

Beginning Hypothesis Testing

October 28–30, 2026 EN.553.211 Zan Ahmad

Chapter 7 used samples to estimate unknown population parameters. Chapter 8 asks whether a particular claim about a parameter is compatible with the observed data. These notes develop the reference model and interpretation of a test, then connect them to confidence intervals through Friday practice. Read the explanations on pp. 1–3; work through the questions on pp. 4–8. The optional continuation on p. 9 introduces critical values and is not required for the Friday questions.

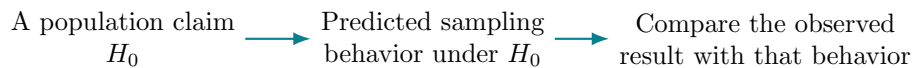
Learning goals

By the end of this introduction, you should be able to:

- Translate a research question into H_0 and H_a before calculating.
- Explain how the null hypothesis supplies a reference distribution, and distinguish a random statistic from its observed value.
- Compute and interpret a large-sample mean-test p -value, choose the correct tail, and state a conclusion at a specified significance level.

1. From an interval to a question about a claim

Recall Chapter 7 Problem 3. When p was unknown, we estimated it and constructed a confidence interval. In the separate model with $p = 0.75$, we could calculate how often a sample proportion would be at most 0.70. Hypothesis testing uses this second kind of reasoning: *temporarily assume a claim, then ask how unusual the observed result would be under that claim.*



The claim supplies the reference model. The data supply the observed result. A small tail probability indicates tension between them, subject to the sampling assumptions.

2. State the hypotheses

The **null hypothesis** H_0 is the claim used to calibrate the test. For a population mean we write $H_0 : \mu = \mu_0$, where μ_0 is the specified **null value**. The **alternative hypothesis** H_a expresses the research question:

Question	Alternative	Test direction
Has the mean increased?	$H_a : \mu > \mu_0$	Right-sided
Has the mean decreased?	$H_a : \mu < \mu_0$	Left-sided
Has the mean changed?	$H_a : \mu \neq \mu_0$	Two-sided

Choose the direction from the scientific question, before seeing the result. The slides compare this to a court's presumption of innocence: insufficient evidence to reject a claim does not prove the claim true. The analogy describes the burden of evidence, not a numerical legal standard.

Check for understanding. A treatment is intended to *reduce* mean recovery time from 12 days. State H_0 and H_a . Would a sample mean above 12 favor this alternative?

The Null Distribution

3. Measure a discrepancy in standard errors

Let $X_1, \dots, X_n$ be independent, identically distributed observations with mean μ and finite variance $0 < \sigma^2 < \infty$. Recall

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

For a normal population, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ exactly. As in our Chapter 7 notes, the second argument of N denotes *standard deviation*. Standardizing gives

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Now impose the null claim. If $H_0 : \mu = \mu_0$ is true and σ is known, the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{under } H_0.$$

For a nonnormal population, the normal statement is an approximation justified by an adequate CLT approximation. When σ is unknown, use

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \stackrel{\text{approx}}{\sim} N(0, 1) \quad \text{under } H_0.$$

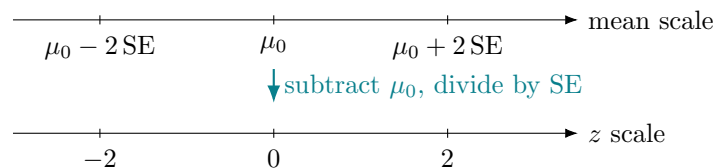
The course uses $n \geq 30$ as a large-sample guideline. It is not a guarantee for every population shape; substantial skewness or heavy tails may need larger samples. A large sample does not repair dependence or biased sampling. Small-sample methods come later.

4. Keep the random quantity and the observation separate

Before sampling, $\bar{X}$ and S vary. After sampling, $\bar{x}$ and s are fixed numbers. With unknown σ , the **observed test statistic** is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{\text{observed estimate} - \text{null value}}{\text{estimated standard error}}.$$

The numerator is a discrepancy in the original units. Dividing by the standard error expresses that discrepancy in units of sampling variability. Thus $z = 2$ means the estimate lies two estimated standard errors above the null value.



