Crash Course in Causality: 15 Quiz Questions

Name: Mubin Modi Nuid:002822945

Question 1: Correlation vs. Causation

Which of the following statements accurately describes the difference between correlation and causation?

A. Correlation implies that one variable directly effects the change in another variable.

B. Causation can be established simply by observing that two variables move in the same direction over time.

C. Correlation means two variables are statistically associated, while causation means one variable directly produces a change in the other.

D. There is no practical difference; the terms can be used interchangeably in data science.

Correct Answer: C

Explanations:

- **A is incorrect:** This describes causation, not correlation. Correlation only implies association, not a direct effect.
- **B is incorrect:** Observation of co-movement (correlation) is not enough to establish causation due to potential confounders or spurious relationships.
- **C is correct:** This accurately distinguishes the two: correlation is a statistical relationship (they move together), while causation is a mechanistic relationship (one drives the other).
- **D is incorrect:** Confusing these two concepts is a fundamental error in data science that leads to incorrect conclusions and failed interventions.

Question 2: The Confounder

In a study finding that people who drink more coffee tend to live longer, researchers realized that smoking was a "confounder." What does this mean?

A. Smoking was the actual cause of living longer, not coffee.

B. Smoking is a third variable that is related to both coffee consumption and lifespan, distorting the true relationship between them.

C. Smoking happened after the coffee drinking, making it a mediating variable.

D. Smoking had no statistical relationship with either variable and was irrelevant to the study.

Correct Answer: B

Explanations:

- **A is incorrect:** A confounder doesn't necessarily mean it is the *sole* cause of the outcome, but rather that it biases the estimated effect of the main variable (coffee).
- **B is correct:** A confounder must influence both the independent variable (coffee) and the dependent variable (lifespan), creating a false or distorted association if not controlled for.
- **C is incorrect:** If it happens *after* the cause, it might be a mediator, not a confounder. Confounders generally exist prior to or alongside the exposure.
- **D is incorrect:** By definition, a confounder *must* have a relationship with the other variables to cause bias.

Question 3: Temporal Precedence

Why is temporal precedence essential for establishing causality?

A. It ensures that the data was collected recently.

B. It guarantees that there are no confounding variables in the dataset.

C. It confirms that the cause happened before the effect, as an effect cannot logically happen before its cause.

D. It increases the correlation coefficient between the two variables.

Correct Answer: C

Explanations:

- **A is incorrect:** The recency of data collection is irrelevant to the logical order of events.
- **B is incorrect:** Temporal precedence does not remove confounders; a third variable could still have caused both events in sequence.

- **C is correct:** This is a fundamental logical requirement for causality; time must flow from cause to effect.
- **D is incorrect:** Temporal order does not inherently strengthen the statistical correlation; it only establishes a logical timeline.

Question 4: Spurious Correlations

Ice cream sales and drowning incidents are highly positively correlated. This is a classic example of a spurious relationship because:

A. Eating ice cream causes cramps, leading directly to drowning.

B. Drowning incidents cause people to buy more ice cream for comfort.

C. A third variable, summer heat, causes increases in both ice cream sales and swimming activity (leading to drownings).

D. The data used in these studies is usually faked.

Correct Answer: C

Explanations:

- **A is incorrect:** While a popular myth, there is no strong evidence that ice cream consumption directly causes drowning.
- **B is incorrect:** This violates temporal precedence and logical mechanisms; drownings do not drive ice cream market trends.
- **C is correct:** Summer heat is the confounder that drives both variables simultaneously, creating a statistical correlation with no direct causal link between the two.
- **D is incorrect:** Spurious correlations often arise from strictly real, accurate data; the *interpretation* is what is flawed, not the data itself.

Question 5: The Counterfactual

What is a "counterfactual" in the context of causal inference?

A. A fact that contradicts established knowledge.

B. The "what if" scenario—what would have happened to the same individual if they had not received the treatment.

C. Data that has been falsified or corrupted during collection.

D. The opposite of the intended outcome of an experiment.

Correct Answer: B

Explanations:

- **A is incorrect:** This is a general definition of something counter-factual in normal speech, but not the technical definition in causal inference.
- **B is correct:** Causal inference often aims to estimate this unobservable state (e.g., "What would the patient's health be today if they *hadn't* taken the pill?").
- **C is incorrect:** Counterfactuals are theoretical constructs needed for analysis, not "fake" or corrupted data.
- **D is incorrect:** It does not refer to a bad outcome, but an alternative unobserved reality.

