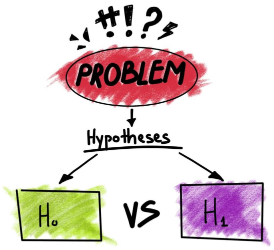


Intro to Hypothesis Testing

Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

1. Motivation
2. Hypothesis Testing
3. Null Hypothesis
4. Alternate Hypothesis
5. P-Value
6. Significance Level
7. Types of Errors
8. Tailed Tests

Motivation

Imagine you are a Data Scientist at YouTube

- Let's say YouTube runs **1 ad** every 15 min in a video
- Now, YouTube wants to run **2 ads** every 15 min

Question: What change would you expect among users if YouTube does this?

Case 1: No change in users subscription memberships or watch time.

Case 2:

a) People may get irritated with more ads

- They may spend less time on YouTube overall
- They might drop off and watch less YouTube videos
- This would be really bad for YouTube

b) Premium subscription memberships may increase

- More people may want to get premium subscriptions
- To remove all ads
- It would be a good thing for YouTube

Let's say, we assume Case 2 happens, but how do we know which of our assumption is **True?**

- We would need to test whether our assumptions/beliefs about those changes are **True or NOT**

Let's take another example

Now, think that you are a Data Scientist working at Flipkart

Let's say Flipkart wants to remove "Add to Cart" button from its e-commerce portal

> **Q. What change would you expect if this is implemented?**

Case 1: No impact on user experience or sales.

Case 2:

a) It may make the user experience better

- Removing the "Add to cart" button may make the checkout process easier
- Buying a product becomes a one-step process

b) It may decrease sales

- It may increase/decrease/have no impact on sales.

How can we check our assumptions and validate them?

- We need a way to test the effectiveness of changes we make in the system
- Before we roll out the changes

This is where Hypothesis Testing comes into the picture

- We need to test for consequences before we can release a change
- Test whether our assumptions/beliefs about those changes are True or NOT
- This becomes our motivation to learn Hypothesis Testing

What is Hypothesis Testing?

Consider the scenario of a coin toss in a cricket match

> **Is there some assumption that we need to make about the coin, before using it for toss?**

Default Assumption: Fair Coin

> **What is the process to check whether this assumption is true or false? / When will you say that assumption is wrong?**

The data should provide some evidence that it is not fair , in order to negate this assumption.

We can obtain data by tossing the coin again and again.

Suppose,

- We tossed a coin 100 times
- Got Heads 90 times
- Then, clearly the coin is biased
- So we negate the assumption.

Ideally the $P(Heads) \approx P(Tails)$, to support the assumption that coin is fair.

So, there are 2 things we start with:-

1. **Assumption**
    - An assumption that we have
2. **Data**
    - Data that will tell if that assumption is correct or negate it otherwise.

Let's start the process of collecting evidence, in a small number of trials.

Consider the following experiments:-

- We toss 10 times and get Heads 7 times
    - Is this possible, or should we be suspicious considering that coin is fair?
    - It can't be expected for a fair coin to always give Heads 5 times out of 10 trials.
    - Data is too less to conclude anything.
- We toss 100 times, and get Heads 70 times
    - Now, can we call it unfair for a fact? Or can it still be a fair coin?
- We toss 1000 times, and get Heads 700 times
    - What about now?

Note that the number of heads was 70% in all these cases.

The only thing that is different in these cases is the number of trials (data).

Ponder the question:

> **Is there a process to start deciding whether, X number of tosses are enough to decide if coin is fair or not?**

For example:

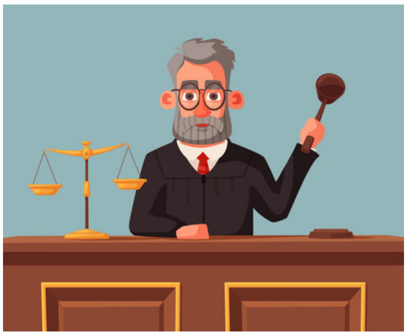
- If coin is tossed twiced, and we get head both times
- Can we simply conclude that coin is biased?

Naturally not.

But as and when we keep tossing 100s or 1000s or 100000s of times, the data becomes more and **more evident**

Hence, we get the intuition that perhaps we need a **quantifying metric**, that can help us determine whether our underlying assumption is true or not (i.e. if coin is fair or biased)

Judge in Court Example



Consider the scenario where a person is brought to court as a suspect for their hearing.

> **Q1. What must be the default assumption of judge about this suspect?**

Innocent until proven guilty.

> **Q2. What is the data in this context, that must be collected?**

Appropriate and adequate evidence, like

- Fingerprints at crime scene
- Murder weapon
- Forensic analysis

...etc

So, when there is enough data, we can pronounce guilty.

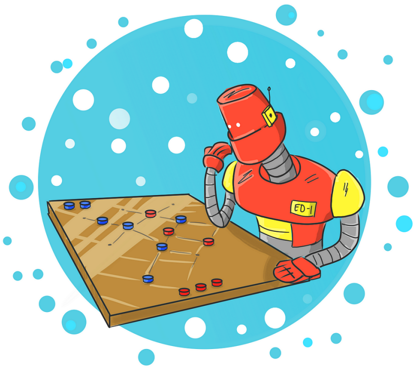
There are also cases where, criminals are released free, because of a lack of evidence.

