

LECTURE 5

Estimation and Hypothesis Testing

Estimators & bias • sampling distributions • bootstrap confidence intervals • permutation tests • testing a difference in means and a correlation • chi-squared tests • errors & power

Principles and Methods of Data Analysis in the Energy Industry

Think Stats, Ch. 8-9 • Elements of Data Science, Ch. 11-13

From describing a sample to trusting what it tells us

A

Estimation

Using a sample statistic to estimate a population parameter, and how sure to be about it

B

Bias & Choosing Estimators

When a sample misrepresents the population, and comparing estimators by their error

C

Classical Hypothesis Testing

Is an observed effect real, or could it have happened by chance alone?

D

More Tests, Errors & Power

Testing a correlation and a proportion, and the two ways a test can be wrong

Running example: still the GECAD photovoltaic dataset (Lectures 1-4, Labs 1-3) — now asking not just "what did we measure?" but "how much would that answer change if we had collected a different sample of days?"

By the end of this lecture, you will be able to

- Explain what an estimator is, and why the sample mean is only one of several possible estimators of a population parameter.
- Describe a sampling distribution, and explain why its spread (the standard error) shrinks as sample size grows.
- Build a bootstrap confidence interval for a statistic by resampling the data itself, with no formula required.
- Recognize sample selection bias, and compare two estimators of the same parameter using mean squared error (MSE).
- State the four-step classical hypothesis-testing framework and compute a p-value with a permutation test.
- Apply a permutation test to a difference in means and to a correlation coefficient, and a chi-squared test to a table of counts.
- Distinguish Type I and Type II errors, describe what statistical power depends on, and explain why replication matters.



PART A

Estimation

Using a sample to estimate a population parameter

A sample statistic standing in for an unknown parameter

- So far, every number we computed (a mean, a median, a correlation) came from the data we HAD — the 25 days of January 2013 that happen to be in this dataset.
- But the question an engineer usually cares about is bigger than those 25 days: "What is the TRUE average peak output of this inverter model, in general?"
- That true, unknown number is a population parameter. The number we compute from our sample — the sample mean — is an estimator of it.
- Estimation is the process of using a sample statistic to infer a population parameter, and it always comes with a second question attached: how far might this estimate be from the truth?

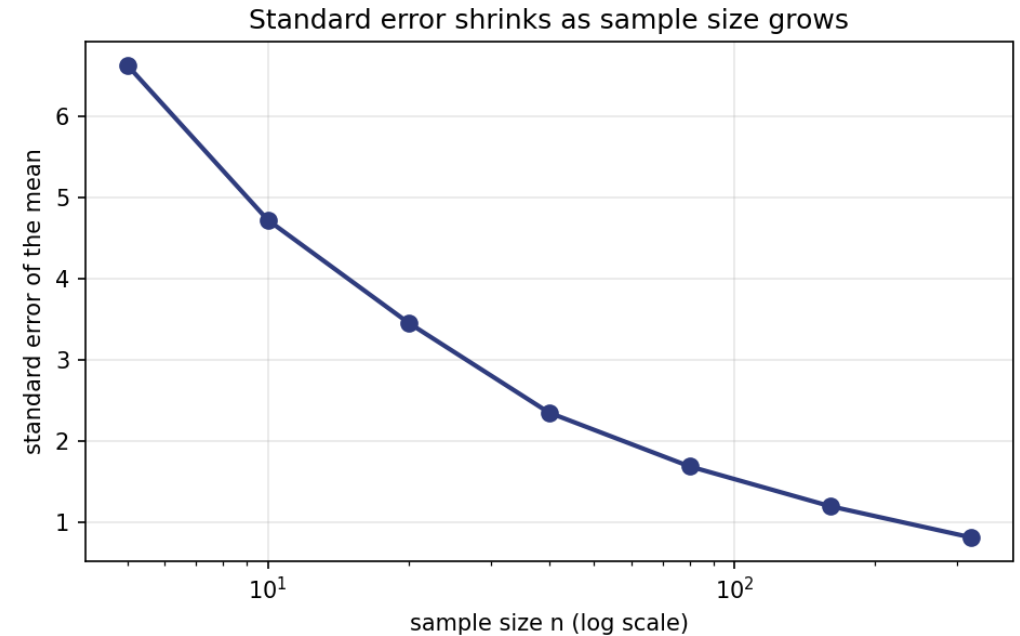
Bias and mean squared error (MSE)

