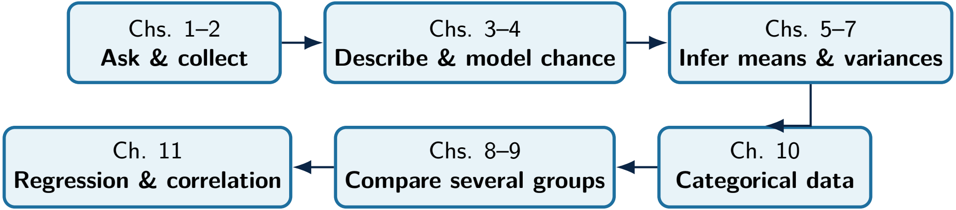
Lecture 1
Statistics and the Scientific Method

STAT 515 — Statistical Methods

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Course arc: Chapters 1–11



Chapter Outline

What Is Statistics?

Definition

Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision making and scientific discovery when the available information is both *limited* and *variable*.

- In short: statistics is the science of **Learning from Data**.
- Everyone deals with data: sales figures, crime rates, contamination levels, survival rates, census figures, consumer choices, . . .

The Four-Step Process of Learning from Data

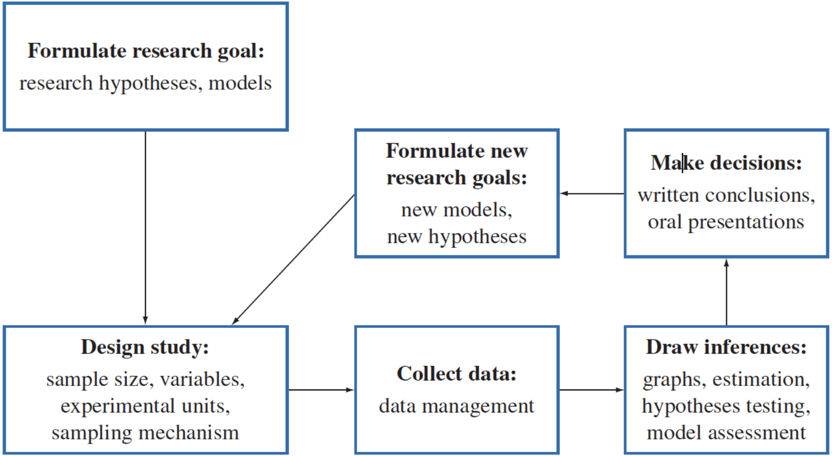
1. **Defining the Problem** — what question are we trying to answer?
2. **Collecting the Data** — designing a study or experiment
3. **Summarizing the Data** — descriptive statistics and graphics
4. **Analyzing the Data, Interpreting the Analyses, and Communicating the Results**

All four steps matter: if the problem isn't clearly defined or the data are collected improperly, even a technically correct analysis can be **misleading**.

Organization of the Textbook

Four-Step Process	Chapters
1. Defining the Problem	Ch. 1
2. Collecting the Data	Ch. 2
3. Summarizing the Data	Ch. 3–4
4. Analyzing, Interpreting, and Communicating the Results	Ch. 5–19 (this course covers Ch. 1–11)

Scientific Method Schematic



The Headline

A campus headline says: *"Students study 18 hours per week on average."* What questions should a statistician ask before believing it?

Questions About the Population and Sample

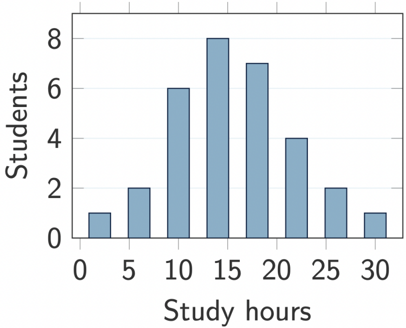
- **Target population**: all college students? Students at this campus?
- **Sampled population**: who was *actually* sampled — does it match the target?
- **Sampling design**: simple random sample, or a convenience sample (library patrons, an online poll anyone could click)?
- **Sample size** n: 20 students vs. 2,000 students give very different credibility.
- **Nonresponse bias**: could only unusually diligent (or unusually lazy) students have responded?

