
Lecture 14: Hypothesis Testing

STAT 515 — Statistical Methods I

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

1. Two methods for statistical inference
2. Point estimation
3. Confidence interval estimation
4. Hypothesis testing
5. Review of estimation and hypothesis testing

Purpose

Hypothesis testing aids the researcher, clinician, or administrator in reaching a **decision** concerning a **population** by examining a **sample** from that population.

An example

A city has certain specifications about breaking strengths of pipes in its sewers. The average break point must be more than 2400 pplf (pounds per linear foot). A certain manufacturer wants the contract for the city's sewer pipes, and claims that their pipes meet the specifications. We want to test this claim. We examine a random sample of 50 sections of pipe from this manufacturer. This sample yields an average break point of 2460 pplf with a standard deviation of 200 pplf.

Given: μ is the true mean break point; $n = 50$, $\bar{x} = 2460$, $s = 200$, and the standard error is $\frac{s}{\sqrt{n}}$.

The Six Elements of a Hypothesis Test

1. **Null hypothesis H_0** — a claim about the value of a parameter; the status quo. By convention we put the *equality* under H_0 .
2. **Alternative hypothesis H_a** — the research hypothesis; the claim we accept only if the data strongly indicate it.
3. **Test statistic** — a sample statistic used to decide whether to reject H_0 , computed *assuming H_0 is true*.
4. **Rejection region** — the values of the test statistic that lead us to reject H_0 in favour of H_a .
5. **Conclusion** — the statistical decision plus a sentence in plain English.
6. **Assumptions** — what we assumed about the population (normality, random sampling, sample size).

The Elements of a Hypothesis Test (1) and (2)

(1) Null hypothesis H_0

A theory or claim about the values of parameters. This claim generally represents the **status quo**, and remains the “truth” unless our analysis indicates otherwise.

For the pipes: $H_0 : \mu = 2400$ (by convention, the equal sign ($=$), as well as directional equality ($\geq$ or $\leq$), is always placed under the null hypothesis (H_0).

(2) Alternative hypothesis H_a

Also called the *research hypothesis*, this is a claim that contradicts the null hypothesis. We only accept it as the “truth” if our analysis strongly indicates to do so.

For the pipes: $H_a : \mu > 2400$.

The Elements of a Hypothesis Test (3)

(3) Test statistic

A sample statistic used to decide whether to reject H_0 .

Test statistic for the pipe example

$$Z = \frac{\bar{X} - 2400}{s/\sqrt{n}}$$

Under H_0 (that is, when $\mu = 2400$), this test statistic follows approximately a $N(0, 1)$ distribution.

Signal-to-noise ratio

The Elements of a Hypothesis Test (4)

(4) Rejection region

The set of numerical values for the test statistic that will lead us to reject the null hypothesis in favor of the alternative hypothesis.

For the pipes, with $\alpha = .05$:

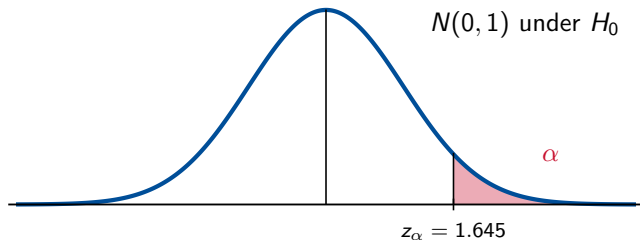
$$RR : \{ Z > z_{.05} \} = \{ Z > 1.645 \}$$

Criterion for setting up the rejection region

Under H_0 , the rejection region (RR) will contain the test statistic with a **small probability** α .

The Rejection Region, Pictured

Under H_0 , the test statistic is $N(0, 1)$; the shaded upper tail has area $\alpha = .05$.



The Level of Significance α

- The values of α are chosen to be small: .01, .05, .10.
- We call α the **level of significance**.
- We usually choose α **before** the test.

For the pipe example we take $\alpha = .05$.

(5) Conclusions

(a) If the numerical value of the test statistic falls in the RR, we **reject** H_0 in favor of H_a . This conclusion will be incorrect (making a type I error) only $(100 \cdot \alpha)\%$ of the time.

(b) If the test statistic is not in the RR, we **fail to reject** H_0 . We may make a type II error.