- An estimator is unbiased if, averaged over many hypothetical samples, it gets the parameter right on average — it is not systematically too high or too low.
- The sample mean is an unbiased estimator of the population mean. The sample variance needs a small correction (dividing by $n-1$, not n) to be unbiased — this is why pandas' `.std()` and `.var()` default to that correction.
- Bias is not the only thing that matters: an estimator can be unbiased but wildly variable from sample to sample, which is just as unhelpful in practice.
- Mean squared error (MSE) combines both problems into one number: $MSE = \text{mean}[(\text{estimate} - \text{true value})^2]$, averaged over many hypothetical samples. A good estimator keeps MSE small.

What would happen if we collected many samples?

- Imagine repeating the data collection many times: 25 NEW days, measured the same way. Each repetition gives a slightly different sample mean.
- The distribution of all those hypothetical sample means is the sampling distribution of the mean. Its spread — the standard error (SE) — tells you how much a single estimate can be expected to wander from the truth.
- You cannot usually collect those extra samples in real life. But simulation (or, as the next slides show, resampling the one sample you DO have) lets you approximate this distribution anyway.

The pattern SE proportional to $1/\sqrt{n}$ is one of the most useful facts in statistics: to cut your uncertainty in half, you need about 4x as much data, not 2x.

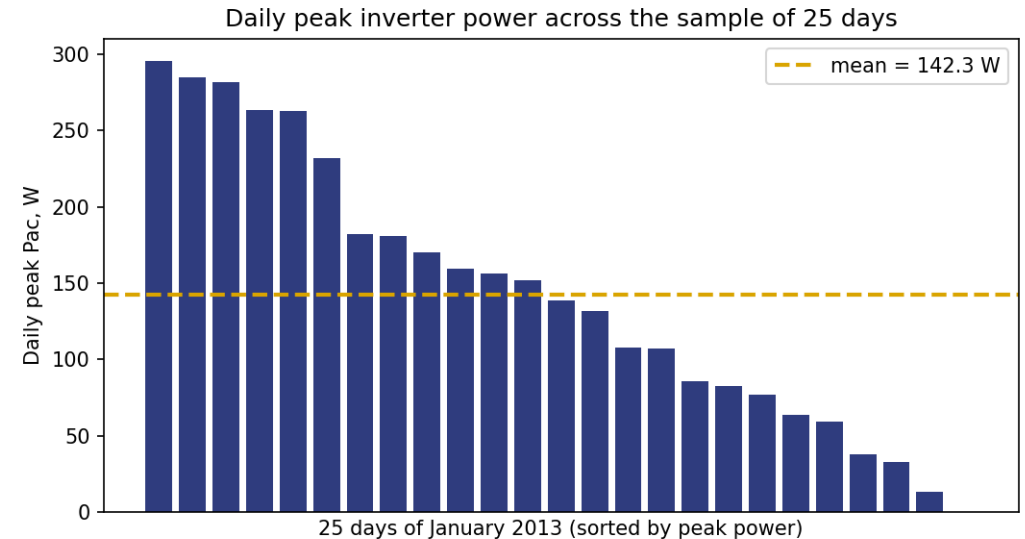


Simulated: standard error of the mean, for samples of different sizes n

WORKED EXAMPLE

Estimating the mean daily peak power of the inverter

- Sample: the maximum Pac reading recorded on each of the 25 available days in January 2013 (n = 25 "daily peak" values).
- Sample mean = 142.26 W, sample standard deviation = 89.26 W.
- Analytic standard error of the mean = $\text{sample_std} / \sqrt{n} = 89.26 / \sqrt{25} \approx 17.85$ W.
- That 25-day sample is itself just one month out of a whole winter — the question for the next slides is how to turn that standard error into a usable confidence interval, using the sample we actually have.



