Statistics and Data Analysis
Unit 03 – Lecture 01 Notes
Population, Sample, Sampling, Estimation and Confidence Intervals

Tofik Ali

February 17, 2026

Topic

Population vs sample; sampling techniques; estimation; confidence intervals (overview).

How to Use These Notes

These notes are written for students who are seeing the topic for the first time. They follow the slide order, but add the missing 'why', interpretation, and common mistakes. If you get stuck, look at the worked exercises and then run the Python demo.

Course repository (slides, demos, datasets): `https://github.com/tali7c/Statistics-and-Data-Analysis`

Time Plan (55 minutes)

- 0–10 min: Attendance + recap of previous lecture
- 10–35 min: Core concepts (this lecture's sections)
- 35–45 min: Exercises (solve 1–2 in class, rest as practice)
- 45–50 min: Mini demo + interpretation of output
- 50–55 min: Buffer / wrap-up (leave 5 minutes early)

Slide-by-slide Notes

Title Slide

State the lecture title clearly and connect it to what students already know. Tell students what they will be able to do by the end (not just what you will cover).

Quick Links / Agenda

Explain the structure of the lecture and where the exercises and demo appear.

- Overview
- Sampling
- Confidence Intervals
- Exercises
- Demo
- Summary

Learning Outcomes

- Differentiate population parameters and sample statistics
- Explain sampling bias vs random error
- Describe common sampling methods (SRS, stratified, cluster)
- Compute and interpret a basic confidence interval for a mean

Why these outcomes matter. A **population** is the entire group we care about (all students, all sensors, all invoices, etc.). A **sample** is the smaller subset we actually observe. Most mistakes in statistics happen when we forget that we only see a sample, but we want to make a statement about the population. A **parameter** is a fixed (but usually unknown) number that describes the population (e.g., μ, σ). A **statistic** is a number computed from the sample (e.g., $\bar{x}$, s). Statistics vary from sample to sample, which is why we talk about uncertainty.

Sampling: Key Points

- Population vs sample
- Bias vs random error
- Representative sampling matters

Explanation. A **population** is the entire group we care about (all students, all sensors, all invoices, etc.). A **sample** is the smaller subset we actually observe. Most mistakes in statistics happen when we forget that we only see a sample, but we want to make a statement about the population. **Bias** is a systematic error: your method tends to be wrong in the same direction again and again (too high or too low). It does not disappear by taking more samples if the sampling process is flawed. Fixing bias usually requires changing the data collection procedure (sampling frame, selection method, non-response handling). **Random error** is natural variation due to sampling and measurement noise. Unlike bias, random error can be reduced by increasing sample size or improving measurement quality. In many formulas you see a $\sqrt{n}$ in the denominator: that is the mathematical reason larger samples give more stable estimates.

Confidence Intervals: Key Points

- Interpretation: long-run coverage
- Width depends on n and variability
- CIs support decision-making with uncertainty

Explanation. A **confidence interval (CI)** is an interval estimate, not a single number. The correct interpretation is long-run: if we repeated the same sampling procedure many times, about 95% of the computed 95% intervals would contain the true population value. It is **not** correct to say there is a 95% probability the parameter lies in a particular computed interval. The **significance level** α is the maximum Type I error rate you are willing to tolerate: the probability of rejecting H_0 when H_0 is actually true. Common choices are 0.05 or 0.01, but the right value depends on consequences of false alarms vs missed detections.

Confidence Intervals: Key Formula

$$\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

How to read the formula. A **confidence interval (CI)** is an interval estimate, not a single number. The correct interpretation is long-run: if we repeated the same sampling procedure many times, about 95% of the computed 95% intervals would contain the true population value. It is **not** correct to say there is a 95% probability the parameter lies in a particular computed interval. The **significance level** α is the maximum Type I error rate you are willing to tolerate: the probability of rejecting H_0 when H_0 is actually true. Common choices are 0.05 or 0.01, but the right value depends on consequences of false alarms vs missed detections.

