Selection Bias, Lurking Variables & Confounding Variables

Statistics Worksheet · Grade 9–12

Name: _____ Date: _____

Learning Objectives

- Identify and distinguish the three types of selection bias: undercoverage, non-response bias, and voluntary response bias
- Define and recognize lurking variables in observational studies
- Define and recognize confounding variables and explain how they differ from lurking variables

Problems

1. Match each definition to its correct term. Which type of selection bias occurs when a certain group is inadequately represented in a sample?

2. A student wants to determine who the most popular teacher is in her school. She only asks students who are in her own friend group. Identify the type of selection bias present and explain why it applies.

3. A health organization mails surveys to 1,000 households asking about exercise habits. Only 200 surveys are returned. What type of bias is most likely present, and why?

4. A radio station asks listeners to call in and vote on whether the city should build a new sports arena. Only 300 people call in out of tens of thousands of listeners. Identify the type of selection bias and explain what characterizes it.

5. Read the descriptions in the table and classify each scenario as Undercoverage (U), Non-response Bias (N), or Voluntary Response Bias (V).

Scenario	Type of Bias
An online poll where anyone can choose to submit a response	
A phone survey where many people do not answer unknown numbers	

Scan to watch

Scenario	Type of Bias
A neighborhood survey conducted only in one zip code of a large city	
A TV show asking viewers to text their vote on a topic	

6. A researcher finds that counties with more hospitals also have higher death rates. She concludes that hospitals cause deaths. What type of variable might explain this association without a causal link between hospitals and deaths? Give an example of such a variable in this context.

7. A study finds a positive association between ice cream sales and the number of drowning incidents. Identify the lurking variable and explain how it accounts for both observations.

8. Explain the difference between a lurking variable and a confounding variable. Then classify the variable described: A researcher studying whether stress causes grumpiness discovers that stressed individuals also drink more coffee and sleep less. Is coffee intake a lurking variable or a confounding variable? Justify your answer.

9. A school district notices that students who own more books at home score higher on reading tests. A researcher concludes that owning books improves reading scores. Analyze this scenario: identify any lurking variable(s), explain whether the conclusion is valid, and describe what type of study design would be needed to establish causation.

10. A political campaign conducts a survey by posting a link on their official social media page, asking followers whether they support the candidate. The results show 92% support. Evaluate this survey by identifying ALL sources of bias present (there may be more than one type), explain how each bias affects the results, and suggest one specific improvement for each bias identified.

Scan to watch

Selection Bias, Lurking Variables & Confounding Variables — Answer Key

Statistics Worksheet · Grade 9–12

Answer Key

1. Answer: Undercoverage

- Undercoverage is defined as the inadequate representation of a certain group in the sample.
- When a sample leaves out a significant portion of the population, the coverage of that group is insufficient — this is undercoverage.

2. Answer: Undercoverage — only a small, non-representative subset of the school (her friend group) was sampled.

- The sample consists only of the student's friend group, which is not representative of the entire school population.
- Because a large portion of the school (students outside her friend group) is excluded, this is undercoverage.

3. Answer: Non-response bias — 800 households did not return the survey, and their habits are unknown and unrepresented.

- Non-response bias occurs when participants are unwilling or unable to respond, leaving their data out of the sample.
- 800 out of 1,000 households did not respond, meaning the results only reflect the 200 who chose to reply — likely a biased subgroup.

4. Answer: Voluntary response bias — only listeners with strong opinions chose to call in, making the sample self-selected.

- Voluntary response bias occurs when participants self-select into a study, typically those with strong opinions.
- In this case, listeners who feel strongly about the arena (for or against) are far more likely to call in than those with mild or no opinions, skewing the results.

5. Answer: See completed table

Scenario	Type of Bias
An online poll where anyone can choose to submit a response	V — Voluntary Response Bias
A phone survey where many people do not answer unknown numbers	N — Non-response Bias
A neighborhood survey conducted only in one zip code of a large city	U — Undercoverage
A TV show asking viewers to text their vote on a topic	V — Voluntary Response Bias

- Online polls where anyone can participate are self-selected — Voluntary Response Bias.

Scan to watch

- Phone surveys where people screen calls result in missing data from non-responders — Non-response Bias.
- Surveying only one zip code excludes the rest of the city — Undercoverage.
- TV text-in votes attract only passionate viewers — Voluntary Response Bias.

6. Answer: Lurking variable — for example, population size. Larger populations have more hospitals AND more deaths, but hospitals do not cause the deaths.

- A lurking variable is a third unseen variable that affects both variables being observed, creating a false impression of causation.
- Population size is a lurking variable here: larger counties have more people, so they naturally have more hospitals AND more deaths.
- The association between hospitals and death rates is spurious — both are driven by population size.

7. Answer: Lurking variable: summer heat (temperature). Hot weather increases both ice cream sales and the number of people swimming, which raises drowning risk.

- The lurking variable is summer heat / high temperature.
- When temperatures rise, more people buy ice cream to stay cool — increasing ice cream sales.
- When temperatures rise, more people also go to beaches and pools — increasing the probability of drowning.
- Ice cream sales do not cause drowning; both are caused by the same unseen variable: hot weather.

8. Answer: Confounding variable. Coffee intake is related to the independent variable (stress) and also affects the response variable (grumpiness), so it meets the definition of a confounding variable.

- A lurking variable is unseen and affects an observational study without being linked to the independent variable.
- A confounding variable is related to the independent variable AND affects the dependent (response) variable.
- In this case, stress (independent variable) causes people to drink more coffee, and the coffee (plus resulting lack of sleep) also contributes to grumpiness (response variable).
- Because coffee intake is connected to the independent variable (stress), it is a confounding variable, not merely a lurking one.

9. Answer: Lurking variable: household income / socioeconomic status. Wealthier families buy more books AND invest more in education. The conclusion is not valid from an observational study alone; a controlled experiment (randomly assigning books to households) would be needed.

- The lurking variable is socioeconomic status (SES) or household income.
- Families with higher income can afford more books AND also provide better educational resources, tutoring, and nutrition — all of which improve reading scores.
- The positive association between books and reading scores may be driven by SES, not the books themselves.
- The conclusion is not valid based on this observational study; correlation does not imply causation.
- To establish causation, a controlled experiment would be needed — randomly assigning extra books to some households and comparing outcomes while controlling for other variables.

10. Answer: 1) Voluntary Response Bias — only followers who care enough to click and respond participate; improvement: use random-digit-dial phone survey. 2) Undercoverage — only the campaign's existing social media followers are reached, excluding the general voting population; improvement: use a random sample from voter registration lists. 3) Non-response Bias — many followers may see but not respond to the link; improvement: follow up with non-responders or use

Scan to watch

mandatory-response design.

- Voluntary Response Bias: The survey is self-selected — only followers motivated to click and respond will participate. These are overwhelmingly likely to be supporters, inflating the approval percentage. Improvement: Use a randomly selected sample of registered voters contacted directly.
- Undercoverage: The sample consists only of the candidate's social media followers, who are not representative of the broader voting population. Many voters — especially older demographics — may not follow the campaign online. Improvement: Sample from voter registration rolls or use a stratified random sample of the electorate.
- Non-response Bias: Even among followers who see the post, many will scroll past without responding. Those who respond may be more enthusiastic supporters. Improvement: Use a probability-based sample with deliberate follow-up for non-responders.
- Combined effect: Each bias works in the same direction (over-representing supporters), which is why the reported 92% support is almost certainly a severe overestimate of actual public support.