

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Hypothesis Testing

(Comparing two samples)

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.

- The previous study was *observational*. A similar comparative analysis for an *RCT* (*Randomized Controlled Trial*) can establish *causality*.
 - *Because of the unbiased assignment of individuals to the treatment and control groups, differences in the outcomes of the two groups can be ascribed to the treatment.*
- Case Study: A randomized controlled trial (RCT) examined the effect of using Botulinum Toxin A (BTA) as a treatment for lower back pain in adults.
 - **31 patients** with low-back pain were randomized into treatment and control groups, with **15 in the treatment group and 16 in control**. The control group was given normal saline (placebo), and the trials were run double-blind so that neither doctors nor patients knew which group they were in.
 - **Eight weeks after the start of the study, 9 of the 15 in the treatment group and 2 of the 16 in the control group had pain relief** (according to a precise definition used by the researchers).

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*

Group	Result average
Control	0.125
Treatment	0.6

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*
- Hypothesis testing main idea:
 - *Consider the possibility that among the 31 patients in the study, some were simply better able to recover than others, even without any help from the treatment.*
 - *What if an unusually large proportion of such patients got assigned to the treatment group, just by chance?*

Group	Result average
Control	0.125
Treatment	0.6

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*
- **Potential outcome** for a patient:
 - *Outcome if assigned to treatment group*
 - *Outcome if assigned to control group*
- There are 31 potential treatment outcomes and 31 potential control outcomes.
- Question: *Are the distributions same or different?*

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*
- **Potential outcome** for a patient:
 - *Outcome if assigned to treatment group*
 - *Outcome if assigned to control group*
- There are 31 potential treatment outcomes and 31 potential control outcomes.
- Question: *Are the distributions same or different?*
 - Issue: We do not get to see all the 62 potential outcomes, but only 31.

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*
- **Potential outcome** for a patient:
 - Outcome if assigned to treatment group
 - Outcome if assigned to control group
- There are 31 potential treatment outcomes and 31 potential control outcomes.
- Question: *Are the distributions same or different?*
 - Issue: We do not get to see all the 62 potential outcomes, but only 31.

Group	Outcome if assigned treatment	Outcome if assigned control
Control	Unknown	1
Control	Unknown	1
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Control	Unknown	0
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	1	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown
Treatment	0	Unknown

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- EDA: *In the treatment group, 60% of the patients had pain relief, compared to only 12.5% in the control group.*
- **Potential outcome** for a patient:
 - *Outcome if assigned to treatment group*
 - *Outcome if assigned to control group*
- **Hypothesis**:
 - **Null Hypothesis**: The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
 - **Alternative Hypothesis**: The distribution of 31 potential “treatment” outcomes is different from that of the 31 control outcomes. The treatment does something different from the control.

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: An RCT with 31 patients

- Treatment group: 9 out of 15
- Control group: 2 out of 16

- **Hypothesis:**

- **Null Hypothesis**: The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
- **Alternative Hypothesis**: The distribution of 31 potential “treatment” outcomes is different from that of the 31 control outcomes. The treatment does something different from the control.

- *There are 31 observed outcomes jointly in the two groups. If the null hypothesis were true, it wouldn't matter which of those 31 outcomes were labeled “treatment” and which “control.”*

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- **Hypothesis**:
 - Null Hypothesis: The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
- **Test statistic**: Absolute value of the difference of outcomes in the two groups.

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- **Hypothesis**:
 - **Null Hypothesis**: The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
- **Test statistic**: Absolute value of the difference of outcomes in the two groups.

