Practical No:03

Correlation & Regression.

Q.1. $X = 65, 45, 50, 60, 40$ } Given.
$Y = 70, 35, 60, 50, 40$

To find $x = ?$

→

	X	Y	X^2	Y^2	XY
	65	70	4225	4900	4550
	45	35	2025	1225	1575
	50	60	2500	3600	3000
	60	50	3600	2500	3000
	40	40	1600	1600	1600
Σ	260	235	13950	13825	13725

$n = 5 \quad \bar{X} = \frac{260}{5} = 52 \quad \bar{Y} = \frac{255}{5} = 51$

$\bar{X}\bar{Y} = \frac{13725}{5} = 2652$

$COV(X, Y) = \frac{\Sigma X_i Y_i}{n} - \bar{X}\bar{Y}$

$= \frac{13725}{5} - 2652$

$\boxed{COV(X, Y) = 93}$

$\sigma X = \sqrt{\frac{\Sigma X_i^2}{n} - \bar{X}^2} = \sqrt{\frac{13950 - 2704}{5}}$

$$\boxed{\sigma x = 9.27}$$

$$\sigma y = \sqrt{\frac{\sigma y_i^2 - \bar{y}^2}{n}} = \sqrt{\frac{13825 - 2601}{5}}$$

$$\boxed{\sigma y = 12.8}.$$

$$\therefore \varepsilon = \frac{cov(x,y)}{\sigma x \, \sigma y} = \frac{93}{9.27 \times 12.8}$$

$$\boxed{\varepsilon = 0.78}.$$

Interpretation as $\varepsilon = 0.78$ there is high +ve correlation bet^n marks in static paper I & II.

Q. 2.

→ Given: $n = 50 \quad \Sigma x = 20 \quad \Sigma y = 25$

$\Sigma x^2 = 85, \quad \Sigma y^2 = 90, \quad \Sigma xy = 75$

To find eq^n of line of Regression.

$y - \bar{y} = b_{yx}(x - \bar{x})$ - (eq^n of line)

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

$$cov(x,y) = \frac{\Sigma xy}{n} - (\bar{x} - \bar{y})$$

$$= \frac{75}{50} - (0.4 \times 0.5)$$

$$= 1.5 - 0.2$$

$$cov(x,y) = 1.3$$

$$\sigma^2 x = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$= \frac{85}{50} - (0.4)^2$$

$$= 1.54$$

Now, $b_{yx} = \frac{Cov(x,y)}{\sigma^2 x}$

$$= \frac{1.3}{1.54}$$

$$= 0.8441$$

∴ The eqⁿ of line will be

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

$$y - 0.5 = 0.8441(x - 0.4) + 0.5$$

$$\therefore y = 0.8441x - 0.1624$$

e.3

→ let us assume that,

$3x + 12y = 19$ is the eqⁿ of line of Regression x on y &

$9x + 3y = 46$ is the eqⁿ of line of Regression y on x.

$\therefore 3x + 12y = 19$

$x = -4y + 19/3$

$\therefore b_{xy} = -4.$

&

$\therefore 9x + 3y = 46$

$y = -3x + 46/3$.

$b_{yx} = -3$.

We know,

$r^2 = b_{xy} \cdot b_{yx}$.

$= -4 \cdot -3$.

$r^2 = 12$

$\therefore r^2$ lies betⁿ 0 & 1

Hence, our assumptⁿ is wrong.

Now, we assume that,

$3x + 12y = 19$ is the eqⁿ of line of regression y on x &

$9x + 3y = 46$ is the eqⁿ of line of regression x on y.

$\therefore 3x + 12y = 19$

$\therefore y = \frac{-3x + 19}{12}$

$\therefore b_{yx} = -3/12 = -1/4$ &

$9x + 3y = 46$.

$\therefore x = \frac{-3y + 46}{9}$

$\therefore b_{xy} = -3/9$

$b_{xy} = -1/3$.

We knows,

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

$= -1/3 \cdot -1/4$.

$r^2 = 1/12 = 0.0833$

$r = 0.2887$ — $\therefore b_{xy}$ & b_{yx} are -ve

The sign of regression coeffient will always be the same, as the sign of the correlation coeffient

11. since both the line of regression posses through the pt. $(\bar{x}, \bar{y})$

The consider

$$3\bar{x} + 12\bar{y} = 19 \quad - (1)$$
$$9\bar{x} + 3\bar{y} = 46 \quad - (2).$$

multiply eqn (1) by 3, we get

$$9\bar{x} + 36\bar{y} = 54 \quad - (3).$$

eqn (3) − eqn (1)

$$\begin{array}{r} 9\bar{x} + 36\bar{y} = 57 \\ 9\bar{x} + 3\bar{y} = 46 \\ - \quad - \\ \hline 33\bar{y} = 11 \end{array}$$

$\therefore \bar{y} = 0.333.$

Substituting $\bar{y} = 1/3$ in eqn (3)

$$3\bar{x} + 12 \times 1/3 = 19$$
$$\bar{x} = 5$$

$\therefore$ mean of $x = \bar{x} = 5$ & mean of $y - \bar{y} = 1/3 = 0.333.$

Practical No: 04
Theory of probability.

Q.1.

→ $\Omega = \{HH, HT, TH, TT\}$

$n = 4.$

(i) A occurance of both the heads

$A = \{HH\}$

$n(A) = 1.$

$P(A) = \frac{n(A)}{n} = 1/4.$

∴ P(occurance of both of heads) = 1/4.

(ii) B = occurance of single head.

$B = \{HT, TH\}$

$n(B) = 2$

$P(B) = \frac{n(B)}{n} = 1/2$

∴ P(occurance of single head) = 1/2

(iii) C = occurance of atleast one head.

$C = \{HH, HT, TH\}$

$n(C) = 3$

$P(C) = \frac{n(C)}{n} = 3/4.$

∴ P(occurance of atleast one head) = 3/4.

Q.2.

→ The total no. of face in which be selected from 52 cards given by

$$^{52}C_4 \text{ r. } e^n C_r$$

Hence Ω written $n = {}^{52}C_4$.

(I) occurance of two red & two black cards.

Two red cards can be drawn in $^{26}C_2$ ways

Two black cards can be drawn in $^{26}C_2$ ways

∴ both the red & black cards are to be selected the No. of favourable cases.

