DegreeLive

BBA | Semester 6 | UK6DSCMGT303

Statistics for Business Decisions

Course Code: UK6DSCMGT303 • Module 3 Notes

degreelive.in

Module 3: Correlation Analysis

Bivariate Association, Causation Distinction, Scatter Diagrams, Karl Pearson r, and Spearman Rank

Chapter 1: The Epistemological Foundations of Bivariate Association

1.1 Definitional Framework and Nature of Correlation

In univariate statistical analysis (measures of central tendency and dispersion), investigations focus exclusively on the behavior, typical value, and variability of a single isolated variable (e.g., examining the distribution of worker wages or the dispersion of product defect rates). However, in modern corporate management and economic reality, commercial phenomena do not exist in isolation; variables interact in continuous, dynamic interdependence. Executives are constantly confronted with bivariate questions: Does an expansion in digital advertising expenditure drive an increase in sales revenue? Does an increase in corporate training hours yield a reduction in manufacturing defect rates? Does employee job tenure correlate with workplace productivity?

The statistical tool dedicated to measuring and analyzing the degree of relationship or association between two or more interrelated variables is **Correlation Analysis**.

According to W. I. King, "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other(s), then they are said to be correlated." A. M. Tuttle defines correlation as "an analysis of the co-variation between two or more variables." L. R. Connor formulates: "If two or more variables move together in sympathy, they are said to be correlated. If the movement of one is accompanied by the movement of the other in the same direction, they are positively correlated; if in opposite directions, negatively correlated."

1.2 Managerial Significance of Correlation in Decision Making

Correlation analysis provides corporate decision-makers with three vital analytical capabilities:

1. **Quantifying the Strength and Direction of Interdependence:** Correlation synthesizes the complex co-movement of two commercial metrics into a single, standardized, dimensionless coefficient ranging from -1.0 to +1.0, revealing instantly whether two variables move in unison or opposition.
2. **Serving as the Prerequisite for Regression and Forecasting:** A high degree of correlation between two variables is the foundational theoretical justification for constructing predictive regression models. If no correlation exists between advertising and revenue, building a regression equation to forecast sales from ad budgets is mathematically futile.
3. **Risk Diversification in Investment Portfolios:** In modern portfolio theory (Markowitz model), investment fund managers deliberately combine financial securities that exhibit negative or low positive correlation. When equity prices drop during an economic downturn, negatively correlated sovereign bonds surge, preserving overall portfolio capital.

Degreelive

Chapter 2: The Critical Philosophical Distinction between Correlation and Causation

2.1 The Postulate: 'Correlation Does Not Imply Causation'

A fundamental epistemological error in corporate analytics is the uncritical assumption that because two variables exhibit a strong mathematical correlation, a direct cause-and-effect relationship exists between them. **Correlation establishes mathematical co-variation, not physical causation.** Two variables may display a statistically significant correlation coefficient under three entirely non-causal scenarios:

1. The Confounding Third Variable (Spurious Correlation)

Two variables may co-vary strongly not because either influences the other, but because both are independently driven by an unobserved common underlying factor (a confounding variable). For example, historical data in a coastal resort city may reveal a high positive correlation ($r = +0.88$) between monthly ice cream sales and monthly accidental drowning deaths. Concluding that consuming ice cream causes swimmers to drown, or that drowning accidents trigger ice cream cravings, is absurd. The co-variation is entirely driven by a third confounding variable: *ambient summer temperature*. Higher temperatures independently drive both ice cream consumption and beach swimming activity.

2. Pure Chance or Accidental Correlation (Nonsense Correlation)

In massive modern databases, algorithms testing thousands of unrelated series inevitably uncover random, bizarre mathematical correlations purely through chance. Classic documented examples include a +0.95 correlation between total US spending on science and suicides by hanging over a ten-year window, or a high correlation between imports of crude oil into Europe and church attendance in Australia. Such associations possess zero theoretical or physical plausibility.