Observed test statistic

$$Z = \frac{\bar{x} - 2400}{s/\sqrt{n}} = \frac{2460 - 2400}{200/\sqrt{50}} = 2.12$$

- Since $2.12 > 1.645$, the calculated test statistic **falls in the rejection region**.
- Therefore we **reject** H_0 and accept H_a .

(6) Assumptions

Clear statements about any assumptions we have made about the population (normality, equal variances, etc.).

Conclusion for the pipe example

We conclude that the true mean break point of the manufacturer's pipes is **significantly larger than 2400 pplf** at the .05 significance level.

Possible Results of Hypothesis Testing

Conclusion based on data (sample)	“Truth”	
	H_0 true	H_0 false
Reject H_0	Type I error	Correct conclusion
Fail to reject H_0	Correct conclusion	Type II error

Type I Error

$$\begin{aligned}\alpha &= P(\text{Type I error}) \\ &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= \text{probability of rejecting a **true** null hypothesis} \\ &= \text{level of significance}\end{aligned}$$

Aim: keep the Type I error small by specifying a **small** rejection region.

Type II Error

$$\beta = P(\text{Type II error}) = P(\text{failing to reject } H_0 \text{ when } H_0 \text{ is false})$$

$$\text{Power} = 1 - \beta = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is false})$$

- **Aim:** keep the Type II error small and the **power** high.
- **Note:** the value of β is usually unknown, since it depends on the specified alternative value.

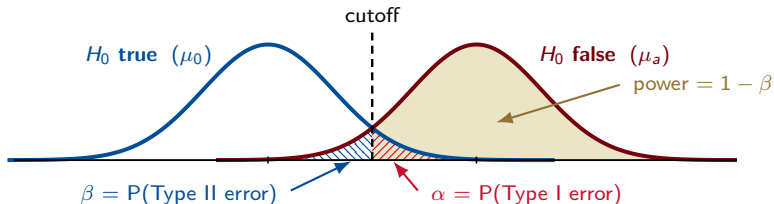
Examples of Decision Situations

Judicial System

- Type I error = **not guilty called guilty**
- Type II error = **guilty called not guilty**

Diagnostic Tests

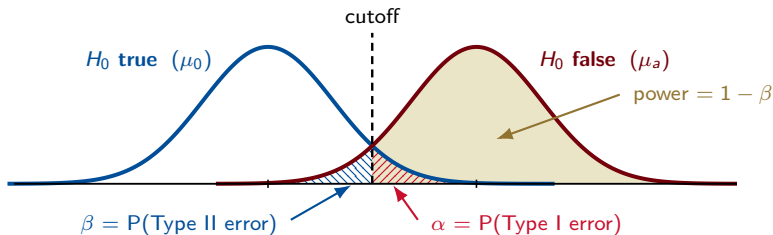
- Type I error = **false positive**
- Type II error = **false negative**



Schematic of α and β

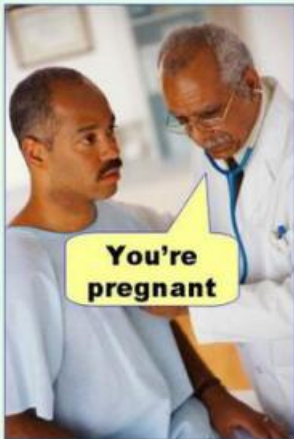
$H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_a$

The Type II error is computed under a *specified* alternative $H_a : \mu = \mu_a$ — move μ_a and β changes.

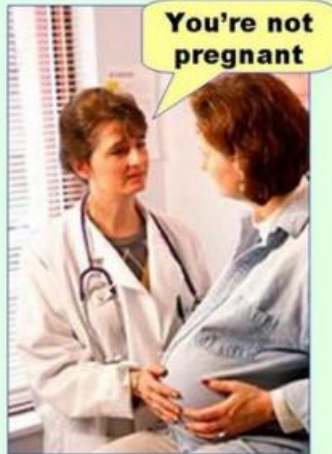


- α area of the H_0 -true curve to the *right* of the cutoff
- β area of the H_0 -false curve to the *left* of the cutoff
- $1 - \beta$ area of the H_0 -false curve to the *right* of the cutoff

