Gandaki University

Rajchautara-32, Pokhara, Gandaki

A

Practical Report On

Fundamentals of Probability and Statistics

Submitted By:

Name: Nirajan Dhakal

Semester: IV

Roll no.: 17

Submitted To:

Ashish Bhujel

Asst. Professor

Gandaki University

1. Given the bivariate data:

X	1	5	3	2	1	1	7	3
y	6	1	0	0	1	2	1	5

a) Fit a regression line of Y on X and hence estimate Y when X=8.

Solution:

Working Expression,

Regression line of Y on X is given by,

$$Y = a + bX$$

Where,

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

b = regression coefficient of Y on X = $\frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.875	1.438		1.999	.093
	X	-.304	.409	-.291	-.744	.485

a. Dependent Variable: Y

From the table above we get,

a = 2.875 and b = -0.304

the estimated model is,

$\hat{Y}$ = a + bX

$\hat{Y}$ = 2.875 - 0.304X

when, x = 8,

$\hat{Y}$ = 2.875 - 0.304 * 8 = 0.443

b. Fit a regression line of X on Y and hence predict X if Y = 3.5.

Solution:

Working Expression,

Regression line of X on Y is given by,

$$X = a + bY$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.431	1.088		3.152	.020
	Y	-.278	.373	-.291	-.744	.485

a. Dependent Variable: X

Where,

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

b = regression coefficient of Y on X = $\frac{n*\sum xy - \sum x * \sum y}{n*\sum x^2 - (\sum x)^2}$

From the table above we get,

A = 3.431 and b = -0.278

The estimated model is,

$\hat{X}$ = a + bY

$\hat{X}$ = 3.431 - 0.278Y

When Y=3.5

$\hat{X}$ = 3.431 - 0.278 * 3.5 = 2.458

2. The data in sales and promotion expenditures on a newly launched product for 6 years are given below:

Year	2003	2004	2005	2006	2007	2008
Sales (in Rs.00,000)	16	20	18	24	20	22
Promotion expenses (Rs. 000)	4	4	6	10	10	12

a) Calculate the two regressions coefficient from the above data of sales and expense.
b) Compute correlation coefficient between sales and expenditure and interpret.
c) Test the significance of the correlation coefficient.
d) Develop the estimating equation that describes the effect if promotional expenses on sales.

e) Explain the meaning of each parameter of the equation, in terms of the above information.

Solution:

Let X represent `Sales` and Y represent `Expenses`.

a. **Solution:**

Working Expression,

Regression line of X on Y is given by,

$$X = a + bY$$

Where,

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

b = regression coefficient of Y on X = $\frac{n*\sum xy - \sum x * \sum y}{n*\sum x^2 - (\sum x)^2}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1534831.461	229202.467		6.696	.003
	expense	60.674	27.660	.739	2.194	.093

a. Dependent Variable: sales

From the table above, we get,

a = 1534831.461 and b = 60.674

b. **Solution:**

Working expression:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Where,

r = Pearson correlation coefficient

n = number of data points

x & y = individual sample points of variables X and Y

Correlations

		sales	expense
sales	Pearson Correlation	1	.739
	Sig. (2-tailed)		.093
	N	6	6
expense	Pearson Correlation	.739	1
	Sig. (2-tailed)	.093	
	N	6	6

From the table above,

The *Pearson Correlation Coefficient,* 'r' = 0.739

It indicates that there is a high degree of positive correlation between sales and promotion expenses. This suggests that as promotion expenses increase, sales tend to increase as well.

c. **Solution**:

The significance (p-value) of the correlation is 0.093, which is greater than the common significance level of 0.05. This means that although there is a strong positive correlation, it is not statistically significant at the 5% level. This could be due to the small sample size (n=6).

d. **Solution**:

We have,

a = 1534831.461 and b = 60.674

Then, estimation equation is given by,

$$X = a + bY$$

sales = 1534831.461 + 60.674 * expenses

when expenses = 20000,

sales = 1534831.461 + 60.674 * 20000 = 2748311.461

Therefore, the sales when promotional expenses are 20000 is expected to be Rs. 2748311.461.

e. **Solution**:

- The constant term (a = 1534831.461) represents the baseline sales without any promotion expenses. It suggests that even with no promotional spending, the company might expect sales of about 1,534,831 units.
- The coefficient of expenses (b = 60.674) indicates that for each additional unit of currency spent on promotion, sales are expected to increase by 60.674 units. This shows a positive return on investment for promotional spending.

3. The weight of fruits were recorded for a sample of 25. The data to be the nearest grams are given:

66 51 92 65 84 57 96 58 56 80 73 55 71 77 78 93 89 61 66 69 96 96 61 75 69

a) Present the data in stem and leaf display.
b) Construct the histogram.
c) Discuss the shape of the data by preparing box-and-whisker plot.

a. **Solution**:

X Stem-and-Leaf Plot

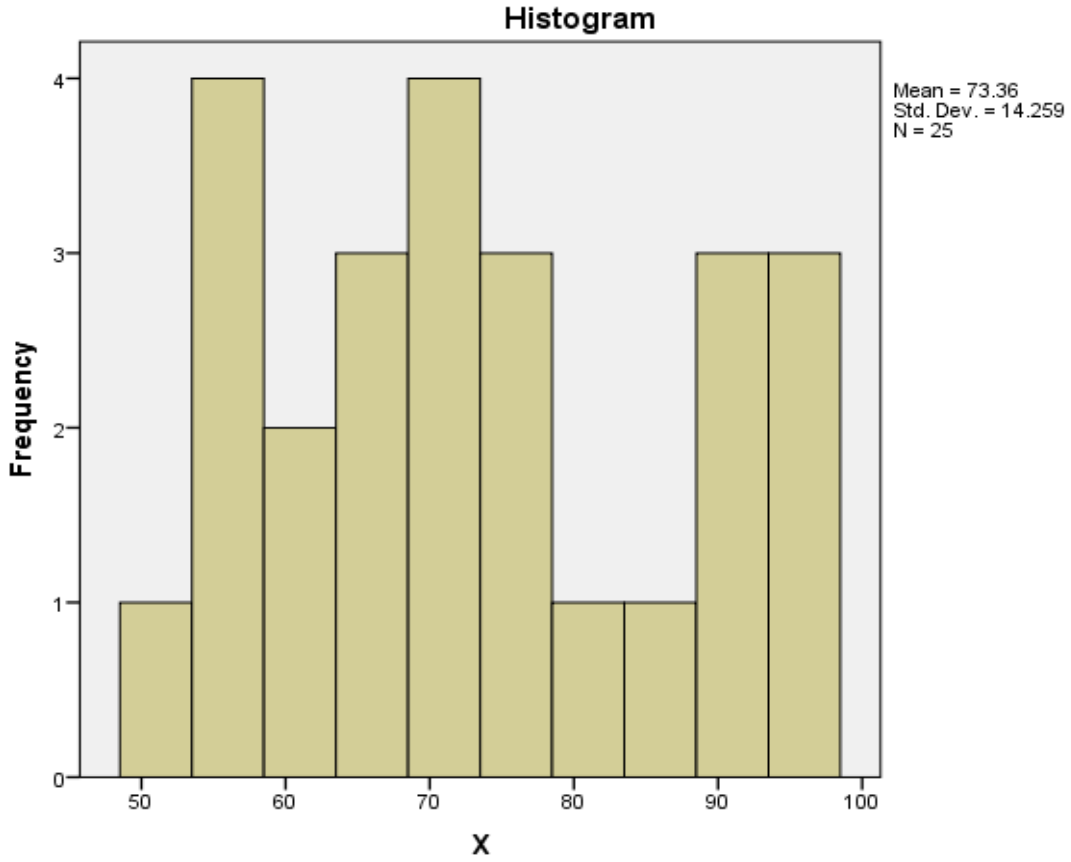
Frequency	Stem &		Leaf
1.00	5	.	1
4.00	5	.	5 6 7 8
2.00	6	.	1 1
5.00	6	.	5 6 6 9 9
2.00	7	.	1 3
3.00	7	.	5 7 8
2.00	8	.	0 4
1.00	8	.	9

