Module - 3

Principle of least square

Curve fitting

Let X, Y be two related variables. Let $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ be n pair of observations. ~~of two related.~~ Finding a relation of the form $\boxed{y = f(x)}$ is called curve fitting.

Here y_i are called observed values corresponding to x_i and $f(x_i)$ are called expected values or predicted values ~~corresponding~~ the difference

$y_i - f(x_i)$ is called the error or residual for ~~the~~ x_i.

[For a best fitting of the form $y = f(x)$ of n pair of observation $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, the sum of the squares of the errors]

ie $E = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \ldots [y_n - f(x_n)]^2$

must be minimum.

Fitting of straight line

Suppose two variables X and Y are linearly related.

Let $Y = aX + b$ be the best fitted straight line to the given data.

Using principle of least squares the best values of a and b are obtained by solving the normal equation:

$$\Sigma y = a\Sigma x + nb$$

$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

Q. Using principle of least squares, fit a straight line for the following data

x	1	5	10	15	20
y	8	12	16	20	26.

x	y	x^2	xy.
1	8	1	8
5	12	25	60
10	16	100	160
15	20	225	300
20	26	400	520

$\Sigma y = 82$

$\Sigma x = 51$

$n = 5$

$\Sigma xy = 1048$

$\Sigma x^2 = 751$.

Let $y = ax+b$ be the best fitted straight line

The normal equations are

$$\Sigma y = a\Sigma x + nb \quad \text{and} \quad \Sigma xy = a\Sigma x^2 + b\Sigma x$$

$$82 = 51a + 5b \quad \text{---(1)}$$

$$1048 = 751a + 51b \quad \text{---(2)}$$

Calculator $\Rightarrow$

we get $a = 0.9168$

$b = 7.048$

$\therefore$ Best fit for straight line is

$$Y = 0.9168X + 7.048$$

Fitting of parabola

Let $Y = ax^2+bx+C$ be the best fitted second degree equation for the given data. Here the normal eqn are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nC$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + C\Sigma x$$

$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + C\Sigma x^2$$

Q. Fit a second degree parabola in the form $y = ax^2 + bx + c$ to the following data

x	0	1	2	3	4	5
y	14	18	22	27	38	40

x	y	x^2	x^3	x^4	xy	x^2y
0	14	0	0	0	0	0
1	18	1	1	1	18	18
2	22	4	8	16	44	88
3	27	9	27	81	81	243
4	38	16	64	256	152	608
5	40	25	125	625	200	1000

The normal eqn are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x$$

$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

$\Sigma x = 15$

$\Sigma y = 159$

$\Sigma xy = 495$

$\Sigma x^2 y = 19347$

$\Sigma x^2 = 55$

$\Sigma x^3 = 225$

$\Sigma x^4 = 979$

$159 = 35a + 15b + 6c$ —— ①

$495 = 225a + 55b + 15c$ —— ②

$1957 = 979a + 225b + 55c$ —— ③.

By solving this

$a = 0.3214$

$b = 3.9642$

$c = 13.6428$.

Best fitted parabola =

$$y = 0.3214x^2 + 3.9642x + 13.6428$$

Note: If the parabola which is to be fitted in the form ~~$y = ax + bx^2$~~ $y = a + bx^2$ the normal equations are

$\Sigma y = na + b\Sigma x^2$ —— ①

$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^4$ —— ②.

Note:- If parabola is like $y = ax + bx^2$, then the normal eqns are

$\Sigma xy = a\Sigma x^2 + b\Sigma x^3$ —— ①

$\Sigma x^2 y = a\Sigma x^3 + b\Sigma x^4$ —— ②

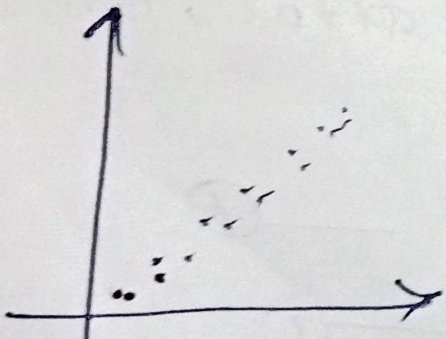
Correlation

Correlation is the method of finding any relation btw 2 variables x and y.