To the event A,

$$m = {}^{26}C_2 \quad {}^{26}C_2$$

$$P(A) = \frac{m}{n}$$

$$= \frac{{}^{26}C_2 \quad {}^{26}C_2}{{}^{52}C_4}$$

$$P(A) = 0.3901$$

(II) B = All cards are of different suits.

$$m = {}^{13}C_4 \cdot {}^{13}C_4 \quad {}^{13}C_4 \cdot {}^{13}C_4$$

$$P(B) = \frac{m}{n}$$

$$= \frac{{}^{13}C_4 \cdot {}^{13}C_4 \cdot {}^{13}C_4 \cdot {}^{13}C_4}{{}^{52}C_4}$$

$$\therefore P(B) = 0.1054$$

(III) C: All cards of same units.

let event C = occurance of all cards of same suits.

i.e. all cards are either diamond, Hearts club spade, using the addition principle of containing.

$\therefore m = {}^{13}C_4 + {}^{13}C_4 + {}^{13}C_4 + {}^{13}C_4$

$= 4\ {}^{13}C_4$.

$P(C) = \frac{m}{n}$.

$= \frac{4\ {}^{13}C_4}{{}^{52}C_4}$.

$= \frac{4\ 13!_0}{\frac{(13-4)!_0 4!_0}{52!_0}}$

$\frac{}{(52-4)!_0 4!_0}$.

$P(C) = 0.01056$.

(IV) D = one king cards.

$m = 4C_1 \cdot 48C_3$

$P(D) = \frac{m}{n}$.

$= \frac{4C_1 \cdot 48C_3}{{}^{52}C_4}$.

$\therefore P(D) = 0.255$.

(i) P(two are red & two are black cards) = 0.3901

(ii) P(all cards are of diff. suits) = 0.1054

(iii) P(all cards are of same suits) = 0.01056.

(iv) P(one is king) = 0.2555.

Q. 3.

→ A commodity consist of 4 members, 3 engineer's, 4 economist, 2 statical & 1 CA.

$$n = \frac{10!}{(10-4)!\,4!} = 210.$$

(i) If event a consist of one engineer, economist, statistican, 1 CA hence, the probability of event A become.

$$A = {}^3C_1 \cdot {}^4C_1 \cdot {}^2C_1 \cdot {}^1C_1$$

$$n(A) = {}^{24}C_1$$

$$P(A) = \frac{{}^{24}C_1}{{}^{4}C_{10}} = \frac{24}{210} = \frac{4}{35}$$

$$\therefore P(A) = 0.1142$$

(ii) The event of B consists of CA & atleast 1 engineer.

a. 1 CA x 1 Eng x 2 other

$$\therefore {}^1C_1 \cdot {}^3C_1 \cdot {}^6C_2$$

$$= 45.$$

b. 1 CA x 2 Eng x 1 other.

$\therefore {}^1C_1 \times {}^3C_2 \times {}^6C_1$

$= 18$

c. 1 CA x 3 Eng x non-other.

$\therefore {}^1C_1 \cdot {}^3C_3 \cdot {}^6C_0$

$= 1$

$\therefore 45 + 18 + 1 = 64.$

$\therefore P(B) = \frac{64}{210} = 0.3048.$

(i) P (each of four categories is included in the commodity) = 0.1142

(ii) P (commodity consists CA & atleast one engineer) = 0.3048

Practical No: 05

Testing of hypothesis.

Q. 1.

→ Hypothesis to be tested:

$H_0 : \mu = 1000$ V/s $H_1 : \mu \neq 1000$.

Given: $\alpha = 5\% = 0.05$, $n = 64$

$\bar{x} = 1038$, $\sigma = 146$.

Now, test statistic.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1038 - 1000}{146/\sqrt{64}}$$

$$= \frac{38}{18.25}$$

$$= 2.0821.$$

$\therefore z_{cal} = 2.0821.$

for given $\alpha = 0.05$ the critical value is

$$z_\alpha = z_{tab} = 0.06 + 1.9$$

$$= 1.96.$$

$\therefore |z_{cal}| > z_\alpha.$

Conclusion:

We reject to at 5% lel 0.5

i.e the hypothesis that population mean is 1000 is rejected.

Q.2.

→ Soln.

let H_0: Digits may occur equally frequency in directory.

H_1: Digits may not occur equally frequently in directory.

Digit	Oi	Ei	$\frac{Oi^2}{Ei}$
0	1026	1000	1052.67
1	1107	1000	1225.44
2	997	1000	994.00
3	996	1000	933.15
4	1075	1000	1117.24
5	933	1000	870.48
6	1107	1000	1225.44
7	972	1000	994.78
8	964	1000	929.29
9	853	1000	729.60
Σ	10,000	10,000	10020.39

$$Ei^0 = \frac{1}{10} \times 10,000$$

$$= 1000.$$

$$Ei^0 = N = 1000.$$

We can use χ^2 statistics as follows,

$$\chi^2 = \sum_{i=1} \left(\frac{0_i^2}{E_i} \right) - \Sigma N$$

$$= 10020.39 - 10,000$$

$$\chi^2 = 20.39.$$

$\therefore \chi^2_{cal} = 20.39.$

for $\alpha = 1\%$ the critical value is.

$\chi^2_{cab} = 21.066.$

$\chi^2_{cal} < \chi^2_{cab}$

Conclusion:

Here, we accept H_0.

i.e Digits may be taken to occur frequently equally in the

Q.3.

→ Hypothesis to be tested:

$H_0 = \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$

Given: $n_1 = 1000$, $n = 800$, $\sigma_1 = 45.2$ cm

$\sigma_2 = 37.3$ cm, $\alpha = 5\% = 0.05$

$\bar{x}_1 = 150$ cm, $\bar{x}_2 = 146$ cm.

Test statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(150-146)}{\sqrt{\frac{(45.2)^2}{1000} + \frac{(37.3)^2}{850}}}$$

$$= 2.056$$

$\therefore z_{cal} \approx 2.06$

$z_\alpha = z_{tab} = 1.684$

$\therefore |z_{cal}| > z_\alpha$

Reject H_0:

Conclusion: we reject H_0

i.e urban area students are taller than other area students.

