Statistics &Data Analysis Concepts for Data Science and ML 5

5

Discrete Probability Distributions

Learning Objectives

The overall objective of this chapter is to help you understand the concepts of probability distributions that are critical to decision making in data analysis. After completing this chapter, you will:

- Understand random variables, probability distribution, and discrete distribution
- Define and calculate the expected value or the mean, variance, and standard deviation of discrete probability distributions
- Learn several discrete probability distributions
- Solve problems involving Binomial distribution using the Binomial formula, Binomial table, and computer packages
- Solve problems involving Poisson distribution using the Poisson distribution formula, table, and computer packages
- Understand and solve problems involving Hypergeometric, Negative Binomial, Geometric, and Discrete Uniform distributions

Random Variables

- A random variable is a variable that takes on different values as a result of the outcomes of a random experiment.

- It is a variable that assumes numerical values governed by chance so that a particular value cannot be predicted in advance.

Random Variables…cont.

- The random variable can be both discrete and continuous.
- If the random variable is either finite or countably infinite, it is a ***discrete random variable***.
- On the other hand, if a random variable takes any value within a given range, it is a ***continuous random variable***.

Discrete Probability Distributions

- A discrete probability distribution is a list of all possible outcomes of a random variable and their probabilities.

METHODS OF DESCRIBING RANDOM VARIABLES AND THEIR PROBABILITIES

There are three methods of describing a discrete random variable:

- (1) List each value of the outcome and the corresponding probability of outcome in a table form,
- (2) Use a histogram that shows the outcome of an experiment on the x-axis and the corresponding probabilities of outcomes on the y-axis, and
- (3) Use a function (or a formula) that assigns a probability to each outcome.

Probability Distribution and Frequency Distribution

- The probability distribution is a model that relates the value of a variable with the probability of occurrence of that value.

The probability distribution describes the frequencies that occur theoretically whereas; the relative frequency distribution describes the frequencies that have actually occurred.

The Theoretical Probability Distribution of Rolling Two dice

The probability Distribution in a tabular form

Table 5.1

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

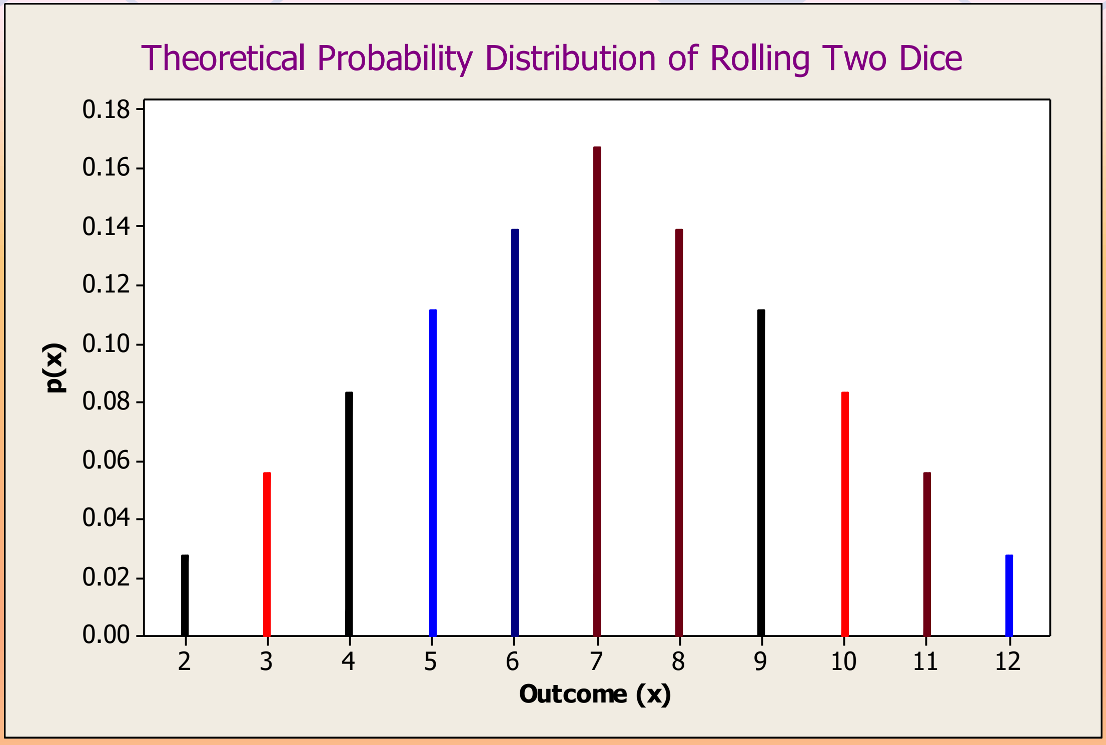
The outcome, x (which is the sum of the numbers on the top faces), is a ***random variable*** as it takes on different values in the population according to some random mechanism

The probability distribution is the outcomes X i , and the probabilities for these outcomes P(Xi) that can be shown both in a tabular or a graphical form.

Requirements of Probability Distributions

1. P(x) is between 0 and 1 (both inclusive) for each x, and
2. $\sum P(x) = 1.0$ (5.1)

The probability distribution of rolling two dice: Graphical Form



This is an example of a discrete probability distribution. In a discrete distribution, the outcomes (X) are integers or whole numbers.

Rolling two dice is a random experiment. The outcome in each throw of two dice is a random phenomenon and assumes values 2 through 12 randomly governed by chance.

Classification of Data: Review

- **Discrete data**

are the result of a counting process and are expressed as whole numbers or integers. Examples: cars sold by Toyota in the last quarter, the number of houses sold last year, or the number of defective parts produced by a company.

Discrete data → Discrete Probability distributions

- **Continuous data**

can take any value within a given range - are measured on a continuum or a scale that can be divided infinitely. Examples: measurements of length, height, diameter, temperature, stock value, sales, etc.

Continuous data → Continuous Probability Distributions

Difference between a Probability Distribution and a Frequency Distribution

Using a statistical computer software (MINITAB) and the theoretical probability distribution (Table 5.1, slide 7), we simulated 1000, 5000 and 10,000 trials of throwing two dice. The Result of 10,000 trials is shown.

Result of simulating 10000 Throws of Two Dice

```
Histogram of trials   N = 10000
Each * represents 35 observation(s)
Outcome         Count
    2.00          251  ********
    3.00          569  *****************
    4.00          829  ************************
    5.00         1079  *******************************
    6.00         1416  *****************************************
    7.00         1652  ************************************************
    8.00         1423  *****************************************
    9.00         1132  *********************************
   10.00          807  ************************
   11.00          580  *****************
   12.00          262  ********
=================================================================
```

Figure 5.4: Result of Simulating 10000 Throws of Two Dice

Comparing the theoretical probability distribution of rolling two dice to Simulated Trials

Table 5.2: Comparing Probability and Frequency Distribution

Row	x Outcome	P(x=x) Prob. Dist.	Rel.Freq1 **1000 Trials**	Rel.Freq2 **5000 Trials**	Rel.Freq3 **10,000 Trials**
	Probability Distribution		**Frequency Distribution**		
1	2	0.0278	0.033	0.0300	0.0251
2	3	0.0560	0.059	0.0582	0.0569
3	4	0.0833	0.088	0.0858	0.0829
4	5	0.1111	0.115	0.1098	0.1079
5	6	0.1389	0.140	0.1402	0.1416
6	7	0.1670	0.147	0.1710	0.1652
7	8	0.1389	0.141	0.1370	0.1423
8	9	0.1111	0.102	0.1038	0.1132
9	10	0.0833	0.089	0.0774	0.0807
10	11	0.0560	0.052	0.0590	0.0580
11	12	0.0278	0.034	0.0278	0.0262=

The results of Table 5.2 indicate that as the number of trials was increased, the relative frequency distributions came closer to the theoretical probability distribution. Also, note that the probability distribution describes the frequencies that occur theoretically, whereas the relative frequency distribution describes the frequencies that have actually occurred.

Expected Value, Variance, And Standard Deviation of A Discrete Distribution

The mean for a discrete random variable is defined mathematically as the expected value and is written as:

$$\mu_x = E(X) = \sum x_i P(x_i)$$

The variance of a discrete random variable is defined as:

$$\sigma^2 = \sum (x_i - \mu)^2 P(x_i)$$

or,

$$\sigma^2 = \sum x^2 P(x) - \mu^2$$

The standard deviation is

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

Example 1

Suppose we want to determine the **mean** and **variance** of the discrete probability distribution given below:

x	0	1	2	3	4	5
P(x)	0.10	0.30	0.25	0.20	0.10	0.05

Calculation of the Mean or Expected Value

x	P(x)	x*P(x)
0	0.10	0.00
1	0.30	0.30
2	0.25	0.50
3	0.20	0.60
4	0.10	0.40
5	0.05	0.25
Total	1.0	2.050

$$\mu_x = E(x) = \sum x_i P(x_i) = 2.050$$

Calculation of the Variance and Standard Deviation

Note that you need to calculate the mean or the expected value of the distribution before you calculate the variance. The variance is calculated using the following formula:

$$\sigma^2 = \sum (x_i - \mu)^2 P(x_i)$$

For this example,

$\sigma^2 = (0 - 2.050)^2 (0.10) + (1 - 2.050)^2 (0.30) + \ldots\ldots\ldots + (5 - 2.050)^2 (0.05) = 1.7475$

and the standard deviation,

$$\sigma = \sqrt{1.7475} = 1.32$$

Example 2

Table 5.3 (next slide) shows the number of cars sold over the past 500 days for a particular car dealership in a certain city.

