Chapter 6: Correlation

Exercises — Correlation

Q1 The unit of correlation coefficient between height in feet and weight in kgs is
(i) kg/feet
(ii) percentage
(iii) non-existent

SOLUTION

The correlation coefficient r is defined as: $r = \Sigma(dx \cdot dy) / \text{sqrt}(\Sigma dx^2 \cdot \Sigma dy^2)$.
In this formula, the units of dx and dy appear in both numerator and denominator and cancel out completely.
r is therefore a pure number — it has no units regardless of the units of the original variables.

Answer: (iii) non-existent — the correlation coefficient is a dimensionless pure number.

Q2 The range of simple correlation coefficient is
(i) 0 to infinity
(ii) minus one to plus one
(iii) minus infinity to infinity

SOLUTION

By definition, Karl Pearson's simple correlation coefficient $r = \Sigma(dx \cdot dy) / \text{sqrt}(\Sigma dx^2 \cdot \Sigma dy^2)$.
Applying the Cauchy-Schwarz inequality: $[\Sigma(dx \cdot dy)]^2 \leq \Sigma dx^2 \cdot \Sigma dy^2$.
Therefore $r^2 \leq 1$, which means $-1 \leq r \leq 1$.
r = +1 indicates perfect positive correlation; r = –1 indicates perfect negative correlation; r = 0 indicates no linear relationship.

Answer: (ii) minus one to plus one — the correlation coefficient always lies between –1 and +1, inclusive.

Q3 If rxy is positive the relation between X and Y is of the type
(i) When Y increases X increases
(ii) When Y decreases X increases
(iii) When Y increases X does not change

SOLUTION

A positive correlation coefficient (rxy > 0) means both variables move in the same direction.
When X increases, Y also tends to increase, and when X decreases, Y also tends to decrease.
Option (i) captures this: when Y increases, X also increases.

Answer: (i) When Y increases, X increases — positive correlation means both variables move in the same direction.

Q4 If rxy = 0 the variable X and Y are
(i) linearly related
(ii) not linearly related
(iii) independent

SOLUTION

The correlation coefficient r measures the degree of linear association between two variables.
If rxy = 0, it means there is no linear relationship between X and Y.
However, a non-linear relationship may still exist (e.g., $Y = X^2$). Therefore X and Y are not necessarily independent.
Option (iii) 'independent' is too strong a conclusion — zero correlation only rules out linear dependence.

Answer: (ii) not linearly related — rxy = 0 means no linear association, but a non-linear relationship may still exist.

Q5 Of the following three measures which can measure any type of relationship?
(i) Karl Pearson's coefficient of correlation
(ii) Spearman's rank correlation
(iii) Scatter diagram

SOLUTION

Karl Pearson's coefficient measures only the degree of linear relationship.
Spearman's rank correlation measures monotonic (both linear and non-linear) relationships by using ranks, making it suitable for any type of relationship whose direction is clear.
Scatter diagram is a graphical tool — it can visually indicate any type of relationship but does not provide a numerical measure.
Among numerical measures, Spearman's rank correlation is applicable to a wider range of relationship types including non-linear ones.

Answer: (iii) Scatter diagram — it can depict any type of relationship (linear, non-linear, positive, negative, no correlation) visually.

Q6 If precisely measured data are available the simple correlation coefficient is
(i) more accurate than rank correlation coefficient
(ii) less accurate than rank correlation coefficient
(iii) as accurate as the rank correlation coefficient

SOLUTION

Karl Pearson's simple correlation coefficient uses the actual numerical values of all observations, extracting full information from the data.
Spearman's rank correlation uses only the ranks (order) of values, discarding the actual magnitudes and thus losing some information.
When data are precisely measured, Karl Pearson's coefficient utilises all available information

and is therefore more accurate.

Answer: (i) more accurate than rank correlation coefficient — when precise data are available, using actual values (Pearson's r) is more informative than using only ranks.

Q7 Why is r preferred to covariance as a measure of association?

SOLUTION

Covariance is defined as: Cov(X, Y) = Σ(dx·dy) / n, where dx = X – X̄ and dy = Y – .

Covariance depends on the units of measurement of X and Y. If the units change, the value of covariance changes.

For example, if height is measured in feet instead of inches, the covariance changes, making comparison difficult.