3. Reverse Causality and Mutual Interaction

Even when a genuine causal relationship exists, correlation cannot identify the direction of causation. Does high advertising spending cause high corporate profits, or do highly profitable firms possess abundant surplus cash that enables lavish advertising budgets? In economics, variables frequently operate in mutual feedback loops (e.g., price and demand; money supply and price levels).

2.2 Criteria for Establishing Genuine Empirical Causation

To establish that Variable X causes Variable Y, business researchers must demonstrate three non-negotiable empirical conditions beyond mathematical correlation: (a) **Concomitant Variation:** A verifiable statistical association between X and Y; (b) **Temporal Precedence:** The cause (X) must chronologically precede the effect (Y) in time; and (c) **Non-Spuriousness:** All conceivable confounding third variables must be controlled or ruled out through randomized controlled experiments or multivariate econometric specifications.

Chapter 3: Comprehensive Taxonomy of Correlation Types

DegreeLive

Correlation is classified across three distinct operational dimensions:

3.1 Classification by Direction of Movement

- **Positive (Direct) Correlation:** When two variables move in the identical direction. An increase in the value of one variable is accompanied by an increase in the other, or a decrease in one is accompanied by a decrease in the other.

  Examples: Advertising expenditure and sales revenue; height and weight; household income and consumer expenditure; temperature and electric air-conditioning power consumption.

- **Negative (Inverse) Correlation:** When two variables move in opposite directions. An increase in the value of one variable is accompanied by a decrease in the other, and vice versa.

  Examples: Price and consumer demand of a normal good; employee training hours and factory accident rates; production automation investment and unit labor costs; altitude and atmospheric temperature.

- **Zero (Uncorrelated) Association:** When movements in one variable exhibit no predictable pattern or relationship with movements in the other ($r = 0$$).

  Examples: Employee shoe size and intelligence quotient (IQ); corporate share price and lunar phases.

3.2 Classification by Constancy of the Ratio of Change

- **Linear Correlation:** When the ratio of change between two variables remains perfectly constant throughout the range of observations. If for every 10% increase in advertising expenditure, sales revenue consistently expands by exactly 5%, the relationship is linear. When plotted on a Cartesian graph, a linear relationship forms a straight line ($Y = a + bX$$).
- **Non-Linear (Curvilinear) Correlation:** When the ratio of change between two variables does not remain constant, fluctuating across different operating ranges. In agriculture, applying chemical fertilizer initially boosts crop yields rapidly; beyond an optimal threshold, additional fertilizer yields diminishing returns, and extreme application eventually poisons the soil, reducing crop output. When plotted, curvilinear data forms a curved trajectory (parabola, exponential curve, or sigmoid curve).

3.3 Classification by Number of Variables Analyzed

- **Simple Correlation:** The study of the relationship between exactly two variables (e.g., rainfall and wheat output).
- **Partial Correlation:** The study of the relationship between two specific variables while holding the effects of one or more other related variables constant (e.g., examining the relationship between advertising and sales while statistically controlling for competitor pricing).
- **Multiple Correlation:** The simultaneous study of the joint relationship between three or more variables (e.g., examining how crop yield is determined jointly by rainfall, fertilizer application, soil quality, and hours of sunshine).

Chapter 4: Exploratory Graphical Methods: Scatter Diagrams and Correlation Graphs

Degreelive

4.1 The Scatter Diagram Method

The **Scatter Diagram** (or Dot Diagram) is the simplest, most intuitive exploratory tool for visually investigating the presence, direction, and form of relationship between two variables. Independent or causal observations (X) are plotted on the horizontal X-axis, and dependent observations (Y) are plotted on the vertical Y-axis, creating a cloud or pattern of coordinate points (X_i, Y_i).