Example 2 …cont.

Table 5.3

(1) No. of Cars Sold, x	(2) Frequency (f_i)	(3) Relative Frequency, P(x)
0	40	40/500=0.08
1	100	0.200
2	142	0.284
3	66	0.132
4	36	0.072
5	30	0.060
6	26	0.052
7	20	0.040
8	16	0.032
9	14	0.028
10	8	0.016
11	2	0.004
Totals	500	1.00

[a] Calculate the relative frequency.

The relative frequencies are shown in column (3) of Table 5.3. Note that the relative frequency distribution is calculated by dividing the frequency of the class by the total frequency which is also the probability, P( ).

[b] Calculate the expected value or the mean number of cars sold.

The expected value is given by

$$\mu_x = E(x) = \sum x_i P(x_i)$$

$$E(x) = (0)(0.08) + (1)(0.200) + (2)(0.284) + (3)(0.132) + (4)(0.072) +$$
$$(5)(0.060) + (6)(0.052) + (7)(0.040) + (8)(0.032) + (9)(0.028) + (10)(0.016)$$
$$+(11)(0.004) = 3.056$$
$$or, E(x) = 3.056$$

[c] Calculate the variance and the standard deviation.

$$\sigma^2 = \sum (x - \mu)^2 P(x)$$

$$\sigma^2 = (0 - 3.056)^2 (0.08) + (1 - 3.056)^2 (0.200) + (2 - 3.056)^2 (0.284)$$
$$+ (3 - 3.056)^2 (0.132) + (4 - 3.056)^2 (0.072) + (5 - 3.056)^2 (0.060)$$
$$+ (6 - 3.056)^2 (0.052) + (7 - 3.056)^2 (0.040) + (8 - 3.056)^2 (0.032)$$
$$+ (9 - 3.056)^2 (0.028) + (10 - 3.056)^2 (0.016) + (11 - 3.056)^2 (0.004)$$
$$= 6.071296$$

The variance can be more easily calculated using the equation below.

$$\sigma^2 = \sum x^2 P(x) - \mu^2$$

The standard deviation for this discrete distribution is

$$\sigma = \sqrt{\sigma^2} = \sqrt{6.071296} = 2.46$$

[d] Find the probability of selling less than four cars.

$$P(x < 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$
$$= 0.08 + 0.200 + 0.284 + 0.132$$
$$= 0.696$$

[e] Find the probability of selling at most four cars.

$$P(x \le 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$
$$= 0.08 + 0.200 + 0.284 + 0.132 + 0.072$$
$$= 0.768$$

[f] What is the probability of selling at least four cars?

$$P(x \geq 4) = P(x=4) + P(x=5) + P(x=6) + P(x=7)$$
$$+P(x=8) + P(x=9) + P(x=10) + P(x=11)$$
$$= 0.072 + 0.060 + 0.052 + 0.040 + 0.032 + 0.028$$
$$+0.016 + 0.004$$
$$= 0.304$$

The above probability can also be calculated as

$$P(x \geq 4) = 1 - P(X < 4)$$
$$= 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)]$$
$$= 1 - [0.08 + 0.200 + 0.284 + 0.132]$$
$$= 0.304$$

[h] What is the probability of selling more than four cars?

$$P(x > 4) = P(x=5) + P(x=6) + P(x=7)$$
$$+P(x=8) + P(x=9) + P(x=10) + P(x=11)$$
$$= 0.060 + 0.052 + 0.040 + 0.032 + 0.028$$
$$+0.016 + 0.004$$
$$= 0.232$$

or,

$$P(x > 4) = 1 - P(X \leq 4)$$
$$= 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$$
$$= 1 - [0.08 + 0.200 + 0.284 + 0.132 + 0.072]$$
$$= 0.232$$

Some Important Discrete Distributions: Bernoulli Process and Binomial Distribution

The Binomial distribution is a very widely used discrete distribution which describes discrete data resulting from an experiment known as a ***Bernoulli*** process.

Bernoulli Trials and Bernoulli distribution:

In many situations, the experiment or the process under study consists of n number of trials. Each trial has only two possible outcomes: success (S) and failure (F). We can denote this as

$x_j = 1$ if the experiment results in a success (S)
$x_j = 0$ if the experiments results in a failure (F)

The Bernoulli distribution is defined as: A random variable that takes only two values 1 and 0 with probabilities p and q respectively; or,

$$P(X=1) = p \text{ and } \quad P(X=0) = q$$

where p and q are Bernoulli variates that follow a Bernoulli distribution. In the above expression, p can also be referred to as the probability of success and q the probability of failure, such that p + q =1 or, p = (1- q).

Example 3

(1) Outcomes of n number of tosses of a fair coin is a Bernoulli process because:

- **each toss has only two possible outcomes, heads (H) and Tails (T), which may be denoted as a success or failure.**
- **probability of outcome remains constant over time, i.e., for a fair coin the probability of success (or probability of getting a head) remains ½ for each toss regardless of the number of tosses.**
- **the outcomes are independent of each other, i.e., the outcome of one toss does not affect the outcome of any other toss.**

(1) Consider a manufacturing process in which the parts produced are inspected for defects. Each part in a production run may be classified as defective or non defective. Each part to be inspected can be considered as a single trial that result in a success (if the part is found to be defective) or a failure (if it is non defective). This is also an example of a Bernoulli trial.

Binomial Distribution

A random variable that denotes x number of successes in n Bernoulli trials is said to have a Binomial distribution in which the probability of successes is given by the following expression:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$ where, x = 0,1,.........,n

In the above expression,

$p(x)$ = probability of x number of successes

n = number of trials

p = probability of success

$(1-p)$ = q is the probability of failure

or,

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

where, x = 0,1,.........,n

Example 4 : Calculating Binomial Probabilities

A product is supposed to contain 5% defective items. Suppose a sample of 10 items is selected. What is the probability of finding (a) exactly 2 (b) more than 2 (c) exactly 3 (d) at least 3, (e) at most 3 and (f) less than 3 defective items?

Solution: To calculate the Binomial probabilities we must know *n* (the number of trials) and *p* (the probability of success). For this problem,

$n = 10 \qquad p = 0.05$

(a) The probability of finding two defects, that is; *p* (x=2). This can be calculated using the Binomial formula

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

$$p(x=2) = \frac{10!}{2!(10-2)!}(0.05)^2(0.95)^8$$

$$= (45)(0.05)^2(0.95)^8$$

= 0.0746 or 7.46%

Calculating Binomial Probabilities using Binomial Distribution Table

In order to calculate the Binomial probabilities using the Binomial table, you must know the number of trials, n, and the probability of success, p.

A Binomial probability distribution table for ***n=10*** and ***p=0.05*** would look like the one shown below. The probability calculated on the previous page is marked in this table

		Probability of success (p)									
n	***x***	***0.05***	***0.1***	***0 .2***	***0.3***	***0 .4***	***0 .5***	***0 .6***	***0 .7***	***0.8***	***0 .9***
10	x=0	0.5987	0.3487	0.1074	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000
	x=1	0.3151	0.3874	0.2684	0.1211	0.0403	0.0098	0.0016	0.0001	0.0000	0.0000
	x=2	→ 0.0746	0.1937	0.3020	0.2335	0.1209	0.0439	0.0106	0.0014	0.0001	0.0000
	x=3	0.0105	0.0574	0.2013	0.2668	0.2150	0.1172	0.0425	0.0090	0.0008	0.0000
	x=4	0.0010	0.0112	0.0881	0.2001	0.2508	0.2051	0.1115	0.0368	0.0055	0.0001
	x=5	0.0001	0.0015	0.0264	0.1029	0.2007	0.2461	0.2007	0.1029	0.0264	0.0015
	x=6	0.0000	0.0001	0.0055	0.0368	0.1115	0.2051	0.2508	0.2001	0.0881	0.0112
	x=7	0.0000	0.0000	0.0008	0.0090	0.0425	0.1172	0.2150	0.2668	0.2013	0.0574
	x=8	0.0000	0.0000	0.0001	0.0014	0.0106	0.0439	0.1209	0.2335	0.3020	0.1937
	x=9	0.0000	0.0000	0.0000	0.0001	0.0016	0.0098	0.0403	0.1211	0.2684	0.3874
	x=10	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0060	0.0282	0.1074	0.3487

n	x	$p = 0.05$
10	0	0.5987
	1	0.3151
	2	0.0746
	3	0.0105
	4	0.0010
	5	0.0001
	6	0.0000
	7	0.0000
	8	0.0000
	9	0.0000
	10	0.0000

The Binomial Table on the left shows the probabilities for n = 10 p = 0.05

Refer to Example 4. The Binomial table on the left can be used to calculate several probabilities shown below. All the probability values can be read from the Binomial table.

Note that if the number of trials n is 10 then the values under the x column will be from 0 through 10. This is because there cannot be more than 10 successes for n=10 trials.

(b) What is the probability of more than two defects?

