**Causal inference:**

1. Explain why observed correlation between X and Y (represented by the estimated coefficient in OLS) does not imply causal relationship X=>Y? What might be the problems that you are facing in identifying the real causal relationship, i.e. treatment effect? (mention min 2 of them)
2. What do you understand under the term “counterfactual”?
3. What is the difference between Average Treatment Effect (ATE) and Average Treatment Effect (ATT). Further, what do we understand under Intention to treat (ITT)?
4. What types of selection mechanisms do you know? (give a real life examples)

**Controlled experiments:**

1. What is randomization in the context of controlled experiments? What assumptions underlie the identification strategy in this case?
2. Discuss the problem of non-compliance in the controlled experiments. What treatment effect do we estimate in this context? Illustrate this problem on the real world example.
3. Discuss the problem of attrition in the controlled experiments. Illustrate this problem on the real world example.
4. Discuss the problem of externalities in the controlled experiments. Illustrate this problem on the real world example.
5. Discuss the problem of randomization on the level of groups (e.g. villages, classes). What implication does it have for our statistical inference? How do we correct for it?
6. Give example of some issues (min 2) that need to be considered in the process of design of experiment.
7. What is Hawthorne effect in the context of measuring treatment effect using controlled experiments?
8. What is John Henry effect in the context of measuring treatment effect using controlled experiments?
9. Give an example of the demand effect in the context of measuring treatment effect using controlled experiments.

**Difference-in-difference:**

1. Provide graphical intuition behind DD estimator and its baseline assumption.
2. What is the assumption that we impose on the unobservables in the DD estimator?
3. Write down and explain models used in the estimation of DD in the panel data and cross-section data. What would happen if we ran both specifications on panel data?
4. Explain the issue of Ashenfelter’s dip in DD estimation. How can you control for this issue?
5. Give an example where expectations of the policy step can affect the validity of results obtained by DD.
6. What do you understand under DDD estimator? Give an example of application.
7. Explain the issue that was raised by Betrand et al (2004) in their paper, which I have presented on the lecture. What is the proposed solution?

**Matching:**

1. Define matching. Explain the baseline assumption of matching (also called unconfoundness assumption). What assumptions on unobservabes does it imply?
2. Give a definition of common support. Why do we need common support in the matching?
3. Explain the method of exact matching. What problems do we encounter in its application?
4. Write down the procedure for the propensity score matching.
5. Explain the first step in the propensity score matching – estimation of propensity score. How do you check balancing property? What other diagnostic/statistics would you run/check? What would you check in terms of common support?
6. How do we obtain treatment effect using nearest neighbor matching method?
7. How do we obtain treatment effect using nearest radius matching method?
8. How do we obtain treatment effect using nearest kernel matching method?
9. How would you choose optimal matching algorithm?
10. What is the implication of the use of propensity score matching model on the validity of statistical inference? How can you deal with this problem?
11. What are the similarities and differences between OLS and propensity score matching?
12. Why and how would one combine propensity score matching with DD approach?

**Instrumental variables**

1. What types of inference problems we aim to solve using IV approach?
2. Provide the intuition behind the use of instrumental variables. What 2 condition must a valid instrumental variable satisfy?
3. Describe the 2 stages of the instrumental variables estimation.
4. What kind of problems do we encounter when IV is weak? (Hint: write down the equation describing the bias of IV estimator) How can we test for weak IV?
5. How and when can we test for the exclusion restriction? What problems you see in this approach?
6. What is the interpretation of treatment effect estimated in the context of heterogenous treatment effect? (hint: 4 groups of population – always takers, never takes, compliers, defyiers)
7. Give examples of situations where a particular IV is valid even after the LATE interpretation of results.

**Regression discontinuity**

1. Explain the main idea behind regression discontinuity design. How does it allow us to identify the causal effect?
2. What is sharp design of regression discontinuity design? (Hint: try to provide graphical illustration of assignment probabilities in this case)
3. What is fuzzy design of regression discontinuity design? (Hint: try to provide graphical illustration of assignment probabilities in this case) What are the problems with this type of design?
4. Compare RDD and IV – what is common and what is different?
5. What would you check in terms of graphical analysis in RD design and why?
6. How do you estimate the treatment effect in RD design? (hint: non-parametric methods)

**Panel data**

1. What are panel data? Give an example of such dataset. What are the advantages over the cross-section data?
2. Explain what is unbalanced panel and what issues does it imply. How can you solve them?
3. Explain what is **between** estimator (provide specification).
4. Explain what is **within** estimator (or fixed effects), provide specification, and state main characteristics/problems of this estimator.
5. Explain what is random coefficients model and what is the main difference compared to fixed effects model.
6. Explain how you would test for FE vs RC model (mainly logic underlying the test).

**Duration models**

1. What are we trying to model using duration analysis? What type of data we need and why?
2. Define survivor function and hazard function. Define Kaplan Meier estimator of survival function
3. What is the difference between truncation and censoring?
4. Explain what types of censoring we observe in data (min 3).
5. Explain how do we apply MLE in the context of duration analysis, if we assume time invariant regressors.
6. What are the properties of proportional hazard models (in terms of interpretation of coefficients).

**Structural models**

1. Explain what we understand under structural modeling.
2. What is the difference between statis and dynamic models? How does this translate into decision making process of agent?
3. What are the usual arguments against use of structural models?