Overview : Use of Probability

In any statistical analysis we use probability in two ways :

1. Probability models describe nondeterministic nature of measurements.
2. Probability is used to quantify the uncertainity in the results (conclusions) of our statistical analysis.

Random Variable:

A random variable is the concept of making some measurement.

- The value of the measurement need not be the same every time, so we speak of the probability distribution of the random variable.

- We know everything there is to know about the random variable once we know the probability that the value of the random variables might be in any specified set.

Examples:

Sex

Age

Height

Blood Pressure

Pulse beat rate

etc

Probability Density Function (p.d.f):

Everything that can be known about a random variable is contained in its probability density function.

Probability density function is a rule that allows us to determine the probability that a particular measurement of a random variable might be within some set of values.

Overview : Use of Probability

In any statistical analysis we use probability in two ways :

1. Probability models describe nondeterministic nature of measurements.
2. Probability is used to quantify the uncertainty in the results (conclusions) of our statistical analysis.

Random Variable:

A random variable is a function on sample space. It is basically a device for transferring probabilities from complicated sample spaces to simple sample spaces.

A random variable X is a function whose domain is the sample space and whose range is the set of real numbers.

Thus a random variable assigns a real value (i.e., a number) to every outcome in the sample space. The particular values are called realisations and are denoted as x.

If the realisations are countable, $x_1, x_2, \ldots$, the random variable is said to be discrete.

In contrast, if there are infinite-many uncountable realisation, the random variable is said to be continuous.

Thus, random variables can be classified according to the types of values they can take on (the range of the random variables).

[Binary, Categorical (ordered or unordered)
Qualitative, Quantitative.]

Random Variable:

A random variable is a function on the sample space. It is basically a device for transferring probabilities from complicated sample spaces to simple sample spaces where the elements are just natural numbers.

Suppose the arrival of telephone calls at an exchange. Modelling this is very complicated as the sample space should include all possible times of arrival of calls and all possible number of calls. If we consider the random variable which counts how many calls arrive before time t (for example) then the sample space becomes $S = \{0, 1, 2, \ldots\}$.

Thus a random variable X is a function whose domain is the sample space and whose range is the set of real numbers. The random variable assigns a real value (i.e a number) to every outcome in the sample space. The particular values are called realizations and are denoted by x.

If the realizations are countable, $x_1, x_2, \ldots$, the random variable is said to be discrete.

In contrast, if there are infinitely-many uncountable realizations, the random variable is said to be continious.

Probability Mass Function:

If X is a discrete random variable, the function given by $f(x) = P(X = x)$ for every x within the range of X, is called Probability mass function or probability function or probability distribution of X.

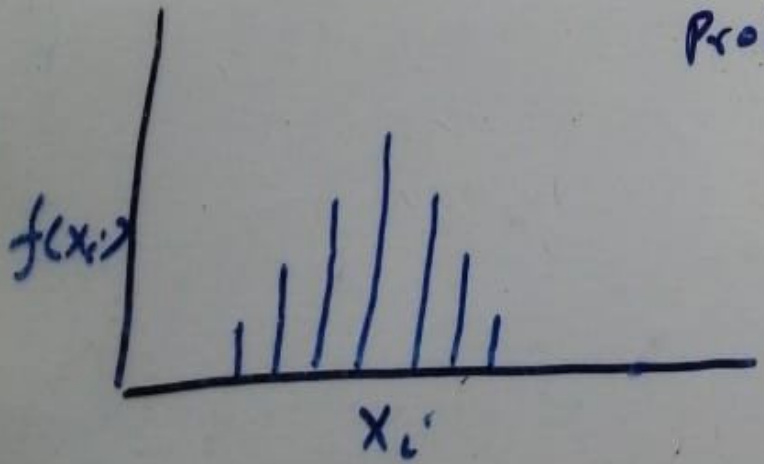
A function can serve as the probability distribution of a discrete random variable X iff its values, $f(x)$, satisfy the conditions.

1. $f(x) \geq 0$ for all x within the range of X.
2. $\sum_x f(x) = 1$, where the summation extends over the values within its domain.

Thus a table, formula or graph representing the probabilities that a random variable X takes, is called a probability distribution.

x_i	x_1	x_2	x_k	Total
$f(x_i)$	$f(x_1)$	$f(x_2)$ ...	$f(x_k)$	1

Probability Histogram.



Probability Distribution: (Continuous r.v.)

