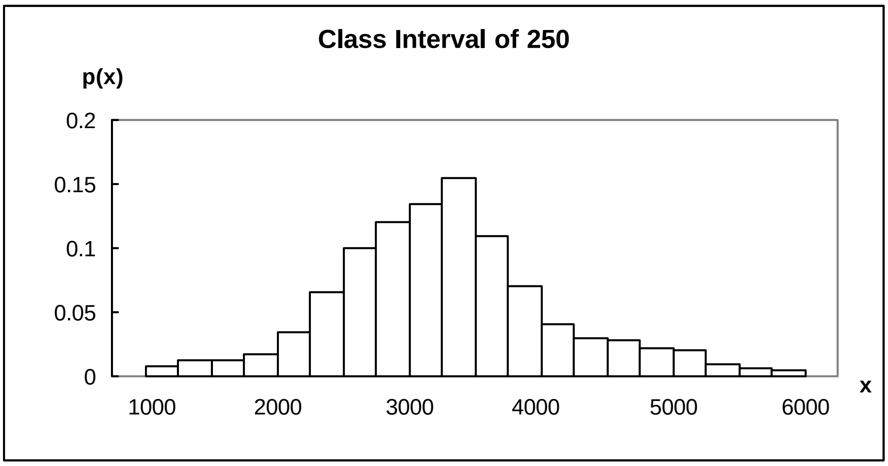
$\sigma = \sqrt{V[X]} = \sqrt{1.41} = 1.19$

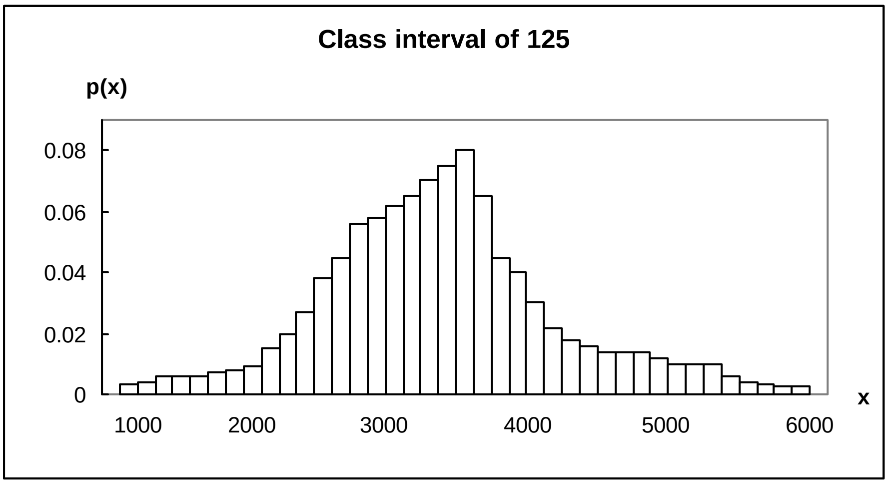
Continuous Random Variables are especially convenient to work with, so even when the set of outcomes is *discrete*, it is often advantageous to use *continuous approximation* to these values.



Suppose we decrease the width of the classes, from a class interval of 1000 to a class interval of say 500, or 250, or even to 1. The polygon begins to look more and more like a *smooth continuous function*.







As the width of the *class interval goes to zero*, the number of class under consideration must increase until, at the limit (number of class = ∞ *), there is an infinite number of such events between any two values of x*.

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$$

P(X = a) = 0	**for any value a**
P(X = b) = 0	**for any value b**

Discrete Probability Distribution: Binomial Distribution

Many experiment share the common element that their outcomes can be classified into one of two events. Each repetition of an experiment involving only two outcomes (for example, each toss of a coin) is called a ***Bernoulli*** *Trial*.

For the purpose of probability theory, interest centers not on a single Bernoulli trial but rather on a *series* of *independent*, *repeated* Bernoulli trial.

Independent : means that the result of any one trial *cannot* ***influence*** *the results of any other trial*.

Repeated : means that the condition under which each trial is *held must be an exact replication of the conditions* underlying all other trials. It implies that the *probability of the two possible outcomes* ***cannot change*** *from trial to trial*.

In a Binomial Distribution, the probabilities of interest are those *of receiving a certain number of success*, **x**, in ***n*** *independent Bernoulli trials*, each trail having only two possible outcomes and the *same probability*, ***p***, *of success*.

Suppose we want x success out of n trials, thus, there will be (n-x) failures. To calculate such probabilities, it is necessary to find the probability of one occurrence of this type, and then multiply this probability by the number of such occurrence.

p(event) = (Number of relevant occurrence) x p(one occurrence)

Since it doesn't make any difference which occurrence we investigate first, let's take the one in which the x successes occur first, and followed by the (n-x) failures.

If we let S = success, and F = failure, then this particular order can be represented as :

S S S S S S . . . S S S S ***F F F F F F . . . F F F F***
x successes ***(n-x) failures*** **n trials**

To determine the joint probability of this particular sequence of success and failures, recall that all trials are *independent*. Since the probability of a success is $p(S) = p$, and the probability of a failure is $p(F) = (1 - p) = q$, the following holds :

$$\begin{aligned} P(SS...SFF...F) &= p(S) \times p(S) \times ... \times p(S) \times p(F) \times p(F) \times ... \times p(F) \\ &= p \times p \times ... \times p \times q \times q \times ... \times q \\ &= \mathbf{p^x q^{n-x}} \end{aligned}$$

Combinations of n objects taken x at a time = $\mathbf{{}_nC_x}$

Therefore, the formula for the Binomial Distribution :

$$\mathbf{P(x \text{ successes in n trials}) = P(X = x \mid n, p) = {}_nC_x\, p^x q^{n-x}} \quad \textbf{for} \begin{cases} n = 1, 2, 3, ... \\ x = 0, 1, 2, ..., n \end{cases}$$

X ~ B (n , p) **or** **b(x : n, p)**

e.g.11 The probability that a certain kind of component will survive a given shock test is 3/4. Find the probability that only 2 of the next 4 components tested survive?

Sol] Let X be the no. of component survive after the shock test.
$n = 4, x = 2, p = 3/4, q = 1-p = 1/4$

$X \sim B(4 , 3/4)$
$P(X=x) = p(x) = {}_nC_x\, p^x q^{n-x}$

$$\begin{aligned} P(X=2) &= p(2) = {}_4C_2\, (3/4)^2 (1/4)^{4-2} \\ &= 6\, (9/16)\, (1/16) \\ &= 54/256 \end{aligned}$$

e.g.12 Find the probability that less than 2 of the next 4 components tested survive ?

Sol] P(X < 2) = P(X=) + P(X=)
=
=
=

Necessary Conditions for Binomial Distribution

1. each observation can be classified as one of *two mutually exclusive events.* (i.e. *success or failure*)

2. the probability for the two possible outcomes must be *constant* from observation to observation.

