

Hypothesis Testing

When is a sample result too unusual to dismiss as sampling variability?

Sergio I. García-Ríos

PA 311N: Data Analysis for Public Policy

LBJ School of Public Affairs

October 19, 2026

Where we left off

Confidence intervals asked:

What values of the population parameter are **plausible** given our sample?

Hypothesis tests turn the logic around:

Suppose a particular population value were true.
Would our sample result be **surprising**?

Same sampling-distribution logic. Different inferential question.

A policy claim we can test

The hypothetical Ames Sustainability Office plans staffing around this claim:

Average residential inspection time is 60 minutes.

A random sample of 40 inspections gives

$$n = 40, \quad \bar{x} = 66, \quad s = 18 \text{ minutes.}$$

Question

Is a six-minute difference something we could reasonably see from sampling variability alone, or is it evidence that inspections take longer than 60 minutes on average?

Two explanations for the same sample

We observed

$$\bar{x} = 66 \text{ minutes.}$$

Explanation 1

The true mean really is 60 minutes.

Our sample happened to run high because of **sampling variability**.

Explanation 2

The true mean is greater than 60 minutes.

The sample is revealing a **real difference** from the planning assumption.

Hypothesis testing asks how compatible the data are with Explanation 1.

Start by pretending the null claim is true.

Null and alternative hypotheses

The **null hypothesis** is the benchmark claim we test:

$$H_0 : \mu = 60.$$

The **alternative hypothesis** is the pattern for which we are seeking evidence:

$$H_A : \mu > 60.$$

This is a **one-sided** test because the policy concern is specifically that inspections take *longer* than planned.

Important

We do not begin by assuming H_A is true. We ask whether the observed data would be unusually extreme *if H_0 were true*.

The alternative determines the tail

Policy question	Alternative	Relevant tail(s)
Longer than planned?	$H_A : \mu > 60$	upper tail
Shorter than planned?	$H_A : \mu < 60$	lower tail
Different from planned?	$H_A : \mu \neq 60$	both tails

The alternative determines the tail

Policy question	Alternative	Relevant tail(s)
Longer than planned?	$H_A : \mu > 60$	upper tail
Shorter than planned?	$H_A : \mu < 60$	lower tail
Different from planned?	$H_A : \mu \neq 60$	both tails

Your Turn

If the city only wants to know whether the average differs from 60 minutes in *either direction*, what should H_A be?

The logic in four steps

1. **State** H_0 and H_A .
2. **Model** what sample statistics should look like if H_0 is true.
3. **Measure** how far the observed statistic is from the null value.
4. **Evaluate** how unusual that result would be under H_0 .

That final probability is the **p-value**.

Hypothesis testing is a structured version of:
“**Could random sampling reasonably have produced this?**”

Conditions still matter

For a one-sample t test of a mean, we need the same logic that supported our confidence interval:

- ▶ **Independence:** observations should come from a random process and not determine one another.
- ▶ **10% condition:** if sampling without replacement, the sample should be less than about 10% of the population.
- ▶ **Shape / sample size:** the population should be roughly normal, or the sample should be large enough that the sampling distribution of $\bar{x}$ is approximately normal.

Conditions justify the reference distribution we use to calculate the p-value.

How unusual is 66 if the true mean is 60?

Under the null hypothesis,

$$\mu_0 = 60.$$

Our observed estimate is

$$\bar{x} = 66.$$

But a six-minute difference means something different when the standard error is 1 minute versus 20 minutes.

We need a standardized distance

How many **standard errors** is the observed sample mean from the null value?

The test statistic

For a one-sample test of a mean,

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

- ▶ Numerator: observed difference from the null claim.
- ▶ Denominator: estimated standard error of the sample mean.
- ▶ Result: distance from the null value measured in **standard errors**.

Large $|t|$ means the sample result is far from what H_0 predicts.

Do not confuse t^* and t

t^* : critical value

Used in a confidence interval:

$$\bar{x} \pm t^* SE.$$

Chosen from the desired confidence level and degrees of freedom.

t : test statistic

Calculated from the observed data:

$$t = \frac{\bar{x} - \mu_0}{SE}.$$

Measures how far the estimate is from the null claim.

The star is not decoration: t^* and t play different roles.

Calculate the standard error

For the inspection-time sample,

$$n = 40, \quad s = 18.$$

So

$$SE = \frac{s}{\sqrt{n}} = \frac{18}{\sqrt{40}} \approx 2.85 \text{ minutes.}$$

A sample mean from repeated samples of 40 inspections would typically vary by roughly 2.85 minutes around the true mean.

