

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Hypothesis Testing

Hypothesis Testing

- Goal: Answer yes/no questions about the world
- Examples:
 - Is coffee good for you?
 - Did water from the Broad Street pump cause cholera?
 - Does coffee cause cancer?
- We are not looking for a scientific reasoning for cause and effect.
- We are looking for reliable evidence in the data to establish causality. If there is indeed evidence to suggest causality, we may want to look for a scientific explanation.
 - Example: John Snow's data analysis only established that the sewage in the water was causing the cholera. It did not tell exactly how. It took separate science experiments to isolate bacterium *Vibrio Cholerae* as the root cause.

Hypothesis Testing

- Goal: Answer yes/no questions about the world
- Examples:
 - Is coffee good for you?
 - Did water from the Broad Street pump cause cholera?
 - Does coffee cause cancer?
- Key idea:
 - Make a hypothesis (typically called the *null hypothesis*)
 - The hypothesis imposes a model on the data
 - Model: *A set of assumptions about the data, including the random process that generates it.*
 - Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.
 - Decide on a *test statistic* and check how likely is the observed (or more extreme) value under the hypothesis.

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution: *“In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an **impartial jury** of the State and district wherein the crime shall have been committed.”*

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution: *“In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an **impartial jury** of the State and district wherein the crime shall have been committed.”*
 - Data:
 - At the time of the trial, only men aged 21 or more were eligible to serve on juries in Talladega County.
 - In that population, 26% of the men were Black.
 - Only eight men among the panel of 100 men (that is, 8%) were Black.

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution:
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed."
 - Data:
 - At the time of the trial, only men aged 21 or more were eligible to serve on juries in Talladega County.
 - In that population, 26% of the men were Black.
 - Only eight men among the panel of 100 men (that is, 8%) were Black.

- Null hypothesis:

- The jury is impartial, i.e., the panel was selected by randomly from the population.
- Test statistic: Number of black men in the panel.
- We want the probability that a randomly selected panel of 100 jurors from the population has less than 8 black men.

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

• Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution:
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed."
 - Data:
 - At the time of the trial, only men aged 21 or more were eligible to serve on juries in Talladega County.
 - In that population, 26% of the men were Black.
 - Only eight men among the panel of 100 men (that is, 8%) were Black.

• Null hypothesis:

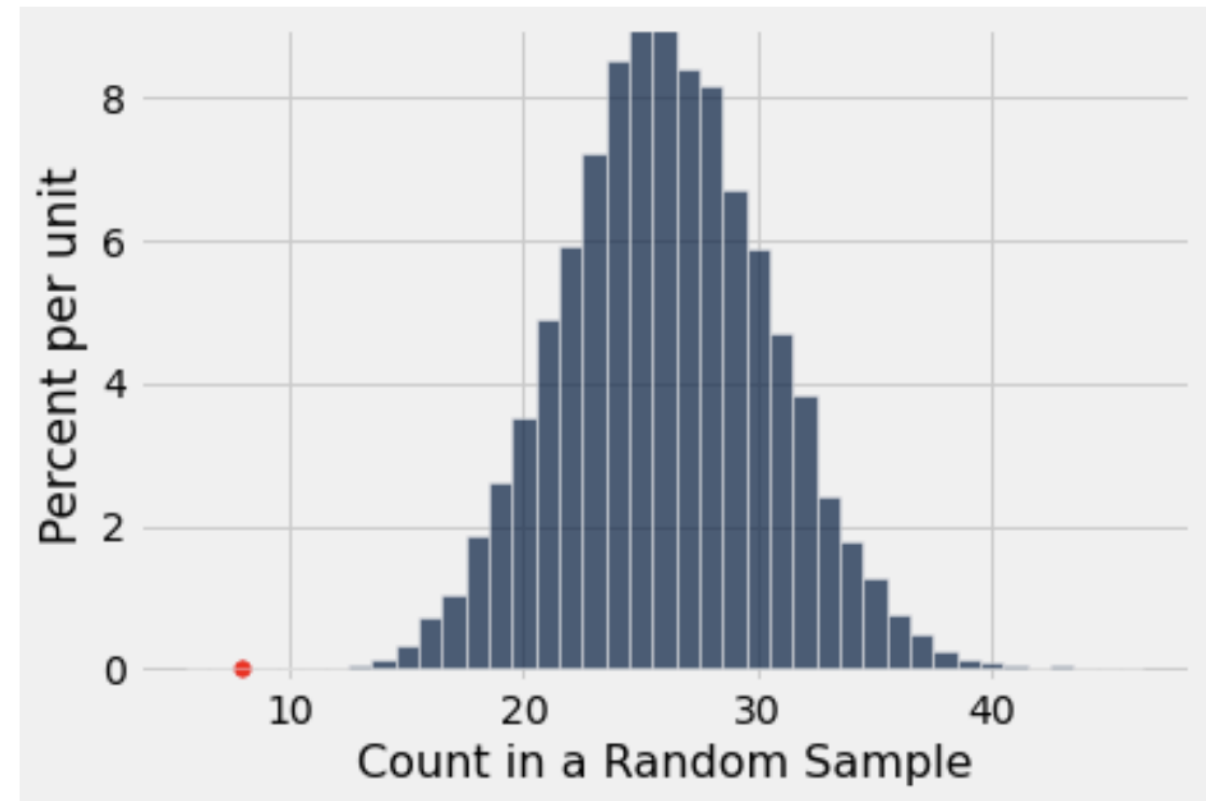
- The jury is impartial, i.e., the panel was selected by randomly from the population.
- Test statistic: Number of black men in the panel.
- We want the probability that a randomly selected panel of 100 jurors from the population has ≤ 8 black men.
 - We can obtain an estimate by *simulating* the test statistic under the null hypothesis.