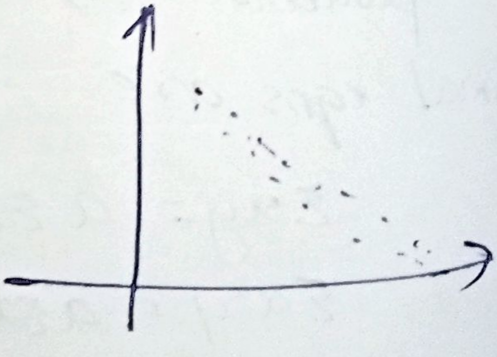
A change in one variable is accompanied by a change in other variable, we say that the two variables are correlated.

If an increment or decrement in one variable is associated with and increment or decrement in other variable, we say that variables are +vely correlated.

If an increment or decrement in one variable is associated with and decrement or increment in other variable, we say that variables are -vely correlated.



+ve



-ve

Karl Pearson's Correlation coefficient.

The degree to which two variables X and Y are correlated can be measured by computing a coefficient called Karl Pearson's correlation coefficient

It is given by

$$\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)} \cdot \sqrt{var(y)}}$$

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

$$\boxed{\rho = \frac{\Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \cdot \sqrt{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}}}$$

$\cancel{-1 \leq \rho \leq 1}$ $-1 \leq \rho \leq 1$

Note

1, Karl Pearson's correlation coefficient lies between -1 and 1.

$$-1 \leq \rho \leq 1$$

2. If $\rho = 1$, the correlation is +ve and perfect.

If $\rho = -1$, the correlation is perfect and -ve.

If $\rho = 0$, there is no correlation b/w the variables.

3. The correlation coefficient is independent of change of origin and change of scale.

ie if we take $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$

where a, b, h and k are suitable numbers, then ~~r(x,y) of~~ $\rho(x,y) =$ ~~r~~ $\rho(u,v)$

Q. Find the coefficient of correlation for the following data

x	20	22	25	26	27	23
y	31	29	32	37	35	34

$$\rho = \frac{\Sigma xy - \frac{1}{n}(\Sigma x)\cdot(\Sigma y)}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}\cdot\sqrt{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}}$$

x	y	xy	x^2	y^2
20	31	620	400	961
22	29	638	484	841
25	32	800	625	1024
26	37	962	676	1369
27	35	945	729	1225
23	34	782	529	~~4862~~ 1156
143	198	4747	3443	~~6682~~ 6576

$$\rho = \frac{4747 - \frac{1}{6}(143)(198)}{\sqrt{3445 - \frac{1}{6}(143)^2} \cdot \sqrt{6576 - \frac{1}{6}(198)^2}}$$

$$\rho = \underline{\underline{0.73}}$$

Q The following table gives the verbal reasoning test score ~~testes for~~ X and eng test score Y for each of a random sample of 8 children who took both test

child	A	B	C	D	E	F	G	H
X	112	113	110	113	112	114	109	113
Y	69	65	75	70	70	75	68	76

calculate the correlation coefficient between the scores of verbal reasoning and english.

$\therefore$

$$\rho = \frac{\Sigma xy - \frac{1}{n}\Sigma x \cdot \Sigma y}{\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \cdot \sqrt{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}}$$

