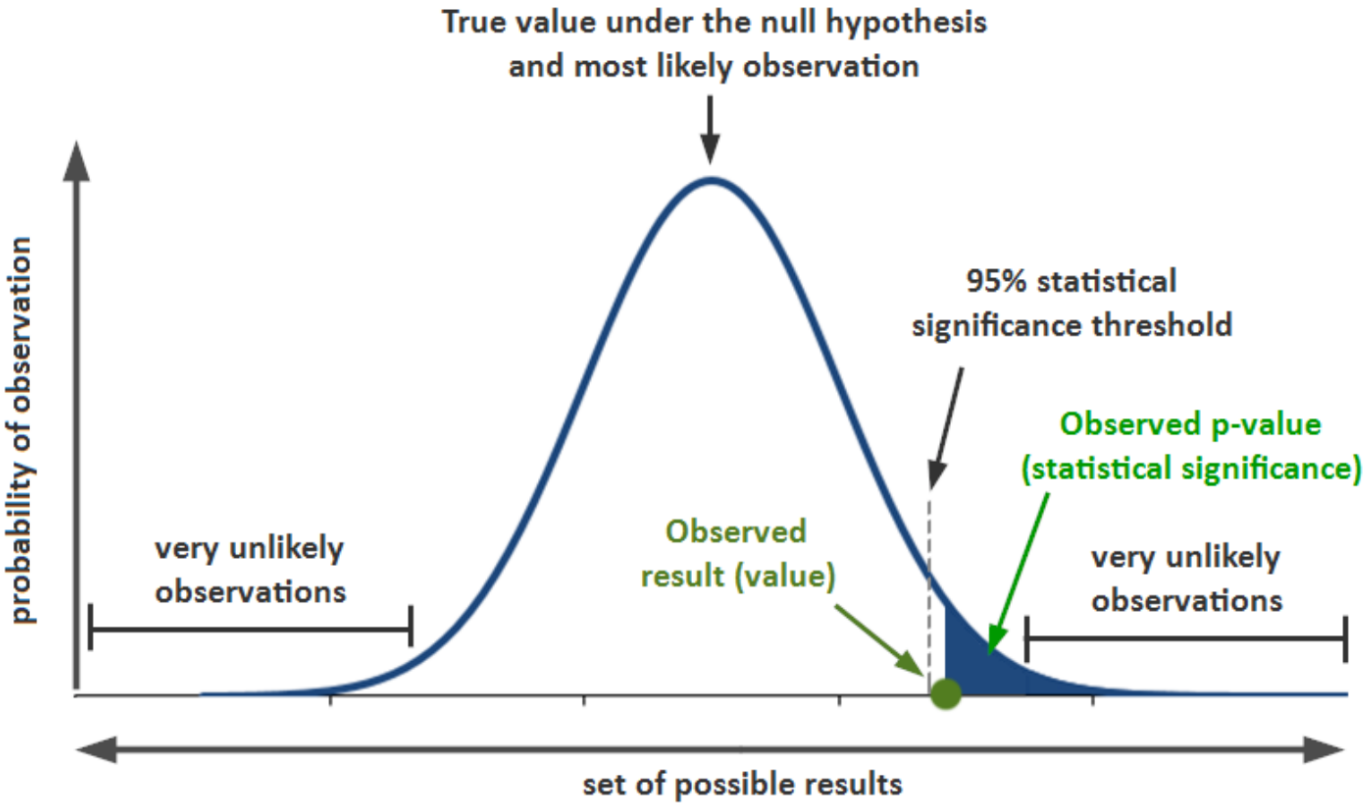
CSCI 3022: Intro to Data Science

Lecture 18: Hypothesis Testing with p–Values

Rachel Cox

Department of Computer Science



Source

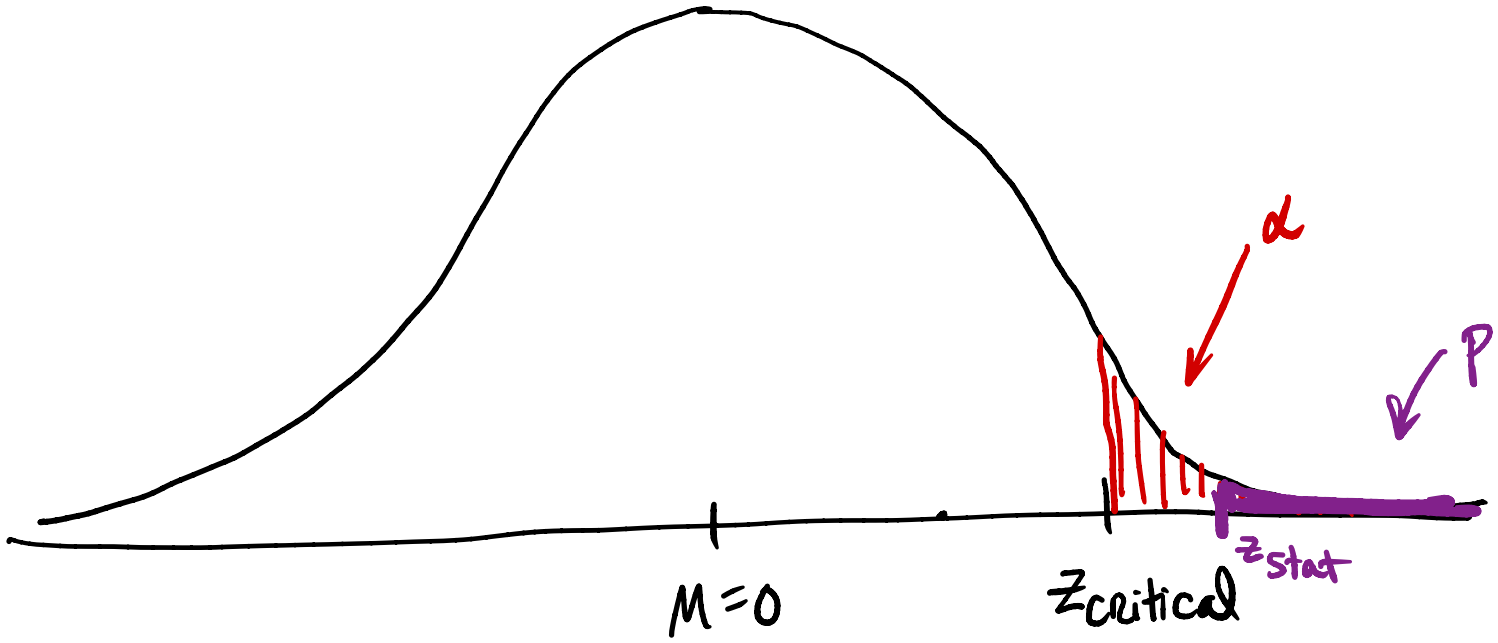
Because $z_{stat} > z_{critical}$

$\Rightarrow$ Reject the null hypothesis

Overview of P-values

upper tail test: $p < \alpha$

α = significance level of the test



A p-value is a probability value

★ p-value = P(observing data at least as extreme as our sample statistic | given H_0 true)

What will we learn today?

- ❑ Hypothesis testing
- ❑ Significance level – how much evidence do you need in order to reject the null hypothesis
- ❑ Rejection regions – if your test statistic falls in here, you have evidence to reject the null hypothesis
- ❑ Type I and Type II errors – false positive and false negatives, respectively

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 25 & 26*

Review from Last Time

A **statistical hypothesis** is a claim about the value of a parameter of a population characteristic.

H_0 , H_A

- The objective of hypothesis testing is to choose, based on sampled data, between two competing hypotheses about the value of a population parameter.

A **test statistic** is calculated from the sample data, assuming that the null hypothesis is true. It is used in the decision about whether or not to reject the null hypothesis.

The **rejection region** is a range of values of the test statistic that would lead you to reject the null hypothesis.

Reminder: A $100 \cdot (1-\alpha)\%$ CI for the mean μ when the value of σ is known is given by:

$$\left[\bar{x} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right] .$$

Critical Region Hypothesis Test

Example: The 1999 Volkswagon Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

→ $n = 100$, large enough to apply CLT

Question: What are the null Hypothesis and Alternative Hypothesis to test the claim that there is statistical evidence that the 1999 Jettas made in Mexico have a lower life expectancy than those made in Germany?