Why use a tail? In a continuous normal model, $\Pr(\bar{X} = \mu_0) = 0$, even when H_0 is true. Exact equality is therefore the wrong criterion. We ask for results *at least as extreme*, in the direction specified by H_a .

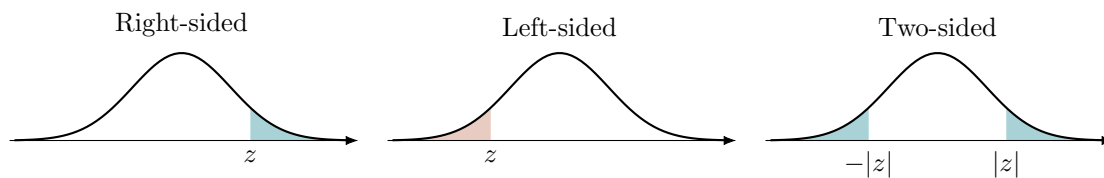
Check for understanding. Suppose the discrepancy points toward H_a . If $\bar{x} - \mu_0$ stays fixed but the standard error doubles, what happens to z ? Explain why the same raw difference now provides weaker evidence.

What Counts as More Extreme

5. The alternative determines the shaded area

The ***p*-value** is the probability, *assuming* H_0 and the *sampling model*, of a test statistic at least as favorable to H_a as the observed statistic. For a standard-normal reference and observed value z :

Alternative	Outcomes at least as extreme	<i>p</i> -value
$\mu > \mu_0$	$Z \geq z$	$1 - \Phi(z)$
$\mu < \mu_0$	$Z \leq z$	$\Phi(z)$
$\mu \neq \mu_0$	$Z \leq - z $ or $Z \geq z $	$2[1 - \Phi(z)]$



The shaded tails illustrate the three rules; these are schematic cutoffs, not a new data set.

Why the factor of two? For a two-sided question, equally large discrepancies in either direction count against H_0 . Symmetry of the standard normal gives

$$\Pr(|Z| \geq |z|) = \Pr(Z \leq -|z|) + \Pr(Z \geq |z|) = 2\Pr(Z \geq |z|).$$

A one-sided *p*-value is not always half a two-sided one: the observed discrepancy must point toward the specified alternative. For example, a large positive z yields a large *left-sided* *p*-value.

6. Two decisions, carefully worded

$$\begin{aligned} p\text{-value} \leq \alpha &\implies \text{reject } H_0; \\ p\text{-value} > \alpha &\implies \text{fail to reject } H_0. \end{aligned}$$

Failure to reject means **insufficient evidence for the specified alternative at this level**. It does not establish equality. Statistical significance also does not measure the scientific importance of a difference; report the observed effect and its context.

Explain before calculating

For each practice question, identify the population parameter, specify H_0 and H_a , state the reference model and its assumptions, and distinguish the statistic from its observed value. Then connect the probability statement to the research question and conclude in context.

Check for understanding. A test yields a *p*-value of 0.08 with $\alpha = 0.05$. State the decision and explain why it does not prove H_0 .

Friday Practice Chapter 7 Review

Fall-break problem 3. An unknown rate and a specified probability

A monitoring program calls a returned water sample *usable* if it meets a prespecified handling checklist. In a random sample of 240 returned kits, 168 samples are usable. Let p be the population proportion of usable samples among this program's returned kits. Assume independent Bernoulli outcomes with a common probability p , and a population much larger than the sample. These data are illustrative.

Use $z_{0.025} = 1.96$ and let Φ denote the standard normal CDF.

- (a) Estimate p and check the observed success and failure counts. Construct an approximate 95% confidence interval for p , showing the estimated standard error and margin of error. Interpret the endpoints as percentages and identify the population they describe.

