22/4/24 **Correlation and Regression:**

Suppose two variables x and y are related in such a way that an increase in one variable is accompanied by an increase or decrease in the other variable. Such a relation is called correlation. If x and y increases / decreases together then x and y are positively correlated. On the other hand, if y increases as x decreases and vice versa, then we say that x and y are negatively correlated.

For eg: Demand and price of a commodity are positively correlated whereas supply and price are negatively correlated.

* **Mean of x & y:**

Let us suppose $x_1, x_2, x_3, \ldots, x_n$ are discrete values of the variable x, then the mean of x is defined as $\bar{x} = \frac{\sum x_i}{n}$

Similarly $\bar{y} = \frac{\sum y_i}{n}$

* **Variance:**

The variance wrt variable x is denoted by V_x and it is defined as

$$V_x = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Similarly $V_y = \frac{1}{n} \sum (y_i - \bar{y})^2$

* **Standard deviation**

The standard deviation is defined as positive square root of variance and denoted by σ_x and σ_y.

$\therefore \sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$ $\quad \sigma_y = \sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}$

* **Correlation Coefficient:**

The measure of correlation b/w two variables x and y is known correlation coefficient and it is defined by

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

where $\sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}$$

If we choose $x_i - \bar{x} = X$ and similarly $y_i - \bar{y} = Y$, then $\sigma_x = \sqrt{\frac{1}{n}\Sigma X^2}$, $\sigma_y = \sqrt{\frac{1}{n}\Sigma Y^2}$

then $r = \frac{\Sigma XY}{n \sigma_x \sigma_y} = \frac{\Sigma XY}{n\sqrt{\frac{1}{n}\Sigma X^2}\sqrt{\frac{1}{n}\Sigma Y^2}}$

$$\boxed{r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}}$$

1) Correlation coefficient (r) always lies between -1 and 1 i.e $-1 \leq r \leq 1$.

2) If $r \to 1$, then x and y are positively correlated.

If $r \to -1$, then x and y are negatively correlated

If $r \to 0$, then no correlation b/w x and y.

① Find the correlation coefficient from the following data.

Husband age (x)	23	27	28	29	30	32
Wife age (y)	18	22	23	24	26	29

Let $n = 6$

$\bar{x} = \frac{\Sigma x_i}{n} = 28.16$

$\bar{y} = \frac{\Sigma y_i}{n} = 23.64$

WKT $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$ where $X = x_i - \bar{x}$, $Y = y_i - \bar{y}$

x	y	X	Y	XY	X^2	Y^2
23	18	-5.16	-5.64	28.896	26.63	31.36
27	22	-1.16	-1.64	1.856	1.35	2.56
28	23	-0.16	-0.64	0.096	0.025	0.36
29	24	0.84	0.36	0.336	0.705	0.16
30	26	1.84	2.36	4.416	3.385	5.76
32	29	3.84	5.36	20.736	14.745	29.16

$\Sigma XY = 56.336$

$\Sigma X^2 = 46.831$

$\Sigma Y^2 = 69.36$

$\therefore r = \frac{56.336}{\sqrt{46.83 \times 69.36}}$

$\therefore r = 0.9884$

~~$r = 0.017$~~

Regression Analysis

Simple Linear Regression

Regression is an estimation of one variable in terms of the other. If That is, it is the representation of y as a function of x and vice versa.

If x and y are correlated, the best fitting straight line in the least square sense gives a reasonably a good relation between x and y.

The best fitting straight line of the form $y = ax + b$ is known as regression line y on x.

Illy the best fitting straight line of the form $x = cy + d$ is known as regression line x on y.

Equations of regression line:

[1] Regression line y on x:

If x and y are correlated, then an equation of regression line y on x is given by

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})} \quad \text{where } b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right)$$

[2] Regression line x on y:

If x and y are correlated, then the regression line x on y is given by

$$\boxed{x - \bar{x} = b_{xy}(y - \bar{y})} \quad \text{where } b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right)$$

i) WKT $b_{yx} \cdot b_{xy} = r^2$

$$\Rightarrow \boxed{r = \pm\sqrt{b_{yx} \cdot b_{xy}}}$$

i) If both b_{yx} and b_{xy} are positive then 'r' is considered to be positive

ii) If both b_{yx} and b_{xy} are negative then 'r' is considered to be negative.

