**For the 2Y3S 29th Tuition Batch**

Slide Mentioned Questions Solved

# Statistics (2105) Kawsar Jahan Ma'am

Collection – Nurun Nabi Mahmud, A&IS 26th

## Questions of Slide

-

Dispersion & Skewness

Correlation Analysis

Regression Analysis

Sampling

Probability

## Dispersion & Skewness

1. In an attempt to estimate potential future demand the national motor company did a study asking married couples how many cars the average energy minded family should own in 2000. for each couple national averaged the husbands and wives responses to get the overall couple response. The answers were

- No of cars 0 0.5 1.0 1.5 2.0 2.5
- Frequency 2 14 23 7 4 2

i) Calculate the variance and SD

ii) Since the distribution is roughly bell shaped, how many of the observations should theoretically fall between .5 and 1.5? Between 0 and 2? How many actually do fall in those intervals?

Solution: i)

| No. of Cars xi | Frequency fi | fixi | $xi^2$ | $fixi^2$ |
|---|---|---|---|---|
| 0 | 2 | 0 | 0 | 0 |
| 0.5 | 14 | 7 | 0.25 | 3.5 |
| 1.0 | 23 | 23 | 1.0 | 23 |
| 1.5 | 7 | 10.5 | 2.25 | 15.75 |
| 2.0 | 4 | 8 | 4 | 16 |
| 2.5 | 2 | 5 | 6.25 | 12.5 |
| | n=52 | $\sum$fixi=53.5 | | $\sum fixi^2$=70.75 |

$\therefore$ Mean, $\bar{x}$= $\frac{\sum fixi}{n} = \frac{53.5}{52}$ = 1.03

$\therefore$ Variance, $\partial^2$= $\frac{\sum fixi^2}{n-1} - \frac{n\bar{x}^2}{n-1} = \frac{70.75}{52-1} - \frac{52(1.03)^2}{52-1}$ = 0.306

$\therefore$ SD, $\partial = \sqrt{0.306} = 0.553$

ii) Since the distribution is roughly bell shaped, we can say that it's almost a symmetrical distribution.
According to the question the observation falls between 0.5 and 1.5 that is $\pm$1SD from the mean. We know that 68% observation lies within $\pm$1SD. So, theoretically (52×68%) or 35.36≅36 observations lie between 0.5 and 1.5.

Again, the observations between 0 to 2 is the range of $\pm$2SD and almost 95% observations lie within this range. So, theoretically (52×95%) or 49.4≅50 observations lie between 0 to 2.
And actually the number of observations falls between 0.5 and 1.5 are,
Range (0.5 – 1.5) = 14+23+7 = 44
Falls between 0 to 2,
Range (0 – 2) = 2+14+23+7+4 =50

2. The following data give the no of passengers travelling by Boeing 747 from one city to another in one week.

320, 290, 265, 300, 270, 315

Calculate the mean and standard deviation.

Solution: i)

| xi | $xi^2$ |
|---|---|
| 320 | 102400 |
| 290 | 84100 |
| 265 | 70225 |
| 300 | 90000 |
| 270 | 72900 |
| 315 | 99225 |
| $\sum$xi= 1760 | $\sum xi^2$= 518850 |

$\therefore$Mean $\bar{x}= \frac{\sum xi}{N} = \frac{17060}{6} = 293.33$
=

$\therefore$SD, $\partial = \sqrt{\frac{\sum xi^2}{N} - \bar{x}^2} = \sqrt{\frac{518850}{6} - 293.33^2} = \sqrt{432.511} = 20.79$

3. From an analysis of monthly wages paid to workers in two companies Beximco and square, we have the following results:

| | Beximco | Square |
|---|---|---|
| No of workers | 500 | 400 |
| Average monthly wage | 1200 | 1100 |
| S.D of wages | 10 | 12 |

i) Compute the combined mean and standard deviation.

ii) Which company shows greater variability in the distribution of wage?

Solution:

i) Given that,

$\bar{X}_B = 1200$ $\bar{X}_S = 1100$

$N_B = 500$ $N_S = 400$

$\partial_B = 10$ $\partial_S = 12$

$\therefore$ Combined mean, $\bar{X}_{BS} = \frac{N_B X_B + N_S X_S}{N_B + N_S} = \frac{(500 \times 1200) + (400 \times 1100)}{500 + 400} = 1155.56$

$\therefore$ Combined SD, $\partial_{BS} = \sqrt{\frac{N_B \partial_B^2 + N_S \partial_S^2 + N_B d_B^2 + N_S d_S^2}{N_b + N_S}}$

$= \sqrt{\frac{500 \times 10^2 + 400 \times 12^2 + 500 \times 44.44^2 + 400 \times 55.56^2}{500 + 400}}$

$= \sqrt{2588.69138} = 50.879$

Here, $d_B = |\bar{X}_B - \bar{X}_{BS}| = |1200 - 1155.56| = 44.44$

$d_S = |\bar{X}_S - \bar{X}_{BS}| = |1100 - 1155.56| = 55.56$

So, the combined mean is 1155.56 and the combined SD is 50.879

ii) To determine which company shows greater variability in distribution of wage, we to compare coefficient of variation,

$CV_{(Beximco)} = \frac{\partial}{\bar{X}_B} \times 100 = \frac{10}{1200} \times 100 = 0.833$

$CV_{(Square)} = \frac{\partial}{\bar{X}_S} \times 100 = \frac{12}{1100} \times 100 = 1.09$

Since the CV is high in Square Company, so Square Company shows greater variability in the distribution of wage.

