Assignment

PRASHANTH T
NUID: 002843755

November 2025

1 Introduction

This document contains 15 multiple-choice questions on fundamental concepts in causal inference and data analysis, covering topics such as correlation vs causation, confounding variables, DAGs, RCTs, propensity score matching, and various causal inference methodologies.

2 Questions and Answers

Question 1: Fundamental Concepts

What is the primary difference between correlation and causation in data analysis?

A) Correlation implies causation in all cases

B) Causation can be inferred from correlation alone

C) Correlation indicates association while causation indicates a direct cause-effect relationship

D) There is no difference between correlation and causation

Correct Answer: C
Explanations:
Option A (Incorrect): This is a common misconception. Correlation does NOT imply causation in all cases. Two variables can be correlated due to coincidence, confounding variables, or reverse causation without having a direct causal relationship.
Option B (Incorrect): Causation cannot be inferred from correlation alone. Additional evidence such as temporal precedence, elimination of confounders, and theoretical mechanisms are required to establish causation.
Option C (Correct): This accurately describes the distinction. Correlation measures statistical association between variables (they move together), while causation indicates that changes in one variable directly cause changes in another.
Option D (Incorrect): There is a fundamental difference. Correlation is symmetric (if X correlates with Y, then Y correlates with X), while causation is directional (X causing Y does not mean Y causes X).

Question 2: Confounding Variables

Which of the following best describes a confounding variable in causal analysis?

A) A variable that is caused by the outcome

B) A variable that affects both the treatment and outcome, creating spurious associations

C) A variable that only affects the treatment variable

D) A variable that mediates the effect of treatment on outcome

Correct Answer: B
Explanations:
Option A (Incorrect): This describes a collider or a consequence of the outcome, not a confounder. A confounder must precede both treatment and outcome.

Option B (Correct): A confounding variable is one that influences both the independent variable (treatment) and dependent variable (outcome), potentially creating a false impression of a causal relationship. For example, ice cream sales and drowning rates are both confounded by temperature/season.

Option C (Incorrect): A variable that only affects the treatment would be an instrumental variable or a predictor of treatment, not a confounder. Confounders must affect both treatment and outcome.

Option D (Incorrect): This describes a mediator variable, which lies on the causal pathway between treatment and outcome, not a confounder which creates an alternate pathway.

Question 3: Directed Acyclic Graphs (DAGs)

In a Directed Acyclic Graph (DAG) for causal inference, what does an arrow from variable X to variable Y represent?

A) X and Y are correlated

B) X directly causes Y

C) Y causes X

D) X and Y share a common cause

Correct Answer: B
Explanations:
Option A (Incorrect): While X and Y would be correlated if X causes Y, the arrow specifically represents causation, not just correlation. Correlation could exist without a direct arrow (through confounders).

Option B (Correct): In DAG notation, an arrow from X → Y explicitly represents a direct causal effect of X on Y. This is the fundamental convention of causal graphs.

Option C (Incorrect): The direction of the arrow matters. X → Y means X causes Y, not the reverse. If Y caused X, the arrow would point from Y to X.

Option D (Incorrect): A common cause would be represented by a third variable with arrows pointing to both X and Y (Z → X and Z → Y), not by a direct arrow between X and Y.

Question 4: Randomized Controlled Trials (RCTs)

Why are Randomized Controlled Trials (RCTs) considered the gold standard for establishing causality?

A) They eliminate all measurement error

B) Random assignment balances both observed and unobserved confounders across treatment groups

C) They are cheaper and faster than observational studies

D) They guarantee external validity

Correct Answer: B
Explanations:
Option A (Incorrect): RCTs do not eliminate measurement error. Instruments can still have precision limitations and human error in data collection can occur. RCTs address confounding, not measurement error.

Option B (Correct): Random assignment ensures that both known and unknown confounding variables are balanced in expectation across treatment and control groups, making the groups comparable except for the treatment. This isolates the causal effect.

Option C (Incorrect): RCTs are typically more expensive and time-consuming than observational studies. They require controlled environments, recruitment, monitoring, and ethical oversight.