Calculate the observed t statistic

Now compare the observed mean with the null value:

$$t = \frac{66 - 60}{2.85} \approx 2.11.$$

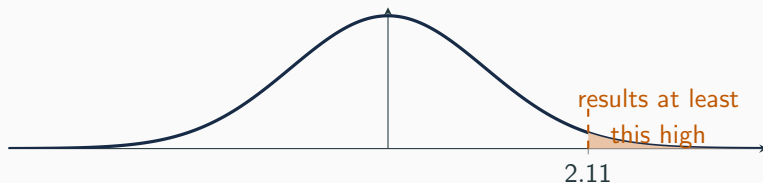
The sample mean is about **2.11 standard errors above** the null value.

Is 2.11 standard errors unusually far into the upper tail? That is exactly what the p-value answers.

What would we see if H_0 were true?

If $H_0 : \mu = 60$ is true, the test statistic follows a t distribution with

$$df = n - 1 = 39.$$



For our one-sided alternative, the p-value is the **upper-tail area beyond $t = 2.11$** .

The p-value

Definition

The p-value is the probability of observing a result at least as extreme as ours, in the direction specified by H_A , **assuming the null hypothesis is true**.

For this example:

$$p = P(T_{39} \geq 2.11) \approx 0.021.$$

Interpretation:

If the true average inspection time were really 60 minutes, only about **2.1%** of random samples like ours would produce a sample mean this high or higher.

Calculate the p-value in R

```
n <- 40
x_bar <- 66
mu_0 <- 60
s <- 18

se <- s / sqrt(n)
t_obs <- (x_bar - mu_0) / se

pt(t_obs, df = n - 1, lower.tail = FALSE)
```

$$p \approx 0.0207.$$

'lower.tail = FALSE' asks R for the upper-tail probability.

From evidence to a decision rule

A p-value is continuous evidence. For a formal decision, we choose a significance level α .

Commonly,

$$\alpha = 0.05.$$

Decision rule:

- ▶ If $p < \alpha$: reject H_0 .
- ▶ If $p \geq \alpha$: fail to reject H_0 .

For our example:

$$0.0207 < 0.05.$$

So we reject H_0 at the 5% significance level.

Interpret the decision in policy language

Do not stop at “reject H_0 .”

Weak conclusion

$p < 0.05$, therefore reject the null.

Policy conclusion

The sample provides statistically significant evidence that the average residential inspection time exceeds 60 minutes.

This does **not** prove the true mean is exactly 66 minutes. It says the 60-minute planning assumption is difficult to reconcile with the observed sample.

Fail to reject is not accept

Suppose instead we had found

$$p = 0.32.$$

The correct conclusion would be:

The sample does **not provide sufficient evidence** that $\mu > 60$.

Not:

“We proved that $\mu = 60$.”

A non-small p-value can reflect a true null, a small effect, noisy data, or insufficient sample size.

What a p-value is not

A p-value is **not**:

- ▶ the probability that H_0 is true;
- ▶ the probability that the result happened “by chance”;
- ▶ the size or practical importance of an effect;
- ▶ the probability that our conclusion is wrong.

It is a conditional probability:

$$P(\text{data as or more extreme than observed} \mid H_0 \text{ true}).$$

Your turn: interpret the p-value

Your turn: interpret the p-value

Your Turn

For the Ames inspection example, $p \approx 0.021$. Which interpretation is best?

1. There is a 2.1% chance that the true mean is 60.
2. There is a 2.1% chance that our sample is wrong.
3. If the true mean were 60, a sample result this high or higher would occur about 2.1% of the time.
4. About 2.1% of inspections take more than 60 minutes.

Your turn: interpret the p-value

Your Turn

For the Ames inspection example, $p \approx 0.021$. Which interpretation is best?

1. There is a 2.1% chance that the true mean is 60.
2. There is a 2.1% chance that our sample is wrong.
3. If the true mean were 60, a sample result this high or higher would occur about 2.1% of the time.
4. About 2.1% of inspections take more than 60 minutes.

Best answer: (3). The probability is calculated assuming H_0 is true.

One-sided versus two-sided p-values

Our original question was directional:

$$H_A : \mu > 60.$$

so we used only the upper tail:

$$p_{\text{one-sided}} \approx 0.021.$$

If instead

$$H_A : \mu \neq 60,$$

then equally extreme deviations in *either direction* count:

$$p_{\text{two-sided}} \approx 2(0.021) = 0.041.$$

Choose the alternative from the substantive question, **not after seeing the data.**

Confidence intervals and two-sided tests agree

Using the same sample,

$$\bar{x} = 66, \quad SE \approx 2.85, \quad df = 39.$$

The 95% confidence interval is approximately

$$66 \pm 2.023(2.85) = (60.24, 71.76).$$

The null value 60 is **outside** the 95% interval.

