🧩 Crash Course in Causality(Written Section) — Titanic Dataset Quiz

Q1. What is the primary *treatment* variable in the Titanic causal analysis?

Options:
A. `sex`
B. `survived`
C. `class`
D. `fare`

✅ **Correct Answers:** A (`sex`)
Explanation:

- `sex` is the treatment variable — we're examining its causal effect on survival.
- `Surviving` is the **outcome**, not the treatment.
- `Class` and `fare` are **confounders** that influence both `sex` and `survival.`

Q2. In the context of this study, what is the *outcome* variable?

Options:
A. `sex`
B. `survived`
C. `age_imputed`
D. `class`

✅ **Correct Answer:** B (`survived`)
Explanation:

- The outcome is whether the passenger survived or not.
- Other variables like `age_imputed` and `class` are predictors or confounders.
- `sex` is the treatment variable whose effect we are estimating on survival.

Q3. Why is imputing missing values in `age` important for causal inference?

Options:
A. It prevents selection bias caused by dropping rows.
B. It increases model accuracy only.
C. It helps maintain the causal structure of the dataset.
D. It introduces new confounders.

✅ **Correct Answers:** A, C
Explanation:

- A: Dropping rows with missing `age` may bias results if missingness relates to survival.
- C: Imputation + adding a missingness indicator (`missing_age`) preserves causal pathways.
- B: Accuracy isn't the main reason — causal validity is.
- D: The missingness flag helps *control* bias, not introduce confounders.

Q4. What role does `missing_age` play in the model?

Options:
A. It is a confounder.
B. It captures bias introduced by non-random missingness.
C. It should always be dropped from the dataset.
D. It acts as an indicator variable.

✅ **Correct Answers:** B, D
Explanation:

- B: `missing_age` flags missing data that could bias causal inference.
- D: It's an indicator variable (1 = missing, 0 = not missing).
- A: It's not a confounder itself — it's a bias-correction variable.
- C: Dropping it removes valuable information about missingness mechanisms.

Q5. Why is one-hot encoding preferred over ordinal encoding in this analysis?

Options:
A. Ordinal encoding imposes an artificial numeric order.
B. One-hot encoding treats categories as independent.
C. Ordinal encoding increases causal interpretability.
D. One-hot encoding avoids introducing false relationships.

✅ **Correct Answers:** A, B, D
Explanation:

- A/D: Ordinal encoding might imply "Third Class > Second Class," which is false.

- B: One-hot encoding properly separates each category.
- C: Ordinal encoding actually *reduces* causal clarity when order is arbitrary.

Q6. What type of bias might occur if we drop rows with missing data?

Options:
A. Selection bias
B. Collider bias
C. Measurement bias
D. Confounding bias

✅ **Correct Answer:** A (Selection bias)
Explanation:

- A: Dropping rows where survival or age is missing skews the sample toward more complete cases.
- B: Collider bias involves conditioning on a variable influenced by both treatment and outcome — not applicable here.
- C: Measurement bias refers to inaccurate data collection.
- D: Confounding bias exists, but dropping rows mainly causes **selection bias**.

Q7. Which variables are likely confounders in the Titanic model?

Options:
A. `class`
B. `fare`
C. `sex`
D. `age_imputed`

✅ **Correct Answers:** A, B, D
Explanation:

- A/B/D: These influence both `sex` and `survived` (e.g., class affects both gender distribution and survival).
- C: `sex` is the treatment, not a confounder.

Q8. In DoWhy, what does the *backdoor criterion* ensure?

Options:
A. All confounders are blocked from influencing the treatment–outcome path.
B. No colliders are conditioned on.
C. All mediators are adjusted for.
D. The model is overfitted.

✅ **Correct Answers:** A, B
Explanation:

- A: The backdoor path blocks spurious correlations via confounders.
- B: Conditioning on colliders is avoided, as it opens biasing paths.
- C: Mediators are *not* adjusted for — that blocks true causal effects.
- D: Overfitting is unrelated.

Q9. What is the estimated direction of the causal effect of being male (`sex_male`) on survival?