The correlation coefficient r = Cov(X, Y) / (σ_x · σ_y) standardises covariance by dividing by the product of the standard deviations.

This makes r dimensionless (unit-free) and bounded between –1 and +1.

r allows comparison across different datasets and variables regardless of their units of measurement.

Covariance only shows direction of association; r shows both direction and strength on a standardised scale.

Answer: r is preferred because it is dimensionless (unit-free) and always lies between –1 and +1, making it comparable across datasets. Covariance depends on units and has no fixed range, so its magnitude is difficult to interpret.

Q8 Can r lie outside the –1 and 1 range depending on the type of data?

SOLUTION

No. The correlation coefficient r is mathematically bounded between –1 and +1 for any dataset.

This follows from the Cauchy-Schwarz inequality: $[\Sigma(dx \cdot dy)]^2 \leq \Sigma dx^2 \cdot \Sigma dy^2$.

Dividing both sides: $r^2 = [\Sigma(dx \cdot dy)]^2 / (\Sigma dx^2 \cdot \Sigma dy^2) \leq 1$.

Therefore $-1 \leq r \leq 1$ always holds, regardless of the type of data.

If a computed value falls outside this range, it signals an arithmetic error in calculation.

Answer: No. r always lies between –1 and +1 for any type of data. A value outside this range indicates a calculation error.

Q9 Does correlation imply causation?

SOLUTION

No. Correlation measures the degree and direction of linear association between two variables; it does not establish cause and effect.

A high positive or negative r between X and Y could arise because: (a) X causes Y, (b) Y

causes X, (c) a third variable Z causes both X and Y, or (d) the correlation is purely coincidental (spurious).

Example: there is a high positive correlation between the number of fire engines at a fire site and the damage caused. This does not mean fire engines cause the damage — both are caused by the severity of the fire.

Another example: ice cream sales and drowning deaths are positively correlated, but only because both increase in summer — neither causes the other.

Establishing causation requires controlled experiments or domain knowledge, not just a high value of r.

Answer: No. Correlation does not imply causation. A high r shows association, not that one variable causes the other — a third variable may drive both, or the correlation may be spurious.

Q10 When is rank correlation more precise than simple correlation coefficient?

SOLUTION

Rank correlation (Spearman's r_s) is more precise than Karl Pearson's r in the following situations:

1. When the data cannot be measured precisely — e.g., attributes like beauty, honesty, or intelligence that can only be ranked, not quantified.
2. When the data contain extreme outliers — Pearson's r is heavily influenced by extreme values, while rank correlation is not, because ranks are bounded.
3. When the relationship is monotonic but non-linear — for example, scatter diagrams showing a curvilinear trend. Rank correlation captures the direction of such relationships; Pearson's r is designed for linear relationships only.
4. When measuring instruments are not available — e.g., ranking students by height or weight without a measuring rod or weighing machine.

Answer: Rank correlation is more precise when data cannot be precisely measured (qualitative attributes), when extreme values are present, or when the relationship is non-linear (monotonic).

Q11 Does zero correlation mean independence?

SOLUTION

No. Zero correlation (r = 0) means there is no LINEAR relationship between the two variables. However, the variables may still be related through a non-linear relationship.

Example: Let X take values –3, –2, –1, 0, 1, 2, 3 and $Y = X^2$. Then r = 0 (a symmetric non-linear relationship), but Y is completely determined by X — they are far from independent.

Zero correlation rules out linear dependence only; it does not establish statistical independence.

Answer: No. Zero correlation (r = 0) means no linear relationship, but a non-linear relationship may still exist. Independence is a stronger condition than zero correlation.

Q12 Can simple correlation coefficient measure any type of relationship?

SOLUTION

No. Karl Pearson's simple correlation coefficient r measures only the degree of LINEAR relationship between two variables.

If the relationship is non-linear (e.g., quadratic, exponential), r will not capture it correctly and may even give a misleading low value of r.

For non-linear but monotonic relationships, Spearman's rank correlation r_s is more appropriate.

For visualising any type of relationship, a scatter diagram is useful.

Therefore, Pearson's r is limited to detecting and measuring linear associations.

Answer: No. Simple correlation coefficient (Pearson's r) measures only linear relationships. For non-linear relationships, rank correlation or scatter diagram should be used.

Q13 Collect the price of five vegetables from your local market every day for a week. Calculate their correlation coefficients. Interpret the result.

