Hypothesis Testing II

Nate Wells

Math 141, 3/24/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will. . .

- Perform hypothesis test to determine whether smiling has effect on leniency of punishment
- Discuss Power! (the statistical kind)

Section 1

Hypothesis Testing Example

Smile!

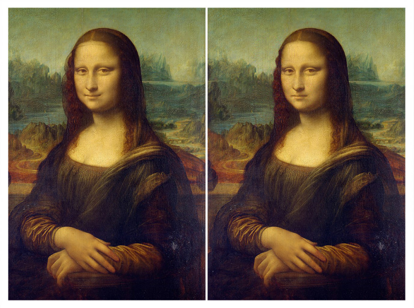
Can a simple smile have an effect on the punishment assigned following an infraction?

Smile!

Can a simple smile have an effect on the punishment assigned following an infraction?

In a 1995 study, Hecht and LeFrance examined the effect of a smile on the leniency of disciplinary action for wrongdoers. Participants in the experiment took on the role of members of a college disciplinary panel judging students accused of cheating.

For each suspect, along with a description of the offense, a picture was provided with either a smile or neutral facial expression. A leniency score was calculated based on the disciplinary decisions made by the participants.



Research Question

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Actual Population:

Response variable:

Explanatory variable:

Statistics:

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population:

Response variable:

Explanatory variable:

Statistics:

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population: intro psych students?

Response variable:

Explanatory variable:

Statistics:

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population: intro psych students?

Response variable: leniency score Y (quantitative)

Explanatory variable:

Statistics:

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population: intro psych students?

Response variable: leniency score Y (quantitative)

Explanatory variable: whether image displayed a smile X (categorical)

Statistics:

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population: intro psych students?

Response variable: leniency score Y (quantitative)

Explanatory variable: whether image displayed a smile X (categorical)

Statistics: Let $\bar{x}_s$ and $\bar{x}_n$ be the mean leniency score among the smile and neutral groups

Parameters:

Research Question

Question: Can a smile have an effect on the punishment assigned following an infraction?

Implied Population: all people

Subjects consisted of 68 intro psych students participating per course requirement

Actual Population: intro psych students?

Response variable: leniency score Y (quantitative)

Explanatory variable: whether image displayed a smile X (categorical)

Statistics: Let $\bar{x}_s$ and $\bar{x}_n$ be the mean leniency score among the smile and neutral groups

Parameters: Let μ_s and μ_n be the theoretical mean leniency score for the population exposed to a smile and neutral epxression

Hypotheses

Hypotheses

Null Hypothesis:

Hypotheses

Null Hypothesis: Leniency and smile are independent.

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis:

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis: Leniency and smile are not independent.

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis: Leniency and smile are not independent.

$$H_0 : \mu_s \neq \mu_n$$

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis: Leniency and smile are not independent.

$$H_0 : \mu_s \neq \mu_n$$

- The hypothesese make no reference to sample statistics.

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis: Leniency and smile are not independent.

$$H_0 : \mu_s \neq \mu_n$$

- The hypothesese make no reference to sample statistics.
- Even if H_0 were true, expect to find a difference between $\bar{x}_s$ and $\bar{x}_n$ (random sampling)

Hypotheses

Null Hypothesis: Leniency and smile are independent.

$$H_0 : \mu_s = \mu_n$$

Alternate Hypothesis: Leniency and smile are not independent.

$$H_0 : \mu_s \neq \mu_n$$

- The hypothesese make no reference to sample statistics.
- Even if H_0 were true, expect to find a difference between $\bar{x}_s$ and $\bar{x}_n$ (random sampling)
- Our goal is to determine how much difference is typical.

Collect Data

```
smiles <- read_csv("Data/Smiles.csv")
```

Collect Data

```r
smiles <- read_csv("Data/Smiles.csv")
```

Let's look at a small section of the data:

```r
sample_n(smiles, 5 )
```

```
## # A tibble: 5 x 2
##   Leniency Group
##      <dbl> <chr>
## 1      3   neutral
## 2      7.5 smile
## 3      6   neutral
## 4      6.5 neutral
## 5      3.5 smile
```

Collect Data

```
smiles <- read_csv("Data/Smiles.csv")
```

Let's look at a small section of the data:

```
sample_n(smiles, 5 )
```

```
## # A tibble: 5 x 2
##   Leniency Group
##      <dbl> <chr>
## 1      3   neutral
## 2      7.5 smile
## 3      6   neutral
## 4      6.5 neutral
## 5      3.5 smile
```

```
smiles %>%
  group_by(Group) %>%
  summarize(avg = mean(Leniency) )
```

```
## # A tibble: 2 x 2
##   Group     avg
##   <chr>   <dbl>
## 1 neutral  4.12
## 2 smile    4.91
```