In that case, we still stick to the default assumption until it is proved otherwise, with concrete evidence.

This example was good to give an intuition/idea that can be used to explain to laymen also.

But since we are studying in context to ML, let's look at a scenario where we may use Hypothesis Testing in ML.

ML Deployment Example [7-8 mins]



When you join a big company like Amazon, they will already have models in place

For example: A Recommendation System

- This will be called an Old / Legacy model

Now suppose you and your team have built a new model.

- When will the company feel confident in deploying your new model in place of the old model?
- They won't directly go and replace it (deploy)
- They will test the new model on a separate system server, without affecting the legacy model

In any company, before pushing a new model for deployment, there will be a product owner who will discuss with you the pros and cons of the legacy model, etc, to see if it is feasible.

> **What would be the default assumption of the product owner / manager?**

They would assume that the legacy model is better, as it is already deployed, and is working, and maintaining certain performance.

They would assume that there might be something wrong with the new model.

> **How will they determine which model is better?**

They would need to be convinced through **proof of concept** to see that new model is better.

- A proof of concept is a simple, initial test to see if an idea is worth pursuing further or if it needs adjustments before moving on to a larger, more expensive project.
- It's like dipping your toes in the water to see if it's warm before jumping into the pool.

> **What would proof of concept look like, in this context?**

Using the new model on a **lot of data**, to prove that it is better than legacy model, by using measures like:

- Performance metrics
- Edge case handling
- Change in sales numbers

...etc

Only then, will the product owner be able to confidently reject the base assumption, and say that new model is better.

Note:

- Here also, the burden of showing proof is on rejecting the assumption
- This means that if there is not enough evidence, you continue with the default assumption.

Third Umpire Example

Consider a Cricket match.

Suppose the on-field umpire has called the Third Umpire and asked for their help.

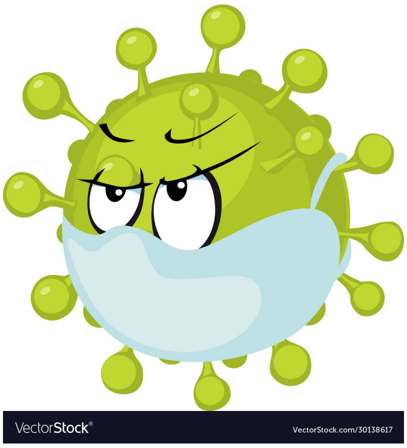
He has given a **soft signal**, meaning that he has given his decision, but is asking for the third umpire's opinion based on checking the camera footages closely, as he is in doubt.

> **Q. What would be the third umpire's default assumption?**

The on-field umpire is correct.

And you can only overturn (reject) their verdict if you have enough data.

Covid Virus Example



When we are performing a COVID test for an individual

> **What should be the default assumption?**

Does not have the virus.

If there is enough data against the assumption, we will say that he/she has the virus.

- Symptoms
- Test result

...etc

Based on the data, we may reject or accept the basic assumption.

So you can see that in the examples we've seen so far, there is a theme of

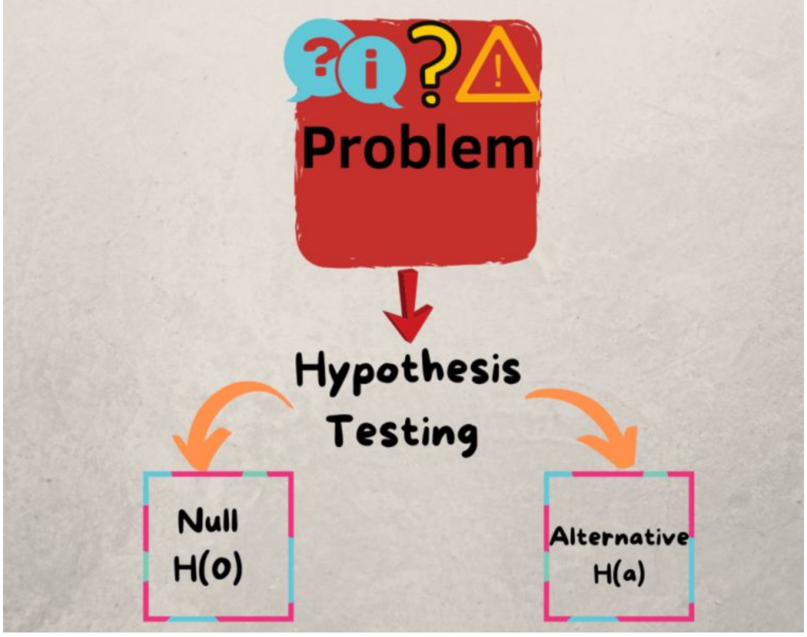
- Default assumption
- Data
- Rejection / acceptance of assumption

This exhibits the idea that in order to accept or reject a defult assumption, we need to conduct some kind of test.