4. A and B are two factory. The average weekly wages and SD of wages of the workers are given below:

| Factory | Average weekly wages | SD | No of workers |
|---|---|---|---|
| A | 1560 | 90 | 200 |
| B | 1580 | 70 | 160 |

Calculate the combined mean and SD of wages of the whole workers in the two factory.

Solution:

Given that,

$\bar{X}_A = 1560$ $\bar{X}_B = 1580$

$\delta_A = 90$ $\delta_B = 70$

$N_A = 200$ $N_B = 160$

$\therefore$ Combined mean, $X_{AB} = \frac{N_A\bar{X}_A + N_B\bar{X}_B}{N_A + N_B} = \frac{200\times1560 + 160\times1580}{200+160} = 1568.89$

$\therefore$ Combined SD, $\partial_{AB} = \sqrt{\frac{N_A\partial_A^2 + N_B\partial_B^2 + N_A d_A^2 + N_B d_B^2}{N_A + N_B}}$

$$= \sqrt{\frac{200\times90^2 + 160\times70^2 + 200\times8.89^2 + 160\times11.11^2}{200+160}}$$

$= \sqrt{6776.54} = 82.31$

Here, $d_A = |\bar{X}_A - \bar{X}_{AB}| = |1560 - 1568.89| = 8.89$

$d_B = |\bar{X}_A - \bar{X}_{AB}| = |1580 - 1568.89| = 11.11$

5. Lives of two models of refrigerators obtained from a survey are presented as follows:

| Life in hours : | 2-4 | 4-6 | 6-8 | 8-10 | 10-12 | 12-14 |
|---|---|---|---|---|---|---|
| Model X : | 1 | 7 | 12 | 10 | 4 | 1 |
| Model Y : | 2 | 7 | 9 | 12 | 2 | 1 |

i. Which model has more average life?
ii. Compute the coefficient of skewness of each model and comment on the results.
iii. Which model is more variable?
iv. How can you compare the kurtosis of the distributions?

Solution: i.

| Life in hours | Model X | | | | | Model Y | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | xi | fi | fixi | Cfi | $fixi^2$ | yi | fi | fiyi | Cfi | $fiyi^2$ |
| 2-4 | 3 | 1 | 3 | 1 | 9 | 3 | 2 | 6 | 2 | 18 |
| 4-6 | 5 | 7 | 35 | 8 | 175 | 5 | 7 | 35 | 9 | 175 |
| 6-8 | 7 | 12 | 84 | 20 | 588 | 7 | 9 | 63 | 18 | 441 |
| 8-10 | 9 | 10 | 90 | 30 | 810 | 9 | 12 | 108 | 30 | 972 |
| 10-12 | 11 | 4 | 44 | 34 | 484 | 11 | 2 | 22 | 32 | 242 |
| 12-14 | 13 | 1 | 13 | 35 | 169 | 13 | 1 | 13 | 33 | 169 |
| | | N=35 | $\sum fixi=269$ | | $\sum fixi^2=2235$ | | N=33 | $\sum fiyi=247$ | | $\sum fiyi^2=2017$ |

Model X, $\bar{x} = \frac{\sum fixi}{N} = \frac{269}{35} = 7.69$

Model Y, $\bar{y}= \frac{\sum fiyi}{N} = \frac{247}{33} = 7.48$

Therefore, the mean of Model X is greater than Model Y. Hence, Model X has more average life.

ii. We know that,

$SK_p= \frac{\bar{x}-Mode}{\partial}$ Here, $\bar{x}= 7.69$

$\therefore$ Mode= $L + \frac{\Delta_1}{\Delta_1+\Delta_2} \times i$

$= 6 + \frac{5}{5+2} \times 2$

$= 7.43$

L= 6

$\Delta_1$= 12 – 7

$\Delta_2$= 12 – 10 = 2

i = 2

$\therefore \partial = \sqrt{\frac{\sum xi^2}{N} - \bar{x}^2} = \sqrt{\frac{2235}{35} - 7.69^2} = 2.17$

$SK_p= \frac{7.69-7.43}{2.17} = 0.12$

Since, $SK_p > 0$, so it is positively skewed.

For Model Y,

$\bar{y}= 7.48$

$\therefore$ Mode= $L + \frac{\Delta_1}{\Delta_1+\Delta_2} \times i$

$= 6 + \frac{5}{5+2} \times 2$

$= 8.46$

L=8

$\Delta_1$= 12 – 9 = 3

$\Delta_2$= 12 – 2 = 10

i = 2

$$\partial = \sqrt{\frac{\sum yi^2}{N} - \bar{y}^2} = \sqrt{\frac{2017}{33} - 7.48^2} = 2.27$$

$SK_p = \frac{7.48 - 8.46}{2.27} = -0.43$

Since, $SK_p < 0$, so the distribution is negatively skewd.

iii. Here,

For Model X, $\bar{x} = 7.48$ $\delta = 2.27$

$CV_x = \frac{\partial}{\bar{x}} \times 100 = \frac{2.17}{7.69} \times 100 = 28.22\%$

For Model Y, $\bar{y} = 7.48$ $\delta = 2.27$

$CV_y = \frac{\partial}{\bar{x}} \times 100 = \frac{2.17}{7.48} \times 100 = 30.35\%$

Since, the coefficient of variation of Model Y is greater, Model Y is more variable.

iv. We can compare the kurtosis of the distributions by calculating $\beta_2$.

If $\beta_2 > 3$; the curve is more peaked and it is called lepto kurtic.

If $\beta_2 < 3$; the curve is less peaked and it is called platy kurtic .

If $\beta_2 = 3$; the curve is neither more or less peaked and it is called meso kurtic.