① Find the coefficient of correlation and hence equations of regression line from the following data:

x	10	14	18	22	26	30
y	18	12	24	6	30	36

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{120}{6} = 20$$

$$\bar{y} = \frac{\Sigma y_i}{n} = 21$$

WKT $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \cdot \Sigma Y^2}}$ where $X = x - \bar{x}$, $Y = y - \bar{y}$

x	y	X	Y	XY	X^2	Y^2
10	18	-10	-3	30	100	9
14	12	-6	-9	54	36	81
18	24	-2	3	-6	4	9
22	6	2	-15	-30	4	225
26	30	6	9	54	36	81
30	36	10	15	150	100	225

$\Sigma XY = 252$ $\Sigma X^2 = 280$ $\Sigma Y^2 = 630$

$$r = \frac{252}{\sqrt{280 \times 630}}$$

$$r = 0.6$$

Regression line y on x:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$b_{xy} = r\left(\frac{\sigma_y}{\sigma_x}\right)$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma X^2} = \sqrt{\frac{1}{6} \cdot 280} = 6.831$$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma y^2} = 10.2469$

$b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right) = 0.6\left(\frac{\sqrt{105}}{2\sqrt{105}} \cdot 3\right) = 0.9$

∴ Regression line y on x is

$y - 21 = 0.9(x - 20) = 0.9x - 18.$

$\Rightarrow \boxed{y = 0.9x + 3}$

Regression line x on y is

$(x - \bar{x}) = b_{xy}(y - \bar{y})$ where $b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right)$

$= 0.6\left(\frac{2\sqrt{105}}{3\sqrt{105}}\right)$

$b_{xy} = 0.4.$

$\therefore (x - 20) = 0.4(y - 21)$

$\boxed{x = 0.4y + 11.6}$

If m_1 and m_2 are the slopes of two intersecting lines then the angle of intersection is

$$\boxed{\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|}$$

Theorem.

If θ is an acute angle b/w two regression lines in case of two variables x and y, show that

$\tan\theta = \left(\frac{1 - r^2}{r}\right)\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$ where σ_x, σ_y and r have their usual meanings. Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

Let the regression line y on x

$(y - \bar{y}) = b_{yx}(x - \bar{x})$

$\therefore$ Slope, $m_1 = b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right)$

Regression line x on y,

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$\therefore$ Slope, $m_2 = \frac{1}{b_{xy}} = \frac{1}{r}\left(\frac{\sigma_y}{\sigma_x}\right)$

WKT, $\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$

$$= \frac{r\left(\frac{\sigma_y}{\sigma_x}\right) - \frac{1}{r}\left(\frac{\sigma_y}{\sigma_x}\right)}{1 + r \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{1}{r} \cdot \frac{\sigma_y}{\sigma_x}}$$

$$= \frac{\frac{\sigma_y}{\sigma_x}\left(r - \frac{1}{r}\right)}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$\tan\theta = \left| \frac{\left(\frac{r^2 - 1}{r}\right)\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \right|$$

If $r = 0$, $\tan\theta \to \infty$

$\therefore \boxed{\theta = \pi/2}$

The two regression lines are perpendicular.

If $r = \pm 1$, $\tan\theta = 0$

$\therefore \boxed{\theta = 0}$

The two regression lines are parallel / coincident

Show that correlation coefficient r lies between -1 and 1.

WKT $r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$

$$r^2 = \frac{(\Sigma XY)^2}{\Sigma X^2 \Sigma Y^2}$$

$$r^2 (\Sigma X^2 \Sigma Y^2) = (\Sigma XY)^2$$

From Schwartz property,

$$(\Sigma XY)^2 \leq \Sigma X^2 \Sigma Y^2$$

$$r^2 (\Sigma X^2 \Sigma Y^2) \leq \Sigma X^2 \Sigma Y^2$$

$$r^2 \leq 1$$

$$\therefore -1 \leq r \leq 1$$

$\bar{x}$ and $\bar{y}$ always satisfies the equations of regression lines

The regression equations of two variables x and y are given by $x = 0.7y + 5.2$, $y = 0.3x + 2.8$. Find the mean of the variables x and y and also find the ~~coorda~~ coefficient of correlation b/w x & y.