Question 6: Randomized Controlled Trials (RCTs)

Why are RCTs considered the "gold standard" for establishing causality?

A. They are cheaper and faster to run than observational studies.

B. Randomization ensures that treatment and control groups are identical in every way, including unobserved confounders, on average.

C. They guarantee a higher correlation coefficient than other methods.

D. They do not require any statistical analysis to interpret.

Correct Answer: B

Explanations:

- **A is incorrect:** RCTs are often far more expensive, time-consuming, and sometimes unethical to run compared to observational studies.
- **B is correct:** By randomly assigning treatment, confounders are equally distributed between groups, isolating the treatment as the only systematic difference.
- **C is incorrect:** They don't guarantee higher correlation; they guarantee that the correlation observed is more likely to be causal.
- **D is incorrect:** RCTs still require rigorous statistical analysis to ensure differences aren't due to random chance.

Question 7: Observational Data

What is the primary weakness of using standard observational data (like typical business sales records) for causal claims?

A. The datasets are usually too small to find statistical significance.

B. The treatment (e.g., a marketing campaign) was not randomly assigned, leading to potential selection bias and confounding.

C. Observational data cannot be used for regression analysis.

D. It is impossible to find correlations in observational data.

Correct Answer: B

Explanations:

- **A is incorrect:** Observational datasets can be massive ("Big Data") and still suffer from causal limitations.
- **B is correct:** In the real world, treatments are chosen for reasons (e.g., running ads only in rich cities), which introduces confounders that bias naive comparisons.
- **C is incorrect:** Regression is frequently applied to observational data, specifically to try and control for these issues.
- **D is incorrect:** Correlations are very easy to find in observational data; the hard part is proving they are causal.

Question 8: Causal Mechanisms

Establishing a "mechanism" helps confirm causality by:

A. Increasing the p-value of the analysis.

B. Providing a plausible theoretical explanation for *how* the cause leads to the effect.

C. Ensuring that the data follows a normal distribution.

D. Removing the need for any quantitative data.

Correct Answer: B

Explanations:

- **A is incorrect:** Mechanisms don't increase p-values (and you generally want low p-values anyway); they are qualitative or theoretical supports.
- **B is correct:** If X causes Y, there should be a logical series of steps explaining that connection (e.g., Smoking -> lung damage -> cancer).
- **C is incorrect:** The distribution of the data (normal or otherwise) is a statistical concern, not a mechanical one.
- **D is incorrect:** A mechanism explains the *why*, but you still need quantitative data to confirm the *if* and *how much*.

Question 9: Reverse Causality

A study finds that companies with high marketing budgets also have high revenue. A researcher suggests "reverse causality" might be at play. What does she mean?

A. Marketing and revenue are completely unrelated.

B. High revenue might be causing the high marketing budgets, rather than the other way around.

C. A third variable is causing both marketing and revenue to increase.

D. The correlation is negative instead of positive.

Correct Answer: B

Explanations:

- **A is incorrect:** Reverse causality admits there is a causal relationship, just in the opposite direction of the initial hypothesis.
- **B is correct:** It suggests the direction of the arrow is Y -> X instead of X -> Y (richer companies can afford more marketing).
- **C is incorrect:** This describes confounding, not reverse causality.
- **D is incorrect:** Reverse causality refers to the *direction of influence*, not the sign (positive/negative) of the correlation.

Question 10: Controlling for Variables

In a regression model, what does it mean to "control for" a variable like Customer Age?

A. To remove all customers over a certain age from the dataset.

B. To mathematically isolate the effect of other variables by holding Age constant in the analysis.

C. To ensure that Age is the only variable affecting the outcome.

D. To randomly assign ages to different customers in the dataset.

Correct Answer: B

Explanations:

- **A is incorrect:** This is filtering or subsetting the data, not controlling for it in a statistical model.
- **B is correct:** "Controlling for" means the model estimates the effect of X on Y *as if* all groups had the same age.
- **C is incorrect:** Controlling for age is meant to *remove* its confounding influence, not make it the only influential variable.
- **D is incorrect:** You cannot randomly assign demographic traits like age to people in an analysis.