Here,

| Life in hours | Model X | | | | | Model Y | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | xi | fi | $(xi-\bar{x})$ | $fi(xi-\bar{x})^2$ | $fi(xi-\bar{x})^4$ | yi | fi | $(yi-\bar{y})$ | $fi(yi-\bar{y})^2$ | $fi(yi-\bar{y})^4$ |
| 2-4 | 3 | 1 | -4.69 | 21.996 | 483.828 | 3 | 2 | -4.48 | 40.141 | 805.642 |
| 4-6 | 5 | 7 | -2.69 | 50.653 | 366.528 | 5 | 7 | -2.48 | 43.053 | 264.792 |
| 6-8 | 7 | 12 | -0.69 | 5.713 | 2.72 | 7 | 9 | -0.48 | 2.074 | 0.479 |
| 8-10 | 9 | 10 | 1.31 | 17.161 | 29.45 | 9 | 12 | 1.52 | 27.725 | 64.055 |
| 10-12 | 11 | 4 | 3.31 | 43.824 | 480.145 | 11 | 2 | 3.52 | 24.781 | 307.44 |
| 12-14 | 13 | 1 | 5.31 | 28.196 | 795.02 | 13 | 1 | 5.52 | 30.47 | 928.445 |
| | | N= 35 | | $\sum fi(xi-\bar{x})^2$ =167.54 | $\sum fi(xi-\bar{x})^4$ =2157.69 | | N= 33 | | $\sum fi(yi-\bar{y})^2$ =168.243 | $\sum fi(yi-\bar{y})^4$ =2370.456 |

$\bar{x}$=7.69 and $\bar{y}$=7.48

$$\mu_{2x}= \frac{\sum fi(xi-\bar{x})^2}{N} = \frac{167.54}{35} = 4.787$$

$$\mu_{4x}= \frac{\sum fi(xi-\bar{x})^2}{N} = \frac{2157.69}{35} = 61.648$$

$$\therefore \beta_{2x}= \frac{\mu_4}{\mu_2^2} = \frac{61.648}{4.787^2} = 2.69$$

Since, $\beta_{2x}>3$, so the kurtosis of Model X is more peaked which is called leptokurtic.

$$\mu_{2y}= \frac{\sum fi(yi-\bar{y})^2}{N} = \frac{168.243}{33} = 5.098$$

$$\mu_{4y}= \frac{\sum fi(yi-\bar{y})^4}{N} = \frac{2370.456}{33} = 71.832$$

$$\beta_{2y}= \frac{\mu_4}{\mu_2^2} = \frac{71.832}{5.098^2} = 2.76$$

Since $\beta_{2y}<3$; So the kurtosis of Model Y is less peaked which is called platykurtic.

6. FundInfo provides information to its subscribers to enable them to evaluate the performance of mutual funds they are considering as potential investment vehicles. A recent survey of funds whose stated investment goal was growth and income produced the following data on total annual rate of return over the past six years:

| Annual Return | 11.0-11.9 | 12.0-12.9 | 13.0-13.9 | 14.0-14.9 | 15.0-15.9 | 16.0-10.9 |
|---|---|---|---|---|---|---|
| Frequency | 2 | 2 | 8 | 10 | 8 | 5 |

i. Calculate the mean, variance and standard deviation of the annual rate of return for this 35 sample funds.

ii. State Chebyshev's theorem. According to Chebyshev's theorem, between what values should at least 75 percent of the sample observations fall? What percentage of the observation actually does fall in that interval?

iii. Because the distribution is roughly bell shaped, between what would you expect to find 68% of the observations? Under empirical rules find the percentage of the observations actually does fall in that interval?

Solution: i.

| Annual rate | xi | fi | fixi | $xi^2$ | $fixi^2$ |
|---|---|---|---|---|---|
| 11-11.9 | 11.45 | 2 | 22.9 | 131.1025 | 262.21 |
| 12-12.9 | 12.45 | 2 | 24.9 | 155.1025 | 310.01 |
| 13-13.9 | 13.45 | 8 | 107.6 | 180.0025 | 1447.22 |
| 14-14.9 | 14.45 | 10 | 144.5 | 208.8025 | 2088.03 |
| 15-15.9 | 15.45 | 8 | 123.6 | 238.7025 | 1909.62 |
| 16-16.9 | 16.45 | 5 | 82.25 | 270.6025 | 1353.01 |
| | | n=35 | $\sum$fixi=505.75 | | $\sum fixi^2$=7370 |

$\therefore \bar{x} = \frac{\sum fixi}{n} = \frac{505.75}{35} = 14.45$

$\therefore$Variance, $\partial^2 = \frac{\sum fixi^2}{n-1} - \frac{n\bar{x}^2}{n-1} = \frac{7370}{35-1} - \frac{35 \times 14.45^2}{35-1} = 1.824$

$\therefore$ SD, $\partial = \sqrt{1.824} = 1.351$

ii. Chebyshev's theorem: According to Chebyshev's theorem, no matter what the shape of the distribution at least 75% observations will fall within $\pm 2\delta$ form the mean and 89% of total observations will fall within $\pm 3\delta$ from the mean.

In this question, According Chebyshev's theorem 75% of observations will within, that $\bar{x} \pm 2\delta$ range is 14.45 – 2(1.351) to 14.45 + 2(1.351) range which is 11.748 to 17.152 .

Here, in 11.748 the frequency would be 1. So, the total frequency from 11.748 to 17.152 is = (1+2+8+10+8+5) = 34

$\therefore$ The actual percentage is = $\frac{34}{35} \times 100 = 97.14\%$

iii. Empirical rule: According to empirical rule for a symmetrical or bell shaped distribution 68.27% of the total observations fall within $\pm 2\delta$ from the mean.