Questions About How the Data Were Collected

- **Observational vs. objective measurement**: self-reported survey, or tracked study-app data?
- **Self-report reliability**: people round up, misremember, or overstate study time.
- **Definition of "studying"**: does it include group work, review videos, reading class emails? Vague definitions inflate or deflate the number.
- **Time frame**: one typical week, or averaged across a semester (including exam weeks)?

Questions About the Number Itself

- **Mean or median?** The mean is sensitive to outliers — a few students reporting 60+ hours/week can pull the mean up substantially.
- **Variability**: what is the standard deviation, range, or IQR? An average of 18 hours can hide enormous spread (some at 2 hours, some at 40).
- **Margin of error / confidence interval**: is any measure of uncertainty reported, or just a bare point estimate?



Questions About the Reporting

- **Who conducted the study**, and is there a motive to report a flattering or alarming number?
- **Is there a comparison**? "Statistically significant" difference from a prior benchmark, or just a number presented in isolation?

The Big Picture

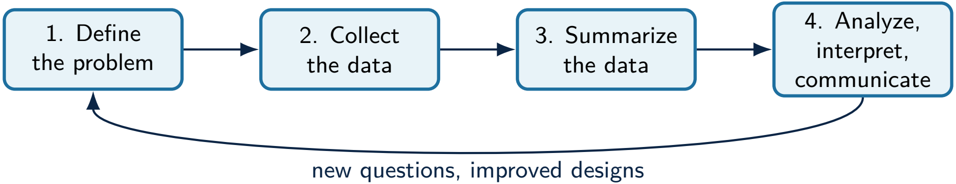
Key Takeaway

A single average, reported without information about **how it was collected**, **from whom**, and **how spread out** the data are, tells you very little.

That is the gap between *“a number”* and *“a trustworthy statistic.”*

This exercise previews ideas from Ch. 1 (population/sample, scientific method), Ch. 2 (sampling design, nonresponse), and Ch. 3 (mean vs. median, variability).

The four-step process: learning from data



Course habit

Every method we learn later should be read through these questions: What was asked? How were data obtained? What was summarized? What conclusion is justified?

Example 2: Turn a vague issue into a data problem

Vague issue

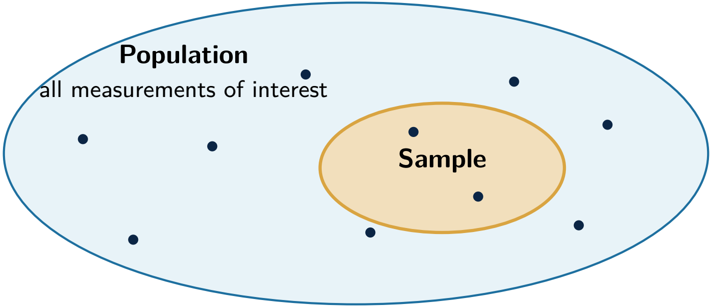
Students say campus buses are unreliable.

Better statistical question

For weekday rides between 8 a.m. and 5 p.m., what proportion of buses arrive more than 5 minutes late?

1. Define **late**: more than 5 minutes after schedule.
2. Define **population**: all eligible weekday rides this term.
3. Choose **sample**: randomly selected routes and times.
4. Summarize: proportion late, delay distribution.
5. Communicate: uncertainty, limitations, action options.

Population and sample



The Population

Set of all measurements.

The Sample

Set of measurements selected from the population.

Four Motivating Problems

1. **Ground beef inspection**: Instead of inspecting all 100,000 packages/day, sample 250 packages to estimate the fraction improperly labeled.
2. **Smoking cessation and weight gain**: 400 participants weighed before/after; average gain of 5 lb — is this a *real* effect or random variation?
3. **Nitrogen fertilizer and wheat yield**: 15 fields, 5 nitrogen rates (3 fields each) — can conclusions generalize beyond the study fields?
4. **Public opinion polls**: How can polling $\sim$1,500 people estimate the opinions of 195$+$ million voters?