Type I error
(false positive)



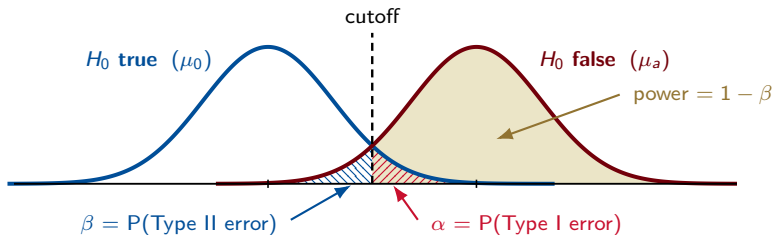
Type II error
(false negative)



Schematic of α and β

$H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_a$

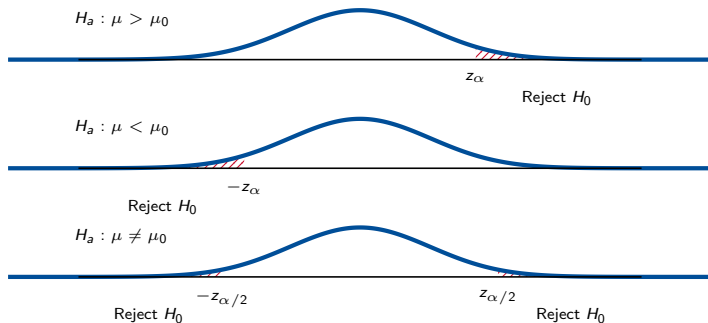
The Type II error is computed under a *specified* alternative $H_a : \mu = \mu_a$ — move μ_a and β changes.



- α area of the H_0 -true curve to the *right* of the cutoff
- β area of the H_0 -false curve to the *left* of the cutoff
- $1 - \beta$ area of the H_0 -false curve to the *right* of the cutoff

- Each probability is measured under **its own curve**: α under the H_0 -true curve, β and $1 - \beta$ under the H_0 -false curve.
- $\alpha + \beta \neq 1$ — they are areas under *different* curves. But $\beta + (1 - \beta) = 1$, since both sit under the same one.
- **Slide the cutoff right:** α shrinks, β grows.
- **Slide the cutoff left:** β shrinks, α grows.
- **Pull the two curves apart** (bigger effect, or larger n): both shrink at once.

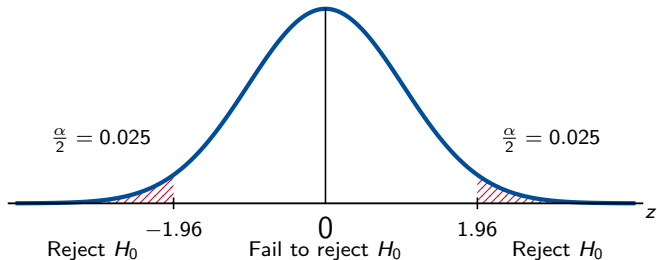
Types of Rejection Regions (Critical Regions)



A one-tailed alternative puts *all* of α in one tail; the two-tailed alternative splits it into two tails of area $\alpha/2$.

Example of a Critical Region

$H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
with $\alpha = 0.05$ split between the two tails.



$z_{\alpha/2}$ is the critical value corresponding to a probability of $\alpha/2$ in one tail of the curve (read from a z table).

p -Value

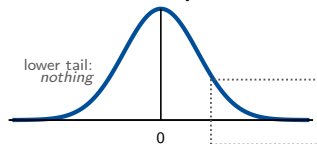
The p -value for a hypothesis test is the probability of obtaining, by chance alone **when H_0 is true**, a value of the test statistic as extreme or more extreme (in the appropriate direction) than the one actually computed.

- The **rejection region** is determined by α , the desired level of significance (the probability of committing a Type I error).
- Reporting the p -value indicates how common or rare the computed value of the test statistic is, **given that H_0 is true**.
- z_{obs} (or t_{obs}) = the computed value of the test statistic.