Ravi Suryawanshi [43]
Kajal Joshi [19]

Simple Linear Regression

New section

Importing the libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline
```

Importing the dataset

In [19]:
```
dataset = pd.read_csv("C:\\Users\\stat\\Simple_Linear_Regression\\Salary_Data.cs
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
dataset.head()
```

Out[19]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Splitting the dataset into the Training set and Test set

In [ ]:
```
```

In [27]:
```
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
from sklearn.metrics import mean_squared_error , r2_score
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, randc
dataset.columns = ['YearsExperience','Salary']
model = smf.ols(formula= 'YearsExperience ~ Salary',
                data=dataset).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	YearsExperience	R-squared:	0.957
Model:	OLS	Adj. R-squared:	0.955
Method:	Least Squares	F-statistic:	622.5
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	1.14e-20
Time:	20:57:38	Log-Likelihood:	-26.168
No. Observations:	30	AIC:	56.34
Df Residuals:	28	BIC:	59.14
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.3832	0.327	-7.281	0.000	-3.054	-1.713
Salary	0.0001	4.06e-06	24.950	0.000	9.3e-05	0.000

Omnibus:	3.544	Durbin-Watson:	1.587
Prob(Omnibus):	0.170	Jarque-Bera (JB):	2.094
Skew:	-0.412	Prob(JB):	0.351
Kurtosis:	2.003	Cond. No.	2.41e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.41e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Training the Simple Linear Regression model on the Training set

In [30]:
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Out[30]:
```
▾ LinearRegression

LinearRegression()
```

Predicting the Test set results

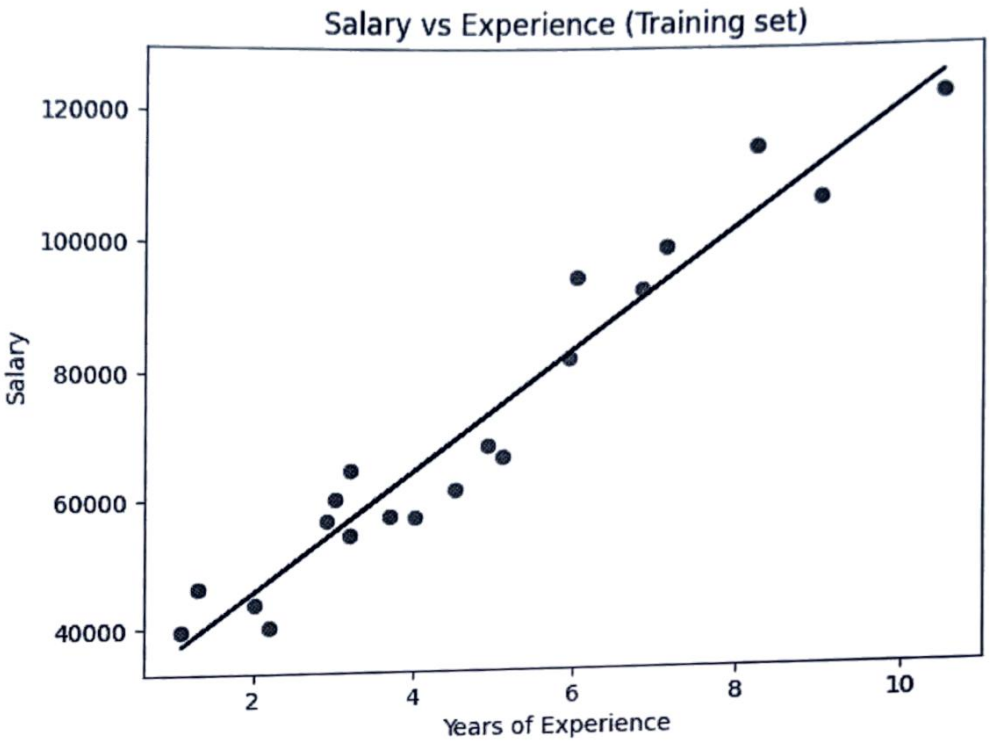
In [32]:
```
y_pred = regressor.predict(X_test)
print(y_pred)
```

```
[ 40835.10590871 123079.39940819  65134.55626083  63265.36777221
 115602.64545369 108125.8914992  116537.23969801  64199.96201652
  76349.68719258 100649.1375447 ]
```

Visualising the Training set results

In [33]:
```
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Training set)')
plt.xlabel('Years of Experience')
```

```
plt.ylabel('Salary')
plt.show()
```