$a = 10 \quad b = 60$
$h = 1$

x	y	$u = \frac{x-a}{h}$	$v = \frac{y-b}{1}$	uv	u^2	v^2
112	69	12	9	108	144	81
113	65	13	5	65	169	25
110	75	10	15	150	100	225
113	70	13	10	130	169	100
112	70	12	10	120	144	100
114	75	14	15	210	196	225
109	68	9	8	72	81	64
113	76	13	16	208	169	256
		96	88	1063	1172	1076

$$\rho = \frac{1063 - \frac{1}{8} \times 96 \times 88}{\sqrt{1172 - \frac{1}{8} \times 96^2} \cdot \sqrt{1076 - \frac{1}{8} \times 88^2}} = \frac{1063 - 1056}{\sqrt{1172 - 1152} \cdot \sqrt{1076 - 968}}$$

$$= \frac{7}{\sqrt{20} \times \sqrt{108}} = \frac{7}{\sqrt{2160}}$$

$$= \frac{7}{46.48}$$

$$= \underline{\underline{0.15}}$$

Spearman's Rank Correlation Coefficient

Instead of taking values of the variable if the ranks of the observations are considered, the correlation coefficient so obtained is called rank correlation coefficient.

Let (x_i, y_i) be the ranks of the i^{th} individual in two characteristics A and B respectively. Assume that no two observations are bracketed equal in either classification, each of the ~~obse~~ variables x and y takes the values 1, 2, 3, 4 . . . n.

Then Spearman's Rank Correlation Coefficient is given by the formula

$$\boxed{1 - \frac{6 \Sigma d^2}{n(n^2-1)}}$$

Q Calculate the rank correlation for the following data.

x	65	63	67	64	68	62	70	66	72
y	68	66	67	65	69	71	70	64	61

:- Rank correlation coefficient is given by

$$1 - \frac{6 \Sigma d^2}{n(n^2-1)} \quad \text{where } d = x_i - y_i$$

X	65	63	67	64	68	62	70	66	72
Y	68	66	67	65	69	71	70	64	61
x_i	6	8	4	7	3	9	2	5	1
y_i	4	6	5	7	3	1	2	8	9
d.	2	2	-1	0	0	8	0	-3	-8
d^2	4	4	1	0	0	64	0	9	64

$= 146$.

$$1 - \frac{6 \times 146}{9(81-1)} = 1 - \frac{6 \times 146}{720} = \cancel{0.798}$$

$$= -0.217$$

If the values of the variables repeat the Spearman's ~~Karl pearson's~~ rank correlation becomes

$$\boxed{\rho(x,y) = 1 - \frac{6\left[\Sigma d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2+1)}{12} \cdots m_k(m_k^2+1)\right]}{n(n^2-1)}}$$

where $m_1, m_2, \ldots m_k$ are the repetition of the ranks.

Q. Calculate the rank correlation for the following data

x	65	63	67	64	68	62	70	66	68	67
y	68	66	68	65	69	66	68	65	71	67
x_i	7	9	4.5	8	2.5	10	1	6	2.5	4.5
y_i	4	7.5	4	9.5	2	7.5	4	9.5	1	6
d	3	1.5	.5	-1.5	.5	2.5	-3	-3.5	1.5	-1.5
d^2	9	2.25	.25	2.25	.25	6.25	9	12.25	2.25	2.25

$\Sigma d^2 = 46$

x_i

$m_1 = 2$ for 68

$\frac{2+3}{2} = 5.5$

$m_2 = 2$ for 67

$\frac{4+5}{2} = 4.5$

y_i

for $68 \Rightarrow 3 = m_3$

$\frac{3+4+5}{3} =$

for $66 \Rightarrow 2 = m_4$

$\frac{7+8}{2} =$

for $65 \Rightarrow 2 = m_5$

$\frac{9+10}{2} =$

$$\frac{m_1(m_1^2-1)}{12} = \frac{2\times(2^2-1)}{12} = \frac{2\times3}{12} = \frac{6}{12} = 0.5$$

$$\frac{m_2(m_2^2-1)}{12} = 0.5$$

$m_4 = m_5 = 0.5$

$$m_3 = \frac{m_3(m_3^2-1)}{12} = \frac{3(3^2-1)}{12} = \frac{3\times8}{12} = \frac{24}{12} = 2$$

$n(n^2-1) = 10(100-1) = 10\times\cancel{100}\,99 = \cancel{1100}\ 990$

$$\rho = 1 - \frac{6[46 + 0.5 + 0.5 + 2 + 0.5 + 0.5]}{990}$$

$$= 1 - \frac{300}{990} = \cancel{0.697} \; \underline{0.697}$$

Regression

Suppose x and y are two related variables the prediction of the value of 1 variable for a particular value of other variable is a part of regression analysis.