SOLUTION

This is a data-collection activity. Sample approach:

Step 1: Record daily prices of five vegetables (e.g., tomato, potato, onion, peas, brinjal) for 7 days.

Step 2: For any two vegetables X and Y, compute: $\bar{X}$, , $\Sigma(dx \cdot dy)$, Σdx^2, Σdy^2, where $dx = X - \bar{X}$ and $dy = Y -$.

Step 3: Apply the formula: $r = \Sigma(dx \cdot dy) / \text{sqrt}(\Sigma dx^2 \cdot \Sigma dy^2)$.

Interpretation: If r is close to +1, the two vegetables' prices tend to rise and fall together (common supply-side factors). If r is close to –1, they move in opposite directions. If r is near 0, there is no linear relationship between their prices.

Note: Seasonal vegetables often show high positive inter-correlations because their supply depends on common weather and transport conditions.

Answer: Activity-based question. Students should collect real data, compute r using the deviation formula, and interpret: r near ±1 means strong correlation; r near 0 means weak or no linear association.

Q14 Measure the height of your classmates. Ask them the height of their benchmate. Calculate the correlation coefficient of these two variables. Interpret the result.

SOLUTION

This is a data-collection activity. Sample approach:

Step 1: Let X = student's own height, Y = height of benchmate (as reported). Record n pairs.

Step 2: Compute $\bar{X} = \Sigma X/n$, $= \Sigma Y/n$.

Step 3: Compute deviations: $dx = X - \bar{X}$, $dy = Y -$.

Step 4: Apply: $r = \Sigma(dx \cdot dy) / \text{sqrt}(\Sigma dx^2 \cdot \Sigma dy^2)$.

Interpretation: If r is positive and high, taller students tend to sit next to taller benchmates (perhaps seating is height-based). If r is near 0, seating is random with respect to height.
A positive r here does not imply causation — the correlation, if found, would reflect a common cause such as the teacher's seating arrangement.

Answer: Activity-based question. Compute r from collected data. A positive r suggests taller students tend to have taller benchmates; r near 0 means seating is height-independent.

Q15 List some variables where accurate measurement is difficult.

SOLUTION

Accurate measurement is difficult for qualitative or subjective variables, including:
1. Intelligence — IQ tests give a proxy but do not capture all dimensions of intelligence.
2. Beauty — subjective; varies across individuals and cultures.
3. Honesty — cannot be directly quantified.
4. Bravery or courage — difficult to put on a numerical scale.
5. Job satisfaction — measured via surveys but is inherently subjective.
6. Creativity — no universally accepted metric.
7. Pain — patients rate it on a scale but it is inherently personal and non-comparable.
8. Happiness or well-being — survey proxies exist but are not precise.

For such variables, ranking (Spearman's method) is preferred over Pearson's correlation, since exact measurement is not possible.

Answer: Variables difficult to measure precisely include: intelligence, beauty, honesty, bravery, job satisfaction, creativity, pain, and happiness. These are better ranked than measured numerically.

Q16 Interpret the values of r as 1, –1 and 0.

SOLUTION

r = +1 (Perfect Positive Correlation): All data points lie exactly on a straight line with a positive slope. As X increases, Y increases in exact proportion. Example: if Y = 2X for all pairs, r = 1.
r = –1 (Perfect Negative Correlation): All data points lie exactly on a straight line with a negative slope. As X increases, Y decreases in exact proportion. Example: if Y = –2X + 10 for all pairs, r = –1.
r = 0 (No Linear Correlation): There is no linear relationship between X and Y. The scatter diagram shows no discernible upward or downward linear trend. Note: a non-linear relationship may still exist.
In practice, r = ±1 almost never occurs; r values closer to ±1 indicate stronger linear association, while values near 0 indicate a weak or absent linear relationship.

Answer: r = 1: perfect positive linear correlation (all points on an upward line). r = –1: perfect negative linear correlation (all points on a downward line). r = 0: no linear relationship between the variables.

Q17 Why does rank correlation coefficient differ from Pearsonian correlation coefficient?