Collect Data

```
smiles <- read_csv("Data/Smiles.csv")
```

Let's look at a small section of the data:

```
sample_n(smiles, 5 )
```

```
## # A tibble: 5 x 2
##   Leniency Group
##      <dbl> <chr>
## 1      3   neutral
## 2      7.5 smile
## 3      6   neutral
## 4      6.5 neutral
## 5      3.5 smile
```

```
smiles %>%
  group_by(Group) %>%
  summarize(avg = mean(Leniency) )
```

```
## # A tibble: 2 x 2
##   Group     avg
##   <chr>   <dbl>
## 1 neutral  4.12
## 2 smile    4.91
```

Each group consisted of 34 students, with $\bar{x}_s = 4.9$ and $\bar{x}_n = 4.1$ and a difference of

$$\bar{x}_s - \bar{x}_n = 0.8$$

Generate Replicates

Assuming that smile has no effect on leniency, then group assignment is superfluous.

Generate Replicates

Assuming that smile has no effect on leniency, then group assignment is superfluous.

We model the effect of sampling by shuffling the `Group` labels among all leniency scores:

Generate Replicates

Assuming that smile has no effect on leniency, then group assignment is superfluous.

We model the effect of sampling by shuffling the `Group` labels among all leniency scores:

Table 1: Original

Leniency	Group
6.5	neutral
6.0	neutral
7.0	smile
4.5	smile
2.0	neutral

Table 2: Shuffled

Leniency	Group
6.5	smile
6.0	neutral
7.0	neutral
4.5	neutral
2.0	smile

$\bar{x}_s - \bar{x}_n = 5.75 - 4.83 = 0.82$

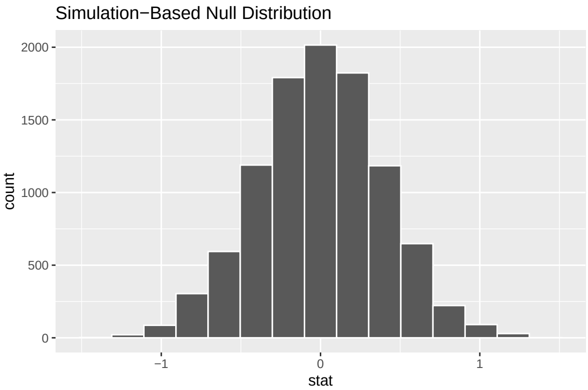
$\bar{x}_s - \bar{x}_n = 5.83 - 4.25 = 1.58$

Simulated Difference

Creating 10,000 shuffled samples should show how difference in leniency scores fluctuates just due to sampling

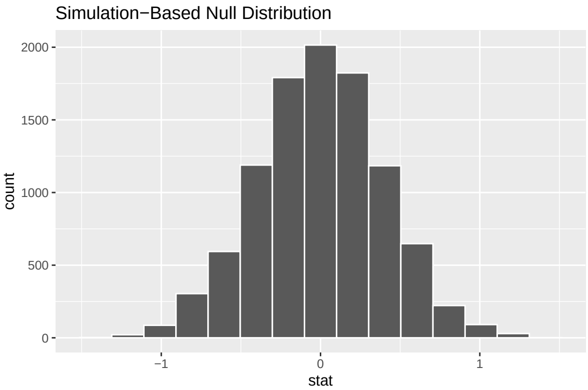
Simulated Difference

Creating 10,000 shuffled samples should show how difference in leniency scores fluctuates just due to sampling



Simulated Difference

Creating 10,000 shuffled samples should show how difference in leniency scores fluctuates just due to sampling



The mean of the simulated Null Distribution is 0.023 and its standard deviation is 0.391

Liklihood of Observed Difference

In our sample, we observed a difference of $\bar{x}_s - \bar{x}_n = 0.8$, which is about 2 standard deviations from the mean.

Liklihood of Observed Difference

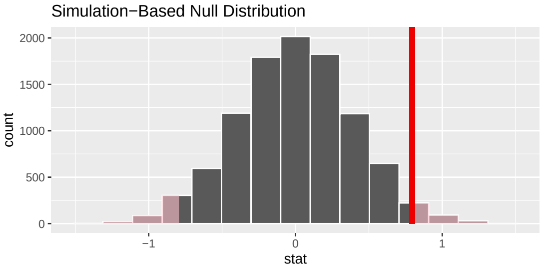
In our sample, we observed a difference of $\bar{x}_s - \bar{x}_n = 0.8$, which is about 2 standard deviations from the mean.