~~the~~ Suppose x and y are linearly correlated then the line which is used to predict the value of y when x is given is called the regression line of y on x and is given by

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

where b_{yx} is called the regression coefficient of y on x which is given by ~~b_{yx}~~ =

$$\boxed{b_{yx} = \frac{\text{cov}(x,y)}{\text{Var}(x^2) \text{ or } \sigma_x^2} = \frac{\Sigma xy - \frac{1}{n}\Sigma x \cdot \Sigma y}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}}$$

Similarly the line which is used to predict the value of x when y is given is called the regression line of x on y which is given by

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

where b_{xy} is called the regression coefficient of x on y and is given by ~~b_{xy} =~~

$$b_{xy} = \frac{\text{cov}(x,y)}{\sigma_y^2} = \frac{\Sigma xy - \frac{1}{n}\Sigma x \cdot \Sigma y}{\Sigma y^2 - \frac{1}{n}(\Sigma y)^2}$$

Note

1. The two regression lines always passes through the point $(\bar{x}, \bar{y})$.
2. We have $b_{yx} = \frac{\text{cov}(x,y)}{\sigma_x^2}$ & $b_{xy} = \frac{\text{cov}(x,y)}{\sigma_y^2}$

Since Var is +ve

b_{xy} and b_{yx} have the same sign.

3. $$b_{yx} \times b_{xy} = \frac{(\text{cov}(x,y))^2}{\sigma_x^2 \cdot \sigma_y^2} = \left(\frac{\text{cov}(x,y)}{\sigma_x \sigma_y}\right)^2 = \rho^2$$

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

we have;

4. $b_{yx} = \frac{cov(x,y)}{\sigma x^2} = \frac{\sigma y}{\sigma x} \cdot \frac{cov(x,y)}{\sigma y \, \sigma x}$

$$= \frac{\sigma y}{\sigma x} \rho$$

$$b_{xy} = \frac{\sigma x}{\sigma y} \rho.$$

Q. In a partially destroyed laboratory record of analysis of correlation of following data the following results are only eligible,

Var $(x) = 9$, regression eqns $8x - 10y + 66 = 0$ and $40x - 18y = 240$.

what are

i) The mean values of x and y

ii) Correlation coefficient between x and y.

iii) the standard deviation of y.

$\because$; $\sigma x^2 = 9$.

$8x - 10y + 66 = 0$

$40x - 18y = 240$

mean

we know the two regression lines always passes through the two regression lines.

$\therefore 8\bar{x} - 10\bar{y} = -66$ —— ①.

$40\bar{x} - 18\bar{y} = 240$ —— ②.

$\bar{x} =$

$\bar{y} =$ calculator

ii we have $\rho = \pm\sqrt{b_{yx} \times b_{xy}}$

Assume that

$8x - 10y + 66 = 0$ is the regression line of x on y.

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

$8x = 10y + 66$

$x = \frac{10}{8}y + \frac{66}{8}$.

$\frac{10}{8} = \frac{5}{4}$

$b_{xy} = \frac{10}{6}$.

Then the regression line of y on x

is

$40x - 18y = 240$

$y = \frac{40}{18}x - \frac{240}{18}$.

$b_{yx} = \frac{40}{18}$.