3. the result of any observation is *independent* to the result of any other observations.

e.g.13 A box contains 4 red balls and 6 black balls. Two balls are drawn from the box *without replacement*. Find the probability that 2 red balls are selected?

Sol] P(1st ball is red) = *4/10*, P(2nd ball is red) = *3/9*

p is not constant from observation to observation, so the probability does not follow Binomial Distribution.

so, P(2 red balls drawn) = P(1st ball is red) x P(2nd ball is red)
= 4/10 x 3/9
= 2/15

e.g.14 Find the probability that 2 red balls are selected from a box containing 4 red balls and 6 black balls, if the 2 balls are *drawn with replacement* ?

Sol] check conditions :
i) two mutually exclusive outcomes only [red or not red]
ii) constant probability of success [p(red ball) = 4/10]
iii) independent observations [the selection of first ball would not affect the selection of second ball]

So, binomial distribution is followed.

Let X be the number of red ball selected.
n = 2, p(red ball) = p = 4/10, x = 2

$X \sim B(n , p) \Rightarrow X \sim B(2 , 0.4)$

$P(\text{2 red balls}) = P(X=2) = {}_2C_2\,(4/10)^2\,(6/10)^{2-2}$
$= 1 \times 4/25 \times 1$
$= 4/25 = 0.160$

e.g.15 The probability that a salesperson will sell a magazine subscription to someone who has been randomly selected from the telephone directory is 0.2. If the salesperson calls 12 individuals this evening, what is the probability that:

a) there are no subscriptions sold?
b) exactly two subscriptions will be sold?
c) at least two subscriptions sold?

Sol] check conditions :

i) two mutually exclusive outcomes only [ ]
ii) constant probability of success [ ]
iii) independent observations [ ]

So, is followed.

Let X be .
n = 12, p = 0.2,

$X \sim B(n, p) \Rightarrow X \sim B(\quad)$

a) P(no subscription) = P(X=0) = C (0.2) (0.8)
=
=

b) P(two subscriptions) = P(X=2) = C (0.2) (0.8)
=
=

c) P(at least two subscriptions) = P(X=2) + P(X=3) + ... + P(X=12)
=
=
=
=

Expected Value and Variance of the Binomial Distribution

The expected value of a **Binomial Distribution b(x : n, p)** is:

$$E[X] = \mu = np$$

Proof: Let the outcome on the k^{th} trial be represented by a Bernoulli random variable I_k, which assumes the values 0 and 1 with the probabilities q and p respectively.

Therefore, in a binomial experiment the number of success can be written as the sum of the n independent indicator variables. Hence,

$X = I_1 + I_2 + I_3 + ... + I_n$

The expected value of any I_k is $E[I_k]$,

$$E[I_k] = (0)\,q + (1)\,p = p$$

Therefore, the expected value of the Binomial Distribution is :

$$\mathbf{E[X]} = E[I_1] + E[I_2] + E[I_3] + ... + E[I_n] = p + p + p + ... + p = \mathbf{np}$$

The variance of a **Binomial Distribution b(x : n, p)** is :

$$\mathbf{V[X] = s^2 = npq}$$

proof]

$$E[I_k^2] = (0)^2 q + (1)^2 p$$
$$= p$$

The variance of any k^{th} trial is given by :

$$V[I_k] = \sigma^2_{I_k} = E[(I_k - p)^2]$$
$$= E[I_k^2] - E[p^2]$$
$$= p - p^2$$
$$= p(1-p)$$
$$= pq$$

therefore, the variance of the Binomial distribution is :

$$\mathbf{V[X]} = \sigma^2_x = \sigma^2_{I_1} + \sigma^2_{I_2} + \ldots. + \sigma^2_{I_n}$$
$$= pq + pq + \ldots + pq$$
$$\mathbf{= npq}$$

and the standard deviation of the Binomial Distribution is :

$$s_x = \sqrt{V[X]}$$

$$s_{\mathbf{x}} = \text{Ö}\mathbf{npq}$$

e.g.16 Refer to example 11, the probability that a certain kind of component will survive a given shock test is 3/4. Find the mean, variance and standard deviation of the distribution?

Sol] $n = 4$, $p = 3/4$

$\mu = E[X] = np$
$= 4(3/4)$
$= 3$

$\sigma^2 = V[X] = npq$
$= 4(3/4)(1-3/4)$
$= 3/4$

$\sigma = \sqrt{V[x]}$
$= \sqrt{3/4}$
$= 0.8660$

e.g.17 Refer to example 15, find the mean, variance and standard deviation of the distribution?

Sol] n = , p = ,

$\mu = E[X] = np$
=
=

$\sigma^2 = V[X] = npq$
=
=

$\sigma = \sqrt{V[x]}$
=
=

Class Exercise 1

For an audio retail shop, there is a 10% chance that an incoming customer would make a purchase before he/she leaves the shop. In a given day, 40 customers visited the shop. What is the probability that not less than 3 customers made purchases?

Discrete Probability Distribution : Poisson Distribution

Another important Discrete Distribution, the *Poisson Distribution*, has Recently found fairly wide application, especially in the area of operations research.

The Poisson Distribution can be used to ***determine the probability of x occurrence <u>per unit time</u>***, if four basic assumption are met :

1. Possible to divide time interval of interest into many sub-intervals.
2. Probability of an occurrence remains constant through the time interval.
3. Probability of two or more occurrences in a sub-interval is small enough to be ignored.
4. Independent of occurrences.

Area of application :

- the number of telephone calls *per hour* received by an office.
- the number of customers *per hour* arrived.

The only parameter necessary to characterize a population described by the Poisson Distribution is the ***Mean Rate*** at which events take place in each unit of time. And it is represented by a Greek letter *lambda (l)*.

Probability Distribution of Poisson Distribution :

P(**x occurrence in a given unit time**)

$$P(X=x) = \begin{cases} e^{-l}\, l^{x} / x! & \text{for } l > 0,\ x = 0, 1, 2, \dots \\ 0 & \text{elsewhere} \end{cases}$$

X ~ P (l) or **p(x : l)**

e.g.18 The average of traffic accidents on a certain section of highway is two per week. Assume that the number of accidents follow a Poisson Distribution.

a) Find the probability of no accident on this section of highway during a one-week period.

b) Find the probability of at most three accidents on this section of highway during a one-week period.

Sol] a) Let X be the no. of accidents per week.