What These Examples Teach Us

- Every one of these problems requires the **four-step process**.
- Sample size and sampling method must be planned *before* data collection.
- Results are only generalizable to a population if the sample is representative of it.
- Reporting must be unambiguous: decisions (fines, drug approvals, policy) depend on a clearly communicated statistical conclusion.
- Later chapters develop the tools to answer: *is an observed difference real, or due to chance?*

Why study statistics?

Consumer

Evaluate claims, charts, polls, risks, and headlines.

Professional

Design studies, monitor quality, forecast, and compare alternatives.

Citizen

Understand evidence used in policy, medicine, law, and public debate.

The goal is not calculation alone.

The goal is better reasoning from data.

Common traps in data stories

- **Undefined population:** Who exactly does this conclusion describe?
- **Convenience sample:** Were easy-to-reach people treated as representative?
- **Missing denominator:** Is a count presented without a rate?
- **Confounding:** Could another variable explain the apparent relationship?
- **Overclaiming:** Does the conclusion go beyond the design?
- **Tiny practical effect:** Is the effect meaningful, not just detectable?

Good statistics starts with good questions.

Activity: identify the pieces

Scenario

A university surveys 300 students from an online volunteer pool and finds that 62% support extending library hours during finals week.

1. What is the likely target population?
2. What is the sampled population?
3. What is the sample?
4. What statistic is reported?
5. What possible bias should be disclosed?

Rewrite the conclusion in a way that is accurate and cautious.

Activity: identify the pieces—answer

1. **Likely target population:** All currently enrolled students at the university, assuming the university wants to make a campus-wide conclusion.
2. **Sampled population:** All university students in the online volunteer pool who had an opportunity to participate in the survey.
3. **Sample:** The 300 students who completed the survey.
4. **Reported statistic:** The sample proportion
$$\widehat{p} = 0.62.$$
Thus, 62% of the surveyed students supported extending library hours.
5. **Possible bias:** Volunteer or self-selection bias, together with possible undercoverage. Students who join the volunteer pool may be more engaged, use the library more frequently, or have stronger opinions than students generally.

Accurate and cautious conclusion

Among the 300 students surveyed from the university's online volunteer pool, 62% supported extending library hours during finals week. Because the respondents came from a volunteer pool, this result may not represent the opinions of the entire university student body.

Quick check: Chapter 1

Question

A sample of 40 water bottles has an average fill of 502.1 mL. The factory wants to know the average fill of all bottles produced today.

- Population?
- Sample?
- Parameter?
- Statistic?

Answer

- All bottles produced today.
- The 40 measured bottles.
- True mean fill of all bottles today, μ.
- Sample mean fill, $\bar{y} = 502.1$ mL.

Current Applications of Statistics

- **Data mining**: extracting patterns from massive data sets — forecasting, risk assessment, sequence identification, clustering.
- **Drug development**: e.g., the 1954 Salk polio vaccine trial — 400,000 children randomized to vaccine vs. placebo; statistics determined sample size and demonstrated efficacy.
- **Forensic science**: statistical design/analysis improves the reliability of evidence (e.g., re-analysis of Comparative Bullet Lead Analysis evidence).
- **Wildlife population estimation**: e.g., estimating bowhead whale population size and growth rate from visual/acoustic census data.
- **Environmental science**: modeling ozone exposure across a city and time to inform public health policy.

Chapter 1 Summary

- Statistics = the science of **Learning from Data** via a **four-step process**: define, collect, summarize, analyze/interpret/communicate.
- A **population** is the full set of measurements of interest; a **sample** is a subset used to make inferences about the population.
- Statistics is essential across science, medicine, law, business, government, and everyday decision making.
- Being statistically literate helps you avoid — and detect — common misinterpretations of data.
- Next: **Chapter 2** — how do we actually collect data (surveys and experiments)?