Unit 3 Kickoff Packet

Collecting Data: Sampling & Study Design

Topics 3.1, 3.2, and 3.3 — Video Follow-Along Assignment

Learning Objectives for This Unit

- **VAR-1.E:** Identify questions to be answered about data collection methods.
- **DAT-2.A:** Identify the type of a study (observational vs. experiment).
- **DAT-2.B:** Identify appropriate generalizations and conclusions based on study type.
- **DAT-2.C:** Identify a sampling method, given a description of a study.
- **DAT-2.D:** Explain why a particular sampling method is or is not appropriate.

Name: ______________________________

Period: __________

PART 1: Unit 3 Cheat Sheet — Key Terms & Definitions

apblue!15 Term	**Definition**
Population	The *entire group* of individuals we want information about. Think: population = *all*.
Sample	A *subset* of the population from which we actually collect data.
Census	Collecting data from *every* individual in the population (often impractical).
Parameter	A number that describes the *population* (often unknown).
Statistic	A number that describes the *sample* (used to estimate parameters).

apblue!15	**Sampling Methods**
Simple Random Sample (SRS)	Every group of n individuals has an *equal chance* of being selected. Use a random number generator.
Stratified Sample	Divide population into *homogeneous* groups (strata), then take an SRS *within each* group.
Cluster Sample	Divide population into *heterogeneous* groups (clusters), then randomly select *entire groups* and sample *all* individuals in chosen groups.
Systematic Sample	Randomly choose a starting point, then sample at *fixed intervals* (e.g., every 20th person).

apblue!15	**Study Types & Conclusions**
Observational Study	No treatments are imposed; we just *observe*. **Cannot determine cause-and-effect.**
Experiment	Treatments *are imposed* on subjects. **Can determine cause-and-effect** (if well-designed).
Bias	Systematic error that causes estimates to be *inaccurate* (not centered at truth).
Variability	How *spread out* estimates are; low variability = high *precision*.

Common Confusion: Cluster vs. Stratified

CLUSTER	**STRATIFIED**
• Ideal groups are **heterogeneous** (mixed) • Select random **groups** • Sample **ALL** in selected groups • Easier but can have high variability	• Ideal groups are **homogeneous** (similar) • Sample **within each** group • Sample **SOME** from every group • More precise but harder to implement

PART 2: Guided Video Notes

Video 1: Topic 3.1 — Do the Data Tell the Truth? [~5 min]

Key Concept: Methods for data collection that do not rely on chance can result in untrustworthy conclusions. The way data is collected *completely informs* how we analyze it.

The Abraham Wald Airplane Problem [0:30–3:30]

1. Abraham Wald worked for the ______________________________, helping the _______________ forces during World War II through mathematical analysis.

2. The British military had a problem: bombers were being shot down over Nazi Germany. They wanted to add heavy ____________ to reinforce planes, but could only put it in ____________ place.

3. Soldiers made bullet hole charts for planes that __________________ from bombing missions and asked Wald where to add armor.

4. **Pause & Think:** Looking at the bullet hole patterns, where would most people initially suggest putting armor? ______________________________

5. **Wald's Key Insight:** The critical realization was about the __________________ in which the data was collected.

6. The planes in the sample were only the ones that ________________________. The bullet holes we see mark places where planes can take hits and still _______________!

7. Therefore, armor should go where we *don't* see bullet holes (like the ____________), because planes hit there never made it back.

Population vs. Sample in the Wald Problem

Population (what we want to know about)	All planes that were hit during bombing missions
Sample (what we actually measured)	Only planes that made it back
Who was missing?	Planes that were shot down
Type of Bias	**Survivorship Bias**

Key Takeaways from Video 1 [4:50–5:12]

8. A proper analysis of data must take into account _________________________________.

9. Sometimes, our samples may not be _____________________ of the whole __________________.

10. The statistician's motto: "Be __________, Be __________, Be ______________. Avoid BS (_________________)."

Video 2: Topic 3.2 — Planning a Study

[∼8 min]

Key Concept: The type of study determines what conclusions we can draw. Observational studies *cannot* establish causation—only experiments can.

Generalization Rules

[3:17–3:57]

1. It is only appropriate to generalize about a population based on samples that are:
   - ______________________________ or otherwise representative, AND
   - Selected from *that specific* ______________________

2. Example: If we study lima beans in humid climates, we ___________________ generalize to black beans, because they are a different population.