Here, $\bar{x} \pm 1\delta$

=> $14.45 \pm 1.351$

So, the range is 13.09 to 15.8.

Here, in 13.09 or 13.1 the frequency would be $\left(\frac{8}{10} \times 9\right)$ or $7.2 \cong 7$ and in 15.8, the frequency would be $\left(\frac{8}{10} \times 9\right)$ or $7.2 \cong 7$

So, the total frequency from 13.1 to 15.8 is (7+10+7) or 24.

$\therefore$ The actual percentage is = $\frac{24}{35} \times 100 = 68.57\%$

## Correlation Analysis

1. Find the Pearsonian correlation coefficient from the following series of marks obtained by 10 students in class test in Mathematics ($X$) and in Statistics ($Y$).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X: | 45 | 70 | 65 | 30 | 90 | 40 | 50 | 75 | 85 | 60 |
| Y: | 35 | 90 | 70 | 40 | 95 | 40 | 60 | 80 | 80 | 90 |

Also calculate the probable error.

Solution:

| Students | X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|---|
| 1 | 45 | 2025 | 35 | 1225 | 1575 |
| 2 | 70 | 4900 | 90 | 8100 | 6300 |
| 3 | 65 | 4225 | 70 | 4900 | 4550 |
| 4 | 30 | 900 | 40 | 1600 | 1200 |
| 5 | 90 | 8100 | 95 | 9025 | 8550 |
| 6 | 40 | 1600 | 40 | 1600 | 1600 |
| 7 | 50 | 2500 | 60 | 3600 | 3000 |
| 8 | 75 | 5625 | 80 | 6400 | 6000 |
| 9 | 85 | 7225 | 80 | 6400 | 6800 |
| 10 | 60 | 3600 | 50 | 2500 | 3000 |
| Total | $\sum X/N=61$ | $\Sigma X^2=40700$ | $\sum Y/N=64$ | $\Sigma Y^2=45350$ | $\Sigma XY=42575$ |

$\therefore$ Coefficient of Correlation $r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$

$$= \frac{42575 - (10 \times 61 \times 64)}{\sqrt{(40700 - 10 \times 61^2)(45350 - 10 \times 64^2)}}$$

$$= -0.9$$

$\therefore$ Probable Error (PE)$r = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$

$$= 0.6745 \times \frac{1 - (-0.9)^2}{\sqrt{10}}$$

$$= 0.04056$$

2. The data on price and quantity purchased relating to a commodity for 5 months given below.

| Months: | January | February | March | April | May |
|---|---|---|---|---|---|
| Prices: | 10 | 10 | 11 | 12 | 12 |
| Quantity: | 5 | 6 | 4 | 3 | 3 |

Find the Pearsonian correlation coefficient between prices and quantity and comment on its sign and magnitude.

Solution: Let's assume that the price is X and quantity is Y.

| Month | X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|---|
| 1 | 10 | 100 | 5 | 25 | 50 |
| 2 | 10 | 100 | 6 | 36 | 60 |
| 3 | 11 | 121 | 4 | 16 | 44 |
| 4 | 12 | 144 | 3 | 9 | 36 |
| 5 | 12 | 144 | 3 | 9 | 36 |
| | $\Sigma$ X/N =11 | $\Sigma X^2$ =609 | $\Sigma$ Y/N = 4.2 | $\Sigma Y^2$ =95 | $\Sigma$ XY= 226 |

$\therefore$ Coefficient of correlation, $r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$

$$= \frac{226 - (5 \times 11 \times 4.2)}{\sqrt{(609 - 5 \times 11^2)(95 - 5 \times 4.2^2)}}$$

$$= -0.96$$

Here, negative sign of r indicate negative correlation & magnitude of the correlation is 0.90.

$\therefore$ Probable Error (PE)$r = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$

$$= 0.6745 \times \frac{1 - (-0.96)^2}{\sqrt{5}}$$

$$= 0.0236$$

3.

| Sales representatives | No. of sales calls | No. of Copiers sold |
|---|---|---|
| Tom Keller | 20 | 30 |
| Jeff Hall | 40 | 60 |
| Brain Virost | 20 | 40 |
| Grey Fish | 30 | 60 |
| Susan Welch | 10 | 30 |
| Carlos Ramirez | 10 | 40 |
| Rich Niles | 20 | 40 |
| Mike Kiel | 20 | 50 |
| Mark Reynolds | 20 | 30 |
| Soni Jones | 30 | 70 |

Using the Copier Sales of America data compute,
i. The correlation coefficient and coefficient of determination. Interpret the result.
ii. Also find the probable error of the data and explain the significance of the data.

Solution: Let's assume that No. of sales calls denoted by X and No. of copiers sold denoted by Y.

| S.R. | X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|---|
| 1 | 20 | 400 | 30 | 900 | 600 |
| 2 | 40 | 1600 | 60 | 3600 | 2400 |
| 3 | 20 | 400 | 40 | 1600 | 800 |
| 4 | 30 | 900 | 60 | 3600 | 1800 |
| 5 | 10 | 100 | 30 | 900 | 300 |
| 6 | 10 | 100 | 40 | 1600 | 400 |
| 7 | 20 | 400 | 40 | 1600 | 800 |
| 8 | 20 | 400 | 50 | 2500 | 1000 |
| 9 | 20 | 400 | 30 | 900 | 600 |
| 10 | 30 | 900 | 70 | 4900 | 2100 |
| Total | $\Sigma X/N = 22$ | $\Sigma X^2 = 5600$ | $\Sigma Y/N = 45$ | $\Sigma Y^2 = 22100$ | $\Sigma XY = 10800$ |