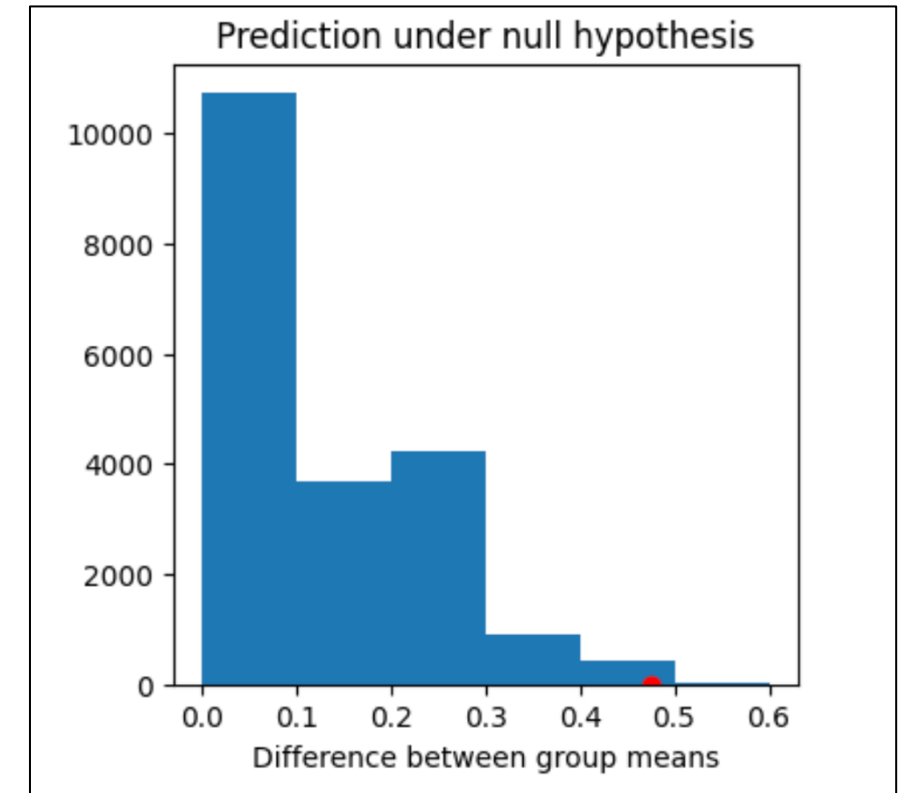
Group	Result average
Control	0.125
Treatment	0.6

$$\begin{aligned}\text{Observed value of test statistic} &= |0.6 - 0.125| \\ &= 0.475\end{aligned}$$

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: An RCT with 31 patients
 - Treatment group: 9 out of 15
 - Control group: 2 out of 16
- **Hypothesis**:
 - **Null Hypothesis**: The distribution of all 31 potential “treatment” outcomes is the same as that of all 31 potential “control” outcomes. Botulinum toxin A does nothing different from saline; the difference in the two samples is just due to chance.
- **Test statistic**: Absolute value of the difference of outcomes in the two groups.
- **Permutation test**: Shuffle the group labels randomly and calculate the difference.

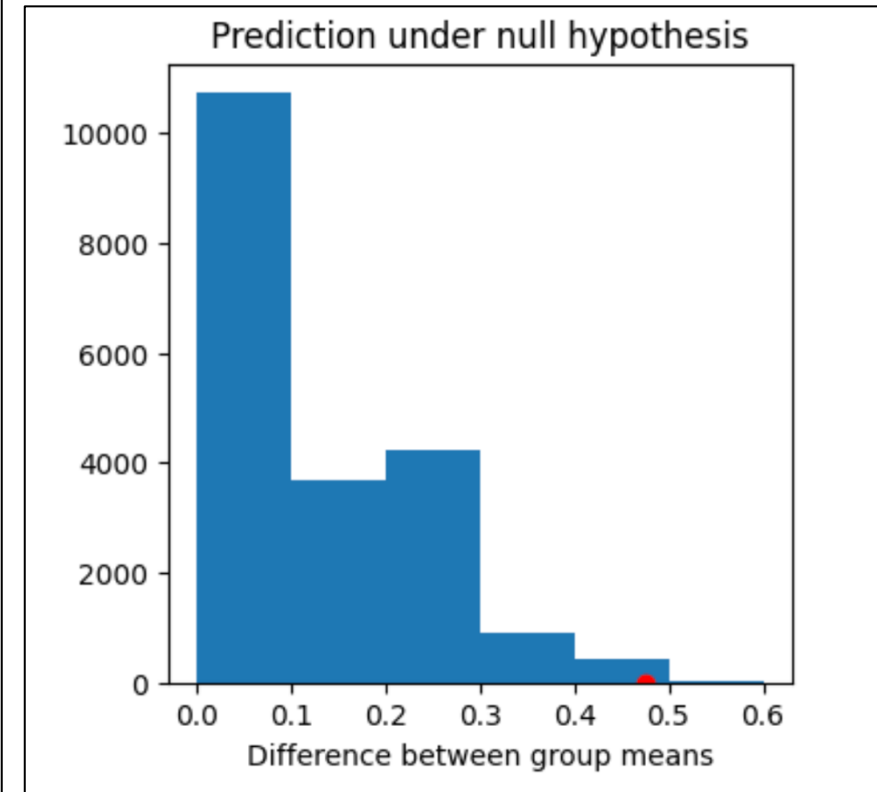


Empirical p-value = 0.97%

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.

