Hypothesis Testing

1

- Definitions + Terminologies
- Framework
- Questions

Default Assumption AND Alternative Hypothesis – H_A

(Null hypothesis – H_0)

Machine Learning Deployment

A machine learning model (legacy) is in production for a few years, and is doing fairly well You and your team have built a new model, and want to claim that it is better

1) What is the default assumption of the product owner? → H_0

The new model is not better than the legacy model

2) When shall we reject this assumption? → H_A

When enough data is given that the new model outperforms the legacy model significantly

Radar example

A radar has to detect a plane

1) What should the default assumption be? → H_0

There is no plane

2) When should the default assumption be rejected?

The default assumption should be rejected only when the data is very conclusive that there is a plane

Judge in Court

H_0 → Not guilty H_A → Guilty

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

Data:

The person has a knife in his pocket
The knife has blood stains
Blood matches that of the victim
His shirt has fingerprints of the victim

It's Highly unlikely that an innocent man has all these data points

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

$P[\text{data} \mid H_0 \text{ is True}]$ is very low

↓

p-value

- If p-value is very low
  We reject H_0 in favour of H_A

- But how much is very low → Significance level (α)

  If p-value $< \alpha$ ⟶ Reject H_0

Hypothesis Testing Setups :

The average height of Indians is 65 inches

- Collected sample heights from your state → (take Average)

(i) Test if your state avg. is greater than population avg.

$H_0 : \mu = 65$

$H_A : \mu > 65$

↳ Right tailed

(ii) Test if your state avg. is less than population avg.

$H_0 : \mu = 65$

$H_A : \mu < 65$

↳ left tailed test

(iii) Test if your state avg. is diff. than population avg.

$H_0 : \mu = 65$

$H_A : \mu \neq 65$

↳ Two Tail

Framework of Hypothesis Testing :

1. Setup Null and Alternative hypothesis ✓
   (H_0) (H_A)
2. Choose the right Test statistic ✓ → point value
3. Setup : Left / Right / Two tailed
4. Compute p-value
5. If · p-value $< \alpha$ → Reject H_0
   · p-value $> \alpha$ → failed to Reject H_0

A snack-food company produces a 450 g bag of pretzels. Although the actual net weights deviate slightly from 450 g and vary from one bag to another, the company insists that the mean net weight of the bags be 450 g and Standard deviation is 3 grams.

As part of its program, the quality assurance department periodically performs a hypothesis test to decide whether the packaging machine is working properly, that is, to decide whether the mean net weight of all bags packaged is 450 g.

They wants to ensure 99% confidence in their every hypothesis test to arrive at statistically significant conclusion.

– Team measures avg. bag weight – collects 30 samples → mean

CASE I: · QA team wants to perform test

↓

Is machine OVERFILLING. ?

Mean wt. of 30 Bags = 452 gm

i) H_0 : $\mu = 450$ gm

H_A : $\mu > 450$ gm ✓

ii) Test statistic

f^m data → mean wt = 452gm. = $\bar{m}$

- $\bar{m}$ is a Random Variable ✓

why? → If we take another random sample of 30 bags → mean wt. changes

∴ $\bar{m}$ is Random Variable

- what distribution 'm' follow?

