Thirty Worked Answers: Correlation and Causation

Topic 77 | Fictional research data | November 2026

Level 1

1. Define correlation

Question: What is a statistical correlation?

Hint: Think about co-variation rather than intervention.

Worked answer: It describes a statistical association between variables, characterized by direction, form and strength. It does not itself identify causal influence.

2. Define causation

Question: What question distinguishes causation from association?

Hint: Imagine intervening on a proposed cause.

Worked answer: A causal question asks whether a specified intervention on X would change Y under a defined context, with research design and assumptions supporting the comparison.

3. Main difference

Question: Can a correlation alone establish what would happen if X were changed?

Hint: Correlation is observational.

Worked answer: No. Co-variation does not establish a counterfactual intervention effect. Confounding, reverse causation and selection are alternative explanations.

4. Positive correlation

Question: A scatterplot slopes upward. Describe its linear association.

Hint: Compare higher X to higher Y.

Worked answer: It suggests positive correlation: higher X tends to accompany higher Y. A scatterplot and calculated r help assess strength.

5. Negative correlation

Question: As X increases, Y tends to decrease. Which direction is this?

Hint: Consider the sign of r.

Worked answer: Negative correlation, with r less than zero if there is negative linear association. This does not by itself mean the relationship is harmful.

6. Zero Pearson correlation

Question: For X=-2,-1,0,1,2 and Y=X squared, what is Pearson r?

Hint: Look for symmetrical U-shape.

Worked answer: Pearson r=0 exactly by symmetry; a deterministic nonlinear relationship exists, so zero linear correlation does not mean independence.

7. Read a scatterplot

Question: What four features should be examined before interpreting r?

Hint: Shape, subgroups, outliers and scale.

Worked answer: Examine direction, shape/linearity, outliers, clusters/subgroups and scale; a summary coefficient can hide structure.

8. Perfect positive

Question: Interpret r=+1 for paired quantitative variables.

Hint: All points on one increasing line.

Worked answer: Perfect positive linear association: all pairs fit an increasing line. It does not prove X causes Y.

9. Perfect negative

Question: Interpret r=-1.

Hint: All points on one decreasing line.

Worked answer: Perfect negative linear association: all paired points fit a decreasing line, without identifying causal direction.

10. Strong not causal

Question: Why does r=.95 alone not prove an intervention effect?

Hint: Consider common causes.

Worked answer: An unmeasured common cause, selection or reverse influence could generate a strong r. Causal evidence requires a justified design and assumptions.

Level 2

11. Two means

Question: For X=1,2,3,4,5 and Y=2,4,6,8,10 calculate means.

Hint: Sum each divided by five.

Worked answer: X mean=3 and Y mean=6. These are the centers used for the Pearson deviation table.

12. Crossproducts

Question: Calculate sum of (x-3)(y-6) for the five positive pairs.

Hint: Deviation vectors -2,-1,0,1,2 and -4,-2,0,2,4.

Worked answer: Products 8,2,0,2,8 sum to 20; sums of squared X and Y deviations are 10 and 40.

13. Pearson calculation

Question: Compute Pearson r from sums 20,10,40.

Hint: 20/sqrt(10*40).

Worked answer: r=20/sqrt(400)=20/20=+1. The coefficient has no units and is symmetric.

14. Moderate association

Question: What does observed r=.60 mean?

Hint: Avoid universal bands.

Worked answer: A positive linear association of magnitude .60 in the analysed data. Its practical strength depends on context, sample, measurement and scatterplot.

15. Confounder

Question: How might prior preparation confound study hours and final grades?

Hint: Prior preparation may affect both.

Worked answer: A hypothetical graph prior preparation -> study hours and -> final score can produce part of the observed association. It is a causal assumption, not proven by r.

16. Reverse causation

Question: A sample shows extra tutoring and lower grades. Does tutoring cause lower grades?

Hint: Who gets offered tutoring?

