Is This Real, or Just Noise?

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 5
February 24 & 26, 2026

Quantifying Uncertainty

Three Statisticians Built the Framework We Still Use

In the 1920s and 30s, three statisticians built the framework we still use today:

- **R.A. Fisher** — developed significance testing, maximum likelihood, and experimental design
- **Jerzy Neyman** and **Egon Pearson** — formalized hypothesis testing with Type I and Type II errors

They disagreed on many things, but shared one core idea:

> We can't know the truth from a single sample.
> But we can design **procedures** whose properties
> we understand across many hypothetical samples.

Frequentist Inference Asks: What If We Repeated This?

Frequentist inference

Drawing conclusions about a population by asking: *what would happen if we repeated the random sampling procedure many times?*

The "many times" is the key idea. Every tool we build this week — standard errors, confidence intervals, hypothesis tests — answers the same question:

How would my results behave under repeated random sampling?

Remember Your Commute Data from PS1?

In PS1 you computed the average commute time for 298 American workers:

Statistic	Value
Sample mean ($\bar{X}$)	31.42 minutes
Standard deviation (s)	21.48 minutes
Sample size (n)	298

The average American commute is sometimes reported as 30 minutes.

New question: Is the average commute *really* different from 30 minutes? Or is 31.42 just noise?

R Can Answer That Question

```r
t.test(commute$commute_time, mu = 30)
```

t	1.14
p-value	0.26
95% confidence interval	[28.98, 33.86]
sample estimate (mean)	31.42

R says: $t = 1.14$, $p = 0.26$, 95% CI $= [28.98,\ 33.86]$

What do these numbers mean? That's this week.

Every Number in That Output Has a Meaning

Estimate = 31.42
best guess

SE = 1.24
precision

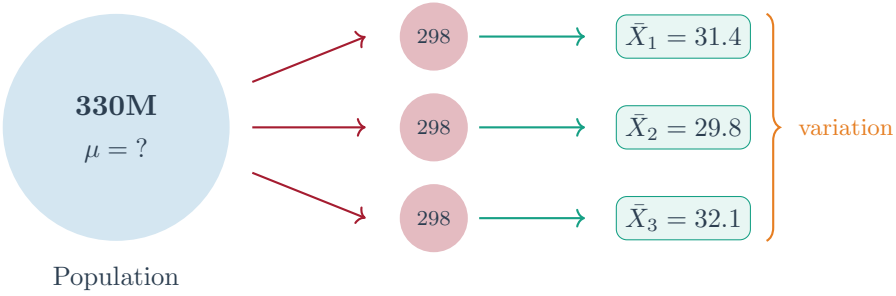
CI = [28.98, 33.86]
plausible range

t = 1.14
surprise

p = 0.26
probability

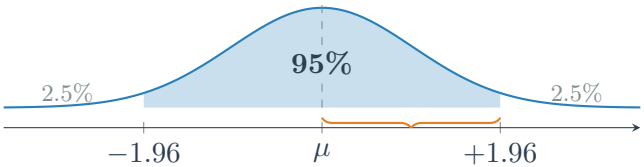
We'll learn each one. By the end of this week, you'll be able to read any R output table and know exactly what every number means.

Each Sample Gives a Different Answer



- The population has a mean μ — but we don't know it
  - We draw a random sample and compute $\bar{X}$
  - Under repeated sampling, each sample gives a different $\bar{X}$
- The **sampling distribution** is the distribution of $\bar{X}$ across all possible samples
- The **standard error** is its standard deviation

95% of Sample Means Fall Within ±1.96 Standard Errors



By the Central Limit Theorem, sample averages are approximately normal

$$95\%\ \text{CI} \quad = \quad \bar{X} \quad \pm \quad 1.96 \times \text{SE}$$

$\bar{X}$ is our best guess; $1.96 \times \text{SE}$ is the **margin of error**

95% Confidence Describes the Procedure, Not the Interval

WRONG: "There is a 95% probability that the true mean is inside this interval."

The true mean μ is a fixed number — it is either inside the interval or it isn't. There is no probability about it.