The probability of having x accidents:

$P(X=x) = p(x) = e^{-\lambda} \lambda^{x} / x!$ and with $\lambda = 2$

$\Rightarrow \quad P(X=x) = e^{-2} 2^{x} / x!$

Probability of no accident = $P(X=0) = p(0)$
$= e^{-2} 2^{0}/0!$
$= e^{-2}$
$= 0.1353$

b) $P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$
$= e^{-2} 2^{0}/0! + e^{-2} 2^{1}/1! + e^{-2} 2^{2}/2! + e^{-2} 2^{3}/3!$
$= 0.1353 + 0.2707 + 0.2707 + 0.1804$
$= 0.8571$

In general, the formula for Poisson Distribution is :

$$\mathbf{P(X=x \mid l, t) = e^{-lt} (l t)^{x} / x!}$$

where:

x - no. of occurrence
λ - average no. of occurrence per unit time
t - no. of unit time measured
e - the base of natural logarithms (e = 2.71828)

e.g.19 Suppose that the number of customer order for a particular computer keyboard is distributed with a Poisson Distribution and the average number of order per hour is 2. What is the probability of getting 20 orders by the end of an eight hours working day?

Sol] λ = average no. of order per hour = 2,
t = no. of working hour in a day = 8,

let X be the no. of order in an eight hours working day
$P(X=x \mid \lambda, t) = e^{-\lambda t} (\lambda t)^x / x!$

$P(X=x \mid 2, 8) = e^{-(2x8)} (2x8)^x / x!$
$= e^{-16} 16^x / x!$

P(getting 20 orders in a day) = P(X=20 |2, 8)
$= e^{-16} 16^{20} / 20!$
= 0.0559

e.g.20 A natural gas exploration company average 4 strikes (that is, natural gas is found) per 100 holes drilled. If 25 holes are to be drilled, Assuming that the number of strike found follows Poisson distribution, what is the probability that :

a) exactly 1 strikes will be found?
b) at least 2 strikes will be found ?

Sol] λ = average no. of strikes found per interval (100 holes) = ,
t = interval measured = ,

let X be .

$P(X=x \mid \lambda, t) = e^{-\lambda t} (\lambda t)^x / x!$

P(X=x | ,) = $e^{-(\;)}$ ($)^x$ / x!
= e ^{x} / x!

a) P(exactly 1 strikes) = P(X= | ,)
=
=

b) P(at least 2 strikes) = P(X=2) + P(X=3) + ...
= 1 -
= 1 -
= 1 -
=

Mean and Variance of Poisson Distribution

For a Poisson Distribution, $p(x \mid \lambda)$,

Mean = m **=** l **Variance =** s^{2} **=** l

For a Poisson Distribution, $p(x \mid \lambda, t)$,

Mean = m **=** l **t** **Variance =** s^{2} **=** l **t**

e.g.21 Refer to example 19, , find the mean, variance and standard deviation of number of order in an eight hours working day?

Sol] mean = $\mu = \lambda t = 2 \times 8 = 16$

variance = $\sigma^2 = \lambda t = 2 \times 8 = 16$

standard deviation = $\sigma = \sqrt{\lambda t} = \sqrt{16} = 4$

e.g.22 Given a Binomial Distribution $b(x \mid 1000, 0.004)$, find the probability that the random variable will assume a value 10.

Sol] $P(X=10) = {}_{1000}C_{10}\,(0.004)^{10}\,(0.996)^{990}$

$= ????$

Approximation to Binomial Distribution

We can use the Poisson Probability distribution to approximate the Binomial Probability Distribution when:

1. **n is large, normally greater than 100**, *and*

2. **p is small, preferably close to zero**.

e.g.23 Given a Binomial Distribution b(x |1000, 0.004), find the *approximate* probability that the random variable will assume a value 10.

Sol] n is large (n = 1000 > 100), and p is small (p = 0.004, it is close to zero). We can approximate by Poisson Distribution using,

$\lambda = \mu = np = 1000 \times 0.004 = 4$

$P(X=10) = {}_{1000}C_{10}\,(0.004)^{10}\,(0.996)^{990}$
$\cong e^{-4}\ 4^{x} / x!$
$= e^{-4}\ 4^{10} / 10!$
$= 0.0053$

e.g.24 In a manufacturing process, it is known that on the average 1 in every 1000 of the items produced has one or more defects. What is the probability that a random sample of 3000 will yield fewer than 3 items with defects?

Sol] It is a Binomial experiment with n = and p = . Since it satisfies the conditions of Binomial distribution.

however, n is large (n = > 100), and p is small (p = , it is close to zero). We can approximate by Distribution using,

$\lambda = \mu = np = \qquad =$

Hence, if X representing .

$P(X < 3) = \sum_{x=0}^{x=2} \mathbf{b}(x \mid \quad , \quad)$

$\approx \sum_{x=0}^{x=2} \mathbf{p}(x \mid \quad)$

$= e^{-3}\ 3 \ / \ \ ! + e^{-3}\ 3 \ / \ \ ! + e^{-3}\ 3 \ / \ \ !$

$= \qquad + \qquad +$

$=$

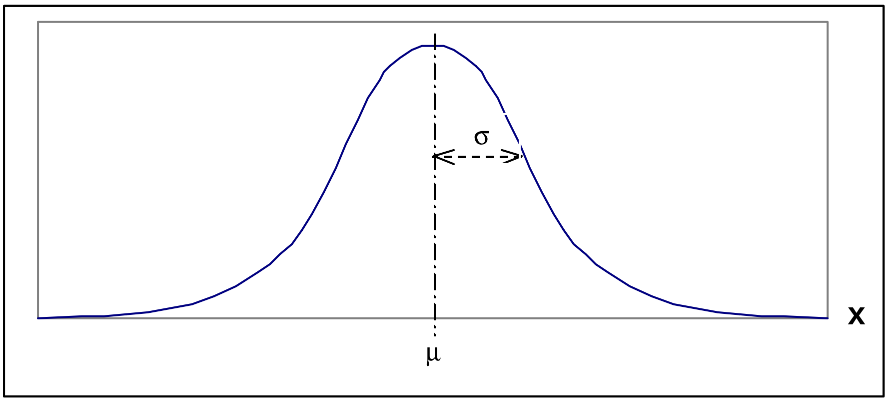
Class Exercise 2:

Suppose that the average number of goals per game scored by the BA soccer team in last season is 2.5. What is the probability that in any one game, three or four goals will be scored?

Continuous Probability Distribution : Normal Distribution

In some cases, measurements can be of infinitely many values corresponding to points on a line interval, e.g. heights and weights. These are known as *continuous random variable*. The most important continuous probability distribution in the field of statistics is the *Normal Distribution*.

A continuous random variable X having the ***bell-shaped distribution*** is called a Normal random variable.



The Normal Distribution is a continuous distribution in which X can *assume any value between* ***minus infinity*** *and* ***plus infinity*** *(-¥ < x < ¥)*. The curve is *completely* ***symmetrical about the mean***.

Two parameters describe the Normal Distribution, *m representing the* ***mean***, and *s representing the* ***standard deviation***.