Option D (Incorrect): RCTs often have limited external validity (generalizability) because they are conducted in controlled settings with specific populations that may not represent the broader population.

Question 5: Propensity Score Matching

What is the primary purpose of propensity score matching in observational studies?

A) To increase sample size

B) To balance covariates between treated and control groups to mimic randomization

C) To identify causal mechanisms

D) To eliminate all selection bias completely

Correct Answer: B
Explanations:

Option A (Incorrect): Propensity score matching typically reduces sample size because unmatched units are discarded. The goal is not to increase but to create more comparable groups.

Option B (Correct): Propensity score matching creates treatment and control groups with similar distributions of observed covariates by matching units with similar probabilities of receiving treatment. This reduces confounding and approximates the balance achieved in RCTs.

Option C (Incorrect): Propensity score matching addresses confounding to estimate causal effects, but it doesn't identify mechanisms (the pathways through which treatment affects outcomes). Mediation analysis addresses mechanisms.

Option D (Incorrect): Propensity score matching only addresses selection bias due to observed confounders. It cannot eliminate bias from unobserved confounders, which is a fundamental limitation of observational studies.

Question 6: Missing Data Mechanisms

Which missing data mechanism poses the most significant threat to valid causal inference?

A) Missing Completely at Random (MCAR)

B) Missing at Random (MAR)

C) Missing Not at Random (MNAR)

D) All mechanisms are equally problematic

Correct Answer: C
Explanations:

Option A (Incorrect): MCAR means missingness is unrelated to any variables (observed or unobserved). While this reduces sample size and statistical power, it doesn't introduce bias in estimates, making it the least problematic mechanism.

Option B (Incorrect): MAR means missingness depends on observed variables but not on unobserved values. This can be handled with appropriate imputation methods (like multiple imputation) that account for observed variables.

Option C (Correct): MNAR means missingness depends on the unobserved values themselves. For example, people with severe health conditions not reporting their status. This creates bias that cannot be corrected with standard methods because the reason for missingness is unknown.

Option D (Incorrect): The mechanisms differ substantially in their impact on causal inference. MNAR is most problematic, MAR is manageable, and MCAR is least concerning for bias.

Question 7: Instrumental Variables

A valid instrumental variable (IV) must satisfy which of the following conditions?

A) It must be correlated with the outcome

B) It must be correlated with the treatment and uncorrelated with unobserved confounders

C) It must be a confounder

D) It must mediate the treatment effect

Correct Answer: B
Explanations:

Option A (Incorrect): An instrumental variable should affect the outcome only through its effect on the treatment (indirect effect), not have a direct effect on the outcome. Direct correlation with the outcome violates the exclusion restriction.

Option B (Correct): A valid IV must satisfy three conditions: (1) relevance - correlated with the treatment, (2) independence - uncorrelated with unobserved confounders, and (3) exclusion restriction - affects outcome only through treatment. This answer captures the key requirements.

Option C (Incorrect): An instrumental variable is explicitly NOT a confounder. Confounders affect both treatment and outcome, while IVs affect only treatment (not outcome directly) and are independent of confounders.

Option D (Incorrect): A mediator lies on the causal pathway between treatment and outcome. An instrumental variable affects the treatment but does not lie on the pathway from treatment to outcome.

Question 8: Counterfactuals

What does the counterfactual outcome represent in causal inference?

A) The observed outcome for the treatment group

B) The outcome that would have occurred under an alternative treatment condition

C) The average outcome across all groups

D) The predicted outcome from a regression model

Correct Answer: B
Explanations:

Option A (Incorrect): The observed outcome is the factual outcome, not counterfactual. The fundamental problem of causal inference is that we can only observe one outcome (treatment or control) for each unit, not both.

Option B (Correct): The counterfactual outcome is the hypothetical outcome that would have been observed if the treatment status had been different. For example, for a treated unit, the counterfactual is what would have happened without treatment. We cannot observe both simultaneously, which is why causal inference requires careful methodology.

Option C (Incorrect): The average outcome across groups is an observed summary statistic, not a counterfactual. Counterfactuals are unit-specific hypothetical outcomes under different treatment conditions.