$x = 0.7y + 5.2$ ~~(y on x)~~ (x on y)

~~$b_{xy} = 0.7$~~ $y = 0.3x + 2.8$ (y on x)

Since $\bar{x}$ and $\bar{y}$ satisfy the two equations,

$\therefore \bar{x} = 0.7\bar{y} + 5.2 \Rightarrow \bar{x} - 0.7\bar{y} = 5.2$

$\bar{y} = 0.3\bar{x} + 2.8 \Rightarrow -0.3\bar{x} + \bar{y} = 2.8$

$\therefore \boxed{\bar{x} = 9.063}$ $\boxed{\bar{y} = 5.519}$

Also, $b_{xy} = 0.7$, $b_{yx} = 0.3$

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

$$= \sqrt{0.7 \times 0.3}$$

$$r = 0.4582$$

② In a partially destroyed laboratory record, only equations of regression line y on x and x on y are available as $4x - 5y = -33$ and $20x - 9y = 107$ respectively. Calculate the mean of x and y and coefficient of correlation between x and y.

Given, $4x - 5y = -33$ (y on x)

$20x - 9y = 107$ (x on y)

Since $\bar{x}$ and $\bar{y}$ satisfy the equations,

$4\bar{x} - 5\bar{y} = -33$

$20\bar{x} - 9\bar{y} = 107$

$\therefore \boxed{\bar{x} = 13} \quad \boxed{\bar{y} = 17}$

Rewrite given eqns,

$y = \frac{4x}{5} + 33 \quad \therefore b_{yx} = 4/5$

$x = \frac{9}{20}y + \frac{107}{20} \quad \therefore b_{xy} = 9/20.$

$\therefore r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{\frac{4}{5} \cdot \frac{9}{20}}$

$\therefore \boxed{r = 0.6}$

Random variable

* **Random experiment:**

An experiment which is done under specified condition with the outcome not known.

For eg: Tossing a fair coin

Throwing a cubical die

Finding temperature of particular place, ...

Sample space

The set of all possible outcome of any random experiment is called the sample space denoted by S.

For eg: If we toss a coin twice, then the

Sample space, $S = \{HH, HT, TH, TT\}$

If we throw two die, then

$S = \{(1,1), (1,2), (1,3), \ldots, (6,6)\}$

Random variable

Suppose that to each outcome of sample space, we assign a number according to some rule, then we have a function defined on the sample space. This function is known as the random variable. and denoted by X, Y, Z, etc.

For eg:

① Consider the experiment of tossing a coin thrice, then the sample space $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

Let X represent No of heads turnup

Then

$X(HHH) = 3$, $X(HHT) = X(HTH) = X(THH) = 2$

$X(HTT) = X(THT) = X(TTH) = 1$, $X(TTT) = 0$

$\therefore$ Range of $X = \{0, 1, 2, 3\}$

② Consider the random experiment of throwing two die. $\therefore S = \{(1,1), (1,2), \ldots, (6,6)\}$

Let X denote sum of the two numbers

Then Range $(X) = \{2, 3, \ldots, 12\}$

③ Consider the random experiment of finding height of a student of RIT. Then

$S = \{$All students of RIT$\}$ and X - Height of student

Range of $X = [4ft, 7ft]$

④ Consider a random experiment of finding duration of telephone ~~convos~~ convo in RIT in one particular day. Then

Range of $X = [1\,sec, 2\,hr]$

Discrete and Continuous Random Variable

If the random variable X takes ~~infinite~~ or countable no. of values, then the random variable X is said to discrete.

eg ① & ② are examples of discrete random variable

If the random variable X takes ~~infin~~ any value in some interval $[a, b]$, then it is said to be continuous random variable.

eg ③ & ④ are examples of continuous random variable

Discrete Probability Distribution

(OR)

* **Probability mass function (pmf)**

Let X be discrete random variable. Then the function $P[X = x_i] = p(x_i)$ is said to discrete probability function of p or (pmf) if it satisfies the following conditions:

i) $p(x_i) \geq 0$

ii) $\sum_i p(x_i) = 1$.

* **Expected Value of Random Variable X (or) Mean of X**

The expected value of random variable X is defined as $\mu = E(x) = \sum_i x_i p(x_i)$

Variance

The variance wrt random variable X is defined as

$$V = E[(x_i - \mu)^2] = \sum_i (x_i - \mu)^2 p(x_i)$$

Alternate:

$$V = E(x^2) - [E(x)]^2 \quad \text{where} \quad E(x^2) = \sum_i x_i^2 p(x_i)$$

Standard Deviation (σ):

The standard deviation wrt random variable X is defined as positive square root of variance i.e.

$$\sigma = \sqrt{V}$$

A fair coin (unbiased coin) is tossed thrice. Let X denote the no. of heads. Find the probability distribution of X. Also find mean, variance, and standard deviation.