2.00	9	.	2 3
3.00	9	.	6 6 6

Stem width: 10 Each leaf: 1 case(s)

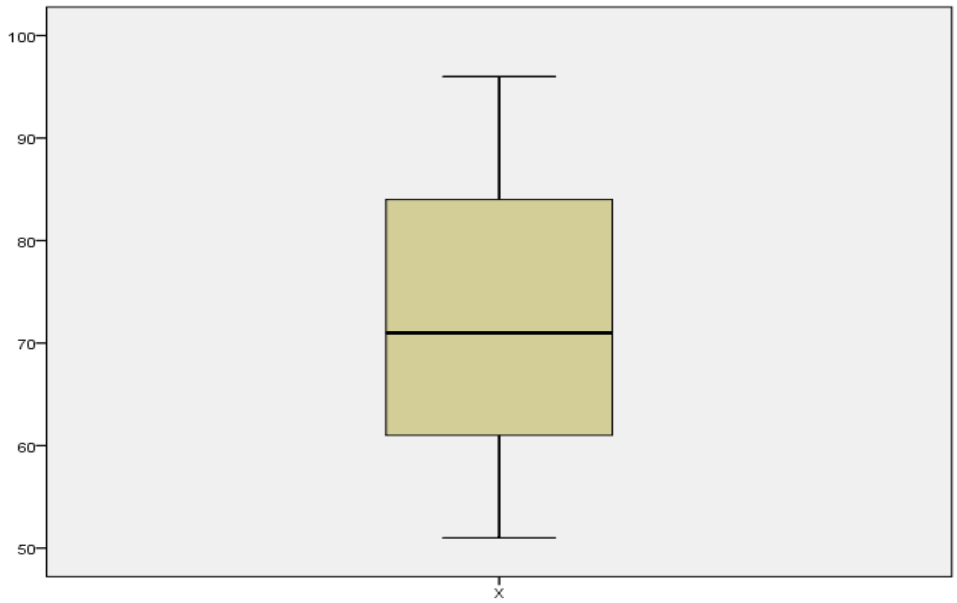
This display shows the distribution of fruit weights. The stem represents the tens digit, and the leaf represents the ones digit.

b. **Solution**:



This histogram visually represents the distribution of fruit weights. The x-axis shows weight ranges, and the y-axis shows the frequency of fruits in each range. We can observe that the distribution is somewhat right skewed, with a concentration of weights between 60 and 80 grams.

c. **Solution**:



The box-and-whisker plot provides several insights about the distribution:

i. The median (middle line in the box) is slightly below the center of the box, indicating a slight right skew.
ii. The whisker on the right side is longer than the left, further confirming the right skew.
iii. There are no outliers visible in this plot.
iv. The box (representing the middle 50% of the data) is relatively compact, suggesting that half of the fruit weights are clustered together.

Overall, the distribution of fruit weights is right-skewed, meaning there are some fruits with notably higher weights pulling the distribution to the right.

4. From the following data, calculate the Spearman's correlation coefficient.

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

Solution:

Spearman's correlation coefficient is used to measure the strength and direction of association between two ranked variables. It's particularly useful when the relationship between variables might be nonlinear but monotonic.