Daily peak Pac, 25 days (02-01-2013 has no daylight readings)

THE BOOTSTRAP

Resampling the sample itself to approximate the sampling distribution

```
import numpy as np

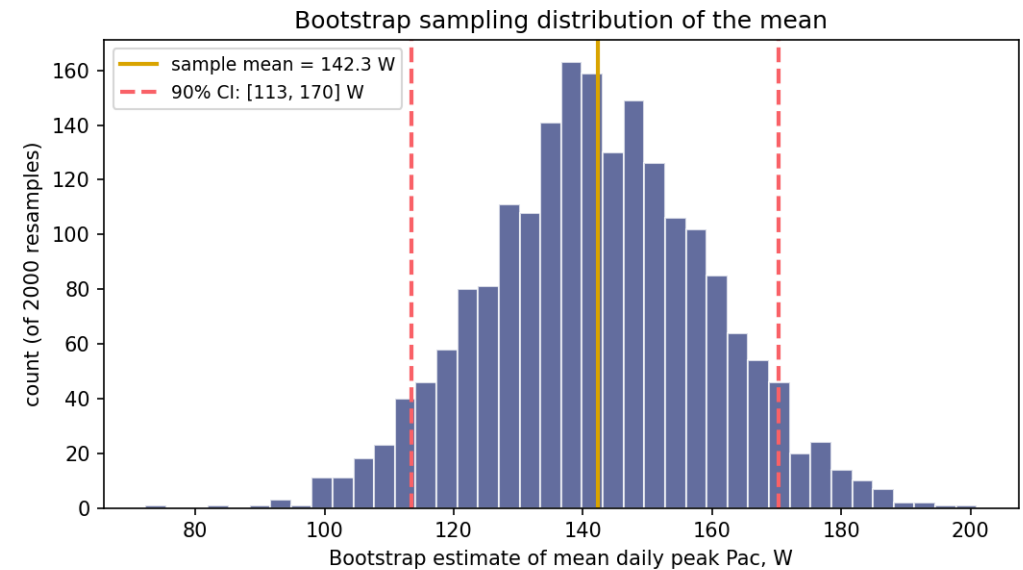
boot_means = []
for _ in range(2000):
    resample = np.random.choice(
        daily_peak, size=len(daily_peak), replace=True)
    boot_means.append(resample.mean())
```

- The bootstrap idea: treat the sample you have as a stand-in for the whole population, and draw new "samples" FROM it, with replacement, the same size as the original.
- Each resample gives a slightly different mean, purely because of which days got picked (some picked twice, some not at all). Repeating this thousands of times builds an approximate sampling distribution — with no formula, and no assumption about the shape of the underlying data.
- This is exactly the technique behind Elements of Data Science, Ch. 11-12 (Resampling and Bootstrap Sampling).

WORKED EXAMPLE

A 90% bootstrap confidence interval for the mean peak Pac

- 2,000 bootstrap resamples of the 25 daily-peak values, each resampled with replacement.
- Bootstrap standard error of the mean = 17.35 W — essentially identical to the analytic 17.85 W from slide 8, which is reassuring: the two independent methods agree.
- 90% confidence interval (5th to 95th percentile of the 2,000 bootstrap means) = [113.4, 170.3] W.
- Interpretation: if we repeated this whole data-collection and resampling procedure many times, about 90% of the intervals built this way would contain the true mean daily peak power.