The formula of Normal Distribution is given as:

$$f(x) = \frac{1}{s\sqrt{2p}} e^{-(1/2)\,[(x-m)/s]^2}$$

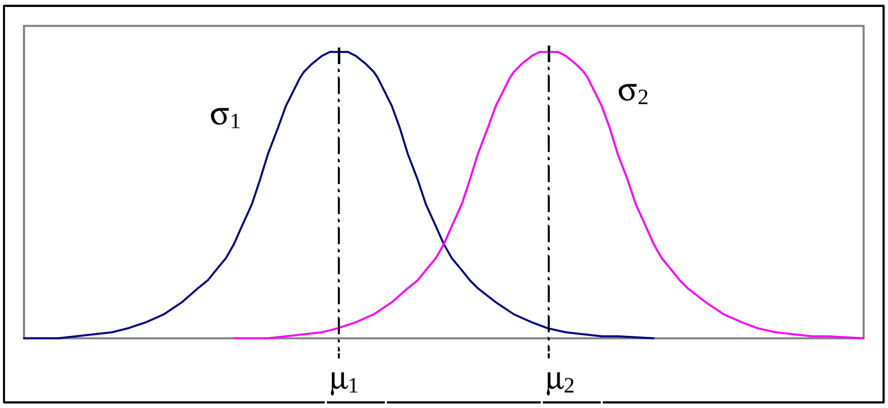
where : π = 3.14159.....
e = 2.71828.....

In labeling the formula, we use the symbol :

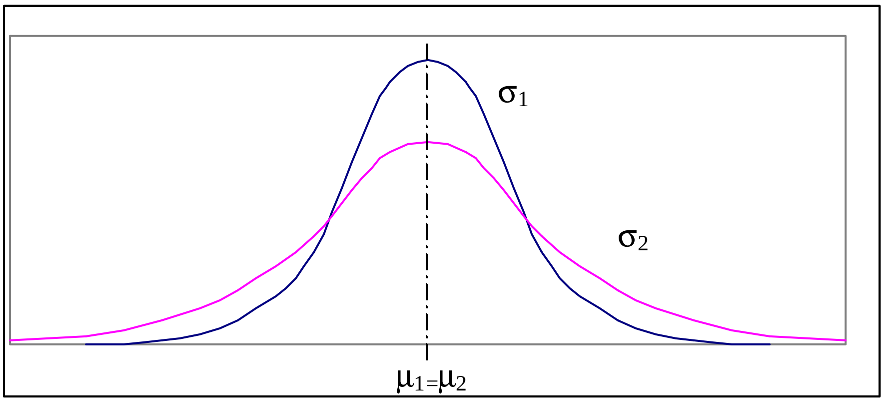
X ~ N(m , s²)	**or**	**n(x: m , s²)**

E[X] = mean = μ

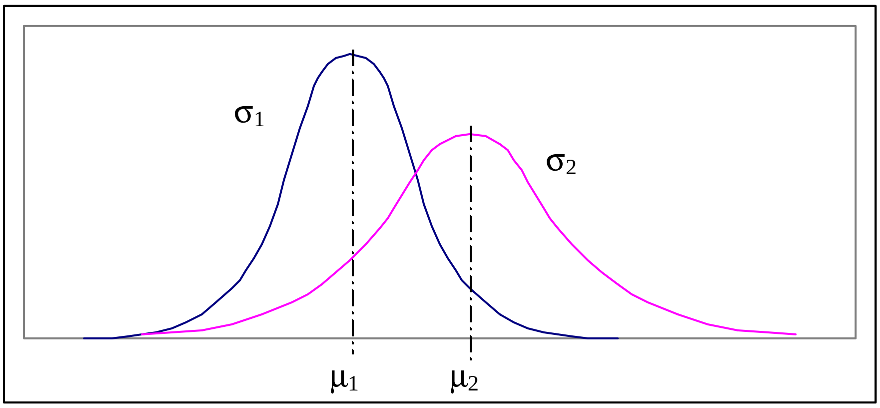
V[X] = variance = σ^2



Normal Curves with $m_1 < m_2$**, and** $s_1 = s_2$



Normal Curves with $m_1 = m_2$**, and** $s_1 < s_2$



Normal Curves with $m_1 < m_2$**, and** $s_1 < s_2$

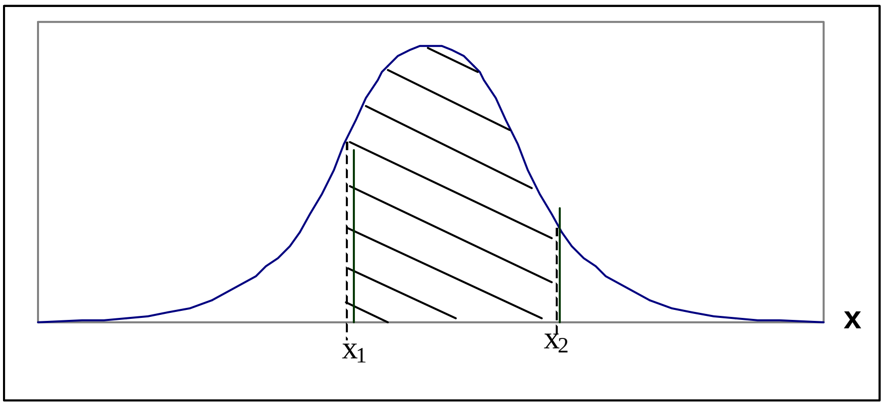
It is important to remember that *all Normal Distributions have the bell-shaped curve regardless of the values of* m *and* s.

Area under the Normal Curve

The area under the curve bounded by the two ordinates $X=x_1$ and $X=x_2$ equals the probability that the random variable X assumes a value between $X=x_1$ and $X=x_2$, (i.e. $P(x_1<X<x_2)$)

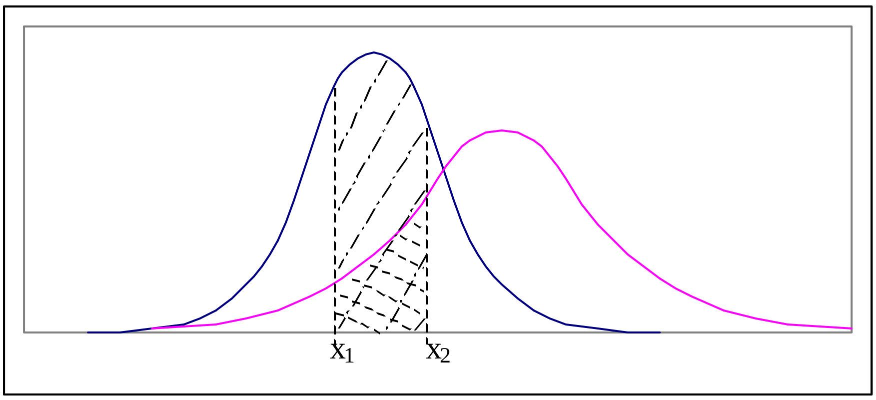
$$P(x_1<X<x_2) = \int_{x_1}^{x_2} f(x)\,dx$$

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



$P(x_1 < X < x_2)$ = **area of the shaded region**.