A random variable 'X' is said to be continuous if there exist a non-negative function f, defined for **all** real $x \in (-\infty, \infty)$, having the property that for **any** set of real numbers

$$P\{X \in A\} = \int_A f(x)\,dx$$

The function f is called the probability density function of the random variable X. Since X must assign some value, f must satisfy

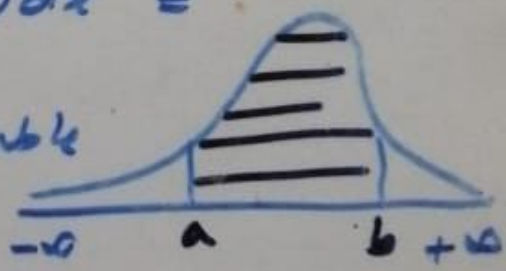
1. $P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx = 1$

If $A = \{a, b\}$, then

$$P\{X \in (a, b)\} = \int_a^b f(x)\,dx = 1$$

2. $P\{X = a\} = \int_a^a f(x)\,dx = 0$

3. $P\{a \le X \le b\} = \int_a^b f(x)\,dx =$



Hence for a continuous random variable

$$P\{X < a\} = P\{X \le a\} =$$

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

is called the cumulative density function or distribution function.

Cumulative Distribution Function (cdf)

The cumulative distribution function (cdf) $F(x)$ of a discrete random variable X with p.m.f $p(x)$ or $f(x)$ is defined for every number x by

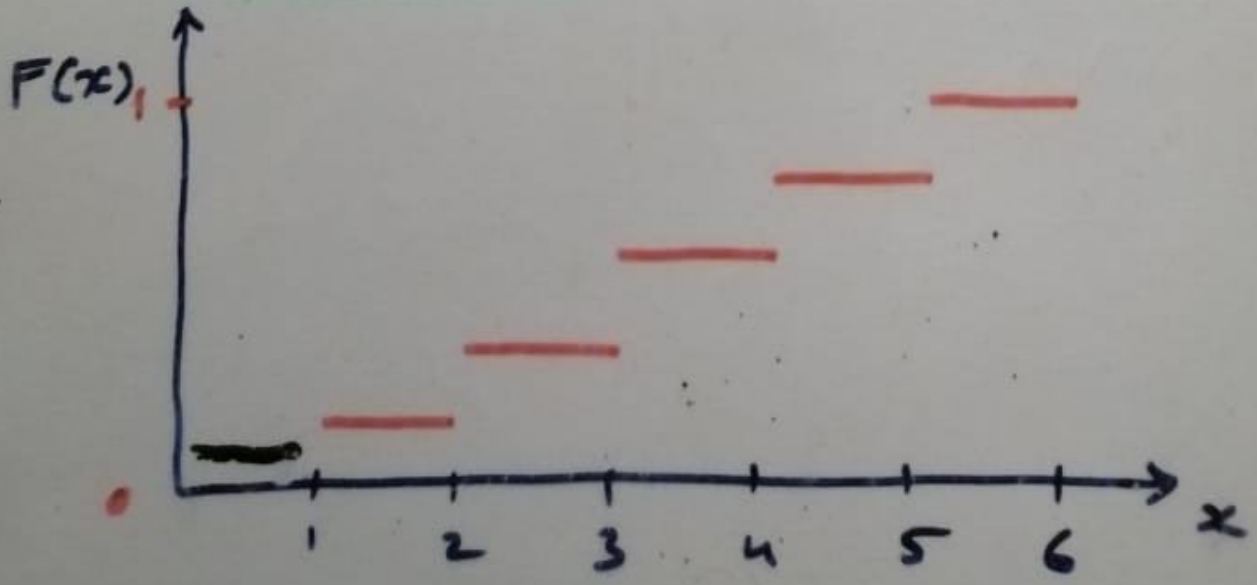
$$F(x) = P(X \leq x) = \sum_{y : y \leq x} p(y) = \sum_{x=0}^{x} p(x)$$

For any number x, $F(x)$ is the probability that the observed value of X will be atmost x.

For a continuous random variable

$$F(a) = P\{X < a\} = P\{X \leq a\} = \int_{-\infty}^{a} f(x)\,dx$$

For X a discrete r.v, the graph will have a jump at every possible value of X and will be flat between any two possible values of X. Such a graph is called a step function



Cumulative Distribution Function

The Cumulative distribution function (cdf) $F(x)$ of a discrete random variable X with p.m.f $p(x)$ is defined for every number x by

a) $$F(x) = P(X \leq x) = \sum_{y : y \leq x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be atmost x.