This is known as **Hypothesis Test** that we talked about.

Let's explore some of it's key terminologies.

Hypothesis Testing Terminologies



Consider the Judge Example [5-7 mins]

When an accused is introduced in front of a judge, the judge doesn't know if the accused is guilty or not

The judge needs to evaluate everything before making a decision

There are 2 possibilities:

1. Accused is innocent
2. Accused is guilty

These are the 2 hypothesis

- A hypothesis is a possibility or assumption of an outcome
- Generally, we have 2 types of hypothesis

1. Null Hypothesis (H_0)

- A Null Hypothesis is like a baseline assumption which we can attempt to disprove
- Here, we assume that any deviation from expected data, is due to chance.

In our Judge Example:

- Before even the trial begins, the judge assumes that the `Accused is innocent`
- This is the judge's Null Hypothesis

Why does the judge assumes that accused is innocent by default?

- Law says "Accused is Innocent till proven Guilty"
- Null Hypothesis is like a default assumption

It is denoted as H_0.

For Ex, in the ML Deployment scenario,

- H_0: Legacy model performs better than the new model.

2. Alternate Hypothesis (H_a)

- Any assumption other than the Null Hypothesis
- It is something we want to prove by disproving the Null Hypothesis
- Therefore, the Alternate Hypothesis becomes: `Accused is guilty`

It is represented as H_a.

For Ex, in the ML Deployment scenario,

- if H_0 is rejected, then we can say the new model performs better than the legacy model.
- This becomes the H_a.

> **How does Hypothesis Testing help the judge?**

- In order to test the null hypothesis, the judge will seek evidence.
- If there is enough evidence, data to conclude that the default assumption was incorrect, the judge will **reject the null hypothesis**.
  - Then that means there is an Alternate hypothesis that is true
- If evidence is NOT enough, the judge will stick to it.

This is **Hypothesis Testing**

- It tells us evaluate whether any deviation in observed effect is significant, or due to random chance.
- If there is sufficient evidence, we reject the Null Hypothesis.
- Otherwise, we stick to it.

3. P-Value

Suppose we have the following data/evidence:-

- **Evidence 1:** Suspect has a knife in their pocket
  - Not enough evidence to say he is guilty
  - Innocent people can also have a knife
- **Evidence 2:** Knife has blood stains on it
  - Not enough evidence.
  - Chef, working with meat / maybe cut himself
- **Evidence 3:** Blood matches victim
  - Perhaps this is too much to be coincidence

    - But still could be planting of evidence
- **Evidence 4:** Suspect fingerprints found on victim
    - Enough evidence
- **Evidence 5:** CCTV catches suspect at scene of crime / Eyewitness

    - Enough evidence

  If we observe all these evidences, then we can reject H_0, because an innocent person is highly unlikely to have this many coincidences.

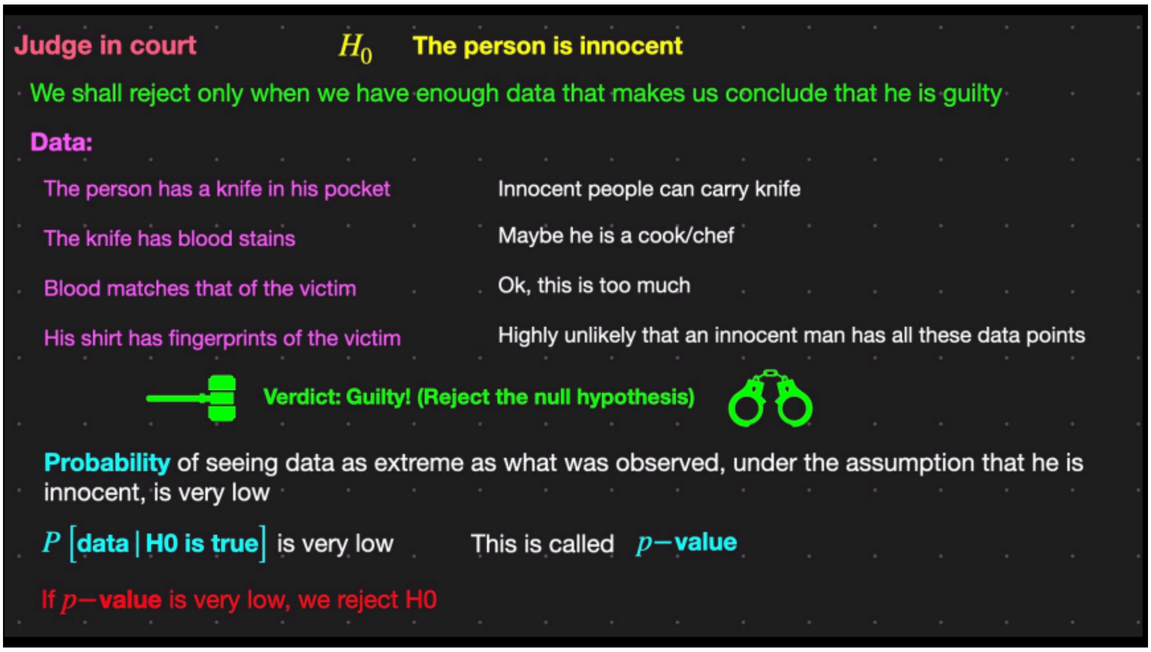
Another way of looking at it is:

`Probability of seeing data as extreme as this, is very low, under the assumption of Null Hypothesis (Innocence)`

- In other words, assuming someone is innocent, the probability of seeing these data is very low.

