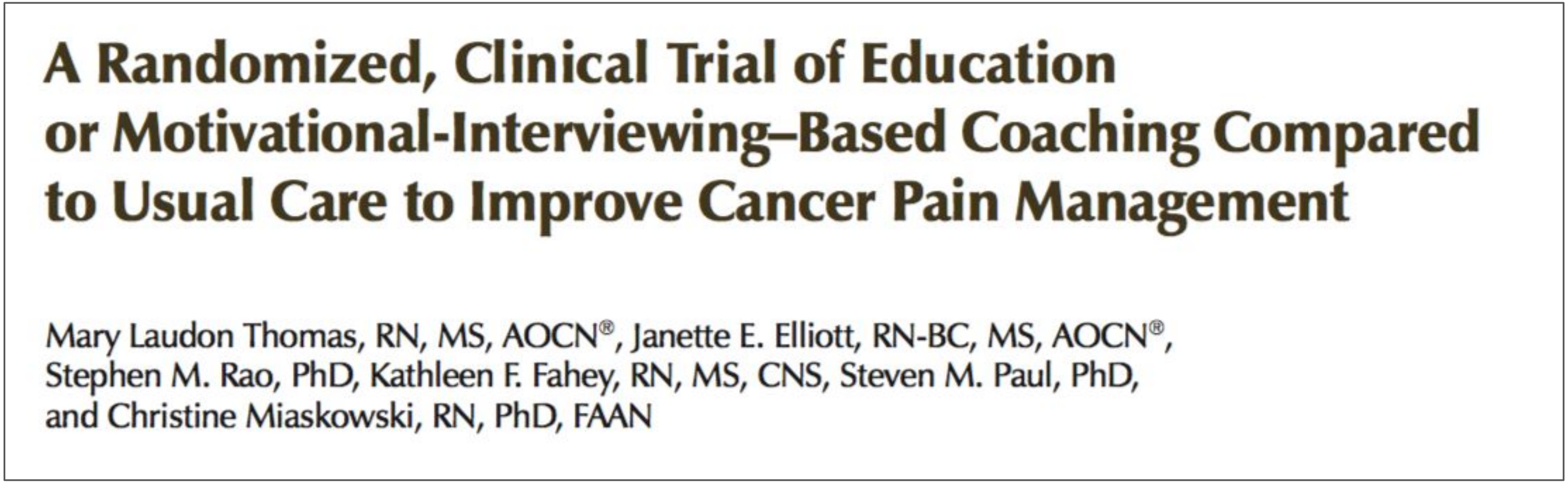
Lecture 17: Hypothesis Testing with p-Values

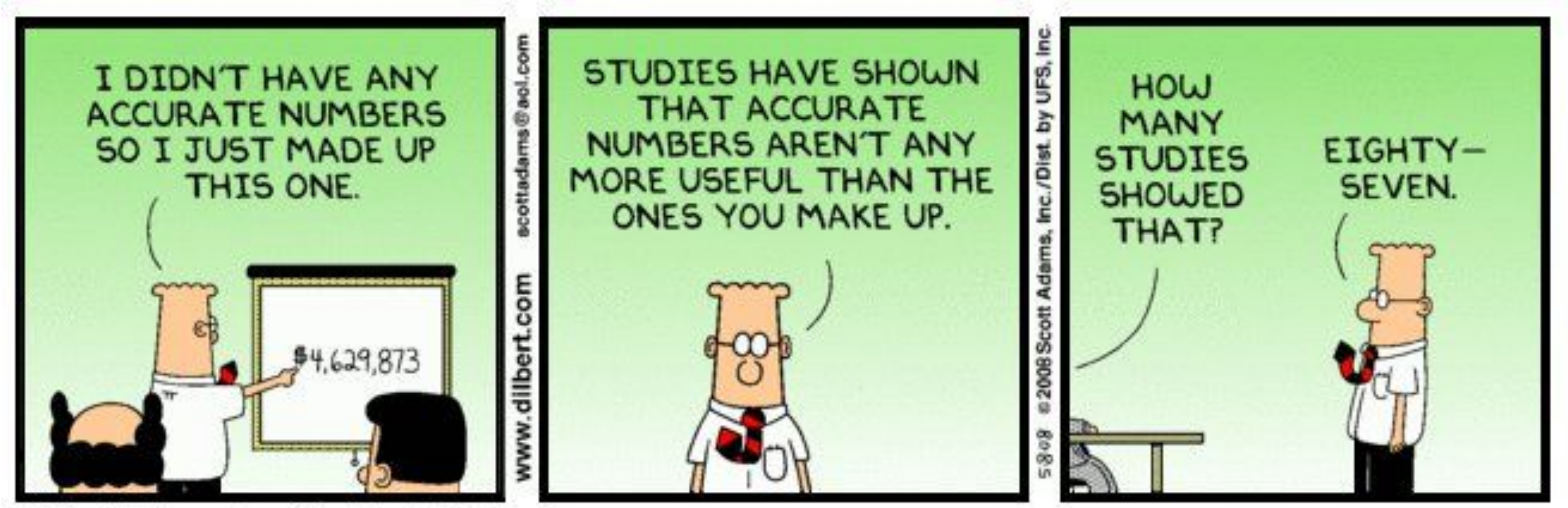


A Randomized, Clinical Trial of Education or Motivational-Interviewing–Based Coaching Compared to Usual Care to Improve Cancer Pain Management

Mary Laudon Thomas, RN, MS, AOCN®, Janette E. Elliott, RN-BC, MS, AOCN®, Stephen M. Rao, PhD, Kathleen F. Fahey, RN, MS, CNS, Steven M. Paul, PhD, and Christine Miaskowski, RN, PhD, FAAN

Announcements and reminders

- HW 3 due **today at 5 PM**
- HW 4 to be posted this afternoon. Due Friday April 5 at 5 PM


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com
scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08
© 2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

© Scott Adams, Inc./Dist. by UFS, Inc.

Previously, on CSCI 3022…

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

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

Definition: 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.

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

We've looked at ways to compute confidence intervals for several different statistics.

For example: a 100·(1-α)% CI for the mean μ when the value of σ is known is given by

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

Critical region HT refresher

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.

μ = mean life exp. of Jetta in Mexico

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?

$H_0: \mu = 300$

$H_1: \mu < 300$

Test stat:

$$\frac{\bar{z}}{\alpha/\sqrt{n}} = \frac{250 - 300}{150/\sqrt{100}}$$

$$= \frac{-50}{15} \Rightarrow z_{TS} = -3.33$$

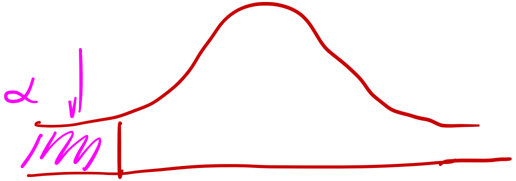
Critical region HT refresher

Example (cont.) 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.

$z_{TS} = -3.33$

$H_1: \mu < 300$ so evidence in favor of H_1 is lower z value

TS $z = -3.33 < z_{crt} = -2.33$

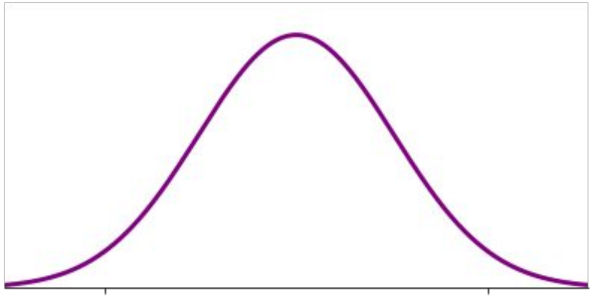


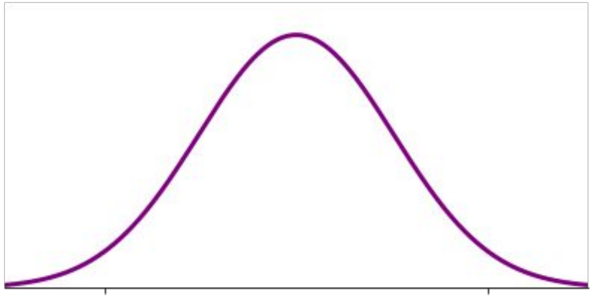
$z_{crt} = \text{ppf}(.01) = -2.33$

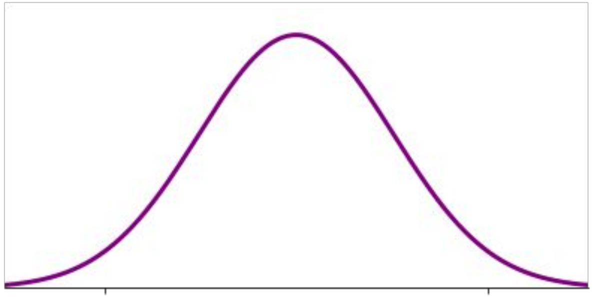
so Reject H_0 & conclude that at the 1% signf. level, there is evidence that the Mexico Jetta mean is < 300k miles

Critical region HT summary

	Alternative hypothesis	Rejection region for level α test
one tailed	$H_1: \theta > \theta_0$	$z \geq z_\alpha$
	$H_1: \theta < \theta_0$	$z \leq -z_\alpha$
two tailed test	$H_1: \theta \neq \theta_0$	$(z \geq z_{\alpha/2})$ or $(z \leq -z_{\alpha/2})$







Critical region HT summary

- Critical region is region where statistic has low probability under Null Hypothesis
- Requires normally distributed data, or large enough sample to use 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

Another way to view the critical region hypothesis test is through a **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: $\mu = \mu_0$
H_1: $\mu < \mu_0$

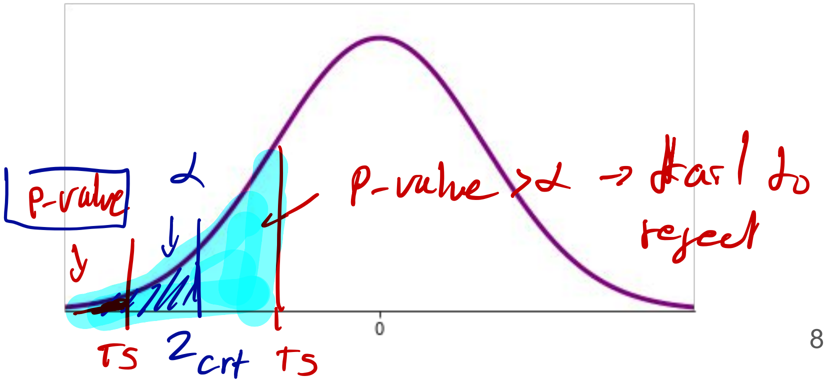
Example:
$H_0 : \mu = 300$
$H_1 : \mu < 300$

$P(TS \leq z_{CTH} \mid H_0 \text{ true}) = \alpha$

$P(z \leq TS \mid H_0 \text{ true}) = p\text{-value}$

The critical region to test is:

Reject H_0 if $\begin{cases} TS \leq z_{CTH} \\ p\text{-value} \leq \alpha \end{cases}$



Introduction to p-values

$p = 0.4$ data $= [H, H, H] \rightarrow$ high p-value

Definition: 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**

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

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…

P-value $= P(\text{data} \mid H_0 \text{ true})$

- … calculated under the assumption that the Null Hypothesis is true
- … always a value between 0 and 1

sanity check

- … **NOT** the probability that the Null Hypothesis is true!

$\hookrightarrow P(H_0 \text{ true} \mid \text{data}) \neq$ P-value $= P(\text{data} \mid H_0 \text{ true})$

Introduction to p-values

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

Concept check: How do we select α? → Based tolerance for type 1 error

Then the decision rule is:

- If p-value ≤ α, then **reject** the Null Hypothesis
- If p-value > α, then **fail to reject** the Null

Thus, if the p-value exceeds the selected significance level, then we cannot reject the Null

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

$\Phi(-3.33) = 0.00043$

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 than those made in Germany? Carry out the p-value test at the 0.01 signif. level.

$TS = -3.33$

$P(\text{data or more extreme} \mid H_0 \text{ true})$

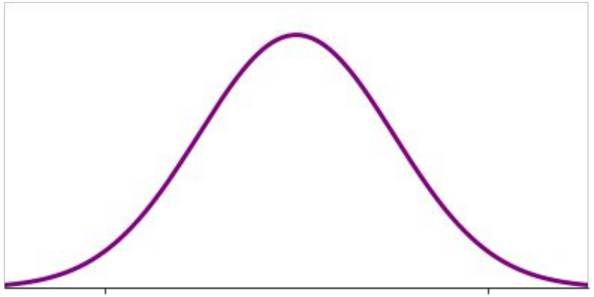
$\text{p-value} = P(z \leq -3.33)$

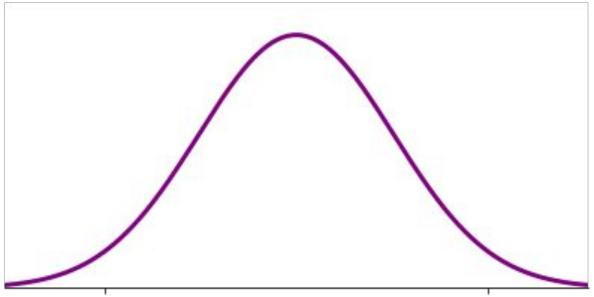
$= \Phi(-3.33) = 0.00043 < 0.01 = \alpha$

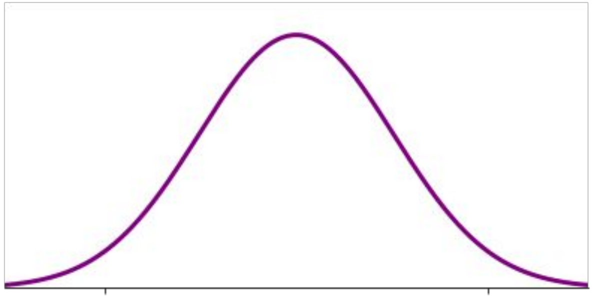
$\therefore$ Reject α.

p-values for different Z-tests

Alternative hypothesis	**Rejection region for 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: $\theta \neq \theta_0$	$(z \geq z_{\alpha/2})$ or $(z \leq z_{\alpha/2})$	

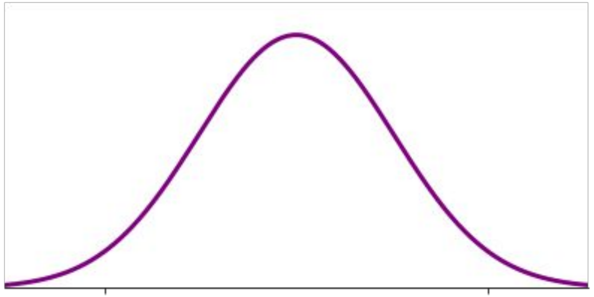


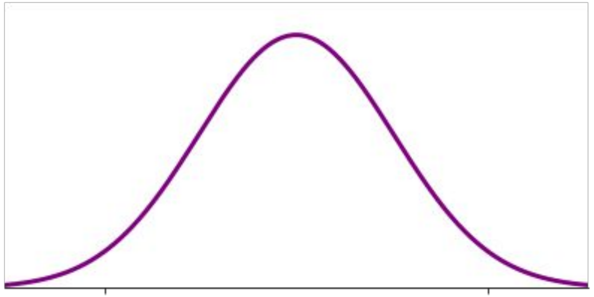


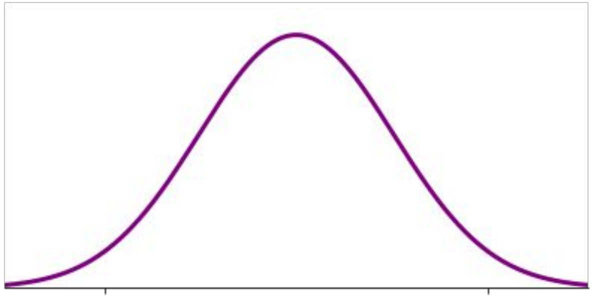


p-values for different Z-tests

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

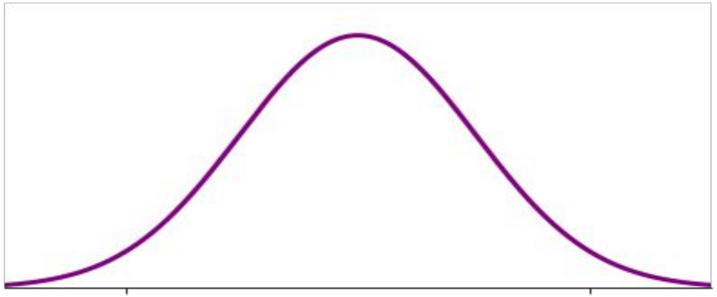






Is the Belgian 1 Euro biased?

