

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Hypothesis Testing

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: Mendel's model

- Claim: For every plant, there is a 75% chance that it will have purple flowers, and a 25% chance that the flowers will be white, regardless of the colors in all the other plants.
- Data: To see whether his model was valid, Mendel grew 929 pea plants of this variety. Among these 929 plants, 705 had purple flowers.

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: Mendel's model

- Claim: For every plant, there is a 75% chance that it will have purple flowers, and a 25% chance that the flowers will be white, regardless of the colors in all the other plants.
- Data: To see whether his model was valid, Mendel grew 929 pea plants of this variety. Among these 929 plants, 705 had purple flowers.

- Null hypothesis:

- The word “null” reinforces the idea that if the data look different from what the null hypothesis predicts, the difference is due to *nothing* but chance.
- From a practical perspective, **the null hypothesis is a hypothesis under which you can simulate data.**
- In Mendel's model, then null hypothesis is that each plant has a 75% chance of having purple flowers, independent of all other plants.
- Alternate hypothesis: This says that some reason other than chance made the data differ from the predictions of the model in 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: Mendel's model

- Claim: For every plant, there is a 75% chance that it will have purple flowers, and a 25% chance that the flowers will be white, regardless of the colors in all the other plants.
- Data: To see whether his model was valid, Mendel grew 929 pea plants of this variety. Among these 929 plants, 705 had purple flowers.

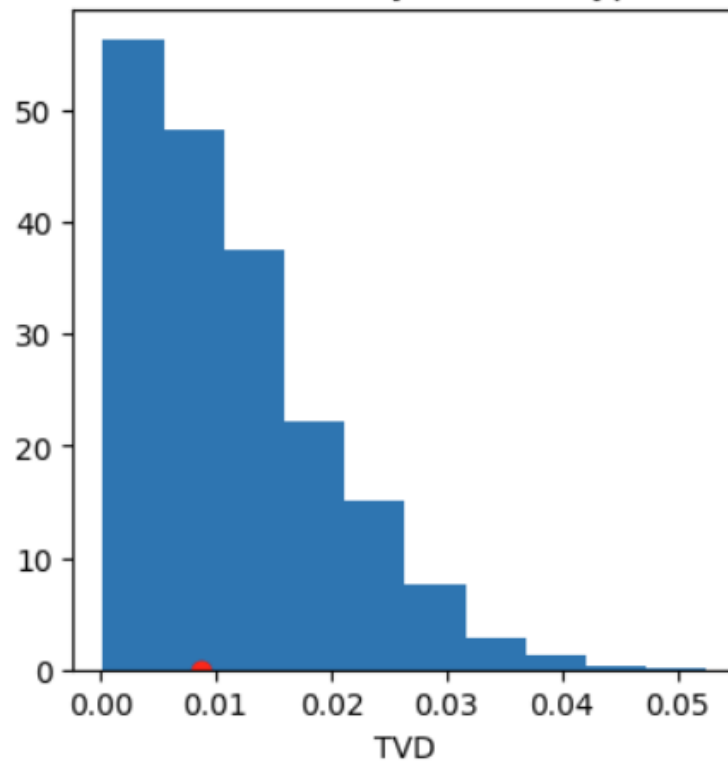
- Null hypothesis: Each plant has a 75% chance of having purple flowers, independent of all other plants.
- Test statistic: Total variation distance
- Observed value of test statistic: $\frac{705}{929} - 0.75 = 0.0088$

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: Mendel's

Prediction made by the null hypothesis



- Null hypothesis: Each plant has a 75% chance of having purple flowers, independent of all other plants.
- Test statistic: Total variation distance
- Observed value of test statistic: $\frac{705}{929} - 0.75 = 0.0088$
- The observed data is consistent with 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.

- When do we reject the null hypothesis?
 - When the observed test statistic (or more extreme value) is unlikely to be observed under the null hypothesis.
 - More extreme: Starting at the observed statistic and looking in the direction that makes us lean toward the alternative.
 - What should be the cut-off probability? Is there some conventional wisdom?
- **p-value**: The p-value of a test is the chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative.
 - How small is “small”? According to the conventions:
 - If the p-value is less than 5%, it is considered small and the result is called “statistically significant.”
 - *There’s a bit of history regarding the 5% cut-off. You can find this in [CIT] section 11.3.*
 - If the p-value is even smaller – less than 1% – the result is called “highly statistically significant.”

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.