Options:
A. Positive — being male increases survival probability.
B. Negative — being male decreases survival probability.
C. Zero — no causal effect detected.
D. Depends on class level.

✅ **Correct Answer:** B (Negative)
Explanation:

- The estimated Average Treatment Effect (ATE) was about **−0.5**, meaning being male reduces survival probability by roughly 50%.
- This aligns with the "women and children first" evacuation policy.

Q10. Why is it important to include `fare` and `class` together?

Options:
A. They jointly influence survival chances and are correlated.
B. One acts as a proxy for the other.
C. They represent different causal paths.
D. They are redundant variables and should not be included together.

✅ **Correct Answers:** A, B, C
Explanation:

- A/B: Higher fare typically corresponds to higher class — both influence survival.
- C: Fare captures economic status, class captures cabin priority; both valid.
- D: Not redundant — they provide complementary information.

Q11. In the fallback regression model (without DoWhy), how is the treatment effect estimated?

Options:
A. As the coefficient of `sex_male` in the logistic regression.
B. As the R^2 value of the model.
C. As the residual variance of `survived`.
D. As the intercept term.

✅ **Correct Answer:** A
Explanation:

- The coefficient of `sex_male` quantifies how being male changes survival probability, controlling for confounders.
- R^2 and intercept don't represent causal effects.

Q12. What does a refutation test (e.g., random common cause) check in DoWhy?

Options:
A. Model's sensitivity to unobserved confounders.
B. Stability of the causal effect.
C. Model convergence speed.
D. Bias due to incorrect DAG direction.

✅ **Correct Answers:** A, B
Explanation:

- A/B: Refutation adds simulated noise/confounders to see if effect holds.
- C/D: These are unrelated to refutation tests.

Q13. What is a *collider* variable in causal inference?

Options:
A. A variable caused by both treatment and outcome.
B. A confounder that causes both treatment and outcome.
C. A mediator between treatment and outcome.
D. A variable to be adjusted for in all models.

✅ **Correct Answer:** A
Explanation:

- Colliders receive arrows from both treatment and outcome; adjusting for them induces bias.
- Confounders *cause* both, not are *caused* by both.

Q14. What makes a dataset "clean but not causal"?

Options:
A. It has no missing values but a mis-specified causal structure.
B. It has high model accuracy but incorrect feature encoding.
C. It violates causal assumptions despite preprocessing.
D. It uses perfect imputation methods.

✅ **Correct Answers:** A, B, C
Explanation:

- A/B/C: Data can be "statistically tidy" but causally wrong if encoding, dropping, or adjustments distort causal links.
- D: Even perfect imputation can't fix poor causal design.

Q15. Why do we include a DAG (Directed Acyclic Graph) before analysis?

Options:
A. It clarifies assumed causal relationships.
B. It determines which variables to adjust for.
C. It's required for DoWhy syntax.
D. It visualizes how data preprocessing impacts causal paths.

✅ **Correct Answers:** A, B, D
Explanation:

- A/B/D: The DAG makes assumptions explicit, guides confounder selection, and shows how preprocessing affects causal paths.
- C: DoWhy can run without a DAG file — it's conceptual, not mandatory

☕ Crash Course in Causality — Case Study 1 Quiz

Dataset: Seaborn Tips Dataset
Treatment: `smoker`
Outcome: `tip` (continuous)
Goal: Estimate how being a smoker influences tipping behavior, adjusting for confounders such as `total_bill`, `size`, `day`, and `time`.

Q1. What is the causal question explored in this case study?

Options:
A. Does being a smoker cause people to tip less?
B. Do higher total bills cause people to smoke more?
C. Does party size cause larger tips?
D. Does smoking status influence tipping after adjusting for confounders?

✅ **Correct Answers:** A, D
Explanation:

- A/D: The study tests the causal impact of smoking on tipping behavior while adjusting for confounders.
- B and C describe other relationships, not the main causal question.

Q2. Which variable is the treatment (`T`)?

