IDHAYA COLLEGE FOR WOMEN, KUMBAKONAM.

DEPARTMENT OF MATHEMATICS

CLASS	**: I BSc., IT**
SUBJECT NAME	**: NUMERICAL ANALYSIS & STATISTICS**
SUBJECT CODE	**: 16SACMA2**
SEMESTER	**: II**
UNIT	**: V (CORRELATION & REGRESSION)**
FACULTY NAME	**: DR. C. KAYALVIZHI**

UNIT V

1. Correlation:

Correlation is a statistical technique which shows the relationship b/w two or more variables. The relationship b/w two series when measured quantitatively is known as correlation.

2. The types of correlation:

3 types:

1. Positive and Negative Correlation
2. Linear and Non-Linear Correlation
3. Simple and Multiple Correlation

3. Properties of correlation coefficient:

1. The correlation coefficient is unaffected by change of origin of reference and scale of reference.
2. The value of correlation coefficient lied between -1 and +1. ($-1 \le r \le +1$)

4. Karl Pearson's Co-efficients of Correlation:

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

5. The merits of coefficient of correlation:

1. It is superior to other methods. It is used calculated directly from the numerical values of each and every pair. Even if one value changes, r changes.
2. The population correlation coefficient can be estimated from the sample value.
3. The significance of the sample correlation coefficient can be tested.
4. Many formulae are available for its calculation. From certain scientific calculators, the value of r can be read out.

6. Spearman's Rank Correlation:

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right].$$

6. Repeated Rank Correlation:

$$\rho = 1 - \left[\frac{6\sum d^2 + \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \dots\dots}{N(N^2-1)}\right]$$

8. Merits and Demerits of Rank Correlation:

Merits:

1. It is easy to calculate.
2. It is simple to understand.
3. It can be applied to any type of data. Qualitative or quantitative. Hence correlation with qualitative data such as honesty, beauty can be found.
4. This is most suitable in case there are two attributes.

Demerits:

1. It is only an approximately calculated measure as actual values are not used for calculations.
2. For large samples it is not convenient method.
3. Combined r of different series cannot be obtained as in case of mean and standard deviation.
4. It cannot be treated further algebraically.

9. Calculate the correlation co-efficient from following data:

X:	10	6	9	10	12	13	11	9
Y:	9	4	6	9	11	13	8	4

Solution:

Coefficient of correlation: $r = \dfrac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$

X	Y	XY	X^2	Y^2
10	9	90	100	81
6	4	24	36	16
9	6	54	81	36
10	9	90	100	81
12	11	132	144	121
13	13	169	169	169
11	8	88	121	64
9	4	36	81	16
$\sum X = 80$	$\sum Y = 64$	$\sum XY = 683$	$\sum X^2 = 832$	$\sum Y^2 = 584$

Here, N = 8

$$r = \frac{8(683) - (80)(64)}{\sqrt{8(832) - (80)^2}\sqrt{8(584) - (64)^2}}$$

$$= \frac{5464 - 5120}{\sqrt{6656 - 6400}\sqrt{4672 - 4096}}$$

$$= \frac{344}{\sqrt{256}\sqrt{576}}$$

$$= \frac{344}{(16)(24)}$$

$$= \frac{344}{384}$$

Therefore $r = 0.896$.

10. Find Karl Pearson's coefficient of correlation from following data:

X:	100	101	102	102	100	99	97	98	96	95
Y:	98	99	99	97	95	92	95	94	90	91

Solution:

Karl Pearson's coefficient of correlation:

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\ \sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

X	Y	XY	X^2	Y^2
100	98	9800	10000	9604
101	99	9999	10201	9801
102	99	10098	10404	9801
102	97	9894	10404	9409
100	95	9500	10000	9025
99	92	9108	9801	8464
97	95	9215	9409	9025
98	94	9212	9604	8836
96	90	8640	9216	8100
95	91	8645	9025	8281
$\sum X = 990$	$\sum Y = 950$	$\sum XY =$ 94111	$\sum X^2 =$ 98064	$\sum Y^2 =$ 90346

Here, N = 10

$$r = \frac{10(94111) - (990)\,(950)}{\sqrt{10(98064) - (990)^2}\ \sqrt{10(90346) - (950)^2}}$$

$$= \frac{941110 - 940500}{\sqrt{980640 - 980100}\ \sqrt{903460 - 902500}}$$

$$= \frac{610}{\sqrt{540}\ \sqrt{960}}$$

$$= \frac{610}{(23.2379)\,(30.9838)}$$

$$= \frac{610}{719.9984}$$

Therefore $r = 0.8472$.

11. Find out the rank correlation co-efficient from following data:

Marks in Maths:	95	55	63	42	72	88	65	49	54	50
Marks in Statistics:	63	55	47	60	48	42	69	70	51	45

Solution:

Rank correlation: $\rho = 1 - \left[\frac{6\sum d^2}{N(N^2 - 1)} \right]$.

