AS2101*: Data and Statistics

12 Lab Groups

Part 1: Things You Need to Know

Why Should You Care?

You test a rocket engine 10 times. You get 10 different thrust values.
Questions you MUST answer:

- What is my **best estimate** of the thrust?
- How much **variability** does this engine have?
- If I test it again, what range will I likely see?

Statistics is how you answer these questions with confidence.

1. Sample Mean ($\bar{x}$) — Your Best Guess

The average of your measurements. The number that best represents your data.

$$\bar{x} = \frac{\text{sum of all measurements}}{\text{number of measurements}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Example: You measure rocket thrust 5 times (kN): 450, 452, 449, 451, 448

$$\bar{x} = \frac{450 + 452 + 449 + 451 + 448}{5} = 450.0 \text{ kN}$$

Interpretation: "Based on these 5 tests, our best guess of the engine's thrust is 450 kN."

2. Sample Standard Deviation (s) — How Spread Out Your Data Is

A measure of how much your measurements bounce around. **Noise. Scatter. Variability.**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

Why $n - 1$? You used the data to calculate the mean first. This "uses up" one degree of freedom. Just use $n - 1$ — it's correct.

Example: Thrust data: 450, 452, 449, 451, 448. $\bar{x} = 450$, $s = 1.58$ kN.

Interpretation: "Individual tests vary by about ±1.6 kN from the average."

Rule of Thumb: For roughly bell-shaped data, about **2/3 of measurements** fall within $\bar{x} \pm 1s$, and **95%** fall within $\bar{x} \pm 2s$.

3. Standard Error of the Mean (SEM) — How Sure You Are About the Mean

Your uncertainty in the **mean value**. If you repeated the whole experiment (5 more tests), how much would the new average bounce around?

$$\boxed{\text{SEM} = \frac{s}{\sqrt{n}}} \tag{3}$$

Example: $s = 1.58$, $n = 5$, $\text{SEM} = 1.58/\sqrt{5} = 0.71$ kN

Interpretation: "My uncertainty in the 450 kN average is about ± 0.7 kN."

Note: s (standard deviation) tells you about individual tests. SEM tells you about the average. **SEM is always smaller than** s (unless $n = 1$).

4. 95% Confidence Interval (CI) for the Mean — Error Bars

A range that **plausibly contains the true mean**. If you could run the experiment 100 times, the true mean would fall inside this interval about 95 times.

$$\boxed{95\% \text{ CI} = \bar{x} \pm t \times \text{SEM}} \tag{4}$$

Note t comes from the **t-table**. [For $n = 5$, $t \approx 2.78$. For $n = 30$, $t \approx 2.04$. For $n > 100$, $t \approx 1.96$].

Example: $\bar{x} = 450$, $\text{SEM} = 0.71$, $n = 5$, $t = 2.78$

$$95\% \text{ CI} = 450 \pm 2.78 \times 0.71 = [448.0, 452.0]$$

Interpretation: "We are 95% confident that the engine's true average thrust is between 448.0 and 452.0 kN."

This is what you put as error bars on graphs!

5. Sample Correlation (r) — Do Two Things Move Together?

A number between -1 and 1 that tells you if two variables are linearly related in your sample. Pearson Correlation can be used to quantify this.

r **value**	**Meaning**	**Example**
0.9	Strong positive	Lift increases with angle of attack
0.0	No relationship	Wing color vs. lift coefficient
-0.9	Strong negative	Drag coefficient decreases with Reynolds number

WARNING: Correlation is **not** causation. Ice cream sales and shark attacks are correlated. Eating ice cream does not cause sharks.

6. Synthetic Data Generation

- You need 1,000 samples for a simulation, but you only tested 10 times.
- You want to see what **might** happen in future tests.

Assume Normal (Gaussian) Distribution (bell-shaped data) or other distributions

1. Calculate $\bar{x}$ and s from your sample.

2. Use Python to generate new random numbers following a Normal distribution with the same $\bar{x}$ and s.

The Workflow

1. **RAW DATA:** You have n measurements (thrust, drag, strain, etc.)
2. **COMPUTE:** $\bar{x}$, s, SEM, 95% CI
3. **ANSWER:** t-tests, correlations etc
4. **GENERATE:** synthetic data using distributions like normal, log-normal or, bootstapping.

Part 2: Aerospace Engineering Lab Problems

Instructions: You will be assigned with ONE dataset. Complete all tasks for your group and submit a lab report.

Part 3: Easy Reference

Descriptive Statistics

Sample Mean:

$$\bar{x} = \frac{1}{n}\sum x_i$$

Sample Std Dev:

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

Standard Error:

$$\text{SEM} = \frac{s}{\sqrt{n}}$$

95% CI for Mean:

$$\bar{x} \pm t_{0.025,n-1} \times \text{SEM}$$

Hypothesis Tests

One-sample t-test (μ_0 expected value or mean):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Two-sample t-test (independent):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Correlation & Regression

Pearson correlation:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Linear regression:

$$y = c + mx$$

Synthetic Data Generation

Using Normal Distribution:

```
np.random.normal(loc=xbar,
    scale=s, size=n)
```

Using Log-normal Distribution:

```
np.random.lognormal(
    mean=ln_xbar, sigma=ln_s,
    size=n)
```

Part 4: t-Table for 95% Confidence Intervals

Degrees of Freedom (df = n-1)	**t-value (two-tailed,** $\alpha = 0.05$**)**
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
6	2.447
7	2.365
8	2.306
9	2.262
10	2.228
11	2.201
12	2.179
13	2.160
14	2.145
15	2.131
16	2.120
17	2.110
18	2.101
19	2.093
20	2.086
25	2.060
30	2.042
40	2.021
50	2.009
60	2.000
80	1.990
100	1.984
∞	1.960

How to use: Find your sample size n, then $df = n - 1$. For $n = 5$, $df = 4$, $t = 2.776$. For $n = 30$, $df = 29$, $t \approx 2.045$ (between 2.042 and 2.060).