H_0: $\mu = 300{,}000$ miles

H_1: $\mu < 300{,}000$ miles

one-tailed test (lower tail)

Critical Region Hypothesis Test

Example: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy than those made in Germany? Carry out a Critical Region test at the 0.01 significance level.

lower-tailed test ✓

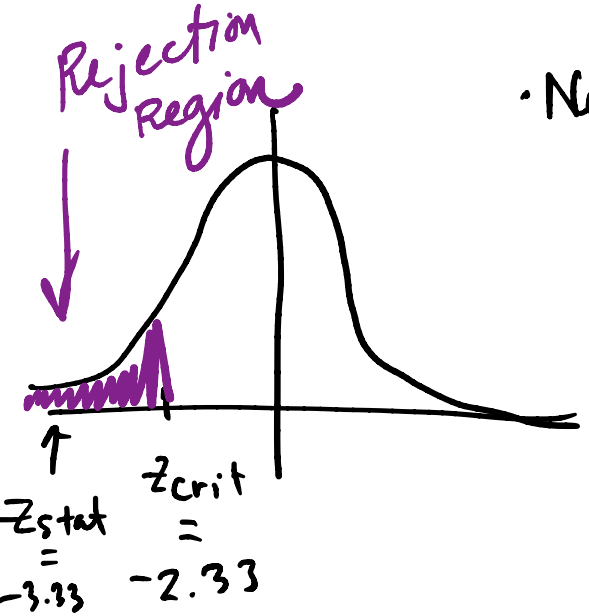
↑ hypothesis test

aka "Z-test"

$\alpha = 0.01$

- Need $z_{critical}$, $z_{critical} = stats.norm.ppf(0.01)$

$= -2.32634787$



- Next $z_{stat} = \dfrac{250{,}000 - 300{,}000}{150{,}000/\sqrt{100}}$

$= -3.33$

Since $z_{stat} < z_{crit}$

$-3.33 < -2.33$ ⇒ Reject the null hypoth.

Critical Region Hypothesis Test Summary

Alternative Hypothesis	Rejection Region for Level α test
$H_1: \ \theta > \theta_0$	$z \geq z_\alpha$
$H_1: \ \theta < \theta_0$	$z \leq -z_\alpha$
$H_1: \ \theta \neq \theta_0$	$\left(z \geq z_{\alpha/2}\right)$ or $\left(z \leq -z_{\alpha/2}\right)$

- Critical region is the region where the statistic has a low probability under the Null Hypothesis
- Requires normally distributed data, or large enough sample to use the Central Limit Theorem
- Under these assumptions, we call this a Z-test

Type I Error – rejecting the Null when the Null is True.
Type II Error – failing to reject the Null when the Null is False.

Introduction to p–values

A **p-value** is the probability, under the Null Hypothesis, that we would get a test statistic at least as extreme as the one we calculated.

observation

For a lower-tailed test with test statistic x, the p-value is equal to $P(X \leq x \,|H_0)$

upper-tailed test: $P(X \geq x \mid H_0)$

Intuition: The p-value assesses the extremeness of the test statistic. The smaller the p-value, the more evidence we have against the Null Hypothesis.

Important Notes:

- The p-value is calculated under the assumption that the Null Hypothesis is true.
- The p-value is always a value between 0 and 1.
- The p-value is NOT the probability that the Null Hypothesis is true.

Introduction to p–values

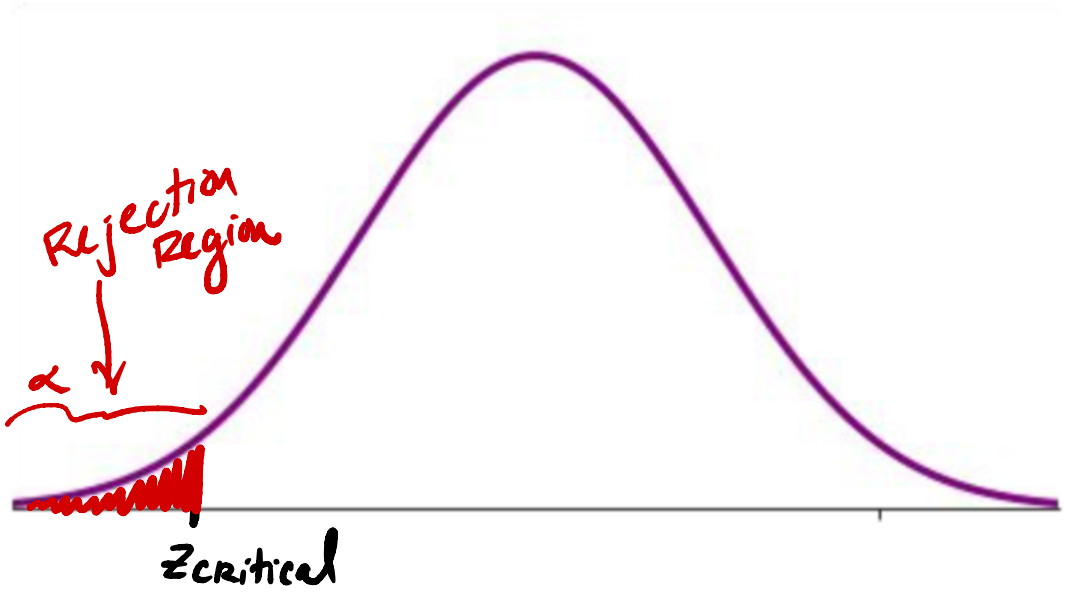
Example: Consider a lower-tail critical region test with the following hypotheses:

$$H_0: \ \mu = \mu_0$$
$$H_1: \ \mu < \mu_0$$

The critical region to test is:

Reject if $z_{stat} \leq z_{crit}$

$p \leq \alpha$



Introduction to p–values

As before, select a significance level α before performing the hypothesis test.

How do we select α?

Then the decision rule is:

- If p-value $\leq \alpha$, then reject the Null Hypothesis.
- If p-value $> \alpha$, then fail to reject the Null Hypothesis.

If the p-value exceeds the selected significance level, then we cannot reject the Null.

- unethical to set α threshold after the test.
- lower α, reduces the number of Type I errors, but it will increase the chance of a type II error.

Introduction to p-values

Example: The 1999 Volkswagon Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

Is there sufficient evidence to conclude that 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out the p-value test at the 0.01 significance level.

$\alpha = 0.01$

we already know $z_{stat} = -3.33$

$p\text{-value} = P(\bar{X} \leq 250{,}000 \mid H_0 \text{ true})$

$= P(z \leq -3.33)$

$= \text{stats.norm.ppf}(-3.33) = 0.00043$

$< \alpha$

Introduction to p–values

Example: The 1999 Volkswagon Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

Is there sufficient evidence to conclude that 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out the p-value test at the 0.01 significance level.

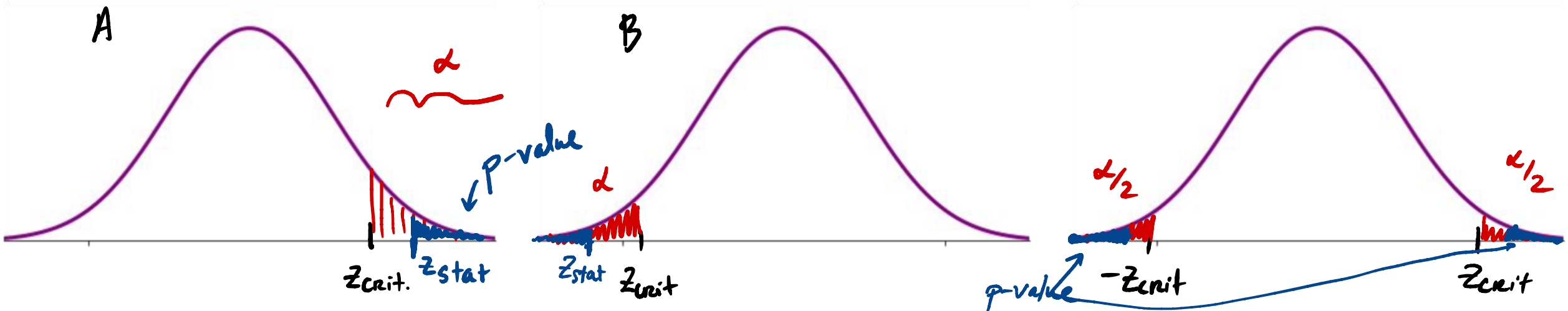
Conclusion : Because p-value $< \alpha$ we reject the null hypothesis.