- *Because the trials were randomized, the test is evidence that the treatment causes the difference. The random assignment of patients to the two groups ensures that there is no confounding variable that could affect the conclusion of causality.*
- *If the treatment had not been randomly assigned, our test would still point toward an association between the treatment and back pain outcomes among our 31 patients.*
- *Without randomization, this association would not imply that the treatment caused a change in back pain outcomes.*
 - *Example: if the patients themselves had chosen whether to administer the treatment, perhaps the patients experiencing more pain would be more likely to choose the treatment and more likely to experience some reduction in pain even without medication. Pre-existing pain would then be a confounding factor in the analysis.*



Empirical p-value = 0.97%

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: (Deflategate)

- *On January 18, 2015, the Indianapolis Colts and the New England Patriots played the American Football Conference (AFC) championship game to determine which of those teams would play in the Super Bowl. After the game, there were allegations that the Patriots' footballs had not been inflated as much as the regulations required; they were softer. This could be an advantage, as softer balls might be easier to catch.*
- *Pressure is often measured in pounds per square inch (psi). NFL rules stipulate that game balls must be inflated to have pressures in the range 12.5 psi and 13.5 psi. Each team plays with 12 balls. Teams have the responsibility of maintaining the pressure in their own footballs, but game officials inspect the balls. Before the start of the AFC game, all the **Patriots' balls were at about 12.5 psi**. Most of the **Colts' balls were at about 13.0 psi**.*
- *During the second quarter, the Colts intercepted a Patriots ball. On the sidelines, they measured the pressure of the ball and determined that it was below the 12.5 psi threshold. Promptly, they informed officials.*
- *At half-time, all the game balls were collected for inspection. Two officials, Clete Blakeman and Dyrrol Prioleau, measured the pressure in each of the balls.*

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: (Deflategate)

- *On January 18, 2015, the Indianapolis Colts and the New England Patriots played the American Football Conference (AFC) championship game to determine which of those teams would play in the Super Bowl. After the game, there were allegations that the Patriots' footballs had not been inflated as much as the regulations required; they were softer. This could be an advantage, as softer balls might be easier to catch.*
- *Pressure is often measured in pounds per square inch (psi). NFL rules stipulate that game balls must be inflated to have pressures in the range 12.5 psi and 13.5 psi. Each team plays with 12 balls. Teams have the responsibility of maintaining the pressure in their own footballs, but game officials inspect the balls. Before the start of the AFC game, all the **Patriots' balls were at about 12.5 psi**. Most of the **Colts' balls were at about 13.0 psi**.*
- *During the second quarter, the Colts intercepted a Patriots ball. On the sidelines, they measured the pressure of the ball and determined that it was below the 12.5 psi threshold. Promptly, they informed officials.*
- *At half-time, all the game balls were collected for inspection. Two officials, Clete Blakeman and Dyrrol Prioleau, measured the pressure in each of the balls.*

Team	Blakeman	Prioleau
Patriots	11.5	11.8
Patriots	10.85	11.2
Patriots	11.15	11.5
Patriots	10.7	11
Patriots	11.1	11.45
Patriots	11.6	11.95
Patriots	11.85	12.3
Patriots	11.1	11.55
Patriots	10.95	11.35
Patriots	10.5	10.9
Patriots	10.9	11.35
Colts	12.7	12.35
Colts	12.75	12.3
Colts	12.5	12.95
Colts	12.55	12.15

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: (Deflategate)

- EDA:

- *Average pressure drop of Colt vs. Patriot ball given that Patriot started with 12.5psi and Colt started with 13.0psi.*

Team	Combined	Pressure Drop
Patriots	11.65	0.85
Patriots	11.025	1.475
Patriots	11.325	1.175
Patriots	10.85	1.65
Patriots	11.275	1.225
Patriots	11.775	0.725
Patriots	12.075	0.425
Patriots	11.325	1.175
Patriots	11.15	1.35
Patriots	10.7	1.8
Patriots	11.125	1.375
Colts	12.525	0.475
Colts	12.525	0.475
Colts	12.725	0.275
Colts	12.35	0.65