Bootstrap sampling distribution of the mean (2,000 resamples)

PART B

Bias & Choosing Estimators

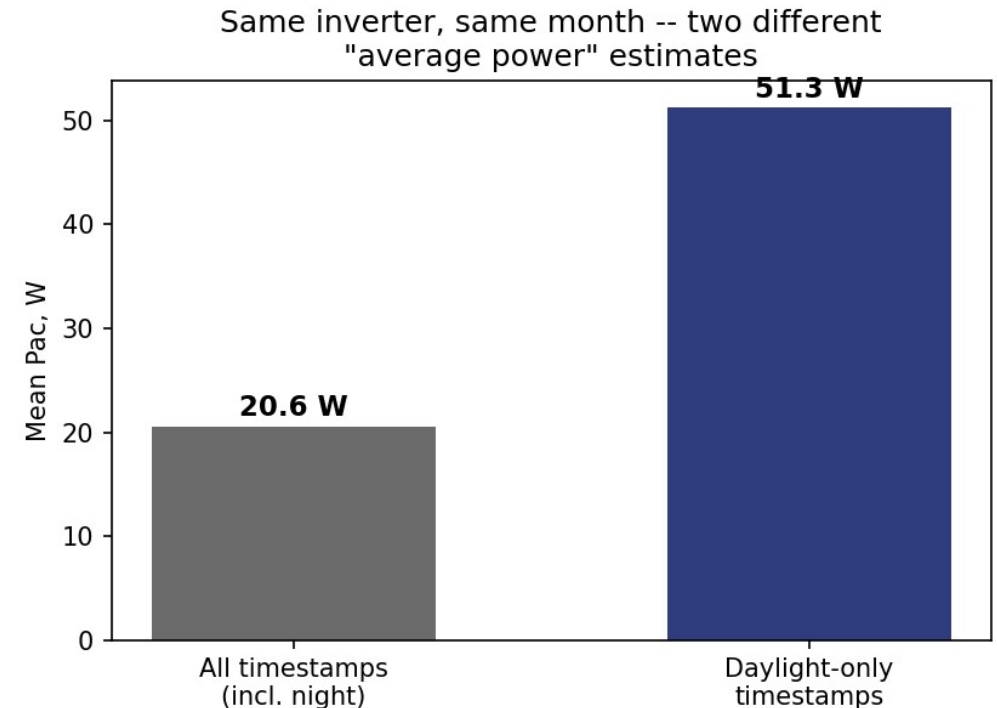
When a sample misrepresents the population, and comparing estimators



SAMPLE SELECTION BIAS

The same inverter, two very different "average power" numbers

- Suppose two students are each asked to estimate the inverter's typical generating power from the same January dataset.
- Student 1 averages Pac over EVERY five-minute timestamp in the dataset, including the roughly two-thirds of readings recorded at night.
- Student 2 first filters to daylight-only timestamps (the same IntSolIrr > 1 rule used in Labs 1-3), then averages.
- Both are "correct" arithmetic — but they are estimating two different things, and the gap between 20.6 W and 51.3 W is entirely about WHICH population each sample actually represents.



Mean Pac: all timestamps vs. daylight-only timestamps

The lesson: before trusting an average, always ask what population your sample actually represents — "average power, day and night" and "average power while generating" are both valid questions, but they are not the same question.

COMPARING TWO ESTIMATORS

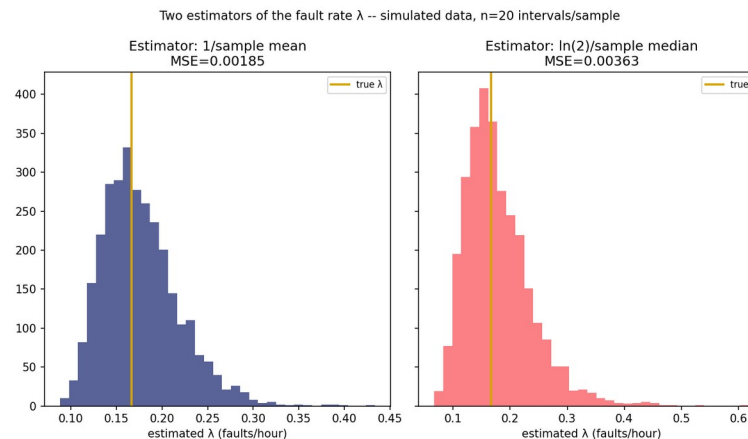
Simulated: estimating a grid-fault rate from inter-fault intervals