- When do we reject the null hypothesis?
 - When the observed test statistic (or more extreme value) is unlikely to be observed under the null hypothesis.
 - More extreme: Starting at the observed statistic and looking in the direction that makes us lean toward the alternative.
 - What should be the cut-off probability? Is there some conventional wisdom?
- ***p-value***: The p-value of a test is the chance, based on the model in the null hypothesis, that the test statistic will be equal to the observed value in the sample or even further in the direction that supports the alternative.
 - How small is “small”? According to the conventions:
 - If the p-value is less than 5%, it is considered small and the result is called “statistically significant.”
 - If the p-value is even smaller – less than 1% – the result is called “highly statistically significant.”
 - ***If you use a $\alpha\%$ cutoff for the p-value, and the null hypothesis happens to be true, then there is about a $\alpha\%$ chance that your test will conclude that the alternative is true.***

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.

- It is important to note that there is a possibility of making a wrong conclusion. We must understand the reasons.
- If you are testing a null hypothesis against the alternative that the null hypothesis isn't true, then there are four ways of classifying reality and the result of the test.

		Truth	
		<i>Null hypothesis (H_0)</i>	<i>Alternate Hypothesis (H_a)</i>
Decision	<i>Reject H_0</i>	Type I error	Correct
	<i>Do not reject H_0</i>	Correct	Type II error

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.

- It is important to note that there is a possibility of making a wrong conclusion. We must understand the reasons.
- If you are testing a null hypothesis against the alternative that the null hypothesis isn't true, then there are four ways of classifying reality and the result of the test.

		Truth	
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)
Decision	Reject H_0	Type I error	Correct
	Do not reject H_0	Correct	Type II error

- Type I error: We erroneously reject the null hypothesis when in fact it is True. This happens due to chance variation when a low probability event happens.
- Type II error: We do not reject the null hypothesis when in fact it is false.

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.

		Truth	
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)
Decision	Reject H_0	Type I error	Correct
	Do not reject H_0	Correct	Type II error

- Type I error: We erroneously reject the null hypothesis when in fact it is True. This happens due to chance variation when a low probability event happens.
 - Type I error rate (α): The probability of rejecting H_0 when H_0 holds.
- Type II error: We do not reject the null hypothesis when in fact it is false.
 - Type II error rate (β): The probability of not rejecting H_0 when it is False.
 - Type II error power ($1 - \beta$): The probability of not making a type II error, i.e., the probability of correctly rejecting H_0 .

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.

		Truth	
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)
Decision	Reject H_0	Type I error	Correct
	Do not reject H_0	Correct	Type II error

- Type I error: We erroneously reject the null hypothesis when in fact it is True. This happens due to chance variation when a low probability event happens.
 - Type I error rate (α): The probability of rejecting H_0 when H_0 holds.
- Type II error: We do not reject the null hypothesis when in fact it is false.
 - Type II error rate (β): The probability of not rejecting H_0 when it is False.
 - Type II error power ($1 - \beta$): The probability of not making a type II error, i.e., the probability of correctly rejecting H_0 .
- The goal should be to reduce both errors, but there is a tradeoff. Setting the right significance level is important in hypothesis testing.

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: *Data Snooping and p-Hacking:*
 - Suppose we test multiple hypothesis using the same data. For example, in a randomized controlled trial about the effect of a drug, researchers might test whether the drug has an effect on various different diseases.
 - Suppose the drug has no effect on any disease. Just by chance, a small number of tests will conclude that the treatment has an effect and that one gets published. This is called data snooping or p-hacking.
 - If the researchers ran multiple different tests before finding one that gave a “*highly statistically significant*” result, use the result with caution. Replicate the experiment testing for that effect alone.

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.

- Until now we have seen examples where there is a single sample and we try to explain how it has been generated.
- Often, we have to compare two samples with each other.
 - Example: *Compare the outcomes of patients who have been assigned at random to a treatment group and a control group.*
- **A/B Testing**: Deciding whether two numerical samples come from the same underlying distribution. (*The name refers to the labels of the two samples, A and B*).
- We'll pick up the concepts using an example:
 - The table births contains the following variables for 1,174 mother-baby pairs.

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 table *births* contains the following variables for 1,174 mother-baby pairs.
- ***Aim:*** Try to understand whether maternal smoking was associated with birth weight.

Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight	Maternal Smoker
120	284	27	62	100	False
113	282	33	64	135	False
128	279	28	64	115	True
108	282	23	67	125	True
136	286	25	62	93	False
138	244	33	62	178	False
132	245	23	65	140	False
120	289	25	62	125	False
143	299	30	66	136	True
140	351	27	68	120	False
... (1164 rows omitted)					

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.