Let the sample space, S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Let X be number of heads.

X(HHH) = 3

X(HHT) = X(HTH) = X(THH) = 2

X(HTT) = X(THT) = X(TTH) = 1

X(TTT) = 0

$P[X=0] = \frac{1}{8}$ $P[X=1] = \frac{3}{8}$ $P[X=2] = \frac{3}{8}$ $P[X=3] = \frac{1}{8}$

∴ The probability distribution function

x_i	0	1	2	3
$p(x_i)$	1/8	3/8	3/8	1/8

Mean, $\mu = \sum E(x_i) = \sum x_i p(x_i)$

$= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8}$

$\boxed{\mu = 1.5}$

Variance, $V = E[(x_i - \mu)^2] = \sum (x_i - \mu)^2 p(x_i)$

$= (0-1.5)^2 \times \frac{1}{8} + (1-1.5)^2 \times \frac{3}{8} + (2-1.5)^2 \times \frac{3}{8} + (3-1.5)^2 \times \frac{1}{8}$

$\boxed{V = 0.75}$

Standard deviation, $\sigma = \sqrt{V}$

$\boxed{\sigma = 0.866}$

Two unbiased dice are thrown and let X represent the sum of two numbers that appear. Find probability distribution, mean, and variance

Let sample space be S

$S = \{(1,1), (1,2), (1,3), \ldots, (6,6)\}$

Let X denote sum of two numbers on dice

$\therefore X(1,1) = 2$

$X(1,2) = X(2,1) = 3$

$X(1,3) = X(3,1) = X(2,2) = 4$

$X(1,4) = X(4,1) = X(2,3) = X(3,2) = 5$

$X(1,5) = X(5,1) = X(2,4) = X(4,2) = X(3,3) = 6$

$X(1,6) = X(6,1) = X(2,5) = X(5,2) = X(3,4) = X(4,3) = 7$

$X(2,6) = X(6,2) = X(3,5) = X(5,3) = X(4,4) = 8$

$X(3,6) = X(6,3) = X(4,5) = X(5,4) = 9$

$X(4,6) = X(6,4) = X(5,5) = 10$

$X(5,6) = X(6,5) = 11$

$X(6,6) = 12$

$P[X=2] = P[X=12] = 1/36$
$P[X=3] = P[X=11] = 2/36 = 1/18$
$P[X=4] = P[X=10] = 3/36 = 1/12$
$P[X=5] = P[X=9] = 4/36 = 1/9$
$P[X=6] = P[X=8] = 5/36$
$P[X=7] = 6/36 = 1/6$

x_i	2	3	4	5	6	7	8	9	10	11	12
$p(x_i)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Mean, $\mu = \sum x_i p(x_i)$

$= \frac{2}{36} + \frac{3}{36} \times 2 + 4 + (4+10) \times \frac{3}{36} + (5+9) \times \frac{4}{36} + (6+8) \times \frac{5}{36}$

$+ 7 \times \frac{6}{36}$

$\mu = 7$

Variance, $V = \sum x^2 p(x) - \mu^2$

$= \frac{23860}{9} - (7)^2 = 2644.11$

Standard deviation, $\sigma = \sqrt{V}$

$=$

124

③ A random variable X has the following density fn.

x_i	1	2	3	4	5	6	7
$p(x_i)$	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2+k$

i) Find k

ii) Probability of $X \geq 6$.

iii) $P[2 \leq X < 6]$

iv) $E(x)$

Given, $p(x_i)$ - probability distribution fn

i) $\sum p(x_i) = 1$

$k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$

$10k^2 + 9k - 1 = 0$

$\boxed{k = \frac{1}{10}, \; k = -1}$

Since $p(x_i)$ cannot be negative, we discard $k = -1$, Therefore, $\boxed{k = \frac{1}{10}}$

ii) $P(X \geq 6)$

$= p(6) + p(7)$

$= 2k^2 + 7k^2 + k$

$= 9k^2 + k$

$= 9\left(\frac{1}{100}\right) + \frac{1}{10}$

$= \frac{19}{100} = 0.19$

iii) $P[2 < X < 6]$

$= p(3) + p(4) + p(5) + p(2)$

$= 2k + 3k + k^2 + 2k$

$= k^2 + 7k$

$= \frac{1}{100} + \frac{7}{10}$

$= \frac{71}{100} = 0.71$

iv) $E(x) = \sum x_i p(x_i)$

$= k + 4k + 6k + 12k + 5k^2 + 12k^2 + 49k^2 + 7k$

$= 66k^2 + 30k$

$= \frac{66}{100} + \frac{30}{10}$

$= \frac{366}{100} = 3.66$

Continuous probability function (OR) Probability Density Function

Let X be a continuous random variable X. Then the function $f(x)$ is said to be a continuous probability distribution function if it satisfies the following two conditions.