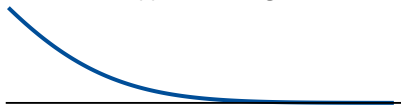
One-Sided: All of α Sits in One Tail

The pipe example: $H_0 : \mu = 2400$ vs. $H_a : \mu > 2400$, $z_{obs} = \frac{2460 - 2400}{200/\sqrt{50}} = 2.12$, $\alpha = .05$.

The whole picture



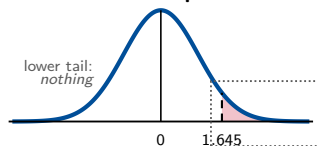
The upper tail, magnified



One-Sided: All of α Sits in One Tail

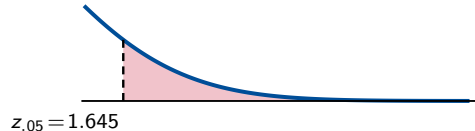
The pipe example: $H_0 : \mu = 2400$ vs. $H_a : \mu > 2400$, $z_{obs} = \frac{2460 - 2400}{200/\sqrt{50}} = 2.12$, $\alpha = .05$.

The whole picture



The upper tail, magnified

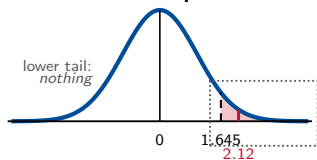
shaded = $\alpha = .05$



One-Sided: All of α Sits in One Tail

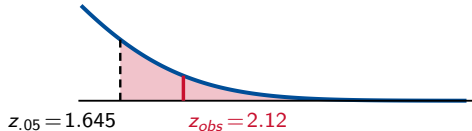
The pipe example: $H_0 : \mu = 2400$ vs. $H_a : \mu > 2400$, $z_{obs} = \frac{2460 - 2400}{200/\sqrt{50}} = 2.12$, $\alpha = .05$.

The whole picture



The upper tail, magnified

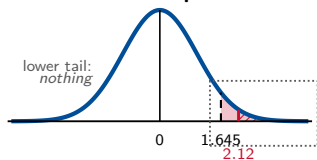
shaded = $\alpha = .05$



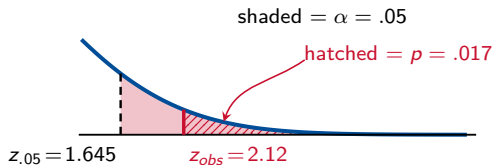
One-Sided: All of α Sits in One Tail

The pipe example: $H_0 : \mu = 2400$ vs. $H_a : \mu > 2400$, $z_{obs} = \frac{2460 - 2400}{200/\sqrt{50}} = 2.12$, $\alpha = .05$.

The whole picture



The upper tail, magnified



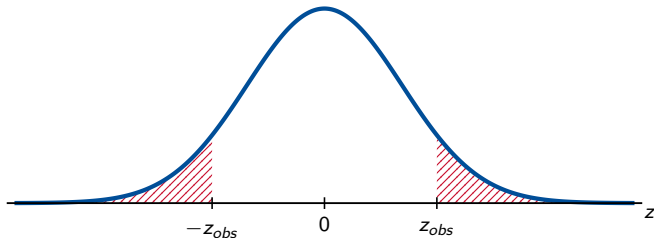
The same comparison, with nothing halved

$$\underbrace{2.12 > 1.645}_{\text{observed vs. cut-off}} \iff \underbrace{.017 < .05}_{\text{tail areas}} \implies \text{reject } H_0$$

Two-Sided p -Value

For a two-sided test,

$$p\text{-value} = P(Z \geq z_{obs}) + P(Z \leq -z_{obs})$$



Convention for judging “statistical significance”

- $p < 0.05$: **statistically significant**
 - $p > 0.05$: NS (not statistically significant)
- **“Statistical” significance is not the same as “practical” significance.**
With a very large n , a tiny and unimportant difference can be statistically significant.

Example 1: Average Ocean Depth

Example 1

An oceanographer wants to test, on the basis of a random sample of size 35, whether the average depth of the ocean in a certain area is 72.4 fathoms. At the .05 level of significance, what will the oceanographer decide if she observed a sample mean of 73.2? Assume that the population standard deviation is 2.1.

Given: $n = 35$, $\bar{x} = 73.2$, $\sigma = 2.1$, $\alpha = .05$, $\mu_0 = 72.4$.