This is known as **p-value**, and can be written as: $P(data|H_0)$

Note that this is the probability of observing the given evidence, **conditioned on** null hypothesis.



Judge in court H_0 The person is innocent

We shall reject only when we have enough data that makes us conclude that he is guilty

Data:

The person has a knife in his pocket	Innocent people can carry knife
The knife has blood stains	Maybe he is a cook/chef
Blood matches that of the victim	Ok, this is too much
His shirt has fingerprints of the victim	Highly unlikely that an innocent man has all these data points

Verdict: Guilty! (Reject the null hypothesis)

Probability of seeing data as extreme as what was observed, under the assumption that he is innocent, is very low

P [data | H0 is true] is very low   This is called $p-$value

If $p-$value is very low, we reject H0

4. Confidence/Significance Level [10 mins]

Now, Let's say we have 100 witnesses in this Trial Court case

- 50 witnesses say the accused is guilty
- 50 witnesses say the accused is innocent

Question: If you are the judge, what will be your decision?

> Innocent

Now, Let's change this scenario a bit

- 75 witnesses say the accused is guilty
- Remaining 25 witnesses say the accused is innocent

Question: Now what will be your decision?

> Guilty?

- No, it will still be innocent

Purposefully did NOT complete the law statement

"Accused is Innocent till proven Guilty... beyond a reasonable doubt"

Now, What is this reasonable doubt?

- Even though 75% of witnesses say the accused is guilty
- There is still a 25% chance that the accused is innocent
- We might be punishing an innocent person with a 25% chance
- We still don't have reasonable confidence that the accused is guilty

This is where the concept of "**Confidence**" & "**Significance Level**" comes into picture

Before we analyze a hypothetical scenario, we need to set a Confidence Level

- It's a minimum criteria based on which we reject or NOT reject the Null Hypothesis
- When our confidence is beyond a threshold, only then do we go with the Alternate Hypothesis
- In all the other cases, we stick with the Null Hypothesis

Question. When can we reject the Null hypothesis? When the value of p-value is high or low?

`pvalue is defined as the Probability of obtaining results at least as extreme as the observed results, under the assumption that the Null Hypothesis (Innocence) is true`

Therefore, **if p-value is low -> Reject H_0**

This is very important to note.

When using in applications, we compare the p-value with a **threshold value (α)**, and if it is less than that, then we reject the H_0.

This threshold is known as **significance level**.

Though different applications may have different value for significance level, if not specified, we use

$\alpha = 0.05$

The judge sets a confidence level of 95%

- Confidence level $(1-\alpha) = 1 - 0.05 = 0.95$
- It means only when 95 out of 100 witnesses say that the accused is guilty, then only the judge will reject the Null Hypothesis of the accused being innocent
- The judge will accept the Alternate Hypothesis of the accused being guilty

- Otherwise, judge will not reject the Null Hypothesis

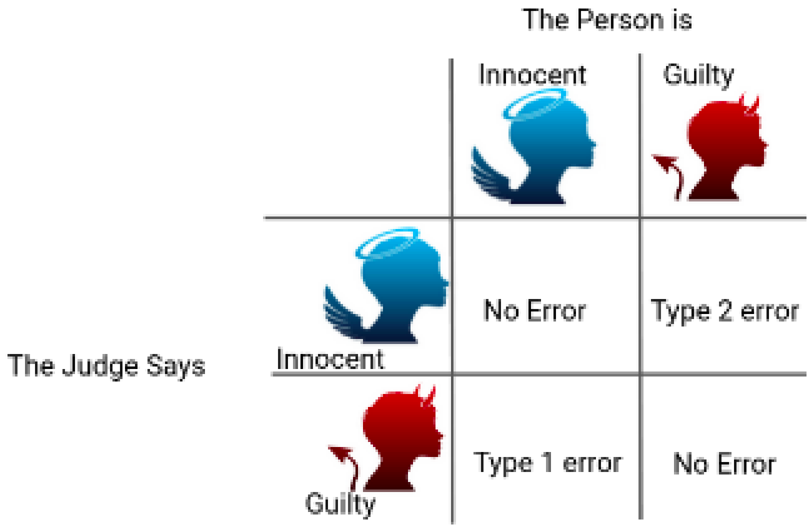
Types of Errors

Let's consider the same judge example.

But why do we need such high confidence?