Hence $\rho = \pm\sqrt{b_{yx} \times b_{xy}} = \pm\sqrt{\frac{40}{18} \times \frac{10}{6}} = \pm\frac{5}{3}$

$-1 < \rho < 1$ $\quad 5/3 > 1.$

Hence our assumption is wrong

$8x - 10y + 66 = 0$ is the regression line of

$10y = 8x + 66$

$y = \frac{8x}{10} + \frac{66}{10}.$

Hence $b_{yx} = \frac{8}{10}$

$40x - 18y = 240$

$40x = 18y + 200$

$x = \frac{18}{40}y + \frac{200}{40}.$

$b_{xy} = \frac{18}{40}.$

$\rho = \sqrt{\frac{8}{10} \times \frac{18}{40}} = \sqrt{\frac{9}{25}} = \frac{3}{5}$

iii) $\sigma y = ?$

$\sigma x^2 = 9$

$b_{yx} = \frac{\sigma y}{\sigma x} \rho$

$\frac{8}{10} = \frac{\sigma y}{\cancel{3}} \times \frac{\cancel{3}}{5}$

$\sigma y = \frac{5 \times 8}{10} = \frac{40}{10} = 4$

Q. The two regression lines are $3x+2y=26$ and $6x+y=31$. Calculate the two regression coefficients and the correlation coefficient.

:- $3x+2y=26$

$6x+y=31$

$3x+2y=26$

$3x=-2y+26$

$x=-\frac{2}{3}y+26$

$6x+y=31$

$y=-6x+31$

Our assumption wrong.

So $3x+2y=26$

$2y=-3x+26$ y on x

$y=-\frac{3}{2}x+26$ $-\frac{3}{2}=b_{yx}$

and

$6x+y=31$ x on y.

$6x=-y+31$

$x=-\frac{1}{6}y+31$ $-\frac{1}{6}\ b_{yx}$

$r=\sqrt{\frac{-3}{2}\times\frac{-1}{6}}=\sqrt{\frac{3}{12}}=-0.5$ because -vely correlated

note=1

$\therefore -1 < r < 1$ our assumption is right.

Hence $\rho = -0.5$

$b_{yx} = -\frac{3}{2}$ and $b_{xy} = -1/6$

Q) Do the equations $y = 2x + 3$ and $x + 3y = 4$ represents the regression of a set of pair of random data. Explain your reason?

:-

$y = 2x + 3$

$x = -3y + 4$

$b_{yx} = 2$, $b_{xy} = -3$ is not possible because sign are different

We know that the two regression coefficient must have the same sign. Hence the given two lines do not represent regression line.

Q The grades of a class of 9 students on a mid term report (x) and on the final examination (y) are as follows.

x	77	50	71	72	81	94	96	99	67
y	82	66	78	54	47	85	99	99	~~65~~ 68

Estimate the final examination grade of a student who received a grade of 85 on the mid term report.

:- To predict the value of y when $x = 85$ we use the regression line of y on x.

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

$$b_{yx} = \frac{\Sigma xy - \frac{1}{n}\Sigma x \, \Sigma y}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}$$

x	77	50	71	72	81	94	96	99	67
y	82	66	78	54	47	85	99	99	68
xy	6314	3300	5538	3888	3807	7990	9504	9801	4556
x^2	5929	2500	3041	3184	6561	8836	9216	9801	4456

$\Sigma x = 707$

$\Sigma y = 698$

$\Sigma x^2 = 57557$

$\Sigma xy = 54698$

$$b_{yx} = \frac{54698 - \frac{1}{9} 707 \cdot 698}{57557 - \frac{1}{9}(707)^2}$$

$$= 0.777$$

$\bar{x} = \frac{\Sigma x}{n} = \frac{707}{9} = 78.5$

$\bar{y} = \frac{\Sigma y}{n} = \frac{678}{9} = 75.33$

$y - 75.33 = 0.777(x - 78.5)$ $x = 85$

$\cancel{y = 0.7776x - 60.9945}$

$y = 75.33 + 0.777(85 - 78.5)$

$y = 80.3505$

Q. The following data are available $\bar{x} = 970$, $\bar{y} = 18$, $\sigma x = 38$, $\sigma y = 2$ and $r = 0.6$

find the value of x when $y = 20$.

∴ To predict the value of x we use regression line of x on y.