$p(x>2) = p(x=3) + p(x=4) + p(x=5) + p(x=6) + p(x=7) + p(x=8) + p(x=9) + p(x=10)$

$= 0.0105 + 0.0010 + 0.0001 + 0.0000 + 0.0000 + 0.0000 + 0.0000 + 0.0000$

$= 0.0116$ **(refer to the table on the left for probability values)**

or,

$p(x>2) = 1 - p(x \le 2)$

$= 1 - [p(x=0) + p(x=1) + p(x=2)]$

$= 1 - [0.5987 + 0.3151 + 0.0746]$

$= 0.0116$

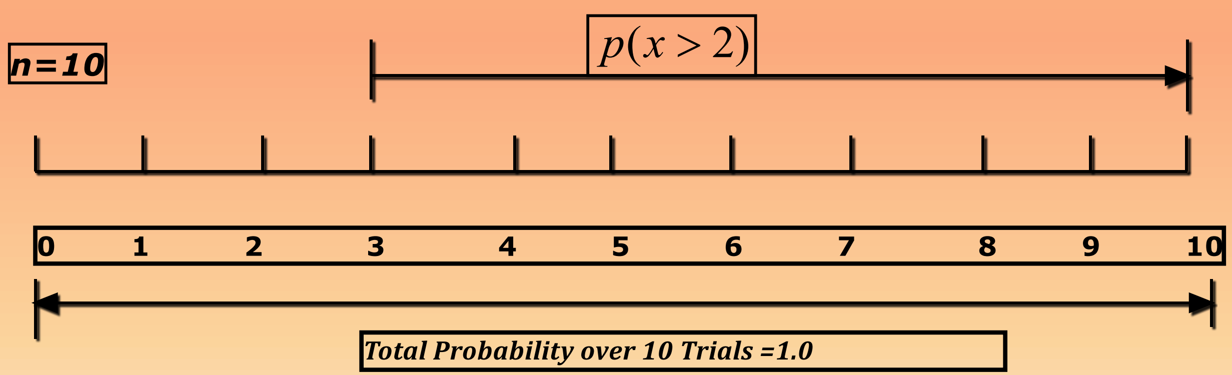
n	x	$p = 0.05$
10	0	0.5987
	1	0.3151
	2	0.0746
	3	0.0105
	4	0.0010
	5	0.0001
	6	0.0000
	7	0.0000
	8	0.0000
	9	0.0000
	10	0.0000

Note: The probability, $p(x > 2) = 1 - p(x \le 2)$ because the sum of the probabilities over 10 trials must add up to 1.0 (verify this by adding the third column of the Binomial table on the left.

the probability, $p(x > 2)$ can be calculated by subtracting the probabilities of $p(x = 0), p(x = 1)$ and $p(x = 2)$ from1.0 or,

$$p(x > 2) = 1 - [p(x = 0) + p(x = 1) + p(x = 2)].$$

Calculating the probability this way reduces the computation significantly. Figure below demonstrates this.



n	x	$p = 0.05$
10	0	0.5987
	1	0.3151
	2	0.0746
	3	0.0105
	4	0.0010
	5	0.0001
	6	0.0000
	7	0.0000
	8	0.0000
	9	0.0000
	10	0.0000

***(c)Find the probability of finding three defects*.**

$p(x=3) = 0.0105$ or, 1.05%

(d) Find the probability of at least three defects

$p(x \geq 3) = p(x=3) + p(x=4) + \ldots\ldots\ldots\ldots + p(x=10)$ or,

$$p(x \geq 3) = 1 - p(x < 3)$$
$$= 1 - [p(=0) + p(=1) + p(=2)]$$
$$= 1 - [0.5987 + 0.3151 + 0.0746]$$
$$= 0.0116$$

(e) ***Find the probability of at most three defects (or no more than three defects)***

$$p(x \leq 3) = p(x=0) + p(x=1) + p(x=2) + p(x=3)$$
$$= 0.5987 + 0.3151 + 0.0746 + 0.0105$$
$$= 0.9989$$

(f)Find the probability of less than three defects

$$p(x<3) = p(x=0) + p(x=1) + p(x=2)$$
$$= 0.5987 + 0.3151 + 0.0746 = 0.9884$$

Example 5

In a binomial situation n = 5 and p = .20. Determine the following probabilities using the binomial formula:

(a) $p(x) = 0$

(b) $p(x) = 1$

Solution:

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

$$(a)\, p(x=0) = \frac{5!}{0!(5-0)!}(0.2)^0(0.8)^{5-0} = (1)(1)(0.8)^5 = 0.3277$$

$$(b)\, p(x=1) = \frac{5!}{1!(5-1)!}(0.2)^1(0.8)^{5-1} = (5)(0.2)\ (0.8)^{5-1} = 0.4096$$

Example 6

Verify the probabilities in Example 5 using the Binomial table

n	x	$p = 0.20$
5	0	0.32768
	1	0.40960
	2	0.20480
	3	0.05120
	4	0.00640
	5	0.00032

Solution:

The Binomial table for n=5 and p=0.2 is shown on the left. For the given n and p the probabilities under x= 0 and x=1 can be read as

p (x= 0) = 0.32768 and

p (x=1) = 0.40960

These probabilities are the same as calculated in Example 5.

n	x	$p = 0.30$
6	0	0.013841
	1	0.071184
	2	0.167790
	3	0.239700
	4	0.231140
	5	0.158496
	6	0.079248
	7	0.029111
	8	0.007798
	9	0.001485
	10	0.000191
	11	0.000015
	12	0.000001

Example 7

A salesperson makes six calls every day and is able to make a sale on 30 percent of the contacts. During the next two days, find:

(a) the probability of making exactly 4 sales.
(b) the probability of making no sales.
(c) the probability of making at least two sales
(d) the probability of making at least one sale.

Solution: The Binomial probabilities for n=12 and p=0.3 are shown on the left. The probabilities can be found using the probabilities in the table

$$(a)\ p(x=4)=0.2311$$

$$(b)\ p(x=0)=0.0138$$

$$(c)\ p(x\geq 2)=1-[p(x<2)]=1-[p(x=0)+p(x=1)]=1-[0.0138+0.0712]=0.915$$

$$(d)\ p(x\geq 1)=1-p(x<1)=1-0.0138=0.9862$$

Calculating Binomial Probabilities Using Excel

The Binomial probabilities can be calculated using the EXCEL function BINOMDIST. The function has the following four arguments:

x (the number of successes), n (the number of trials), p (the probability of success), and cumulative

For the fourth argument (cumulative), FALSE is used if the probability of successes is desired for example, if the probability of an individual value such as, $x=2$ is to be calculated. A TRUE is used for the fourth argument if the cumulative probability of x or fewer successes is required.

Book1

	A	B	C	D	E	F	G	H
1	Number of Trials (n)		10					
2	Probability of Success		0.05					
3								
4								
5			x					
6			0	=BINOMDIST(C6,C1,C2,FALSE)				
7			1					
8			2					
9			3					
10			4					
11			5					
12			6					
13			7					
14			8					
15			9					
16			10					
17								

To calculate the Binomial probabilities when the number or trials (n) is 10 and the probability of success (p) equals 0.05 or, n=10 and p=0.05, set up a the worksheet shown on the left. Type the argument in cell D6, hit the enter key and copy the function to the last value or to x=10.

Mean, Variance, And Standard Deviation Of Binomial Distribution

The mean or expected value of the Binomial distribution is given by

$$E(x) = \mu = np$$

where, n= number of trials, and p = probability of success

Variance of a Binomial Distribution:

$$\sigma^2 = npq \qquad \text{where,} \qquad q = (1 - p)$$

Therefore,

$$\sigma^2 = np(1 - p)$$

Note: p = probability of success, q= probability of failure such that, p+q=1

Standard Deviation of a Binomial distribution:

$$\sigma = \sqrt{np(1 - p)}$$

Example 8

Find the mean and standard deviation of a Binomial random variable with
[a] n=100, p = 0.5

$$\mu = np = (100)(0.5) = 50$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{100(0.5)(0.5)} = 5.0$$

Example 9

If a coin is tossed 1000 times, what is the mean and standard deviation of the number of heads?

$$E(x) = \mu = np = (1000)(0.5) = 500$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{(1000)(0.50)(0.50)} = 15.81$$

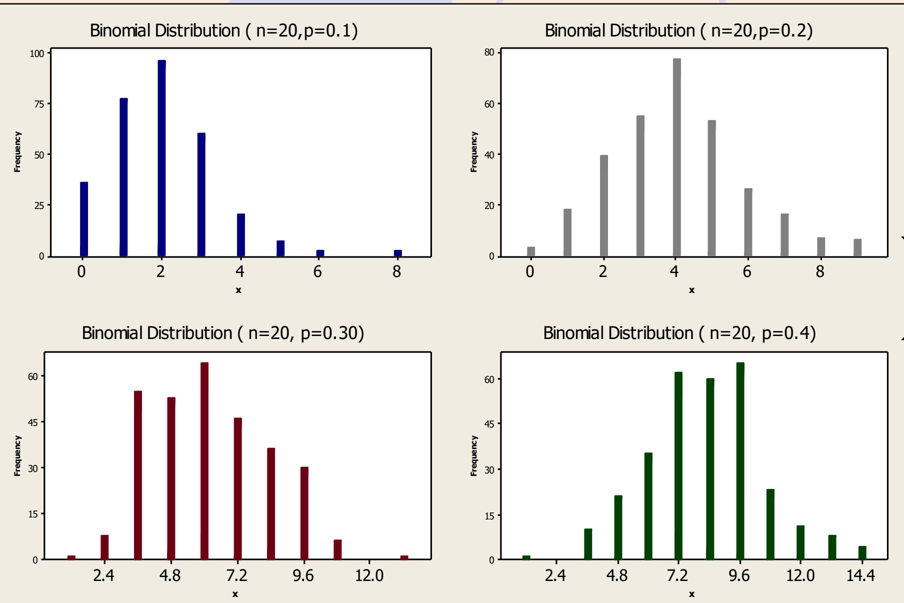
Exploring The Binomial Distribution Using Computer Simulation