4.2 Visual Diagnostic Interpretation of Scatter Patterns

1. **Perfect Positive Correlation ($r = +1.0$):** All plotted points fall precisely upon a straight line sloping upward from bottom-left to top-right.
2. **Perfect Negative Correlation ($r = -1.0$):** All plotted points fall precisely upon a straight line sloping downward from top-left to bottom-right.
3. **High Degree of Positive Correlation ($r \approx +0.8$ to $+0.95$):** The plotted points do not form a strict line but cluster tightly along an upward-sloping diagonal corridor from lower-left to upper-right.
4. **High Degree of Negative Correlation ($r \approx -0.8$ to -0.95):** The plotted points cluster tightly along a downward-sloping diagonal corridor from upper-left to lower-right.
5. **Low Degree of Correlation ($r \approx \pm 0.2$ to ± 0.4):** The points show a faint diagonal tendency but are widely scattered and dispersed across the graph.
6. **Zero (No) Correlation ($r = 0$):** The plotted points form an amorphous circular cloud or are scattered completely at random across the entire graph, showing no discernible upward, downward, or linear pattern.

Managerial Merits and Demerits: Scatter diagrams provide immediate visual clarity, quickly exposing non-linear patterns and extreme outlier anomalies that would distort mathematical formulas. However, their primary limitation is that they do not supply a precise numerical coefficient; visual interpretation remains subjective.

Chapter 5: Karl Pearson's Product-Moment Coefficient of Correlation

5.1 Definitional Architecture and Theoretical Assumptions

Formulated by Karl Pearson in 1896, the **Product-Moment Coefficient of Correlation** (conventionally denoted by the lower-case letter r) is the preeminent mathematical measure of linear association between two quantitative continuous variables. It is defined as the quotient obtained by dividing the covariance between two variables by the product of their individual standard deviations:

$$r = \frac{\text{Covariance}(X,Y)}{\sigma X \cdot \sigma Y} = \frac{\frac{\sum (X-\bar{X})(Y-\bar{Y})}{N}}{\sqrt{\frac{\sum (X-\bar{X})^2}{N}} \cdot \sqrt{\frac{\sum (Y-\bar{Y})^2}{N}}}$$

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 \cdot \sum (Y-\bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

Where: x = X - \bar{X and y = Y - \bar{Y.

Core Theoretical Assumptions of Pearson's Coefficient:

1. **Bivariate Normal Distribution:** Both variables X and Y are continuous and normally distributed in the underlying population.
2. **Strict Linearity:** The relationship between X and Y is linear. If a relationship is non-linear, Pearson's formula will produce a misleadingly low coefficient, even if the non-linear relationship is 100% deterministic.
3. **Homoscedasticity:** The variance of Y values around the regression line is constant across all levels of X\$.
4. **Absence of Extreme Outliers:** Because Pearson's r relies on squared deviations, anomalous outliers exert disproportionate leverage, pulling the coefficient away from its authentic value.

5.2 Alternative Computational Formulations

1. Direct Raw Data Formulation (Machine Formula)

When arithmetic means are fractional or non-integers, using deviations produces tedious decimals. The raw data formulation avoids computing deviations entirely:

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

2. Short-Cut Method (Deviations from Assumed Means)

Where $d_x = X - A_x$ and $d_y = Y - A_y$, with A_x and A_y representing assumed means:

$$r = \frac{N\sum(d_x d_y) - (\sum d_x)(\sum d_y)}{\sqrt{[N\sum d_x^2 - (\sum d_x)^2][N\sum d_y^2 - (\sum d_y)^2]}}$$

5.3 Mathematical Properties of Pearson's Correlation Coefficient

Pearson's r exhibits five foundational algebraic theorems:

1. **Bounded Strict Limits:** The numerical value of the correlation coefficient lies strictly within the closed interval of negative one and positive one:

$$-1.0 \leq r \leq +1.0$$

2. **Dimensionless and Pure Number:** Pearson's r is an absolute, unit-free dimensionless ratio. Whether variables are measured in rupees, grams, hours, or degrees, the correlation coefficient is directly comparable across datasets.
3. **Symmetry of Correlation:** The correlation between X and Y is mathematically identical to the correlation between Y and X\$:

$$r_{xy} = r_{yx}$$

4. **Complete Invariance to Change of Origin and Scale:** The value of Pearson's r is completely unaffected by adding, subtracting, multiplying, or dividing the data by arbitrary constants. If $u = \frac{X-A}{c1}$ and $v = \frac{Y-B}{c2}$, where $c1 > 0$ and $c2 > 0$:

$$r_{xy} = r_{uv}$$

Computational Power: Large, complex figures can be simplified by subtracting large base values and dividing by common multiples without altering the computed correlation coefficient.