Worked answer: No such conclusion follows. Students with earlier academic difficulty may be selected for tutoring. Measure earlier grades and assignment mechanisms.

17. Spurious relationship

Question: Why could two unrelated variables trending upward over time appear correlated?

Hint: Shared time trend.

Worked answer: Common trends, seasonality or other factors can produce large raw correlations without a direct causal pathway. Time series need design-aware evaluation.

18. Pearson versus Spearman

Question: How do Pearson and Spearman differ?

Hint: Original values versus ranks.

Worked answer: Pearson measures linear association between quantitative values; Spearman is Pearson correlation of ranks and reflects monotonic association, with ties handled by rank conventions.

19. Outlier influence

Question: In x=1..10,y=2x, replace final y=20 by -15. What happens to r?

Hint: Recalculate both coefficients.

Worked answer: Original r=1; altered r approximately 0.0291. One changed pair almost eliminates the linear correlation, though other pairs remain aligned.

20. Nonlinear dependence

Question: How can r=0 coexist with y=x squared?

Hint: Check symmetry and shape.

Worked answer: For symmetric X=-2,-1,0,1,2, positive and negative crossproducts cancel; the U-shaped deterministic dependence remains.

Level 3

21. Interpret the main data

Question: What is the sample size and Pearson study-hours/final-score coefficient?

Hint: Use the fictional CSV and CORREL.

Worked answer: N=240 distinct fictional learners, Pearson r=0.7567. Describes observed association, not a causal estimate.

22. Stratified comparisons

Question: After splitting by median prior score 58.7, what are hours/score correlations?

Hint: Compute within each 120-record group.

Worked answer: Lower prior preparation n=120: r=0.6711, higher prior preparation n=120: r=0.6550. Neither conditional correlation identifies a causal effect.

23. Simpson reversal

Question: Find Method A and B overall pass rates given A 43/100 and B 77/100.

Hint: Divide passed by total.

Worked answer: A 43%, B 77%. Despite A being higher in easy (95 vs 90) and hard (30 vs 25) groups, B is higher overall due to unequal difficulty composition.

24. Sampling versus assignment

Question: Explain random sampling versus random treatment assignment.

Hint: Selection versus allocation.

Worked answer: Sampling chooses which units enter a study and may help generalization; assignment chooses intervention conditions among included units and supports causal comparisons.

25. Evaluate randomized trial

Question: Can random assignment alone guarantee valid causal results?

Hint: Consider follow-up losses and compliance.

Worked answer: No. Noncompliance, attrition, interference, poor measurement, deviations and external validity need evaluation.

26. Confounder, mediator, collider

Question: Given C->X,C->Y; X->M->Y; X->S<-Y, identify C,M,S.

Hint: Draw each causal pattern.

Worked answer: C is a hypothesized common cause, M a mediator, S a collider. Adjusting for every variable indiscriminately may introduce bias.

27. Selection bias

Question: Why might surveying only students who voluntarily join a research club bias a campus estimate?

Hint: Selection is nonrandom.

Worked answer: Club volunteers can differ systematically in prior achievement, interest and study time; the observed association may not generalize to all learners.

28. Statistical significance

Question: If Pearson test reports p=.01, does that prove study hours cause score improvement?

Hint: A p-value conditions on null model.

Worked answer: No. Small p suggests observed test statistic is unusual under the null and assumptions; it does not eliminate confounding or determine causal direction.

29. Better study design

Question: Propose a design to evaluate an optional practice resource.

Hint: Specify intervention and assignment.

Worked answer: Recruit eligible participants ethically, randomly assign practice resource versus suitable comparison where feasible, measure outcomes consistently, monitor attrition and estimate group effects with uncertainty.

30. Evidence-based conclusion

Question: Write a responsible summary of the 240-student synthetic project.

Hint: Include r, causal limitation and possible next steps.

Worked answer: The fictional observational dataset shows study hours and scores associated (r=0.757). Prior preparation and other generated factors affect the data. It cannot quantify a real-world causal effect; a properly designed study would be needed.