However, the probability $P(x_1 < X < x_2)$,(i.e. area of the shaded region) would be different for different curves, either different in μ or σ or both.



The probability can be determined in two ways !!!

1. use calculus to integrate the Normal function with the specified μ and σ from x_1 to x_2.

2. use a table of probability already calculated for the precise parameters μ and σ.

But both of these methods are unsatisfactory, since it is too tedious to evaluate the integral, and these is no such table exist.

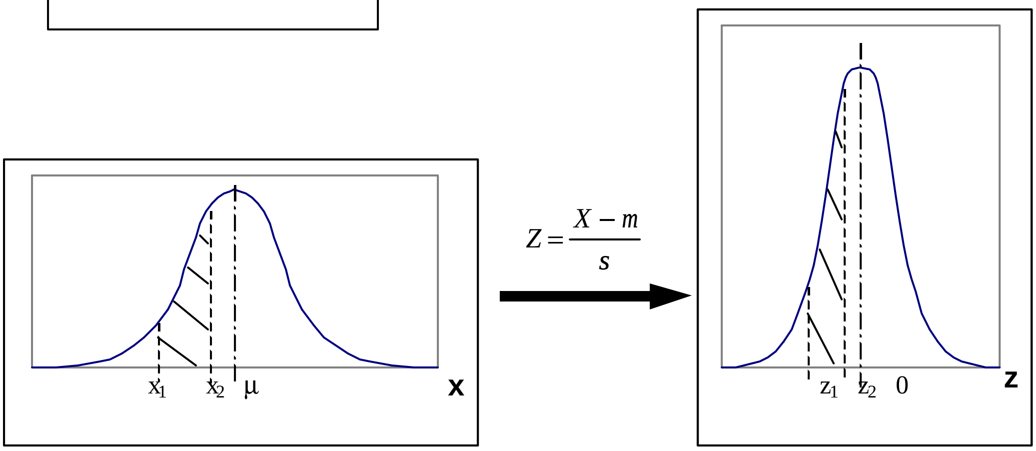
However, there is a table calculated for Normal function with μ = 0 and σ = 1.

$N(0, 1^2)$

The distribution of a Normal random variable with ***mean 0 and variance 1*** is called a ***Standard Normal Distribution***.

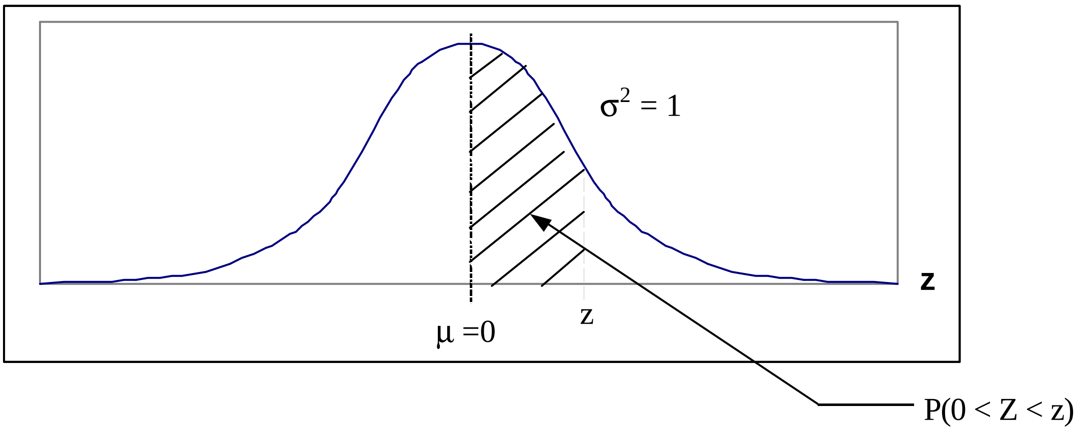
We are able to transform all the observations of ***any normal random variable X*** to a new set of observation of a ***normal random variable Z with mean 0 and variance 1***. By the transformation :

$$Z = \frac{X - m}{s}$$



Use of the Standard Normal Table

Table 3 gives area under the standard normal curve corresponding to P(0 < Z < z) for the values of z ranging from –3.09 to +3.09



e.g.25 Given a Standard Normal Distribution, find,

a) P(Z > 1.84),
b) P(Z > -1.84),
c) P(-1.97 < Z < 0.86)

Sol] a) P(Z > 1.84) = 0.5 – 0.4671 = 0.0329 (from Mason appendix D)

b) P(Z > -1.84) = 0.5 +0.4671= 0.9671 (from Mason appendix D)

c) P(-1.97 < Z < 0.86) = 0.3051 + 0.4756 (from Mason appendix D)
= 0.7807

e.g.26 Given a Standard Normal Distribution, find,

a) P(Z > 2.05),
b) P(Z > -1.5),
c) P(1.1 < Z < 2.2),
d) P(Z > -5)

Sol] a) P(Z > 2.05) = = (from Mason appendix D)

b) P(Z > -1.5) = = (from Mason appendix D)

c) P(1.1 < Z < 2.2) = (from Mason appendix D)
=

d) P(Z > -5) =

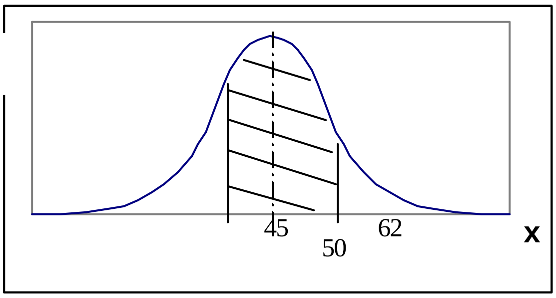
e.g.27 Given a Normal Distribution with $\mu = 50$ and $\sigma = 10$, find the probability that X assumes a value between 45 to 62.

Sol] Transform the distribution to Standard Normal Distribution by $Z = (X-m)/s$.

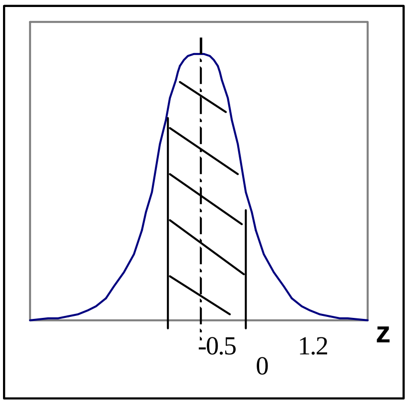
The Z-values corresponding to $x_1 = 45$ and $x_2 = 62$ are :

$z_1 = (45 - 50)/10 = -0.5$
$z_2 = (62 - 50)/10 = 1.2$

Therefore, $\mathbf{P(45 < X < 62)}$



$\mathbf{= P(-0.5 < Z < 1.2)}$



= 0.1915 + 0.3849 (from Mason appendix D)
= 0.5764

e.g.28 Given a Normal Distribution with $\mu = 28$ and $\sigma = 3.4$, find the probability that X assumes a value between 30 to 36.5.