i) $f(x_i) \geq 0 \quad \forall x_i$

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

Consequently the probability that a random variable X lies between a and b is

$$P[a<x<b] = \int_a^b f(x)\,dx.$$

Expected Value of X (or) Mean of X

The expected value of a continuous random variable X is defined as $E(x) = \mu = \int_{-\infty}^{\infty} x f(x)\,dx$

Variance of X

The variance wrt a continuous random variable X is defined as $V = E(X^2) - (E(x))^2$

where $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx.$

Standard Deviation of X

The standard deviation of X is defined as the positive square root of variance i.e $\sigma = \sqrt{V}$

Cumulative distribution function

Cumulative distribution function c.d.f is denoted

by $F(x)$ and is defined by

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

cdf helps to determine the probabilities in different intervals.

① A random variable X has the density function $f(x) = \frac{c}{1+x^2}$ where $-\infty < x < \infty$.

i) Find the value of constant c

ii) Find the probability that x lies b/w $1/3$ and 1

iii) Find the corresponding cdf $F(x)$.

Given that, $f(x) = \frac{c}{1+x^2}$ is the probability density function.

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

$$\int_{-\infty}^{\infty} \frac{c}{1+x^2}\,dx = 1$$

$$c \tan^{-1}x \Big|_{-\infty}^{\infty} = 1$$

$$c(\tan^{-1}\infty - \tan^{-1}(-\infty)) = 1$$

$$c\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = 1$$

$$\therefore \boxed{c = \frac{1}{\pi}}$$

ii) $P\left[\frac{1}{3} < x < 1\right] = \int_{1/3}^{1} f(x)\,dx = \int_{1/3}^{1} \frac{c}{1+x^2}\,dx$

$$= c\left[\tan^{-1}x\right]_{1/3}^{1}$$

$$= \frac{1}{\pi}\left[\tan^{-1}1 - \tan^{-1}1/3\right]$$

$$= \frac{1}{\pi}\left[\frac{\pi}{4} - \tan^{-1}1/3\right] = 0.1476$$

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

$= \int_{-\infty}^{x} \frac{c}{1+x^2}\,dx$

$= c\,[\tan^{-1}x]_{-\infty}^{1}$

$= \frac{1}{\pi}\left[\tan^{-1}x + \frac{\pi}{2}\right]$

The diameter of an electric cable is assumed to a continuous random variable with pdf $f(x) = \begin{cases} 6x(1-x) & 0<x<1 \\ 0 & \text{otherwise} \end{cases}$

Verify that the above function a valid probability function. Also find the mean and variance.

i) To verify, $\int_0^1 f(x)\,dx = \int_0^1 6x(1-x)\,dx$.

$= \int_0^1 (6x - 6x^2)\,dx$

$= \left[\frac{6x^2}{2} - \frac{6x^3}{3}\right]_0^1$

$= 3 - 2$

$= 1$. Hence verified

ii) $\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x\,6x(1-x)\,dx$.

$= \int_0^1 (6x^2 - 6x^3)\,dx$.

$= \frac{6x^3}{3} - \frac{6x^4}{4}\Big|_0^1 = 2 - \frac{3}{2} = \frac{1}{2}$

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

$= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$

$= \int_{0}^{1} x^2 \cdot 6x(1-x)\,dx - \frac{1}{2}$

$= \int_{0}^{\infty} 6x^3 - 6x^4\,dx - \frac{1}{2}$

$= \left[\frac{6}{4}x^4 - \frac{6}{5}x^5\right]_0^1 - \frac{1}{2}$

$= \frac{3}{2} - \frac{6}{5} - \frac{1}{2} = \frac{15 - 12 - 5}{10}$

$V = -0.2$

6/5/24

③ The pdf of a random variable X is given by

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

Find i) cumulative distribution function $F(x)$

ii) Probability of $X \geq 1.5$. $P[X \geq 1.5]$

Given $f(x) = \begin{cases} x & 0 \le x \le 1 \\ 2-x & 1 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases}$