- Scenario (simulated, clearly labeled, following Think Stats Ch. 8's exponential-distribution example): time between grid faults follows an exponential distribution with a true rate $\lambda = 1/6$ faults per hour (one fault every 6 hours, on average).
- Given a sample of $n = 20$ observed intervals, two different estimators of λ are possible:

$$\lambda_{1} = 1 / \text{sample_mean}(\text{intervals})$$

$$\lambda_{2} = \ln(2) / \text{sample_median}(\text{intervals})$$

- Simulating 3,000 samples of $n = 20$ and computing both estimators' MSE: the mean-based estimator's MSE (0.00181) is about half the median-based estimator's MSE (0.00380) — the mean-based estimator is the better choice here.



PART C

Classical Hypothesis Testing

Is an observed effect real, or could chance alone explain it?

Given a sample and an apparent effect...

- The central question: "Given a sample and an apparent effect, what is the probability of seeing an effect this big by chance alone?" (Think Stats, Ch. 9)
- Step 1 - Quantify the apparent effect with a single number, the test statistic (e.g. a difference in means, a correlation coefficient).
- Step 2 - Define a null hypothesis: a model of the system in which the effect you saw is NOT real (e.g. "these two days' power readings are really drawn from the same underlying distribution").
- Step 3 - Simulate many datasets under that null hypothesis, and compute the test statistic for each one. This builds the null distribution.
- Step 4 - The p-value is the fraction of the null distribution at least as extreme as your observed statistic. A small p-value means the observed effect would be surprising if the null hypothesis were true.

A null hypothesis you can simulate with one line of code

```
pooled = np.concatenate([day_a, day_b])
n_a = len(day_a)

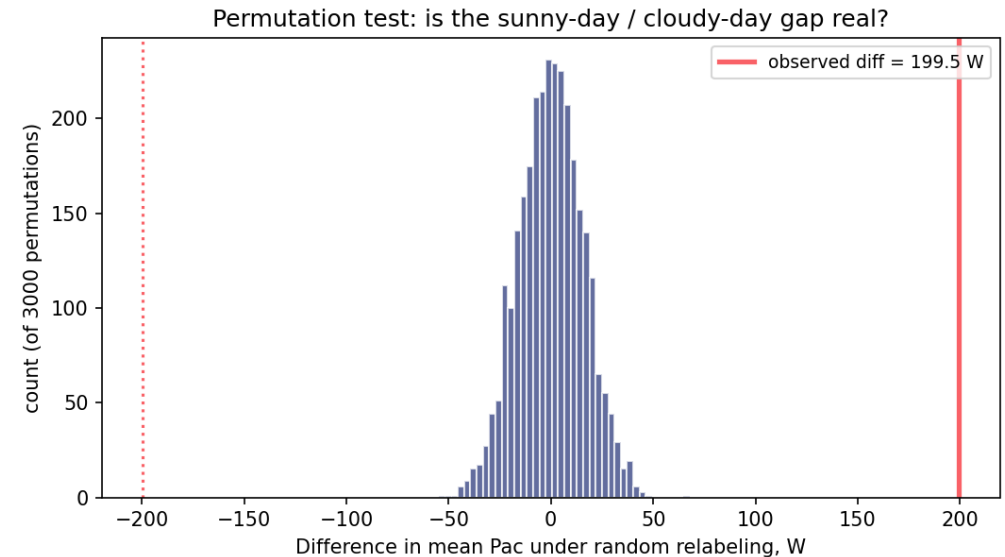
perm_diffs = []
for _ in range(3000):
    np.random.shuffle(pooled)
    diff = pooled[:n_a].mean() - pooled[n_a:].mean()
    perm_diffs.append(diff)
```