$\therefore$ Coefficient of correlation, $r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$

$$= \frac{10800 - (10 \times 22 \times 45)}{\sqrt{(5600 - 10 \times 22^2)(22100 - 10 \times 45^2)}}$$

$$= 0.76$$

$\therefore$ Coefficient of determination, $r^2 = (0.76)^2 = 0.58$

Here, positive sign of r indicate positive correlation & magnitude of correlation is 0.70 Coefficient of determination ($r^2$) is the square of the coefficient of correlation. Here, 54% of variation in the dependent variable (Y) has been explained by independent variable (X).

ii. Here, Probable Errorb(PE)r = $0.6745 \times \frac{1-r^2}{\sqrt{N}} = 0.6745 \times \frac{1-(-0.76)^2}{\sqrt{10}} = 0.09$

Since, r = 0.79 > 6PE(r) =(6×0.09)= 0.54 the correlation is significant.

4. Pran juice is studying the effect of its latest advertising campaign. People chosen at random are called and asked how many juice they had bought in the past week and how many advertisement they have either seen in the past week.

| Months | ; Jan | Feb | Mar | Apr | May | Jun | Jul | Aug |
|---|---|---|---|---|---|---|---|---|
| Number of Ads : | 3 | 7 | 4 | 2 | 0 | 4 | 1 | 2 |
| Juice purchased: | 11 | 18 | 9 | 4 | 7 | 6 | 3 | 8 |

i) Allowing two months' time lag calculate coefficient of correlation
ii) Calculate the sample coefficient of determination. Interpret the result.

i) Allowing 2 months' time lag calculating coefficient of correlation (No. of ads in January reflects on juice purchase in March)
Here, X represent No. of ads & Y represent No. of Juice purchased

| Months | X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|---|
| Jan | 3 | 9 | 9 | 81 | 27 |
| Feb | 7 | 4 | 49 | 16 | 28 |
| Mar | 4 | 7 | 16 | 49 | 28 |
| Apr | 2 | 6 | 4 | 36 | 12 |
| May | 0 | 3 | 0 | 9 | 0 |
| Jun | 4 | 8 | 16 | 64 | 32 |
| | $\sum X/N = 3.33$ | $\sum Y = 6.167$ | $\sum X^2 = 94$ | $\sum Y^2 = 255$ | $\sum XY = 127$ |

∴ Coefficient of Correlation, $r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$

$= \frac{127-(6\times3.33\times6.167)}{\sqrt{(94-6\times3.33^2)(255-6\times6.167^2)}} = 0.135$

∴ There is a positive correlation between No. of ads and No. of juice purchased.

ii. Here, Coefficient of correlation, r = 0.135

∴ Coefficient of determination, $r^2 = (0.135)^2 = 0.0182$

Here, 1.82% of the variation in the dependent variable (Y) has been explained by the independent variable (X).

5. 12 entries in a painting competition were ranked by two judges as shown below.

| Entry | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Judge I | 5 | 2 | 3 | 4 | 1 | 6 | 8 | 7 | 10 | 9 |
| Judge II | 4 | 5 | 2 | 1 | 6 | 7 | 10 | 9 | 3 | 8 |

Find the coefficient of rank correlation.

Solution:

| Rank of Judge I | Rank of Judge II | d= Judge I – Judge II | $d^2$ |
|---|---|---|---|
| 5 | 4 | 1 | 1 |
| 2 | 5 | −3 | 9 |
| 3 | 2 | 1 | 1 |
| 4 | 1 | 3 | 9 |
| 1 | 6 | −5 | 25 |
| 6 | 7 | −1 | 1 |
| 8 | 10 | −2 | 4 |
| 7 | 9 | −2 | 4 |
| 10 | 3 | 7 | 49 |
| 9 | 8 | 1 | 1 |
| | | $\sum d = 0$ | $\sum d^2 = 104$ |

Now,
Coefficient of rank correlation will be,

$$P = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6\times 104}{10(10^2-1)} = 1 - 0.6303 = 0.369$$

Hence, there is a medium degree of positive correlation.

6. Calculate Karl Perason's coefficient f correlation between expenditure on advertising (X) and sales(Y) from the data given below:

| X | 39 | 65 | 62 | 90 | 82 | 75 | 25 | 98 | 36 | 78 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 47 | 53 | 58 | 86 | 62 | 68 | 60 | 91 | 51 | 84 |

Solution:

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 39 | 47 | 1521 | 2209 | 1833 |
| 65 | 53 | 4225 | 2809 | 3445 |
| 62 | 58 | 3844 | 3364 | 3596 |
| 90 | 86 | 8100 | 7396 | 7740 |
| 82 | 62 | 6724 | 3844 | 5084 |
| 75 | 68 | 5625 | 4624 | 5100 |
| 25 | 60 | 625 | 3600 | 1500 |
| 98 | 91 | 9604 | 8281 | 8918 |
| 36 | 51 | 1296 | 2601 | 1836 |
| 78 | 84 | 6084 | 7056 | 6552 |
| $\sum X/N = 65$ | $\sum Y/N = 61$ | $\sum X^2 = 47648$ | $\sum Y^2 = 45784$ | $\sum XY = 45604$ |

$\therefore$ Coefficient of correlation, $r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$

$$= \frac{45604 - (10 \times 65 \times 61)}{\sqrt{(47648 - 10 \times 65^2)(45784 - 10 \times 61^2)}}$$

$$= 0.0001286$$

$\therefore$ There is a low degree of positive correlation.