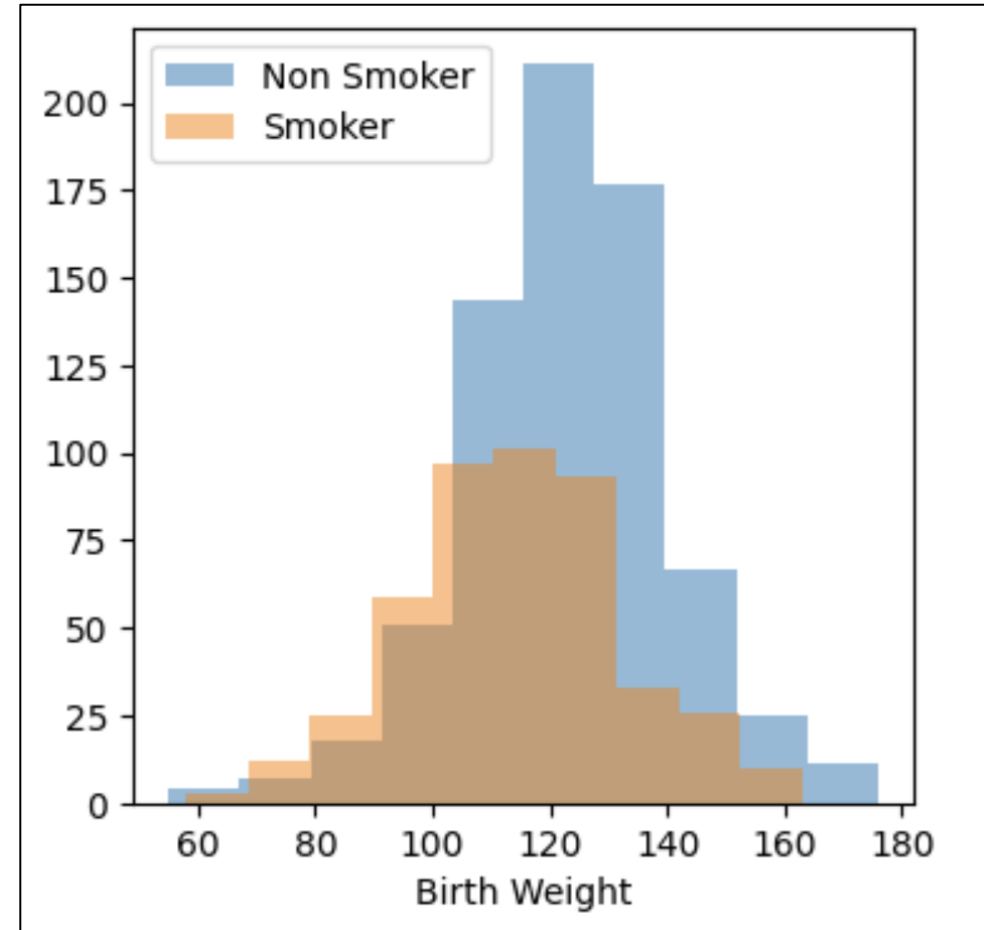
- **Aim**: *Try to understand whether maternal smoking was associated with birth weight.*
- Let's start with some EDA.
 1. **Count of smokers and non-smokers.**

Maternal Smoker	count
False	715
True	459

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.

- **Aim**: *Try to understand whether maternal smoking was associated with birth weight.*
- Let's start with some EDA.
 1. Count of smokers and non-smokers.
 2. **Distribution of birth weights.**
 - Observation: *There seems to be some difference.*



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.

- **Aim:** *Try to understand whether maternal smoking was associated with birth weight.*
- **Hypothesis:** *There is no underlying difference in the populations; the distributions in the samples are different just due to chance.*
 - **Null Hypothesis:** *In the population, the distribution of birth weights of babies is the same for mothers who don't smoke as for mothers who do. The difference in the sample is due to chance.*
 - **Alternate Hypothesis:** *In the population, the babies of the mothers who smoke have a lower birth weight, on average, than the babies of the non-smokers.*
- **Test Statistic:** *Difference in the mean.*
 - What is the **observed** difference?

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.

- **Aim**: *Try to understand whether maternal smoking was associated with birth weight.*
- **Hypothesis**: *There is no underlying difference in the populations; the distributions in the samples are different just due to chance.*
 - **Null Hypothesis**: *In the population, the distribution of birth weights of babies is the same for mothers who don't smoke as for mothers who do. The difference in the sample is due to chance.*
 - **Alternate Hypothesis**: *In the population, the babies of the mothers who smoke have a lower birth weight, on average, than the babies of the non-smokers.*
- **Test Statistic**: *Difference in the mean.*
 - What is the **observed** difference? (-9.266)

Maternal Smoker	Birth Weight average
False	123.085
True	113.819

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.