- (b) Now consider a *separate hypothetical model* in which the population proportion is known to be $p = 0.75$. For a new independent sample of $n = 240$, find the approximate probability $\Pr(\hat{p} \leq 0.70)$. Check the expected counts, standardize, and use a direct normal approximation without continuity correction.

- (c) Explain why the standard errors in parts (a) and (b) use different proportions. Which question estimates a fixed unknown parameter, and which assigns a probability to a random sample result?

- (d) *Think ahead using the sample-size planning section.* A new program wants a planned 95% normal-interval margin of error of at most 4 percentage points. No reliable advance estimate of its population proportion is available. Find the smallest integer sample size that meets this target under the worst-case choice of p . Justify that choice and state whether your answer is a required *total* or a number to add to 240.

Problem 1 The GPA Example

Hypotheses and the statistic

Use the example from the recorded Chapter 8 lecture. The historical reference mean GPA is 3.37. An independent random sample of 50 current students has $\bar{x} = 3.47$ and $s = .31$. Before collecting the sample, the research question was whether the population mean GPA had increased. The chosen significance level is $\alpha = .05$.

Use the lecture's large-sample normal approximation, assuming its sampling conditions are adequate. The population standard deviation is unknown. Write Φ for the standard normal CDF.

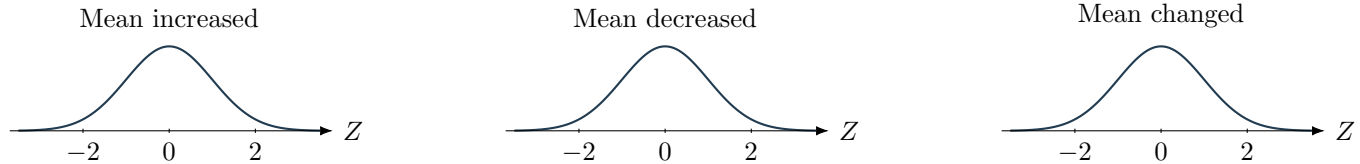
- (a) **Hypotheses and parameters.** Define the population parameter μ . Explain the roles of the *null hypothesis* and *alternative hypothesis*, and state both for this question. Identify the *null value* μ_0 . Is a hypothesis a statement about μ or about the observed $\bar{x}$?

- (b) **Statistic and reference distribution.** Write the test statistic in terms of the random quantities $\bar{X}$ and S , then calculate its observed value z using $\bar{x}$ and s . Explain what the numerator and denominator measure. Define the *null distribution*; state the distribution used here and the assumption under which it applies. Which is random before sampling: the statistic or the null value?

Problem 1 Continued

Data reference: $n = 50$, $\bar{x} = 3.47$, $s = .31$; historical mean 3.37.

- (c) **The p -value and its tail.** Define a p -value in words, including both what is assumed and what counts as “at least as favorable.” For each separately specified research question below, state H_a , write a probability involving Z and the observed z , shade its region, and calculate the p -value. The data stay the same; these are three alternative questions, each chosen *before* seeing data.



Mark the observed boundary or boundaries and shade the relevant region. Use a normal table or calculator.

- (d) **Significance and conclusion.** Define the significance level as used so far in this lecture. For the original *increase* question, state the decision at .05 and the conclusion about the population. Reassess the same evidence using a prechosen .01 level. Which of the data, z , p -value, threshold, and decision change?

- (e) **Interpretation check.** Explain what is wrong with each statement and replace it with a correct one:
 (i) “The p -value is the probability that our null hypothesis is wrong.” (ii) “If we fail to reject, we have proved that the population mean is 3.37.”

Problem 2 The Chapter 7 Connection

One sample and two research questions

Return to Chapter 7 Problem 3. In a random sample of 240 returned water-sample kits, 168 are usable. Let p be the population proportion of usable samples *among this program's returned kits*. Assume independent Bernoulli outcomes with a common p and a population much larger than the sample. These data are illustrative.

New question. Before sampling, the program asks whether the usable-sample proportion *differs from .75*. The decline comparison below is a second, separately preplanned question. Use $\alpha = .05$.