Options:
A. `smoker`
B. `tip`
C. `total_bill`
D. `day`

✅ **Correct Answer:** A
Explanation:

- `smoker` (yes/no) is the treatment whose causal effect on `tipping` we estimate.
- `tip` is the outcome; `total_bill`, `day` are controls/confounders.

Q3. What is the outcome (`Y`) variable in this analysis?

Options:
A. `tip`
B. `smoker`
C. `day`
D. `size`

✅ **Correct Answer:** A
Explanation: `tip` is the numerical outcome predicted by smoking status after adjusting for confounders.

Q4. Which variables act as potential confounders in this study?

Options:
A. `total_bill`
B. `size`
C. `day`
D. `time`
E. `sex`

✅ **Correct Answers:** A, B, C, D, E
Explanation:
All these variables can influence both `smoker` status and `tipping`:

- Bigger parties (size) may have more smokers and higher tips.
- Weekend days and night times change both smoking and tipping patterns.
- Sex may affect social behavior and tip amounts.

Q5. Why do we encode categorical variables like `smoker`, `day`, and `time` before analysis?

Options:
A. To numerically represent categories for modeling.
B. Because machine-learning models cannot handle strings directly.
C. To impose artificial ordering between days of the week.
D. To avoid bias introduced by label encoding when no order exists.

✅ **Correct Answers:** A, B, D
Explanation:
One-hot encoding represents categories without imposing order.
Option C is incorrect because artificial order should be avoided, not added.

Q6. What does including `total_bill` as a control variable achieve in the causal model?

Options:
A. Blocks a backdoor path between smoking and tipping.
B. Accounts for spending habits that influence tips.
C. Acts as a mediator between smoking and tipping.
D. Removes spurious correlation due to different bill sizes.

✅ **Correct Answers:** A, B, D
Explanation: `total_bill` is a confounder — larger bills and smoking behavior both affect tips.
It's not a mediator (C is incorrect).

Q7. If we fail to adjust for party size, what type of bias can occur?

Options:
A. Confounding bias
B. Selection bias
C. Collider bias
D. Measurement bias

✅ **Correct Answer:** A
Explanation: Party size affects both smoking and tipping — not adjusting for it introduces confounding bias.

Q8. In DoWhy terms, what does the "backdoor criterion" mean here?

Options:
A. We must block all paths from smoker to tip that go through confounders.
B. We should adjust for variables that cause both treatment and outcome.
C. We include all mediators between smoker and tip.
D. We avoid conditioning on colliders.

✅ **Correct Answers:** A, B, D
Explanation: The backdoor criterion ensures causal identification by blocking spurious paths through confounders and avoiding colliders.
Mediators (C) should *not* be controlled.

Q9. What encoding technique did we use for categorical features in the Tips dataset?

Options:
A. One-Hot Encoding
B. Ordinal Encoding
C. Target Encoding
D. Frequency Encoding

✅ **Correct Answer:** A
Explanation: One-Hot Encoding was used to avoid imposing order on categorical variables like `day` and `time`.

Q10. What does a negative causal effect of `smoker` on `tip` imply?

Options:
A. Smokers tip less than non-smokers on average.
B. Being a smoker causes a decrease in tips after adjusting for confounders.
C. Smokers have larger total bills and thus tip more.
D. There is a positive relationship between smoking and tipping.

✅ **Correct Answers:** A, B
Explanation: A negative ATE means smokers tip less causally, not just correlationally.
C and D contradict the estimated direction of effect.

Q11. What does including `sex` and `time` as control variables help achieve?

Options:
A. Removes confounding due to social or temporal patterns.
B. Blocks bias from differences in behavior by gender or time of day.
C. Adds redundant features to the model.
D. Improves causal interpretation by adjusting for context.

✅ **Correct Answers:** A, B, D
Explanation: Sex and time affect both smoking and tipping behaviors, so adjusting for them clarifies the causal effect of smoking.

Q12. Why is DoWhy used in this study?

Options:
A. To formally identify and estimate the causal effect using graphical criteria.
B. To perform black-box prediction only.
C. To test robustness through refutation methods.
D. To compare linear vs non-linear causal models.