RIGHT: If we repeated the study many times, 95% of the resulting intervals would contain the true μ. The "95%" describes the *method's long-run performance*, not any single interval.

Think of it as a batting average: the procedure "hits" (contains μ) 95% of the time

Larger Samples Buy Precision

Sample Size (n)	SE	Margin of Error
100	5.0%	$\pm$10%
1,000	1.6%	$\pm$3%
10,000	0.5%	$\pm$1%

Margin of error $= 1.96 \times$ SE — the half-width of a 95% confidence interval

SE shrinks with $\sqrt{n}$: to halve the margin of error, quadruple the sample

Back to the Commute Data

Now we can read the R output from earlier:

Number	Meaning
$\bar{X} = 31.42$	Sample mean commute time
$\text{SE} = 21.48/\sqrt{298} = 1.24$	Precision of the estimate
95% CI $= [28.98,\ 33.86]$	Plausible range for μ

The CI contains 30. We **can't rule out** that the true average commute is 30 minutes.

The $t = 1.14$ and $p = 0.26$ said the same thing — we'll formalize why shortly.

We saw a difference of 1.42 minutes in the commute data.

But how do we formally decide whether a pattern is real or just noise?

The Lady Tasting Tea

R.A. Fisher Invented the Answer at a Tea Party

The scene: Cambridge, 1920s. Muriel Bristol claims she can tell whether **milk** or **tea** was poured first.

Fisher's experiment:

- 8 cups of tea: 4 milk-first, 4 tea-first
- Presented in **random order**
- Bristol told there are exactly 4 of each
- She must identify which 4 are milk-first

Fisher's insight: design the test *before* seeing the result. The randomization creates the probability — no population needed.

If She's Guessing, How Many Ways Could She Choose?

Assume she's guessing (the null). She picks 4 cups at random. Let's count every possible selection.

#	1	2	3	4	5	6	7	8	Match
1	●	●	●	●	○	○	○	○	2/4
2	●	●	●	○	●	○	○	○	2/4
3	●	●	●	○	○	●	○	○	3/4
4	●	●	●	○	○	○	●	○	2/4
5	●	●	●	○	○	○	○	●	3/4
6	●	●	○	●	●	○	○	○	1/4
7	●	●	○	●	○	●	○	○	2/4
8	●	●	○	●	○	○	●	○	1/4
9	●	●	○	●	○	○	○	●	2/4
10	●	●	○	○	●	●	○	○	2/4
11	●	●	○	○	●	○	●	○	1/4
12	●	●	○	○	●	○	○	●	2/4
13	●	●	○	○	○	●	●	○	2/4
14	●	●	○	○	○	●	○	●	3/4
15	●	●	○	○	○	○	●	●	2/4
16	●	○	●	●	●	○	○	○	2/4
17	●	○	●	●	○	●	○	○	3/4
18	●	○	●	●	○	○	●	○	2/4

18 of 70 possible selections — none match so far

If She's Guessing, How Many Ways Could She Choose?

#	1	2	3	4	5	6	7	8	Match
19	●	○	●	●	○	○	○	●	3/4
20	●	○	●	○	●	●	○	○	3/4
21	●	○	●	○	●	○	●	○	2/4
22	●	○	●	○	●	○	○	●	3/4
23	●	○	●	○	○	●	●	○	3/4
24	●	○	●	○	○	●	○	●	4/4
25	●	○	●	○	○	○	●	●	3/4
26	●	○	○	●	●	●	○	○	2/4
27	●	○	○	●	●	○	●	○	1/4
28	●	○	○	●	●	○	○	●	2/4
29	●	○	○	●	○	●	●	○	2/4
30	●	○	○	●	○	●	○	●	3/4
31	●	○	○	●	○	○	●	●	2/4
32	●	○	○	○	●	●	●	○	2/4
33	●	○	○	○	●	●	○	●	3/4
34	●	○	○	○	●	○	●	●	2/4
35	●	○	○	○	○	●	●	●	3/4
36	○	●	●	●	●	○	○	○	1/4

36 of 70 — only 1 match so far (row 24, highlighted)

Fisher's Exact p-Value: 1 in 70

We enumerated all 70 ways to pick 4 cups from 8.