Visualising the Test set results

In [34]:
```
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

Practical No:03

Correlation & Regression.

Q.1. $X = 65, 45, 50, 60, 40$ } Given.
$Y = 70, 35, 60, 50, 40$

To find $x = ?$

→

	X	Y	X^2	Y^2	XY
	65	70	4225	4900	4550
	45	35	2025	1225	1575
	50	60	2500	3600	3000
	60	50	3600	2500	3000
	40	40	1600	1600	1600
Σ	260	235	13950	13825	13725

$n = 5 \quad \bar{X} = \frac{260}{5} = 52 \quad \bar{Y} = \frac{255}{5} = 51$

$\bar{X}\bar{Y} = \frac{13725}{5} = 2652$

$COV(X, Y) = \frac{\Sigma X_i Y_i}{n} - \bar{X}\bar{Y}$

$= \frac{13725}{5} - 2652$

$\boxed{COV(X, Y) = 93}$

$\sigma X = \sqrt{\frac{\Sigma X_i^2}{n} - \bar{X}^2} = \sqrt{\frac{13950 - 2704}{5}}$

$$\boxed{\sigma x = 9.27}$$

$$\sigma y = \sqrt{\frac{\sigma y_i^2 - \bar{y}^2}{n}} = \sqrt{\frac{13825 - 2601}{5}}$$

$$\boxed{\sigma y = 12.8}.$$

$$\therefore \varepsilon = \frac{cov(x,y)}{\sigma x \, \sigma y} = \frac{93}{9.27 \times 12.8}$$

$$\boxed{\varepsilon = 0.78}.$$

Interpretation as $\varepsilon = 0.78$ there is high +ve correlation bet^n marks in static paper I & II.

Q. 2.

→ Given: $n = 50 \quad \Sigma x = 20 \quad \Sigma y = 25$

$\Sigma x^2 = 85, \quad \Sigma y^2 = 90, \quad \Sigma xy = 75$

To find eq^n of line of Regression.

$y - \bar{y} = b_{yx}(x - \bar{x})$ - (eq^n of line)

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

$$cov(x,y) = \frac{\Sigma xy}{n} - (\bar{x} - \bar{y})$$

$$= \frac{75}{50} - (0.4 \times 0.5)$$

$$= 1.5 - 0.2$$

$$cov(x,y) = 1.3$$

$$\sigma^2 x = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$= \frac{85}{50} - (0.4)^2$$

$$= 1.54$$

Now, $b_{yx} = \frac{Cov(x,y)}{\sigma^2 x}$

$$= \frac{1.3}{1.54}$$

$$= 0.8441$$

∴ The eqⁿ of line will be

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

$$y - 0.5 = 0.8441(x - 0.4) + 0.5$$

$$\therefore y = 0.8441x - 0.1624$$

e.3

→ let us assume that,

$3x + 12y = 19$ is the eqⁿ of line of Regression x on y &

$9x + 3y = 46$ is the eqⁿ of line of Regression y on x.

$\therefore 3x + 12y = 19$

$x = -4y + 19/3$

$\therefore b_{xy} = -4.$

&

$\therefore 9x + 3y = 46$

$y = -3x + 46/3$.

$b_{yx} = -3$.

We know,

$r^2 = b_{xy} \cdot b_{yx}$.

$= -4 \cdot -3$.

$r^2 = 12$

$\therefore r^2$ lies betⁿ 0 & 1

Hence, our assumptⁿ is wrong.

Now, we assume that,

$3x + 12y = 19$ is the eqⁿ of line of regression y on x &

$9x + 3y = 46$ is the eqⁿ of line of regression x on y.

$\therefore 3x + 12y = 19$

$\therefore y = \frac{-3x + 19}{12}$

$\therefore b_{yx} = -3/12 = -1/4$ &

$9x + 3y = 46$.

$\therefore x = \frac{-3y + 46}{9}$

$\therefore b_{xy} = -3/9$

$b_{xy} = -1/3$.

We knows,

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

$= -1/3 \cdot -1/4$.

$r^2 = 1/12 = 0.0833$

$r = 0.2887$ — $\therefore b_{xy}$ & b_{yx} are -ve

The sign of regression coeffient will always be the same, as the sign of the correlation coeffient

11. since both the line of regression posses through the pt. $(\bar{x}, \bar{y})$

The consider

$$3\bar{x} + 12\bar{y} = 19 \quad - (1)$$
$$9\bar{x} + 3\bar{y} = 46 \quad - (2).$$

multiply eqn (1) by 3, we get

$$9\bar{x} + 36\bar{y} = 54 \quad - (3).$$

eqn (3) − eqn (1)

$$\begin{array}{r} 9\bar{x} + 36\bar{y} = 57 \\ 9\bar{x} + 3\bar{y} = 46 \\ - \quad - \\ \hline 33\bar{y} = 11 \end{array}$$

$\therefore \bar{y} = 0.333.$

Substituting $\bar{y} = 1/3$ in eqn (3)

$$3\bar{x} + 12 \times 1/3 = 19$$
$$\bar{x} = 5$$

$\therefore$ mean of $x = \bar{x} = 5$ & mean of $y - \bar{y} = 1/3 = 0.333.$

Practical No: 04
Theory of probability.

Q.1.

→ $\Omega = \{HH, HT, TH, TT\}$

$n = 4.$

(i) A occurance of both the heads

$A = \{HH\}$

$n(A) = 1.$

$P(A) = \frac{n(A)}{n} = 1/4.$

∴ P(occurance of both of heads) = 1/4.

(ii) B = occurance of single head.

$B = \{HT, TH\}$

$n(B) = 2$

$P(B) = \frac{n(B)}{n} = 1/2$

∴ P(occurance of single head) = 1/2

(iii) C = occurance of atleast one head.

$C = \{HH, HT, TH\}$

$n(C) = 3$

$P(C) = \frac{n(C)}{n} = 3/4.$

∴ P(occurance of atleast one head) = 3/4.

Q.2.

→ The total no. of face in which be selected from 52 cards given by

$$^{52}C_4 \text{ r. } e^n C_r$$

Hence Ω written $n = {}^{52}C_4$.

(I) occurance of two red & two black cards.

Two red cards can be drawn in $^{26}C_2$ ways

Two black cards can be drawn in $^{26}C_2$ ways

∴ both the red & black cards are to be selected the No. of favourable cases.

To the event A,

$$m = {}^{26}C_2 \quad {}^{26}C_2$$

$$P(A) = \frac{m}{n}$$

$$= \frac{{}^{26}C_2 \quad {}^{26}C_2}{{}^{52}C_4}$$

$$P(A) = 0.3901$$

(II) B = All cards are of different suits.

$$m = {}^{13}C_4 \cdot {}^{13}C_4 \quad {}^{13}C_4 \cdot {}^{13}C_4$$

$$P(B) = \frac{m}{n}$$

$$= \frac{{}^{13}C_4 \cdot {}^{13}C_4 \cdot {}^{13}C_4 \cdot {}^{13}C_4}{{}^{52}C_4}$$

$$\therefore P(B) = 0.1054$$

(III) C: All cards of same units.

let event C = occurance of all cards of same suits.

i.e. all cards are either diamond, Hearts club spade, using the addition principle of containing.

$\therefore m = {}^{13}C_4 + {}^{13}C_4 + {}^{13}C_4 + {}^{13}C_4$

$= 4\ {}^{13}C_4$.

$P(C) = \frac{m}{n}$.

$= \frac{4\ {}^{13}C_4}{{}^{52}C_4}$.

$= \frac{4\ 13!_0}{\frac{(13-4)!_0 4!_0}{52!_0}}$

$\frac{}{(52-4)!_0 4!_0}$.

$P(C) = 0.01056$.

(IV) D = one king cards.

$m = 4C_1 \cdot 48C_3$

$P(D) = \frac{m}{n}$.

$= \frac{4C_1 \cdot 48C_3}{{}^{52}C_4}$.

$\therefore P(D) = 0.255$.

(i) P(two are red & two are black cards) = 0.3901

(ii) P(all cards are of diff. suits) = 0.1054

(iii) P(all cards are of same suits) = 0.01056.

(iv) P(one is king) = 0.2555.

Q. 3.

→ A commodity consist of 4 members, 3 engineer's, 4 economist, 2 statical & 1 CA.

$$n = \frac{10!}{(10-4)!\,4!} = 210.$$

(i) If event a consist of one engineer, economist, statistican, 1 CA hence, the probability of event A become.

$$A = {}^3C_1 \cdot {}^4C_1 \cdot {}^2C_1 \cdot {}^1C_1$$

$$n(A) = {}^{24}C_1$$

$$P(A) = \frac{{}^{24}C_1}{{}^{4}C_{10}} = \frac{24}{210} = \frac{4}{35}$$

$$\therefore P(A) = 0.1142$$

(ii) The event of B consists of CA & atleast 1 engineer.

a. 1 CA x 1 Eng x 2 other

$$\therefore {}^1C_1 \cdot {}^3C_1 \cdot {}^6C_2$$

$$= 45.$$

b. 1 CA x 2 Eng x 1 other.

$\therefore {}^1C_1 \times {}^3C_2 \times {}^6C_1$

$= 18$

c. 1 CA x 3 Eng x non-other.

$\therefore {}^1C_1 \cdot {}^3C_3 \cdot {}^6C_0$

$= 1$

$\therefore 45 + 18 + 1 = 64.$

$\therefore P(B) = \frac{64}{210} = 0.3048.$

(i) P (each of four categories is included in the commodity) = 0.1142

(ii) P (commodity consists CA & atleast one engineer) = 0.3048

Practical No: 05

Testing of hypothesis.

Q. 1.

→ Hypothesis to be tested:

$H_0 : \mu = 1000$ V/s $H_1 : \mu \neq 1000$.

Given: $\alpha = 5\% = 0.05$, $n = 64$

$\bar{x} = 1038$, $\sigma = 146$.

Now, test statistic.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1038 - 1000}{146/\sqrt{64}}$$

$$= \frac{38}{18.25}$$

$$= 2.0821.$$

$\therefore z_{cal} = 2.0821.$

for given $\alpha = 0.05$ the critical value is

$$z_\alpha = z_{tab} = 0.06 + 1.9$$

$$= 1.96.$$

$\therefore |z_{cal}| > z_\alpha.$

Conclusion:

We reject to at 5% lel 0.5

i.e the hypothesis that population mean is 1000 is rejected.

Q.2.

→ Soln.

let H_0: Digits may occur equally frequency in directory.

H_1: Digits may not occur equally frequently in directory.

Digit	Oi	Ei	$\frac{Oi^2}{Ei}$
0	1026	1000	1052.67
1	1107	1000	1225.44
2	997	1000	994.00
3	996	1000	933.15
4	1075	1000	1117.24
5	933	1000	870.48
6	1107	1000	1225.44
7	972	1000	994.78
8	964	1000	929.29
9	853	1000	729.60
Σ	10,000	10,000	10020.39

$$Ei^0 = \frac{1}{10} \times 10,000$$

$$= 1000.$$

$$Ei^0 = N = 1000.$$

We can use χ^2 statistics as follows,

$$\chi^2 = \sum_{i=1} \left(\frac{0_i^2}{E_i} \right) - \Sigma N$$

$$= 10020.39 - 10,000$$

$$\chi^2 = 20.39.$$

$\therefore \chi^2_{cal} = 20.39.$

for $\alpha = 1\%$ the critical value is.

$\chi^2_{cab} = 21.066.$

$\chi^2_{cal} < \chi^2_{cab}$

Conclusion:

Here, we accept H_0.

i.e Digits may be taken to occur frequently equally in the

Q.3.

→ Hypothesis to be tested:

$H_0 = \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$

Given: $n_1 = 1000$, $n = 800$, $\sigma_1 = 45.2$ cm

$\sigma_2 = 37.3$ cm, $\alpha = 5\% = 0.05$

$\bar{x}_1 = 150$ cm, $\bar{x}_2 = 146$ cm.

Test statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(150-146)}{\sqrt{\frac{(45.2)^2}{1000} + \frac{(37.3)^2}{850}}}$$

$$= 2.056$$

$\therefore z_{cal} \approx 2.06$

$z_\alpha = z_{tab} = 1.684$

$\therefore |z_{cal}| > z_\alpha$

Reject H_0:

Conclusion: we reject H_0

i.e urban area students are taller than other area students.

Ravi Suryawanshi [43]
Kajal Joshi [19]

Simple Linear Regression

New section

Importing the libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline
```