Team	Pressure Drop average
Colts	0.46875
Patriots	1.20227

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: (Deflategate)
- EDA: *Average pressure drop*
- Null Hypothesis: *There is no difference in the distributions.*
- Test statistic:
 - average drop for Patriots - average drop for Colts

Team	Combined	Pressure Drop
Patriots	11.65	0.85
Patriots	11.025	1.475
Patriots	11.325	1.175
Patriots	10.85	1.65
Patriots	11.275	1.225
Patriots	11.775	0.725
Patriots	12.075	0.425
Patriots	11.325	1.175
Patriots	11.15	1.35
Patriots	10.7	1.8
Patriots	11.125	1.375
Colts	12.525	0.475
Colts	12.525	0.475
Colts	12.725	0.275
Colts	12.35	0.65

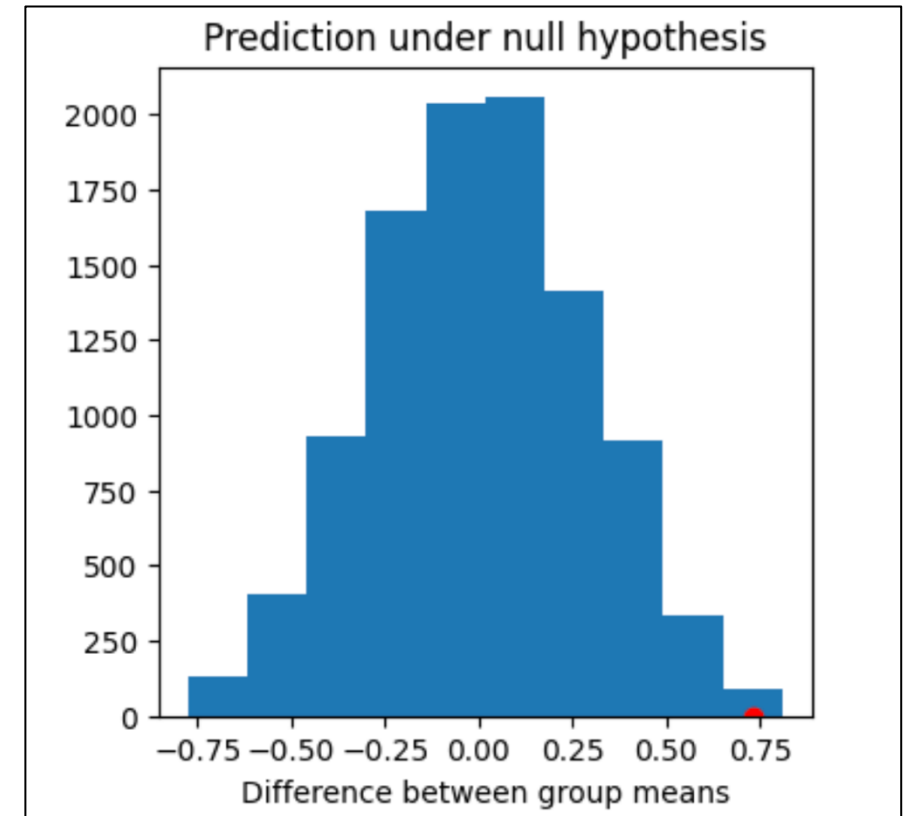
Team	Pressure Drop average
Colts	0.46875
Patriots	1.20227

Observed value = 0.73352

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: (Deflategate)
- EDA: *Average pressure drop*
- Null Hypothesis: *There is no difference in the distributions.*
- Test statistic:
 - average drop for Patriots - average drop for Colts
- Permutation test: Shuffle the Teams and calculate the test statistic.



Empirical p-value = 0.28%

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.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
- For every H_{0i} , it is rejected if the p-value is below the significance level α .

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- The **Family Wise Error Rate (FWER)** is defined as:

$$FWER = \Pr[V \geq 1]$$

Hypothesis Testing

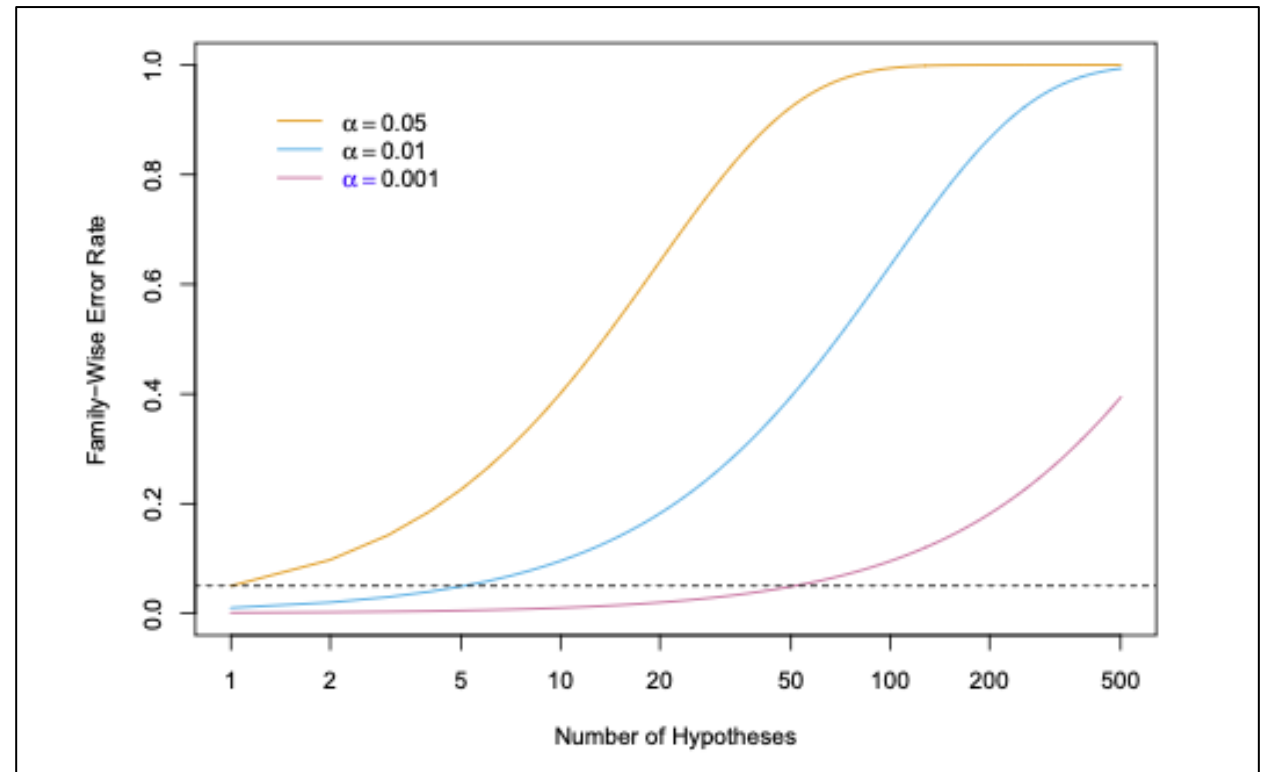
- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
- For every H_{0i} , it is rejected if the p-value is below the significance level α .

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- The **Family Wise Error Rate (FWER)** is defined as:
$$FWER(\alpha) = \Pr[V \geq 1] = 1 - \Pr[V = 0] = 1 - (1 - \alpha)^m \text{ (independence and all null)}$$



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.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
 - For every H_{0i} , it is rejected if the p-value is below the significance level α .
- *Family Wise Error Rate (FWER)*: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?

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.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
- For every H_{0i} , it is rejected if the p-value is below the significance level α .

- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$

- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?

$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] \quad (\text{from union bound}) \\ &= m \cdot \alpha \end{aligned}$$

- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.

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.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
- For every H_{0i} , it is rejected if the p-value is below the significance level α .

- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$

- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?

$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] && \text{(from union bound)} \\ &= m \cdot \alpha \end{aligned}$$

- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.

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.

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
 - For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] && \text{(from union bound)} \\ &= m \cdot \alpha \end{aligned}$$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.
- **Holm's step-down method**: Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \min\{j: p_j > \frac{\alpha}{m-j+1}\}$; reject all null hypothesis where $p_j \leq p_L$

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.

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypothesis $H_{01}, \dots, H_{0m}$
- For every H_{0i} , it is rejected if the p-value is below the significance level α .

- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$

- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?

$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] \quad (\text{from union bound}) \\ &= m \cdot \alpha \end{aligned}$$

- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.

- The Bonferroni method may be too conservative. So, lesser power.

- **Holm's step-down method**: Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \min\{j: p_j > \frac{\alpha}{m-j+1}\}$; reject all null hypothesis where $p_j \leq p_L$.

- Show that $\Pr[V \geq 1] \leq \alpha$.

End