Only **1** selection gets all 4 milk-first cups correct.

$$P(\text{all correct} \mid \text{guessing}) = \frac{1}{70} \approx 0.014$$

This is an **exact p-value** — the probability of a result *as extreme as or more extreme than* what we observed, assuming the null hypothesis is true.

Fisher was computing p-values before anyone called them that. The formal definition comes Thursday.

If the Null Were True, This Would Almost Never Happen

If she were just guessing, there's only a 1.4% chance she'd get all 8 right.

She got all 8 right.

So she probably wasn't just guessing.

This is the **entire logic** of hypothesis testing:

Assume nothing is happening. Ask: how surprising is the data? If very surprising $\rightarrow$ something real is happening.

Formalizing the Framework

Every Hypothesis Test Has a Null and Alternative

Hypotheses

Null hypothesis (H_0): The "nothing is happening" claim

Alternative hypothesis (H_1): Something *is* happening

Examples:

- H_0: Bristol is guessing vs. H_1: Bristol can taste the difference
- H_0: Average commute $= 30$ min vs. H_1: Average commute $\neq 30$ min
- H_0: College and non-college commutes are equal vs. H_1: They differ

The Test Statistic Measures Surprise in Units of SE

We need a single number that measures how far our data is from what H_0 predicts

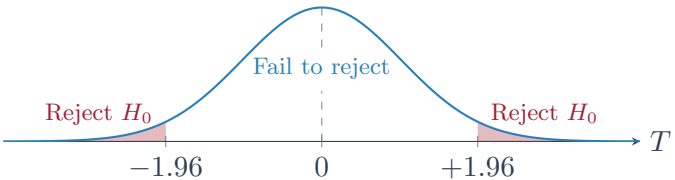
$$T = \frac{\text{estimate} - \text{null value}}{\text{standard error}}$$

- **Numerator:** How far is our estimate from H_0?
- **Denominator:** How much variation do we expect from sampling?
- **Ratio:** How many "standard errors" away from H_0?

In the commute data: $T = (31.42 - 30)/1.24 = 1.14$ — only 1.14 SEs from the null

If $|T| > 1.96$, We Reject the Null at 5%

We set $\alpha = 0.05$: accept at most a 5% chance of a false alarm. Under H_0, T follows a standard normal (mean = 0, SD = 1):



Commute data: $|T| = 1.14 < 1.96 \rightarrow$ **fail to reject**. The 1.42-minute difference is not surprising enough.

CI–Test Duality: You Already Knew How to Test

θ_0 is inside the 95% CI

$$\Longleftrightarrow$$

We *fail to reject* $H_0 \;:\; \theta \;=\; \theta_0$ at $\alpha \;=\; 0.05$

Commute example:

- 95% CI = $[28.98,\ 33.86]$
- Is 30 inside? **Yes** $\rightarrow$ fail to reject $H_0 : \mu = 30$
- Same conclusion as $|T| = 1.14 < 1.96$

Quick shortcut: want to test if an effect is zero? Check if 0 is inside the CI.

Tuesday summary

1. Hypothesis testing asks: "Is this pattern real or just noise?"
2. The logic: assume nothing, check how surprising the data is
3. The test statistic T measures surprise in units of SE
4. Reject H_0 when $|T| > 1.96$ (equivalently, when the null value is outside the CI)

Thursday: t-statistics in R, p-values, and common mistakes

Errors and Tradeoffs

Two Types of Mistakes

	H_0 **is TRUE**	H_0 **is FALSE**
We reject H_0	**Type I Error** False alarm — "crying wolf"	**Correct!** We detected a real effect
We fail to reject	**Correct!** No false alarm	**Type II Error** Missed detection — "missing the wolf"

Alpha Is a Choice, Not a Law of Nature

$\alpha = 0.05$ means: we accept at most a 5% false positive rate

This is a **choice**, not a law of nature

- Social science: $\alpha = 0.05$ (by convention)
- Particle physics: $\alpha = 0.0000003$ ("five sigma")
- Some exploratory work: $\alpha = 0.10$