- To avoid making a critical error

Based on what evidence we get, and what the judge decides, One of the 4 cases can happen:



The Judge Says		The Person is Innocent	The Person is Guilty
	Innocent	No Error	Type 2 error
	Guilty	Type 1 error	No Error

There are 2 possible results to a Hypothesis Test:

- We reject H0
- We fail to reject H0

Indicating that the ultimate objective is to try and check if we can reject the H0. Hence, we regard

- Rejecting H0 (calling them guilty) as **Positive**, and
- Failing to reject H0 (calling them innocent) as **Negative**.

Case 1: We decide that the accused is innocent, and he is actually innocent (i.e. **True Negative**)

Case 2: We decide that the accused is guilty, but he is actually innocent

Case 3: We decide that the accused is innocent, but he is actually guilty

Case 4: We decide that the accused is guilty, and he is actually guilty (i.e. **True Positive**)

Note:

- Our decision is correct in Case 1 and Case 4
- We are making an error in decisions in Case 2 and Case 3
- The above matrix is called **Confusion Matrix**

Case 2 is a Type - I Error

- Also known as **False Positive**
- We call the probability of making a Type-I Error as **alpha (α)**

Case 3 is a Type - II Error

- Also known as **False Negative**
- We call the probability of making a Type-II Error as **beta (β)**

Now, think about **Type - I error**

- We are charging the accused as guilty when he is actually innocent
- We are rejecting the Null Hypothesis when it is actually TRUE

The relationship between Significance Level(α) & Type I Error(α) :

1.**Significance Level:** The significance level is the predetermined threshold at which you decide whether to reject the null hypothesis.

2.**Type I Error:** When performing a hypothesis test, a Type I error occurs when you reject a true null hypothesis.

- In other words, it's the probability of observing a significant result (rejecting the null hypothesis) when there is no real effect or difference.

- The chosen significance level sets the boundary for the acceptable probability of making a Type I error.
- If you set (α = 0.05), you're saying you're willing to accept a 5% chance of incorrectly rejecting a true null hypothesis.

    - If the p-value (the probability of obtaining the observed results under the assumption that the null hypothesis is true) is less than or equal to α, you reject the null hypothesis.
- Therefore, **to decrease the Type I error rate, we can choose a lower significance level.**

    - α value, such as 0.01 instead of 0.05.
    - This makes the criteria for rejecting the null hypothesis stricter and reduces the likelihood of making a Type I error.

But why don't we control Type-II Error as well?

- NOT Rejecting the Null Hypothesis when it is actually FALSE is critical too.

- We'll talk about that in the next lecture when we discuss the Power of Test.

Errors while conducting a Covid Virus Test

Suppose you are conducting a test for COVID virus.

> **What would be your hypothesis?**
>
> - Null Hypothesis H_0: No Virus
> - Alternate Hypothesis H_a: Has COVID virus

So, in reality, there are only 2 possibilities:

- Person has virus
- Person does not have virus

Based on the evidence or data collected, we will either accept (no virus) or reject (has virus) the null hypothesis to make a final decision / result.

This gives us 4 possible cases:-

- Case A
    - Does not have virus in reality
    - Decision is also that no virus
- Case B
    - Does not have the virus in reality
    - Decision is that person has virus
- Case C
    - Person has the virus in reality
    - The decision is that no virus
- Case D
    - Person has the virus in reality
    - Decision is also that they have a virus

> **Which of these 4 cases are errors?**

B, C.

- In B, the test result/decision is +ve (has virus), but it is an error, so we call it **False Positive**
  - This is also known as **Type 1 Error**
- In C, the test result/decision is -ve (does not have virus), but it is an error, so we call it **False Negative**
  - This is also known as **Type 2 Error**

> **What would be the appropriate names for cases A and D?**

- A: True Negative
- D: True Positive

Tailed Tests

Left-Tailed Test

We saw Null and Alternate Hypothesis and also that the burden of proof lies on Alternate.

Let's consider another example

Burger Example

> Suppose there is a burger place that claims that all their burgers are **200 grams**.
> A customer who consumed their burger is still hungry after eating, and wants to prove that their burgers are lighter, and not as much as promised.