- For "are these two groups really different?" questions, the natural null hypothesis is: the group label (which day a reading came from) makes no difference — all the pooled readings come from one and the same distribution.
- If that null hypothesis is true, shuffling the labels at random should produce differences in means that look just like the real data's grouping, on average.
- The p-value is then: (number of shuffles whose $|\text{difference}| \geq \text{the real observed } |\text{difference}|$) / (total number of shuffles).

WORKED EXAMPLE

Is the sunny-day / cloudy-day power gap real?

- 04-01-2013 (sunniest day, by total daily energy): $n = 114$ daylight readings, mean Pac = 199.79 W.
- 18-01-2013 (cloudiest day with a full set of daylight readings): $n = 101$ daylight readings, mean Pac = 0.29 W.
- Observed difference in means = 199.50 W.
- 3,000 permutations of the pooled, relabeled readings: 0 of 3,000 produced a difference as large as 199.50 W in either direction $\rightarrow$ p-value $< 1/3,000$ (approx 0.0003).
- Welch's t-test cross-check gives the same conclusion: $t = 26.5$, $p < 0.000001$ — this gap is not something 215 random five-minute readings would produce by chance.



Null distribution of the permuted difference in means (3,000 shuffles)

What the number does, and does not, tell you

- A two-sided test (used on the previous slide) asks whether the difference is unusually large in EITHER direction; a one-sided test asks only about one direction, and is only appropriate when you committed to that direction before looking at the data.
- A small p-value means the effect is unlikely to be a fluke of random sampling — it does NOT by itself mean the effect is large, important, or caused by what you think caused it (the correlation-vs-causation lesson from Lecture 4 applies here too).
- A large p-value does not prove the null hypothesis is true — it only means this sample did not provide strong evidence against it. With a small sample, even a real effect can produce a large p-value (low statistical power — Part D).