Given by

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

$b_{xy} = \frac{\sigma x}{\sigma y} r = \frac{38}{2} \times 0.6 = 11.4$

$\therefore x - 970 = 11.4\,(y - 18)$ when $y = 20$

$x = 970 + 11.4\,(20 - 18)$

$= 970 + 11.4 \cdot 2$

$= 992.8$

Two Dimensional Random Variable

Let X and Y be two discrete random variables associated with a random experiment. If there exist a function p such that

$$p(x,y) = P(X = x, ~~and~~ Y = y),$$

then p is called the joint probability mass function of x and y.

If $p(x,y)$ is a joint pmf then $\sum_x \sum_y p(x,y) = 1$.

Marginal probability Distributions

The fn $p_x(x)$ defined by

$P_x(x) = \sum_y p(x,y)$ is called the marginal probability distribution of x.

Similarly the fn $P_Y(y) = \Sigma_x p(x,y)$ is called the maginal probability distribution of y.

Independent Random Variables

Two random variable X, Y are said to be independent.

$p(x,y) = P_x(x) \cdot P_y(y)$

Q. The joint probability mass function of a random variable x and y is given by

$p(x,y) = k(x+2y)$, $x = 0,1,2$ $y = 0,1,2,3$,

find the following.

i) The value of k

ii) The maginal distribution of x and y.

iii) Probability than x is $<= 1$

iv) " $y <= 2$

v) $x <= 1, y <= 2$.

vi) Are x and y independent

$x \backslash y$	0	1	2	3	$P_x(x)$
0	0	2K	4k	6K	12K
1	K	3k	5k	7K	16K
2	2k	4k	6k	8K	20k
$P_y(y)$	3k	9k	15k	21k	48k.

$\frac{1}{48}(x+2y)$

i) Total probability = 1

So

$$12K + 16K + 20K + 3K + 9K + 15K + 21K = 1$$

$$48K = 1$$

$\because$ p is a joint pmf.

$$\sum_x \sum_y p(x,y) = 1.$$

$$48K = 1$$

$$K = 1/48$$

ii) Marginal distribution of x is given by.

x	0	1	2
$P_x(x)$	$\frac{12}{48}$	$\frac{16}{48}$	$\frac{20}{48}$.

Similarly the marginal distribution of y is given by.

y	0	1	2	3
$P_y(y)$	$\frac{3}{48}$	$\frac{9}{48}$	$\frac{15}{48}$	$\frac{21}{48}$

iii $p(x<=1)$

$= p(x=0) + p(x=1)$

$= \frac{12}{48} + \frac{16}{48} = \frac{28}{48}$

iv) $p(Y<=2)$

$p(y=0) + p(y=1) + p(y=2)$

$\frac{3}{48} + \frac{9}{48} + \frac{15}{48} = \frac{27}{48}$

v) $p(x<=1, y<=2)$

$x = 0, 1$ y can be $0, 1, 2$

$p(1,0) + p(1,1) + p(1,2)$

$p(0,0) + p(0,1) + p(0,2)$

$= k + 3k + 5k + 0 + 2k + 4k = 15k$

$= 15 \times \frac{1}{48} = \frac{15}{48}$

vi) $p(1,2) = 5k = \frac{5}{48}$

$P_x(1) = \frac{16}{48}$ $\quad p(1,2) \neq P_x(1)\, P_y(2)$

$P_y(2) = \frac{15}{48}$

Another method

To prove independency

$$P_x(x) = \Sigma_y p(x,y)$$

$$= \Sigma_y \frac{1}{48}(x+2y)$$

$$= \frac{1}{48}\left[(x+2\times0)+(x+2\times1)+(x+2\times2)+(x+2\times3)\right]$$

$$= \frac{1}{4}\left[4x+12\right] \quad x = 0,1,2.$$

$$P_y(y) = \Sigma_x p(x,y)$$

$$= \Sigma_x \frac{1}{48}(x+2y)$$

$$= \frac{1}{48}[0+2y] + \frac{1}{48}[1+2y] + \frac{1}{48}[2+2y]$$

$$= \frac{1}{48}\left[3+4y\right] \quad y = 0,1,2,3$$

$$P_x P_y = \frac{1}{48}(4x+12) \times \frac{1}{48}(3+6y) \neq p(x,y).$$