Working Expression,

$$\rho = \frac{6\sum d^2}{n(n^2-1)}$$

Where,

ρ = Spearman's rank correlation coefficient

Correlations

			x	y
Spearman's rho	x	Correlation Coefficient	1.000	.747*
		Sig. (2-tailed)	.	.013
		N	10	10
	y	Correlation Coefficient	.747*	1.000
		Sig. (2-tailed)	.013	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

d = difference between the two ranks of each observation

n = number of observations

The Spearman correlation coefficient between x and y is 0.747 i.e. there is a high degree of positive correlation.

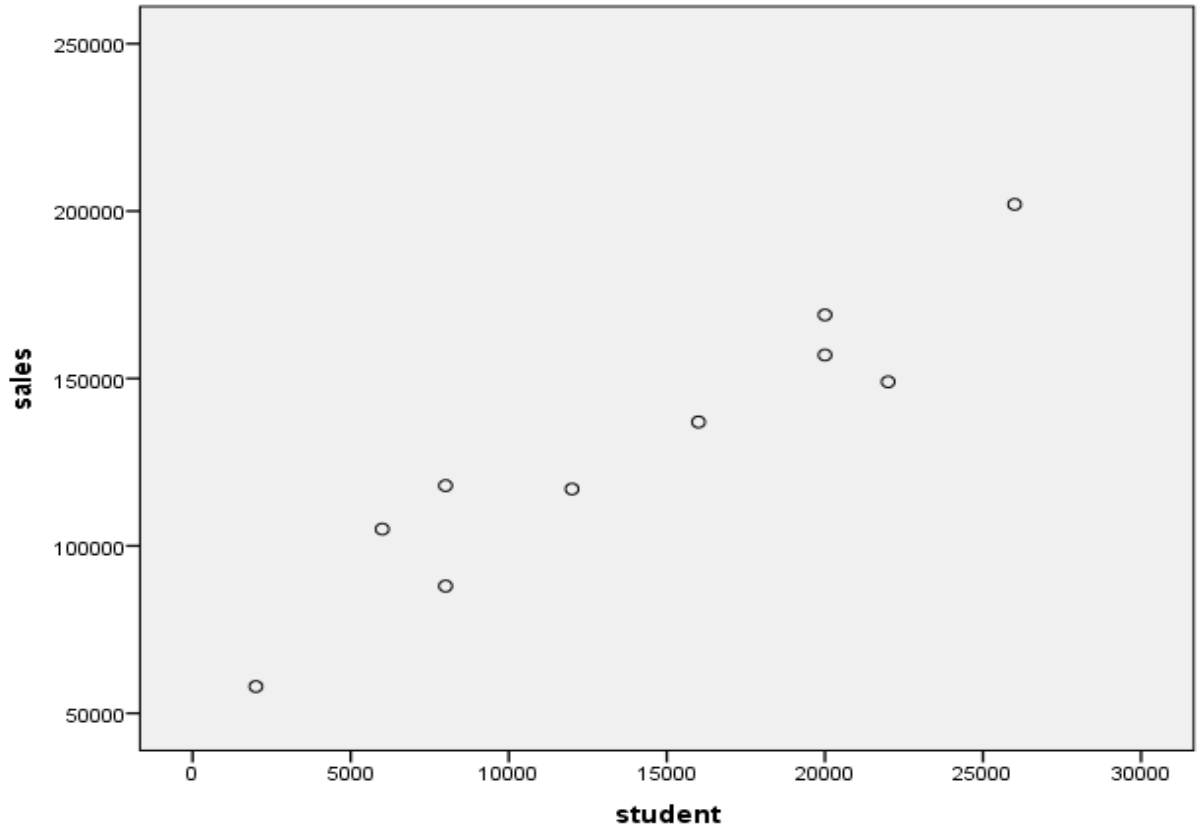
In practical terms, this means that as the ranks of variable x increase, the ranks of variable y tend to increase as well, and this relationship is unlikely to have occurred by chance. However, remember that correlation does not imply causation.

5. The following is the hypothetical data representing student population and quarterly sales for 10 busy restaurants of Pokhara nearby PN Campus.