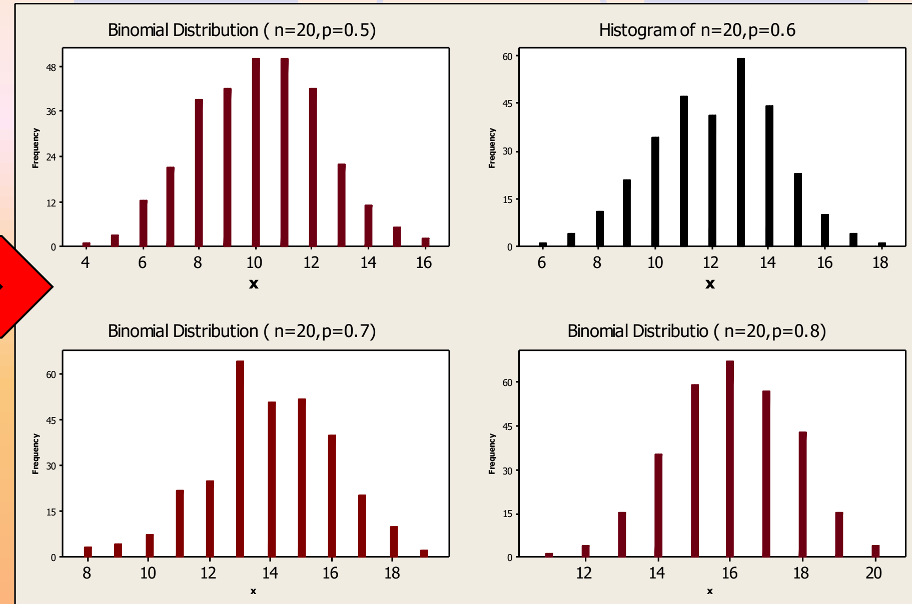
The parameters of the Binomial Distribution are the number of trials (n), and the probability of success (p). In this section we will demonstrate how the shape of the Binomial Distribution changes when we change the characteristic parameters of the distribution.

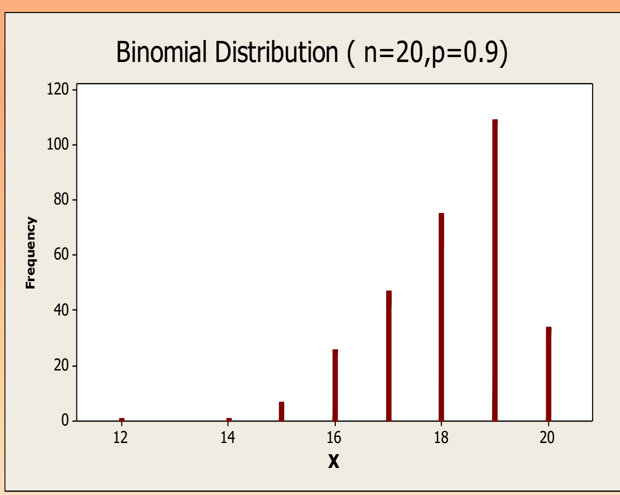
Experiment 1:

In this experiment, we generated 500 random numbers from the Binomial distribution with number of trials, n=20 and various values of probability of success, p (p=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) using MINITAB .

Observe how the shape of the distribution changes (next slide).



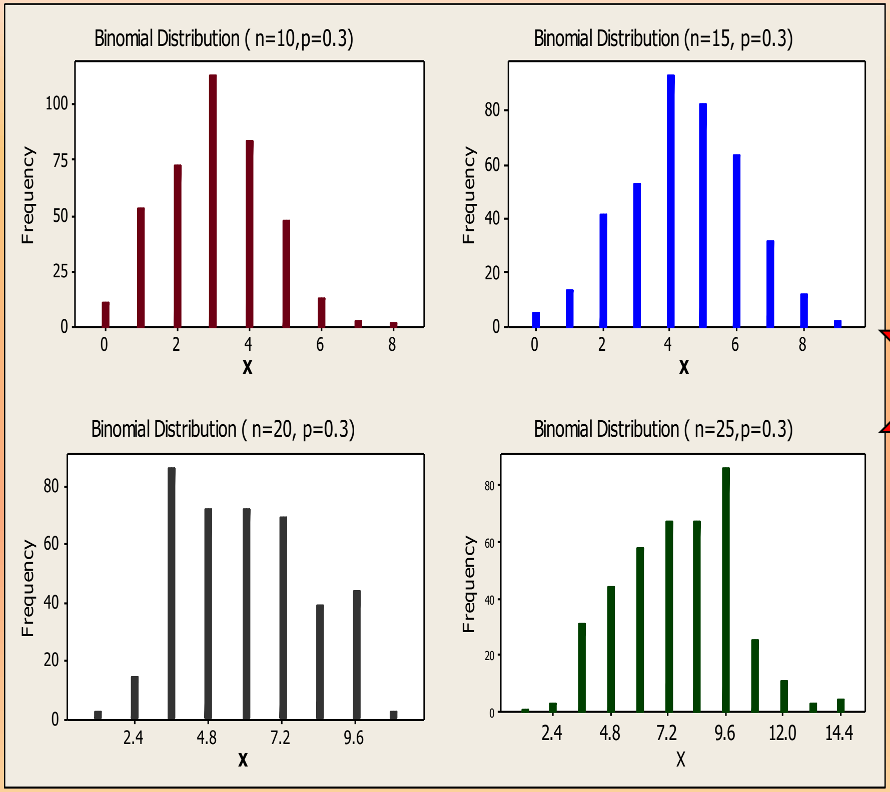


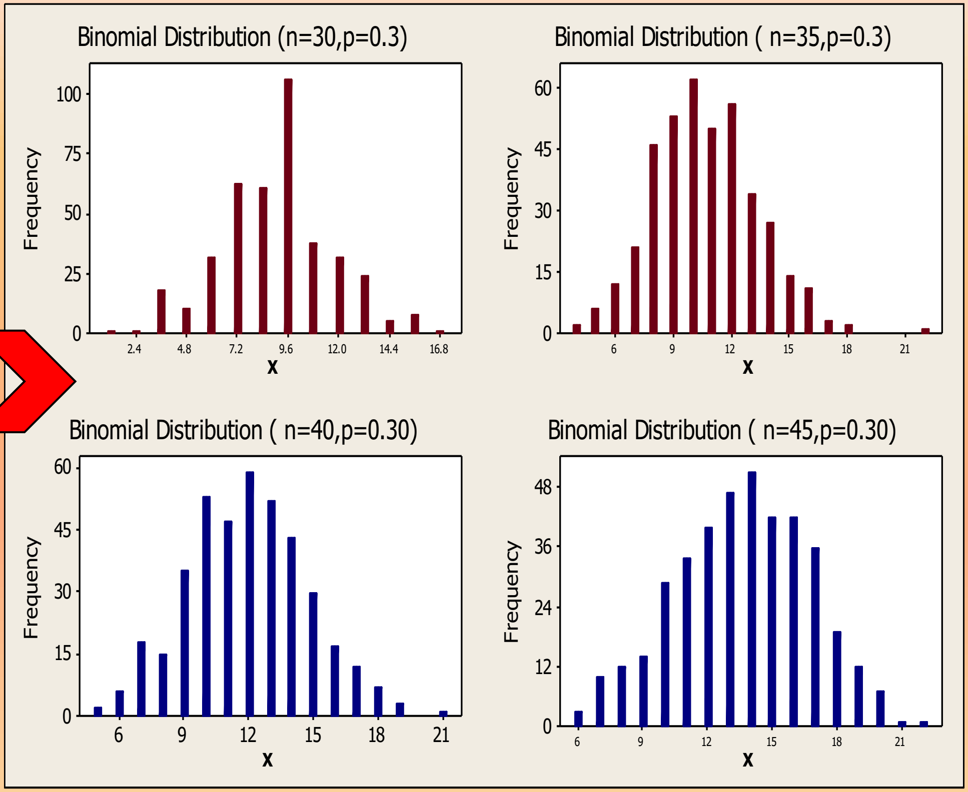


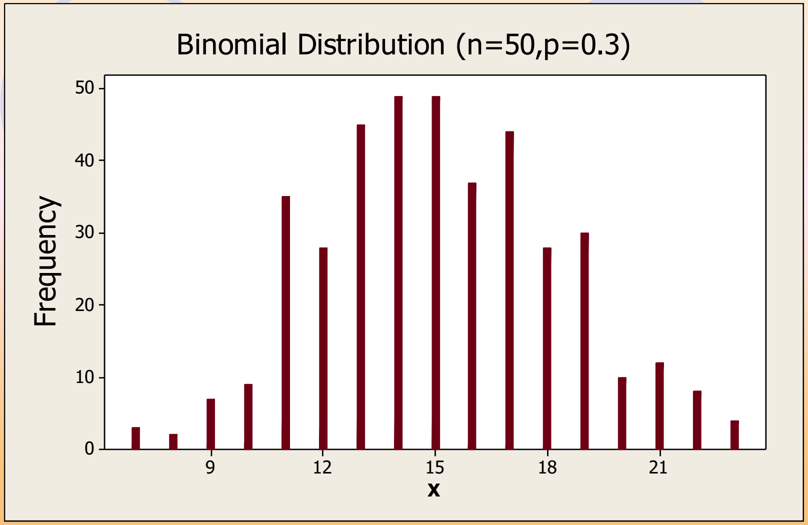
Conclusion: you can see that when

- **the parameter p is less than 0.5, the shape of the distribution is skewed to the right.**
- **p exceeds 0.5, the shape of the distribution is skewed to the left, and**
- **p=0.5, the shape of the distribution is approximately symmetrical.**

Experiment 2: ***Generate 300 random numbers from the Binomial distribution with the probability of success p=0.3, and various values for the number of trials n, (n = 10, 15,20,25,30, 35, 40, 50). Observe how the shape of the distribution changes (see Figure s below).***







***Conclusion:* Comparing the shape of the above distributions for the different values of n, it can be seen that if the probability of success p is held constant and the sample size n is increased, the sum of Bernoulli variables increases.**

***The Binomial distribution becomes more and more symmetrical by virtue of the central limit theorem. Also, as n increases, it is not easy to work with the Binomial distribution (the probability calculations become difficult and time consuming). Thus, for a large n the Binomial distribution may be approximated by a Normal distribution*.**

The Poisson Distribution

: A random variable x is said to follow a Poisson distribution if it assumes only nonnegative values and its probability density function is given by:

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$ where, x = 0,1,2,......,n

Where, μ represents the mean and variance of the distribution where $\mu > 0$.

