CAAP Statistics - Lec12

Jul 25, 2022

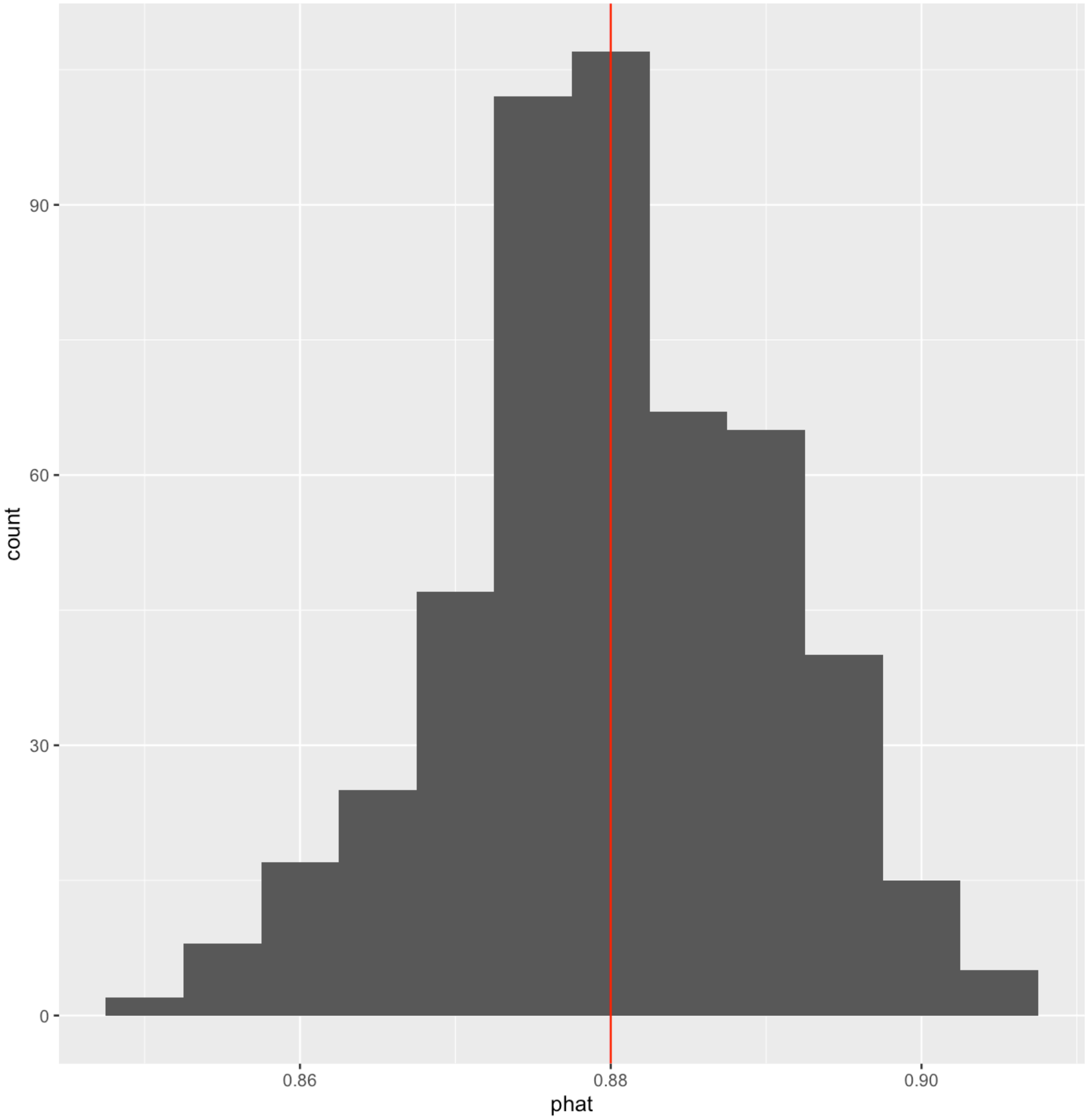
Review: Point Estimate and Sampling Variability

- Point estimates and sampling variability
  - What is sampling distribution?
    - Point estimate is not fixed quantity, but random quantity associated with the variance!
  - Central Limit Theorem: sample mean(or sample proportion) follows normal distribution **as the sample size increases**
    - Recall Law of Large Numbers : sample mean approximates the population mean **as the sample size increases**

Sampling distribution

What is the shape and center of this distribution?

The distribution looks symmetric and somewhat bell-shaped.



Learning Objectives

- Point estimates and sampling variability
  - What is sampling distribution
  - Central Limit Theorem
- Confidence intervals for a proportion
  - Interpreting the confidence interval
- Hypothesis testing for a proportion
  - Null hypothesis vs. Alternative hypothesis
  - Decision Error(Type I error, Type II error)

Confidence Intervals for a Proportion

Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462)
and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Facebook’s categorization of user interests

Most commercial websites (e.g. social media platforms, news out- lets, online retailers) collect a data about their users’ behaviors and use these data to deliver targeted content, recommendations, and ads. To understand whether Americans think their lives line up with how the algorithm-driven classification systems categorizes them, Pew Research asked a representative sample of 850 American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests. 67% of the respondents said that the listed categories were accurate. Estimate the true proportion of American Facebook users who think the Facebook categorizes their interests accurately.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook’s categorization of user interests

$$\hat{p} = 0.67 \qquad n = 850$$

The approximate 95% confidence interval is defined as

$$point\ estimate \pm 1.96 \times SE$$

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.67 \times 0.33}{850}} \approx 0.016$$

$$\begin{aligned} \hat{p} \pm 1.96 \times SE &= 0.67 \pm 1.96 \times 0.016 \\ &= (0.67 - 0.03, 0.67 + 0.03) \\ &= (0.64, 0.70) \end{aligned}$$

Facebook’s categorization of user interests

Which of the following is the correct interpretation of this confidence interval? We are 95% confident that...

(a) 64% to 70% of American Facebook users in this sample think Facebook categorizes their interests accurately.

(b) 64% to 70% of all American Facebook users think Facebook categorizes their interests accurately

(c) there is a 64% to 70% chance that a randomly chosen American Facebook user’s interests are categorized accurately.

(d) there is a 64% to 70% chance that 95% of American Facebook users’ interests are categorized accurately.

Facebook's categorization of user interests

Which of the following is the correct interpretation of this confidence interval? We are 95% confident that...

(a) 64% to 70% of American Facebook users in this sample think Facebook categorizes their interests accurately.

(b) 64% to 70% of all American Facebook users think Facebook categorizes their interests accurately

(c) there is a 64% to 70% chance that a randomly chosen American Facebook user's interests are categorized accurately.

(d) there is a 64% to 70% chance that 95% of American Facebook users' interests are categorized accurately.

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation

point estimate ± 1.96 × SE

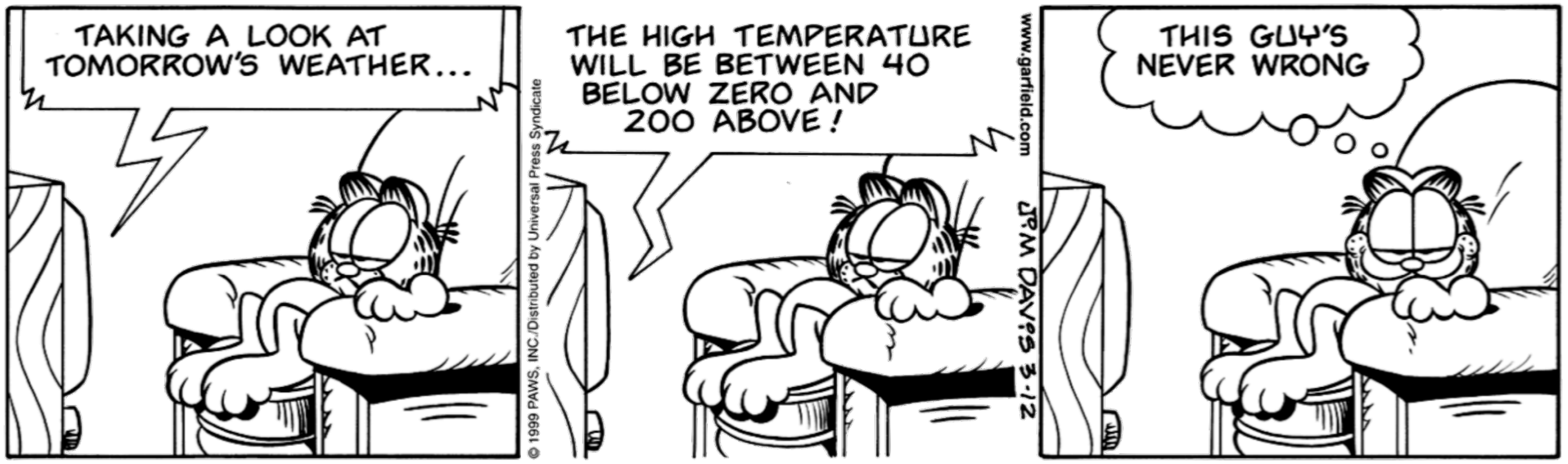
Then about 95% of those intervals would contain the true population proportion (p).

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVPS 3-12

If the interval is too wide it may not be very informative.

Changing the confidence level

point estimate ± $z^\star$ × SE

- In a confidence interval, $z^\star$ × SE is called the **margin of error**, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = 1.00

(e) Z = 1.65

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

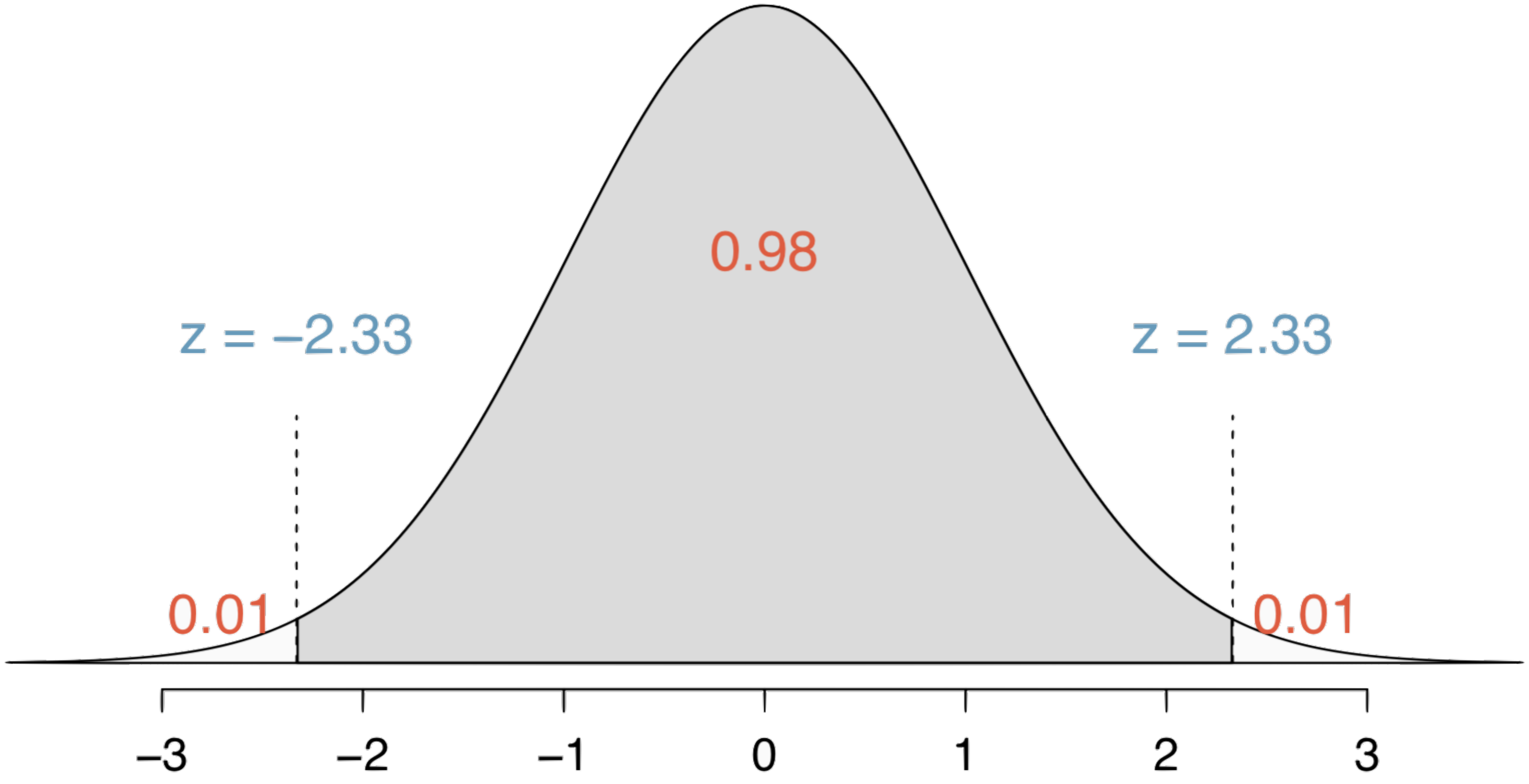
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = 1.00

(e) Z = 1.65



Interpreting confidence intervals

Confidence intervals are ...

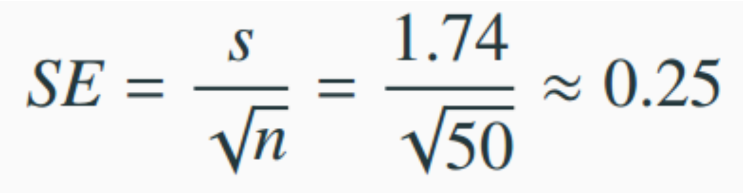
- always about the population
- are not probability statements
- only about population parameters, not individual observations
- only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$\bar{x} = 3.2$ $s = 1.74$

The approximate 95% confidence interval is defined as point estimate ± 2 x SE



$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$\Rightarrow$ $3.2 \pm 2 \times 0.25$

$\Rightarrow$ $(3.2 - 0.5, 3.2 + 0.5)$

$\Rightarrow$ $(2.7, 3.7)$

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Practice

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

A more accurate interval

Confidence interval, a general formula

point estimate ± z x SE*

A more accurate interval

Confidence interval, a general formula

point estimate ± z x SE*

Conditions when the point estimate = $\bar{x}$

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

A more accurate interval

Confidence interval, a general formula

point estimate ± z x SE*

Conditions when the point estimate = $\bar{x}$

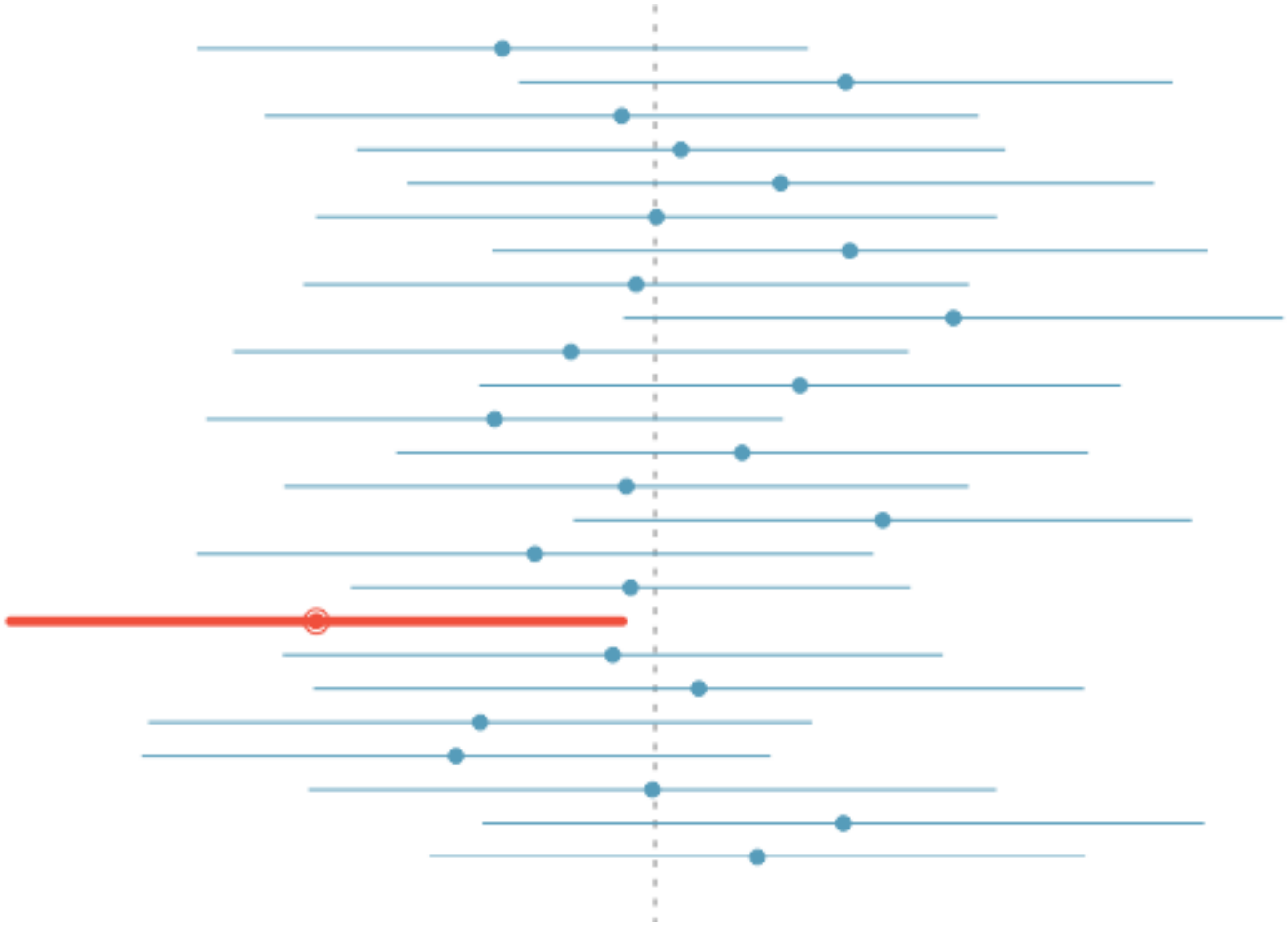
1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ in the next chapter.

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* ± *2 x SE*.
- Then about 95% of those intervals would contain the true population mean (μ).

- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

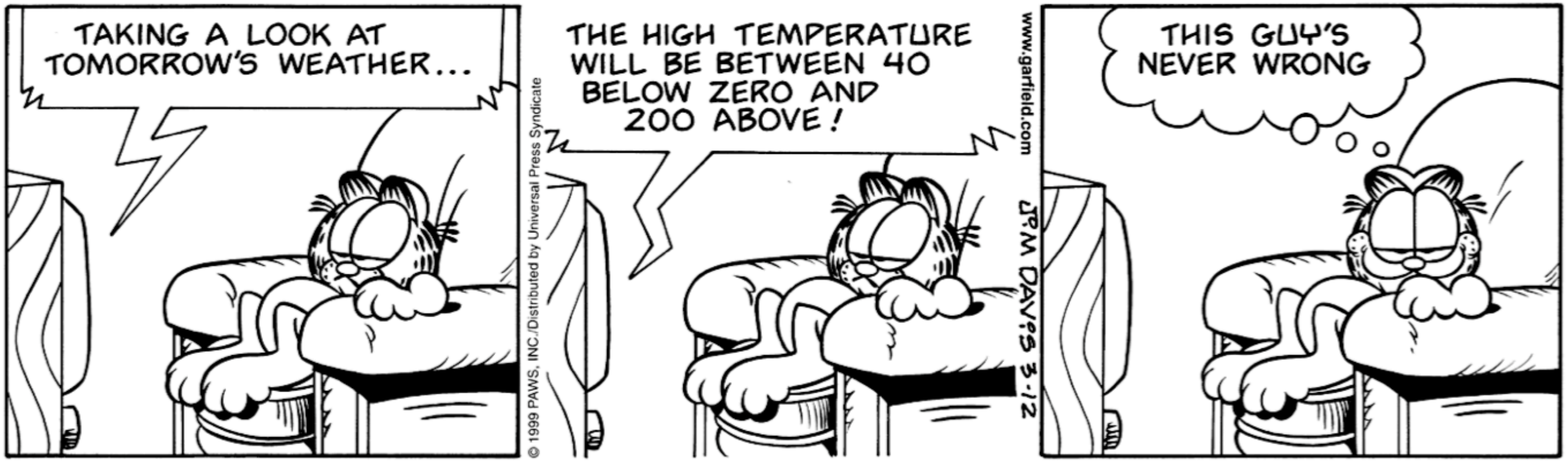

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

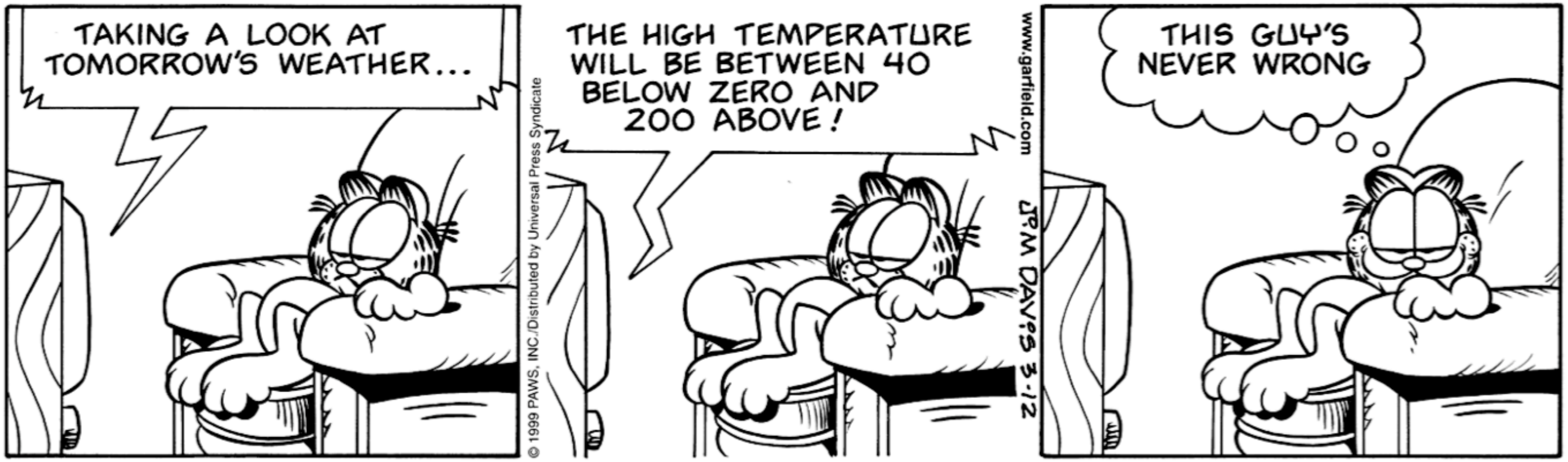
Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

If the interval is too wide it may not be very informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

point estimate ± z x SE*

- In a confidence interval, *z* x SE* is called the *margin of error*, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust *z** in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, *z** = 1.96.
- However, using the standard normal (*z*) distribution, it is possible to find the appropriate *z** for any confidence level.

Practice

Which of the below Z scores is the appropriate z^* when calculating a 98% confidence interval?

(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$

Practice

Which of the below *Z* scores is the appropriate *z** when calculating a 98% confidence interval?

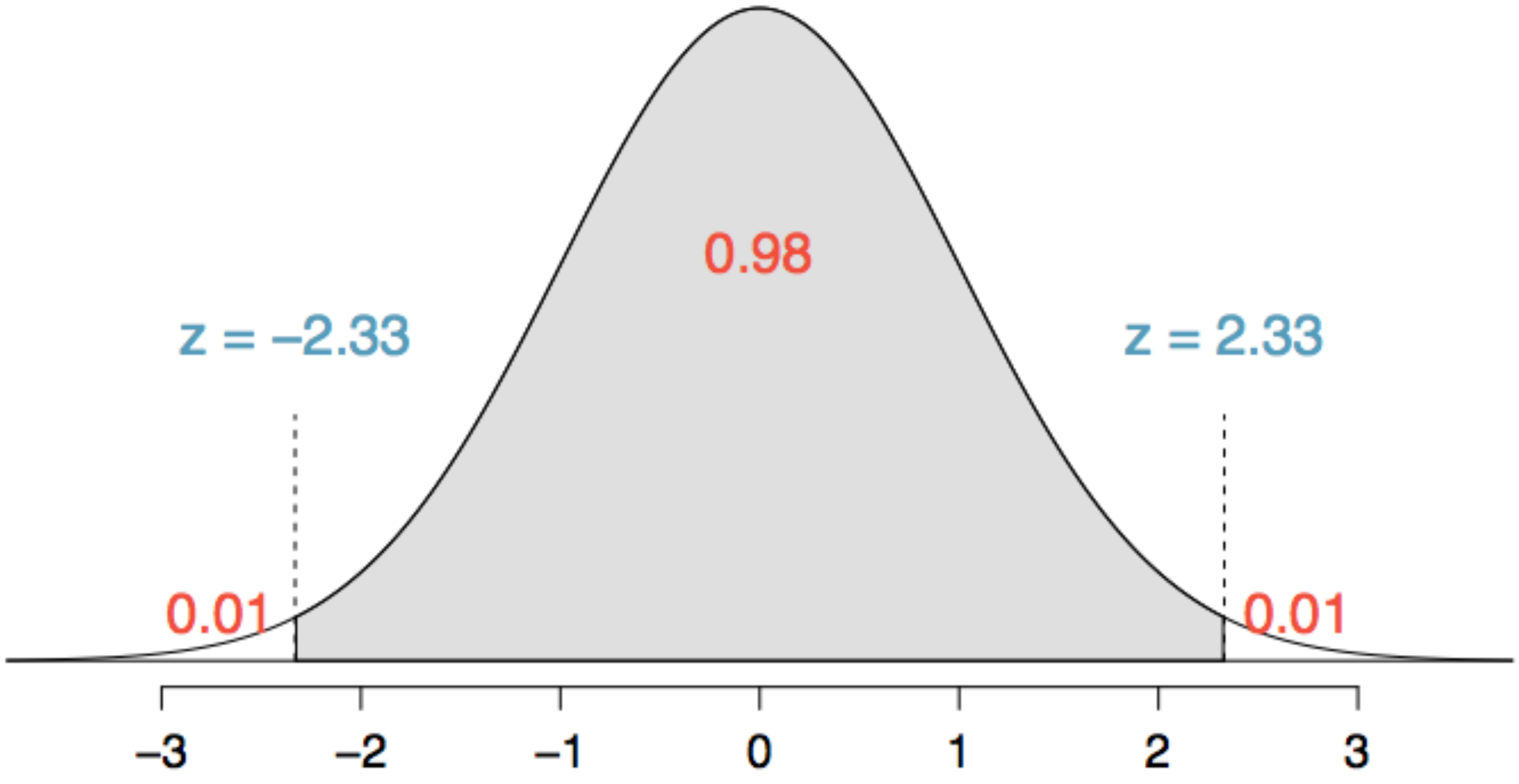
(a) *Z* = 2.05

(b) *Z* = 1.96

(c) *Z* = 2.33

(d) *Z* = -2.33

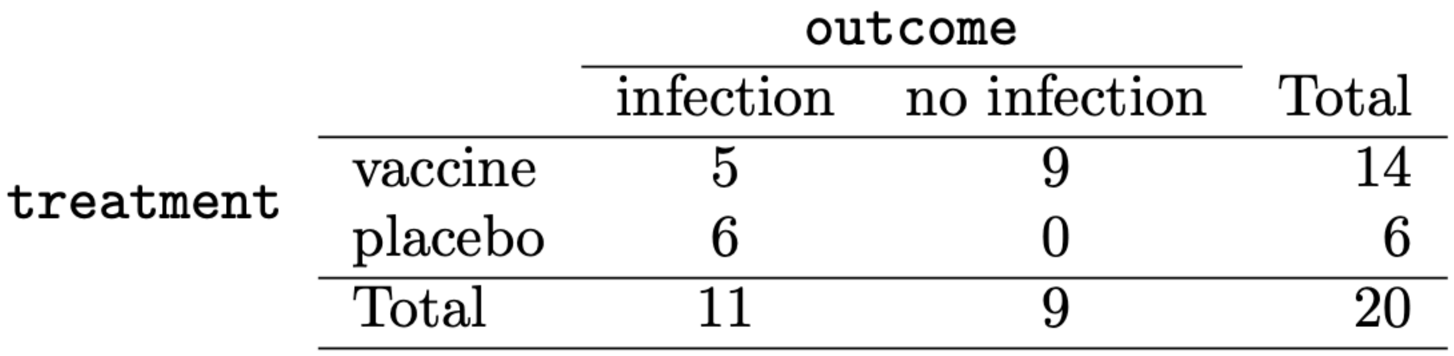
(e) *Z* = -1.65



Hypothesis Testing for a Proportion

Remember when...

Malaria Vaccine Experiment



		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

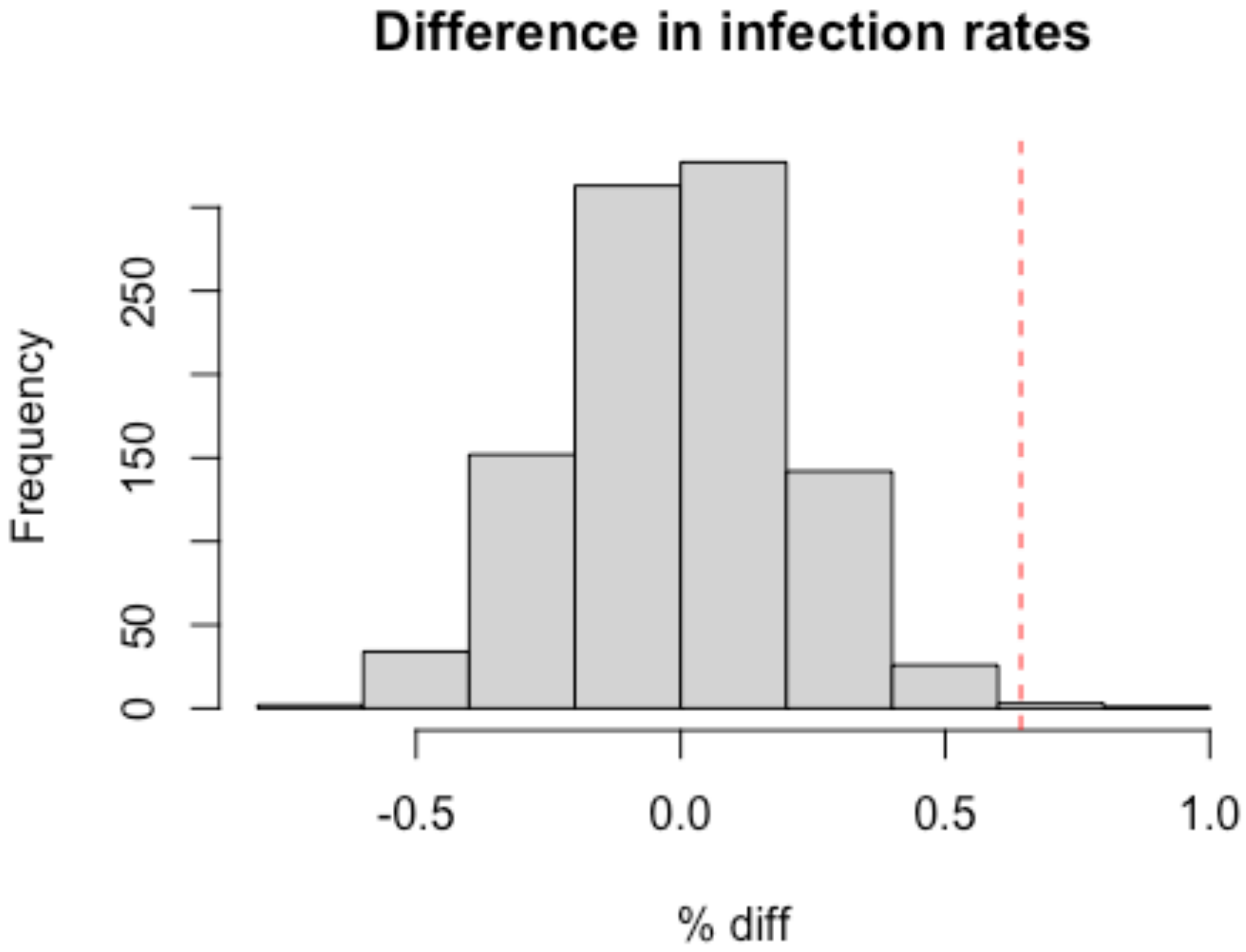
$$\hat{p}_{\text{trt}} = 5 / 14 = 0.357$$

$$\hat{p}_{\text{control}} = 6 / 6 = 1.00$$

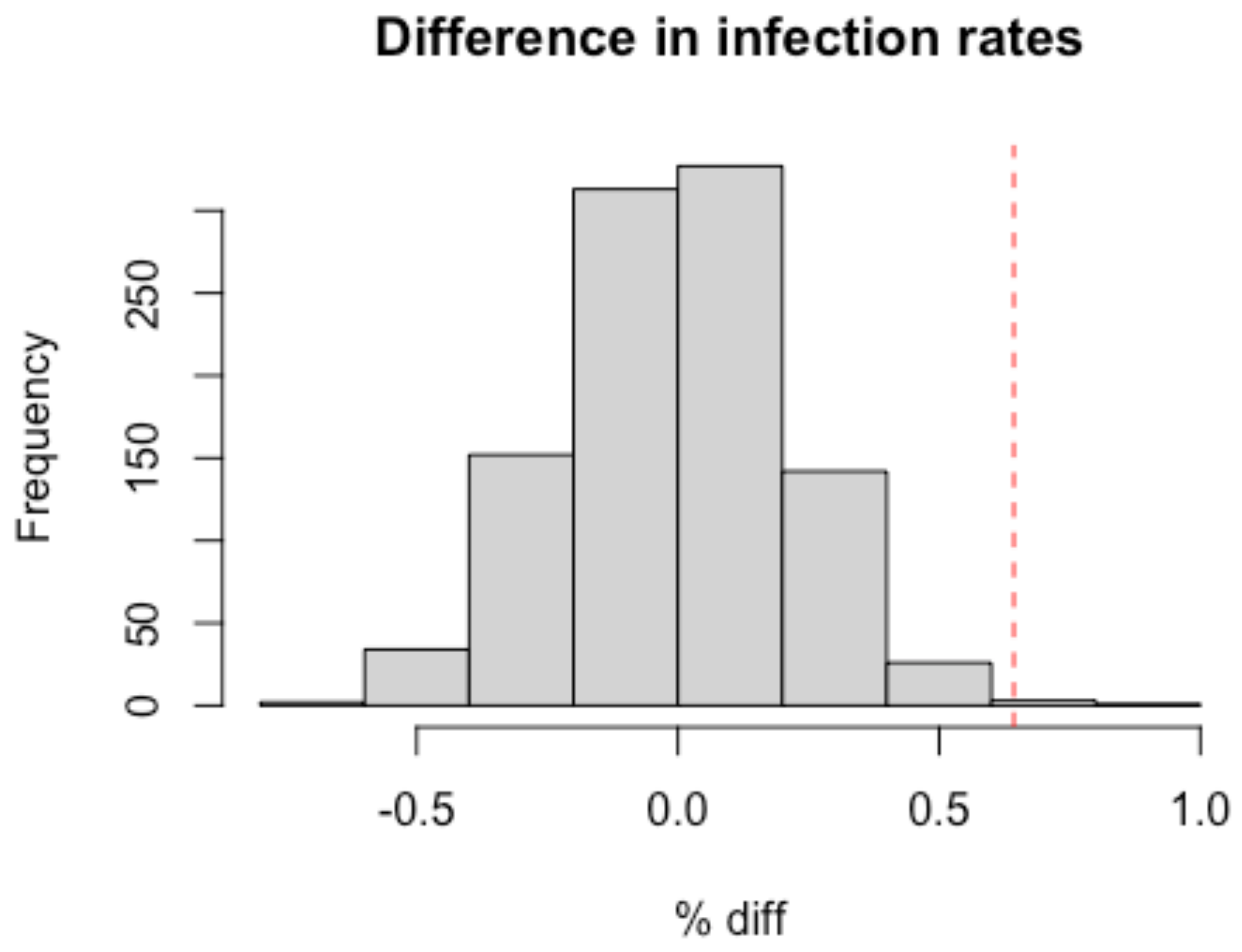
Possible explanations:

- Vaccine and Infection rate are *independent*, no effectiveness in vaccine, observed difference in proportions is simply due to chance.
  → null (nothing is going on)
- Vaccine and Infection rate are *dependent*, the vaccine is actually effective, observed difference in proportions is not due to chance.
  → alternative (something is going on)

Result


Difference in infection rates
Frequency
0
50
150
250
-0.5
0.0
0.5
1.0
% diff

Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (control group's infection rate is higher than the vaccinated group by 64.3%p), we decided to reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

- We start with a *null hypothesis* *(H_0)* that represents the status quo.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the proporton of American Facebook users who think Facebook categorizes their interests accurately as 64% to 67%. Based on this confidence interval, do the data support the hypothesis that majority of American Facebook users think Facebook categorizes their interests accurately.

The associated hypotheses are:

H_0: $p = 0.50$: 50% of American Facebook users think Facebook categorizes their interests accurately

H_A: $p > 0.50$: More than 50% of American Facebook users think Facebook categorizes their interests accurately

Null value is not included in the interval → reject the null hypothesis.

This is a quick-and-dirty approach for hypothesis testing, but it doesn’t tell us the likelihood of certain outcomes under the null hypothesis (p-value)

Decision errors

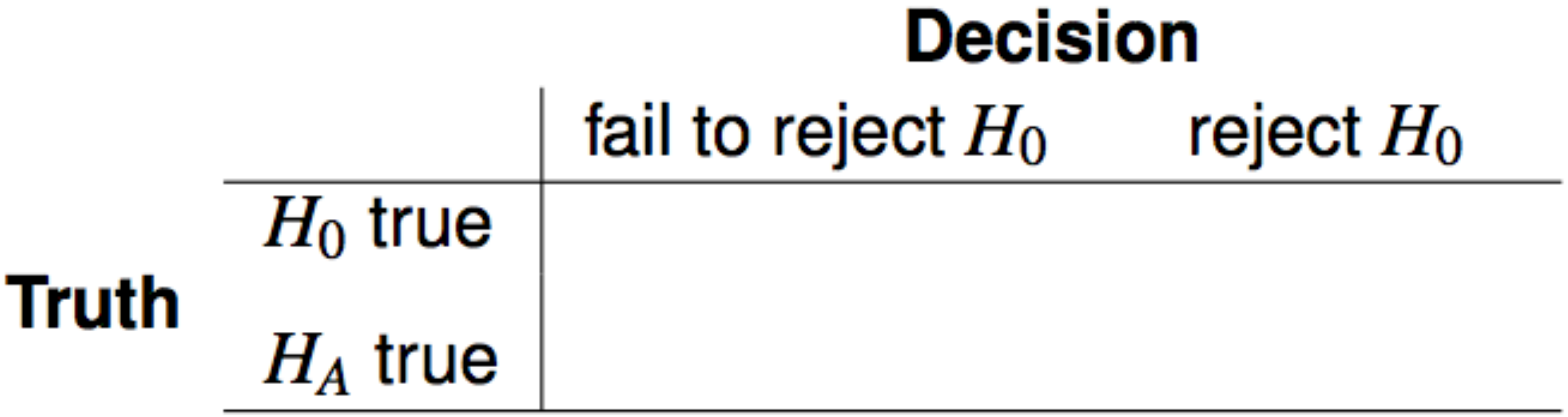
- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

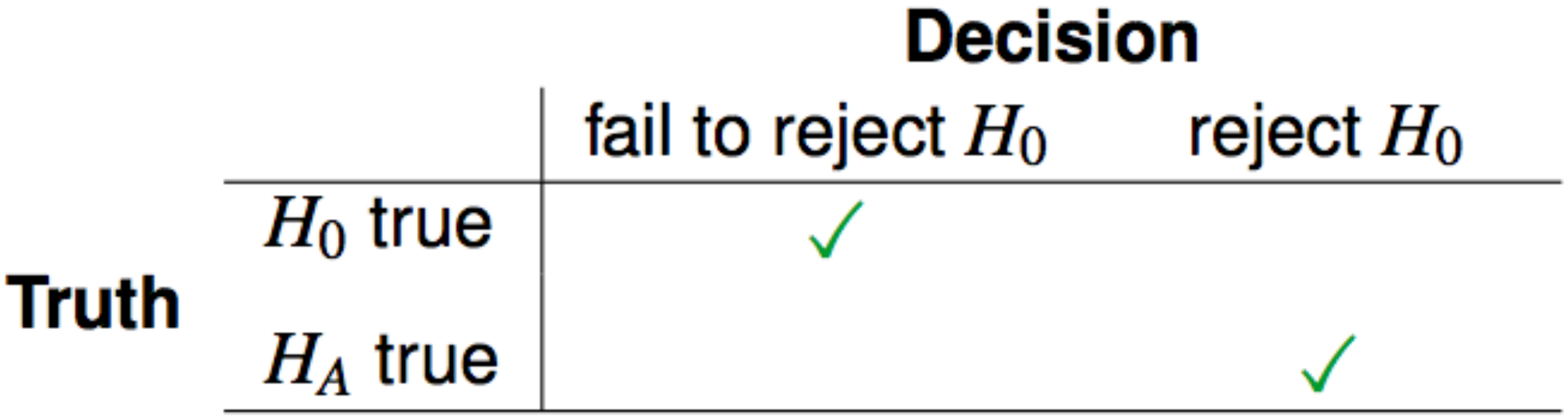
Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

Decision errors (cont.)

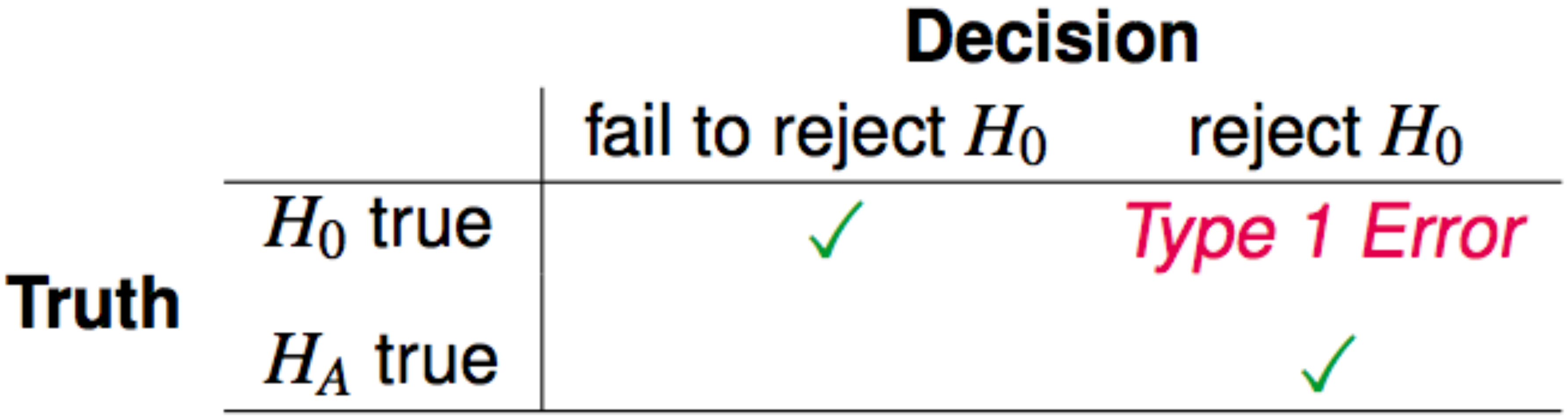
There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

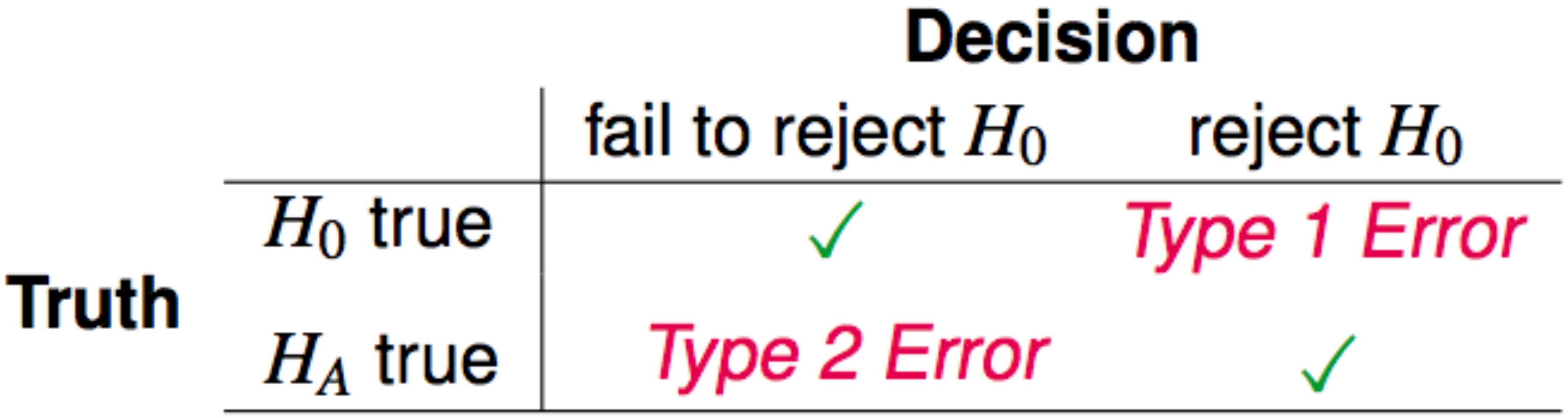


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

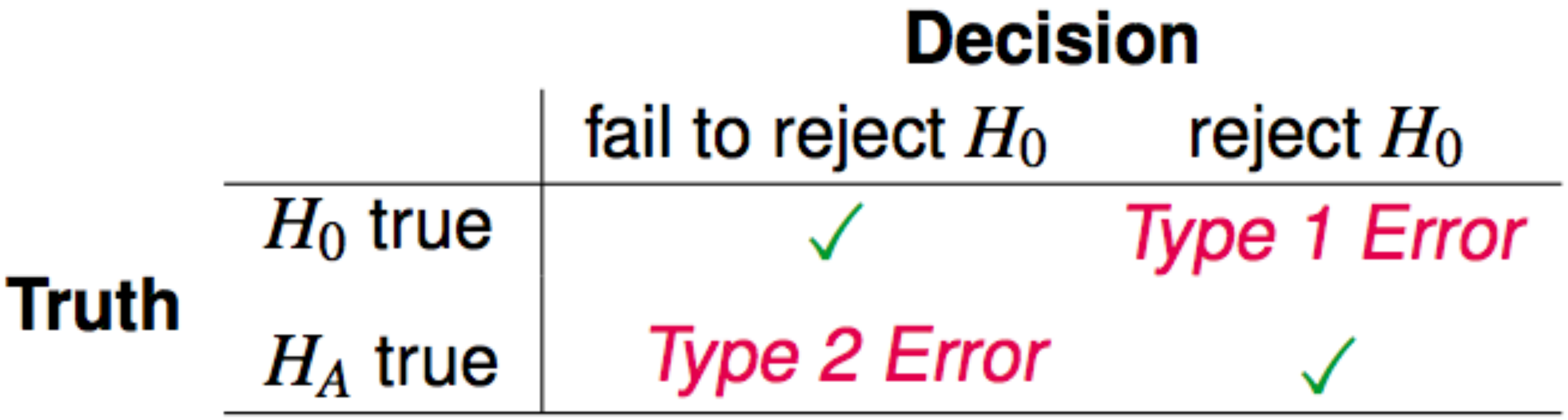


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

  Type 2 error

- Declaring the defendant guilty when they are actually innocent

  Type 1 error

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"
- William Blackstone

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

Setting the hypotheses

The *parameter of interest* is the proportion of all American Facebook users who are comfortable with Facebook creating categories of interests for them.

There may be two explanations why our sample proportion is lower than 0.50 (minority).

- The true population proportion is different than 0.50.
- The true population mean is 0.50, and the difference between the true population proportion and the sample proportion is simply due to natural sampling variability.

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

Setting the hypotheses

We start with the assumption that 50% of American Facebook users are comfortable with Facebook creating categories of interests for them

H_0: $p = 0.50$

We test the claim that the proportion of American Facebook users who are comfortable with Facebook creating categories of interests for them is different than 50%.

H_A: $p \neq 0.50$

Facebook interest categories - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Respondents in the sample should be independent of each other with respect to whether or not they feel comfortable with their interests being categorized by Facebook.

(b) Sampling should have been done randomly.

(c) The sample size should be less than 10% of the population of all American Facebook users.

(d) There should be at least 30 respondents in the sample.

(e) There should be at least 10 expected successes and 10 expected failure.

Facebook interest categories - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Respondents in the sample should be independent of each other with respect to whether or not they feel comfortable with their interests being categorized by Facebook.

(b) Sampling should have been done randomly.

(c) The sample size should be less than 10% of the population of all American Facebook users.

(d) There should be at least 30 respondents in the sample.

(e) There should be at least 10 expected successes and 10 expected failure.

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.

$$\hat{p} \sim N\left(\mu = 0.50, SE = \sqrt{\frac{0.50 \times 0.50}{850}}\right)$$

$$Z = \frac{0.41 - 0.50}{0.0171} = -5.26$$

The sample proportion is 5.26 standard errors away from the hypothesized value. Is this considered unusually low? That is, is the result *statistically significant*?

Yes, and we can quantify how unusual it is using a p-value.

p-values

We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject H_0*.

If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject H_0*.

Facebook interest categories - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample proportion lower than 0.41), if in fact H_0 were true (the true population proportion was 0.50).

$$P(\hat{p} < 0.41 \; or \; \hat{p} > 0.59 \mid p = 0.50) = P(|Z| > 5.26) < 0.0001$$

Facebook interest categories - Making a decision

p-value < 0.0001

- If 50% of all American Facebook users are comfortable with Facebook creating these interest categories, there is less than a 0.01% chance of observing a random sample of 850 American Facebook users where 41% or fewer or 59% of higher feel comfortable with it.
- This is a pretty low probability for us to think that the observed sample proportion, or something more extreme, is likely to happen simply by chance.

Since p-value is *low* (lower than 5%) we *reject H_0*.

The data provide convincing evidence that the proportion of American Facebook users who are comfortable with Facebook creating a list of interest categories for them is different than 50%.

The difference between the null value of 0.50 and observed sample proportion of 0.41 is *not due to chance* or sampling variability.

Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.

We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

One vs. two sided hypothesis tests

In two sided hypothesis tests we are interested in whether p is either above or below some null value p_0: $H_A : p \neq p_0$.

In one sided hypothesis tests we are interested in p differing from the null value p_0 in one direction (and not the other):

If there is only value in detecting if population parameter is less than p_0, then H_A: $p < p_0$.

If there is only value in detecting if population parameter is greater than p_0, then $H_A : p > p_0$.

Two-sided tests are often more appropriate as we often want to detect if the data goes clearly in the opposite direction of a hypothesis direction as well.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

- The associated hypotheses are:
  H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average
  H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

- The associated hypotheses are:
  H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average
  H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average
- Since the null value is included in the interval, we do not reject the null hypothesis in favor of the alternative.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

- The associated hypotheses are:
  $H_0: \mu = 3$: College students have been in 3 exclusive relationships, on average
  $H_A: \mu > 3$: College students have been in more than 3 exclusive relationships, on average
- Since the null value is included in the interval, we do not reject the null hypothesis in favor of the alternative.
- This is a quick-and-dirty approach for hypothesis testing. However it doesn't tell us the likelihood of certain outcomes under the null hypothesis, i.e. the p-value, based on which we can make a decision on the hypotheses.

Number of college applications

A similar survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by <u>all</u> Duke students.

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
- We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended)

$$H_0 : \boldsymbol{\mu} = 8$$

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
- We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended)

$$H_0 : \boldsymbol{\mu} = 8$$

- We test the claim that the average number of colleges Duke students apply to is greater than 8

$$H_A : \boldsymbol{\mu} > 8$$

Number of college applications - conditions

Which of the following is *not* a condition that needs to be met to proceed with this hypothesis test?

a) Students in the sample should be independent of each other with respect to how many colleges they applied to.
b) Sampling should have been done randomly.
c) The sample size should be less than 10% of the population of all Duke students.
d) There should be at least 10 successes and 10 failures in the sample.
e) The distribution of the number of colleges students apply to should not be extremely skewed.

Number of college applications - conditions

Which of the following is *not* a condition that needs to be met to proceed with this hypothesis test?

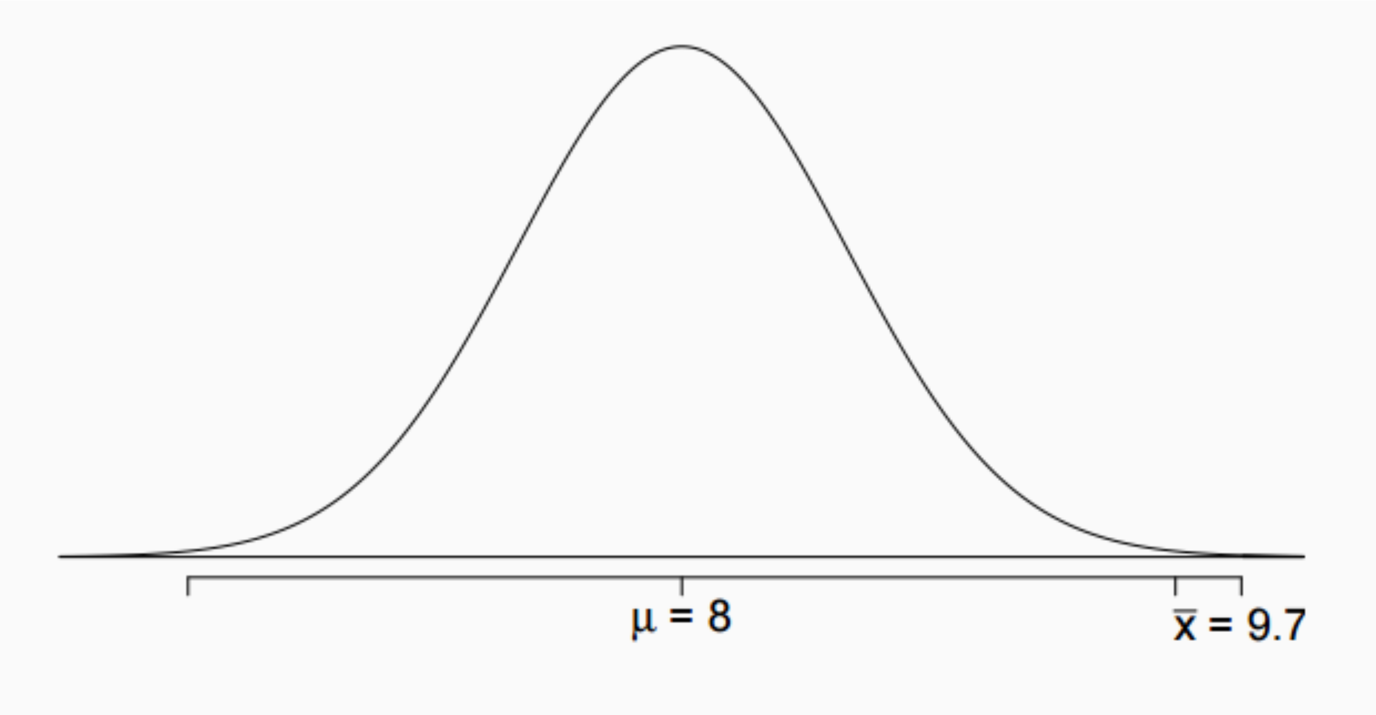
a) Students in the sample should be independent of each other with respect to how many colleges they applied to.
b) Sampling should have been done randomly.
c) The sample size should be less than 10% of the population of all Duke students.
d) There should be at least 10 successes and 10 failures in the sample.
e) The distribution of the number of colleges students apply to should not be extremely skewed.

Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.

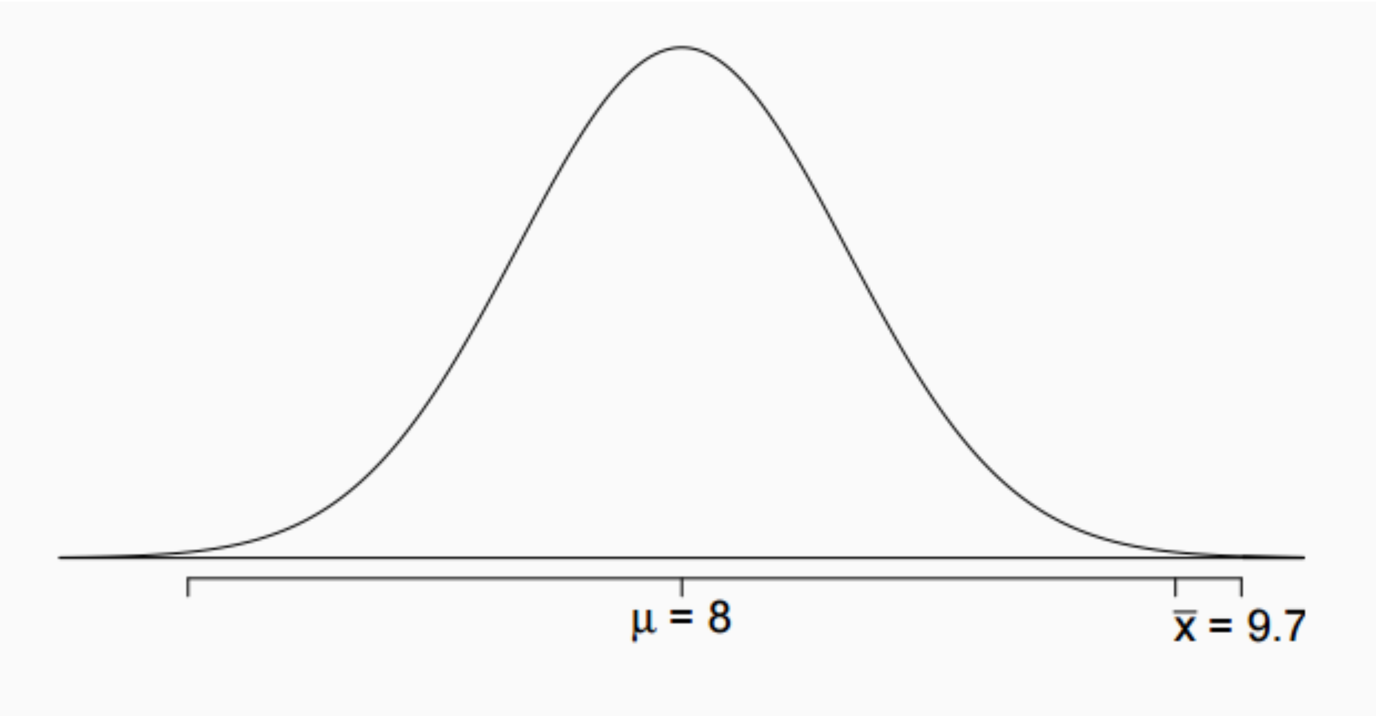
Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



Test Statistic

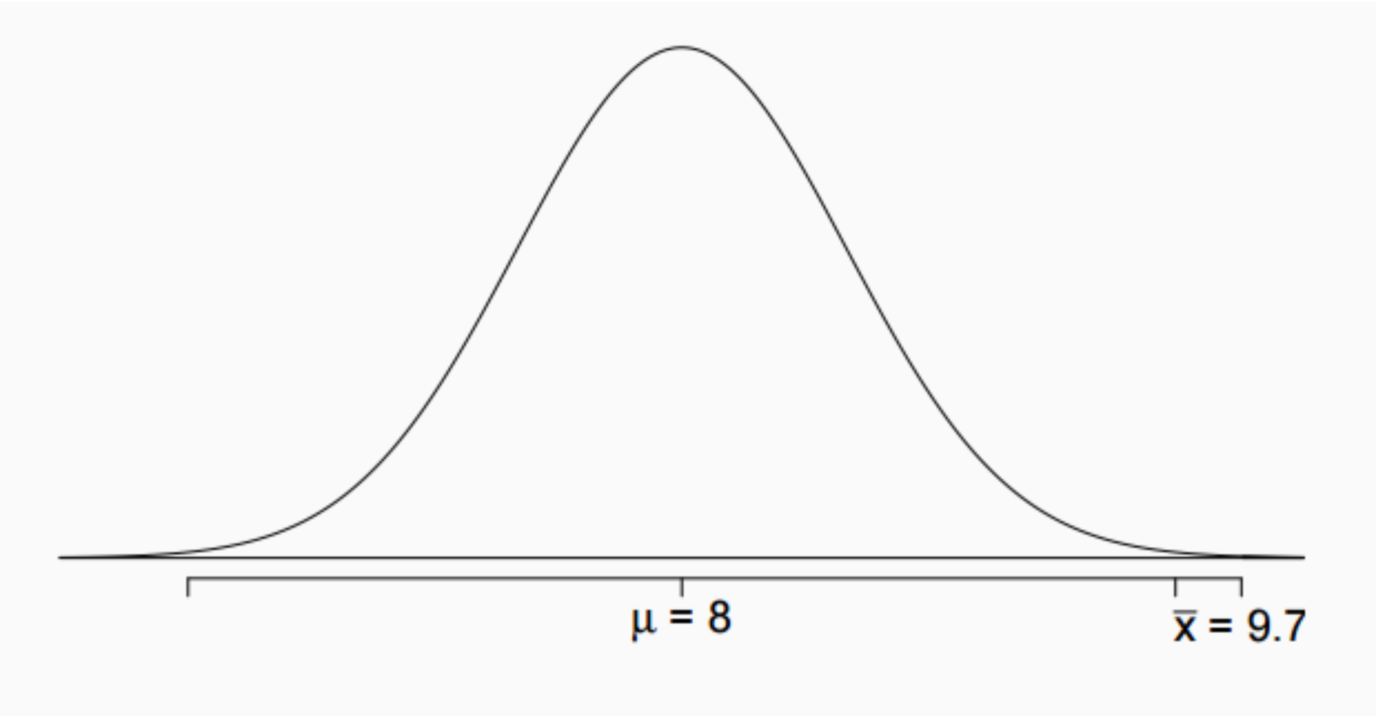
In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.

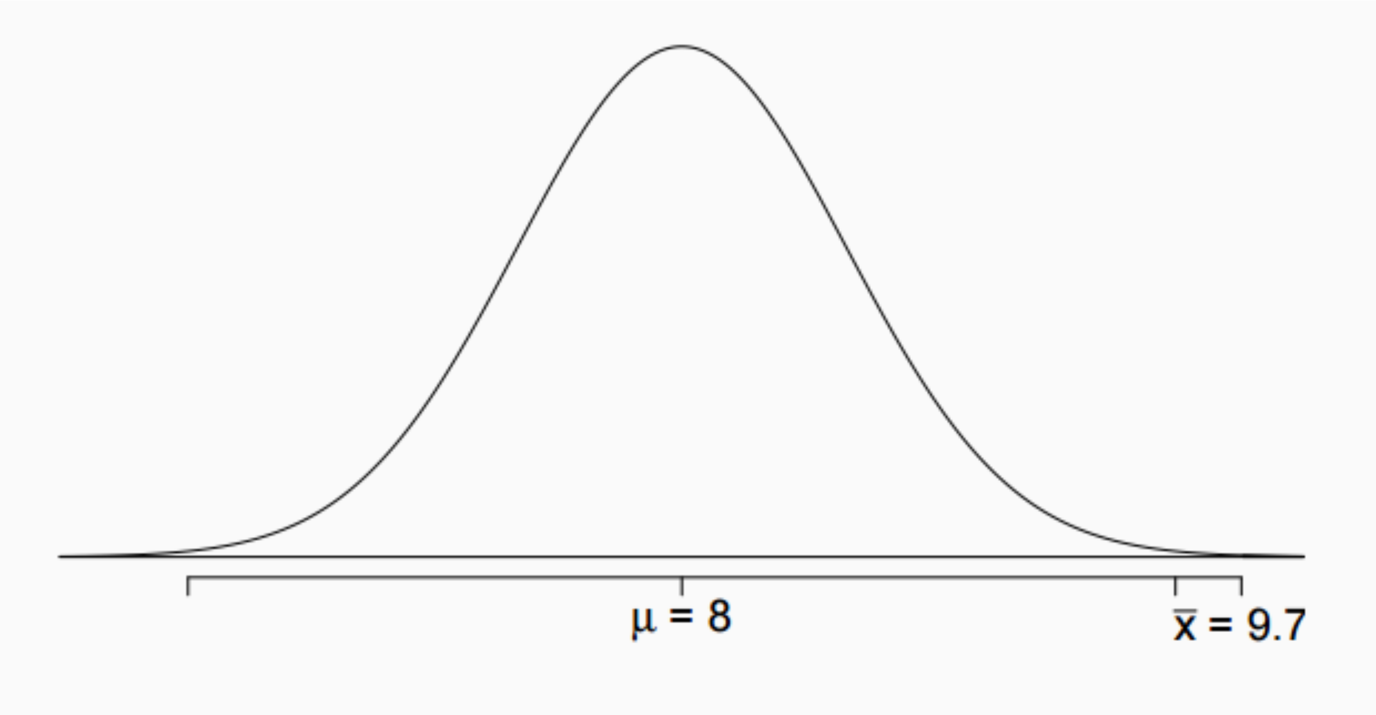


$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



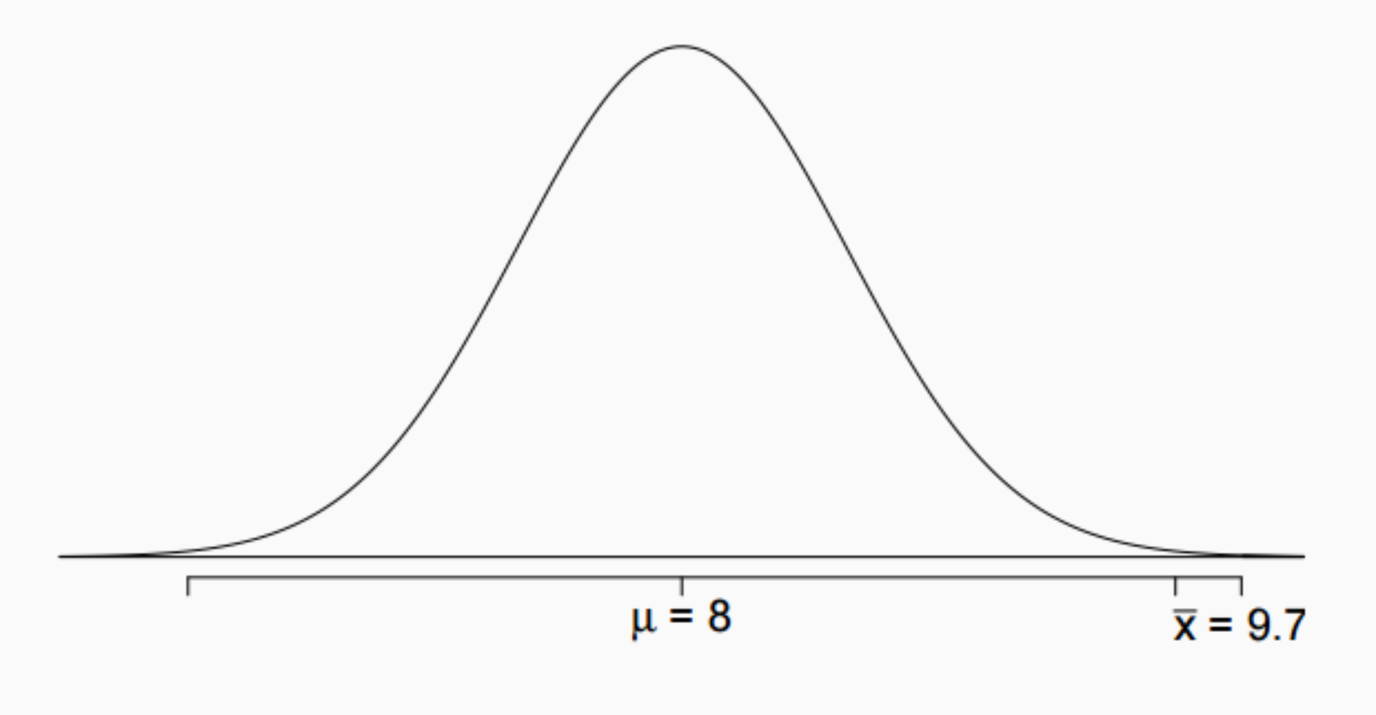
The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

Test Statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

Yes, and we can quantify how unusual it is using a p-value.

p-values

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

p-values

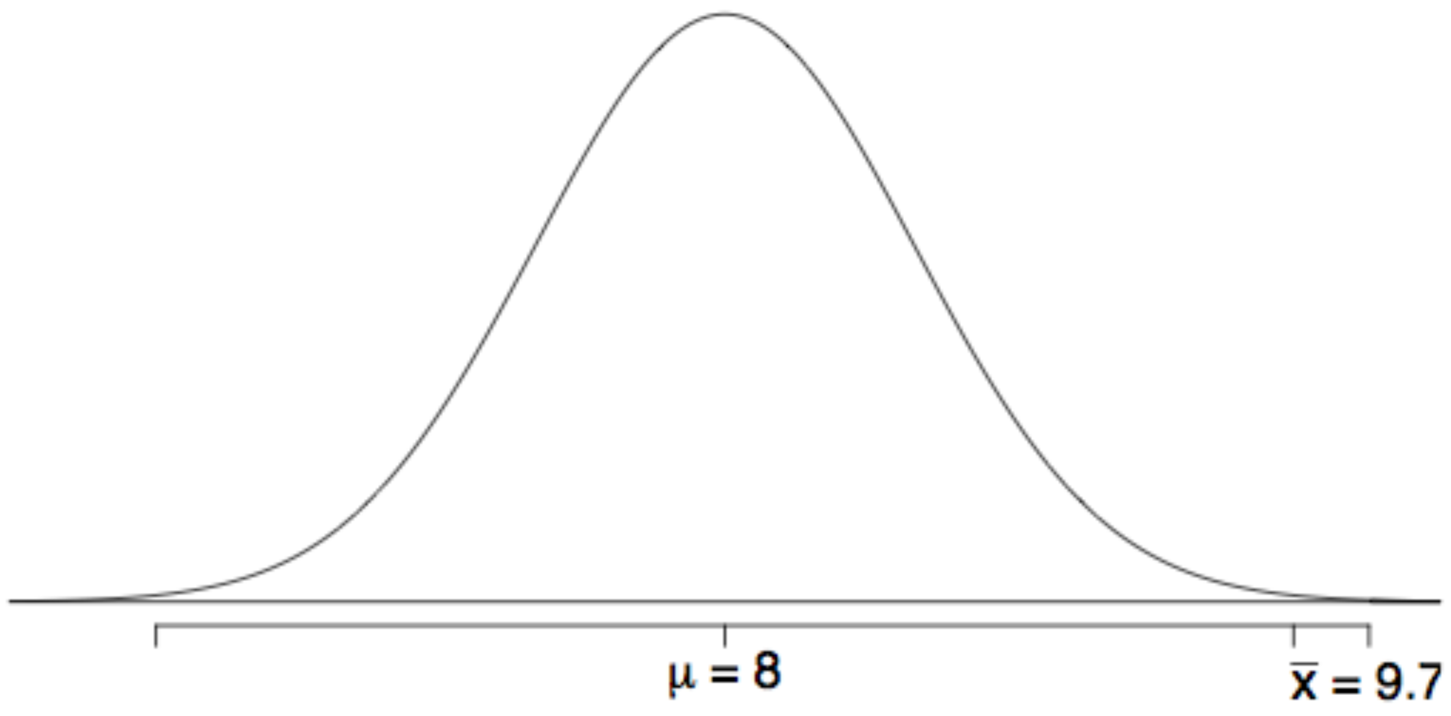
- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject H_0*.

p-values

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject H_0*.
- If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject H_0*.

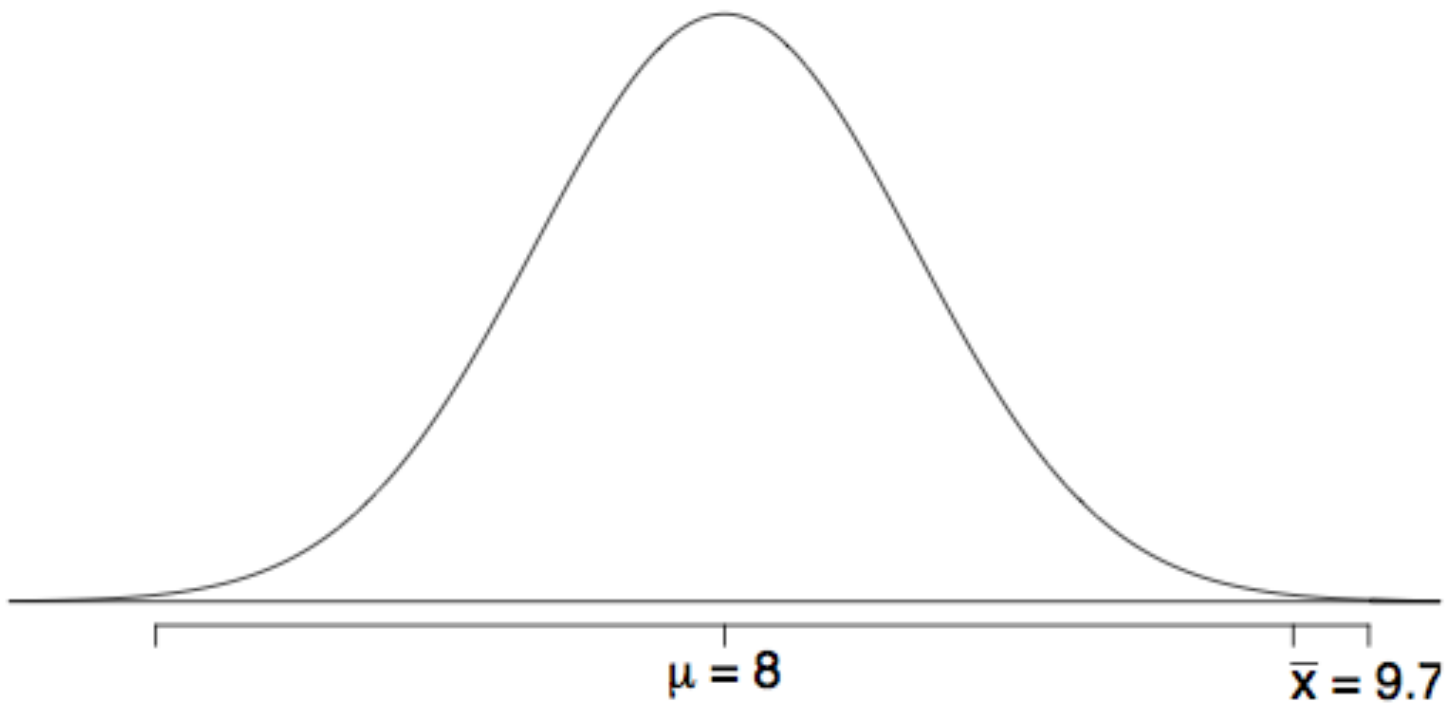
Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of college applications - Making a decision

- p-value = 0.0003

Number of college applications - Making a decision

- p-value = 0.0003
    - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject* H_0.

Number of college applications - Making a decision

- p-value = 0.0003
    - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
    - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject* H_0.
- The data provide convincing evidence that Duke students apply to more than 8 schools on average.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject H_0*.
- The data provide convincing evidence that Duke students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Practice

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

b) *Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.*

c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?”, the alternative hypothesis would be different

$$H_0: \mu = 7$$

$$H_A: \mu \neq 7$$

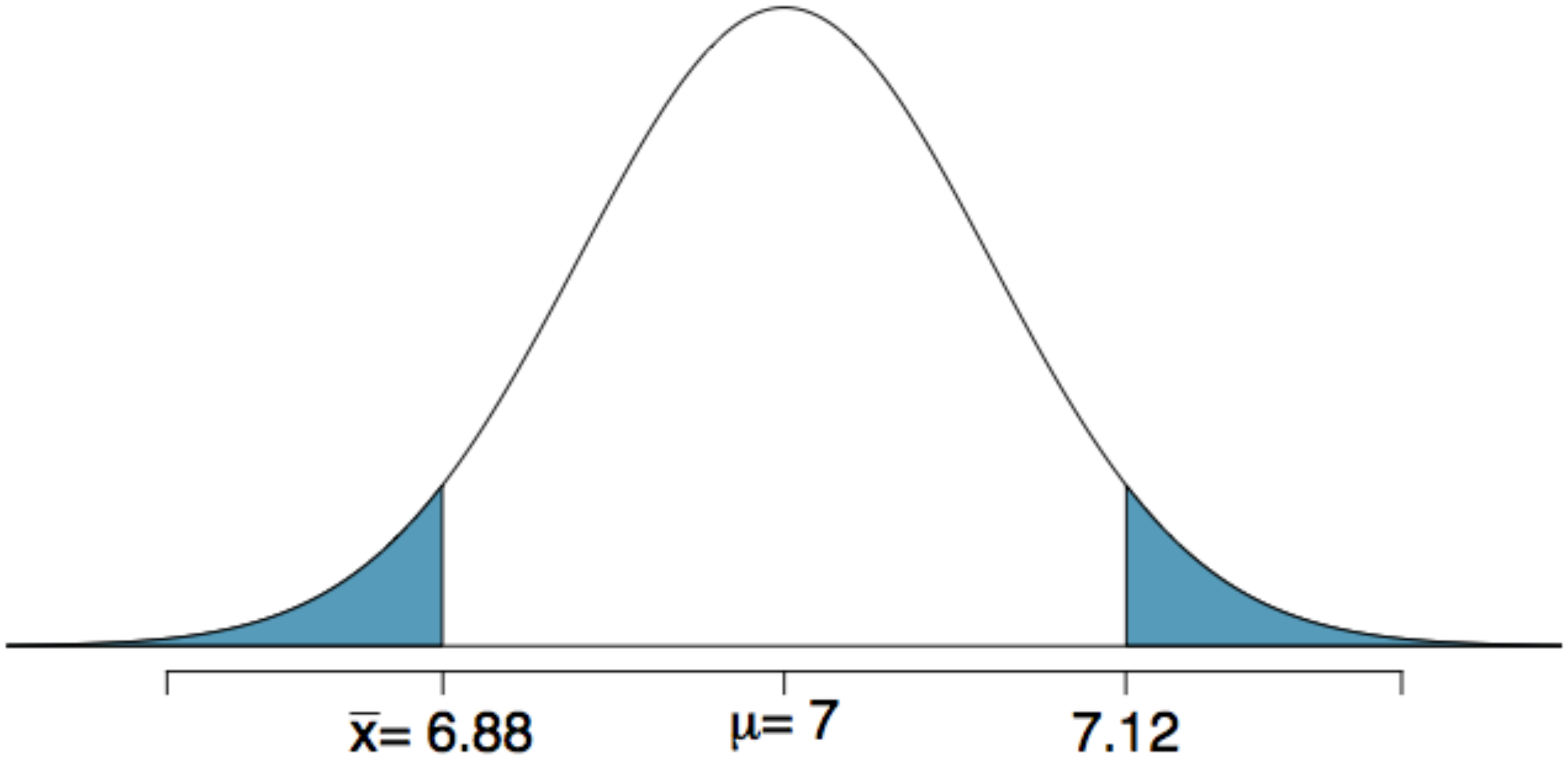
Two-sided hypothesis testing with p-values

- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?”, the alternative hypothesis would be different

$$H_0: \mu = 7$$

$$H_A: \mu \neq 7$$

- Hence the p-value would change as well:



p-value

$$= 0.0485 \times 2$$

$$= 0.097$$

Choosing a significance level

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.
- We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

the next two slides provide a brief summary of hypothesis testing...

Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

Recap: Hypothesis testing for a population mean

1. Set the hypotheses

- H_0: μ = null value
- H_A: μ < or > or ≠ null value

2. Calculate the point estimate

3. Check assumptions and conditions

- Independence: random sample/assignment, 10% condition when sampling without replacement
- Normality: nearly normal population or $n \geq 30$, no extreme skew -- or use the *t* distribution (Ch 5)

4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \ where\ SE = \frac{s}{\sqrt{n}}$$

5. Make a decision, and interpret it in context

- If p-value < α, reject H_0, data provide evidence for H_A
- If p-value > α, do not reject H_0, data do not provide evidence for H_A

Let’s discuss!

Testing for food safety

A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector uses a hypothesis testing framework to evaluate whether regulations are not being met. If he decides the restaurant is in gross violation, its license to serve food will be revoked.

(a) Write the hypotheses in words.

(b) What is a Type 1 Error in this context?

(c) What is a Type 2 Error in this context?

(d) Which error is more problematic for the restaurant owner? Why?

(e) Which error is more problematic for the diners? Why?

True or False

Determine if the following statements are true or false, and explain your reasoning. If false, state how it could be corrected.

(a) If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval.

(b) Decreasing the significance level, α, will increase the probability of making a Type 1 Error.

(c) Suppose the null hypothesis is $p = 0.5$ and we fail to reject H_0. Under this scenario, the true population proportion is 0.5.

(d) With large sample sizes, even small differences between the null value and the observed point estimate, a difference often called the effect size, will be identified as statistically significant.

Practical vs. statistical significance

Determine whether the following statement is true or false, and explain your reasoning: “With large sample sizes, even small differences between the null value and the observed point estimate can be statistically significant.”

Tomorrow is R Session!

- The first half of the lecture will be R session
- The second half of the lecture will be the time for the project discussion
  - Objective: Find the dataset that interests you the most!
  - Try to digest the project as much as possible during the lecture time!