P–values for different Z–tests

	Alternative hypothesis	Rejection region for level α test	p-value level α test
A	$H_1: \theta > \theta_0$ (one-tailed)	$z \geq z_\alpha$	p-value = $1 - \Phi(z)$
B	$H_1: \theta < \theta_0$ (one-tailed)	$z \leq z_\alpha$	p-value = $\Phi(z)$
C	$H_1: \theta \neq \theta_0$ (two-tailed test)	$(z \geq z_{\alpha/2})$ or $(z \leq z_{\alpha/2})$	p-value = $2 \cdot \Phi(-\lvert z \rvert)$

z-test

p-value



Introduction to p–values

Example: To test if the Belgian 1 Euro coin is fair, you flip it 100 times and observe 38 heads. Perform a p-value test at the 5% significance level.

$H_0: p = 0.5$

$H_1: p \neq 0.5$ ← $\alpha = .05$

Compute the test statistic.

$$z_{stat} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{38}{100} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -2.4$$

$$\text{p-value} = 2 \cdot \Phi(-|z_{stat}|)$$

$$= 0.0164$$

$$< 0.05 = \alpha \Rightarrow \text{Reject the null.}$$

Two-sample testing for difference of means

Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value.

What kind of Null and Alternative Hypotheses might we want to test?

$H_0: \quad \mu_1 = \mu_2 \longrightarrow H_0: \mu_1 - \mu_2 = 0$

(or $\mu_1 - \mu_2 = c$)

$H_1: \quad \mu_1 > \mu_2 \longrightarrow H_1: \mu_1 - \mu_2 < 0$

or

$\mu_1 - \mu_2 > 0$

$\mu_1 < \mu_2$

$\mu_1 \neq \mu_2 \longrightarrow \mu_1 - \mu_2 \neq 0$

Two–sample testing for difference of means

Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value.

Assuming that our sample sizes are large enough, we can standardize our test statistics as:

↘ Apply the Central Limit Theorem

Holds for samples ≥ 30

- Test statistic = z = $\dfrac{(\overline{x_1} - \overline{x_2}) - c}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Two-sample testing for difference of means

Example: Data on calorie intake (per day) for a sample of teens who reported they do not typically eat fast food and another sample of teens who said they do, is as follows:

Fast Food?	Sample Size	Sample Mean	Sample SD
no	663	2258	1519
yes	413	2637	1138

Do these data provide statistical evidence at the 0.05 significance level that the true average calorie intake for teens who typically eat fast food exceeds that of teens who do not, by more than 200 calories per day?

μ_1 = mean for those who do eat fast food

μ_2 = mean for those who don't

$H_0: \mu_1 - \mu_2 = 200$

$H_1: \mu_1 - \mu_2 > 200$

Two-sample testing for difference of means

Example: (continued) Data on calorie intake (per day) for a sample of teens who reported they do not typically eat fast food and another sample of teens who said they do, is as follows:

Fast Food?	Sample Size	Sample Mean	Sample SD
no	663	2258	1519
yes	413	2637	1138

$\bar{x}_1$ $\bar{x}_2$

$\alpha = 0.05$

Do these data provide statistical evidence at the 0.05 significance level that the true average calorie intake for teens who typically eat fast food exceeds that of teens who do not, by more than 200 calories per day?

$$Z_{stat} = \frac{(2637 - 2258) - 200}{\sqrt{\frac{1138^2}{413} + \frac{1519^2}{663}}} \approx 2.20$$

Next, we compute the p-value.

p-value $= 1 - \Phi(2.20)$

$= 1 - \text{stats.norm.cdf}(2.20)$

$= 0.014$

$< 0.05 = \alpha$

$\Rightarrow$ Reject the null hypothesis

Two–sample testing for difference of means

Example: (continued) Data on calorie intake (per day) for a sample of teens who reported they do not typically eat fast food and another sample of teens who said they do, is as follows:

Fast Food?	Sample Size	Sample Mean	Sample SD
no	663	2258	1519
yes	413	2637	1138

Do these data provide statistical evidence at the 0.05 significance level that the true average calorie intake for teens who typically eat fast food exceeds that of teens who do not, by more than 200 calories per day?

Common p–value misunderstandings

Misconception #1: If p = 0.05, then Null Hypothesis has a 5% chance of being true.

$$\text{p-value} = P(\text{observing data at least as extreme as your test statistic} \mid \text{Null Hypothesis is True}) \neq P(H_0 \text{ true})$$

Misconception #2: If p is very small, then your alternate hypothesis is very likely to be significant.