The Poisson distribution occurs when there are events which do not occur as outcomes for a fixed number of trials of an experiment (unlike that of the Binomial distribution), but which occur at random points of time and space.

The Poisson distribution is the correct distribution to apply when n is very large (that is, the area of opportunity is very large) and an event has a constant and very small probability of occurrence.

The Poisson distribution may be viewed as a limiting case of the Binomial distribution when the following conditions are satisfied:

[a] the number of trials, n is infinitely large, i.e., $n \to \infty$
[b] the probability of success p for each trial is infinitely small, and
[c] np = μ is finite. Thus, p = μ/n and q = 1- μ/n (where μ is a positive real number)

The Poisson distribution is completely described by only one parameter, μ. This distribution is always skewed to the right because in equation

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

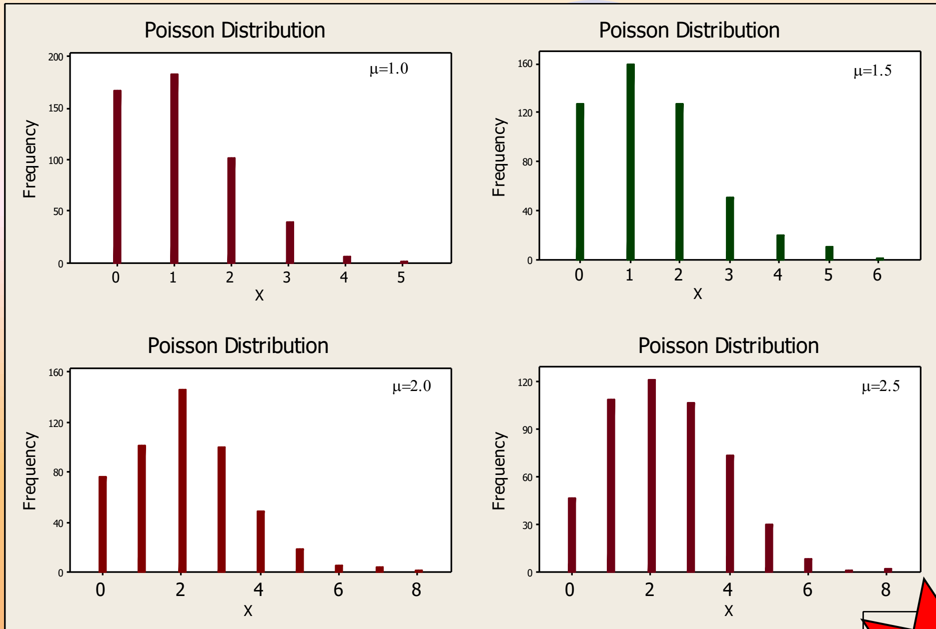
x is never less than zero; and x is always a positive integer. The shape of the Poisson distribution becomes more and more symmetrical if the parameter. $\mu \geq 6$

Investigating The Poisson Distribution Using Computer Simulation

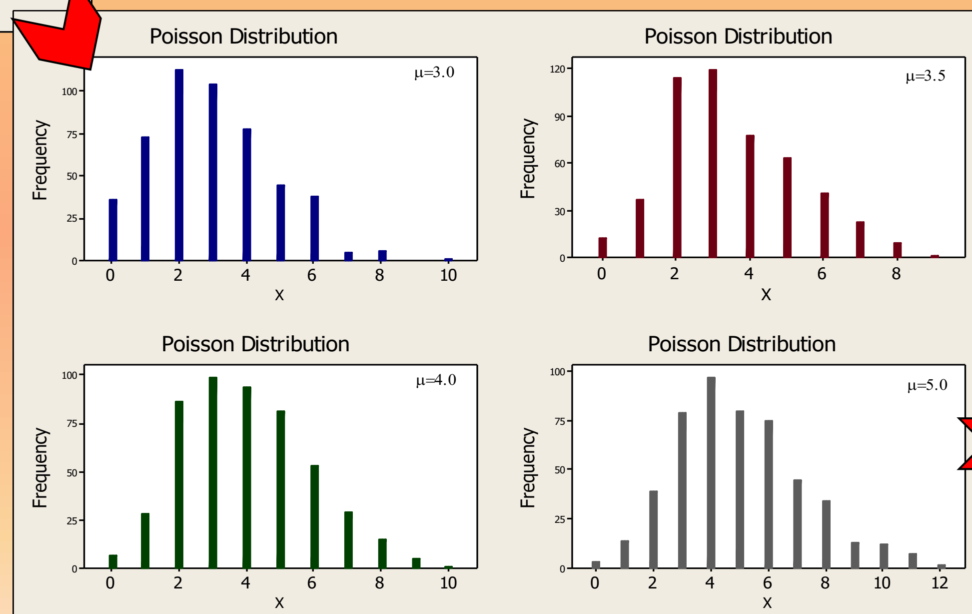
Problem Statement: We will demonstrate that the Poisson distribution, in general, is skewed to the right. However, the distribution becomes more and more symmetrical for $\mu \geq 6$.

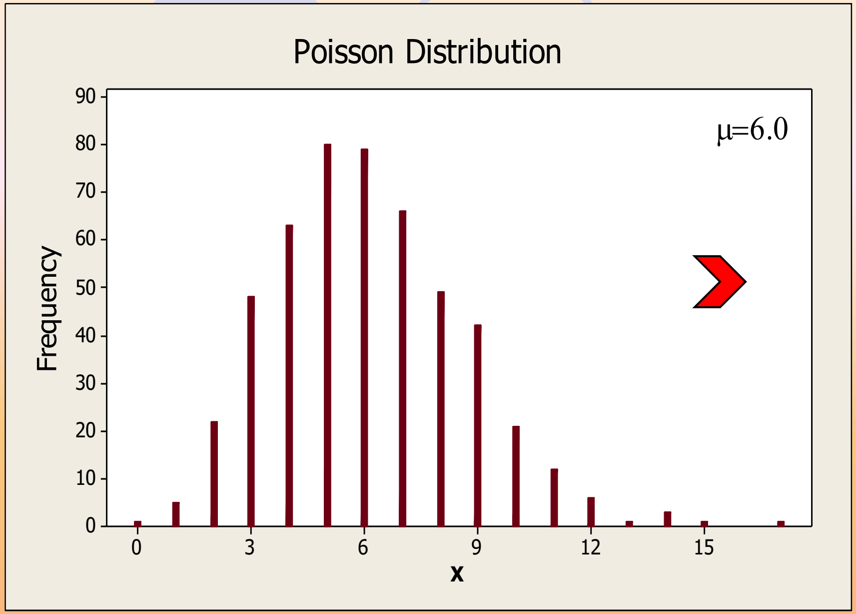
Description: Observe how the shape of the Poisson distribution changes as we change the value of the characteristic parameter of the distribution, μ. To observe the general shape, we generated 500 random numbers using the Poisson distribution with various values of mean, μ.

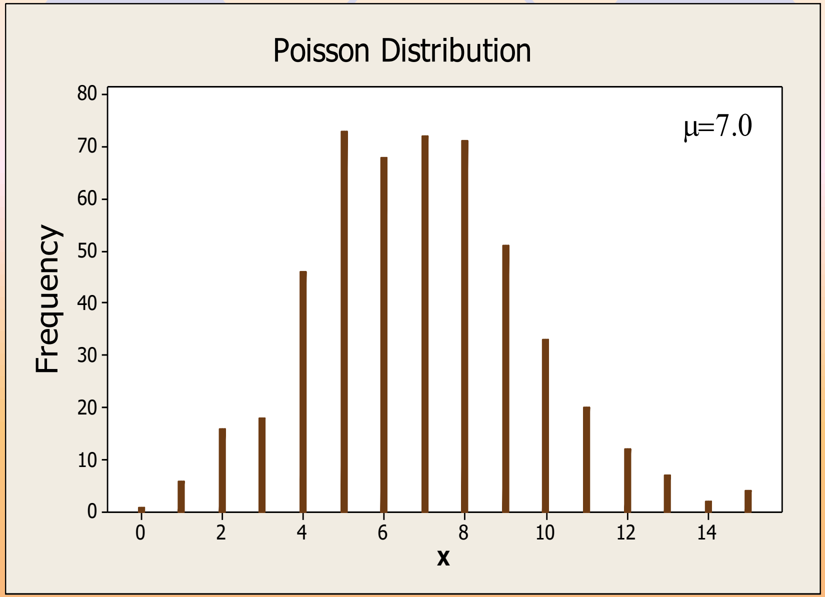
We generate the first set of 500 numbers with $\mu = 1$ and then generate the other sets with μ = 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, and 8. Each time, we construct a histogram of the generated data and observe the shape of the histogram. The results are shown in the next slide.