Restaurant	1	2	3	4	5	6	7	8	9	10
Student population (1000s)	2	6	8	8	12	16	20	20	22	26
Quarterly sales (Rs. 1000s)	58	105	88	118	117	137	157	169	149	202

a) Develop a scatter diagram for these data.

Solution:



This scatter plot visualizes the relationship between student population (x-axis) and quarterly sales (y-axis) for the 10 restaurants. Each point represents a single restaurant.

Observations from the scatter plot:

1. There appears to be a strong positive linear relationship between student population and quarterly sales.
2. As the student population increases, quarterly sales tend to increase as well.
3. The points form a fairly tight cluster around an imaginary straight line, suggesting a strong correlation.
4. There don't appear to be any obvious outliers or anomalies in the data.

b) Try to approximate the relationship between student population (X) and quarterly sales (Y) through the data.

Solution:

To approximate the relationship, we'll use a linear regression model. The strength of this relationship can be assessed using the coefficient of determination (R^2).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.950[a]	.903	.891	13829.317

a. Predictors: (Constant), student

The R^2 value of 0.903 indicates that 90.3% of the sales is accounted for by the student.

Interpretation:

1. R = 0.950 indicates a very strong positive correlation between student population and quarterly sales.
2. R^2 = 0.903 means that approximately 90.3% of the variation in quarterly sales can be explained by the variation in student population.
3. The adjusted R^2 (0.891) is close to R^2, suggesting that the model isn't overfitted.
4. The standard error of the estimate (13829.317) represents the average distance that the observed values fall from the regression line. Smaller values indicate a better fit.

c) Develop the estimated regression equation by computing the values of a and b, and interpret a and b.

Solution:

Let X represent student and Y represent sales,

Working Expression,

Regression line of Y on X is given by,

$$Y = a + bX$$

Where,

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

b = regression coefficient of Y on X = $\frac{n*\sum xy - \sum x * \sum y}{n*\sum x^2 - (\sum x)^2}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60000.000	9226.035		6.503	.000
	student	5.000	.580	.950	8.617	.000

a. Dependent Variable: sales

From the table above we get,

a = 60000 and b = 5

Therefore, estimated regression equation is,

$$Y = a + bX$$

Y = 60000 + 5X

i.e. sales = 60000 + 5 * student

d) Predict the sales for a restaurant to be located with 16000 students.

Solution:

When student(X) = 16000

Then,

sales = 60000 + 5 * 16000 = 140000

Hence, the estimated sales with 16000 students are Rs. 140000.

6. The store in Kathmandu has realized that the sales of the product depend entirely on the advertising and pricing. The store has collected the following data.

Daily sales (in 000 unit sold)	10	12	14	16	18	20	22
Advertising Rs. 000	1	2	3	4	5	6	7
Price Rs.	5	8	10	12	6	10	5

a) Develop the multiple regression equation for the above data.

Solution:

Using multiple linear regression, we obtain:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.000	.000		.	.
	ad	2.000	.000	1.000	.	.
	price	.000	.000	.000	.	.

a. Dependent Variable: sales

The multiple regression equation is:

$$Sales = 8 + 2(ad) + 0(price)$$

Interpretation:

1. The constant term (8) suggests that even with no advertising and regardless of price, the base sales would be 8 units.
2. The coefficient for advertising (2) indicates that for each unit increase in advertising, sales are expected to increase by 2 units, holding price constant.
3. The coefficient for price (0) suggests that price has no effect on sales, which is unusual and may warrant further investigation.
4. The lack of standard errors and significance values ('.') is concerning and might indicate perfect multicollinearity or other data issues.

b) Compute coefficient of multiple determination and interpret the result.

Solution:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.000

a. Predictors: (Constant), price, ad

The coefficient of multiple determination (R^2) is 1.000.