Option D (Incorrect): While regression models can help estimate counterfactual outcomes by prediction, the counterfactual itself is a conceptual framework about potential outcomes, not simply a model prediction.

Question 9: Feature Selection in Causal Analysis

When performing feature selection for causal analysis, which approach is most appropriate?

A) Include all variables with high correlation to the outcome

B) Use domain knowledge to identify confounders and include them while excluding colliders

C) Select features based solely on predictive performance metrics like R^2

D) Randomly select features to avoid bias

Correct Answer: B
Explanations:
Option A (Incorrect): High correlation with the outcome doesn't determine causal relevance. Including colliders (variables caused by both treatment and outcome) can introduce bias. Mediators might also be inappropriate to control for depending on the research question.

Option B (Correct): Causal feature selection requires theoretical understanding of the causal structure. Confounders must be included to block backdoor paths, while colliders should be excluded to avoid introducing spurious associations. This requires domain knowledge and causal reasoning, not just statistical criteria.

Option C (Incorrect): Predictive performance metrics are appropriate for prediction tasks but not for causal inference. Models with high R^2 might control for mediators or colliders, biasing causal estimates even while improving prediction.

Option D (Incorrect): Random feature selection would lead to omitted variable bias (missing confounders) and potentially include colliders. Causal inference requires thoughtful, theory-driven selection, not randomness.

Question 10: Treatment Effect Heterogeneity

What does treatment effect heterogeneity mean in causal analysis?

A) The treatment has no effect on any individuals

B) The treatment effect varies across different subgroups or individuals

C) The treatment effect is the same for everyone

D) The treatment causes heterogeneous outcomes to become homogeneous

Correct Answer: B
Explanations:
Option A (Incorrect): Treatment effect heterogeneity means effects vary across individuals, not that there is no effect at all. Even if the average treatment effect is zero, there could be heterogeneity (positive effects for some, negative for others).

Option B (Correct): Heterogeneity means the causal effect differs across units or subgroups. For example, a medication might be highly effective for younger patients but ineffective or harmful for older patients. Understanding heterogeneity is crucial for personalized treatments and policy targeting.

Option C (Incorrect): This describes homogeneous treatment effects, the opposite of heterogeneity. In reality, most interventions have heterogeneous effects due to individual differences in biology, context, or characteristics.

Option D (Incorrect): Treatment effect heterogeneity refers to variation in the treatment's causal impact across individuals, not to changes in outcome variance. A treatment can have heterogeneous effects while increasing, decreasing, or not affecting outcome variance.

Question 11: Backdoor Criterion

The backdoor criterion in causal inference helps identify:

A) Variables that mediate the treatment effect

B) Sets of variables that, when controlled for, block all spurious paths between treatment and outcome

C) Variables that are caused by both treatment and outcome

D) Variables that should never be included in the analysis

Correct Answer: B
Explanations:
Option A (Incorrect): The backdoor criterion identifies confounders to control for, not mediators. Mediators are on the causal path from treatment to outcome, while backdoor paths are non-causal associations through confounders.

Option B (Correct): The backdoor criterion provides a graphical rule for identifying sets of variables that, when conditioned on, block all non-causal ("backdoor") paths from treatment to outcome while leaving the causal path open. This isolates the causal effect.

Option C (Incorrect): Variables caused by both treatment and outcome are colliders. The backdoor criterion explicitly states that these should NOT be controlled for, as conditioning on colliders opens spurious paths.

Option D (Incorrect): The backdoor criterion identifies variables that SHOULD be included (controlled for) to achieve causal identification, not variables to exclude.

Question 12: Handling Categorical Variables

When encoding categorical variables for causal analysis, which consideration is most important?

A) Always use label encoding to reduce dimensionality

B) Choose encoding that preserves causal interpretation and doesn't create artificial ordering

C) Use the encoding method that maximizes predictive accuracy

D) Randomly assign numerical values to categories

Correct Answer: B
Explanations:
Option A (Incorrect): Label encoding (assigning arbitrary integers like 0, 1, 2 to categories) imposes an artificial ordering that doesn't exist in nominal variables (like color or country). This can distort causal estimates by implying ordinal relationships.