Example 1: Hypotheses and Test Statistic

Hypotheses

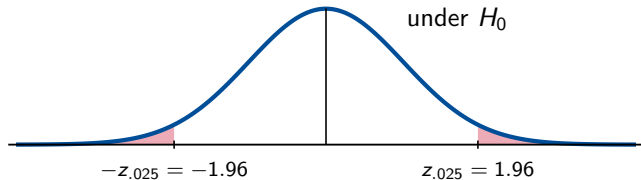
$$H_0 : \mu = 72.4 \quad \text{vs.} \quad H_a : \mu \neq 72.4$$

Test statistic

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{73.2 - 72.4}{2.1/\sqrt{35}} = \mathbf{2.254}$$

Example 1: The Rejection Region

$$\begin{aligned} RR : \{Z > z_{\alpha/2} \text{ or } Z < -z_{\alpha/2}\} &= \{|Z| > z_{\alpha/2}\} \\ &= \{Z > z_{.025} \text{ or } Z < -z_{.025}\} = \{Z > 1.96 \text{ or } Z < -1.96\} \end{aligned}$$



Example 1: Decision and Conclusion

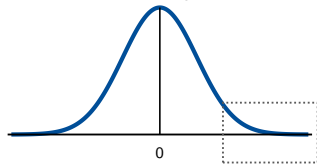
- Since $|2.254| > 1.96$, the test statistic **falls in the rejection region**.
- Therefore we **reject H_0** at $\alpha = .05$ (at the .05 significance level).

Conclusion

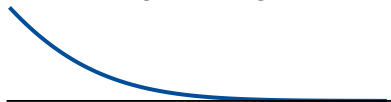
We conclude that the true mean depth is **significantly different from 72.4 fathoms**.

The Cut-off and the p -Value Are One Comparison

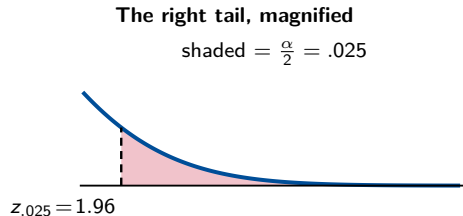
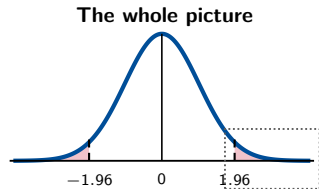
The whole picture



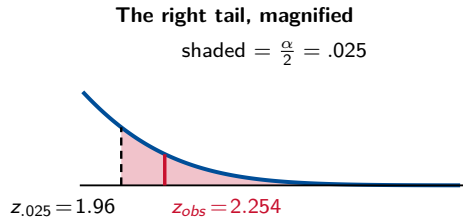
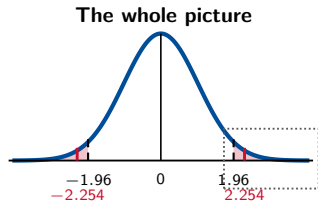
The right tail, magnified



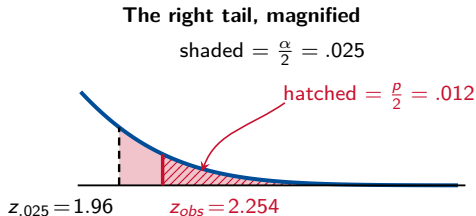
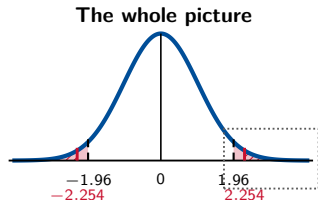
The Cut-off and the p -Value Are One Comparison



The Cut-off and the p -Value Are One Comparison



The Cut-off and the p -Value Are One Comparison



One comparison, read three ways

$$\underbrace{2.254 > 1.96}_{\text{observed vs. cut-off}} \iff \underbrace{.012 < .025}_{\text{tail areas}} \iff \underbrace{.024 < .05}_{p \text{ vs. } \alpha} \implies \text{reject } H_0$$

The hatched region sits *inside* the shaded one — that is what $p < \alpha$ looks like. Slide z_{obs} back toward 0 and the hatching spills past the dashed line: then $p > \alpha$, and we fail to reject.