Importing the dataset

In [19]:
```
dataset = pd.read_csv("C:\\Users\\stat\\Simple_Linear_Regression\\Salary_Data.cs
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
dataset.head()
```

Out[19]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Splitting the dataset into the Training set and Test set

In [ ]:
```
```

In [27]:
```
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
from sklearn.metrics import mean_squared_error , r2_score
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, randc
dataset.columns = ['YearsExperience','Salary']
model = smf.ols(formula= 'YearsExperience ~ Salary',
                data=dataset).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	YearsExperience	R-squared:	0.957
Model:	OLS	Adj. R-squared:	0.955
Method:	Least Squares	F-statistic:	622.5
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	1.14e-20
Time:	20:57:38	Log-Likelihood:	-26.168
No. Observations:	30	AIC:	56.34
Df Residuals:	28	BIC:	59.14
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.3832	0.327	-7.281	0.000	-3.054	-1.713
Salary	0.0001	4.06e-06	24.950	0.000	9.3e-05	0.000

Omnibus:	3.544	Durbin-Watson:	1.587
Prob(Omnibus):	0.170	Jarque-Bera (JB):	2.094
Skew:	-0.412	Prob(JB):	0.351
Kurtosis:	2.003	Cond. No.	2.41e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.41e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Training the Simple Linear Regression model on the Training set

In [30]:
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Out[30]:
```
▾ LinearRegression

LinearRegression()
```

Predicting the Test set results

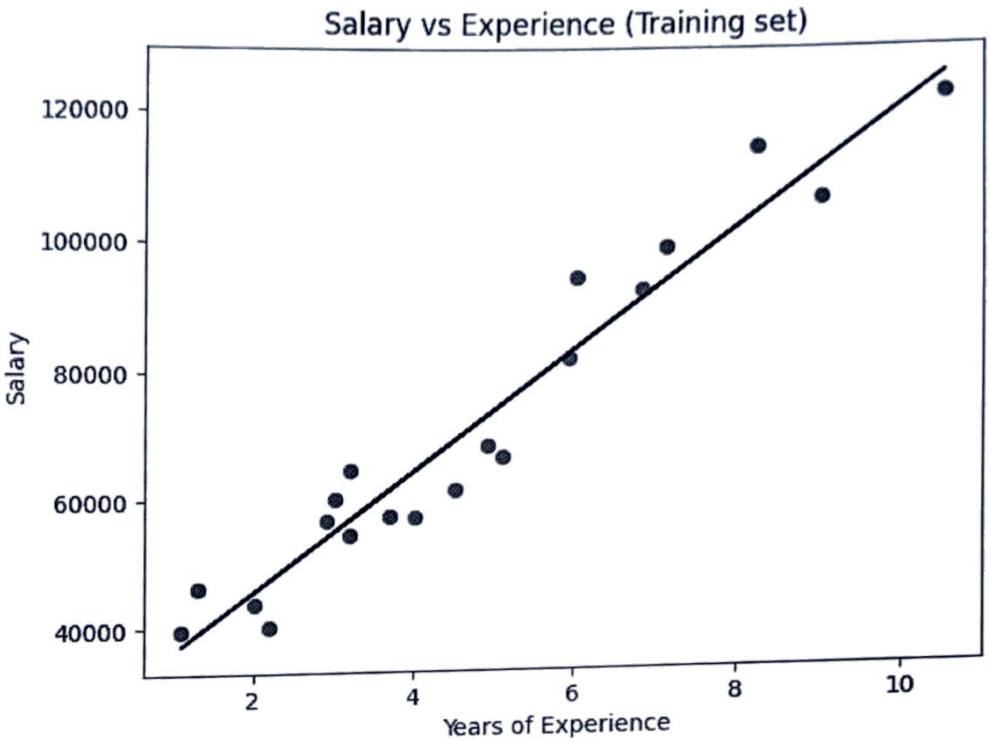
In [32]:
```
y_pred = regressor.predict(X_test)
print(y_pred)
```