Sol] Transform the distribution to Standard Normal Distribution by $Z = (X-m)/s$.
The Z-values corresponding to $x_1 = 30$ and $x_2 = 36.5$ are :

$z_1 =$ $\quad = \quad \cong$

$z_2 =$ $\quad =$

P(30 < X < 36.5)
= P(< Z <)
= (from Mason appendix D)
=

e.g.29 Given a Normal Distribution with mean 40 and variance 36. Find the value of x that has :

a) 45% of the area to the left,

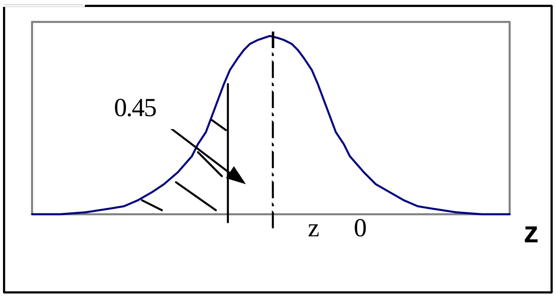
b) 14% of the area to the right.

Sol] a) $X \sim N(40, 6^2)$

$P(X < x) = 0.45$

$\Rightarrow P((X-\mu)/\sigma < (x-40)/6) = 0.45$

$\Rightarrow P(Z < (x-40)/6) = 0.45$



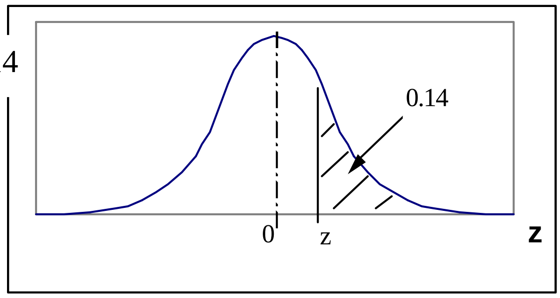
From table 3, z = ***-0.13*** (from Mason appendix D)

$\Rightarrow (x - 40)/6 = -0.13$

$\Rightarrow x = 39.22$ (correct to 2 decimal places)

b) $P(X > x) = 0.14$

$\Rightarrow P(Z > (x-40)/ 6) = 0.14$



From table 3, z = ***1.08*** (from Mason appendix D)

$\Rightarrow (x - 40)/ 6 = 1.08$

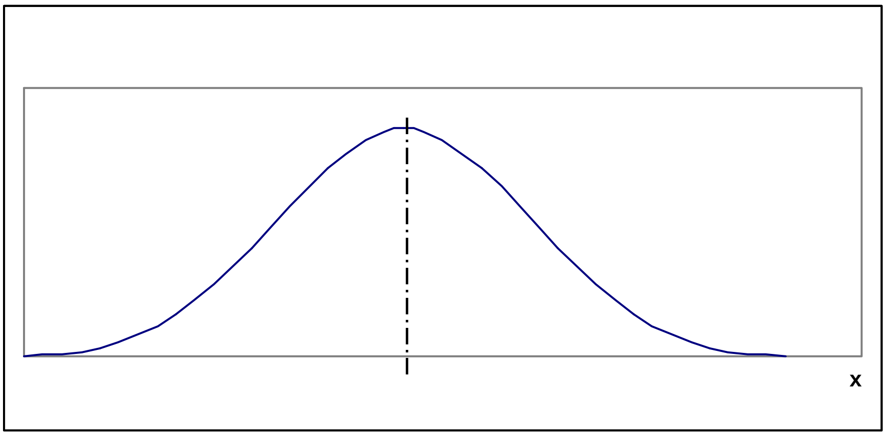
$\Rightarrow x = 46.48$ (correct to 2 decimal places) e.g.30 Given a Normal Distribution with mean 20 and variance 10. Find the value of x that has 33% of the area to the left,

Sol] X ~ N(, ^{2})

P(X < x) = 0.33

$\Rightarrow$ P((X-μ)/σ < ( )/) = 0.33

$\Rightarrow$ P(Z < ( )/) = 0.33



Sketch a normal curve, and identify the required area

From table 3, z = (from Mason appendix D)

$\Rightarrow$ ( )/ =

$\Rightarrow$ x = (correct to 2 decimal places)

e.g.31 Given a Binomial Distribution b(x |100, 0.5), find the probability that the random variable will assume a value more than 40 ?

Sol] n = 100, p = 0.5, q = 1 - 0.5 = 0.5

P(X>40) = P(X=41) + ... + P(X=100)

or = 1 - [P(X=0) + P(X=1) + ... + P(X=39) + P(X=40)]

= ????

Normal Approximation to the Binomial Distribution

Normal Distribution provides a very close approximation to the Binomial Distribution when :

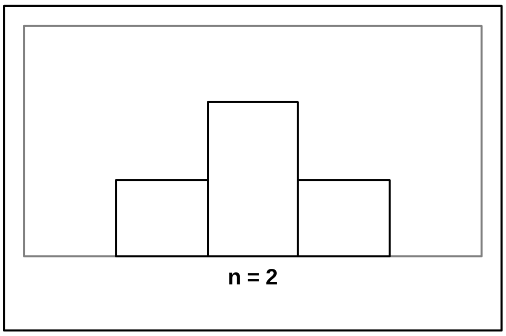
1) n, the number of trials, is very ***large*** , and
2) p, the probability of a success on an individual trial, is *close to* ***0.5***.