SOLUTION

The two coefficients differ for the following reasons:

1. Data used: Pearson's r uses actual numerical values; Spearman's r_s uses only the ranks (order) of the values.
2. Information loss: By converting values to ranks, Spearman's method loses information about the actual magnitudes of differences between values. Pearson's method retains all this information.
3. Relationship type: Pearson's r measures only linear association; Spearman's r_s measures any monotonic (consistently increasing or decreasing) association, including non-linear ones.
4. Effect of outliers: Pearson's r is sensitive to extreme values because it uses actual magnitudes. Spearman's r_s is not, since extreme values are merely given extreme ranks.
5. Result: Since r_s uses less information than r, it is generally less precise when precise measurement is available. r_s ≤ r when data are precisely measured.
6. Equal only when: r_k = r only if the first differences of the ranked series are constant — a condition rarely met in practice.

Answer: Rank correlation differs from Pearson's correlation because it uses only ranks (losing magnitude information), is unaffected by outliers, and captures monotonic (not just linear) relationships. As a result, r_s is generally less precise than r when precise data are available.

Q18 Calculate the correlation coefficient between the heights of fathers in inches (X) and their sons (Y):
X: 65, 66, 57, 67, 68, 69, 70, 72
Y: 67, 56, 65, 68, 72, 72, 69, 71

SOLUTION

n = 8. Karl Pearson's coefficient (direct method): r = (n*SXY - SX*SY) / sqrt[(n*SX^2 - (SX)^2)(n*SY^2 - (SY)^2)]

Computation table:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	56	4356	3136	3696
57	65	3249	4225	3705
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830

X	Y	X^2	Y^2	XY
72	71	5184	5041	5112

SX = 534, SY = 540, SXY = 36118, SX^2 = 35788, SY^2 = 36644
r = (8x36118 - 534x540) / sqrt[(8x35788 - 534^2)(8x36644 - 540^2)]
= (288944 - 288360) / sqrt[(286304 - 285156)(293152 - 291600)]
= 584 / sqrt(1148 x 1552) = 584 / 1334.8 = 0.44

Note: the NCERT answer key prints r = 0.603, but the data as given in the textbook works out to r = 0.44 (the sons'/fathers' height figures carry a suspected misprint in the key). The value 0.44 is the correct coefficient for the printed data.

Answer: r = 0.44 (a low positive correlation for the data as printed). The NCERT key states 0.603; see the note on the suspected data misprint.

Q19 Calculate the correlation coefficient between X and Y and comment on their relationship:
X: –3, –2, –1, 1, 2, 3
Y: 9, 4, 1, 1, 4, 9

SOLUTION

n = 6. Using: $r = (n\cdot\Sigma XY – \Sigma X\cdot\Sigma Y) / \text{sqrt}[(n\cdot\Sigma X^2 – (\Sigma X)^2)(n\cdot\Sigma Y^2 – (\Sigma Y)^2)]$

Computation table:

X	Y	X^2	Y^2	XY
–3	9	9	81	–27
–2	4	4	16	–8
–1	1	1	1	–1
1	1	1	1	1
2	4	4	16	8
3	9	9	81	27
Σ		28	196	0

ΣX = 0, ΣY = 28, ΣXY = 0
Numerator = 6 × 0 – 0 × 28 = 0
r = 0 / sqrt[...] = 0

Answer: r = 0. Although r = 0 (no linear relationship), $Y = X^2$ shows a perfect non-linear (quadratic) relationship. This illustrates that zero correlation does not imply independence — a strong non-linear relationship exists.

Q20 Calculate the correlation coefficient between X and Y and comment on their relationship:
X: 1, 3, 4, 5, 7, 8
Y: 2, 6, 8, 10, 14, 16

SOLUTION

n = 6. Observe that Y = 2X for every pair. We verify and compute r formally.

Computation table:

X	Y	X^2	Y^2	XY
1	2	1	4	2
3	6	9	36	18
4	8	16	64	32
5	10	25	100	50
7	14	49	196	98
8	16	64	256	128
Σ		164	656	328

ΣX = 28, ΣY = 56
Numerator = 6 × 328 – 28 × 56 = 1968 – 1568 = 400
Den_A = 6 × 164 – 28^2 = 984 – 784 = 200
Den_B = 6 × 656 – 56^2 = 3936 – 3136 = 800
Denominator = sqrt(200 × 800) = sqrt(160000) = 400
r = 400 / 400 = 1

Answer: r = 1. There is perfect positive linear correlation between X and Y. Since Y = 2X for all pairs, every increase in X is accompanied by an exactly proportional increase in Y.