```
[ 40835.10590871 123079.39940819  65134.55626083  63265.36777221
 115602.64545369 108125.8914992  116537.23969801  64199.96201652
  76349.68719258 100649.1375447 ]
```

Visualising the Training set results

In [33]:
```
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Training set)')
plt.xlabel('Years of Experience')
```

```
plt.ylabel('Salary')
plt.show()
```



Visualising the Test set results

In [34]:
```
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

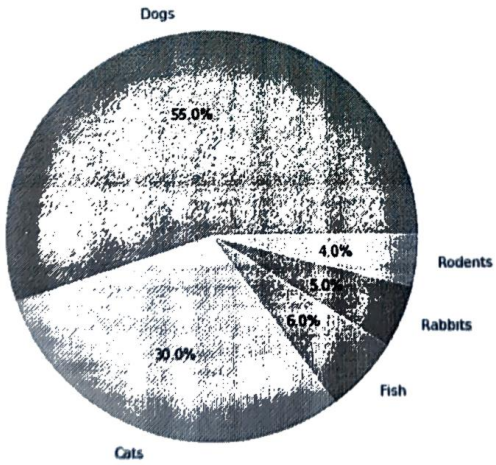
Practical No. 01

Name: Ravi Suryawanshi
Roll No. 43

Q.6 Draw a Pie Diagram for the following data.

Dogs	55%
Cats	30%
Fish	6%
Rabbits	5%
Rodents	4%

```
#Pie Chart 1

from matplotlib import pyplot as plt


animals = ["Dogs", "Cats", "Fish", "Rabbits", "Rodents"]

data = [55, 30, 6, 5, 4]


fig = plt.figure(figsize =(10, 7))

plt.pie(data, labels = animals, autopct = '%1.1f%%')


plt.show()
```

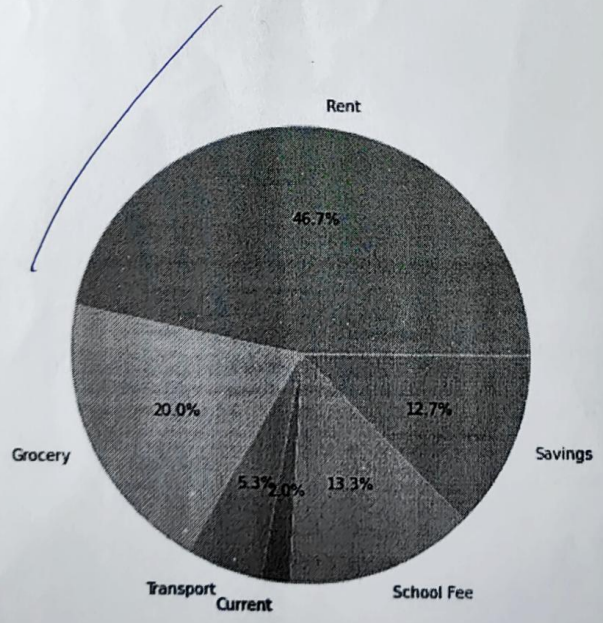


Q.7 Draw a pie diagram for the following data.

Expenses	Rent	Grocery	Transport	Current	School Fee	Savings
Amount	7000	3000	800	300	2000	1900

#Pie Chart 2

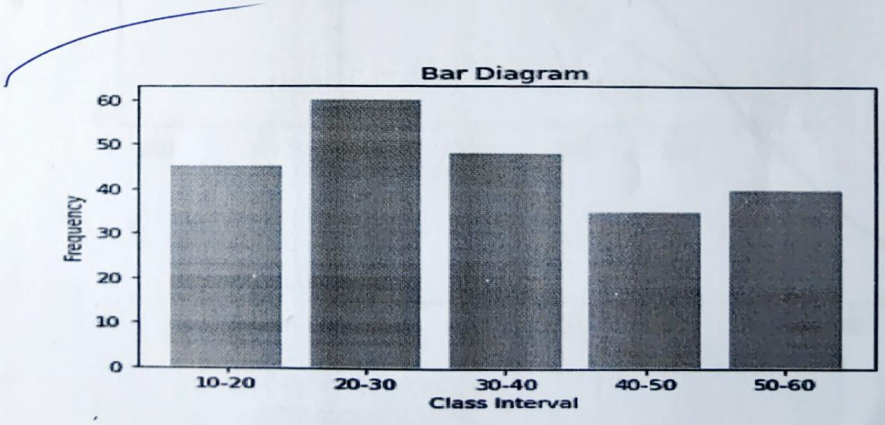
```
import matplotlib.pyplot as plt

expenses = ["Rent", "Grocery", "Transport", "Current", "School Fee", "Savings"]
amount = [7000, 3000, 800, 300, 2000, 1900]

fig = plt.figure(figsize =(10, 7))
plt.pie(amount, labels = expenses, autopct = '%1.1f%%')

plt.show()
```



Q.4 Represent the following data using simple bar diagram.