The Racial Income Gap: An Observational Study

[3:57–6:10]

3. The 2018 Current Population Survey found:
   - White households median: $70,642 Black households median: $58,665
   - Difference: ≈ $_______________/year → $___________________ over a 45-year career

4. Possible causes include: inequity of schools, familial connections, or direct discrimination. These are called __________________________ variables.

5. This survey is an ______________________________ because no treatments were imposed. Therefore, we ___________________ infer cause and effect.

6. A **retrospective** study examines ___________________ data; a **prospective** study follows individuals into the _______________.

The Résumé Experiment

[6:10–7:47]

7. Researchers sent identical résumés to employers, ______________________ assigning either a commonly white or commonly Black name.

8. They measured the ______________________ rates for both groups.

9. This is an ___________________ because treatments (name type) are ___________________ upon subjects.

10. If well-designed, an experiment _______________ determine a causal relationship.

Fill In: Observational Study vs. Experiment

	Observational Study	**Experiment**
Treatments imposed?	__________	__________
Can show causation?	__________	__________
Example from video	____________________	____________________

Video 3: Topic 3.3 (Part 1) — Random Sampling Methods [~9 min]

Context: San Antonio is one of the most economically **segregated** cities in the U.S.—incomes are similar *within* neighborhoods but vary greatly *between* neighborhoods. We want to estimate the median household income using different sampling methods.

Why Sample? [0:00–2:24]

1. A **census** collects data from ______________ individuals, but censuses are ____________________ to conduct.

2. Due to the right-skewed nature of income data, the ______________ is preferred over the mean.

3. If done well, a random sample should be ________________________ of the general population.

Simple Random Sample (SRS) [2:24–5:00]

4. An SRS is a sample where every ______________ of size n has an ______________ chance of being chosen.

5. **How to conduct:** Number all individuals 1 to N, then use a ___________________________ to select n numbers ____________________ (don't repeat).

6. The SRS sample median was **$50,500**.

Cluster Random Sample [5:00–6:13]

7. Divide the population into _________________ that are near one another.

8. Take an SRS of entire ______________, then sample __________ individuals in selected clusters.

9. In San Antonio, 2 clusters were randomly selected out of 100 regions. The cluster sample median was **$110,350**.

10. Advantage: Easier to collect (only need to visit ____________________ areas).

Stratified Random Sample [6:13–7:20]

11. Divide the population into ______________ based on a similar characteristic.

12. Take an SRS ______________ each stratum, then combine all selected individuals.

13. In San Antonio, 100 homes were randomly selected from *each* of the 100 regions. The stratified sample median was **$51,025**.

Systematic Random Sample

[8:09–8:48]

14. Randomly choose a ________________, then sample at a fixed ________________ interval.

15. Example: In a lunch line, randomly select a number 1–20, then sample every __________ person.

16. Advantage: Very ________________, especially when individuals are "lined up."

Video 4: Topic 3.3 (Part 2) — Evaluating Sampling Methods [~8 min]

Key Question: The three methods gave different estimates ($50,500, $110,350, $51,025). Which should we trust? We evaluate using **bias** (accuracy) and **variability** (precision).

Understanding Bias & Variability [1:42–2:52]

1. **Bias** measures ________________: Are estimates centered at the ________________?
2. **Variability** measures ________________: How spread out are the estimates?
3. The "gold standard" is ________________ bias and _____________ variability.
4. The *true* median household income in San Antonio (from the census) is $_____________.

Label the Dartboards

Match each description: Biased + High Var, Biased + Low Var, Unbiased + High Var, Unbiased + Low Var

Target A (off-center, spread)	**Target B** (off-center, clustered)	**Target C** (centered, spread)	**Target D** (centered, clustered)
______________ ______________	______________ ______________	______________ ______________	______________ ______________

Simulation Results for Each Method [2:52–7:42]

5. **Non-Random Sample** (students picking "representative" areas):
   - Result: _____________—students systematically overestimated.
   - Lesson: Non-random samples can lead to _____________.

6. **Simple Random Sample (SRS):**
   - Bias: ________________ (centered at true median)
   - Variability: ________________
   - Advantage: Unbiased, easy to _____________
   - Disadvantage: Can be difficult to ________________; may not be as precise

7. **Cluster Random Sample:**
   - Bias: ________________ (uses random selection)
   - Variability: _____________ (estimates spread widely)

- Why? Income is homogeneous ______________ regions but varies ______________ regions.
- Works best when clusters are _____________________ and similar to one another.