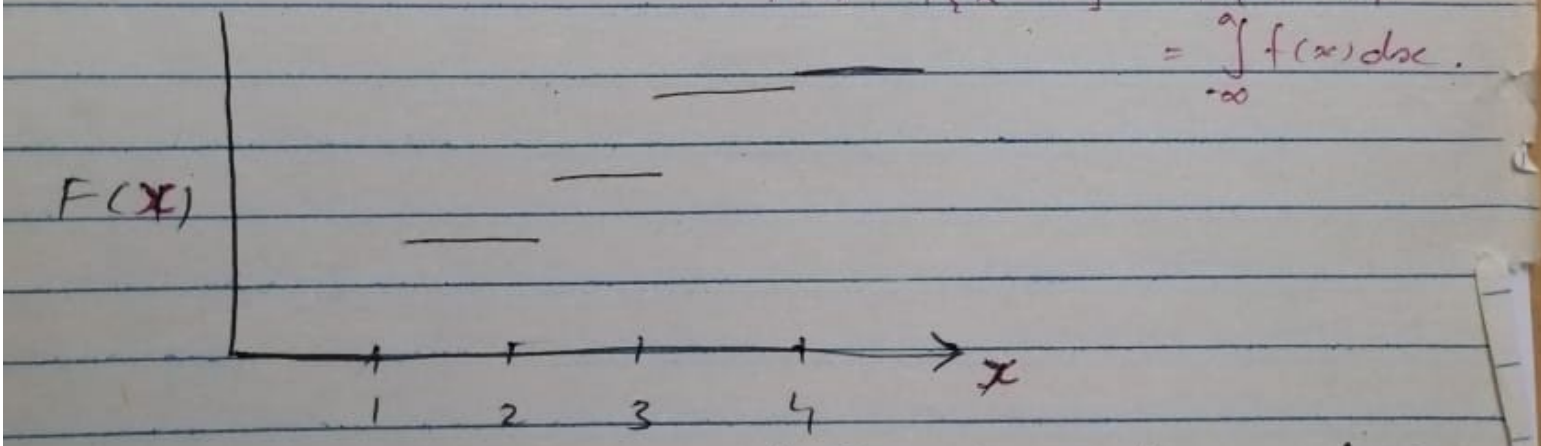
Example:

Suppose a p.m.f of Y is

Y	1	2	3	4
$p(y)$	.4	.3	.2	.1

$$F(y) = \begin{cases} 0 & \text{if } y < 1 \\ 0.4 & \text{if } 1 \leq y < 2 \\ 0.7 & 2 \leq y < 3 \\ .9 & 3 \leq y < 4 \\ 1 & 4 \leq y \end{cases}$$

A graph of $F(x)$ is



b) Hence for a continuous r.v

$$F(a) = P\{X < a\} = P\{X \leq a\} = \int_{-\infty}^{a} f(x)\,dx.$$

For X a d.r.v, the graph of $F(x)$ will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a step-function.

Problem:

A gas station operates two pumps, each of which can pump up to 10,000 gallons of gas in a month. The total amount of gas pumped at the station in a month is a random variable Y (measured in 10,000 gallons) with a p.d.f given by

$$f(y) = \begin{cases} y & , \ 0 < y < 1 \\ 2-y & , \ 1 \leq y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

a) Graph $f(y)$.

b) Find $F(y)$ and graph it

c) Find the probability that the station will pump between 8,000 and 12,000 gallons in a particular month.

d) Given that the station pumped more than 10,000 gallons in a particular month, find the probability that the station pumped more than 15,000 gallons during the month.

Problem:

A gas station operates two pumps, each of which can pump up to 10,000 gallons of gas in a month. The total amount of gas pumped at the station in a month is a random variable Y (measured in 10,000 gallons) with a p.d.f given by

$$f(y) = \begin{cases} y & , \ 0 < y < 1 \\ 2-y & , \ 1 \leq y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