The shape of the Poisson Distribution for different values of the characteristic parameter, μ







Conclusion: ***From the above Figures we can see that the Poisson distribution becomes more and more symmetrical as the value of the mean μ, becomes 6 and higher.***

For the values of μ below 6, the shape is skewed to the right.

The parameter μ is also known as the shape parameter; a change in this parameter changes the shape of the probability density function of the Poisson distribution. Thus, for the Poisson distribution:

Applications /Characteristics Of Poisson Distribution

Applications : The Poisson distribution is used to describe a number of processes, such as,

- **arrival and distribution of calls coming to a call center.**
- **arrivals of cars at a carwash.**
- **number of accidents at a particular intersection.**
- **number of customers arriving at bank.**
- **number of planes arriving or departing in an airport during peak hours**

Characteristics of Poisson Distribution

Consider the arrival of cars at a particular intersection during a rush hour. In this case, the average (mean) arrivals of cars per rush hour can be estimated from the past data. If the rush hour is divided into periods or intervals, such as one second each, the following statements are true:

- ***the probability that exactly one car will arrive at the intersection in any given second is very small and is constant for every one second interval.***
- ***the probability that two or more cars will arrive within one second of each other is even smaller, the probability of which can be assigned as zero.***
- ***the number of cars that arrive in a given one second interval is independent of the time when that one second interval occurs.***
- ***the number of arrivals in any one second interval is not dependent on the number of arrivals in any other one second interval.***

Poisson Probabilities

The Poisson distribution is used to calculate the probability of x number of occurrences. The probability of x number of occurrence is given by

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!} \qquad \text{where, } x = 0,1,2,......,n$$

p(x) = probability of x occurrences
μ = mean or average number of occurrences
e = base of natural logarithm = 2.718281828...

Difference between the Binomial and Poisson Distributions Distribution

Binomial distribution is the result of a fixed number of trials whereas; the Poisson distribution occurs when there are events which do not occur as outcomes of fixed number of trials.

Poisson distribution does not have a fixed number of trials and the number of occurrences is independent in time and space.

Calculating Poisson Probabilities

Example 10: The numbers of accidents at a particular intersection are up. An investigation on the safety of the intersection indicates that the average number of accidents is 5 per month. Suppose that the number of accidents is distributed according to Poisson distribution.

(a) For any given month, calculate the probability of 0, 1, 2, 3, and 4 accidents.

The average number of accidents per month is μ = 5. The required probabilities can be calculated as

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}$$

$$p(x=0) = \frac{(5)^0 e^{-5}}{0!} = 0.00674$$

$$p(x=1) = \frac{(5)^1 e^{-5}}{1!} = 0.0337$$

$$p(x=2) = \frac{(5)^2 e^{-5}}{2!} = 0.0843$$

$$p(x=3) = \frac{(5)^3 e^{-5}}{3!} = 0.1404$$

$$p(x=4) = \frac{(5)^4 e^{-4}}{4!} = 0.1755$$

[b] What is the probability of 0 or one or two accidents per month?

p (0 or 1 or 2) = p (x=0) + p (x=1) + p (x=2)
= 0.00674+0.0337+0.0843
= 0.1247

[c] What is the probability of four or more accidents (at least four accidents) in a given month?

$$p(x \geq 4) = 1 - p(x < 4)$$
$$= 1 - [p(x=0) + p(x=1) + p(x=2) + p(x=3)]$$
$$= 1 - [0.00674 + 0.0337 + 0.0843 + 0.1404]$$
$$= 0.7349$$

For the probabilities p (x=0), p (x=1), p (x=2) and p(x=3), see part (a)

[d] What is the probability of more than four accidents?

$$p(x > 4) = 1 - p(x \leq 4)$$
$$= 1 - [p(x=0) + p(x=1) + p(x=2) + p(x=3) + p(x=4)]$$
$$= 1 - [0.00674 + 0.0337 + 0.0843 + 0.1404 + 0.1755]$$
$$= 0.6998$$

Calculating Poisson Probabilities using Poisson Table

The Poisson distribution has only one parameter; the mean or the average. The average must be known or determined in order to calculate the Poisson probabilities. Using the process mean or the average, we can calculate the probabilities using either the formula or the table of Poison distribution.

Example 11: (Calculating probabilities using Poisson table)

According to the police records, the average number of two-car accidents in a certain city is 3.1 per day. What is the probability that

[a] there will be fewer than three accidents on any given day?

In this example, the average, $\mu = 3.1$ accidents per day. Using this value, we will use the Poisson table to calculate the probabilities. The next slide shows a partial Poisson table.

Note that the Poisson distribution does not have fixed number of trials. Therefore, the probabilities of x occurrences continue till the probability becomes zero or close to zero [see the Poisson Table on the next page].

A Partial Poisson Probability Table for different values of μ

x	μ = 3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.0984	0.0915	0.0850
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046
4	0.1733	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477
6	0.0555	0.0608	0.0662	0.0716	0.0771	0.0826	0.0881	0.0936
7	0.0246	0.0278	0.0312	0.0348	0.0385	0.0425	0.0466	0.0508
8	0.0095	0.0111	0.0129	0.0148	0.0169	0.0191	0.0215	0.0241
9	0.0033	0.0040	0.0047	0.0056	0.0066	0.0076	0.0089	0.0102
10	0.0010	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039
11	0.0003	0.0004	0.0005	0.0006	0.0007	0.0009	0.0011	0.0013
12	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004
13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

In our example, we are interested in the probabilities for mean, μ =3.1. The Poisson probabilities for μ=3.1 and various values of x are shown from the Table on the previous.

x	μ =3.1
0	0.0450
1	0.1397
2	0.2165
3	0.2237
4	0.1734
5	0.1075
6	0.0555
7	0.0246
8	0.0095
9	0.0033
10	0.0010
11	0.0003
12	0.0001
13	0.0000
14	0.0000

Using the values from the table above, the probability of fewer than three accidents can be calculated as

$$p(x<3)=p(x=0)+p(x=1)+p(x=2)$$
$$=0.0450+0.1397+0.2165$$
$$=0.4012$$ or, 40.12%

To read the probabilities from the Poisson table refer to the column $\mu=3.1$ on the left and then read the probabilities for x=0, 1, 2,..,etc.

[b] What is the probability of exactly three accidents?

p(x=3) = 0.2237

[c] What is the probability of at least three accidents?

$$p(x\geq 3)=1-p(x<3)$$
$$=1-[p(x=0)+p(x=1)+p(x=2)]$$
$$=1-[0.0450+0.1397+0.2165]$$
$$=0.5988$$

x	μ =3.1
0	0.0450
1	0.1397
2	0.2165
3	0.2237
4	0.1734
5	0.1075
6	0.0555
7	0.0246
8	0.0095
9	0.0033
10	0.0010
11	0.0003
12	0.0001
13	0.0000
14	0.0000

[d] What is the probability of more than four accidents?

$$p(x>4)=1-p(x\leq 4)$$
$$=1-[p(x=0)+p(x=1)+p(x=2)+p(x=3)+p(x=4)]$$
$$=1-[0.0450+0.1397+0.2165+0.2237+0.1734]$$
$$=1-0.7983$$
$$=0.2017$$

Refer to the Poisson Table on the left

μ =3.0

x	P(X = x)
0	0.049787
1	0.149361
2	0.224042
3	0.224042
4	0.168031
5	0.100819
6	0.050409
7	0.021604
8	0.008102
9	0.002701
10	0.000810
11	0.000221
12	0.000055
13	0.000013
14	0.000003
15	0.000001
16	0.000000

Example 12:

Phone calls at a call center arrive at the rate of 36 per hour.

(a) Find the probability of receiving 4 calls in a 5 minute interval of time.

The average number of calls every hour is 36; that is, the call center gets 36 calls every 60 minutes. Therefore, the average for 5 minute is (36)(5)/60 = 3 or μ=3 for five minute interval.

$$(a)\, p(x=4)=0.1680$$

See the table on the left for the probability

(b) Find the probability of receiving more than 4 calls in a 5 minute interval of time.

$$(b)\, p(x>4)=1-p(x\le 4)=1-[p(x=0)+p(x=1)+p(x=2)+p(x=3)+p(x=4)$$

= 1-[0.0498+0.1494+0.2240+0.2240+0.1680) = 0.1848

(c) Find the probability of receiving exactly 10 calls in 15 minutes

The average number of calls per hour is 36. Therefore, the average number of calls in 15 minutes is (36)(15)/60= 9, or

Look into the Poisson table for μ =9.0 and x=10. The required probability is

$$p(x=10)=0.1186$$

(f) Suppose no calls are on hold. If the agent takes 5 minutes to complete the current call, how many callers do you expect to be waiting by that time? What is the probability that none will be waiting?