Since the Null Distribution is approximately Normal, a difference like this or more extreme occurs about 5% of the time.

Liklihood of Observed Difference

In our sample, we observed a difference of $\bar{x}_s - \bar{x}_n = 0.8$, which is about 2 standard deviations from the mean.

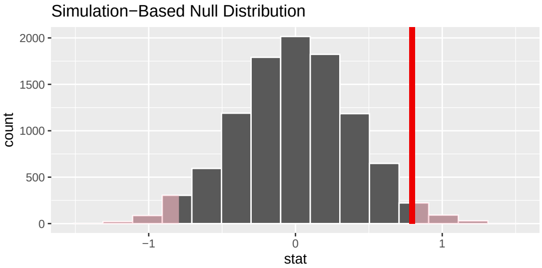
Since the Null Distribution is approximately Normal, a difference like this or more extreme occurs about 5% of the time.



Liklihood of Observed Difference

In our sample, we observed a difference of $\bar{x}_s - \bar{x}_n = 0.8$, which is about 2 standard deviations from the mean.

Since the Null Distribution is approximately Normal, a difference like this or more extreme occurs about 5% of the time.

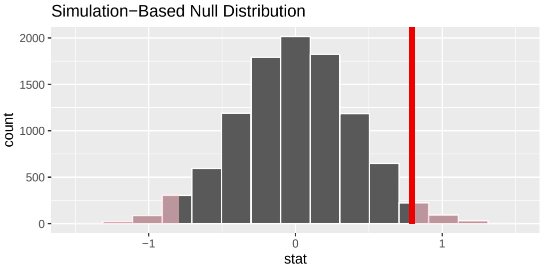


Since we have a two-sided alternate hypothesis, we consider the area in **both** tails when calculating our P-value.

Liklihood of Observed Difference

In our sample, we observed a difference of $\bar{x}_s - \bar{x}_n = 0.8$, which is about 2 standard deviations from the mean.

Since the Null Distribution is approximately Normal, a difference like this or more extreme occurs about 5% of the time.



Since we have a two-sided alternate hypothesis, we consider the area in **both** tails when calculating our P-value.

The precise P-value (area in both tails) is 0.0474

Conclusion

If the null hypothesis were true (and leniency were independent of smile), then we would observe a sample difference as extreme as this one less than 5% of the time.

Conclusion

If the null hypothesis were true (and leniency were independent of smile), then we would observe a sample difference as extreme as this one less than 5% of the time.

This is good evidence that our observation is inconsistent with the Null Hypothesis model

Conclusion

If the null hypothesis were true (and leniency were independent of smile), then we would observe a sample difference as extreme as this one less than 5% of the time.

This is good evidence that our observation is inconsistent with the Null Hypothesis model

Using the standard significance level $\alpha = 0.05$ and noting $\text{P-Value} < \alpha$, we reject the null hypothesis in favor of the alternative.

Conclusion

If the null hypothesis were true (and leniency were independent of smile), then we would observe a sample difference as extreme as this one less than 5% of the time.

This is good evidence that our observation is inconsistent with the Null Hypothesis model

Using the standard significance level $\alpha = 0.05$ and noting $\text{P-Value} < \alpha$, we reject the null hypothesis in favor of the alternative.

Conclusion

Based on an experiment performed on 68 intro psych students, a smile likely does have an effect on the punishment assigned following an infraction.

Interpreting Hypothesis Tests

Suppose we were to replicate the Smile experiment on a different group of psych students, obtaining a P-value of 0.001, which is statistically significant at the standard $\alpha = 0.05$ level.

Interpreting Hypothesis Tests

Suppose we were to replicate the Smile experiment on a different group of psych students, obtaining a P-value of 0.001, which is statistically significant at the standard $\alpha = 0.05$ level.

Consider the following two interpretations of this result:

Interpreting Hypothesis Tests

Suppose we were to replicate the Smile experiment on a different group of psych students, obtaining a P-value of 0.001, which is statistically significant at the standard $\alpha = 0.05$ level.

Consider the following two interpretations of this result:

1. "Based on the study, there is at most a 5% chance that the null hypothesis is correct (i.e. that a smile has no effect on leniency)." Is this correct?