## Regression Analysis

1.

| X | 1 | 3 | 4 | 8 | 9 | 4 | 14 |
|---|---|---|---|---|---|---|---|
| Y | 1 | 2 | 4 | 5 | 7 | 8 | 9 |

Hence obtain:

a) The regression coefficient of X on Y and of Y on X.
b) $\bar{X}$ and $\bar{Y}$
c) Coefficient of correlation between X and Y.
d) What is the estimated value of Y when X=10 and of X when Y=5.

Solution: (a) & (b)

| X | Y | $(X-\bar{X})$ | $(Y-\bar{Y})$ | $(X-\bar{X})(Y-\bar{Y})$ | $(X-\bar{X})^2$ | $(Y-\bar{Y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | -6.143 | -4.143 | 25.450 | 37.736 | 17.164 |
| 3 | 2 | -4143 | -3.143 | 13.021 | 17.164 | 9.878 |
| 4 | 4 | -3.143 | -1.143 | 3.592 | 9.879 | 1.306 |
| 8 | 5 | 0.857 | -0.143 | -0.123 | 0.734 | 0.020 |
| 9 | 7 | 1.857 | 1.857 | 3.448 | 3.448 | 3.448 |
| 11 | 8 | 3.857 | 2.857 | 11.019 | 14.876 | 8.162 |
| 14 | 9 | 6.857 | 3.857 | 26.447 | 47.018 | 14.876 |
| $\sum X= 50$ | $\sum Y=36$ | | | $\sum(X-\bar{X})(Y-\bar{Y})$ =82.854 | $\sum(X-\bar{X})^2$ =130.855 | $\sum(Y-\bar{Y})^2$ = 54.854 |

Here,

$\bar{X}=\frac{\sum X}{N} = \frac{50}{7} = 7.143$

$\bar{Y}=\frac{\sum Y}{N} = \frac{36}{7} = 5.143$

$\therefore$ Regression coefficient of Y on X, $b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{82.854}{130.855} = 0.633$

$\therefore$ Regression coefficient of X on Y, $b_{xy}= \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{82.854}{54.854} = 1.51$

c) ∴ Coefficient of Correlation, $r = \pm \sqrt{(b_{yx} \times b_{xy})}$

$= \pm\sqrt{(0.633 \times 1.51)}$

$= 0.978$

d) When X is independent,

$\hat{Y} = a + bx$

$= 0.62 + (0.633)x$

$= 0.62 + (0.633)10$

$= 6.95$

$a = \bar{Y} - b\bar{X}$

$= 5.143 - (0.633)7.143$

$= 0.62$

(When x=10)

So, When X=10, value of Y= 6.95

When Y is independent,

$\hat{X} = a + by$

$= -0.62 + (1.51)y$

$= -0.62 + (1.51)5$

$= 6.93$

$a = \bar{X} - b\bar{Y}$

$= 7.143 - (1.51)5.143$

$= -0.62$

(When y=5)

So, When Y=5, value of X=6.93

2. What is regression coefficient? Show that $r^2 = bxy + byx$ where symbols have their usual meanings. What can you say about the angle between the regression line when?
i. r=0
ii. r=1
iii. r increases from 0 to 1

Solution:
In the regression equation the quantity b is called the "Regression Coefficient". There are two regression equation. Regression equation of X on Y and regression equation of Y on X.
Regression equation of X on Y,
This denoted by bxy. It measures the amount of changes in x corresponding to a unit change in y. the regression coefficient of x on y is given by,

$$b_{xy} = r\frac{\partial_x}{\partial_y}$$

When deviation is taken from the means of X and Y, the regression coefficient is obtained by,

$$b_{xy}= \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}$$

Regression equation of y on x,
This denoted by byx. It measures the amount of changes in y corresponding to a unit change in y. the regression coefficient of y on x is given by,

$$b_{yx}= r\frac{\partial_y}{\partial_x}$$

When deviation is taken from the means of X and Y, the regression coefficient is obtained by,

$$b_{yx}= \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

We know that,

$b_{xy}= r\frac{\partial_x}{\partial_y}$ & $b_{yx}= r\frac{\partial_y}{\partial_x}$

$\therefore b_{xy}\times b_{yx}= r\frac{\partial_x}{\partial_y} \times r\frac{\partial_y}{\partial_x}$

$=> b_{xy}\times b_{yx} = r^2$ (shown)

As there are two regression equation, there are two regression lines. Angle between two regression lines depends on correlation of two variables X and Y. The further two lines to each other, the lesser is the degree of correlation and nearer the two regression lines to each other, higher is the degree of correlation. So, when the value of correlation,

i. r=0, angle between two regression line is right angle or 90 degree
ii. r=1, angle between two regression line is 0 degree as the two regression lines coincide.
iii. r increases from 0 to 1, the regression lines will gradually come closer. So, the angle between two regression lines will be greater than 0 degree but less or equal to 90 degree.

3. Obtain the equation of lines of regression of Y on X from the following data.

| X | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
|---|---|---|---|---|---|---|---|
| Y | 5.27 | 5.68 | 6.25 | 7.21 | 8.02 | 8.71 | 8.42 |

Estimate the most probable value of Y when X=40.