The cdf cumulative distribution function cdf,

$F(x) = \int_{-\infty}^{x} f(x)\,dx = P[X \le x]$

I. In $(-\infty, 0)$

$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx = 0$

II. In $[0, 1]$

$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{x} f(x)\,dx$

$= 0 + \int_{0}^{x} x\,dx$

$$= \frac{x^2}{2}\Big|_0^x$$

$$= \frac{x^2}{2}$$

II. In $[1,2]$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{x} f(x)\,dx.$$

$$= 0 + \frac{x^2}{2} + \int_1^x (2-x)\,dx + \int_0^1 x\,dx.$$

$$= \frac{x^2}{2}\Big|_0^1 + 2x - \frac{x^2}{2}\Big|_1^x + \frac{x^2}{2}$$

$$= \frac{1}{2} - 0 + 2x - \frac{x^2}{2} - 2 + \frac{1}{2} + \frac{x^2}{2} - \frac{1}{2}$$

$$= 2x - \frac{x^2}{2} - 1$$

III. In $(2, -\infty)$

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx + \int_{2}^{x} f(x)\,dx.$$

$$= \int_0^1 x\,dx + \int_1^2 2-x\,dx$$

$$= \frac{x^2}{2}\Big|_0^1 + \left[2x - \frac{x^2}{2}\right]_1^2$$

$$= \frac{1}{2} + 4 - 2 - 2 + \frac{1}{2}$$

$$= 1$$

$$\therefore P[X \geq 1.5] = \int_{1.5}^{\infty} f(x)\,dx = \int_{1.5}^{2} f(x)\,dx + \int_{2}^{\infty} f(x)\,dx$$

$$= \int_{1.5}^{2} (2-x)\,dx + 0$$

$$= 2x - \frac{x^2}{2}\Big|_{1.5}^{2} = 4 - 2 - 3 + \frac{9}{8} = \frac{1}{8}$$

Bernoullie's trial (or) Repeated trial

A random experiment with only two possible outcomes namely success and failure is called Bernoullie's trial.

where the probability of success is denoted by p and that of failure is denoted by q.

Clearly $p+q=1$

For eg. ① in the exp. of tossing a fair coin, the head denotes success.

Here $p=\frac{1}{2}$ $q=\frac{1}{2}$ $\Rightarrow p+q=1$

② in the random exp. of throwing two dice, appearance of sum 10 is success and not appearing the sum 10 is failure

Here, $p=\frac{3}{36}=\frac{1}{12}$, $q=\frac{33}{36}=\frac{11}{12}$

Binomial distribution function

Let p be the probability of success and q is the probability of failure. The probability of x successes out of n trials is given by

$$p(x) = {}^nC_x\, p^x\, q^{n-x}$$

where $x = 0, 1, 2, \ldots, n$

This is called Bernoullie distribution. The above distribution can also be represented by

x	0	1	2	3	....	n
p(x)	q^n	${}^nC_1\, p\, q^{n-1}$	${}^nC_2\, p^2\, q^{n-2}$	${}^nC_3\, p^3\, q^{n-3}$		p^n

note: ① $(a+b)^n = a^n + {}^nC_1 a^{n-1} b + {}^nC_2 a^{n-2} b^2 + {}^nC_3 a^{n-3} b^3 + \ldots + b^n$ is known as binomial expansion of $(a+b)^n$

① Obtain the mean, & variance and standard deviation of binomial distribution function.

Let the pdf of binomial distribution is

$$p(x) = {}^nC_x p^x q^{n-x}$$

x	0	1	2	3	....	n
$p(x)$	q^n	${}^nC_1 p q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	${}^nC_3 p^3 q^{n-3}$		p^n

