DA1 Exam Prep

---

1. What are in the rows and columns of a data table?

\*ANSWER:\* Rows represent observations, while columns represent variables. Rows contain data points, and columns represent characteristics or properties of observations.

---

2. What are ID variables?

\*ANSWER:\* ID variables uniquely identify each entity in a data table, ensuring consistency and integrity across tables in a dataset.

---

3. What are xsec, tseries, and xt panel data? What's an observation in each? Give an example for each.

\*ANSWER:\*

- xsec (cross-sectional data): Observations of different units at the same time (e.g., people in the same year).

- tseries (time series data): Observations of the same unit at different time periods (e.g., a person observed over months).

- xt panel data (cross-section time series data): Multiple units observed across multiple time periods (e.g., yearly financial data for all companies). Each observation is identified by indices i (cross-section) and t (time series).

---

4. What's the validity and what's the reliability of a variable? Give an example of a variable with high validity and one with low validity.

\*ANSWER:\*

- Validity: Measures if a variable accurately represents the intended concept. Example: A thermometer measuring body temperature is valid.

- Reliability: Assesses the consistency of a measurement. Example: A self-report questionnaire on sleep quality may have low validity if responses are subjective and biased.

---

5. What's selection bias? Give an example of data with selection bias and one without.

\*ANSWER:\*

- Selection Bias: Systematic error due to non-random selection.

- Example with Bias: Studying a new drug's effectiveness where only voluntary participants are included.

- Example without Bias: Studying disease prevalence with random sampling, ensuring a representative sample.

---

6. List two common advantages of admin data and two potential disadvantages.

\*ANSWER:\*

- Advantages:

1. High reliability of measured variables.

2. High, often complete, coverage.

- Disadvantages:

1. Limited variables.

2. Low validity of some important variables.

---

7. How can we tell if a sample is representative of a population?

\*ANSWER:\* Assess through benchmarking and evaluating the sampling process. Benchmarking compares variable distributions, while evaluating the process involves understanding the random or non-random selection.

---

8. List two sampling rules that likely lead to a representative sample and two that don't.

\*ANSWER:\*

- Likely Representative:

1. Random sampling.

2. Fixed rules unrelated to data distribution.

- Not Likely Representative:

1. Non-random methods related to important variables.

2. Biased selection due to nonresponse.

---

9. List three common features of Big Data. Why does each feature make data analysis difficult?

\*ANSWER:\*

1. Volume: Large datasets exceed standard capacities.

2. Complexity: Diverse data structures require specialized tools.

3. Continuous collection: Real-time data flow demands ongoing analysis management. These features make analysis challenging due to technical, structural, and temporal complexities.

---

10. An important principle for research is maintaining confidentiality. How can we achieve that when we collect survey data?

\*ANSWER:\* Maintain confidentiality by de-identifying data, obtaining informed consent, adhering to ethical and legal principles, and consulting experts. These practices ensure privacy and ethical data use.

---

11. You want to collect data on the learning habits of students in your data analysis class. List two survey methods and highlight their advantages and disadvantages.

\*ANSWER:\*

1. Self-administered web survey:

- Advantages: Convenient, anonymous.

- Disadvantages: Low response rates, potential tech issues.

2. Personal interview:

- Advantages: Flexible, allows clarification.

- Disadvantages: Time-consuming, potential interviewer bias.

---

12. You want to collect data on the friendship network of students in a class. List arguments for each option, paying attention to representation, costs, and ethical issues.

\*ANSWER:\*

1. Collecting networks from Facebook:

- Representation: Comprehensive, privacy concerns.

- Costs: Potentially low.

- Ethical Issues: Informed consent, privacy.

2. Conducting an online survey:

- Representation: Equal, captures diversity.

- Costs: Moderate.

- Ethical Issues: Informed consent, privacy.

---

13. You consider surveying a sample of employees at a large firm. List four selection methods and assess whether each would result in a representative sample.

\*ANSWER:\*

1. Random Sampling: Potentially representative if every employee has an equal chance.

2. Stratified Sampling: Representative if subgroups are proportionally included.

3. Cluster Sampling: May not be representative if there's high variability within clusters.

4. Convenience Sampling: Prone to selection bias, likely not representative.

---

14. You want to examine the growth of manufacturing firms in a country. Discuss the potential issues of coverage and its consequences. Does it matter which country it is?

\*ANSWER:\*

- Potential Issues:

1. Exclusion of non-listed firms: Bias due to excluding non-listed firms with potentially different growth patterns.

2. Limited sample size: Smaller sample may impact statistical power and generalizability.

3. Selection bias: Listed firms may systematically differ from non-listed ones.

4. Market-driven growth: Biased towards firms more market-driven, overlooking alternative strategies.

- Consequences:

1. Limited insights into overall sector performance.

2. Biased conclusions about growth factors.

3. Misrepresentation of policy implications.

- Country Matters:

- Yes, as factors like stock market development and regulatory frameworks vary, affecting representativeness.

---

15. 1000 firms are randomly selected from all the SMEs (small and medium enterprises) in a country. What is the population, what is the sample, and is the sample representative?

\*ANSWER:\*

- Population: All small and medium enterprises (SMEs) in the country.

- Sample: The 1000 firms randomly selected from the SME population.

- Representativeness: Not determined without information on benchmarking or distribution comparison.

---

16. You are doing a survey about the smoking habits of the students of your university and want to reach a 20% sample. Here are some potential sampling rules. Would each lead to a representative sample? Why or why not?

\*ANSWER:\*

1. Stand at the main entrance and select every fifth entering student: Non-representative, systematic bias.

2. Get the students' email list and select every fifth person in alphabetic order: Non-representative, systematic bias.

3. Select the first fifth of students in alphabetic order: Non-representative, systematic bias.

4. Sort students according to a random number and select the first fifth: Likely representative, random and unbiased.

---