Interpreting Hypothesis Tests

Suppose we were to replicate the Smile experiment on a different group of psych students, obtaining a P-value of 0.001, which is statistically significant at the standard $\alpha = 0.05$ level.

Consider the following two interpretations of this result:

1. "Based on the study, there is at most a 5% chance that the null hypothesis is correct (i.e. that a smile has no effect on leniency)." Is this correct?
2. "Since the P-Value so much smaller than the significance level, we conclude that a smile must have a **strong** effect on leniency." Is this correct?

Section 2

Power and Errors

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.
- A **Type 2 Error** occurs when we fail to reject H_0 when it is in fact false.

Types of Errors

- Hypothesis Tests give framework for comparing uncertainty, but do not guarantee that our conclusion will never be in err.
  - Remember: Unlikely things happen. All of the time.
- There are four possible outcomes to a hypothesis test, summarized below:

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	Correct Decision	Type 1 Error
	H_A true	Type 2 Error	Correct Decision

- A **Type 1 Error** occurs when we reject H_0 when it is actually true.
  - The coin is actually fair. But we saw an unlikely event and claimed the coin was biased.
- A **Type 2 Error** occurs when we fail to reject H_0 when it is in fact false.
  - The coin was indeed biased. But we withheld judgment since unlikely events do happen from time to time.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

With great power comes...

Significance Level and Power

- The significance level of a hypothesis test corresponds to our willingness to make Type I errors.
- Decreasing the significance level decreases the number of Type I errors made across a large number of experiments.
- Is there a cost to decreasing significance level to ensure we do not make Type I errors?
  - Yes! Because decreasing the significance level also makes it less likely we will reject H_0, and so usually increases the chance of making a Type 2 error.
- The **power** of a statistical test is the probability of correctly rejecting the null hypothesis when it is false. That is

$$\text{Power} = 1 - \text{Probability of Type II Error}$$

  - In general, computing power can be difficult, and requires we investigate the distribution of a sample statistic under the alternative hypothesis.

With great power comes...greater chance of Type I error.

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are the Null and Alternate Hypotheses in this case?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has COVID

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has COVID
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has COVID
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?

Rapid COVID test

A quick and accessible (but unreliable) test for COVID-19 is to match a patient's symptoms to the 10 most common symptoms exhibited by victims of COVID.

Suppose a person walks into a medical clinic with 6 of the 10 symptoms of COVID, and medical personnel are concerned the person may have COVID.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has COVID
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?
5. What significance level are you willing to use for this COVID test? *Remember, decreasing significance level also decreases the power of the test.*

DNA Tests

DNA testing allows researchers to compare markers in a person's DNA to those found at crime scene. Suppose the DNA found at a crime scene will **always** match the perpetrator of the crime. However, there is a small chance that the crime scene DNA will also match the markers for another innocent person.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA matches that found at the crime scene.

1. What are the Null and Alternate Hypotheses in this case?

DNA Tests

DNA testing allows researchers to compare markers in a person's DNA to those found at crime scene. Suppose the DNA found at a crime scene will **always** match the perpetrator of the crime. However, there is a small chance that the crime scene DNA will also match the markers for another innocent person.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA matches that found at the crime scene.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has committed the crime.

DNA Tests

DNA testing allows researchers to compare markers in a person's DNA to those found at crime scene. Suppose the DNA found at a crime scene will **always** match the perpetrator of the crime. However, there is a small chance that the crime scene DNA will also match the markers for another innocent person.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA matches that found at the crime scene.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has committed the crime.
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?

DNA Tests

DNA testing allows researchers to compare markers in a person's DNA to those found at crime scene. Suppose the DNA found at a crime scene will **always** match the perpetrator of the crime. However, there is a small chance that the crime scene DNA will also match the markers for another innocent person.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA matches that found at the crime scene.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has committed the crime.
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?

DNA Tests

DNA testing allows researchers to compare markers in a person's DNA to those found at crime scene. Suppose the DNA found at a crime scene will **always** match the perpetrator of the crime. However, there is a small chance that the crime scene DNA will also match the markers for another innocent person.

Suppose a person is on trial for a crime. Forensic scientists attest that the person's DNA matches that found at the crime scene.

1. What are the Null and Alternate Hypotheses in this case?
2. What 'statistic' is being used to determine whether the person has committed the crime.
3. In the context of this problem, what does a Type I error represent? What are some possible consequences of a Type I error?
4. Similarly, what does a Type II error represent? What are some possible consequences of a Type II error?
5. What significance level are you willing to use for this DNA test? *Remember, decreasing significance level also decreases the power of the test.*