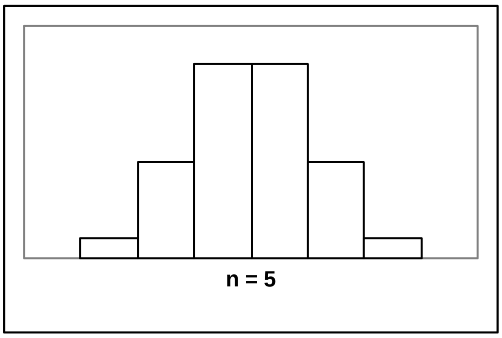
The Normal curve will have :

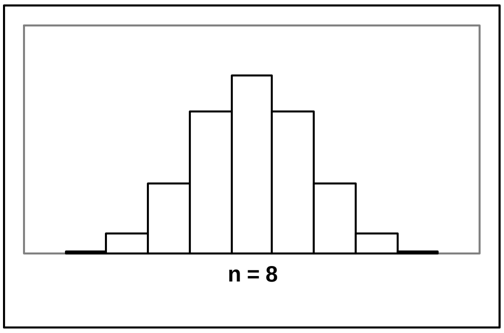
> **the mean** = μ = **np**
>
> **the standard deviation** = σ = $\sqrt{npq}$

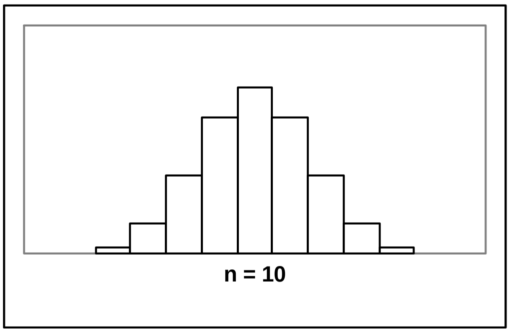
Note : a good rule of thumb is to use this approximation only when *np and n(1-p) are both greater than or equal to* ***5*****.**

The following figures show that, for a Binomial Distribution with p = 0.5, larger the value of n, the distribution will be smoother and be bell-shaped.

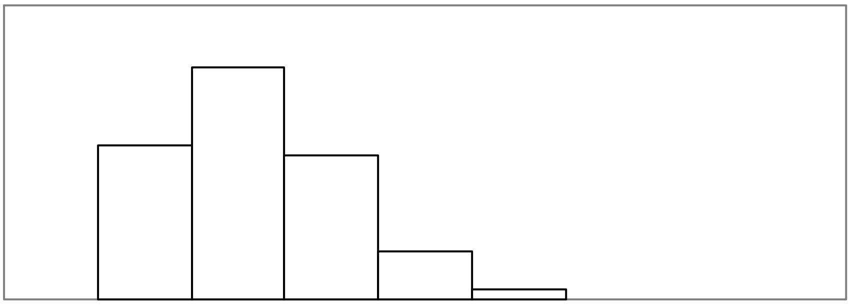




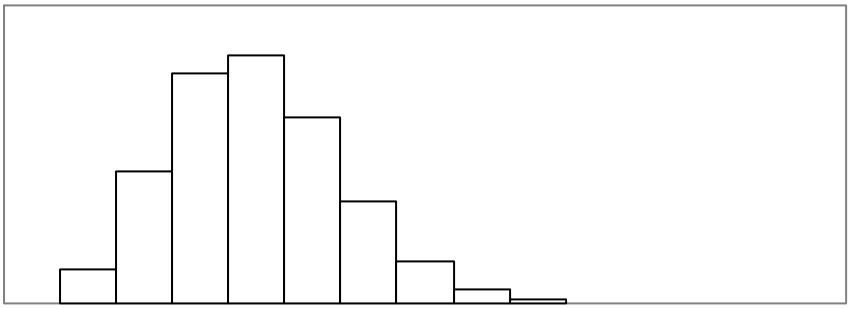




And even when the probability of success of the Binomial Distribution is *not close to 0.5*, but *with large n*, the distribution will also looks like bell-shape.



p = 0.2, n = 6



p = 0.2, n = 15

Therefore, before doing Normal approximation to Binomial distribution, the conditions must be checked.

$np \geq 5$	and	$nq \geq 5$

Another important point to be noted for Normal Approximation is the transformation of the distribution *from* ***discrete to continuous***. Let' s look at the example below:

e.g.32 Given a random variable X follows Binomial Distribution with n=10, and p=0.5. Using Normal Approximation to find the probability P(4≤ X ≤ 7).

Sol] check conditions:

Since p is close to 0.5, and both np and nq equal to 5, so Normal approximation can be used.

From the first figure, we find that the probability P(4≤ X ≤ 7) should equals to the total area of the 4 bars (i.e. the shaded region 1).

However, if we just find the probability P(4≤ X ≤ 7) for the Normal Distribution, the probability found would be equal to the area of shaded region 2.

And this approximated probability (shaded region 2) is different from the actual probability (shaded region 1) significantly.

So we have to modify the approximation by *adding a correctional factor*, which move the lower and upper boundaries for the Normal approximation,

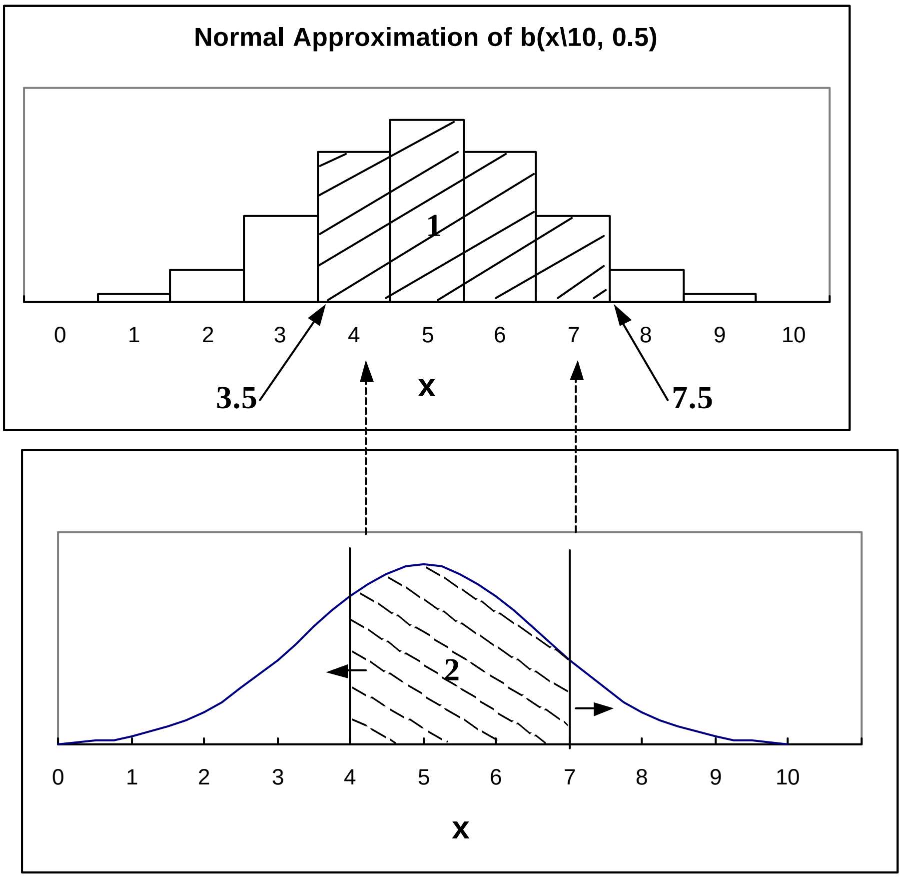
μ = np = 10 x 0.5 = 5, $\sigma = \sqrt{npq} = \sqrt{10x0.5x0.5}$ = 1.581

P(4 £ X £ 7 | X~ b(n, p)) (X follows Binomial Distribution)

= P(4 **- 0.5 <** X < 7 **+ 0.5** | X ~N(μ, σ^2)) **(add continuity adjustment)**

= P(3.5 **<** X < 7.5 | X ~N(μ, σ^2)) (X follows Normal Distribution)

= P(-0.9487 < Z < 1.5811)
= P(-0.95 < Z < 1.58)
= 0.3289 + 0.4429 (from Mason appendix D)
= 0.7718



e.g.33 Given a Binomial Distribution b(x |100, 0.5), find the probability that the random variable will assume a value more than 40?

