CSCI 3022

intro to data science with probability & statistics

October 26, 2018

Introduction to *p*-values and hypothesis testing

- Check in using App
- your topic here data blog
- Next Weds, no office hrs.

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Switching advertising strategies

$H_0: \mu = 200$ $H_1: \mu > 200$

- **Example**: Suppose a company is considering hiring a new outside advertising company to help generate traffic to their website. Under their current advertising they get, on average, 200 thousand hits per day with a standard deviation of 50 thousand hits per day. You decide to hire the new ad company for a 30 day trial. During those 30 days, your website gets 210 thousand hits per day. Perform a hypothesis test to determine if the new ad campaign outperforms the old one at the .05 significance level.

$z_\alpha = z_{0.05} = 1.645$

CLT $N\left(\mu, \frac{\sigma^2}{n}\right)$

If null H_0 were true

$\bar{X} \sim N\left(200, \frac{50^2}{30}\right)$

$Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$

$\frac{210 - 200}{50/\sqrt{30}} = 1.1$

Important assumptions

- **Question**: What assumptions did we make in the previous example?

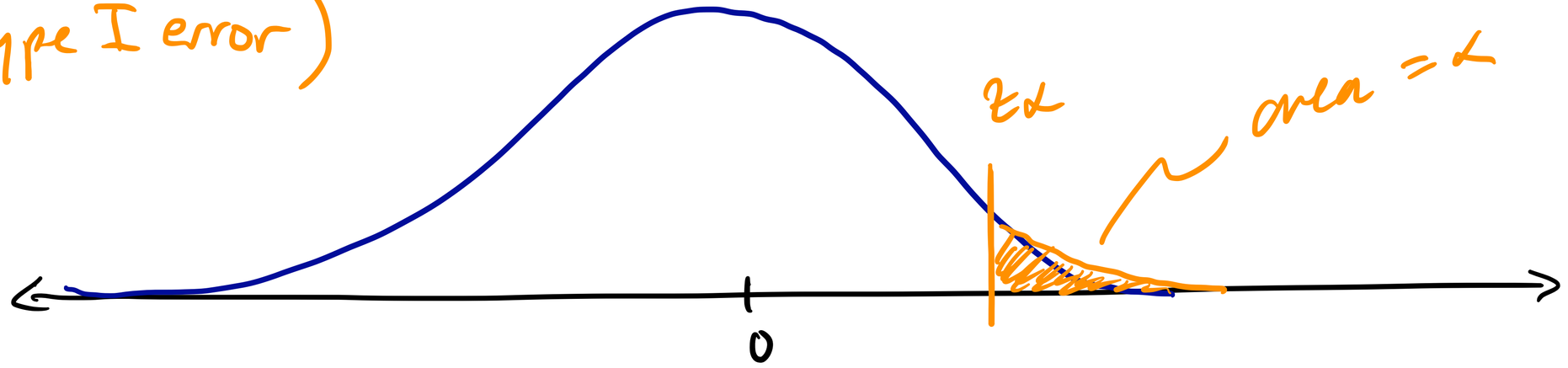
(1) Assumed that the CLT would hold. $n=30$ samples (days)

(2) Assumed that we can represent the involved distributions as normals.

Errors in hypothesis testing

- **Definitions**:
- A **Type I Error** occurs when the Null hypothesis is rejected, but the Null hypothesis is in fact true (**False Positive**)
- A **Type II Error** occurs when the Null hypothesis is not rejected, but the Null hypothesis is in fact false (**False Negative**)
- **Question**: What is the probability that we commit a **Type I Error**?

$\alpha = Pr(\text{Type I error})$



Errors in hypothesis testing

- **Definitions**:
- A **Type I Error** occurs when the Null hypothesis is rejected, but the Null hypothesis is in fact true (**False Positive**)
- A **Type II Error** occurs when the Null hypothesis is not rejected, but the Null hypothesis is in fact false (**False Negative**)
- **Question**: What is the probability that we commit a **Type I Error**?
- **Answer**: this is exactly the significance level α
- **Consequence**: choose α by considering willingness to risk a Type I error.

Think "wallet inspector" or see paper from MSR Cormac Hurley?

Rejection region refresher

- **Example**: The 1999 Volkswagen 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 300K miles and standard deviation 150K 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 250K miles.
- **Question 1**: What are the Null hypothesis and alternative hypothesis to test the claim that there is statistical evidence that 1999 Jettas made in Mexico have a smaller life expectancy that those made in Germany?

$H_0: \mu = 300k$

$H_1: \mu < 300k$

$\mu =$ life expectancy of Jetta made in Mexico.

Rejection region refresher

- **Example**: The 1999 Volkswagen 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 300K miles and standard deviation 150K 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 250K miles.

- **Question 2**: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a rejection region test at the 0.01 significance level.

$n=100$ $\bar{X}=250$

$\mu=300$ under H_0

$\sigma=150$ under H_0

$\alpha=0.01$

CLT: $\bar{X} \sim N\left(300, \frac{150^2}{100}\right)$

Box Muller:

$$\text{test stat} = \frac{250-300}{150/\sqrt{100}} = -3.33$$

stats.norm.ppf( ) = $z_?$

Rejection region refresher

- **Example**: The 1999 Volkswagen 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 300K miles and standard deviation 150K 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 250K miles.

- **Question 2**: Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a rejection region test at the 0.01 significance level.

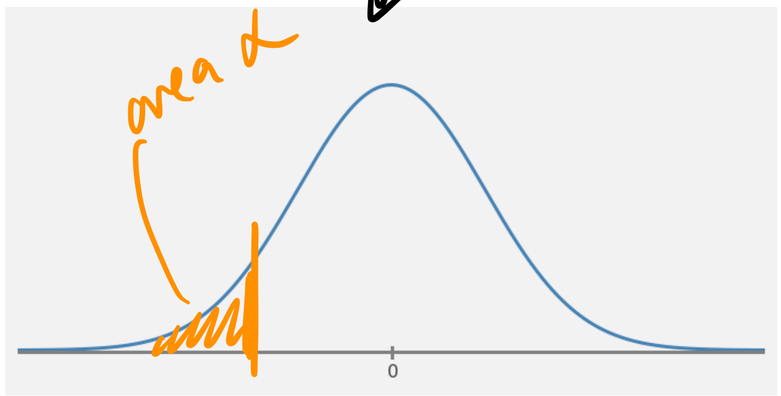
$z_{critical} = stats.norm.ppf(0.01)$

$= stats.norm.ppf(0.99) \times (-1)$

$= -2.33$

Rejection region & critical value summary

Alternative Hypothesis	Rejection Region
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$



Critical region HT summary

- **Critical Region** is region where test statistic has low probability under Null Hypothesis.
- Requires normally distributed data, or large enough sample for Central Limit Theorem.
- Under these assumptions we call this a Z-Test
- Rejecting the Null when the Null is true is called a Type I Error
- The probability of committing a Type I Error is α , the significance level of the test.
- Failing to reject the Null when the Null is false is called a Type II Error

Introduction to p-values

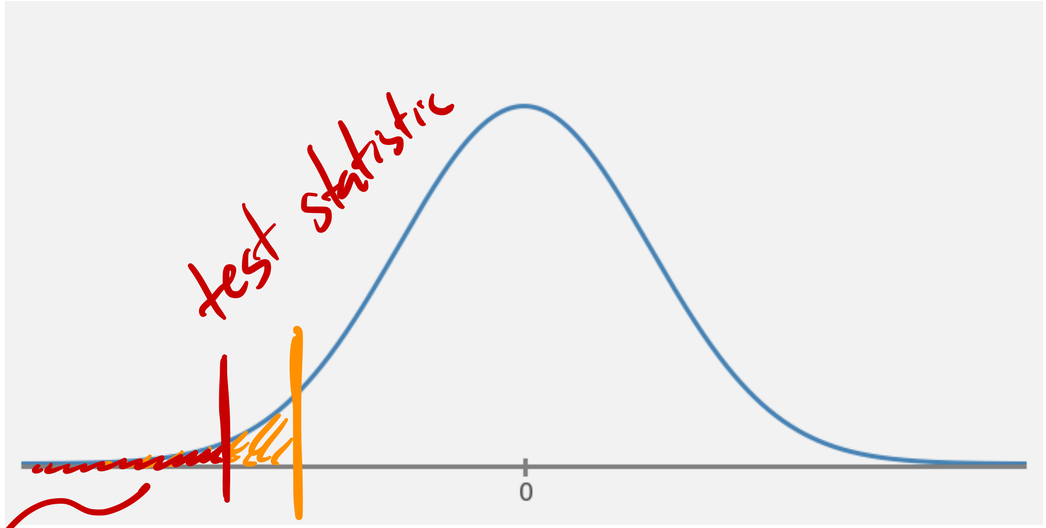
- Another way to view the critical region hypothesis test is through a so-called p-value
- This framework for HT is very popular in scientific study and reporting
- **Example**: Consider a lower-tail critical region test with the following hypotheses.

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

- The critical region test is:

Comparing test statistic to $z_{critical}$.

Comparing p-value to α.



p-values for various hypothesis tests

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

The p-value decision rule

- As before, select a significance level α before performing the hypothesis test
- 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
- Thus if the p-value exceeds the selected significance level then we cannot reject the Null hypothesis.

e.g. if p-value = 0.1, and $\alpha = 0.05$, then we cannot reject the null hypothesis.

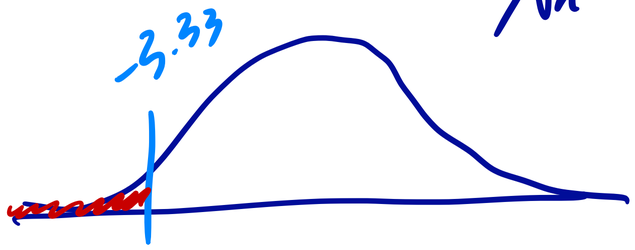
- Note: The p-value can be thought of as the smallest significance level at which the Null hypothesis can be rejected.

Jetta life expectancy with p-values

- **Example**: The 1999 Volkswagen 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 300K miles and standard deviation 150K 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 250K 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 SL.

μ_0 is mean under H_0

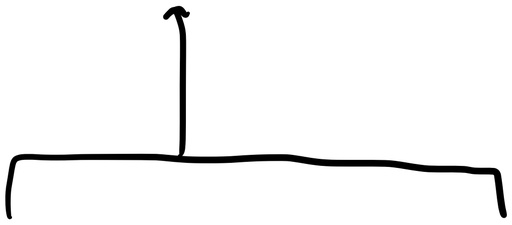
test stat: $\dfrac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} = \dfrac{250 - 300}{150/\sqrt{100}} = -3.33$



pval = CDF(-3.33)

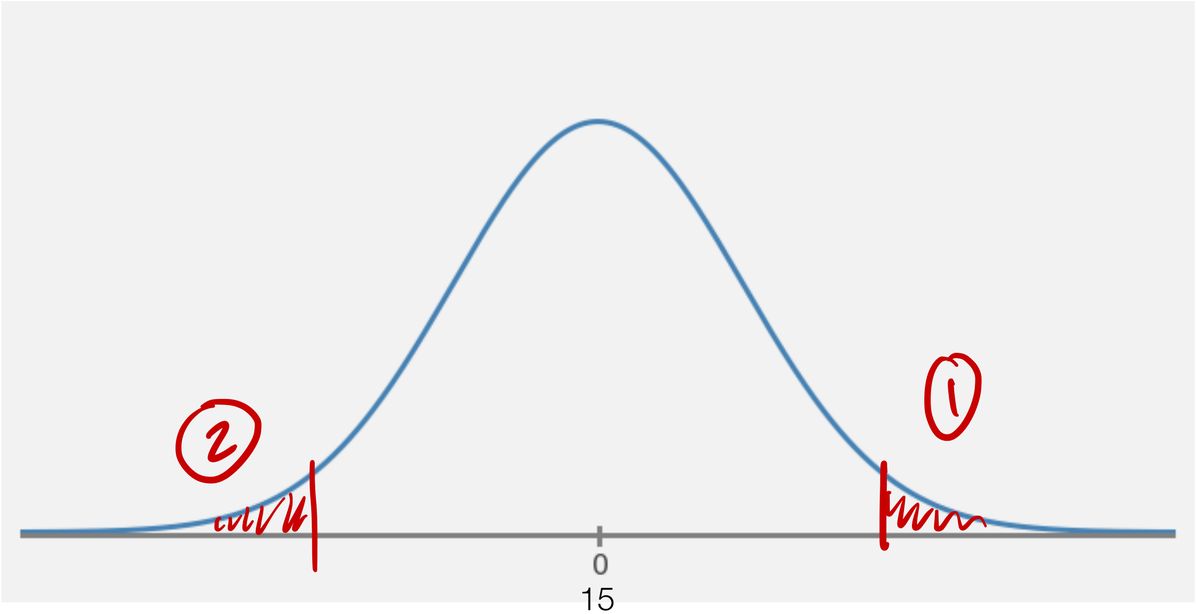
$= \Phi(-3.33)$ = stats.norm.cdf(-3.33) = 0.00043 ≤ 0.01

Reject H_0!



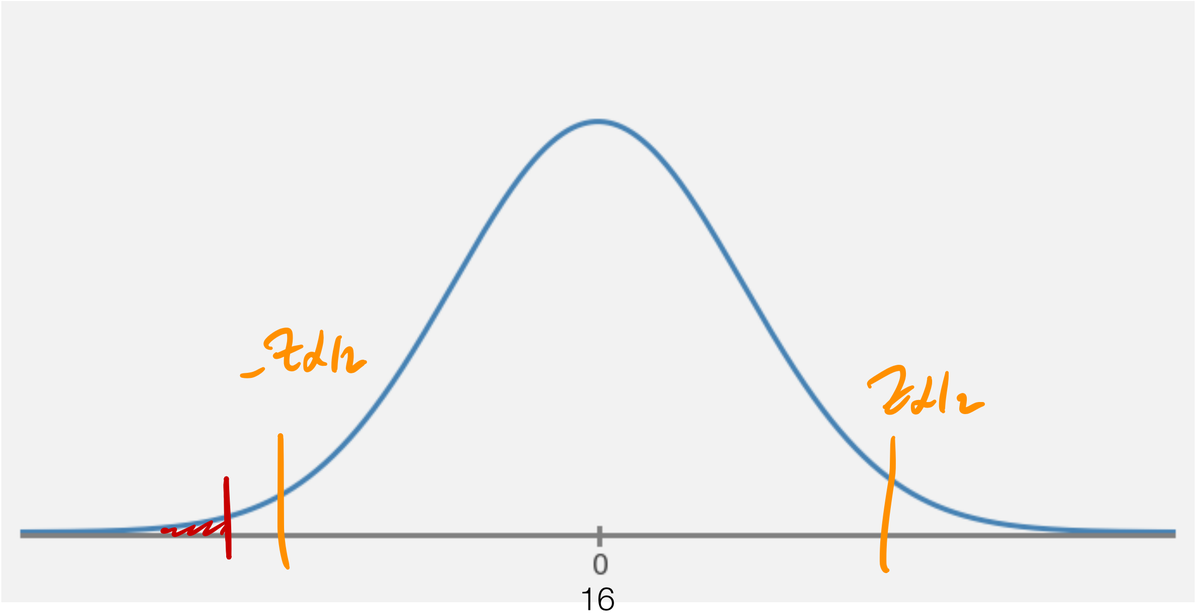
p-values for different z-tests

Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
① $H_1 : \theta > \theta_0$	$z \geq z_\alpha$	$1 - \Phi(z) \leq \alpha$
② $H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	$\Phi(z) \leq \alpha$
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	Next Slide.



p-values for different z-tests

Alternative Hypothesis	Critical Region Level α Test	p-value Level α Test
$H_1 : \theta > \theta_0$	$z \geq z_\alpha$	
$H_1 : \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1 : \theta \neq \theta_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2\Phi(-\lvert z \rvert) \leq \alpha$



Is the Belgian 1 Euro biased?

- Example: To test if the Belgian 1 Euro coin is fair you flip it 100 times and observe 38 Heads. Perform a p-value Z-test at the .05 significance level.

$\alpha = 0.05$

$\hat{p} = \frac{38}{100} = 0.38$

$\sigma = \sqrt{\frac{p(1-p)}{n}}$

$H_0: p = 0.5$

two-sided.

$H_1: p \neq 0.5$ biased!

p_0

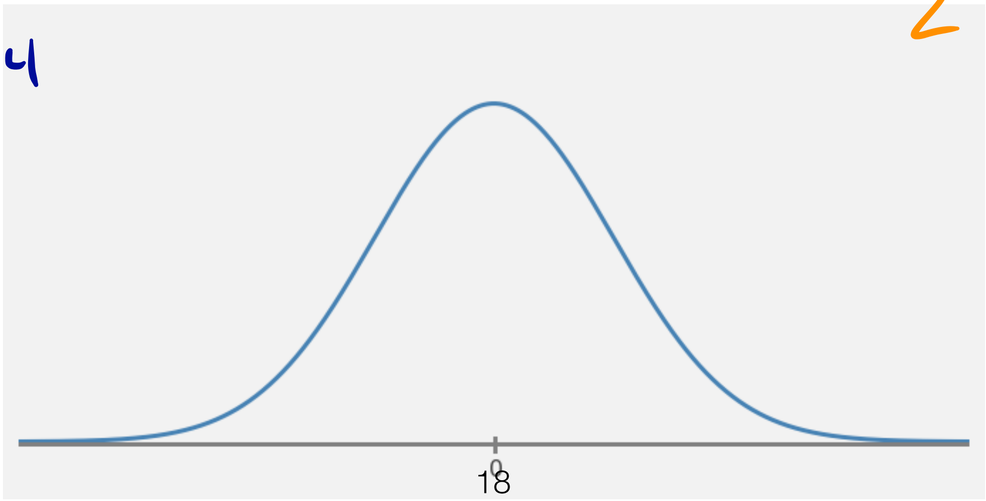
$$z = \frac{0.38 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -2.4$$

$\hat{p}$

σ_0

check pval: $2 \times \Phi(-|z|)$

$2 \times \Phi(-2.4) = 0.0164 < 0.05$



Reject H_0!

Coin is Biased!

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.
- **Question**: What kinds of Null and alternative hypotheses might we want to test?

some constant

$H_0: \mu_1 - \mu_2 = C$

$H_1: \mu_1 - \mu_2 \neq C$

$H_1: \mu_1 - \mu_2 < C$

$H_1: \mu_1 - \mu_2 > C$

CLT

$$\frac{(\mu_1 - \mu_2) - C}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

New Ad is better than old Ad by C clicks per day → for instance.

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:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - C}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

- We can then compute an appropriate p-value in the usual way!

Yay!

Two-Sample Testing for Difference of Means

- **Example**: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
pop 2 No	663	2258	1519
pop 1 Yes	413	2637	1138

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

$H_0: \mu_1 - \mu_2 = 200$

$H_1: \mu_1 - \mu_2 > 200$

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

$p\text{ value} = 1 - \Phi(2.20) \;?\; 0.05$

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?

We found: p-value = 1 - $\Phi(2.20)$ = 1-stats.norm.cdf(2.20) = 0.014 < 0.05 → **reject H_0**

Concept check: What about at the 0.01 (1%) significance level?

Common p-value misunderstandings

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

The p-value is the probability of observing your data (or more extreme), IF H_0 was true.

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

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

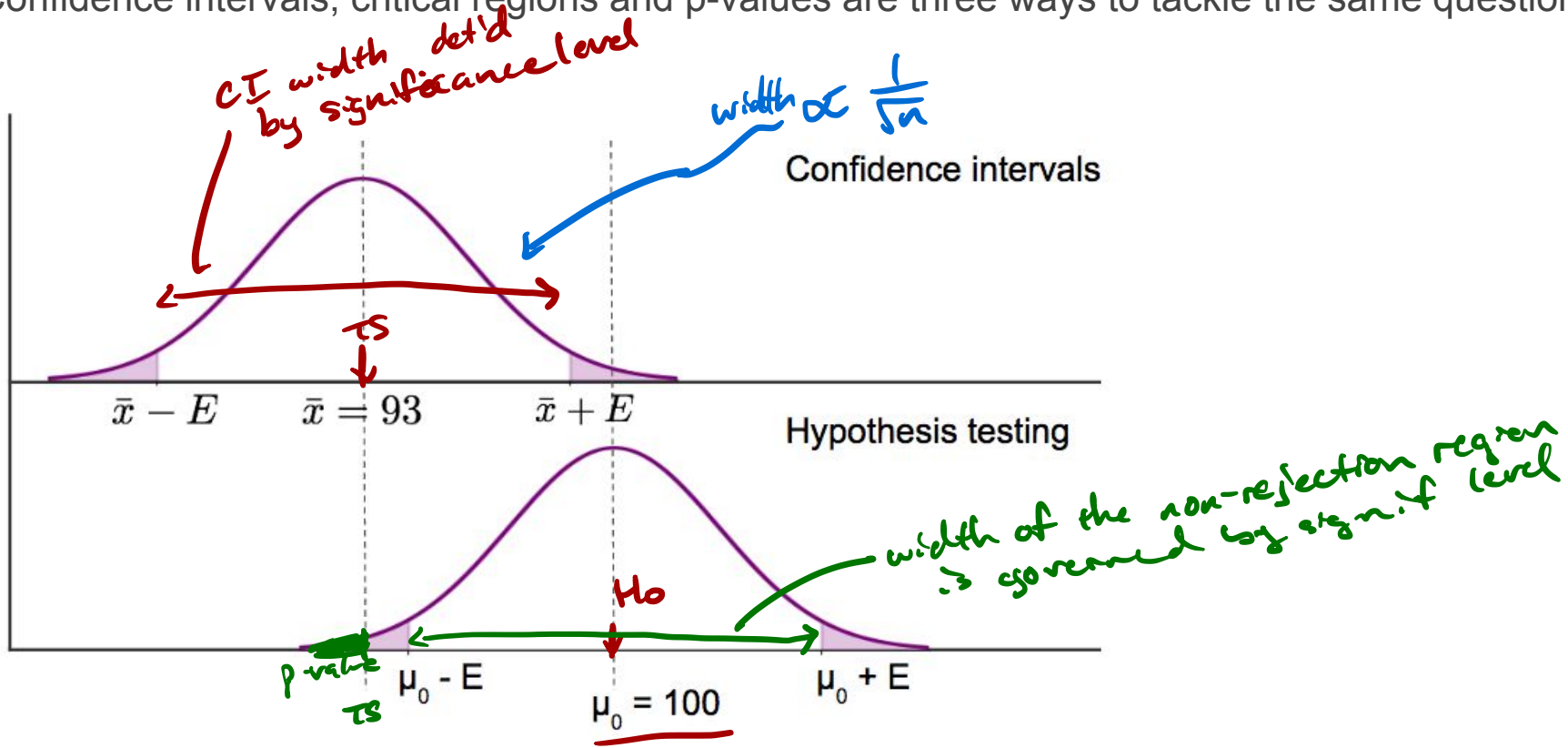
$\bar{X}$ of size 10,000 & $s = 0.0000001$ $\bar{X} = 10.001$

$\bar{X}$ signif. diff from 10 but not substantially

$$\text{CI width} = 2 \times z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

CIs vs Critical Regions vs p-Values

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



ERRORS, AND NOT-ERRORS!

	H_0 true	H_0 false
Reject H_0	Type I error (False positive)	✓ (correct action)
Fail to reject H_0	✓	Type II error (False negative)

What just happened?

- **Hypothesis testing** happened!
  - A way to formally ask questions like:

    $\mu_A \neq \mu_B$ or $\mu_A < \mu_B$

- **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 positives and negatives, respectively)