- (a) **Retrieve the confidence interval.** Use your Chapter 7 work, or recompute the estimate $\hat{p}$, estimated standard error, and approximate 95% confidence interval. Use $z_{.025} = 1.96$. Does the interval include .75? State what population quantity the interval estimates.

- (b) **Build a null reference model.** State H_0 and H_a for the new question. Check the expected success and failure counts under H_0 . The model from Chapter 7 Problem 3(b) supplies

$$Z = \frac{\hat{p} - .75}{\sqrt{.75(.25)/240}} \stackrel{\text{approx}}{\sim} N(0, 1) \quad \text{when } p = .75.$$

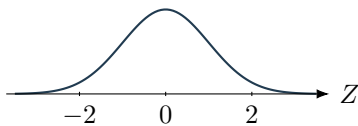
Use this supplied model to calculate the observed z . Explain why the interval in (a) uses an estimated standard error but this reference calculation uses .75. In each calculation, identify what is fixed and what varies across repeated samples.

Problem 2 Continued

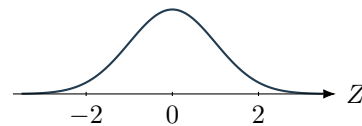
Data reference: 168 usable samples among 240 returned kits; reference proportion .75; $\alpha = .05$. Reuse your calculations from the preceding page.

- (c) **Turn the probability into evidence.** For the *differs* question, mark the observed boundaries and shade the appropriate region on the first curve. Calculate the p -value using the direct normal approximation, without continuity correction. Give a decision and a contextual conclusion. Does this decision agree with the fact you observed about the 95% interval?

Proportion differs from .75



Proportion is below .75



- (d) **Compare a planned decline question.** A second analyst had specified $H_a : p < .75$ *before* seeing the same data. Shade that analyst's region on the second curve. Use your earlier calculation to obtain the p -value and decision at .05. Explain how the two analysts can reach different decisions without changing the data or the null distribution.

- (e) **Reconcile the interval and the tests.** A classmate says, "A two-sided 95% confidence interval decides every 5% test of its null value." Use your results to assess the claim. Identify both the difference in tail questions and the difference between the two standard-error formulas. Should an analyst choose the direction of H_a after seeing which tail gives the smaller p -value? Explain.

Optional Extension: Critical Values

Read after the core lesson. This page develops the rejection-region material in slides 21–23. It is not needed to complete the core example by its p -value.

7. A decision rule has an error rate

A **Type I error** is rejecting H_0 when H_0 is true. For an exactly calibrated continuous test of size α ,

$$\Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha.$$

For our large-sample tests, the calibration is approximate. A significance level of .05 describes a long-run false-rejection rate under the null model; it is not the probability that this particular conclusion is wrong.

8. Convert a tail threshold to a statistic threshold

The course uses **upper-tail** critical values: $\Pr(Z \geq z_\alpha) = \alpha$. For a right-sided test, monotonicity of Φ gives

$$1 - \Phi(z) \leq \alpha \iff \Phi(z) \geq 1 - \alpha \iff z \geq z_\alpha.$$

The **rejection region** R_0 is the set of statistic values for which we reject H_0 :

Alternative	Rejection region R_0
$\mu > \mu_0$	$[z_\alpha, \infty)$
$\mu < \mu_0$	$(-\infty, -z_\alpha]$
$\mu \neq \mu_0$	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$

For a two-sided test, split the total null rejection probability α equally between the two tails. This is the same critical-value convention used for Chapter 7 confidence intervals.

9. Connect the two descriptions

The rejection threshold is fixed before sampling. The observed statistic and its p -value depend on the sample. The critical-value method compares a statistic with a fixed boundary; the p -value method compares the corresponding tail probability with α .

Check for understanding. Return to the GPA example in Problem 1. For $\alpha = .05$, use $z_{.05} \approx 1.645$ to specify the rejection region. Compare your observed statistic with it. Does this agree with your p -value decision?

Check for understanding. Why does a two-sided 5% test use 1.96 rather than 1.645?