✅ **Correct Answers:** A, C, D
Explanation: DoWhy identifies effects, estimates them, and supports refutations for robustness.
It's not a generic predictive tool (B is wrong).

Q13. What does the refutation test (add random common cause) verify?

Options:
A. Whether the effect changes significantly after introducing noise.
B. Model's sensitivity to unobserved confounding.
C. Model's accuracy on the test set.
D. Robustness of the estimated causal effect.

✅ **Correct Answers:** A, B, D
Explanation: Refutation adds synthetic confounders to check if the effect is robust.
C relates to predictive evaluation, not causal refutation.

Q14. What can happen if we accidentally adjust for a collider in this dataset?

Options:
A. We introduce spurious correlations between treatment and outcome.
B. We block the true causal path.
C. We open a backdoor path that biases estimates.
D. We reduce the variance of the effect estimate.

✅ **Correct Answers:** A, C
Explanation: Adjusting for a collider creates spurious association between smoking and tipping, biasing the causal effect. Variance (D) is not directly impacted in that way.

Q15. Why is data preparation crucial for causal inference in this study?

Options:
A. Encoding and missing data handling affect causal paths.
B. Causal inference depends on accurate variable relationships.

C. Clean data automatically ensures causal validity.
D. Improper encoding can change the direction or magnitude of the causal effect.

✅ **Correct Answers:** A, B, D
Explanation: Causal validity relies on sound data preparation. C is wrong — clean data alone is not necessarily causal.

🚗 Crash Course in Causality — Case Study 2 Quiz

Dataset: Seaborn Car Crashes Dataset
Treatment: `alcohol`
Outcome: `total`
Goal: Estimate how alcohol involvement influences the number of car crashes while controlling for confounders such as speeding, distraction, prior violations, insurance losses, and state differences.

Q1. What is the central causal question in this case study?

Options:
A. Does higher alcohol involvement cause more total crashes?
B. Do more crashes cause higher alcohol involvement?
C. Does speeding increase alcohol use among drivers?
D. Does alcohol involvement affect total crashes after adjusting for confounders?

✅ **Correct Answers:** A, D
Explanation:

- A & D express the intended causal direction — alcohol → crashes, adjusted for confounders.
- B reverses causality; C is unrelated to the causal question.

Q2. What is the *treatment variable (T)* in this analysis?

Options:
A. `alcohol`
B. `speeding`
C. `total`
D. `ins_premium`

✅ **Correct Answer:** A
Explanation:
`alcohol` is the treatment whose causal impact on `total` (outcome) is being estimated.

Q3. Which variable is the *outcome (Y)*?

Options:
A. `alcohol`
B. `total`

C. `speeding`
D. `ins_losses`

✅ **Correct Answer:** B
Explanation:
`total` represents the total number of crashes — the causal outcome of interest.

Q4. Which variables were included as *confounders* in the causal model?

Options:
A. `speeding`
B. `not_distracted`
C. `no_previous`
D. `ins_premium`
E. `ins_losses`
F. `abbrev` (state)

✅ **Correct Answers:** A, B, C, D, E, F
Explanation:
All these can influence both alcohol involvement and crash totals, thus acting as confounders that must be adjusted for.

Q5. Why do we one-hot encode the `abbrev` (state) variable?

Options:
A. To prevent imposing a numeric order on states.
B. To let each state have its own baseline risk level.
C. To improve interpretability of regional effects.
D. To create artificial ranking of states.

✅ **Correct Answers:** A, B, C
Explanation:
Encoding states as dummy variables allows separate intercepts per state.
Option D is incorrect because we explicitly *avoid* artificial ordering.

Q6. Why is adjusting for `speeding` essential in this causal setup?

Options:
A. Speeding influences both alcohol involvement and crash totals.
B. Speeding is a mediator between alcohol and crashes.
C. Speeding is a confounder that could bias the alcohol effect.
D. Speeding causes alcohol involvement.

✅ **Correct Answers:** A, C
Explanation:
Speeding correlates with both treatment and outcome, so failing to adjust for it introduces confounding. It's not a mediator (B ✗) or a cause of alcohol use (D ✗).