Question 11: Selection Bias

If a voluntary survey about a new product is only answered by people who loved the product, the results will suffer from:

A. Selection Bias.

B. Reverse Causality.

C. Temporal Precedence.

D. Random Noise.

Correct Answer: A

Explanations:

- **A is correct:** The sample "selected" itself (volunteered), meaning it is not representative of the whole population.
- **B is incorrect:** There is no indication that the effect is causing the cause here; it's a sampling problem.
- **C is incorrect:** This is a requirement for causality, not a type of bias.
- **D is incorrect:** The error here is systematic (only happy people), not random.

Question 12: Causal Inference in Business

Why is distinguishing correlation from causation critical for business decision-making?

A. It allows businesses to use more complicated-looking math in their presentations.

B. Intervening based on simple correlation (e.g., forcing employees to wear suits because well-dressed employees are productive) may be useless or harmful if there is no causal link.

C. Causal inference is the only way to predict future sales accurately.

D. It isn't critical; correlation is usually enough for most business decisions.

Correct Answer: B

Explanations:

- **A is incorrect:** The goal is better outcomes, not just complex appearances.
- **B is correct:** If you mistake correlation for causation, you might spend money changing something (like dress code) that has no actual effect on the outcome you want (productivity).
- **C is incorrect:** Pure correlation (standard machine learning) is often excellent for *prediction* (passive), even if it fails at *intervention* (active).
- **D is incorrect:** While sometimes acceptable for low-stakes decisions, relying on correlation for major strategic interventions can lead to massive failures.

Question 13: Directed Acyclic Graphs (DAGs)

In a causal diagram (DAG), what does an arrow drawn from variable A to variable B (A -> B) represent?

A. That A and B are correlated, but we don't know why.

B. That B causes A.

C. A hypothesized direct causal effect of A on B.

D. That A happens later in time than B.

Correct Answer: C

Explanations:

- **A is incorrect:** An arrow in a DAG is a strong claim of assumed causality, not just correlation.
- **B is incorrect:** The arrow points from cause to effect; A -> B means A causes B.
- **C is correct:** DAGs are used to map out assumed causal relationships to help identify confounders.
- **D is incorrect:** Arrows generally follow time (cause before effect), so A must happen before or at the same time as B.

Question 14: Simpson's Paradox

What phenomenon describes a situation where a trend appears in several different groups of data but disappears or reverses when these groups are combined?

A. The Placebo Effect.

B. Simpson's Paradox.

C. The Law of Large Numbers.

D. Standard Deviation.

Correct Answer: B

Explanations:

- **A is incorrect:** The placebo effect concerns perceived improvements due to belief in a treatment, not statistical trend reversals.
- **B is correct:** Simpson's Paradox is a classic causal problem where failing to control for the group structure leads to the wrong conclusion.
- **C is incorrect:** The Law of Large Numbers states that larger samples get closer to the true average, not that trends reverse.

- **D is incorrect:** Standard deviation is a measure of data spread, not a paradoxical trend phenomenon.

Question 15: ATE (Average Treatment Effect)

What is the "Average Treatment Effect" (ATE) attempting to measure?

A. The average difference in outcomes between the treated and control groups, assuming well-balanced data.

B. The correlation coefficient between the treatment and the outcome.

C. The average of all variables in the dataset before the treatment was applied.

D. The number of people who received the treatment.

Correct Answer: A

Explanations:

- **A is correct:** ATE measures the actual causal impact of an intervention on the entire population on average (e.g., "On average, this drug increases lifespan by 2 years").
- **B is incorrect:** ATE is a measure of *impact size* (e.g., +$500 sales), not just a correlation score (-1 to 1).
- **C is incorrect:** This would just be baseline descriptive statistics, not a treatment effect.
- **D is incorrect:** This is just a count of the sample size, not a measure of the effect itself.

End of Quiz