1. Two methods for statistical inference
2. Point estimation
3. Confidence interval estimation
4. Hypothesis testing
5. Review of estimation and hypothesis testing

■ Estimation

- What is the best estimate?
- Construct a 95% confidence interval — does the interval cover the hypothesized value?

■ Hypothesis testing

- Does the observed test statistic fall within the rejection region of the sampling distribution?
- Is the associated p -value less than the significance level (usually $\alpha = 0.05$)?

Key point

These procedures lead to the same inference.

The duality between a test and an interval

For a **two-sided** test of $H_0 : \mu = \mu_0$ at level α :

reject $H_0 \iff$ the $100(1 - \alpha)\%$ CI **excludes** μ_0

fail to reject $H_0 \iff$ the $100(1 - \alpha)\%$ CI **covers** μ_0

- Both use the **same** $\bar{x}$, the same SE , and the same critical value — so they *must* agree.
- $|z_{obs}| > z_{\alpha/2}$ says $\bar{x}$ is more than $z_{\alpha/2}$ standard errors from μ_0 ; that is exactly the statement that μ_0 lies outside $\bar{x} \pm z_{\alpha/2} SE$.

Exercise A: Ocean Depth — the Test

Exercise A

An oceanographer tests, from a random sample of $n = 35$, whether the average depth in a certain area is 72.4 fathoms. She observes $\bar{x} = 73.2$ and knows $\sigma = 2.1$. Use $\alpha = .05$.

$$H_0 : \mu = 72.4 \quad \text{vs.} \quad H_a : \mu \neq 72.4$$
$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2.1}{\sqrt{35}} = 0.355, \quad z_{obs} = \frac{73.2 - 72.4}{0.355} = \mathbf{2.254}$$

Since $|2.254| > 1.96$, **reject** H_0 at the .05 level.

$$p = 2P(Z > 2.254) = \mathbf{.024} < .05. \quad \checkmark$$

Exercise A: Now the Confidence Interval

Same $\bar{x}$, same SE , same critical value — just rearranged:

$$\bar{x} \pm z_{.025} \left(\frac{\sigma}{\sqrt{n}} \right) = 73.2 \pm 1.96(0.355) = 73.2 \pm 0.696 = (\textbf{72.504}, \textbf{73.896})$$

Same conclusion

The hypothesized value $\mu_0 = 72.4$ is **not inside** the interval. So 72.4 is not a plausible value for μ — which is precisely what “reject H_0 ” said.

Bonus: the interval also tells us *where* μ probably is (72.5 to 73.9 fathoms) — information the p -value alone does not give.

Exercise B: A Different Sample Mean

Exercise B

Everything as in Exercise A, except the oceanographer observes $\bar{x} = 72.9$ instead of 73.2.

The test:

$$z_{obs} = \frac{72.9 - 72.4}{0.355} = \mathbf{1.409}, \quad |1.409| < 1.96 \implies \mathbf{\text{fail to reject } H_0}$$

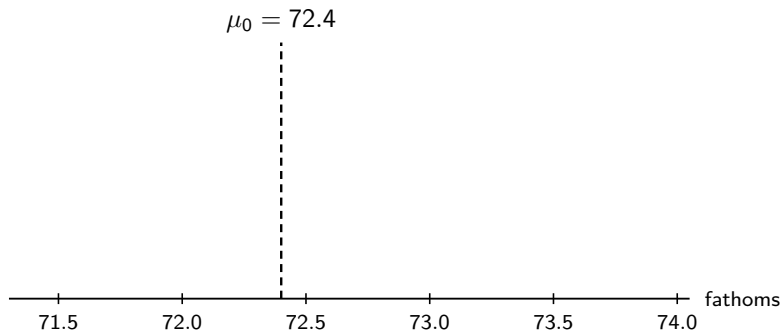
$$p = 2P(Z > 1.409) = \mathbf{.159} > .05. \quad \checkmark$$

The interval:

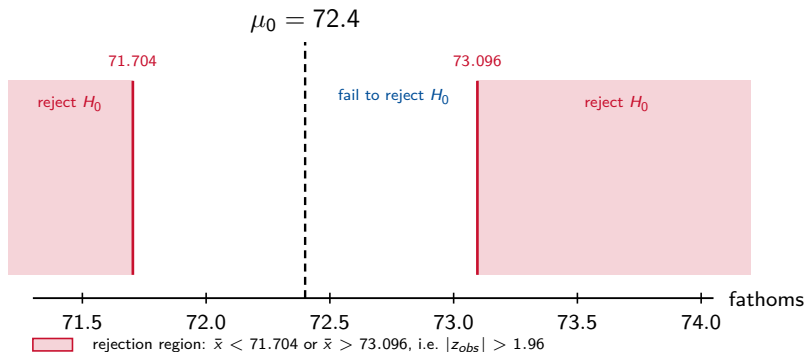
$$72.9 \pm 1.96(0.355) = 72.9 \pm 0.696 = (\mathbf{72.204}, \mathbf{73.596})$$

$\mu_0 = 72.4$ **is inside** the interval — 72.4 remains plausible. Same conclusion again.

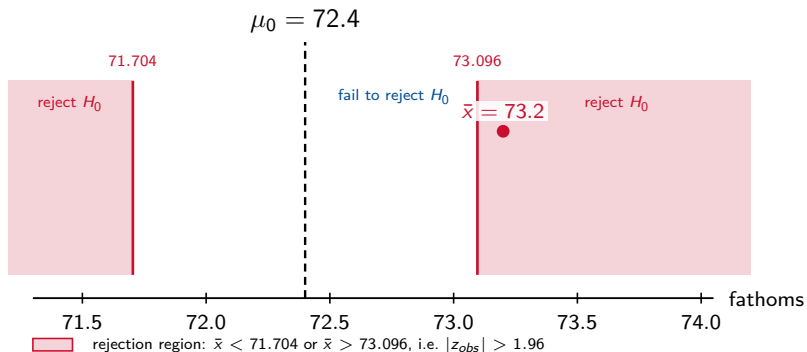
Seeing Both Exercises at Once



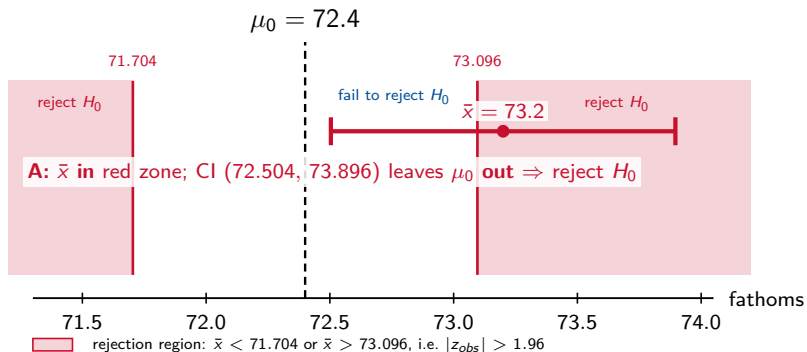
Seeing Both Exercises at Once



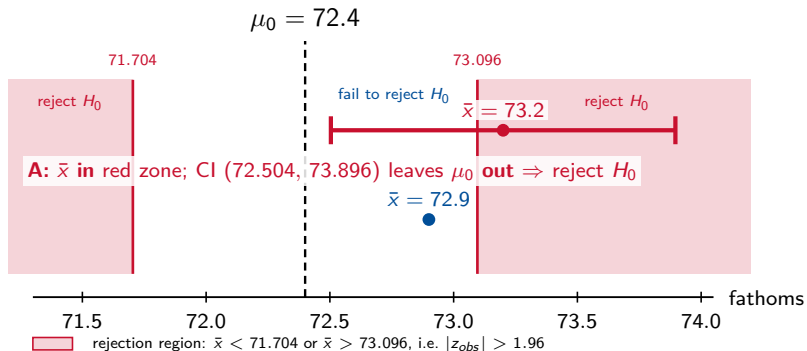
Seeing Both Exercises at Once



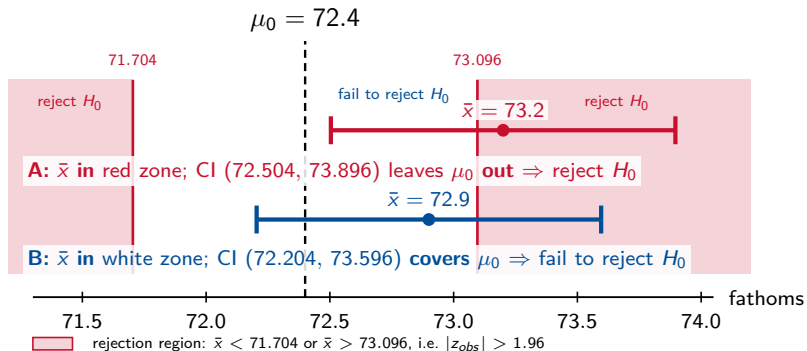
Seeing Both Exercises at Once



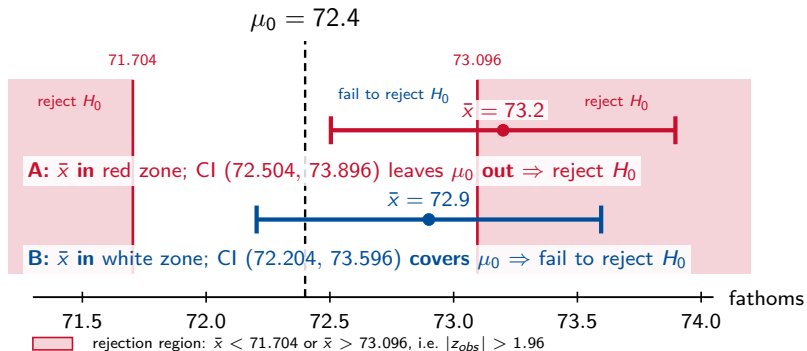
Seeing Both Exercises at Once



Seeing Both Exercises at Once



Seeing Both Exercises at Once



The red cutoffs sit 0.696 from μ_0 — exactly the CI half-width. So “ $\bar{x}$ lands in the red zone” and “the interval misses μ_0 ” are the *same* statement, read from opposite ends.

Three Procedures, One Inference

	Exercise A ($\bar{x} = 73.2$)	Exercise B ($\bar{x} = 72.9$)
Rejection region	$ 2.254 > 1.96$ reject	$ 1.409 < 1.96$ fail to reject
p-value	$.024 < .05$ reject	$.159 > .05$ fail to reject
95% CI	$(72.504, 73.896)$ μ_0 outside $\Rightarrow$ reject	$(72.204, 73.596)$ μ_0 inside $\Rightarrow$ fail to reject

Key point

All three procedures result in the same inference — every time.

Two Cautions

- The equivalence requires **matching** pieces: a **two-sided** test at level α pairs with a **two-sided** $100(1 - \alpha)\%$ CI.