On the average the call center receives 3 calls every 5 minutes (from the data in the problem, the call center gets 36 calls every hour).
If no calls are on hold, and the agent takes 5 minutes to help a customer, there will be 3 calls waiting.
The probability that none will be waiting is

$$p(x=0)=0.0498$$

See the table on the previous page for the probability

Calculating Poisson Probabilities using EXCEL

The Poisson probabilities are calculated using the EXCEL function POISSON. The function has the following four arguments:

x **(the number of occurrences),** μ **(the mean), and cumulative**

For the third argument (cumulative), FALSE is used if the probability of occurrences is desired for example, if the probability of an individual value such as, is to be calculated. A TRUE is used for the third argument if the cumulative probability of or fewer successes is required.

Book1

	A	B	C	D	E	F
1						
2	Mean	3.1				
3						
4		x				
5		0	=POISSON(B5,B2,FALSE)			
6		1				
7		2				
8		3				
9		4				
10		5				
11		6				
12		7				
13		8				
14		9				
15		10				
16		11				
17		12				
18		13				
19						

To calculate the Probabilities probabilities when the average, μ = 3.1, set up a the worksheet shown on the left. Type the argument in cell C5, hit the enter key and copy the function to the last value or to x=13. The probabilities will be as shown on page 48, Poisson table.

Approximating The Binomial Distribution With The Poisson Distribution (Poisson's Approximation)

If the number of trials (n) is very large, and the probability of success (p) is small, it is usually difficult to calculate the Binomial probability.

For example, suppose that a researcher has determined that the probability of getting a certain type of blood disorder in humans is 0.00004 or 0.004%. If a sample of 100,000 people (n=100,000) is selected, determine the probability of at least 5 of them having this particular type of blood disorder.

In this case, it will be difficult to find this probability using the Binomial distribution for such a large sample size, n. In such cases, we can find the probability by approximation; that is, we approximate the Binomial probability using the Poisson distribution.

Using Poisson's Approximation

To calculate the Poisson probability, the only parameter we need is the mean or the average. The mean of the Binomial is calculated by

$$\mu = np$$

that is, the mean can be obtained by multiplying ***n*** and ***p***. In this example, n=100,000 and p = 0.00004. Therefore,

$$\mu = np = (100,000)(0.00004) = 4$$

We can now use this value in the Poisson formula:

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

to calculate the required probability. The probability we get is approximate but the problem is simplified.

When the number of trials *n* in the Binomial distribution is large and the probability of success *p* is small, so that $np \leq 7$ then the Binomial distribution can be approximated by the Poisson distribution. We can also use the Poisson's approximation when the number of trials *n > 20* and the probability of success, $p < 0.05$

Example on Poisson's Approximation

Suppose a random sample of 100 computer chips is taken from a production process that historically produces 2 in 100 defective products. Find the probability of finding using the Binomial distribution and Poisson's approximation.

(a) no defective chip (b) one defective chip (c) two defective chips.
(d) more than two defective chips.

In this case, n=100 and p= 2/100 =0.02. Note that n is large and p is small. The probabilities can be calculated using the Binomial distribution but it will be easier to use the Poisson's approximation to calculate the probabilities.

Solution: ***Probabilities using Binomial distribution***

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

$$given: n = 100; p = 0.02$$

$$(a)\, p(x=0) = \frac{100!}{0!(100-0)!}(0.02)^0(0.98)^{100-0} = (1)(1)(0.98)^{100} = 0.1326$$

Binomial distribution

$$p(x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}\ given: n=100; p=0.02$$

$$(b)\,p(x=1)=\frac{100!}{1!(100-1)!}(0.02)^1(0.98)^{100-1}=100(0.02)^1(0.98)^{99}=0.2707$$

$$(c)\,p(x=2)=\frac{100!}{2!(100-2)!}(.02)^2(0.98)^{100-2}=4950(0.02)^2(0.98)^{98}=0.2734$$

$$(d)\,p(x>2)=1-p(x\le 2)=1-[p(x=0)+p(x=1)+p(x=2)=1-[0.1326+0.2707+0.2734]=0.3233$$

Poisson with mean = 2

x	P(X = x)
0	0.135335
1	0.270671
2	0.270671
3	0.180447
4	0.090224
5	0.036089
6	0.012030
7	0.003437
8	0.000859
9	0.000191
10	0.000038
11	0.000007
12	0.000001
13	0.000000

Probabilities using Poisson's approximation

Since n is large and p is small, the above probabilities can also be calculated using Poisson's distribution. Note that to use the Poisson probability formula, we must have the mean, μ. This is calculated by

$$\mu = np = (100)(0.02) = 2$$

Refer to the Poisson table on the left and read the probabilities for μ=2 and x=0, 1, and 2. These values are shown on the next slide.

$$(a)\, p(x=0) = 0.1353$$
$$(b)\, p(x=1) = 0.2707$$
$$(c)\, p(x=2) = 0.2707$$
$$(d)\, p(x>2) = 1 - p(x \le 2) = 1 - [p(x=0) + p(x=1) + p(x=2)]$$
$$= 1 - [0.1353+0.2707+0.2707] = 0.3233$$

The above probabilities are close to Binomial probabilities obtained earlier.

The probabilities can also be calculated using the Poisson's formula

Compare the Binomial and Poisson's approximation probabilities.

	Binomial Probability n= 100, p=0.02	Poisson's Approximation μ = np = (100)(0.02)= 2.0
p (x=0)	0.1326	0.1353
p(x=1)	0.2707	0.2707
p(x=2)	0.2734	0.2707
p(x>2)	0.3233	0.3233

Note how close the probabilities are.

The Hypergeometric Distribution

The hypergeometric probability distribution is the other discrete distribution closely related to the binomial distribution.

Comparing the Binomial and Hypergeometric Distributions

The binomial distribution is the result of a fixed number of trials. It is applicable in cases where

- ***the trials are independent,***
- ***only two outcomes are possible on each trial: "success" or "failure," and***
- ***the probability of success remains constant for each trial. For example, in the toss of a single coin, the probability of getting a head is 0.5, and it remains the same for any number of tosses.***

Hypergeometric Distributions

In hypergeometric distribution,

- the trials are not independent and the probability of success changes from trial to trial
- For example, suppose a population consists of 10 items. The probability of selecting a particular item from the population is 1/10. If the item is not returned before selecting the second item, that is, if the sampling is done without replacement, then after the first draw there are only 9 items remaining, and the probability of selecting the second item is only 1/9. Likewise, the probability of selecting an item in the third draw is 1/8. This is based on the assumption that the population is finite.

If the probability of success is not constant from trial to trial and the sampling is done from a population without replacement, the appropriate distribution is the hypergeometric distribution.

Hypergeometric Distributions

The conditions for the hypergeometric distribution are:
(a) the population size is finite,
(b) the sampling is done from the finite population without replacement, and
(c) the sample size n is greater than 5% of the population size, N.

- Suppose there is a finite population of size N and some number D ($D \leq N$) is of interest.
- The number D can be the number of nonconforming items in a population that contains nonconforming and conforming items in a production lot. This can also be the number of persons belonging to certain race from a population of finite size.
- We select a sample of size n without replacement, then the random variable of interest x is the number of items in the sample that falls into a class of interest.
- Thus, the hypergeometric distribution calculates the probability of x , the specified number of successes.

The hypergeometric probability function, p(x) is used to determine the probability of x successes in a sample size of n selected without replacement and is given by:

$$p(x) = \frac{(C_x^D)(C_{n-x}^{N-D})}{C_n^N} = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

Where:

$p(x)$ = probability of x number of successes in a sample of size n

D = the number of successes in the population or the number of interest

N = population size

x = the number of successes of interest, x = 0, 1, 2, ...

n = sample size

C_n^N = the combinations of all units

C_x^D = the combinations of x successes from D successes

C_{n-x}^{N-D} = the combinations of (n–x) failures from (N–D) failures

Example 1

A shipment of 20 items has three nonconforming units. What is the probability of drawing one nonconforming unit in a random sample of five? ***N=20*** ***D=3*** ***n=5*** ***=1***

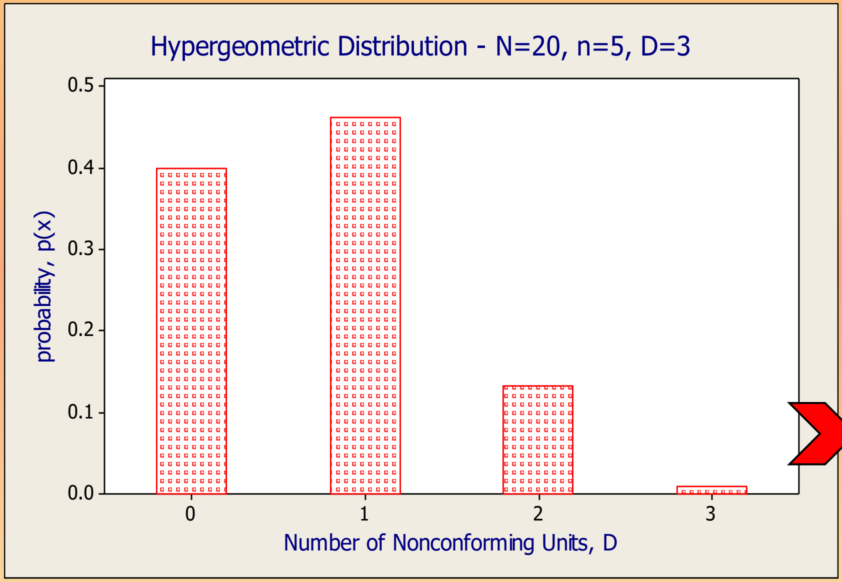
$$p(x) = \frac{(C_x^D)(C_{n-x}^{N-D})}{C_n^N}$$

or,

$$p(x=1) = \frac{(C_1^3)(C_{5-1}^{20-3})}{C_5^{20}} = 0.4605$$

The probabilities of x = 0 (no nonconforming), x =1, x=2, and =3 are shown below. These probabilities are also plotted. Note that there are only 3 nonconforming units. Therefore, p(x=4) is not possible. Also, the sum of the probabilities in the table is equal to 1.0.

x	P(x)
0	0.399123
1	0.460526
2	0.131579
3	0.008772



The shape of the hypergeometric distribution changes as the parameters of the distribution change. ***The parameters of the distribution are the population size N, the sample size n, and the number of successes D.*** The next example shows how the shape of the hypergeometric distribution changes as these parameters change.