Q7. What kind of bias might occur if we fail to control for `ins_premium` and `ins_losses`?

Options:
A. Confounding bias
B. Selection bias
C. Collider bias
D. Sampling bias

✅ **Correct Answer:** A
Explanation:
Insurance factors affect crash rates and may correlate with alcohol usage patterns across states. Ignoring them leads to confounding bias.

Q8. What does the *backdoor criterion* ensure in this context?

Options:
A. That all non-causal paths between `alcohol` and `total` are blocked.
B. That colliders are not conditioned on.
C. That the model includes every variable in the dataset.
D. That causal identification is valid given observed confounders.

✅ **Correct Answers:** A, B, D
Explanation:
The backdoor criterion identifies a valid adjustment set; C is incorrect because not all variables are needed.

Q9. Suppose the estimated causal effect of `alcohol` on `total` is positive. What does this imply?

Options:
A. Higher alcohol involvement increases total crashes.
B. Alcohol does not affect total crashes.
C. States with more drinking tend to have fewer crashes.
D. There is a direct causal relationship: alcohol → crashes.

✅ **Correct Answers:** A, D
Explanation:
A positive ATE means that as alcohol involvement increases, crash totals rise, supporting a causal interpretation.

Q10. Why is the DoWhy library (or its fallback regression) appropriate for this analysis?

Options:
A. It formalizes causal assumptions and estimation.
B. It identifies confounders via the DAG/backdoor criterion.
C. It predicts future crash counts without any assumptions.
D. It can test robustness through refutation methods.

✅ **Correct Answers:** A, B, D
Explanation:
DoWhy structures causal reasoning and enables effect estimation with robustness checks.
C ✗ — predictive modeling alone is not causal inference.

Q11. What does the *refutation test (add random common cause)* accomplish?

Options:
A. Tests sensitivity of the causal effect to hidden confounders.
B. Adds a random variable to see if the effect changes significantly.
C. Checks multicollinearity among predictors.
D. Assesses robustness of causal estimates.

✅ **Correct Answers:** A, B, D
Explanation:
The test introduces noise to verify whether the causal effect remains stable; it's not about multicollinearity.

Q12. What could happen if we controlled for a mediator, such as driver fatigue, in this model?

Options:
A. The total causal effect of alcohol on crashes would be underestimated.
B. We would block part of the true causal pathway.
C. We would improve precision of estimation.
D. We would introduce post-treatment bias.

✅ **Correct Answers:** A, B, D
Explanation:
Adjusting for mediators blocks the causal effect and biases results; it rarely improves precision.

Q13. Which statement best describes a *collider* in the car-crash causal graph?

Options:
A. A variable caused by both alcohol involvement and total crashes.
B. A variable that causes both alcohol involvement and crashes.
C. A variable completely unrelated to either.
D. A variable we should condition on for better accuracy.

✅ **Correct Answer:** A
Explanation:
A collider is *caused by* both treatment and outcome; conditioning on it induces spurious correlation.

Q14. Why is data preparation (e.g., encoding, imputation, normalization) crucial before causal estimation?

Options:
A. Because causal relationships depend on how variables are represented.
B. Because preprocessing affects which paths appear active in the DAG.
C. Because causal inference is purely statistical and unaffected by data prep.
D. Because correct preprocessing prevents distortion of effect estimates.

✅ **Correct Answers:** A, B, D
Explanation:
Encoding and scaling choices can alter relationships or introduce artificial correlations.
C ✗ — causal inference *is* sensitive to representation.

Q15. Why is this causal framework relevant for public-policy decision-making?

Options:
A. It quantifies how much alcohol restrictions could reduce crashes.
B. It distinguishes correlation from causation, guiding interventions.
C. It helps states design evidence-based road-safety policies.
D. It focuses only on predicting crash counts without interpretation.

✅ **Correct Answers:** A, B, C
Explanation:
Causal analysis informs *actionable* policy (e.g., limiting alcohol consumption).
D ✗ — predictive models alone don’t reveal causal levers.