Sol] n = , p = , q = =

check conditions: np = and nq = , so approximation can be used.

μ = np = = , $\sigma = \sqrt{npq}$ = =

P(X > 40 | X~ b(n, p)) (X follows Binomial Distribution)

= P(X > | X ~N(μ, σ^2)) **(add continuity adjustment)**

= P(X > | X ~N(μ, σ^2)) (X follows Normal Distribution)

= P(Z >)
= P(Z >)
= (from Mason appendix D)
=

Normal Approximation to the Poisson Distribution

The Normal Distribution may also be used to approximate the Poisson Distribution whenever **the parameter λ, the expected number of successes in a given unit time, greater than or equal to 5**.

The Normal curve will have:

> **the mean** = $\mu = \lambda$
> **the standard deviation** = $\sigma = \sqrt{\lambda}$

before doing Normal approximation to Poisson distribution, the condition must be checked.

> $\lambda \geq$ **5**

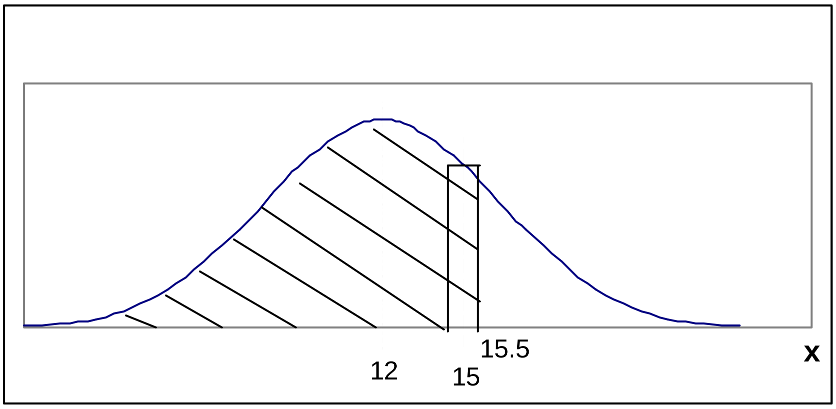
e.g.34 Suppose that at an automobile plant, the average number of work stoppages per day due to equipment problems during the production process is 12.0. What then is the approximate probability of having 15 or fewer work stoppages due to equipment problems on any given day?

Sol] Let X be the number of stoppages in a given day,
and X follows Poisson Distribution with mean rate $\lambda = 12$.

check conditions: Since $\lambda = 12 > 5$, Normal approximation can be used,

mean = $\mu = \lambda = 12$
standard deviation = $\sigma = \sqrt{\lambda} = \sqrt{12}$

$P(X \leq 15 \mid X \sim p(\lambda))$ (X follows Poisson Distribution)
$= P(X < 15 + \mathbf{0.5} \mid X \sim N(\mu,\sigma^2))$ **(add continuity adjustment)**
$= P(X < 15.5 \mid X \sim N(\mu,\sigma^2))$ (X follows Normal Distribution)



$= P(Z < (15.5 - 12)/\sqrt{12})$
$= P(Z < 1.01)$
$= 0.5 + 0.3438$ (from Mason appendix D)
$= 0.8438$

e.g.35 Given a random variable X follows Poisson Distribution with mean rate 10. Find the approximate probability $P(7 \le X < 13)$.

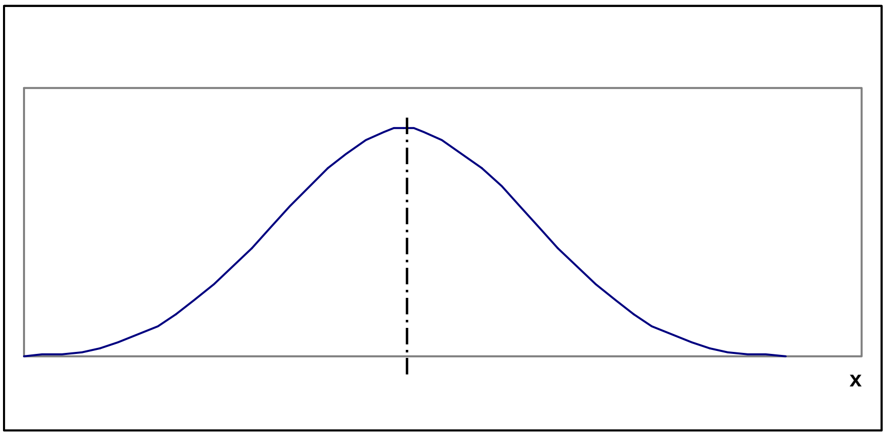
Sol] check conditions: Since $\lambda =$ $>$, approximation can be used.

mean $= \mu = \lambda =$; standard deviation $= \sigma = \sqrt{\lambda} =$

$P(7 \le X < 13 \mid X \sim p(\lambda))$ (X follows Poisson Distribution)

$= P(\quad < X < \quad \mid X \sim N(\mu,\sigma^2))$ **(add continuity adjustment)**

$= P(\quad < X < \quad \mid X \sim N(\mu,\sigma^2))$ (X follows Normal Distribution)



Sketch a normal curve, and identify the required area

$= P(\quad < Z < \quad)$

$= P(\quad < Z < \quad)$

$=$ $+$ (from Mason appendix D)

$=$

Class Exercise 3:

The monthly entertainment expenditure of Hong Kong secondary school students follows a normal distribution with mean \$500 and standard deviation \$100.

What is the probability that a randomly selected student will spend more than \$450 and less than \$600 in a month?

Class : ____________ **Name** : __________________ **No.** : ___________

Class Exercise 1 (Solution)

Let X be

X ~ B(,)

P(X ≥)

= 1 - [P(X=) + P(X=) + P(X=)]

= 1 - (+ +)

= 1 -

=

Class Exercise 2 (Solution)

Let X be

λ = X~ P( )

P(X = or X =)
= +

= +

=

Class Exercise 3 (Solution)

Let X be

X ~ N(, ^{2})

P(≤ X ≤)

= P((-)/ ≤ Z ≤ (-)/)

= P(< Z <)

= + (from mason appendix D)

=