Exercises (with Solutions)

Attempt the exercise first, then compare with the solution. Focus on interpretation, not only arithmetic.

Exercise 1: Parameter vs Statistic

Give one example of a parameter and one of a statistic.

Solution

- Parameter: population mean
- Statistic: sample mean

Walkthrough. A **population** is the entire group we care about (all students, all sensors, all invoices, etc.). A **sample** is the smaller subset we actually observe. Most mistakes in statistics happen when we forget that we only see a sample, but we want to make a statement about the population. A **parameter** is a fixed (but usually unknown) number that describes the population (e.g., μ, σ). A **statistic** is a number computed from the sample (e.g., $\bar{x}$, s). Statistics vary from sample to sample, which is why we talk about uncertainty.

Exercise 2: CI Interpretation

In one sentence, what does a 95% CI mean (correctly)?

Solution

- About 95% of such intervals contain the true mean in repeated sampling.

Exercise 3: Bias Scenario

Why does convenience sampling create bias?

Solution

- Because some groups are over/under-represented systematically.

Walkthrough. **Bias** is a systematic error: your method tends to be wrong in the same direction again and again (too high or too low). It does not disappear by taking more samples if the sampling process is flawed. Fixing bias usually requires changing the data collection procedure (sampling frame, selection method, non-response handling).

Mini Demo (Python)

Run from the lecture folder:

```
python demo/demo.py
```

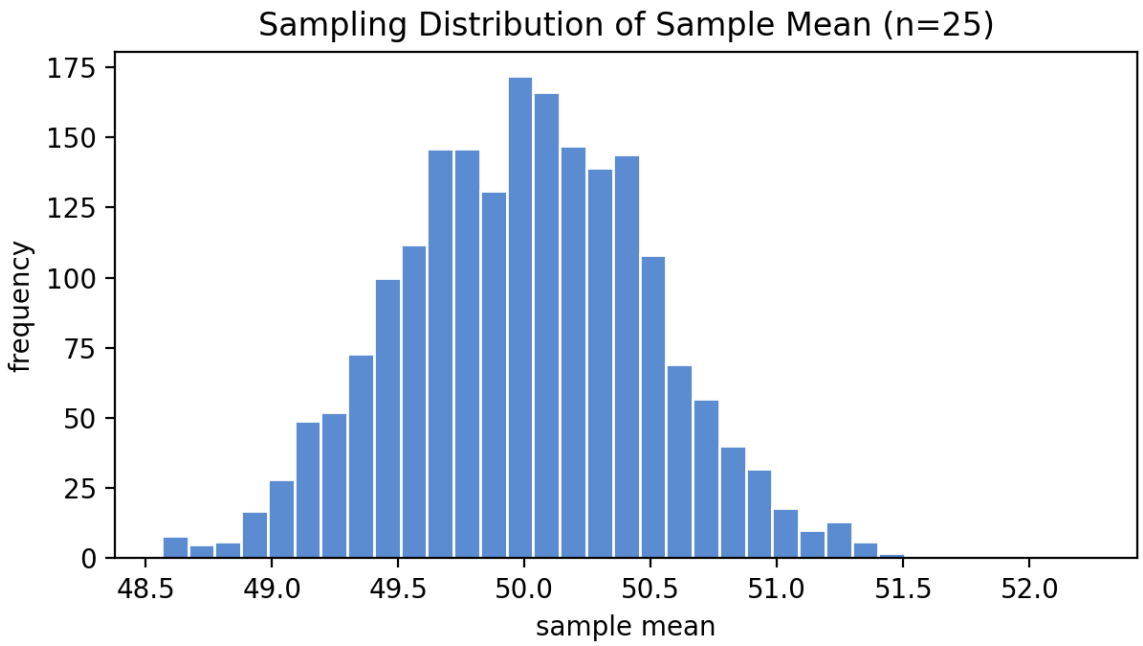
Output files:

- `images/demo.png`
- `data/results.txt`

What to show and say.