Lower α = fewer false alarms, but harder to detect real effects

We Can't Minimize Both Errors at Once

- Making our test **stricter** (lower α):
  - Fewer false alarms (less Type I error)
  - But more missed detections (more Type II error)
- Making our test **looser** (higher α):
  - Fewer missed detections (less Type II error)
  - But more false alarms (more Type I error)

The only way to reduce *both* errors: collect more data (larger n)

In Science, Type I Errors Are Especially Dangerous

A Type I error means **publishing a false finding**

- Other researchers build on it
- Policies get designed around it
- Textbooks teach it to students
- It takes *years* to undo the damage

A Type II error (missing a real effect) is disappointing—but it doesn't mislead the field

This asymmetry is why science sets α low and tolerates missing some real effects

The Replication Crisis Is Accumulated Type I Errors

In 2015, the Open Science Collaboration tried to replicate 100 published psychology studies

Result: Only 36% replicated successfully

Why so many false positives?

- **Publication bias**: journals publish significant results, not null results
- **Small samples**: small n means noisy estimates and inflated effects
- **p-hacking**: running many tests and reporting the one that "works"
- Type I errors **accumulate** across thousands of published studies

The t-Statistic

The t-Statistic Formula

For testing whether a population mean equals μ_0:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

- $\bar{X}$ = sample mean
- μ_0 = the value under H_0
- s = sample standard deviation
- $s/\sqrt{n}$ = the standard error

Under H_0, $T \approx N(0,1)$ by the CLT. Reject H_0 when $|T| > 1.96$

One-Sample Test: Is the Average Commute Different from 30?

Data from PS1: $n = 298$, $\bar{X} = 31.42$, $s = 21.48$. H_0: $\mu = 30$ vs. H_1: $\mu \neq 30$

Step 1: $\text{SE} = \dfrac{s}{\sqrt{n}} = \dfrac{21.48}{\sqrt{298}} = 1.24$

Step 2: $T = \dfrac{31.42 - 30}{1.24} = \dfrac{1.42}{1.24} = 1.14$

Step 3: Compare to 1.96

$|T| = 1.14 < 1.96 \rightarrow$ **Fail to reject** H_0

The sample mean is 1.42 minutes above 30, but that's only 1.14 standard errors away — not surprising enough

R Computes the t-Test Automatically

```r
t.test(commute$commute_time, mu = 30)
```

```
One Sample t-test
t = 1.14          df = 297
p-value = 0.26
95% CI:           [28.98, 33.86]
mean of x:        31.42
```

Same numbers we computed by hand: $t = 1.14$, $p = 0.26$, CI contains 30

R uses the t-distribution ($\text{df} = n - 1 = 297$), not the normal — but for large n, they give the same answer

Two-Sample Test: Do College Graduates Commute Longer?

	College+	No College
n	176	122
$\bar{X}$	33.47 min	28.47 min

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}_{\text{diff}}} = \frac{33.47 - 28.47}{2.92} = \frac{5.00}{2.92} = 1.71$$

$|T| = 1.71 < 1.96 \rightarrow$ **Fail to reject**

5 minutes sounds like a lot, but with this sample size it's only 1.71 SEs — not enough to rule out chance

(R computes SE_{diff} from the group standard deviations: `t.test(commute_time ~ college, data = commute)`)

The Same Test in R

```
t.test(commute_time ~ college, data = commute)
```

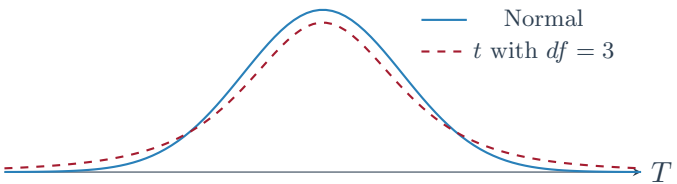
```
Welch Two Sample t-test
t = 1.71          df = 157.4
p-value = 0.089
95% CI:           [-0.73, 10.73]
mean college:     33.47
mean no college:  28.47
```

The CI $[-0.73,\ 10.73]$ contains $0 \rightarrow$ fail to reject. Same answer. $p = 0.089$ — close to 0.05 but not quite.