Misconception #3: A statistically significant effect is equivalent to a substantial effect.

$$\text{CI width: } 2 * Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

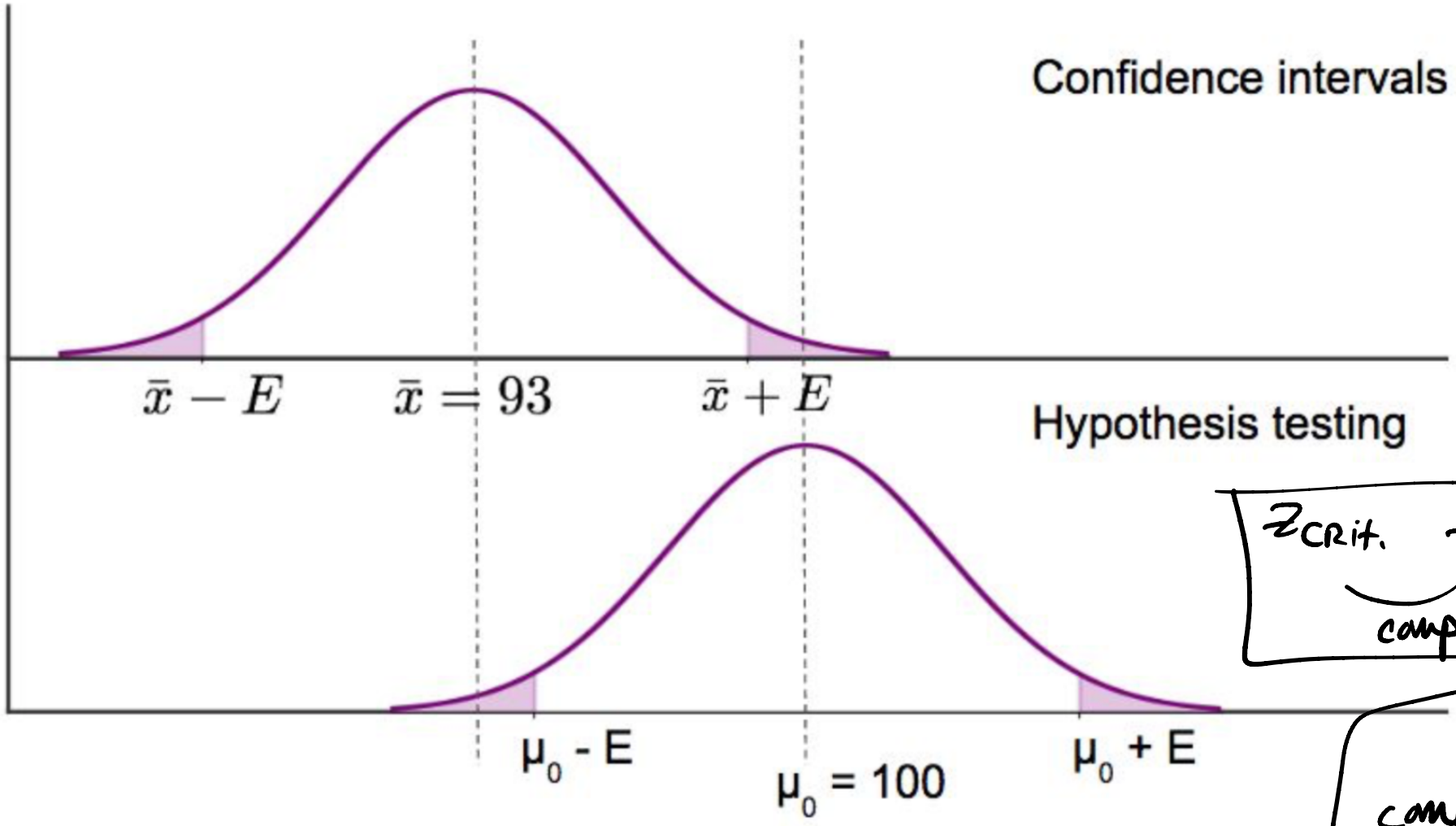
let $n \to \infty$

$\Rightarrow$ CI width $\to 0$

with larger + larger amounts of data, everything becomes statistically significant

CIs vs. Critical Regions vs. p–Values

Confidence intervals, critical regions, and p-values are three ways to tackle the same question.



$z_{crit.}$ z_{stat}
compare

p-values
compare p-value, α

Next Time:

- Small Sample Hypothesis Testing