PART D

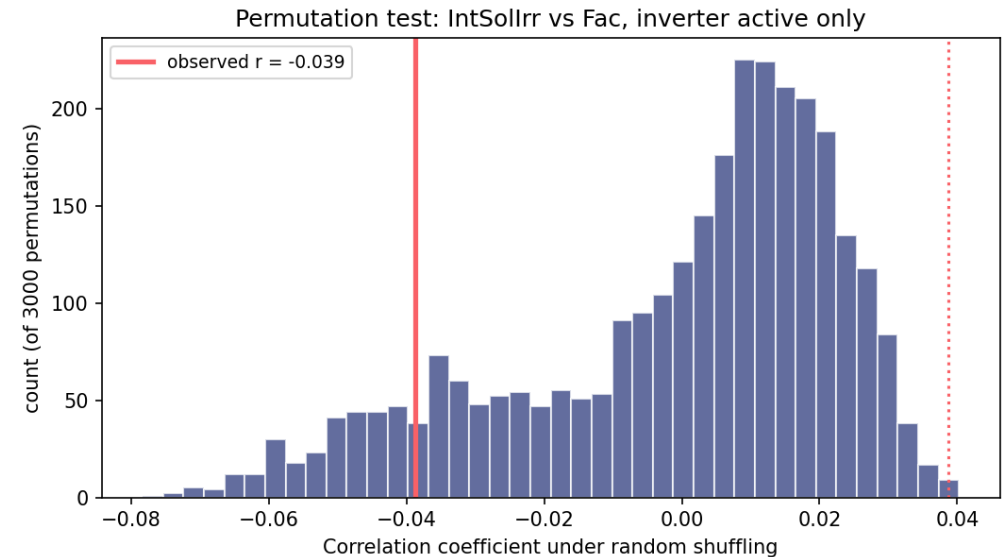
More Tests, Errors & Power

Testing a correlation and a proportion, and the two ways a test can be wrong

TESTING A CORRELATION

Revisiting Lecture 4's spurious IntSolIrr / Fac correlation

- Lecture 4 found that, restricted to moments the inverter is actually on ($\text{Fac} > 0$, $n = 1,805$ readings), the IntSolIrr / Fac correlation collapses from $r = 0.72$ to essentially nothing.
- The permutation test for a correlation shuffles one variable relative to the other, breaking any real association while keeping each variable's own distribution intact.
- Observed $r = -0.0387$ on the active-only readings. 3,000 shuffles give a p-value of 0.0997 — ABOVE the conventional 0.05 threshold.
- Formal conclusion: this residual correlation is not distinguishable from pure chance. It reinforces Lecture 4's conclusion, now backed by a proper significance test rather than eyeballing the size of r .



Null distribution of the permuted correlation (3,000 shuffles)

CHI-SQUARED TEST

Testing whether two days' idle/active proportions really differ

Day	Active (Pac > 0)	Idle (Pac = 0)	Total
04-01-2013 (sunny)	108	178	286
18-01-2013 (cloudy)	5	237	242

- A chi-squared test of independence compares the OBSERVED counts in a table like this to the counts you would EXPECT if the row (day) and column (active/idle) were independent.
- $\chi^2 = \sum (\text{observed} - \text{expected})^2 / \text{expected}$, summed over every cell of the table — large deviations from the "independent" expectation in any cell inflate χ^2 .
- For this table: $\chi^2 = 97.2$, degrees of freedom = 1, p-value ≈ 0.000000 — overwhelming evidence that the fraction of time the inverter is active genuinely differs between these two days.

The two ways a hypothesis test can be wrong

	Null hypothesis is actually TRUE	Null hypothesis is actually FALSE
Test says "significant"	Type I error (false positive)	Correct detection
Test says "not significant"	Correct	Type II error (false negative)

- Type I error rate is controlled by your chosen significance threshold (commonly 0.05) — it is the chance of calling a real-nothing effect "significant."
- Type II error rate depends on the TRUE size of the effect, the sample size, and the noise in the data — a real effect can still be missed if the sample is too small.
- Statistical power is the probability of correctly detecting a real effect of a given size — power increases with sample size, which is exactly why the obvious sunny/cloudy gap (slide 17, $n \sim 215$) was so easy to detect, while the near-zero correlation (slide 20, $n = 1,805$) still left some uncertainty.

One month of one inverter is not the whole story

- Every result in this lecture comes from one inverter's readings across one January. A real engineering conclusion should be checked against other months, other inverters, and other sites before it is trusted.
- Think Stats makes the same point by revisiting its "first babies" pregnancy-length example with a second, independently collected dataset — a result that holds up under replication is far more trustworthy than one that does not.
- Practical habit: whenever a test gives a surprising or highly convenient result, ask what a second, independent sample of days (or a second inverter) would be expected to show.

This is the same "verify your data" discipline from Lab 1 (the dead ExlSolIrr column) and Lecture 4 (the TmpAmb / WindVel placeholder sensors) — now applied to whole conclusions, not just individual columns.

WRAP-UP

From a single number to how much to trust it

- An estimator turns a sample statistic into a guess at a population parameter; bias and MSE let you judge how good that guess is.
- The bootstrap builds a sampling distribution — and a confidence interval — by resampling the data you already have, no formula required.
- Classical hypothesis testing turns "does this effect look real?" into a p-value, computed by simulating a null world with a permutation test.
- The same permutation idea tests a difference in means, a correlation, or (via chi-squared) a table of counts.
- Every test can fail in two directions (Type I / Type II error); power and replication are how you guard against trusting a fluke.

Coming up in Lecture 6: building on these foundations as the course moves toward multiple regression and more advanced modeling of energy data.