- **Aim**: *Try to understand whether maternal smoking was associated with birth weight.*
- **Hypothesis**: *There is no underlying difference in the populations; the distributions in the samples are different just due to chance.*
 - **Null Hypothesis**: *In the population, the distribution of birth weights of babies is the same for mothers who don't smoke as for mothers who do. The difference in the sample is due to chance.*
 - **Alternate Hypothesis**: *In the population, the babies of the mothers who smoke have a lower birth weight, on average, than the babies of the non-smokers.*
- **Test Statistic**: *Difference in the mean.*
 - What is the **observed** difference? (-9.266)
- How do we simulate this test statistic under the null hypothesis?

Hypothesis Testing

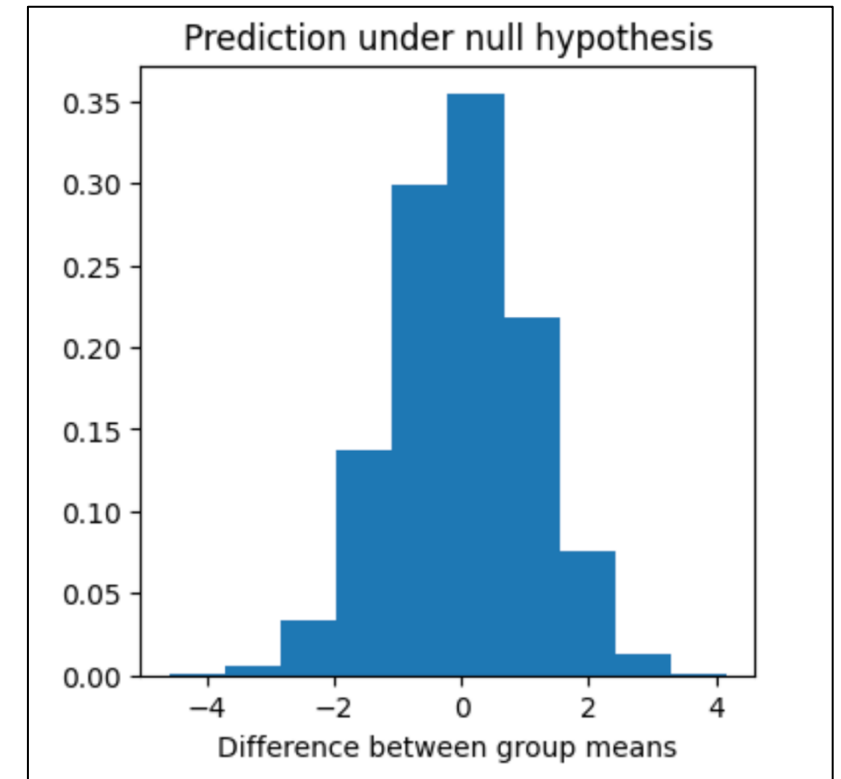
- Make a hypothesis
- The hypothesis imposes a model on the data
- Check if the data does not fit the model. If not, then reject the hypothesis, otherwise accept.

- **Aim:** *Try to understand whether maternal smoking was associated with birth weight.*
- **Hypothesis:** *There is no underlying difference in the populations; the distributions in the samples are different just due to chance.*
 - **Null Hypothesis:** *In the population, the distribution of birth weights of babies is the same for mothers who don't smoke as for mothers who do. The difference in the sample is due to chance.*
 - **Alternate Hypothesis:** *In the population, the babies of the mothers who smoke have a lower birth weight, on average, than the babies of the non-smokers.*
- **Test Statistic:** *Difference in the mean.*
 - What is the **observed** difference? **(-9.266)**
- How do we simulate this test statistic under the null hypothesis?
 - If there were no difference between the two distributions in the underlying population, then whether a birth weight has the label True or False with respect to maternal smoking should make no difference to the average.
 - **Idea:** Shuffle all the labels randomly among the mothers. This is called **random permutation**. Calculate the difference of the means of 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.

- **Aim:** Try to understand whether maternal smoking was associated with birth weight.
- **Hypothesis:** There is no underlying difference in the populations; the distributions in the samples are different just due to chance.
 - Null Hypothesis: In the population, the distribution of birth weights of babies is the same for mothers who don't smoke as for mothers who do. The difference in the sample is due to chance.
 - Alternate Hypothesis: In the population, the babies of the mothers who smoke have a lower birth weight, on average, than the babies of the non-smokers.
- **Test Statistic:** Difference in the mean.
 - What is the **observed** difference? (-9.266)
- How do we simulate this test statistic under the null hypothesis?
 - If there were no difference between the two distributions in the underlying population, then whether a birth weight has the label True or False with respect to maternal smoking should make no difference to the average.
 - Idea: Shuffle all the labels randomly among the mothers. This is called **random permutation**. Calculate the difference of the means of the two groups.



Empirical p -value = 0

End