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution:
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed."
 - Data:
 - At the time of the trial, only men aged 21 or more were eligible to serve on juries in Talladega County.
 - In that population, 26% of the men were Black.
 - Only eight men among the panel of 100 men (that is, 8%) were Black.

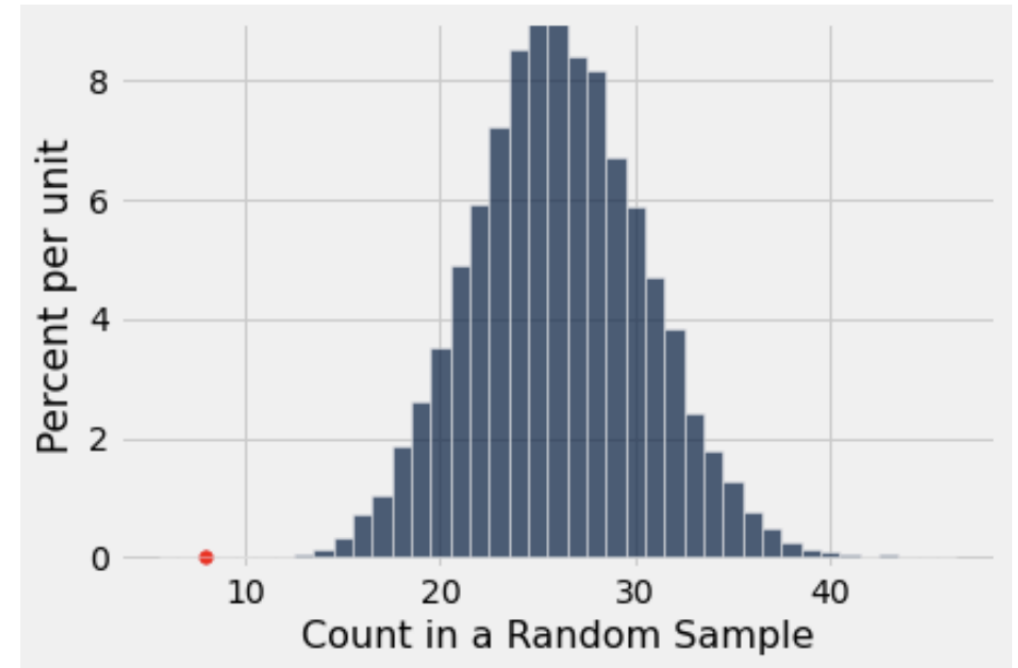


Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

• Case study: Jury selection

- USA Supreme Court case of Robert Swain, a Black man convicted in Talladega County, Alabama, in 1962.
- Robert Swain appealed his indictment on the grounds that Black people were systematically excluded from juries in Talladega County.
 - Amendment VI of the United States Constitution:
"In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed."
 - Data:
 - At the time of the trial, only men aged 21 or more were eligible to serve on juries in Talladega County.
 - In that population, 26% of the men were Black.
 - Only eight men among the panel of 100 men (that is, 8%) were Black.