5. **Geometric Mean of Regression Coefficients:** The correlation coefficient is the geometric mean of the two linear regression coefficients (b_{yx} and b_{xy}\$):

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

The algebraic sign of r is identical to the common sign of both regression coefficients.

Chapter 6: Inferential Reliability and Probable Error of the Correlation Coefficient

6.1 Standard Error of the Correlation Coefficient

When correlation is calculated from a random sample of N observations drawn from a bivariate normal population, the calculated sample coefficient (r\$) is subject to sampling fluctuations. The **Standard Error of** r (SE_r\$) quantifies this sampling dispersion:

$$SE_r = \frac{1-r^2}{\sqrt{N}}$$

6.2 Probable Error (PE) and Decision Criteria

In classical statistical theory, the **Probable Error** of the correlation coefficient (PE_r\$) is defined as that value which, when added to and subtracted from the population parameter, encompasses exactly 50% of the sample distribution. It is formulated as:

$$PE_r = 0.6745 \times SE_r = 0.6745 \times \left(\frac{1-r^2}{\sqrt{N}}\right)$$

Statisticians deploy Probable Error to test whether an observed sample correlation coefficient is statistically significant or merely the product of random sampling chance:

Mathematical Condition	Statistical Inference & Decision Rule
$r < PE_r$	**Not Significant:** The correlation is statistically insignificant; there is no evidence of genuine association in the population.
$PE_r < r < 6 \cdot PE_r$	**Inconclusive:** Correlation exists in the sample, but evidence of genuine population association is weak or ambiguous.
$r > 6 \cdot PE_r$	**Highly Significant:** The correlation is statistically significant and dependable; genuine population correlation is definitively established.

> Confidence Limits of Population Correlation $(\rho) = r \pm PE_r$

Chapter 7: Spearman's Rank Correlation Coefficient

7.1 Philosophical Rationale and Operational Scope

Karl Pearson's product-moment coefficient requires continuous, interval-or-ratio numerical data and strictly linear bivariate normality. However, in many corporate, psychological, and behavioral management investigations, variables cannot be measured on quantitative physical scales. Attributes such as managerial leadership quality, advertising visual aesthetics, brand prestige, employee honesty, or beauty contest performance are fundamentally **Qualitative**.

In 1904, Charles Edward Spearman formulated **Spearman's Rank Correlation Coefficient** (conventionally denoted by the Greek letter rho, ρ or R\$). Rank correlation measures the degree of association between two sets of qualitative ranks assigned to individuals or items.

7.2 Computational Formulations across Data Scenarios

1. Case A: Untied Ranks (No Ties / Ranks Not Repeated)

When each item occupies a distinct rank position with zero duplicate scores:

> $$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)}$$
>
> **Where:** $D = R_1 - R_2$ represents the difference between the two ranks assigned to an item, and N is the total number of paired items ranked.

2. Case B: Tied or Repeated Ranks

When two or more individuals receive identical scores, assigning unique ranks is impossible. Statistical theory assigns each tied item the **Average of the Ranks** they would have occupied if separated. For example, if two employees tie for the 3rd and 4th positions, each is assigned the average rank of $\frac{3+4}{2} = 3.5$; the subsequent employee receives rank 5.

When ties occur, Spearman's formula requires adding an algebraic correction factor of $\frac{m(m^2-1)}{12}$ to $\sum D^2$ for every group of tied values, where m is the number of items tied in that specific group:

$$\rho = 1 - \frac{6\left[\sum D^2 + \sum \frac{m_i(m_i^2-1)}{12}\right]}{N(N^2-1)}$$

7.3 Merits, Demerits, and Applications of Rank Correlation

Primary Merits: (1) perfectly suited for ordinal qualitative data; (2) robust against extreme outliers; (3) non-parametric (makes zero assumptions regarding population normality); and (4) simple to compute for modest sample sizes.