Example 2

A machine shop produces certain type of shaft that is inspected in lots. From a lot of 50 shafts, a sample of 5 is selected without replacement. The lot is accepted if the sample has no more than one defective shaft. Otherwise, it is rejected. Suppose that a lot containing 4% defective is selected for inspection. What is the probability that the lot will be accepted?

N=50, n=5, D=100(0.04)=4, the probability of accepting the lot

$$p(x \le 1) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} = \frac{\sum_{0}^{1}\binom{4}{x}\binom{50-x}{5-x}}{\binom{50}{5}} = 0.9550$$

Other Discrete Probability Distribution

Understand and solve problems involving

Negative Binomial Distribution

Geometric Distribution, and

Discrete Uniform distributions

For a detailed discussion on the above distributions, see Chapter 5.

Random Variables

A random variable is a variable that takes on different values as a result of the outcomes of a random experiment. It can also be a variable that assumes numerical values governed by chance so that a particular value cannot be predicted in advance. There are two types of random variables.

- ***Discrete (Countable)***
  - ***Random variable X is either finite or countably infinite***
- ***Continuous(Uncountable)***
  - ***The random variable takes any value within a given range***

Probability Distribution and Frequency Distribution

The probability distribution is a model that relates the value of a variable with the probability of occurrence of that value. The probability distribution describes the frequencies that occur theoretically; whereas, the relative frequency distribution describes the frequencies that have actually occurred.

Expected Value, Variance, and Standard Deviation of a Discrete Distribution

Expected Value $\mu_x = E(X) = \sum X_i P(X_i)$

Variance $\sigma^2 = \Sigma (x_i - \mu)^2 P(x_i)$

Standard Deviation $\sigma = \sqrt{\sigma^2}$

Chapter 5: Discrete Probability Distributions - Flow Chart (1)

Discrete Probability Distributions

Binomial Distribution

Binomial Distribution

The experiment or the process under study consists of n number of trials. Each trial has only two possible outcomes; success (S) and failure (F). The following properties hold:

- **there are n number of trials**
- **each trial has only two possible outcomes; success (S) and failure (F)**
- **the probability of success and the probability of failure remains constant across trials**
- **the outcomes are independent of each other**

A random variable X that denotes x number of successes in n Bernoulli trials is said to have a Binomial distribution.

The Binomial distribution calculates the probability of x successes in n trials using the following expression:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{where, } x = 0,1,\ldots\ldots\ldots,n$$

p(x) = probability of x number of successes, n = number of trials, p = probability of success, (1-p) = q is the probability of failure

Poisson Distribution

The Poisson Distribution

A random variable X is said to follow a Poisson distribution if it assumes only nonnegative values and its probability density function is given by:

$$p(x) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{where, } x = 0,1,2,\ldots\ldots,n$$

where μ represents the mean and variance of the distribution.

Note that $\mu > 0$

The Poisson distribution occurs when there are events which do not occur as outcomes for a fixed number of trials of an experiment (unlike that of the Binomial distribution), but which occur at random points of time and space. The Poisson distribution is the correct distribution to apply when n is very large (that is, the area of opportunity is very large) and an event has a constant and very small probability of occurrence. The Poisson distribution calculates the probability of X number of occurrences.

Mean, Variance, and Standard Deviation of Binomial Distribution

The mean or expected value of the Binomial distribution is given by

$$E(x) = \mu = np$$

where, n= number of trials, and p = probability of success

Variance of a Binomial Distribution:

$$\sigma^2 = np(1-p)$$

Standard Deviation of a Binomial distribution:

$$\sigma = \sqrt{np(1-p)}$$

Chapter 5: Discrete Probability Distributions - Flow Chart (2)

Discrete Probability Distributions...continued

Hypergeometric Distribution

Hypergeometric Distribution

- In hypergeometric distribution, the trials are not independent and the probability of success changes from trial to trial.
- If the probability of success is not constant from trial to trial and the sampling is done from a population without replacement, the appropriate distribution is the hypergeometric distribution.
- The hypergeometric distribution calculates the probability of x, the specified number of successes.
- The conditions for the hypergeometric distribution are: (a) the population size is finite, (b) the sampling is done from the finite population without replacement, and (c) the sample size n is greater than 5% of the population size, N.

The hypergeometric probability function, p(x) that is used to determine the probability of x successes in a sample size of n selected without replacement and is given by:

$$p(x) = \frac{(C_x^D)(C_{n-x}^{N-D})}{C_n^N} = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

p(x) = probability of x number of successes in a sample of size n

D = the number of successes in the population or the number of interest

N = population size, x = the number of successes of interest, x = 0, 1, 2, ...

n = sample size

C_n^N = the combinations of all units

C_x^D = the combinations of x successes from D successes

C_{n-x}^{N-D} = the combinations of (n–x) failures from (N–D) failures

Negative Binomial or Pascal Distribution

Negative Binomial or Pascal Distribution

The basic difference between the binomial and the negative binomial distribution is that the binomial distribution is used to calculate the probability of x number of successes out of n trials where the number of trials is fixed, whereas, in a negative binomial distribution the trials are repeated until a fixed number of successes occur. We are interested in finding the probability that the r^{th} success occurs on the x^{th} trial.

Suppose we have a succession of n Bernoulli trials. Assume that the (i) trials are independent, (ii) the probability of success p in the trial remains constant from trial to trial, and (iii) the probability of failure is q = 1-p. Then the probability distribution of the random variable X, the number of trial on which the r^{th} success occurs, is given by

$$b^-(x, r, p) = \binom{x-1}{r-1} p^r q^{x-r};\ x = r, r+1, r+2, \ldots\ldots$$

The probability of r^{th} success occurring on the x^{th} trial is also written as

$$p(x) = P(X = x) = \binom{x-1}{r-1} p^r q^{x-r};\ x = r, r+1, r+2, \ldots\ldots$$

Note: For the Poisson distribution; the mean and the variance are equal. The equality of mean and the variance is an important characteristic of the Poisson distribution. For the binomial distribution, the mean is always greater than the variance. In some cases, the observable phenomenon gives rise to empirical distributions in which the variance is larger than the mean. In cases where the variance is larger than the mean, the negative binomial distribution provides a good model.

Chapter 5: Discrete Probability Distributions - Flow Chart (3)

Discrete Probability Distributions...continued

Geometric Distribution

The Geometric Distribution

- *The geometric distribution is related to a sequence of Bernoulli trials in which the random variable X takes two values 0 and 1 with the probability q and p respectively, that is, p(X=1) = p, p(X=0)=q, and q = 1-p.*
- *In the geometric distribution, the number of trials is not fixed, and the random variable of interest X, is defined as the number of trials required to achieve the first success.*
- *The geometric distribution can be derived as a special case of negative binomial distribution. In the probability density function of negative binomial distribution above, if r = 1, we get the probability distribution for the number of trials required to achieve the first success.*
- *If we have a series of independent trials that can result in a success with probability p and a failure with probability q where, q=1-p, then the random variable X that denotes the number of trials on which the first success occurs, is given by*

$$p(x;p) = pq^{x-1} \quad \text{where, } x = 1,2,\ldots$$
$$= 0 \quad \text{otherwise.}$$

The distribution is called geometric because the probabilities for x=0, 1, 2,...., are the various terms of geometric progression.

All the assumptions of Binomial distribution apply to negative binomial distribution. Unlike the binomial distribution, the geometric distribution does not have a fixed sample size because the sampling continues until the first success is observed. Also, the variable x cannot be 0 because at least one trial is needed for the first success to occur.

Discrete Uniform Distribution

Discrete Uniform Distribution

In a discrete uniform distribution, the probability for a discrete variable is equal for all values. For example, in a toss of a six-sided fair dice, the probability of occurrence of each of the possible outcome 1 through 6 is 1/6. The probability density of a discrete uniform distribution is given by

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

for x =a, a+1, ...,b-1, b (and zero elsewhere)

The mean and variance of the discrete uniform random variable is given by

$$\mu = \frac{a+b}{2}$$

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

Multinomial Distribution

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. If a trial results in more than two mutually exclusive outcomes, then this leads to a multinomial distribution. Suppose $E_1, E_2, E_3, \ldots\ldots, E_k$ are k mutually exclusive outcomes of a trial with probabilities $p_1, p_2, p_3, \ldots\ldots, p_k$ then the probability that E_1 occurs x_1 times, E_2 occurs x_2 times..., and E_k occurs x_k times in n independent observations can be given by

$$p(x_1, x_2, x_3, \ldots) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}; 0 \le x \le n$$

Note that

$$x_1 + x_2 + \ldots + x_k = n \quad \text{and} \quad p_1 + p_2 + \ldots + p_k = 1$$

Chapter 5: Discrete Probability Distributions - Flow Chart (4)