Conclusion: *There's quantitative evidence of unfairness.*

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, *"All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court."*
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, *"All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court."*
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

Ethnicity	Eligible	Panels
Asian/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04

Visualization helps to understand data.

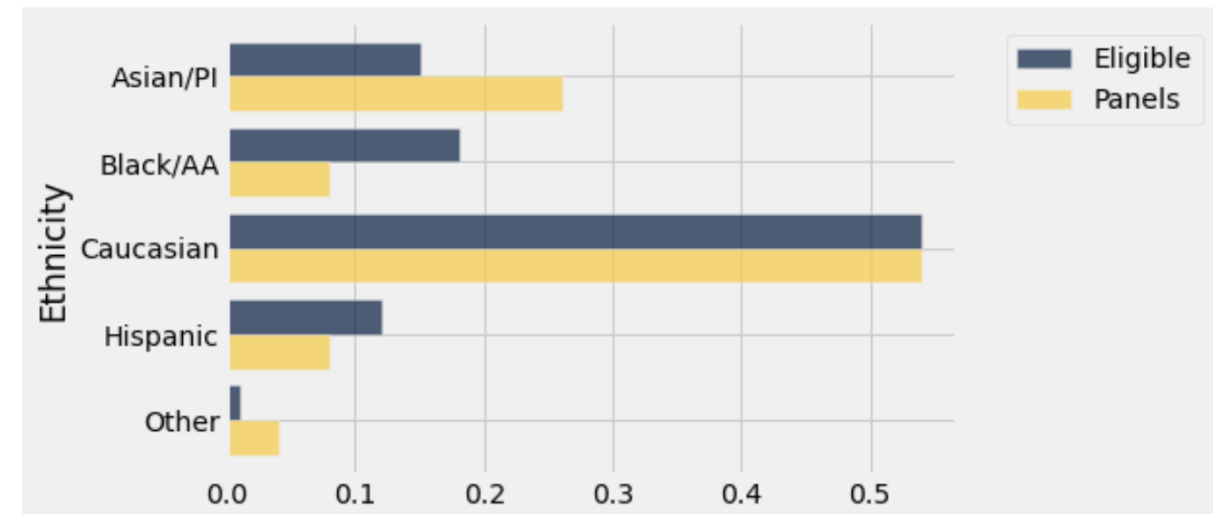
Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, *"All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court."*
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

Ethnicity	Eligible	Panels
Asian/PI	0.15	0.26
Black/AA	0.18	0.08
Caucasian	0.54	0.54
Hispanic	0.12	0.08
Other	0.01	0.04



Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, *"All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court."*
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

- Null hypothesis: Jury were selected at random
- Test statistic: Some quantitative measure of difference between two distributions:

- Total variation distance:

$$\frac{1}{2} \cdot \sum_x |P(x) - Q(x)|$$

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, *"All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court."*
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

- Null hypothesis: Jury were selected at random
- Test statistic: Some quantitative measure of difference between two distributions:
 - Total variation distance:
$$\frac{1}{2} \cdot \sum_x |P(x) - Q(x)|$$
- What is the observed value?

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

• Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, "*All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court.*"
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

• Null hypothesis: Jury were selected at random

• Test statistic: Some quantitative measure of difference between two distributions:

- Total variation distance (TVD):

$$\frac{1}{2} \cdot \sum_x |P(x) - Q(x)|$$

- What is the observed value? **0.14**
- What is the probability that a jury panel selected at random has TVD 0.14 or worse?

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

• Case study: Jury selection (Multiple categories)

- Section 197 of California's Code of Civil Procedure says, "*All persons selected for jury service shall be selected at **random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court.*"
- In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report on jury selection in Alameda County, California.
- Claim: Certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.
- Data: Demographic data for all the 1453 people who reported for jury duty (*from the composition of the jury panels in 11 felony trials in Alameda County in the years 2009 and 2010*).