Demerits: (a) cannot be applied to continuous grouped frequency distributions; (b) computationally tedious for vast datasets ($N > 30$); and (c) discards raw metric magnitude information, using only relative rank ordering.

Chapter 8: Comprehensive Worked Practical Problems in Correlation Analysis

8.1 Worked Problem 1: Pearson's Correlation Coefficient from Raw Data

A regional consumer electronics brand launched a targeted promotional campaign across eight representative retail districts. Marketing analysts recorded the monthly promotional expenditure (X in lakhs of rupees) and corresponding monthly sales volume (Y in crores of rupees):

District	Promotional Spend X (Rs. L)	Sales Volume Y (Rs. Cr)	x = X - 6	y = Y - 8	x^2	y^2	x · y
1	2	4	-4	-4	16	16	+16
2	3	5	-3	-3	9	9	+9
3	4	7	-2	-1	4	1	+2
4	5	6	-1	-2	1	4	+2
5	6	8	0	0	0	0	0

District	Promotional Spend X (Rs. L)	Sales Volume Y (Rs. Cr)	x = X - 6	y = Y - 8	x^2	y^2	x · y
6	8	10	+2	+2	4	4	+4
7	9	11	+3	+3	9	9	+9
8	11	13	+5	+5	25	25	+25
Total	**Sum X = 48**	**Sum Y = 64**	**Sum x = 0**	**Sum y = 0**	**Sum x^2 = 68**	**Sum y^2 = 68**	**Sum xy = 67**

Step-by-Step Mathematical Resolution:

1. Calculate Arithmetic Means:

$$\bar{X} = \frac{\sum X}{N} = \frac{48}{8} = 6.0 \text{ lakhs}$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{64}{8} = 8.0 \text{ crores}$$

Both means are exact whole integers, enabling computation via actual deviations x and y.

2. Compute Pearson's Product-Moment Coefficient:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} = \frac{67}{\sqrt{68 \times 68}} = \frac{67}{68} \approx +0.9853$$

Managerial Interpretation: The coefficient $r = +0.985$ reflects an exceptionally strong, nearly perfect positive linear correlation between promotional expenditure and sales revenue.

3. Compute Probable Error and Test Statistical Significance:

$$r^2 = (0.9853)^2 = 0.9708 \implies 1 - r^2 = 1 - 0.9708 = 0.0292$$

$$PE_r = 0.6745 \times \left(\frac{1 - r^2}{\sqrt{N}}\right) = 0.6745 \times \left(\frac{0.0292}{\sqrt{8}}\right) = 0.6745 \times \left(\frac{0.0292}{2.8284}\right) \approx 0.00697$$

$$6 \times PE_r = 6 \times 0.00697 = 0.0418$$

Inference: Because $r = 0.9853$ vastly exceeds $6 \times PE_r = 0.0418$, the correlation is **definitively statistically significant** and dependable for corporate budget allocation.

8.2 Worked Problem 2: Spearman's Rank Correlation with Tied Ranks

Eight candidates appearing for an executive sales leadership role were evaluated independently by two senior human resource interviewers, Panel Judge A and Panel Judge B. The interview panel awarded

the following assessment scores (out of 100):

Candidate	Judge A Score	Judge B Score	Rank A (R1)	Rank B (R2)	D = R1 - R2	D^2
1	80	70	2	3	-1	1.00
2	65	55	5.5	6	-0.5	0.25
3	90	85	1	1	0	0.00
4	65	60	5.5	5	+0.5	0.25
5	75	75	3	2	+1	1.00
6	50	45	8	8	0	0.00
7	70	65	4	4	0	0.00
8	60	50	7	7	0	0.00
Total					**Sum D = 0**	**Sum D^2 = 2.50**

Step-by-Step Mathematical Resolution:

1. Analyze Rank Ties:

- In Judge A's evaluations, two candidates share the score of 65. If separated, they would occupy ranks 5 and 6. Therefore, each receives average rank $\frac{5+6}{2} = 5.5$. Number of tied items $m1 = 2$.
- In Judge B's evaluations, there are no ties ($m2 = 0$).