Interpretation:

i. An R^2 of 1.000 indicates that 100% of the variation in sales is explained by advertising and price in this model.
ii. This perfect fit is extremely unusual in real-world data and suggests potential issues:
    a. The model might be overfitted.
    b. There could be perfect multicollinearity between variables.
    c. The sample size might be too small (equal to the number of predictors plus one).
    d. There might be errors in data entry or calculation.
iii. The standard error of the estimate being 0.000 further confirms the perfect fit, which is rarely seen in practice.

While the results show a perfect fit, they are likely too good to be true. In real-world scenarios, such perfect relationships are extremely rare. It's crucial to review the data collection process, check for data entry errors, and consider collecting more data points if possible. The model, as it stands, may not be reliable for making predictions or drawing conclusions about the relationships between advertising, pricing, and sales.

Conclusion

This practical report has explored various statistical techniques and their applications in analyzing different datasets. Here's a summary of our key findings and their implications:

1. **Bivariate Data Analysis**:

We examined the relationship between two variables using linear regression. The negative slopes in both regression equations (Y on X and X on Y) suggest an inverse relationship between the variables. This demonstrates how the choice of dependent and independent variables can affect the interpretation of results.

2. **Sales and Promotion Expenditure Analysis**:

We found a strong positive correlation ($r = 0.739$) between promotional expenses and sales. The regression equation suggests that increased promotional spending is associated with higher sales. However, the relationship was not statistically significant at the 5% level, possibly due to the small sample size. This highlights the importance of considering sample size in statistical analyses.

3. **Weight of Fruits Analysis**:

The stem-and-leaf plot, histogram, and box-and-whisker plot all indicated a right-skewed distribution of fruit weights. This suggests that while most fruits cluster around a lower weight range, there are some notably heavier fruits pulling the distribution to the right. Such insights can be valuable in quality control or pricing strategies in fruit production or retail.

4. **Spearman's Correlation Coefficient Analysis**:

We found a strong positive correlation ($\rho = 0.747$) between the two variables, which was statistically significant. This demonstrates the utility of Spearman's correlation in analyzing ranked data or when the relationship between variables might be nonlinear but monotonic.

5. **Student Population and Quarterly Sales Analysis**:

The scatter plot and regression analysis revealed a very strong positive relationship between student population and restaurant sales ($R^2 = 0.903$). The model suggests that for every additional student, quarterly sales increase by $5. This information could be extremely valuable for restaurant owners considering locations near educational institutions.

6. **Multiple Regression Analysis**:

The perfect fit ($R^2 = 1.000$) in the advertising and pricing model is highly unusual and suggests potential issues with the data or model specification. This serves as a reminder of the importance of critical thinking and data validation in statistical analysis.

To sum up, this report illustrates the potency of statistical methods in revealing connections and trends from diverse sets of data. Statistics offers useful

instruments for making decisions based on data, starting from business activities such as sales prediction and evaluation of promotional strategies to quality control in agricultural production.

This report also highlights some important things to consider in statistical analysis:
1. **Sample Size**: As seen in the promotion expenditure analysis, small sample sizes can limit the statistical significance of findings.

2. **Data Quality**: The perfect fit in the multiple regression analysis reminds us of the importance of data validation and the need to approach unusually perfect results with skepticism.

3. **Context**: While statistical tools are powerful, their results must always be interpreted within the context of the problem and with consideration of external factors that may not be captured in the data.

4. **Model Assumptions**: The appropriateness of linear regression, correlation analyses, and other techniques depends on certain assumptions about the data. It's crucial to verify these assumptions for reliable results.

5. **Causation vs. Correlation**: While we found several strong correlations, it's important to remember that correlation does not imply causation. Additional research and controlled experiments would be needed to establish causal relationships.

In conclusion, this practical report not only demonstrates the application of various statistical techniques but also emphasizes the importance of critical thinking and cautious interpretation in data analysis. As we continue to navigate an increasingly data-driven world, these statistical skills and considerations will be invaluable across various fields and industries.