> **Q1. What will be H_0 and H_a?**

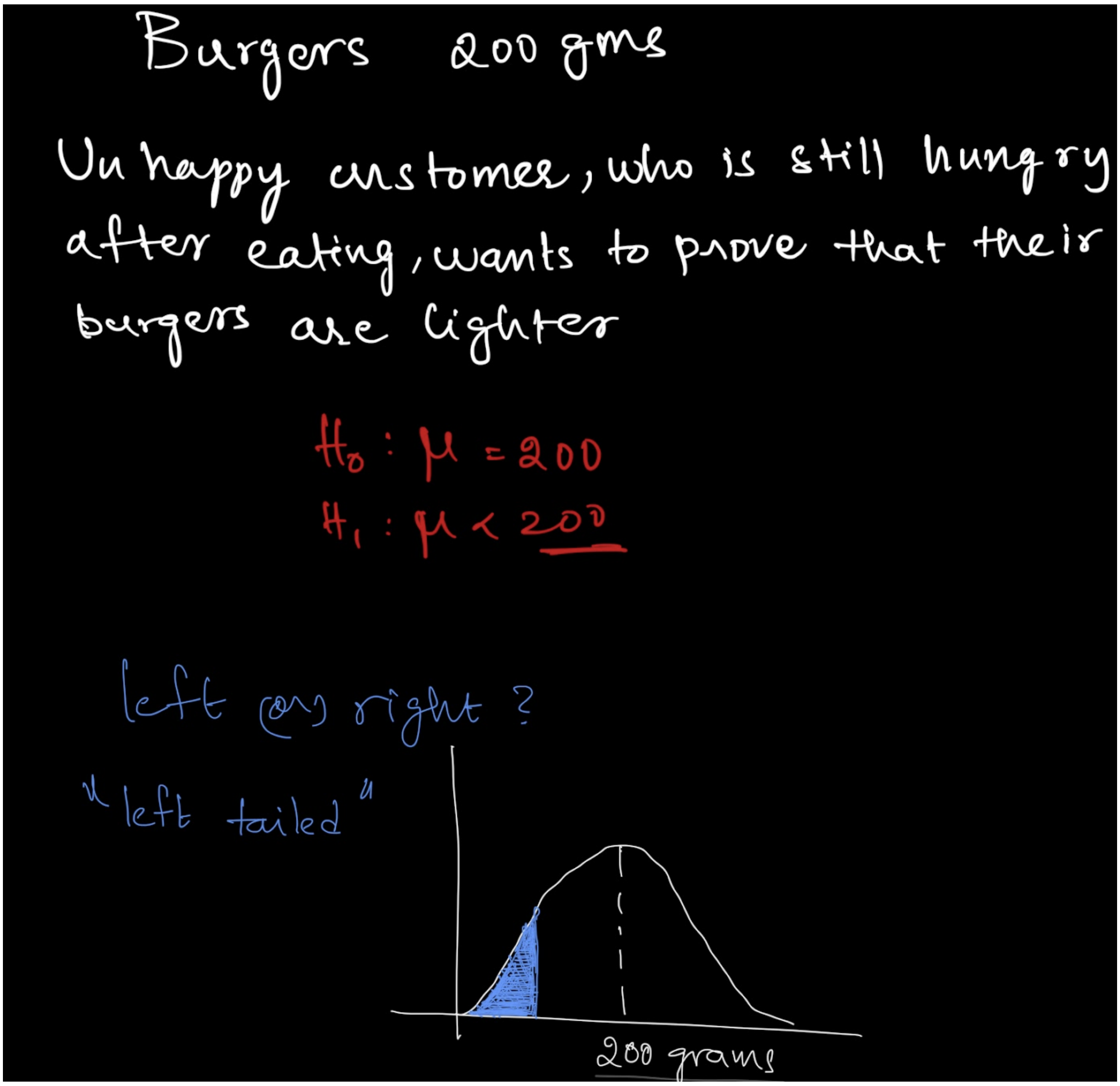
- H_0: Average burger weights 200g ($\mu = 200$)
- H_a: Average burger weight is less than 200g ($\mu < 200$)

So, the burden of proof lies on the person wanting to prove H_a

> **Q2. In the alternate that this customer is setting, is the evidence on the left side or the right?**

Left side.

For the given null and alternate hypothesis, the test we want to do is called the **left-tailed test**, since we want to show the average weight of the burger to be on the left side of the claimed weight (less than).



Burgers 200 gms

Un happy customer, who is still hungry after eating, wants to prove that their burgers are lighter

$H_0 : \mu = 200$

$H_1 : \mu < 200$

left (or) right ?

" left tailed "

Right-Tailed Test

> Consider the example of the Legacy Model, which had an accuracy of 90%. You want to claim that your new model is better.

So, naturally, the burden of proof is on you.