Maths: X	Statistics: Y	Rank X	Rank Y	$d = X - Y$	d^2
95	63	1	3	-2	4
55	55	6	5	1	1
63	47	5	8	-3	9
42	60	10	4	6	36
72	48	3	7	-4	16
88	42	2	10	-8	64
65	69	4	2	2	2
49	70	9	1	8	64
54	51	7	6	1	1
50	45	8	9	1	1

Here N = 10, $\sum d^2 = 200$

$$\rho = 1 - \left[\frac{6(200)}{10(100^2 - 1)} \right]$$

$$= 1 - \left[\frac{1200}{10(100 - 1)} \right]$$

$$= 1 - \left[\frac{1200}{10(99)} \right]$$

$$= 1 - \left[\frac{1200}{990} \right]$$

$$= 1 - 1.2121$$

Thererfore $\rho = 0.2121$.

12. Find out the rank correlations co-efficient from following data:

x:	48	33	40	9	16	16	65	24	16	57
y:	13	13	24	6	15	4	20	9	6	19

Solution:

Repeated Rank correlations: $\rho = 1 - \left[\dfrac{6\sum d^2 + \dfrac{m_1(m_1^2-1)}{12} + \dfrac{m_2(m_2^2-1)}{12}}{N(N^2-1)} \right]$

X	Y	Rank X	Rank Y	$d = X - Y$	d^2
48	13	3	5.5	-2.5	6.25
33	13	5	5.5	-0.5	0.25
40	24	4	1	3	9
9	6	10	8.5	1.5	2.25
16	15	8	4	4	16
16	4	8	10	-2	4
65	20	1	2	-1	1
24	9	6	7	-1	1
16	6	8	8.5	-0.5	0.25
57	19	2	3	-1	1

-

Here N = 10, $\sum d^2 = 41$, $m_1 = 3$, $m_2 = 2$

$$\rho = 1 - \left[\frac{6(41) + \dfrac{3\left[(3)^2-1\right]}{12} + \dfrac{2\left[(2)^2-1\right]}{12}}{10(10^2-1)} \right]$$

$$= 1 - \left[\frac{256 + \dfrac{3[9-1]}{12} + \dfrac{2[4-1]}{12}}{10(100-1)} \right]$$

$$= 1 - \left[\frac{256 + \dfrac{3(8)}{12} + \dfrac{2(3)}{12}}{10(99)} \right]$$

$$= 1 - \left[\frac{256 + \dfrac{24}{12} + \dfrac{6}{12}}{990} \right]$$

$$= 1 - \left[\frac{256 + 2 + \dfrac{1}{2}}{990} \right]$$

$$= 1 - \left[\frac{\dfrac{512+4+1}{2}}{990} \right]$$

$$= 1 - \left[\frac{\frac{517}{2}}{990} \right]$$

$$= 1 - \left[\frac{258.5}{990} \right]$$

$$= 1 - 0.2611$$

Therefore $\rho = 0.7389$.

13. Regression:

The term regression analysis refers to the methods by which estimates are made of the value of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process.

14. The uses of Regression:

1. Prediction of unknown value is made possible by regression analysis.
2. Nature of relationship: The regression device is useful in establishing and estimating the nature of the relationship b/w two variables.
3. Policy formulation.
4. Touch stone of hypothesis.

15. Write regression equations:

(1) The regression equation of Y on X:

$$(Y - \overline{Y}) = b_{YX}(\mathrm{X} - \overline{X})$$

Where b_{YX} is called the regression co-efficient of Y on X.

$$b_{YX} = r\frac{\sigma_X}{\sigma_Y}$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2}$$

(2) The regression equation of X on Y:

$$(\mathrm{X} - \overline{X}) = b_{XY}(Y - \overline{Y})$$

Where b_{XY} is called the regression co-efficient of X on Y.

$$b_{XY} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2}$$

16. Properties of Rregression Co-efficients:

1. The correlation coefficient is the geometric mean between the two regression coefficients.

$$b_{YX} = r\frac{\sigma_X}{\sigma_Y}$$

$$b_{XY} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{YX}\,.\,b_{XY} = r^2$$

$$r = \pm\sqrt{b_{YX}\,.\,b_{XY}}$$

r is positive if the two regression coefficient are positive.

r is negative if the two regression coefficient are negative.

2. If one regression coefficient is greater than one, the other regression coefficient is less than one

$$r^2 \le 1$$
$$b_{YX}\,.\,b_{XY} \le 1$$
$$b_{YX} \le \frac{1}{b_{XY}}$$

Therefore, if one regression coefficient is greater than one, the other regression coefficient is less than one.

17. When $\overline{X} = 40, \overline{Y} = 60, \sigma_X = 10, \sigma_Y = 15$ and $r = 0.7$ give the regression equation of Y on X.

Solution:

Given: $\overline{X} = 40, \overline{Y} = 60, \sigma_X = 10, \sigma_Y = 15$ and $r = 0.7$.

To find: The regression equation of Y on X:

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X})$$

Where $b_{YX} = r\frac{\sigma_X}{\sigma_Y}$

$$= 0.7\frac{10}{15}$$
$$= 0.7(0.6666)$$
$$= 0.4666$$

$$(Y - 60) = 0.4666(X - 40)$$
$$(Y - 60) = 0.4666X - 18.664$$
$$Y = 0.4666X - 18.664 + 60$$
$$Y = 0.4666X - 41.336$$