- Creates a random sample and computes a 95% CI for the mean (t-interval).
- Plots a histogram of many sample means to show the sampling distribution.
- Use it to discuss why larger n gives a narrower CI (smaller standard error).

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

If your CI is too wide, what two actions reduce its width (without cheating)?

Suggested answer (for revision). Increase sample size n and reduce variability s (better measurement or a more homogeneous/stratified sample).

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview **Sampling** **Confidence Intervals** **Exercises** **Demo** **Summary**

Agenda

- Overview
- Sampling
- Confidence Intervals
- Exercises
- Demo
- Summary

Learning Outcomes

- Differentiate population parameters and sample statistics
- Explain sampling bias vs random error
- Describe common sampling methods (SRS, stratified, cluster)
- Compute and interpret a basic confidence interval for a mean

Sampling: Key Points

- Population vs sample
- Bias vs random error
- Representative sampling matters

Confidence Intervals: Key Points

- Interpretation: long-run coverage
- Width depends on n and variability
- CIs support decision-making with uncertainty

Confidence Intervals: Key Formula

$$\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

Exercise 1: Parameter vs Statistic

Give one example of a parameter and one of a statistic.

Solution 1

- Parameter: population mean
- Statistic: sample mean

Exercise 2: CI Interpretation

In one sentence, what does a 95% CI mean (correctly)?

Solution 2

- About 95% of such intervals contain the true mean in repeated sampling.

Exercise 3: Bias Scenario

Why does convenience sampling create bias?

Solution 3

- Because some groups are over/under-represented systematically.

Mini Demo (Python)

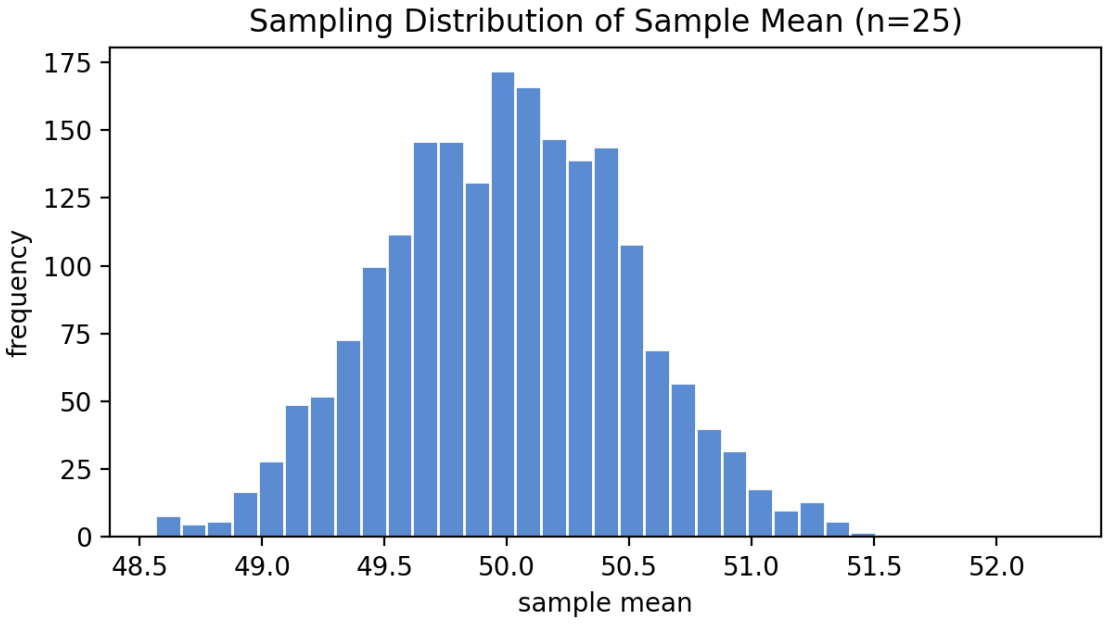
Run from the lecture folder:

```
python demo/demo.py
```

Outputs:

- `images/demo.png`
- `data/results.txt`

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

If your CI is too wide, what two actions reduce its width (without cheating)?