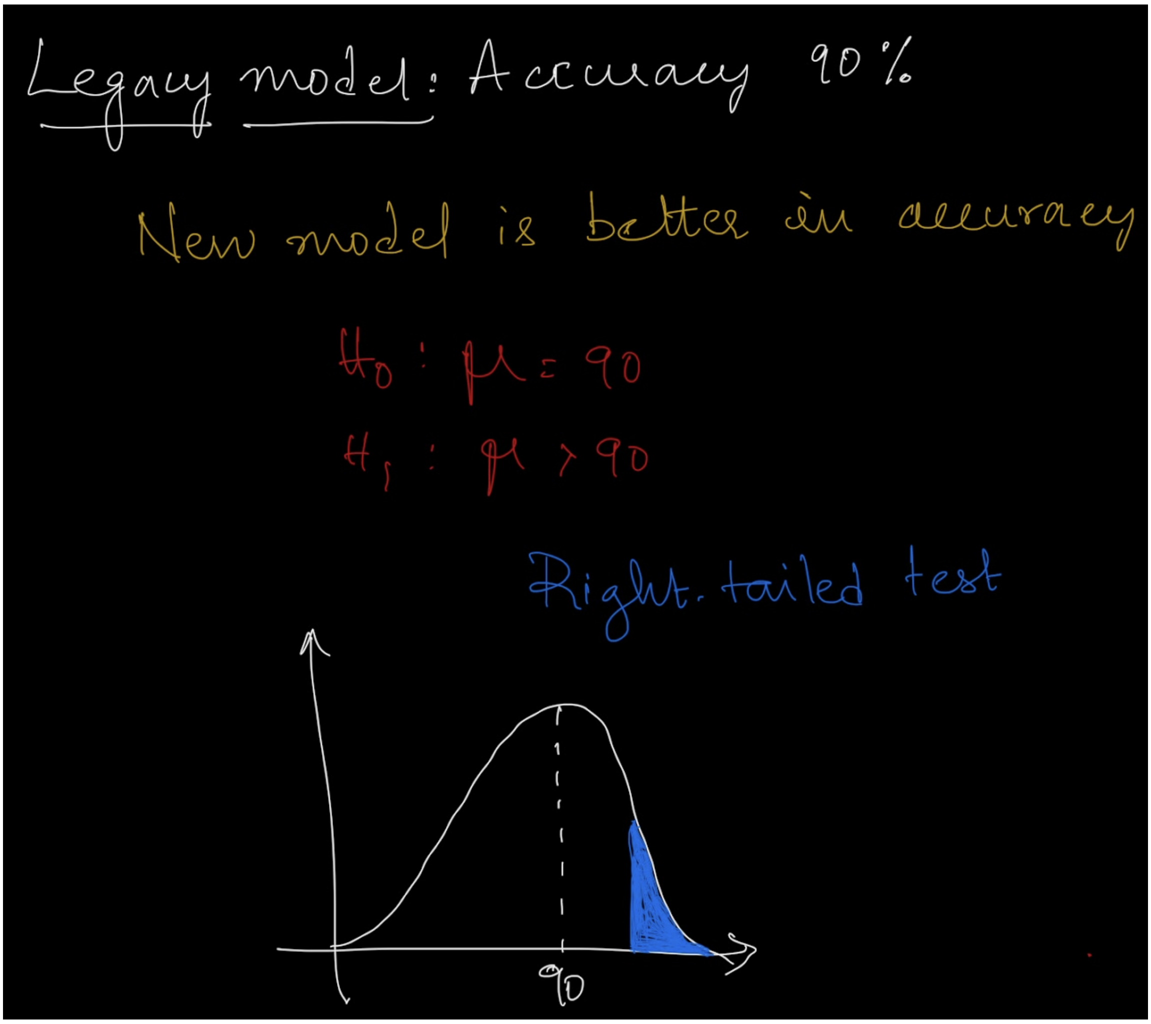
> **What will be the hypothesis?**

- H_0: $\mu = 90$
- H_a: $\mu > 90$

> **Here, which direction do you want to show your evidence in?**

Right side.

So, you'd need to perform the Right tailed test.



Legacy model: Accuracy 90%

New model is better in accuracy

$H_0 : \mu = 90$

$H_1 : \mu > 90$

Right-tailed test

Two-Tailed Test

> Suppose you are looking at the height of people in India. It is believed that the average height of Indians is 65 inches. You want to find out if that holds true for the people of your state. Are they taller or shorter?
>
> **What will be the hypothesis?**