I. Mean, $\mu = E(x) = \sum x p(x)$

$$= 0(q^n) + 1({}^nC_1 p q^{n-1}) + 2({}^nC_2 p^2 q^{n-2}) + \ldots + n p^n$$

$$= np q^{n-1} + \frac{n(n-1)}{2} p^2 q^{n-2} + \ldots + np^n$$

$$= np\left(q^{n-1} + (n-1) p q^{n-2} + \ldots + p^{n-1}\right)$$

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

$$= np(1)^{n-1}$$

$$\boxed{\mu = np}$$

II. Variance, $V = E(X^2) - (E(X))^2$

$$V = \sum x^2 p(x) - \mu^2$$

$$V = \sum x^2 p(x) - (np)^2 \quad \text{---} (1)$$

$$\sum x^2 p(x) = \sum (x^2 - x + x) p(x)$$

$$= \sum x(x-1) p(x) + \sum x p(x)$$

$$= \sum x(x-1) p(x) + np$$

$$\therefore (1) \Rightarrow V = \sum x(x-1) p(x) + np - (np)^2$$

$$= 0 + 0 + 2\,{}^nC_2 p^2 q^{n-2} + 6\,{}^nC_3 p^3 q^{n-3} + \ldots + n(n-1) p^n + np - (np)^2$$

$$= 2\frac{n(n-1)}{2} p^2 q^{n-2} + 6\frac{n(n-1)(n-2)}{6} p^3 q^{n-3} + \ldots + n(n-1) + np - (np)^2$$

$= n(n-1)p^2\left(q^{n-2} + (n-2)pq^{n-3} + \dots + p^{n-2}\right) + np - (np)^2$

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

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

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

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

$= np - p^2n$

$= np(1-p)$

$\boxed{V = npq}$

III. Standard deviation

$\sigma = \sqrt{V}$

$\boxed{\sigma = \sqrt{npq}}$

The mean and variance of binomial distribution are 16 and 8. Find the distribution of x.

Given, $\mu = 16 \quad V = 8$.

$\mu = np$ and $V = npq$.

$npq = 8$.

$16q = 8$

$\boxed{q = \frac{1}{2}}$ $\therefore \boxed{p = 1 - q = \frac{1}{2}}$

$\mu = np = 16$.

$n = \frac{16}{1/2}$

$\boxed{n = 32}$

$\therefore p(x) = {}^{32}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{32-x}$.

$p(x) = {}^{32}C_x \left(\frac{1}{2}\right)^{32}$

The probability that a man aged 60 will live upto 70 is 0.65. What is the probability that out of 10 men, now aged 60 atleast 7 will live upto 70.

given the probability a man aged 60 will live till 70, $p = 0.65$ (success)

Failure, $q = 1 - 0.65 = 0.35$

No. of trials, $n = 10$

$p(x) = {}^{n}C_x \, p^x \, q^{n-x}$.

$p(x) = {}^{10}C_x \, (0.65)^x \, (0.35)^{10-x}$.

Probability that atleast 7 men aged 60 will live till 70,

$P[X \geq 7] = P[X=7] + P[X=8] + P[X=9] + P[X=10]$

$= {}^{10}C_7 \, 0.65^7 \, 0.35^3 + {}^{10}C_8 \, 0.65^8 \, 0.35^2 + {}^{10}C_9 \, 0.65^9 \, 0.35$

$+ {}^{10}C_{10} \, 0.65^{10}$

$= \underline{0.5138}$

The probability that a pen manufactured by a company is defective is 0.1. If 12 such pens are selected, find the probability that i) exactly two will be defective

ii) atleast two will be defective

iii) atmost two will be defective

iv) non of them will be defective.

Probability of a pen being defective, $p = 0.1$

$q = 1 - p = 0.9$, $n = 12$

i) $p(x) = {}^{n}C_x \, p^x \, q^{n-x}$

$p(x) = {}^{12}C_x \, (0.1)^x \, (0.9)^{12-x}$

i) $P[X=2] = {}^{12}C_2\, 0.1^2\, 0.9^{10}$

$= 0.2301$

ii) $P[X \geq 2] = 1 - (P[X=0] + P[X=1])$

$= 1 - ({}^{12}C_0\, 0.1^0\, 0.9^{12} + {}^{12}C_1\, 0.1^1\, 0.9^{11})$

$= 1 - (0.659)$

$= 0.341$

iii) $P[X \leq 2] = P[0] + P[X=0] + P[X=1] + P[X=2]$

$= 0.659 + 0.2301$

$= 0.8891$

iv) $P[X=0] = {}^{12}C_0\, 0.1^0\, 0.9^{12}$

$= 0.2824$

(5) Determine the probability of getting 9 exactly twice in four throws of two dice.