Class Interval	010-20	20-30	30-40	40-50	50-60
Frequency	45	60	48	35	40

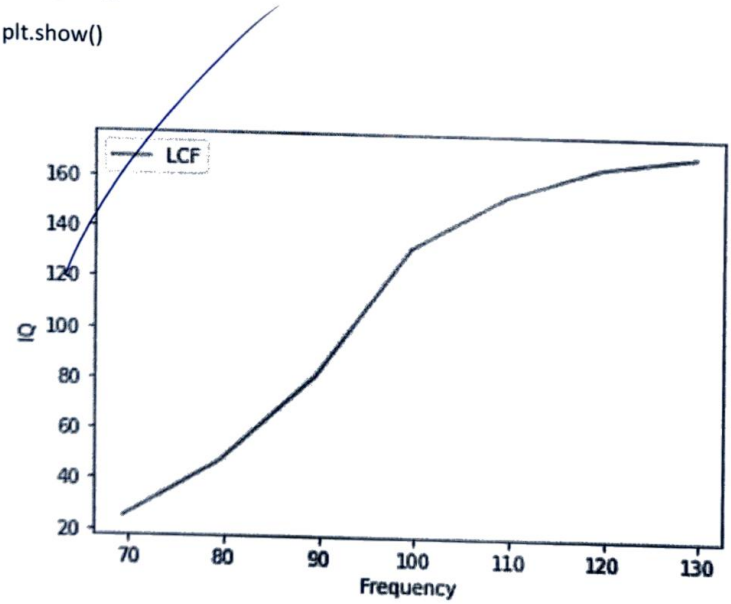
```
#Bar Plot

import matplotlib.pyplot as plt


x_ranges = ['10-20', '20-30', '30-40', '40-50', '50-60']

y_frequency = [45, 60, 48, 35, 40]


plt.bar(x_ranges, y_frequency, color='skyblue')


plt.xlabel('Class Interval')

plt.ylabel('Frequency')

plt.title('Bar Diagram')


plt.show()
```



Q.2 Draw the less than cumulative frequency curve from the following frequency distribution.

IQ	Frequency
60-69	25
70-79	22
80-89	34
90-99	51
100-109	21
110-119	12
120-129	5

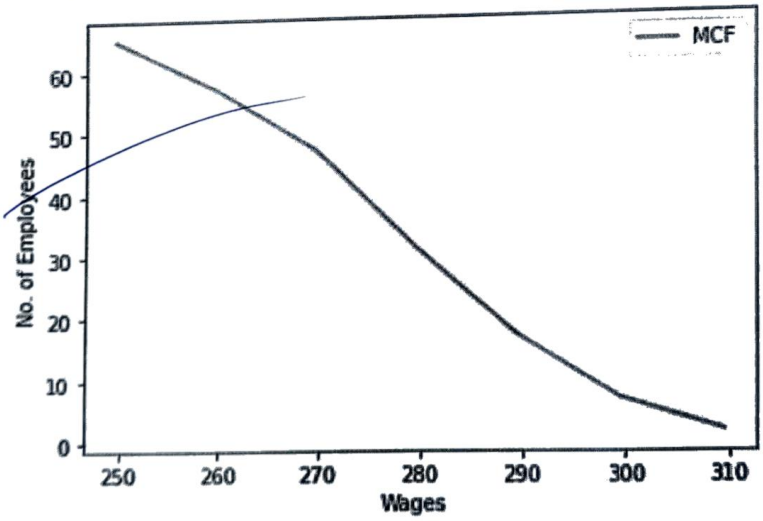
```
#LCF
import matplotlib.pyplot as plt
freq = [25, 47, 81, 132, 153, 165, 170]
iq = [69.5, 79.5, 89.5, 99.5, 109.5, 119.5, 129.5]
plt.plot(iq, freq, label = "LCF")
plt.xlabel("Frequency")
plt.ylabel("IQ")
plt.legend()
plt.show()
```



Q.3 The following table gives the frequency distribution of weekly wages of 65 employees of a company. Draw more than frequency curve.

Wages (Rs)	250-259	260-269	270-279	280-289	290-299	300-309	310-319
No of Employees	8	10	16	14	10	5	2

```
#MCF
import matplotlib.pyplot as plt
emp = [65, 57, 47, 31, 17, 7, 2]
wages = [249.5, 259.5, 269.5, 279.5, 289.5, 299.5, 309.5]
plt.plot(wages,emp, label = "MCF")
plt.xlabel("Wages")
plt.ylabel("No. of Employees")
plt.legend()
plt.show()
```



Name :-Ravi Suryawanshi (43)

Practical No. 02

Q1)Calculate the four central moments for the following data and also comment on nature ofdistribution.

Xi	1	2	3	4	5	6	7	8	9
Fi	1	16	13	25	30	22	9	5	2

=>

```
frequency_distribution = {1: 3, 2: 4, 3: 2, 4: 1}

total_frequency = sum(frequency_distribution.values())

mean = sum(x * (freq / total_frequency) for x, freq in frequency_distribution.items())

n = 2

central_moment = sum(((x - mean) ** n) * (freq / total_frequency) for x, freq in
frequency_distribution.items())

print(f"Mean (First Central Moment): {mean}")

print(f"{n}th Central Moment: {central_moment}")
```

O/p=>

```
Mean (First Central Moment): 2.1
2th Central Moment: 0.89
```

Q2) Compute the i) Karl Pearson's Coefficient of Skewness . ii) Bowley's Coefficient of Skewness and iii)Pearsonian Coefficient of Skewness from the following data:

Daily Expenditure (Rs.)	0-20	20-40	40-60	60-80	80-100
No. of Families.	13	19	25	27	16

i)=>

```
importnumpy as np

classes = [10, 20, 30, 40, 50]  # Define the class boundaries

frequencies = [5, 12, 20, 8, 5]  # Define the corresponding frequencies

midpoints = [(classes[i] + classes[i + 1]) / 2 for i in range(len(classes) - 1)]
```

```
mean = sum(midpoints[i] * frequencies[i] for i in range(len(midpoints))) / sum(frequencies)

variance = sum(((midpoints[i] - mean) ** 2) * frequencies[i] for i in range(len(midpoints))) /
sum(frequencies)

std_deviation = np.sqrt(variance)

if mean < median:

skewness_type = "positive"

elif mean > median:

skewness_type = "negative"

else:

skewness_type = "no skew"

skewness = 3 * (mean - median) / std_deviation

print("Mean:", mean)

print("Standard Deviation:", std_deviation)

print("Skewness Type:", skewness_type)

print("Pearson's Coefficient of Skewness:", skewness)
```

O/p=>

```
Mean: 28.7
Standard Deviation: 8.968890678339212
Skewness Type: negative
Pearson's Coefficient of Skewness: 1.906590303447253
```

ii)=>

```
importnumpy as np
fromscipy import stats
data = np.array([12, 15, 17, 18, 20, 21, 22, 23, 25, 28, 30, 32, 35, 40, 45])
median = np.median(data)
q1, q3 = np.percentile(data, [25, 75])
iqr = q3 - q1
mode_result = stats.mode(data)
mode = mode_result.mode
bowley_skewness = (median - mode) / iqr
print("Median:", median)
```

```
print("Mode:", mode)
print("Interquartile Range (IQR):", iqr)
print("Bowley's Coefficient of Skewness:", bowley_skewness)
```

O/p=>

```
Median: 23.0
Mode: 12
Interquartile Range (IQR): 12.0
Bowley's Coefficient of Skewness: 0.9166666666666666
```

iii)=>

```
importnumpy as np

data = np.array([1.2, 2.5, 3.7, 4.1, 5.8, 6.2, 7.9])

mean = np.mean(data)

std_dev = np.std(data)

central_moment3 = np.mean((data - mean) ** 3)

pearsonian_skewness = central_moment3 / (std_dev ** 3)

print("Pearsonian Coefficient of Skewness:", pearsonian_skewness)
```

O/p=>

```
Pearsonian Coefficient of Skewness: 0.04810830532968162
```

Q3)Compute the first four central moments for the following frequency distribution of wages of workerin a factory.