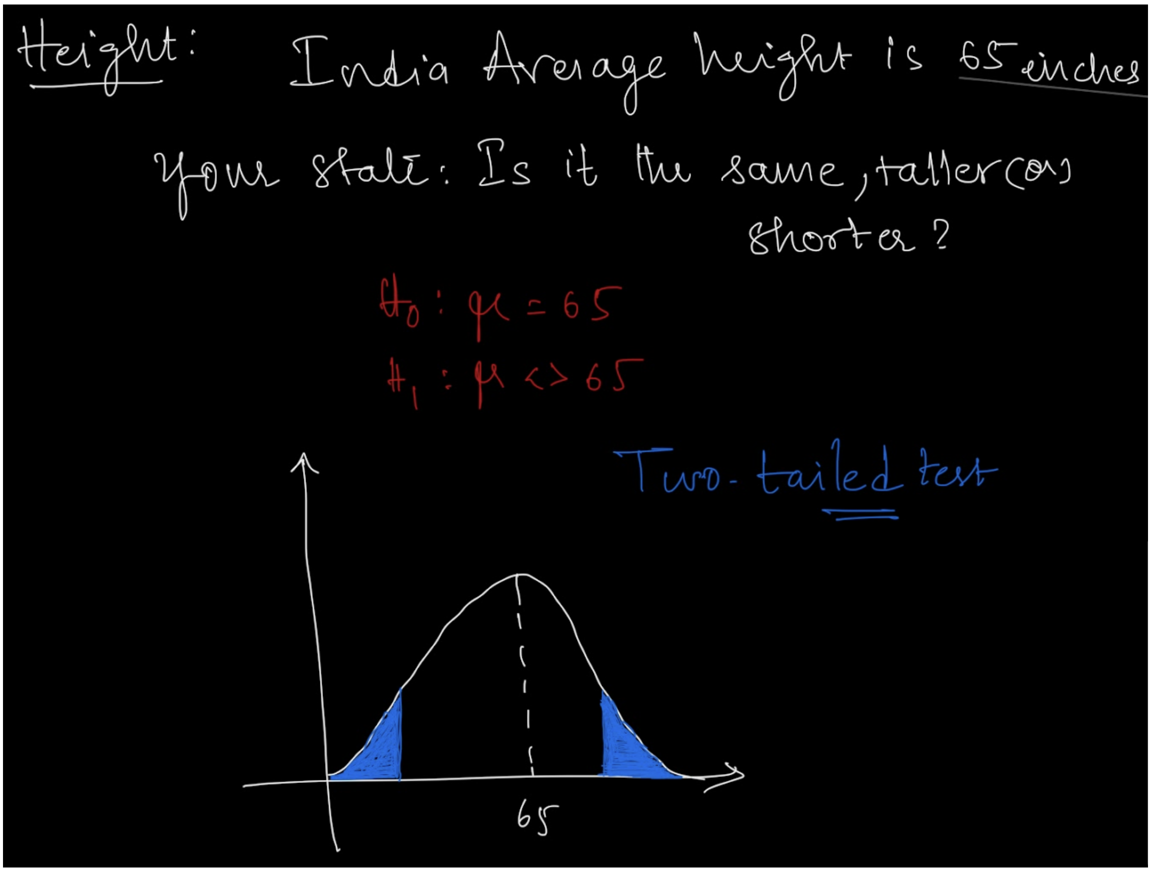
- H_0: $\mu = 65$
- H_a: $\mu \neq 65$

Since we just want to see if it is 65 or not, we don't care if it is greater or lesser, just should not be equal.

> **Here, which direction do you want to show your evidence in?**

Both.

Hence you'd need to perform a Two-Tailed Test.


Height:
India Average height is 65 inches
Your state: Is it the same, taller (or) shorter?
$H_0 : \mu = 65$
$H_1 : \mu <> 65$
Two-tailed test
65

Deep Dive in the Coin Toss Example

Recall that our default assumption was:

- H_0: Fair coin

We had the following observations:

Scenario 1: We toss 10 times, and get Heads 7 times

> **If it was a fair coin, then what distribution would this follow?**

Binomial Distribution.

When we are computing the p-value, we want to look at the data believing that the H_0 is true.

> **What is the Alternate hypothesis we're trying to prove/check?**

IF the coin is biased towards heads

- $H_a : P(heads) > 0.5$

Therefore,

> **Which tailed test do we need to perform?**

Right tailed test

- Since we're looking for more heads than expected with a fair coin, we need a right-tailed test.
  - This means we'll check how probable it is to get 7, 8, 9, or even 10 heads if the coin is fair.

i.e. $P(No\ of\ heads = 7\ or\ 8\ or\ 9\ or\ 10 | Fair\ Coin)$

This is the **P-Value**.

> **How can we calculate this? Do we use pmf or cdf or pdf?**

`1 - binom.cdf(k=6, n=10, p=0.5)`

- `p=0.5` because we are looking at it under the assumption H_0 is true (i.e. fair coin)

In [ ]:

```
from scipy.stats import binom

1 - binom.cdf(k=6, n=10, p=0.5)
```

Out[ ]:

```
0.171875
```

In [ ]:

```
binom.pmf(k=7, n=10, p=0.5) + binom.pmf(k=8, n=10, p=0.5) + binom.pmf(k=9, n
```

Out[ ]:

```
0.17187500000000003
```

> **When would we say that coin is not fair?**

When the pvalue is less than significance level (0.05)

- Since $0.171875 > 0.05$, we **fail to reject the Null Hypothesis.**

Scenario 2: We toss 100 times, and get Heads 70 times

The null and alternate hypothesis still remain the same for this case.

> **Which test do we need to perform?**

- Now our p-value will be obtained by evaluating the cummulative probability to the right of what was observed which represents the probability of getting 70 or more heads with a fair coin.

i.e. $P(No\ of\ heads = 70\ or\ 71\ or\ 72\ .....\ or\ 100 | Fair\ Coin)$

Hence **Right-tailed test**.

- We're interested in finding out how likely it is to get more heads than expected under the assumption of a fair coin.

In [ ]:

```
print(1 - binom.cdf(k=69, n=100, p=0.5))
print("Since this pvalue is less than significance level, we reject the Nul
3.925069822796612e-05
Since this pvalue is less than significance level, we reject the Null Hypot
hesis.
```

The more tosses we do (bigger sample size), the clearer the picture gets.

- With just a few tosses, even a biased coin could fool us by showing mostly heads by chance.
- But as we keep flipping, the true nature of the coin starts to shine through.

Hypothesis Testing Framework

So, now we have a framework to compute a quantifiable metric that will help us decide if we should accept or reject our null hypothesis.

We start any Hypothesis-testing problem with 2 things:

- Assumption
- Data

Let's summarise the framework into steps:-

1. Setup Null and Alternate Hypothesis
2. Choose the distribution (Gaussian, Binomial, etc)
3. Select the Left vs Right vs Two-Tailed test, as per the hypothesis
4. Compute p-value
5. Compare the p-value to the significance level (α) and accept/reject the Null Hypothesis accordingly.