2. Compute the Tie Correction Factor:

$$\text{Correction Factor} = \frac{m1(m1^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{6}{12} = 0.50$$

3. Calculate Spearman's Rank Correlation Coefficient:

$$\text{Adjusted } \sum D^2 = \sum D^2 + 0.50 = 2.50 + 0.50 = 3.00$$

$$N = 8 \implies N(N^2-1) = 8(64-1) = 8(63) = 504$$

$$\rho = 1 - \frac{6 \times 3.00}{504} = 1 - \frac{18}{504} = 1 - 0.0357 = +0.9643$$

Managerial Assessment: The high rank correlation coefficient ($\rho = +0.964$) establishes exceptional inter-rater reliability between both human resource evaluators, confirming high objectivity and consensus in candidate scoring.

8.3 Worked Problem 3: Pearson's Correlation via the Short-Cut (Assumed Mean) Method

An automotive component manufacturer records the machine operating temperature (X in degrees Celsius) and the corresponding lubricant viscosity loss (Y in Centistokes) across seven industrial test cycles:

Cycle	Temperature X	Viscosity Loss Y	dx = X - 70	dy = Y - 25	dx^2	dy^2	dx · dy
1	65	21	-5	-4	25	16	+20
2	66	23	-4	-2	16	4	+8
3	68	24	-2	-1	4	1	+2
4	70 (Ax)	25 (Ay)	0	0	0	0	0
5	72	26	+2	+1	4	1	+2
6	74	28	+4	+3	16	9	+12
7	75	28	+5	+3	25	9	+15
Total			**Sum dx = 0**	**Sum dy = 0**	**Sum dx^2 = 90**	**Sum dy^2 = 40**	**Sum dx·dy = 59**

Step-by-Step Mathematical Resolution:

$$r = \frac{N\sum(dx\,dy) - (\sum dx)(\sum dy)}{\sqrt{[N\sum d_x^2 - (\sum dx)^2][N\sum d_y^2 - (\sum dy)^2]}}$$

Because $\sum dx = 0$ and $\sum dy = 0$: $\quad r = \frac{7(59) - (0)(0)}{\sqrt{[7(90) - 0][7(40) - 0]}}$

$$r = \frac{413}{\sqrt{630 \times 280}} = \frac{413}{\sqrt{1,76,400}} = \frac{413}{420} = +0.9833$$

Engineering Insight: There is a direct, near-perfect positive linear relationship ($r = +0.983$) between operating temperature and lubricant degradation, establishing the operational imperative for auxiliary cooling systems.

Chapter 9: The Coefficient of Determination and Business Forecasting Models

9.1 The Coefficient of Determination (r^2)

While Karl Pearson's correlation coefficient (r\$) measures the direction and strength of linear association, it does not directly communicate the proportion of variation shared between the two variables. To quantify explanatory power, econometricians and business analysts compute the **Coefficient of Determination** (r^2\$).

The coefficient of determination is defined as the square of the correlation coefficient. It represents the proportion of the total variation in the dependent variable (Y\$) that is mathematically explained by or associated with variation in the independent variable (X\$):

> $$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$
>
> **Practical Illustration:** If the correlation between advertising expenditure and sales revenue is $r = +0.80$, the coefficient of determination is $r^2 = (0.80)^2 = 0.64$ (or **64%**). This proves that 64% of the total fluctuation in corporate sales revenue is accounted for by changes in advertising spending.

9.2 The Coefficient of Non-Determination (k^2) and Alienation (k)

The residual unexplained variation is quantified through complementary metrics:

- **Coefficient of Non-Determination (k^2):** Represents the proportion of total variation in Y that is left unexplained by X\$, attributable to other unobserved factors or random noise:

  > $$k^2 = 1 - r^2 = \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$
  >
  > In the advertising example above, $k^2 = 1 - 0.64 = 0.36$ (or **36%**), indicating that 36% of sales variance is driven by non-advertising factors (pricing, competitor actions, distribution reach).

- **Coefficient of Alienation (k):** Defined as the square root of the coefficient of non-determination:

  > $$k = \sqrt{1 - r^2}$$
  >
  > Measures the lack of association or degree of estrangement between the two variables.