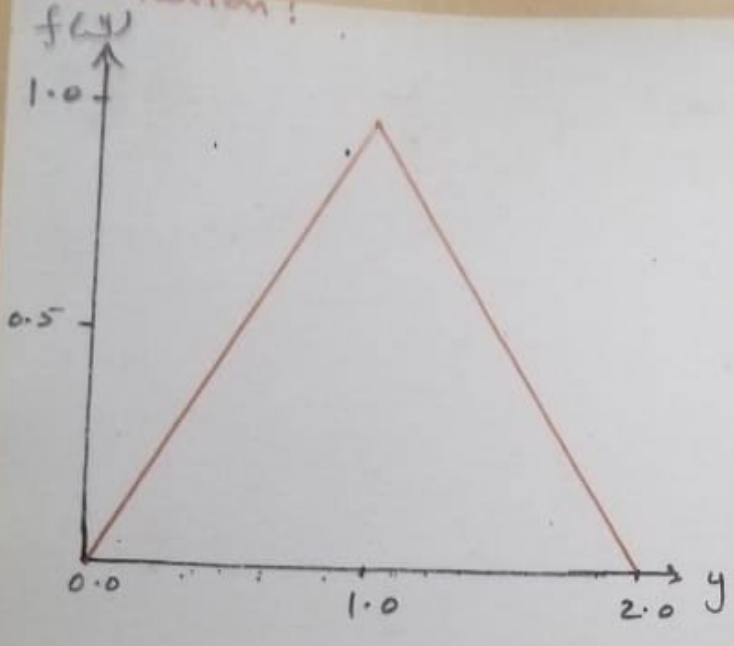
a) Graph $f(y)$.

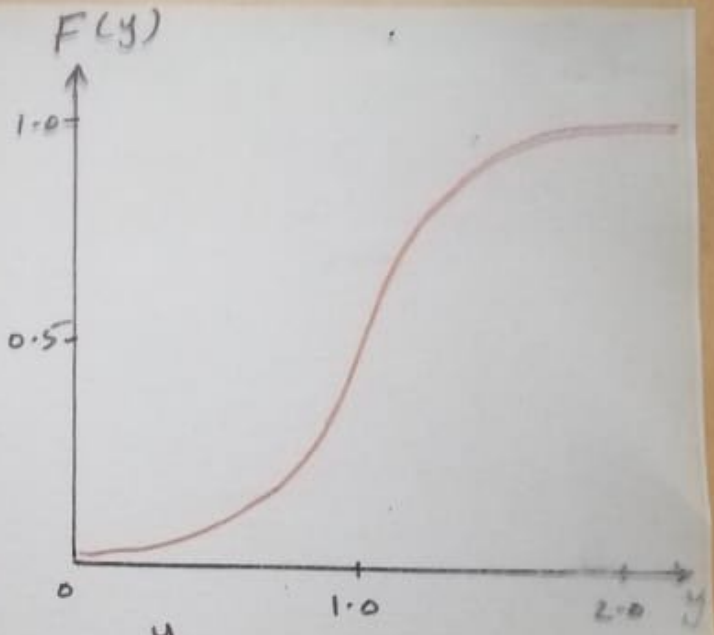
b) Find $F(y)$ and graph it

c) Find the probability that the station will pump between 8,000 and 12,000 gallons in a particular month.

d) Given that the station pumped more than 10,000 gallons in a particular month, find the probability that the station pumped more than 15,000 gallons during the month.

Solution:





b) $F(y) = P(Y \le y) = \int_{-\infty}^{y} f(y)\,dy$

$$= \begin{cases} \int_0^y y\,dy ; & 0 < y < 1 \\ \int_0^1 y\,dy + \int_1^y (2-y)\,dy ; & 1 \le y < 2 \\ 1 ; & y \ge 2 \end{cases}$$

$$= \begin{cases} \frac{y^2}{2} ; & 0 < y < 1 \\ \frac{y^2}{2}\Big|_0^1 + 2y\Big|_1^y - \frac{y^2}{2}\Big|_1^y ; & 1 \le y < 2 \\ 1 ; & y \ge 2 \end{cases}$$

$$= \begin{cases} \frac{y^2}{2} ; & 0 < y < \\ \frac{1}{2} + 2(y-1) - \frac{1}{2}(y^2-1) ; & 1 \le y < 2 \\ 1 ; & y \ge 2 \end{cases}$$

$$F(y) = \begin{cases} \frac{y^2}{2} & ; \quad 0 < y < 1 \\ 2y - \frac{y^2}{2} - 1 & ; \quad 1 \leq y < 2 \\ 1 & ; \quad y \geq 2 \end{cases}$$