The sample space, $S = \{(1,1), (1,2), \ldots, (6,6)\}$

Favorable cases for 9 are $\{(3,6), (6,3), (4,5), (5,4)\}$

∴ The probability of getting 9 is

$p = \frac{4}{36} = \frac{1}{9}$

$q = \frac{8}{9}$

$n = 4$

$p(x) = {}^nC_x\, p^x\, q^{n-x} = {}^4C_x \left(\frac{1}{9}\right)^x \left(\frac{8}{9}\right)^{4-x}$

$\therefore$ P[getting 9 exactly twice]

$= P[X = 2]$

$= {}^4C_2 \left(\frac{1}{9}\right)^2 \left(\frac{8}{9}\right)^2$

$= 0.0585$

(6) In 100 sets of 10 ter tosses of unbiased coin, in how many cases should or we expect

i) 7 heads and 3 tails

ii) atleast 7 heads

Since the coin is unbiased, the probability of getting head, $p = 1/2$ (success) $q = 1/2$ (failure)

But $n = 10$,

$$p(x) = {}^nC_x p^x q^{n-x}$$

$$p(x) = {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x}$$

$$p(x) = {}^{10}C_x \left(\frac{1}{2}\right)^{10}$$

i) P[7 heads and 3 tails in 1 set]

$= P[X = 7]$

$= {}^{10}C_7 \frac{1}{2^{10}}$

$= 0.1172$

Approximate

$\therefore$ No. of sets which gives us seven heads and three tails, $= 100 \times 0.117$

$= 11.7$

≈ 12.

ii) P[atleast 7 heads in one set]

$= P[X \geq 7]$

$= P[X=7] + P[X=8] + P[X=9] + P[X=10]$

$= {}^{10}C_7 \frac{1}{2^{10}} + {}^{10}C_8 \frac{1}{2^{10}} + {}^{10}C_9 \frac{1}{2^{10}} + {}^{10}C_{10} \frac{1}{2^{10}}$

$= \frac{1}{2^{10}} (120 + 45 + 10 + 1)$

$= 0.1718$.

$\therefore$ Approx. no. of set $= 100 \times 0.1718$

$= 17.18$

$\approx \underline{\underline{18}}$.

Poisson Distribution

The poisson distribution is another important discrete probability function where the probability of success is very small and number of trials is very large i.e. $p \to 0$ and $n \to \infty$

Since p is very small, only rare events follow this law. We sometimes come across a situation where the probability p of an event is very small but the no. of trials n is so large that the event may occur several times.

An example is

No. of people born blind in a large city every year. Individually being born blind is a rare event and its probability is very small. But in a large city, the total no. of blind births every year is significant. As a net result, we may get several blind births.

Emission and disintegration of radioactive rays. no of occurrance of a rare disease and the no of deaths by horse kicked in an army corps are some of the phenomena which follows the poisson distribution

Stating the assumptions, derive the poisson distribution as a limiting case of binomial distribution. Also find mean, variance and standard deviation.

$$\lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n = e^{-x}$$

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^{+x}$$

Since poisson distribution is limiting case of binomial distribution where $p \to 0$ and $n \to \infty$

$\therefore np = \mu$ (finite quantity)

WKT $p(x) = {}^nC_x\, p^x\, q^{n-x}$ ——— (1)

But,

WKT ${}^nC_x = \frac{n!}{x!(n-x)!}$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)(n-x)!}{x!(n-x)!}$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}$$

$$\therefore p(x) = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}\, p^x\, q^{n-x}$$

$$= \frac{n\cdot n\left(1-\frac{1}{n}\right)n\cdot\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{(x-1)}{n}\right)}{x!}\, p^x\, q^{n-x}$$

$$= \frac{n^x\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)}{x!}\, p^x\, q^{n}\cdot q^{-x}$$

$$= (np)^x \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right) q^n q^{-x}}{x!}$$

Since $n \to \infty$,

$$p(x) = \frac{\mu^x \cdot q^n q^{-x}}{x!} \quad \text{---(2)}$$

But $q^n = (1-p)^n$

$$= \left(1-\frac{\mu}{n}\right)^n$$

$$\therefore \lim_{n\to\infty} q^n = \lim_{n\to\infty} \left(1-\frac{\mu}{n}\right)^n = e^{-\mu}$$

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

Once again $q^{-x} = (1-p)^{-x}$

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

$$\therefore \lim_{p\to 0} q^{-x} = \lim_{p\to 0} \frac{1}{(1-p)^x} = \underline{1}$$

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