Option B (Correct): For causal analysis, encoding should reflect the true nature of the variable. Nominal variables should use one-hot encoding to avoid imposing false ordering. The encoding should maintain interpretability of causal effects (e.g., the effect of being in category A vs. category B).

Option C (Incorrect): While predictive accuracy matters in prediction tasks, causal inference prioritizes unbiased causal estimates and interpretability. An encoding that boosts R^2 but distorts causal interpretation is inappropriate.

Option D (Incorrect): Random assignment of numerical values to categories would be arbitrary and non-reproducible, making results uninterpretable and potentially introducing bias through false assumptions about variable relationships.

Question 13: Regression Discontinuity Design (RDD)

What is the key identifying assumption in a Regression Discontinuity Design?

A) Treatment is randomly assigned

B) Units just above and below the threshold are similar except for treatment status

C) All confounders are observed

D) The outcome is continuous

Correct Answer: B
Explanations:
Option A (Incorrect): In RDD, treatment is NOT randomly assigned but rather determined by whether a running variable crosses a threshold. For example, students scoring above 60 receive a scholarship. RDD exploits this discontinuity, not randomization.

Option B (Correct): RDD assumes that units just on either side of the threshold are comparable in all respects except treatment status (local randomization). For example, a student scoring 59.9 should be similar to one scoring 60.1, making the discontinuous jump in outcomes at the threshold attributable to treatment.

Option C (Incorrect): RDD does not require observing all confounders globally. The design assumes local similarity near the threshold, which provides causal identification even with unobserved confounders elsewhere in the distribution.

Option D (Incorrect): While continuous outcomes are common in RDD, the outcome can be binary or categorical. The key requirement is discontinuity in treatment probability or receipt at the threshold, not outcome continuity.

Question 14: Difference-in-Differences (DiD)

The parallel trends assumption in Difference-in-Differences means:

A) Treatment and control groups must have identical levels before treatment

B) Treatment and control groups would have followed the same trend in the absence of treatment

C) The treatment effect is the same for all time periods

D) Both groups must have the same sample size

Correct Answer: B
Explanations:
Option A (Incorrect): DiD does NOT require groups to have the same pre-treatment levels. In fact, DiD is designed to handle situations where treatment and control groups differ in levels, as long as their trends are parallel.

Option B (Correct): The parallel trends assumption is that, in the absence of treatment, the difference between treatment and control groups would have remained constant over time (parallel trajectories). Violation of this assumption (diverging trends) would bias DiD estimates.

Option C (Incorrect): This describes treatment effect homogeneity over time, which is not required for basic DiD. Event study designs can examine time-varying effects without violating the parallel trends assumption.

Option D (Incorrect): Sample size balance is not part of the parallel trends assumption. Groups can have different sizes as long as the counterfactual trends would have been parallel.

Question 15: Data Preprocessing for Causal Analysis

Which data preprocessing step requires special consideration in causal inference that differs from standard machine learning?

A) Standardizing numerical features

B) Deciding whether to control for post-treatment variables

C) Removing duplicate observations

D) Converting text to numerical representations

Correct Answer: B
Explanations:

Option A (Incorrect): Standardization (scaling features to similar ranges) is a common preprocessing step for both causal inference and machine learning. While it affects coefficient interpretation, it doesn't fundamentally change causal identification strategies.

Option B (Correct): Post-treatment variables (measured after treatment assignment) can be mediators, colliders, or consequences of treatment. Controlling for them can block the causal path of interest (mediators) or introduce bias (colliders). This requires causal reasoning about temporal ordering and the causal graph, unlike standard ML where any predictive feature might be included.

Option C (Incorrect): Removing duplicates is important for data quality in both causal inference and machine learning equally. It doesn't require special causal considerations.

Option D (Incorrect): Text encoding (like TF-IDF or embeddings) is a technical preprocessing step similar across causal and predictive tasks. The causal consideration is whether to include text features (based on the causal graph), not how to encode them.