c) $P(8000 \leq y \leq 12000) = P(0.8 \leq y \leq 1.2)$

$$= \int_{0.8}^{1} y\,dy + \int_{1}^{1.2} (2-y)\,dy$$

$$= \frac{1}{2}[y^2]_{0.8}^{1} + 2[y]_1^{1.2} - \frac{1}{2}[y^2]_1^{1.2}$$

$$= \frac{1}{2}(1-0.64) + 2(1.2-1) - \frac{1}{2}(1.44-1)$$

$$= 0.36$$

d) $$\frac{P(Y > 15000 \cap X > 10000)}{P(X > 10,000)} = \frac{P(Y > 15000)}{P(X > 10000)}$$

$$= \frac{P(Y > 1.5)}{P(X > 1.0)} = \frac{1 - P(Y \leq 1.5)}{1 - P(X \leq 1.0)}$$

$$P(Y \leq 1.5) = \int_0^1 y\,dy + \int_1^{1.5} (2-y)\,dy$$

$$= \frac{1}{2}[y^2]_0^1 + 2[y]_1^{1.5} - \frac{1}{2}[y^2]_1^{1.5}$$

$$= \frac{1}{2}(1-0) + 2(1.5-1) - \frac{1}{2}(2.25-$$

$$= 0.875$$

Date:

PRACTICAL 2: Probability Distributions

1. A random variable 'x' has a probability distribution

$$P(x) = \begin{cases} k/x & ; x = 2, 4, 8 \\ 0 & ; \text{otherwise} \end{cases}$$

a) Determine the value of the constant k
b) What is P[X=4] ?
c) What is P[X<4]
d) What is P[3≤X≤9] ?

2. A random variable 'X' has a CDF

$$F(x) = \begin{cases} 0 & x < -3 \\ 0.4 & -3 \leq x < 5 \\ 0.8 & 5 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

a) Draw a graph of the CDF
b) Write P(x), the p.d.f of X.

3. Show that the following function are p.d.f for some values of k and determine k. Also sketch the functions

a) $f(x) = kx^2 \quad ; \quad 0 \leq x \leq 2$
b) $f(x) = k(2-x), \quad 0 \leq x \leq 1$
c) $f(x) = ke^{-x} \quad 0 < x$

4. A continious random variable has p.d.f

$$f(x) = kx, \quad 0 \leq x \leq 4$$

a) Find the value of the constant k.
b) Sketch the function f(x)
c) find $P[1 \leq x \leq 2\frac{1}{2})$ and sketch

Assignment 2

Q1. (6 points) The sample space of a random experiment is $\{a, b, c, d, e, f\}$ and each outcome is equally likely. A random variable is defined as follows

outcome	a	b	c	d	e	f
x	0	0	1.5	1.5	2	3

Determine the probability mass function of X. Determine the following probabilities:

(a) $P(X = 1.5)$ (b) $P(0.5 < X < 2.7)$ (c) $P(X > 3)$

(d) $P(0 \leq X < 2)$ (e) $P(X = 0 \text{ or } X = 2)$

Solution to Q1:

Probability mass function is

$$P(X = 0) = P(\{a, b\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6};\ P(X = 1.5) = \frac{2}{6},\ P(X = 2) = \frac{1}{6},\ P(X = 3) = \frac{1}{6}.$$

(a) $P(X = 1.5) = \frac{2}{6}$

(b) $P(0.5 < X < 2.7) = P(X = 1.5) + P(X = 2) = \frac{3}{6}$

(c) $P(X > 3) = 0$

(d) $P(0 \leq X < 2) = P(X = 0) + P(X = 1.5) = \frac{4}{6}$

(e) $P(X = 0 \text{ or } X = 2) = \frac{3}{6}$

Marking scheme for Q1:

Completely correct p.m.f. - 1 point, correct answer for each part - 1 point. Total - 6 points.

Q2. (2 points) Determine the mean and the variance in Question Q1

Solution to Q2:

$\mathrm{E}(X) = 1.33$, $\mathrm{Var}(X) \approx 1.15$

Marking scheme for Q2:

1 point for the mean and the variance. Total - 2 points.

Q3. We say that X has *uniform distribution* on a set of values $\{x_1, \ldots, x_k\}$ if

$$P(X = x_i) = \frac{1}{k}, \qquad i = 1, \ldots, k.$$

The thickness measurements of a coating process are *uniformly distributed* with values 0.15, 0.16, 0.17, 0.18, 0.19. Determine the mean and variance.

Solution to Q3:

We have $P(X = 0.15) = \cdots = P(X = 0.19) = 1/5$. Mean: 0.17; Variance:

$$\frac{1}{5}\left(0.01^2 + 0.02^2 + 0.02^2 + 0.01^2\right)$$

Marking scheme for Q3:

This question will not be marked

Date:

2

Binomial Experiment:

A binomial experiment is one that possess the following properties.