Wages (In Rs.)	100-200	200-300	300-400	400-500	500-600
No. of Employees	8	30	10	9	3

=>

```
importnumpy as np

data = np.array([1.2, 2.5, 2.7, 3.1, 3.5, 4.0, 4.2, 4.8, 5.0])

mean = np.mean(data)
```

```
variance = np.var(data)

skewness = np.mean((data - mean) ** 3) / (variance ** (3/2))

kurtosis = np.mean((data - mean) ** 4) / (variance ** 2)

print("Mean:", mean)

print("Variance:", variance)

print("Skewness:", skewness)

print("Kurtosis:", kurtosis)
```

O/p=>

```
Mean: 3.4444444444444446
Variance: 1.3046913580246913
Skewness: -0.42829398867132956
Kurtosis: 2.359431132277121
```

Ravi Suryawanshi [437]
Kajal Joshi [19]

Simple Linear Regression

New section

Importing the libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline
```

Importing the dataset

In [19]:
```
dataset = pd.read_csv("C:\\Users\\stat\\Simple_Linear_Regression\\Salary_Data.cs
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
dataset.head()
```

Out[19]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Splitting the dataset into the Training set and Test set

In [ ]:

In [27]:
```
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
from sklearn.metrics import mean_squared_error , r2_score
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, randc
dataset.columns = ['YearsExperience','Salary']
model = smf.ols(formula= 'YearsExperience ~ Salary',
                data=dataset).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	YearsExperience	R-squared:	0.957
Model:	OLS	Adj. R-squared:	0.955
Method:	Least Squares	F-statistic:	622.5
Date:	Mon, 27 Nov 2023	Prob (F-statistic):	1.14e-20
Time:	20:57:38	Log-Likelihood:	-26.168
No. Observations:	30	AIC:	56.34
Df Residuals:	28	BIC:	59.14
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.3832	0.327	-7.281	0.000	-3.054	-1.713
Salary	0.0001	4.06e-06	24.950	0.000	9.3e-05	0.000

Omnibus:	3.544	Durbin-Watson:	1.587
Prob(Omnibus):	0.170	Jarque-Bera (JB):	2.094
Skew:	-0.412	Prob(JB):	0.351
Kurtosis:	2.003	Cond. No.	2.41e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.41e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Training the Simple Linear Regression model on the Training set

In [30]:
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Out[30]:
```
▾ LinearRegression
LinearRegression()
```

Predicting the Test set results

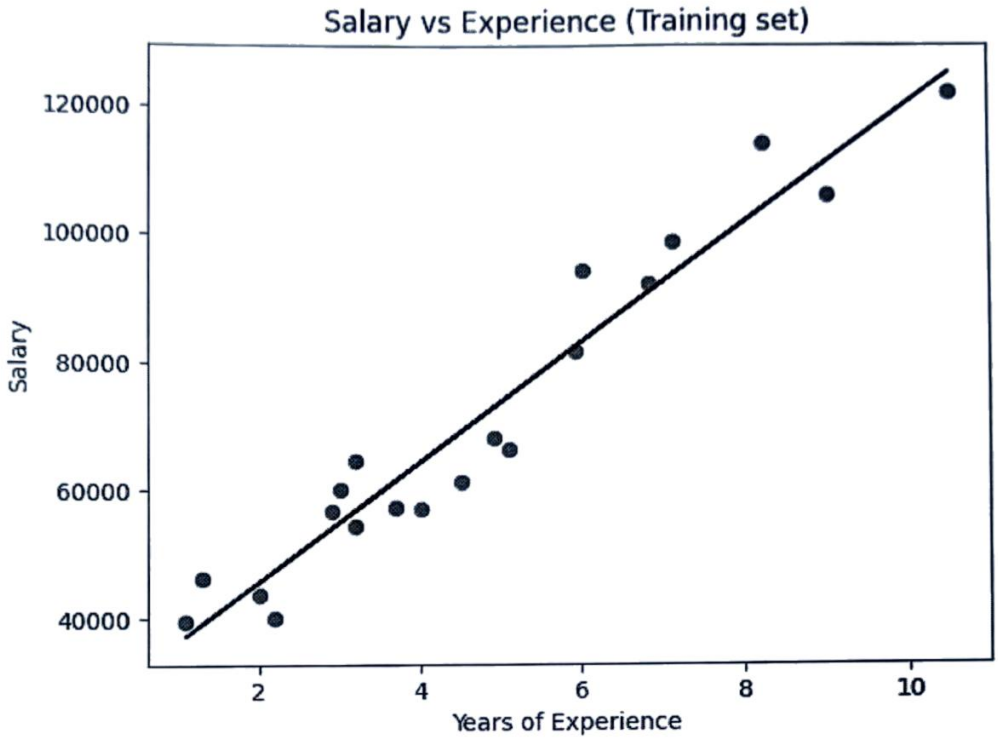
In [32]:
```
y_pred = regressor.predict(X_test)
print(y_pred)
```

```
[ 40835.10590871 123079.39940819  65134.55626083  63265.36777221
 115602.64545369 108125.8914992  116537.23969801  64199.96201652
  76349.68719258 100649.1375447 ]
```

Visualising the Training set results

In [33]:
```
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Training set)')
plt.xlabel('Years of Experience')
```

```
plt.ylabel('Salary')
plt.show()
```



Visualising the Test set results

In [34]:
```
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

In [35]:
```
#prediction of testing data using a training model

y_pred = regressor.predict(X_test)
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df
```

Out[35]:

	Actual	Predicted
0	37731.0	40835.105909
1	122391.0	123079.399408
2	57081.0	65134.556261
3	63218.0	63265.367772
4	116969.0	115602.645454
5	109431.0	108125.891499
6	112635.0	116537.239698
7	55794.0	64199.962017
8	83088.0	76349.687193
9	101302.0	100649.137545

In [ ]: