Hypothesis testing

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$$\hat{p}_{males} = 21/24 \approx 0.88$$

$$\hat{p}_{females} = 14/24 \approx 0.58$$

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$$\hat{p}_{males} = 21/24 \approx 0.88$$

$$\hat{p}_{females} = 14/24 \approx 0.58$$

Possible explanations:

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

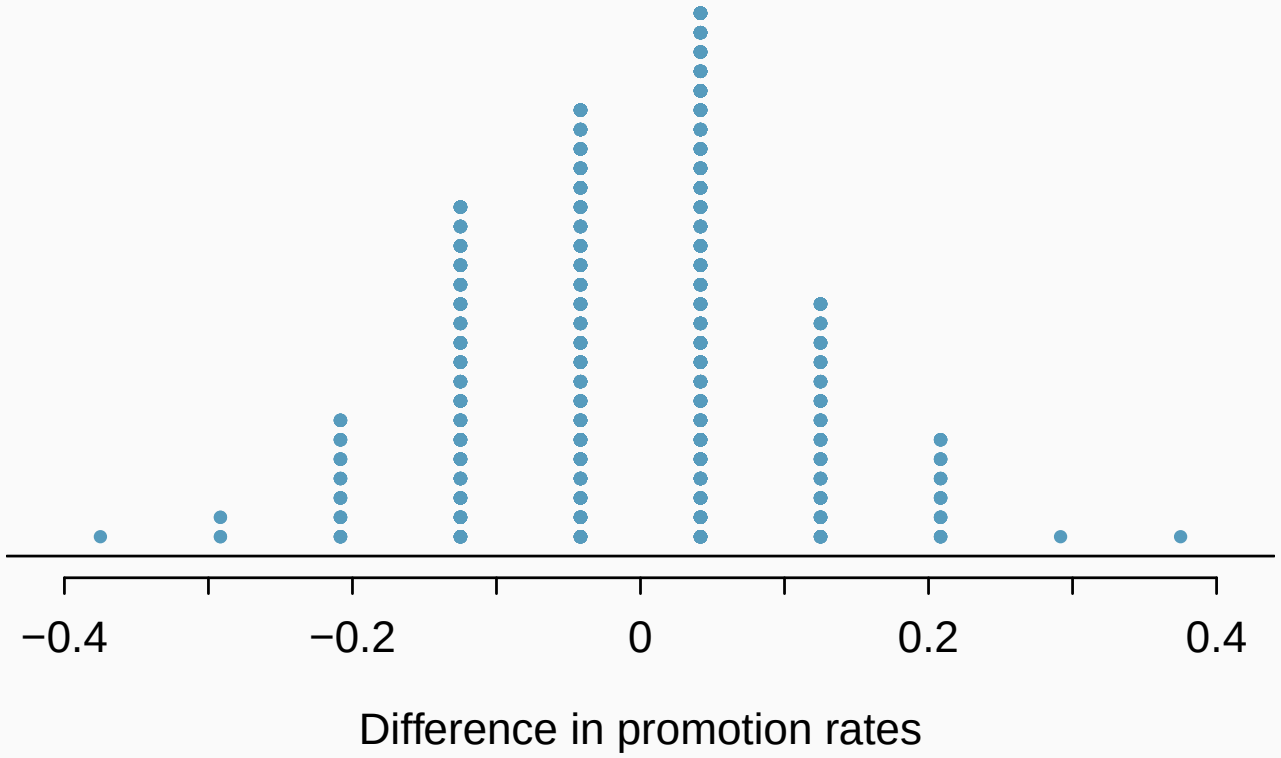
$$\hat{p}_{males} = 21/24 \approx 0.88$$

$$\hat{p}_{females} = 14/24 \approx 0.58$$

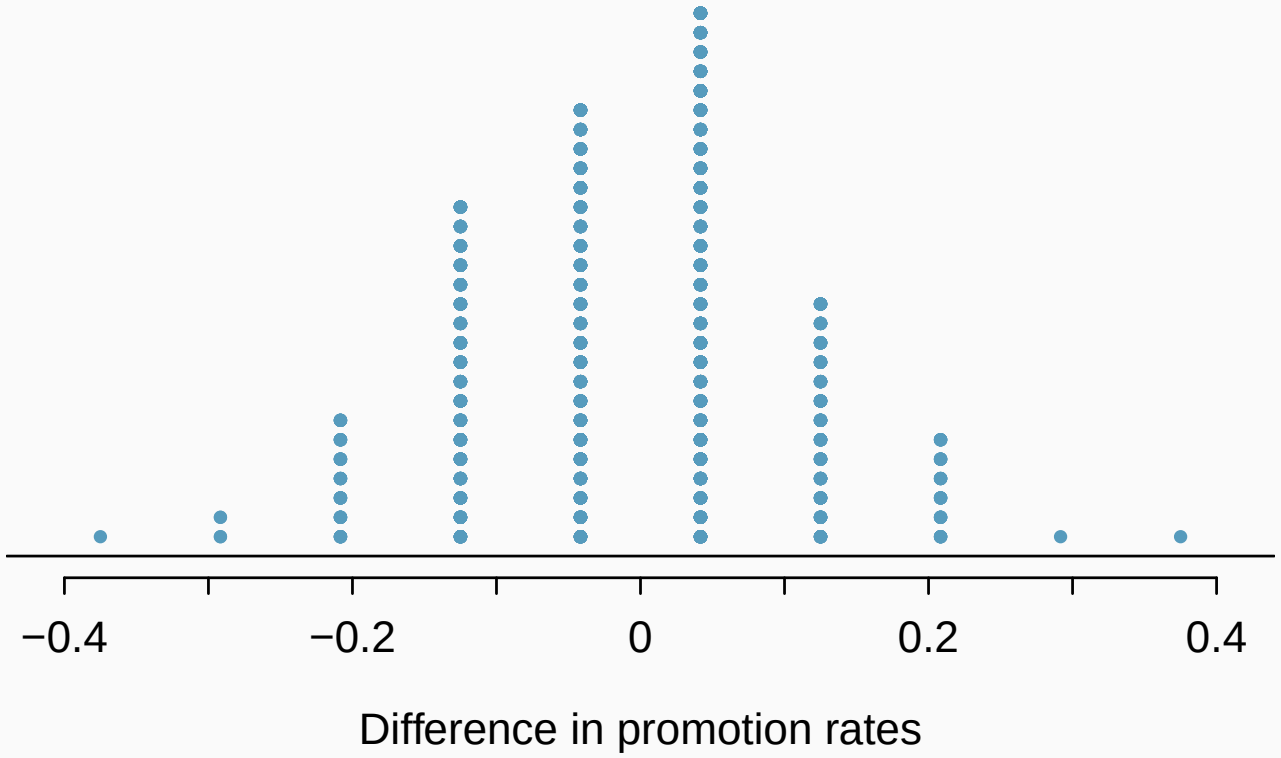
Possible explanations:

- Promotion and gender are *independent*; there is no gender discrimination; observed difference in proportions is simply due to chance. $\rightarrow$ *null* - (nothing is going on)
- Promotion and gender are *dependent*; there is gender discrimination, observed difference in proportions is not due to chance. $\rightarrow$ *alternative* - (something is going on)

Result



Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being at least 30 percentage points higher than female promotions), 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. (Difference [of means or proportions] due to chance.)

Recap: hypothesis testing framework

- We start with a *null hypothesis (H_0)* that represents the status quo. (Difference [of means or proportions] due to chance.)
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for. (Difference due to association between variables.)

Recap: hypothesis testing framework

- We start with a *null hypothesis (H_0)* that represents the status quo. (Difference [of means or proportions] due to chance.)
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for. (Difference due to association between variables.)
- 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. (Difference [of means or proportions] due to chance.)
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for. (Difference due to association between variables.)
- 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 difference is unusual under the null, we reject the null. (Unusual is measured with z score and tail area.) Otherwise, we retain the null.

Recap: hypothesis testing framework

- We start with a *null hypothesis (H_0)* that represents the status quo. (Difference [of means or proportions] due to chance.)
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for. (Difference due to association between variables.)
- 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 difference is unusual under the null, we reject the null. (Unusual is measured with z score and tail area.) Otherwise, we retain the null.

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 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.

example study

On the Longfellow bridge, the speed limit is 25 miles per hour. A cycling advocate hopes to find evidence that on average cars exceed the limit on the bridge. The advocate decides to measure a random sample of car speeds and to apply a one-tailed test.

example study

On the Longfellow bridge, the speed limit is 25 miles per hour. A cycling advocate hopes to find evidence that on average cars exceed the limit on the bridge. The advocate decides to measure a random sample of car speeds and to apply a one-tailed test.

What are the relevant hypotheses?

example study

On the Longfellow bridge, the speed limit is 25 miles per hour. A cycling advocate hopes to find evidence that on average cars exceed the limit on the bridge. The advocate decides to measure a random sample of car speeds and to apply a one-tailed test.

What are the relevant hypotheses?

null hypothesis, $H_0: \quad \mu = 25$

example study

On the Longfellow bridge, the speed limit is 25 miles per hour. A cycling advocate hopes to find evidence that on average cars exceed the limit on the bridge. The advocate decides to measure a random sample of car speeds and to apply a one-tailed test.

What are the relevant hypotheses?

null hypothesis, $H_0: \quad \mu = 25$

alternative hypothesis, $H_A: \quad \mu > 25$

example study (fake data)

The advocate takes a sample of 80 cars, which yield a sample mean of 28 miles per hour and a standard deviation of 15 miles per hour. Does the advocate have significant evidence of speeding?

example study (fake data)

The advocate takes a sample of 80 cars, which yield a sample mean of 28 miles per hour and a standard deviation of 15 miles per hour. Does the advocate have significant evidence of speeding?

We wonder... maybe the true population mean is 25 mph and this sample mean of 28 mph is just due to natural fluctuation.

example study (fake data)

The advocate takes a sample of 80 cars, which yield a sample mean of 28 miles per hour and a standard deviation of 15 miles per hour. Does the advocate have significant evidence of speeding?

We wonder... maybe the true population mean is 25 mph and this sample mean of 28 mph is just due to natural fluctuation.

So, we want to know how common/uncommon this difference is from random fluctuation alone.

example study (fake data)

The advocate takes a sample of 80 cars, which yield a sample mean of 28 miles per hour and a standard deviation of 15 miles per hour. Does the advocate have significant evidence of speeding?

We wonder... maybe the true population mean is 25 mph and this sample mean of 28 mph is just due to natural fluctuation.

So, we want to know how common/uncommon this difference is from random fluctuation alone.

Luckily, sampling distributions are normal! So if we estimate that $\sigma \approx 15$ mph, we have everything we need to answer that question.

Build the sampling distribution of the null hypothesis

The null hypothesis suggests that the population parameters are $\mu = 25$ and $\sigma \approx 15$.

Build the sampling distribution of the null hypothesis

The null hypothesis suggests that the population parameters are $\mu = 25$ and $\sigma \approx 15$.

What are the parameters of the null's **sampling distribution**?

$$\mu = 25$$

Build the sampling distribution of the null hypothesis

The null hypothesis suggests that the population parameters are $\mu = 25$ and $\sigma \approx 15$.

What are the parameters of the null's **sampling distribution**?

$$\mu = 25$$

$$SE = \frac{15}{\sqrt{80}} = 1.68$$

Test statistic

The **test statistic** is how many standard errors the null's population mean and the observed sample mean are from each other.

Test statistic

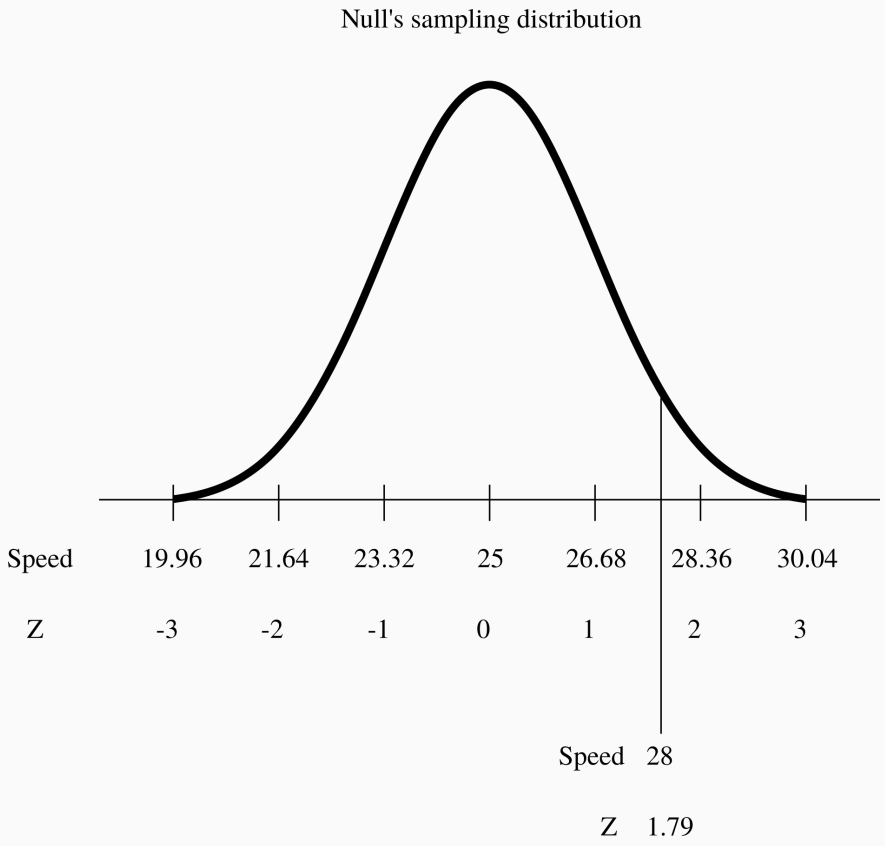
The **test statistic** is how many standard errors the null's population mean and the observed sample mean are from each other.

The test statistic is a measure of how unusual the observed sample mean would be if the null's hypothesis were true.

Test statistic

The **test statistic** is how many standard errors the null's population mean and the observed sample mean are from each other.

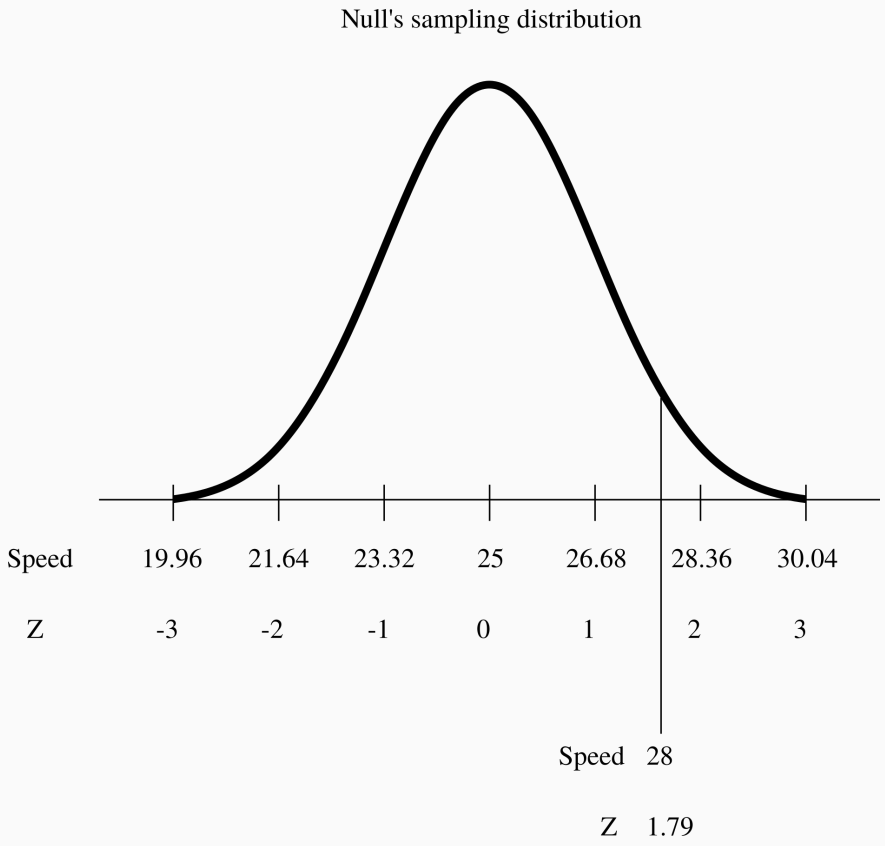
The test statistic is a measure of how unusual the observed sample mean would be if the null's hypothesis were true.



Test statistic

The **test statistic** is how many standard errors the null's population mean and the observed sample mean are from each other.

The test statistic is a measure of how unusual the observed sample mean would be if the null's hypothesis were true.

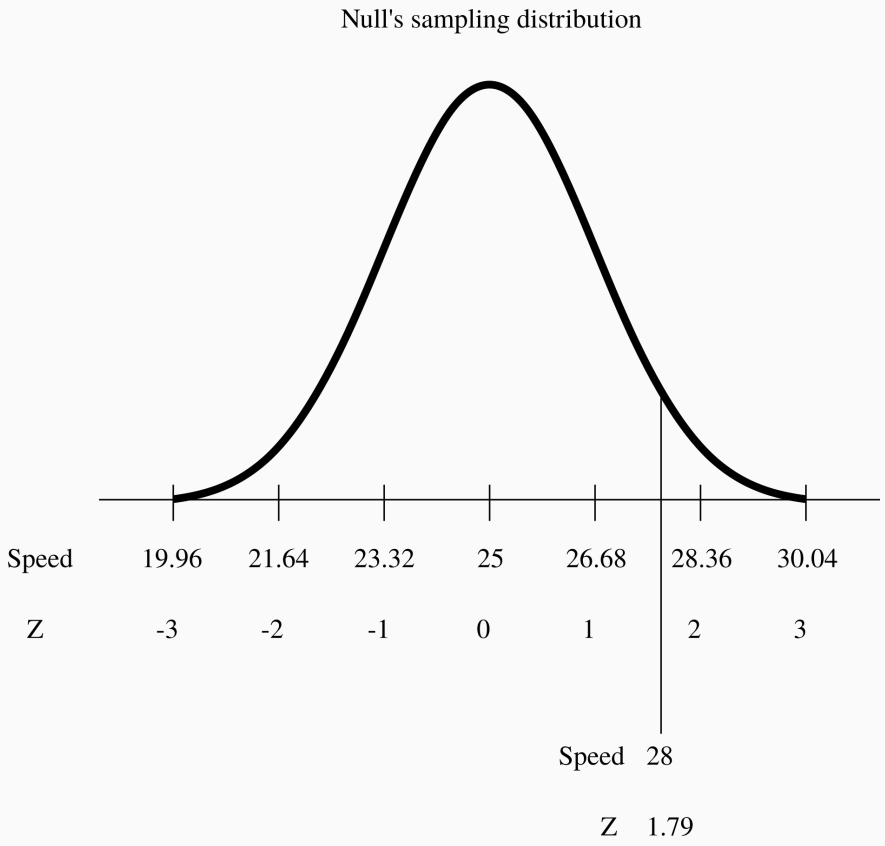


In this case, the test statistic is $z = \frac{28-25}{1.68} = 1.79$.

Test statistic

The **test statistic** is how many standard errors the null's population mean and the observed sample mean are from each other.

The test statistic is a measure of how unusual the observed sample mean would be if the null's hypothesis were true.



In this case, the test statistic is $z = \frac{28-25}{1.68} = 1.79$.

We quantify how unusual this is with 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

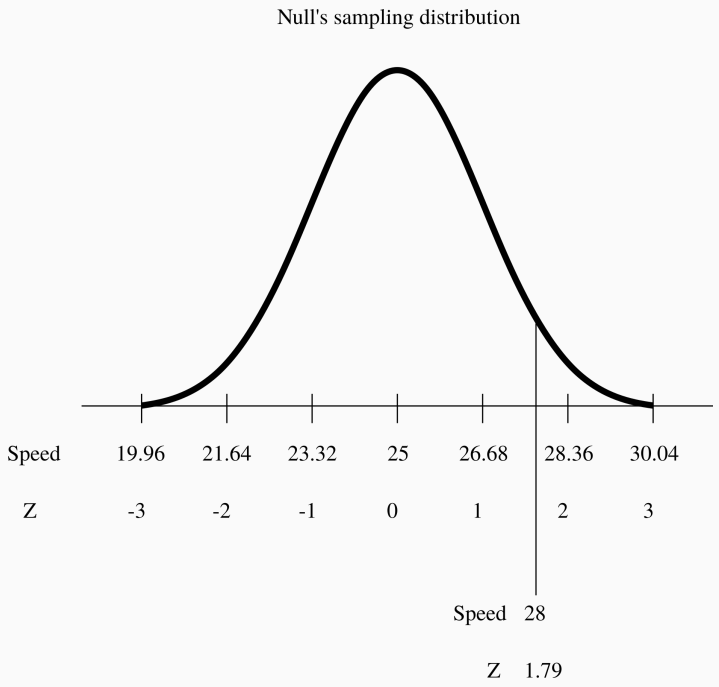
- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that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

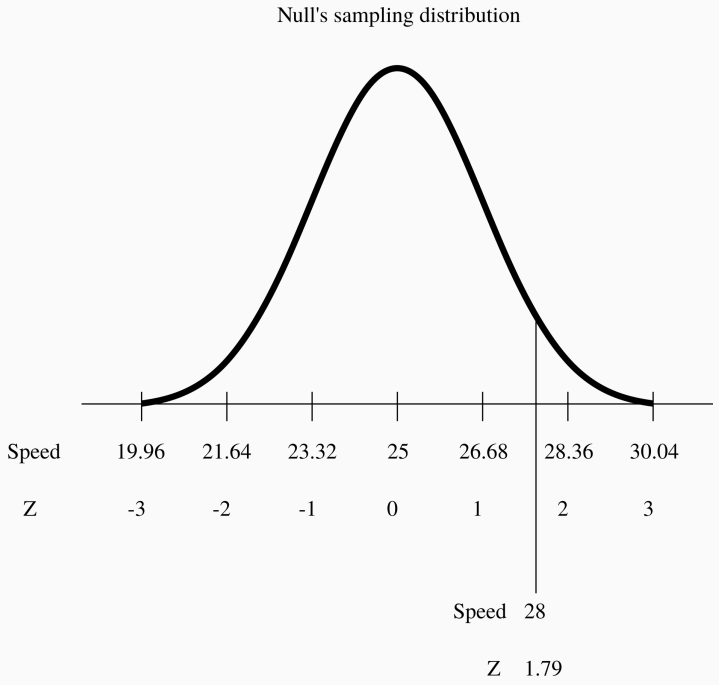
Speeding example

p-value: The probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 28) if in fact H_0 were true (the population mean was 25).