 - A one-sided test at α pairs with a *one-sided* bound, not with the usual two-sided interval.
 - A test at $\alpha = .01$ pairs with a 99% CI, not a 95% one.
- The two outputs answer *different questions*:
 - The p -value says **how surprising** the data would be if H_0 were true.
 - The CI says **which values of μ are plausible**, and how precisely we have pinned μ down.

Practical advice

Report **both**. “Significantly different from 72.4 ($p = .024$); 95% CI (72.50, 73.90) fathoms” tells the whole story in one line.

The Three Forms of the Alternative

Alternative H_a	Rejection region
$H_a : \mu > \mu_0$	$\{ Z > z_\alpha \}$
$H_a : \mu < \mu_0$	$\{ Z < -z_\alpha \}$
$H_a : \mu \neq \mu_0$	$\{ Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2} \}$

- One-tailed alternatives put all of α in a single tail.
- The two-tailed alternative splits α into two tails of area $\alpha/2$ each.

- Confidence intervals for μ (with σ known or unknown), for p , and for σ — always followed by a plain-language interpretation.
- Needed sample size: $n = \left(\frac{z_{\alpha/2}\sigma}{SE}\right)^2$ for a mean and $n = \frac{z_{\alpha/2}^2 pq}{(SE)^2}$ for a proportion — always round up.
- The six elements of a hypothesis test: H_0 , H_a , test statistic, rejection region, conclusion, assumptions.
- The large-sample z test for μ , with one- and two-tailed rejection regions.