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 0.95 significance level.



Two-sample testing for difference of means

S’pose 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?

Two-sample testing for difference of means

S’pose 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?

H_0: $\mu_1 - \mu_2 = c$

H_1: $\mu_1 - \mu_2 > c$

H_1: $\mu_1 - \mu_2 < c$

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

Two-sample testing for difference of means

S’pose 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:

Then, we can compute an appropriate p-value in the usual way (using Φ(z))

Two-sample testing for difference of means

S’pose 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:

$$\text{test statistic} \;=\; z \;=\; \frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \;\sim\; N(0,1)$$

Then, we can compute an appropriate p-value in the usual way (using $\Phi(z)$)

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?

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?

H_0: $\mu_1 - \mu_2 = 200$
H_1: $\mu_1 - \mu_2 > 200$

$$z = \frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} = \frac{(2637 - 2258) - 200}{\sqrt{\frac{1138^2}{413} + \frac{1519^2}{663}}} \approx 2.20$$

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

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 - Φ(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

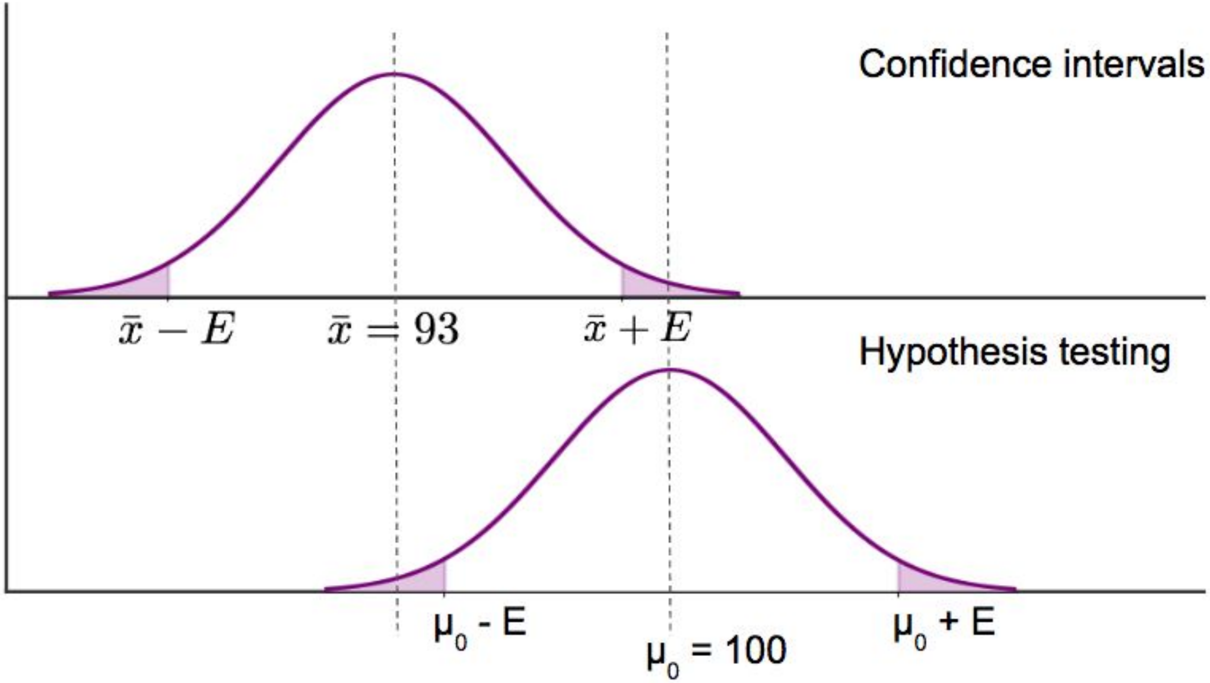
Misconception #1: If p = 0.05, the Null Hypothesis has a 5% chance of being true.

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

CIs vs Critical Regions vs p-Values

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



What just happened?

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

$$\mu_A \neq \mu_B \quad \text{or} \quad \mu_A < \mu_B$$

- **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)

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.