Producing Data

Data Collection, Sampling and Experimental Design

Shahkar Ahmad Nahvi

DoEE, IUST

Learning Outcomes

- distinguish between observational studies and experiments,
- explain principles of experimental design,
- understand sampling techniques,
- identify bias in surveys and experiments,
- explain sampling variability and inference,
- discuss ethical issues in data collection.

Big Picture

Data Production

$\Downarrow$

Randomness and Sampling

$\Downarrow$

Probability

$\Downarrow$

Statistical Inference

Good inference requires good data production.

1. Sources of Data

Anecdotal Data

- Based on isolated experiences.
- Often unreliable and biased.

Example: "This medicine worked for my friend."

Available Data

- Previously collected data.
- Examples: databases, APIs, institutional records.

Limitations of Available Data

- Data may not match research objectives.
- Missing values may exist.
- Some groups may be underrepresented.

Example
Fitness app data may overrepresent health-conscious users.

Large datasets are not automatically reliable.

2. Observational Studies

Definition
Researchers observe subjects without imposing treatments.

Examples

- Smoking and health
- Traffic monitoring
- Social media analysis

Observational studies show association, not causation.

3. Experiments

Definition

Researchers impose treatments and observe responses.

Examples

- Drug testing
- Fertilizer comparison
- Teaching method evaluation

Properly designed experiments can establish causation.

Observational Study vs Experiment

Aspect	Observational	Experiment
Treatment imposed	No	Yes
Researcher control	Low	High
Conclusion	Association	Causation
Randomization	Rare	Important

Correlation does not imply causation

Why Observational Studies Still Matter

Experiments are not always possible.

Reasons

- Ethical limitations
- High cost
- Practical difficulties

Example

Researchers cannot force people to smoke for decades to study cancer risk.

4. Basic Experimental Terms

Experimental Unit

- Individual or object studied.

Treatment

- Specific condition applied.

Response Variable

- Measured outcome.

Comparative Experiment

- Compares multiple treatments.

5. Control and Placebo

Control Group

- Baseline group for comparison.

Placebo

- Fake treatment resembling the real one.

Placebo Effect

- Improvement caused by belief in treatment.

6. Blinding

Blinding

- Subjects do not know treatment assignment.

Double-Blind Design

- Neither subjects nor evaluators know assignments.

Double-blind experiments reduce both subject and observer bias.

7. Randomization

Definition
Assigning treatments using chance.

Why It Works

- Balances hidden variables
- Prevents favoritism
- Reduces systematic bias

Randomization is the foundation of valid inference

Matched Pairs Design

- Similar subjects paired together.
- Reduces variability.

Block Design

- Subjects grouped into similar blocks.
- Randomization performed within blocks.

9. Population and Sampling

Population

- Entire group of interest.

Sample

- Subset selected from the population.

Why Sample?

- Saves time
- Reduces cost
- Practical for large populations

10. Simple Random Sampling

Simple Random Sample (SRS)
Every possible sample of a given size has equal probability of selection.

Advantages

- Fair selection
- Reduces bias
- Supports probability methods

Example
Selecting students using random roll numbers.

11. Stratified and Multistage Sampling

Stratified Sampling

- Population divided into strata.
- Sample drawn from each group.

Multistage Sampling

- Sampling performed in stages.

Example:

$$\text{Country} \rightarrow \text{State} \rightarrow \text{School}$$

Sampling vs Experimental Randomization

Sampling Randomization

- Used to select individuals from population.

Experimental Randomization

- Used to assign treatments.

Students commonly confuse these two concepts.

12. Errors in Surveys

Undercoverage

- Some groups excluded.

Nonresponse

- Selected individuals do not respond.

Response Bias

- Incorrect or dishonest responses.

Wording Effects

- Question wording influences answers.

Survey Example

Question
"Do you support strict laws to improve public safety?"

Issue
Question wording encourages agreement.

Improved Version
"Do you support or oppose stricter public safety laws?"

13. Sampling Variability

Different random samples from the same population produce different results.

Example
Two opinion polls may report slightly different percentages.

Variability is natural and unavoidable in sampling.

14. Sampling Distribution

Definition
Distribution of a statistic across repeated samples.

Importance
Foundation for:

- confidence intervals,
- hypothesis testing,
- statistical inference.

15. Bias vs Variability

Aspect	Bias	Variability
Nature	Systematic error	Random fluctuation
Cause	Poor design	Random sampling
Large sample effect	Not reduced	Reduced

Large samples reduce variability, NOT bias

16. Ethics in Data Collection

Institutional Review Board (IRB)

- Reviews studies involving humans.

Informed Consent

- Participants voluntarily agree.

Confidentiality

- Protect participant information.

17. Case Study: Drug Testing

Scenario
New drug tested only on healthier patients.

Problem
Selection bias produces misleading conclusions.

Better Design
Randomized comparative experiment with control group.

18. Case Study: Teaching Method

Scenario
Students voluntarily choose a new teaching method.

Problem
Motivated students may self-select.

Solution
Random assignment of students to teaching methods.

19. Case Study: Election Polling

Scenario
Poll conducted only using landline phones.

Problem
Younger voters underrepresented.

Type of Error
Undercoverage bias.

20. Case Study: Online Reviews

Scenario
Company analyzes only positive online reviews.

Problem
Data are not representative.

Better Approach
Use all reviews or random sampling.

21. Important Definitions

Population: Entire group of interest.

Sample: Subset selected from population.

Randomization: Assignment using chance.

Bias: Systematic deviation from truth.

Sampling Variability: Natural variation among samples.

22. Numerical Example 1

Problem

A company has:

- 60% male employees
- 40% female employees

Sample size required $= 100$

Solution

Male:

$$0.60 \times 100 = 60$$

Female:

$$0.40 \times 100 = 40$$

Sampling method: Stratified sampling.

23. Numerical Example 2

Problem
Population size $= 10$
Sample size $= 2$

Total possible samples:

$$\binom{10}{2} = 45$$

Probability of selecting any specific sample:

$$P = \frac{1}{45}$$

24. Concept Check

Question 1
Is this an observational study or experiment?

Researchers observe sleep patterns of students during exams.

Question 2
Why does increasing sample size not remove bias?

Question 3
Why are double-blind experiments preferred?

25. Key Takeaways

- Good data production is essential.
- Randomization reduces bias.
- Experiments support causation.
- Representative sampling is critical.
- Variability is natural in sampling.
- Large samples reduce variability, not bias.
- Ethics must always be maintained.

Thank You