Solution:

| X | Y | $(X-\bar{X})$ | $(Y-\bar{Y})$ | $(X-\bar{X})(Y-\bar{Y})$ | $(X-\bar{X})^2$ |
|---|---|---|---|---|---|
| 12 | 5.27 | -18 | -1.81 | 32.58 | 324 |
| 18 | 5.68 | -12 | -1.40 | 16.80 | 144 |
| 24 | 6.25 | -6 | -0.83 | 4.98 | 36 |
| 30 | 7.21 | 0 | 0.13 | 0 | 0 |
| 36 | 8.02 | 6 | 0.94 | 5.64 | 36 |
| 42 | 8.71 | 12 | 1.63 | 19.56 | 144 |
| 48 | 8.42 | 18 | 1.34 | 24.12 | 324 |
| $\sum X=210$ | $\sum Y=49.56$ | | | $\sum(X-\bar{X})(Y-\bar{Y})=103.68$ | $\sum(X-\bar{X})^2=1008$ |

$\bar{X}=\frac{\sum X}{N}=\frac{210}{7}=30$

$\bar{Y}=\frac{\sum Y}{N}=\frac{49.56}{7}=7.08$

$b_{yx}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}=\frac{103.58}{1008}=0.1029$

So, the regression equation will be, $(Y-\bar{Y})=b_{xy}(X-\bar{X})$

$$\Rightarrow Y-7.08=0.1029(X-30)$$
$$\Rightarrow Y=0.1029X-3.087+7.08$$
$$\Rightarrow Y=0.1029X+3.993 \therefore$$

Probable value of Y will be when X=40,

$$\therefore Y=0.1029\times 40+3.993$$
$$=8.109$$

## Sampling

The following data are the results of a market survey with a sample size 50.Regarding the acceptability of a new product which the company wants to launch. The scores of the respondent on the appropriate scales are as follows:

40 45 41 45 45 30 39 8 48 12

26 23 24 26 29 8 40 41 42 36

27 35 18 25 35 40 42 43 44 09

28 27 32 28 27 25 26 38 37 25

29 35 32 28 40 41 43 44 45 40

i) Draw a simple random sample of size 10 using the above data, then find their mean and variance.

ii) Draw a stratified random sample of 15 from the above population. Stratified the population

iii) Draw a systematic sample of 6 from the population.

Solution: i)
Here, N= 50 and n=10
Simple random sampling table

| Serial no. | Population | Serial no. | Population | Serial no. | Population |
|---|---|---|---|---|---|
| 1 | 40 | 18 | 41 | 35 | 27 |
| 2 | 45 | 19 | 42 | 36 | 25 |
| 3 | 41 | 20 | 36 | 37 | 26 |
| 4 | 45 | 21 | 27 | 38 | 38 |
| 5 | 45 | 22 | 35 | 39 | 37 |
| 6 | 30 | 23 | 18 | 40 | 25 |
| 7 | 39 | 24 | 25 | 41 | 29 |
| 8 | 8 | 25 | 35 | 42 | 35 |
| 9 | 48 | 26 | 40 | 43 | 32 |
| 10 | 12 | 27 | 42 | 44 | 28 |
| 11 | 26 | 28 | 43 | 45 | 40 |
| 12 | 23 | 29 | 44 | 46 | 41 |
| 13 | 24 | 30 | 9 | 47 | 43 |
| 14 | 26 | 31 | 28 | 48 | 44 |
| 15 | 29 | 32 | 27 | 49 | 45 |
| 16 | 8 | 33 | 32 | 50 | 40 |
| 17 | 40 | 34 | 28 | | |

∴selecting sample by using simple random number table.

| Random no. | Sample(xi) | $xi^2$ |
|---|---|---|
| 15 | 29 | 841 |
| 09 | 48 | 2304 |
| 41 | 29 | 841 |
| 35 | 27 | 729 |
| 20 | 36 | 1296 |
| 45 | 40 | 1600 |
| 38 | 38 | 1444 |
| 01 | 40 | 1600 |
| 39 | 37 | 1369 |
| 29 | 44 | 1936 |
| | $\sum xi = 368$ | $\sum xi^2 = 13960$ |

∴ Mean, $\bar{x} = \frac{\sum X}{N} = \frac{368}{10} = 36.8$

Variance of samples, $s^2 = \frac{\sum xi^2}{n-1} - \frac{n\bar{x}^2}{n-1} = \frac{13960}{10-1} - \frac{10 \times 36.8^2}{10-1} = 46.4$

ii) Here, N=50 and n= 15

There are two strata here, a) x<30 & b) x≥30

| Strata 1 (x<30) | | | | Strata 2 (x≥30) | | | |
|---|---|---|---|---|---|---|---|
| Serial no. | Population | Serial no. | Population | Serial no. | Population | Serial no. | Population |
| 1 | 08 | 16 | 27 | 1 | 40 | 16 | 42 |
| 2 | 12 | 17 | 25 | 2 | 45 | 17 | 43 |
| 3 | 26 | 18 | 26 | 3 | 41 | 18 | 44 |
| 4 | 23 | 19 | 25 | 4 | 45 | 19 | 32 |
| 5 | 24 | 20 | 29 | 5 | 45 | 20 | 38 |
| 6 | 26 | 21 | 28 | 6 | 30 | 21 | 37 |
| 7 | 29 | | | 7 | 39 | 22 | 35 |
| 8 | 08 | | | 8 | 48 | 23 | 32 |
| 9 | 27 | | | 9 | 40 | 24 | 40 |
| 10 | 18 | | | 10 | 41 | 25 | 41 |
| 11 | 25 | | | 11 | 42 | 26 | 43 |
| 12 | 09 | | | 12 | 36 | 27 | 44 |
| 13 | 28 | | | 13 | 35 | 28 | 45 |
| 14 | 27 | | | 14 | 35 | 29 | 40 |
| 15 | 28 | | | 15 | 40 | | |