• Null hypothesis: Jury were selected at random

• Test statistic: Some quantitative measure of difference between two distributions:

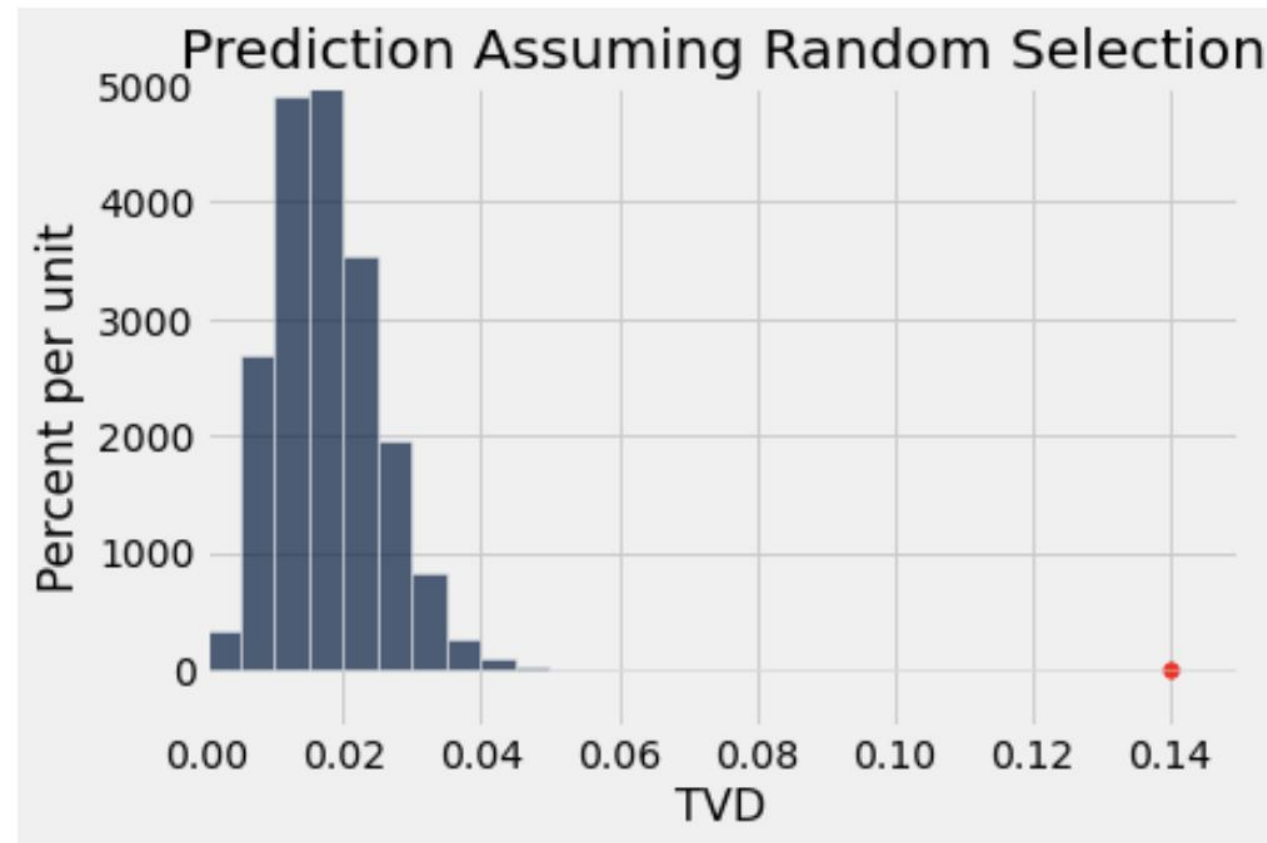
- Total variation distance (TVD):

$$\frac{1}{2} \cdot \sum_x |P(x) - Q(x)|$$

- What is the observed value? **0.14**
- What is the probability that a jury panel selected at random has TVD 0.14 or worse?
 - Simulation allows us to estimate this probability. Very low probability is grounds for rejection of null hypothesis.

Hypothesis Testing

- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.



- Null hypothesis: Jury were selected at random
- Test statistic: Some quantitative measure of difference between two distributions:
 - Total variation distance (TVD):
$$\frac{1}{2} \cdot \sum_x |P(x) - Q(x)|$$
- What is the observed value? **0.14**
- What is the probability that a jury panel selected at random has TVD 0.14 or worse?
 - Simulation allows us to estimate this probability. Very low probability is grounds for rejection of null hypothesis.

End