Small Samples Need Fatter Tails

For small samples, the normal approximation is too generous

William Gosset ("Student") at the Guinness brewery showed: when you estimate the SE from small samples, you need **fatter tails**



As n grows, $t \to$ normal. R uses the t-distribution automatically — you don't need to choose

The p-Value

The p-Value Is the Probability of Data This Extreme Under H_0

p-value

The probability of observing data **as extreme as or more extreme than** what we got, **if the null hypothesis were true**.

- Small p = data is very surprising under H_0 = strong evidence against H_0
- Large p = data is not surprising under H_0 = weak evidence against H_0

$$p \quad < \quad 0.05 \Rightarrow \text{reject } H_0 \text{ at } \alpha \quad = \quad 0.05$$

Commute data: $p = 0.26$ (one-sample) and $p = 0.089$ (two-sample) — neither crosses 0.05

The p-Value Gives Continuous Evidence

Instead of just reject/don't reject, the p-value tells you *how much* evidence you have

p-value	Rough interpretation
> 0.10	Little evidence against H_0
0.05–0.10	Weak evidence
0.01–0.05	Moderate evidence
0.001–0.01	Strong evidence
< 0.001	Very strong evidence

But remember: 0.05 is just a convention, not a bright line between truth and falsehood

Two-Sided vs. One-Sided Tests

Two-sided (default):

- $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
- Reject if $|T| > 1.96$ (evidence in either tail)

One-sided (use with caution):

- $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$
- Reject if $T > 1.645$ (one tail only)
- Requires a *pre-specified* directional hypothesis

Default to two-sided unless you have a
strong, pre-registered reason for a direction

What These Numbers Do NOT Mean

WRONG: "$p = 0.03$ means there's a 3% chance H_0 is true"

WRONG: "We are 95% confident the true value is in the CI"

RIGHT: $p = 0.03$ means if H_0 were true, there's a 3% chance of data this extreme. It says nothing about the probability H_0 is true.

RIGHT: 95% of CIs constructed this way would contain μ. Any single CI either does or doesn't.

Both are about the *procedure's long-run performance*, not about any single result

Statistical Significance $\neq$ Scientific Significance

A **statistically significant** result can be tiny and irrelevant

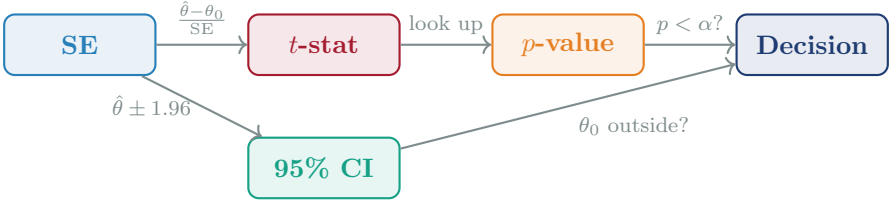
Example: A study of 1 million voters finds that a campaign ad increases turnout by 0.02 percentage points ($p < 0.001$)

- Statistically significant? Yes—p is tiny
- Scientifically significant? No—0.02 points is meaningless

With a large enough sample, **everything** is statistically significant. What matters is the **size** of the effect, not just whether it differs from zero.

The Big Picture

Everything Connects



These are all the **same framework** viewed from different angles:

- **SE** measures how much estimates vary
- t**-stat** counts SEs from the null; p**-value** converts to probability
- **CI** gives a range of plausible values
- All lead to the same reject/fail-to-reject decision

R Gives You Everything in One Table

```
summary(lm(commute_time ~ college, data = commute))
```

	Estimate	Std. Error	t value	p value
(Intercept)	28.47	1.95	14.60	< 0.001
college	5.00	2.92	1.71	0.089

Everything from this week — SE, t-statistic, p-value — appears in one table

Week 6: what those numbers mean and how to interpret them

Every number you compute
from data has uncertainty.
Hypothesis testing
is the discipline of
taking that uncertainty seriously.

Questions?

Reading: QSS 7.1–7.2