Here, $N = N_1 + N_2$ $K = \frac{n}{N} = \frac{15}{50} = 0.3$

$= 21 + 29 = 50$

$n = 15$

$n_1 = K \times N_1$ $n_2 = K \times N_2$

$= 0.3 \times 21 = 6.3 \cong 6$ $= 0.3 \times 29 = 8.7 \cong 9$

| Strata 1(x<30) | | Strata 2(x≥30) | |
|---|---|---|---|
| Random no. | Sample | Random no. | Sample |
| 15 | 28 | 07 | 39 |
| 09 | 27 | 01 | 40 |
| 20 | 29 | 04 | 45 |
| 01 | 08 | 15 | 40 |
| 12 | 09 | 08 | 48 |
| 03 | 26 | 10 | 40 |
| | | 17 | 43 |
| | | 16 | 42 |
| | | 23 | 32 |

iii)

Here, N= 50 n=6

$\therefore K = \frac{N}{n} = \frac{50}{6} = 8.3 \cong 8$

Let, the 1st unit, i = 5 ($i \leq K$)

| Random no. | Sample |
|---|---|
| i=5 | 45 |
| i+k=13 | 24 |
| i+k=21 | 27 |
| i+k=29 | 44 |
| i+k=37 | 26 |
| i+k=45 | 40 |

## Probability

1. The human resource department of a Bank Asia has the following records of its 200 employees:

| Age | BBA degree | MBA degree | Total |
|---|---|---|---|
| Less than 30 | 90 | 10 | 100 |
| 30 to 40 | 20 | 30 | 50 |
| Greater than 40 | 40 | 10 | 50 |
| Total | 150 | 50 | 200 |

If an employee is selected At random, find the probability that, She/he

i) Has only BBA degree,

ii) Has only BBA degree given that she /he is 30 to 40,

iii) Is under 30 given that has only a MBA degree.

Solution: Let's assume that,

A = an employee under 30 year old
B = an employee between 30 to 40 year old
C = an employee with BBA degree
D = an employee with MBA degree

i) The probability of an employee with BBA degree only, P(C) = $\frac{150}{200}$ = 0.75

ii) The probability of an employee with BBA degree given that 30 to 40 old,

$$P(C/B) = \frac{P(C \cap B)}{P(B)} = \frac{\frac{20}{200}}{\frac{50}{200}} = 0.40$$

iii) The probability of an employee under 30 given that with only MBA degree,

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{\frac{10}{200}}{\frac{50}{200}} = 0.20$$

2. A manufacturer firm produces steel pipes in three plants with daily production volumes of 500, 1000 and 2000 units respectively. According to past experience it is known that the fractions of defective output produced by the three plants are respectively at random 0.005, .008 and 0.010. If a pipe is selected from a day's total production and found to be defective, find out, (i) from which plant for this defective pipe, the probability is highest?

Solution:
Total Production = 500+1000+2000= 3500
$A_1$= Production units of 1st plant
$A_2$= Production units of 2nd plant
$A_3$= Production units of 3rd plant
D = a defective item.

∴The probability of selecting a unite from 1st, 2nd, 3rd plants are,

$P(A_1)= \frac{500}{3500} = 0.1428$ $P(A_3)= \frac{2000}{3500} = 0.5714$

$P(A_2)= \frac{1000}{3500} = 0.2857$

Now the joint probabilities are,

$P(A_1 \cap D)= P(A_1)\times P(D/ A_1)= 0.1428\times 0.005= 0.0007$

$P(A_2 \cap D)= P(A_2)\times P(D/A_2)= 0.2857\times 0.008= 0.0028$

$P(A_3 \cap D)= P(A_3)\times P(D/A_3)= 0.5714\times 0.01= 0.057$

Using Bay's theorem to determine required probability,

$$P(A_1/D)= \frac{P(A_1 \cap D)}{P(A_2 \cap D)+P(A_3 \cap D)+P(A_3 \cap D)} = \frac{0.0007}{0.0007+0.0028+0.057} = \frac{0.0007}{0.05998} = 0.0116$$

Similarly,

$P(A_2/D)= \frac{0.0028}{0.05998} = 0.038$ $P(A_3/D)= \frac{0.057}{0.05998} = 0.95$

Since, the $P(A_3/D)$ has the higest probability among three. So, defective unite will be most likely from 3rd plant.

3. A company is planning its company picnic. The only thing that will cancel the picnic is a thunderstorm. The weather service has predicted dry conditions, with probability 0.2, moist conditions with probability 0.45, and wet conditions with probability 0.35. If the probability of a thunderstorm given dry conditions is 0.3, given moist conditions is 0.6, and given wet conditions is 0.8. What is the probability of a thunderstorm? If we know the picnic was indeed canceled, what is the probability moist conditions were in effect?

Let's assume that event of thunderstorm is T, dry condition is D, moist condition is M and wet condition is W.

Here,

$P(D)= 0.2$ , $P(M)= 0.45$ , $P(W)= 0.35$

The conditional probabilities are, $P(T/D)= 0.3$, $P(T/M)= 0.6$, $P(T/W)= 0.8$

Now, the probability of thunderstorm is sum of $P(D\cap T)$, $P(M\cap T)$ and $P(W\cap T)$.

Here,

$P(D\cap T)= P(D/T).P(D)= 0.3\times 0.2= 0.06$

$P(M\cap T)= P(M/T).P(M)= 0.6\times 0.45= 0.27$

$P(W\cap T)= P(W/T).P(W)= 0.8\times 0.35=0.28$

So, $P(T)= P(D\cap T) + P(M\cap T) + P(W\cap T)$

$= 0.06 + 0.27 + 0.28$

$= 0.61$

By using Bay's thorium we can determine the probability of moist condition if picnic is canceled.

$$P(T/M)= \frac{P(M\cap T)}{P(D\cap T)+P(M\cap T)+P(W\cap D)}$$

$$= \frac{0.27}{0.61} = 0.443$$