8. **Stratified Random Sample:**

   - Bias: _________________
   - Variability: ______________ (very precise!)
   - Why? Each sample has a similar ______________ of incomes.
   - Works best when strata are _____________________.
   - Disadvantage: Can be ________________________ to implement.

Fill In: Sampling Methods Comparison

Method	**Bias?**	**Variability**	**Best When...**
Simple Random Sample	__________	__________	_____________________
Cluster Sample	__________	__________	Clusters are ___________
Stratified Sample	__________	__________	Strata are ___________

Key Takeaways from Video 4

[7:42–7:58]

9. Random sampling tends to provide _____________________ estimates.

10. Cluster sampling is effective when clusters are _____________________ and ______________ to each other.

11. Stratified sampling is most effective when strata are _____________________.

PART 3: Check for Understanding

Directions: Answer the following questions to check your understanding of Topics 3.1–3.3. Show your reasoning!

Section A: Observational Studies vs. Experiments

A1. An observational study found that sleep is associated with job performance ($r = 0.86$). A reader concluded that more sleep *causes* better performance. Why is this conclusion **not** correct?

(A) The correlation value should equal 1 for such a conclusion.

(B) The correlation value should be negative for such a conclusion.

(C) The sample was not representative of the population.

(D) Causation cannot be determined from an observational study.

(E) The correlation value implies less than 75% of variability is explained.

Answer: ______ **Explain why:** ____

A2. A researcher studies the effects of an herbal supplement on colds. From 50 people with colds, 25 are assigned to take the supplement and 25 are asked to drink water. The duration of each cold is recorded. What are the **experimental units**?

(A) All people with a cold

(B) The sample of 50 people who had a cold

(C) The 25 people who were given the supplement

(D) The 25 people who drank water

(E) The recorded number of days that the cold lasted

Answer: ______

A3. A researcher selects a simple random sample of 1,200 women who are students at Midwestern colleges. To which population can results be generalized?

(A) All students in the United States

(B) All college students in the United States

(C) All women who are students in the United States

(D) All students at Midwestern colleges in the United States

(E) All women who are students at Midwestern colleges in the United States

Answer: ______ **Why?** ____

Section B: Sampling Methods (Focus on Cluster vs. Stratified)

B1. At a clothing store, clothes are displayed on racks. Clothes on each rack have *similar prices*, but prices *vary greatly between racks.* A consumer randomly selects 4 pieces from *each* rack to estimate the typical price. What type of sample is this?

(A) A census

(B) A cluster sample

(C) A simple random sample

(D) A stratified random sample

(E) A systematic random sample

Answer: ______ **Explain:** ____

B2. A school district has 15 high schools. Seniors' post-graduation plans *vary greatly from one school to the next.* The superintendent selects 5 schools at random and surveys *all* seniors at those schools. What is a disadvantage of this cluster sample?

(A) Cluster sampling is usually too expensive.

(B) The sample will be too large to yield accurate results.

(C) The schools in the sample might not represent all seniors.

(D) Absent students on survey day could affect results.

(E) There is no disadvantage.

Answer: ______ **Explain in terms of bias/variability:**

B3. Margo wants to estimate the percent of red marbles in a bag. She randomly selects a marble, records its color, *puts it back*, shakes the bag, and repeats. What type of sampling is this?

(A) Cluster sampling

(B) Stratified random sampling

(C) Systematic random sampling

(D) Random sampling with replacement

(E) Random sampling without replacement

Answer: ______

Section C: You Try It! — Choosing the Right Method

C1. Scenario: A company has offices in 5 different cities. Commute times are *similar within each city* but *vary greatly between cities*. A researcher wants to estimate the average commute time for all employees.

(a) Should the researcher use **cluster** or **stratified** sampling? Circle one.

(b) Explain your reasoning using the concepts of homogeneity/heterogeneity:

(c) What would happen if the researcher used the *wrong* method? (Discuss variability.)

C2. Scenario: Wildlife biologists want to estimate the deer population in a region with three distinct areas: a forest, a lake, and a town. Deer populations differ by area.

(a) Why would **stratified** sampling be better than a simple random sample for placing observation stations?

(b) If researchers marked deer and recounted them later, what problem arises if the same deer are counted multiple times? How would this affect the population estimate?

Exit Ticket: Unit 3 Kickoff Summary

In 2–3 sentences, explain the **two most important ideas** from Unit 3 so far. Include at least one idea about sampling methods and one about study design.