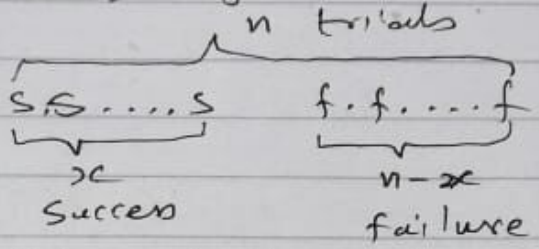
a) The experiment consists of n repeated trials
b) Each trial results in an outcomes that may be classified as success or a failure.
c) The probability of success, denoted by p remains constant from trial to trial.
d) The repeated trials are independent.

Date:

Binomial Distribution:

Consider an experiment with two possible outcomes call them success (s) and failure (f) with $P(s) = p$ and $P(f) = q$ such that $p+q=1$. Let 'X' be a random variable denotes the number of successes in n independent repeated trials, e.g.

Consider



The probability distribution of the particular sequence (by multiplicative Law of independent events) is

$$p^x q^{(n-x)} \quad \text{or} \quad p^x (1-p)^{n-x}$$

The number of sequence in which 'x' successes and n-x failures are observed in some order is $\binom{n}{x}$ ways. which is binomial coefficient.

Thus the probability Distribution that exactly x successes and n-x failures occur in n independent trials is

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x} \; ; \; x = 0, 1, 2, \ldots, n$$

which is known as binomial distribution with index 'n' and parameter p.

$$\mu_4' = E[x^4] = \sum_{x=0}^{n} x^4 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x] \times \binom{n}{x} p^x q^{n-x}$$

$$\boxed{\mu_4' = n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np.}$$

Moment about the mean :

$$\mu_r = \frac{1}{n} \sum (x_i - \bar{x})^r$$

$$\mu_1 = \mu_1' - \mu_1' = 0.$$

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= n(n-1)p^2 + np - n^2p^2$$

$$= n^2p^2 - np^2 + np - n^2p^2$$

$$= np(1-p)$$

$$\sigma^2 = \boxed{\mu_2 = npq} \Rightarrow \sigma = \sqrt{\mu_2} = \sqrt{npq}$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= [n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np] - 3[n(n-1)p^2 + np]np + 2n^3p^3$$

$$= n[n^2 - 3n + 2]p^3 + 3[n^2 - n]p^2 + np - 3n^2p^2 - 3n^2p^3 + 3np^2 + 2n^3p^3$$

$$= n^3p^3 - 3n^2p^3 + 2np^3 + 3n^2p^2 - 3np^2 + np - 3n^2p^2 - 3n^3p^3 + 3n^2p + 2n^3$$

$$= 2np^3 - 3np^2 + np = np[1 - 3p + 2p^2]$$

$$\mu_1' = np\left[q^{n-1} + \binom{n-1}{1} p q^{n-2} + \cdots + p^{n-1}\right]$$

$$= np\,[q+p]^{n-1}$$

$\because q+p=1$

$$\boxed{\mu_1' = np}$$

at $r=2$

$$\mu_2' = E[x^2] = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x + x(x-1)] \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} + \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x q^{n-x}$$

$$= np + 0 + 0 + 2\cdot 1 \binom{n}{2} p^2 q^{n-2} + 3\cdot 2 \binom{n}{3} p^3 q^{n-3} + \cdots$$

$$\cdots + n(n-1) \binom{n}{n} p^n q^{n-n}$$

$$= np + n(n-1)p^2\left[q^{n-2} + \binom{n-2}{1} p q^{n-3} + \cdots + p^{n-2}\right]$$

$$= np + n(n-1)p^2\,[q+p]^{n-2}$$

$$= np + n(n-1)p^2$$

$$= np\,[1 + (n-1)p] = np\,[1 + np - p]$$

$$= np\,[q + np]$$

$$\boxed{\mu_2' = npq + n^2p^2}$$

Similarly Proceeding we get

$$\mu_3' = E[x^3] = \sum_{x=0}^{n} x^3 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} [x(x-1)(x-2) + 3x(x-1) + x] \binom{n}{x} p^x q^{n-x}$$

$$= n(n-1)(n-2)p^3(q+p)^{n-3} + 3n(n-1)p^2(p+q)^{n-2}$$

$$+ np(q+p)^{n-1}$$

$$\boxed{\mu_3' = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np}$$