Speeding example

p-value: The probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 28) if in fact H_0 were true (the population mean was 25).



$$P(\bar{x} > 28 | \mu = 25) = P(Z > 1.79) = 0.0367$$

Making a decision

- p-value = 0.0367

Making a decision

- p-value = 0.0367
  - If the true average speed 25, there is only 3.67% chance of observing a random sample of 80 drivers who on average drive 28 mph or more.

Making a decision

- p-value = 0.0367
  - If the true average speed 25, there is only 3.67% chance of observing a random sample of 80 drivers who on average drive 28 mph or more.
  - This is a pretty low probability for us to think that a sample mean of 28 mph is likely to happen simply by chance.

Making a decision

- p-value = 0.0367
  - If the true average speed 25, there is only 3.67% chance of observing a random sample of 80 drivers who on average drive 28 mph or more.
  - This is a pretty low probability for us to think that a sample mean of 28 mph is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject* H_0.

Making a decision

- p-value = 0.0367
  - If the true average speed 25, there is only 3.67% chance of observing a random sample of 80 drivers who on average drive 28 mph or more.
  - This is a pretty low probability for us to think that a sample mean of 28 mph 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 cars speed.

Making a decision

- p-value = 0.0367
  - If the true average speed 25, there is only 3.67% chance of observing a random sample of 80 drivers who on average drive 28 mph or more.
  - This is a pretty low probability for us to think that a sample mean of 28 mph 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 cars speed.
- The difference between the null value of 25 mph and observed sample mean of 28 mph is *not due to chance* or sampling variability.

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.

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.

$$H_0 : \mu = 7$$

$$H_A : \mu \neq 7$$

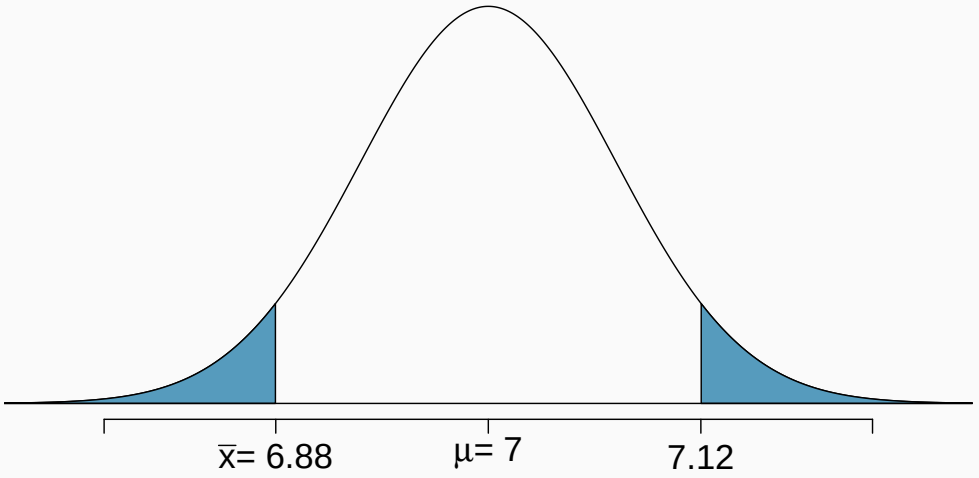
Two-sided hypothesis testing with p-values

- 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
$= 0.0485 \times 2$
$= 0.097$

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:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

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:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

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:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

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

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:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

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:

$$H_0: \text{ Defendant is innocent}$$
$$H_A: \text{ Defendant is guilty}$$

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?

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.

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

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.

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

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

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 are provided as 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
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$
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
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 $< \alpha$, reject H_0, data provide evidence for H_A
   - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A