That matches the two-sided test:

$$p \approx 0.041 < 0.05.$$

Connection

A two-sided test at $\alpha = 0.05$ rejects exactly those null values excluded by the corresponding 95% confidence interval.

What does $\alpha = 0.05$ mean?

The significance level controls how willing we are to reject H_0 when it is actually true.

$\alpha = 0.05$ sets a 5% Type I error rate under repeated testing.

It is a decision threshold chosen *before* looking at the result.

Smaller α requires stronger evidence to reject H_0 .

Two ways a decision can be wrong

		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0 true	Correct	Type I error
	H_A true	Type II error	Correct

- ▶ Type I: false positive.
- ▶ Type II: false negative.

There is no decision rule that eliminates both errors in every sample.

Type I error in the inspection example

Recall:

$$H_0 : \mu = 60, \quad H_A : \mu > 60.$$

A **Type I error** would mean:

Concluding that inspections take longer than 60 minutes on average
when the true mean is actually 60.

Possible policy consequence: the city expands staffing or budgets additional time when that change was not actually necessary.

Type II error in the inspection example

A **Type II error** would mean:

Failing to detect that inspections really do take longer than 60 minutes on average.

Possible policy consequence: the city keeps an unrealistic staffing model, causing delays, overtime, or missed appointments.

Type II error in the inspection example

A **Type II error** would mean:

Failing to detect that inspections really do take longer than 60 minutes on average.

Possible policy consequence: the city keeps an unrealistic staffing model, causing delays, overtime, or missed appointments.

Your Turn

Which error would be more costly in this policy setting? What information would you need to decide?

Power: the ability to detect a real effect

Power is the probability that a test correctly rejects H_0 when a meaningful alternative is true.

Power generally increases when:

- ▶ the true effect is larger;
- ▶ the sample size is larger;
- ▶ the data are less variable;
- ▶ we use a less stringent α .

A study can fail to find statistical significance simply because it had too little information to detect the effect reliably.

Statistical significance is not policy significance

Suppose a huge study estimates that a new scheduling system changes average wait time by

0.4 minutes

with

$$p < 0.001.$$

The result is statistically distinguishable from zero.

But is a 24-second improvement worth implementation costs?

Two separate questions

Is the effect statistically detectable?

Is the effect substantively important for the policy decision?

A hypothesis test is not a verdict machine

A good policy analysis reports more than “significant” or “not significant.”

Report:

- ▶ the estimate and its units;
- ▶ uncertainty, preferably a confidence interval;
- ▶ the test statistic and p-value when relevant;
- ▶ the size and practical meaning of the effect;
- ▶ assumptions and design limitations.

Evidence informs a decision. It does not make the decision for us.

The full workflow

1. **Question:** what policy claim are we evaluating?
2. **Hypotheses:** state H_0 and H_A .
3. **Conditions:** is the inferential model defensible?
4. **Estimate uncertainty:** $SE = s/\sqrt{n}$.
5. **Standardize:** $t = (\bar{x} - \mu_0)/SE$.
6. **Tail probability:** calculate the p-value.
7. **Decision:** compare p with α if a formal rule is needed.
8. **Interpret:** translate the evidence into policy language.

Your turn: set up the test

A city claims the average permit-processing time is 30 days. A random sample suggests the process may be slower.

Your turn: set up the test

A city claims the average permit-processing time is 30 days. A random sample suggests the process may be slower.

Your Turn

Write the hypotheses for a test of whether the true mean processing time exceeds 30 days.

Your turn: set up the test

A city claims the average permit-processing time is 30 days. A random sample suggests the process may be slower.

Your Turn

Write the hypotheses for a test of whether the true mean processing time exceeds 30 days.

Our Turn

$$H_0 : \mu = 30 \quad \text{vs.} \quad H_A : \mu > 30.$$

The relevant p-value will be an **upper-tail** probability.

What to remember

- ▶ Hypothesis tests ask whether the observed data are compatible with a specific null claim.
- ▶ The test statistic measures distance from the null value in **standard errors**.
- ▶ The p-value asks how unusual the observed result would be **if H_0 were true**.
- ▶ Reject H_0 when evidence is sufficiently inconsistent with it; otherwise **fail to reject**.
- ▶ Statistical significance is not the same as practical importance.

**The p-value is not the probability
that the null is true.**

It is a measure of how surprising the data would be if the null were true.