Data Collection and Sampling

Lecture 7

QBA201 – Quantitative Business Analysis

Shi Qiu – Fall 2026

Observation, experiment and survey

Observation

Record events or characteristics without assigning a treatment.

Experiment

Assign treatments to study their effects. Random assignment helps separate treatment from other influences.

Survey

Ask people to report information.

Discuss: Which method would you use to compare two randomly assigned checkout designs?

A survey has both a sample and respondents

The selected sample is everyone invited under the sampling plan.

Respondents are the people who provide usable answers.

$$\text{Simple response rate} = \frac{\text{completed responses}}{\text{eligible invited units}}.$$

Report how the denominator and a completed response were defined.

Take a short class survey

Two questions ask about your recent QBA201 study time and other courses.

Think about the past 7 days. Answer independently before we discuss the questions together.

Participation is voluntary and does not affect your grade. You may skip a question.

Open the class survey. The same link is below the online slide.

An original question: what exactly is being measured?

Original classroom poll

During a typical week this semester, approximately how many hours do you spend on QBA201 outside of scheduled class time (e.g., studying, homework...)?

Original choices: less than 2; 2–4; 4–6; 6–8; more than 8 hours.

Discuss: Where should someone who spent exactly 4 hours answer? What does "typical" mean?

A second question: the comparison is more subtle

Original classroom poll

Given the quantitative rigor and challenging nature of QBA201, how many hours per week do you typically need to devote outside of class to keep up?

Original choices: 2–4; 4–6; 6–8; 8–10; more than 10 hours.

Discuss: Besides the opening phrase, what else changed?

The original polls changed several things at once

- Framing: a description of difficulty was added.
- Quantity: actual time spent became time needed to keep up.
- Scale: low and high categories changed; boundaries overlapped.

The source polls also had different response counts: 75 and 79.

Their difference does not isolate a wording effect. Hold the other features fixed in a new comparison.

Questionnaire design is part of measurement

- Use short, neutral questions about one clearly defined quantity.
- Specify the time period, unit and activities to include.
- Give nonoverlapping, reasonably complete response options.
- Pretest: ask people what they understood before using the answers.

Plan the analysis while writing the questions. A polished form does not guarantee comparable measurements.

Our class survey holds the measurement fixed

Both versions ask about the past 7 days, use the same two core questions, and offer identical answer choices.

The effort categories are: under 2; 2 to under 4; 4 to under 6; 6 to under 8; 8 to under 10; and 10 or more hours.

Each student sees one randomly assigned version. One version adds framing about difficulty and competing workloads.

Compare answer shares, including the number who answered. A small class may show little or no difference.

Simple random sampling

For a fixed size n, every possible sample of n distinct population members has the same chance of selection.

Example: randomly choose three distinct names from a complete class list.

Equal individual selection probabilities alone do not fully define a simple random sample.

Select 40 distinct tax returns from 1,000

1. Assign one row and unique ID to each eligible return.
2. Generate an independent random number for each row.
3. Sort by the random numbers.
4. Select the first 40 rows.

In Excel, freeze generated values when saving the selected sample.

Discuss: Why is selecting the same return twice unsuitable for a 40-return audit?

Stratified sampling uses every stratum

Partition the population into nonoverlapping groups, then randomly sample within each group.

Example: divide employees into income bands and sample from each band.

Discuss: Why must every employee belong to exactly one stratum?

Proportional allocation follows population shares

Income band	Share	$n = 400$	$n = 1,000$
Under 25,000	25%	100	250
25,000–39,999	40%	160	400
40,000–60,000	30%	120	300
Over 60,000	5%	20	50
Total	100%	400	1,000

Stratum sample size = population share $\times$ n.

Cluster sampling selects groups

Select some groups at random, then observe all units in those selected groups in a one-stage cluster sample.

Example: randomly select apartment buildings and survey every household in those buildings.

Stratified sampling takes units from every stratum. Cluster sampling selects only some clusters.

Systematic sampling uses a random start

To select every tenth record:

1. Choose a starting position at random from 1 to 10.
2. Select that record and every tenth record afterward.

Discuss: What could go wrong if a repeating ten-record pattern is built into the list?

Sampling error arises because a sample is selected

Two valid random samples can have different means even when both procedures are correct.

That difference reflects sampling variability.

Discuss: Does a sample mean different from the population mean prove that data were collected incorrectly?

Nonsampling errors have other sources

- Incorrect measurement, recording or coding.
- Nonresponse that changes the composition of respondents.
- Coverage or selection mechanisms related to the outcome.

Increasing sample size does not by itself correct these problems.

Data-acquisition errors

A result can be wrong because:

- Equipment measures inaccurately.
- A value is copied into the wrong row.
- Respondents interpret the question differently.
- The recorded unit is incorrect.

Discuss: How could a mix of minutes and hours affect a study-time average?

Nonresponse can change who is represented

Suppose busy students are less likely to answer a study-time survey.

The respondents may differ systematically from the selected sample.

Discuss: Would a high response rate prove that nonresponse bias is zero?

Assess who did not respond and how their outcomes might differ.

Selection bias concerns the selection mechanism

A survey of only students present at one optional event can miss important parts of the target population.

More generally, selection related to the outcome can bias an unadjusted estimate.

Discuss: What target group would a survey recruited only through a car park underrepresent?

Explain the recruitment and coverage limits before generalizing.