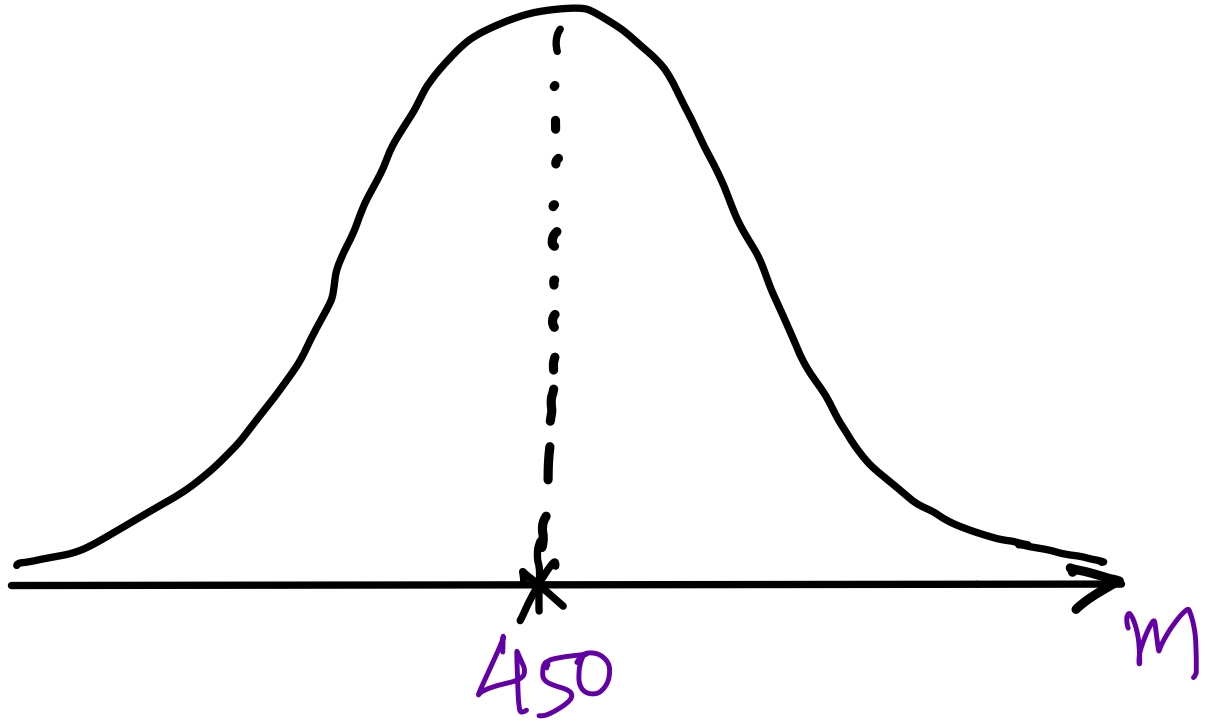
$\bar{m}$ → Sample mean

From CLT; $\bar{m}$ ~ Normal distribution ✓

- under the assumptions H_0 is True (i.e $\mu = 450$)

$\bar{x} \sim$ Normal distribution

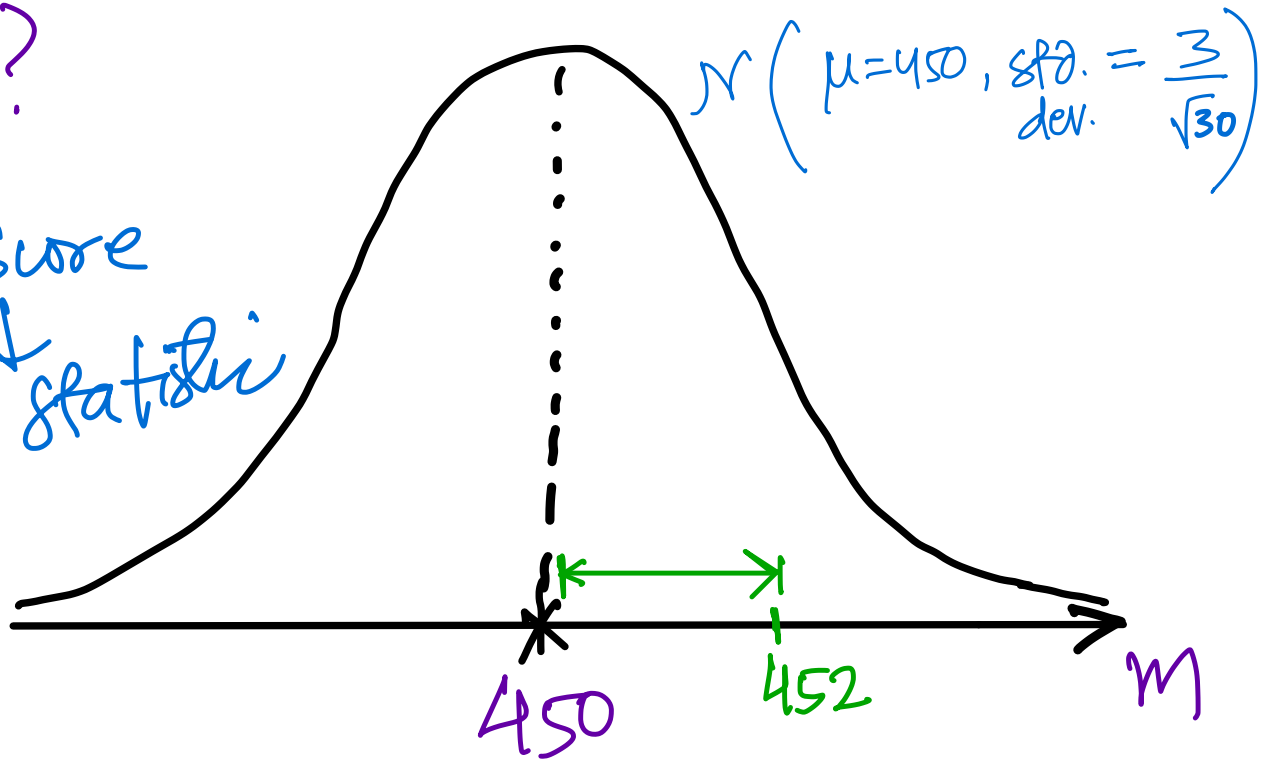
$$\sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \sim N\left(\mu = 450, \text{std. dev.} = \frac{3}{\sqrt{30}}\right)$$



Observed Value of m ?

$m = 452$
$n = 30$

$\Rightarrow$ Z score $\rightarrow$ Test statistic



- how many std. dev. away '$m = 452$' of m mean is?

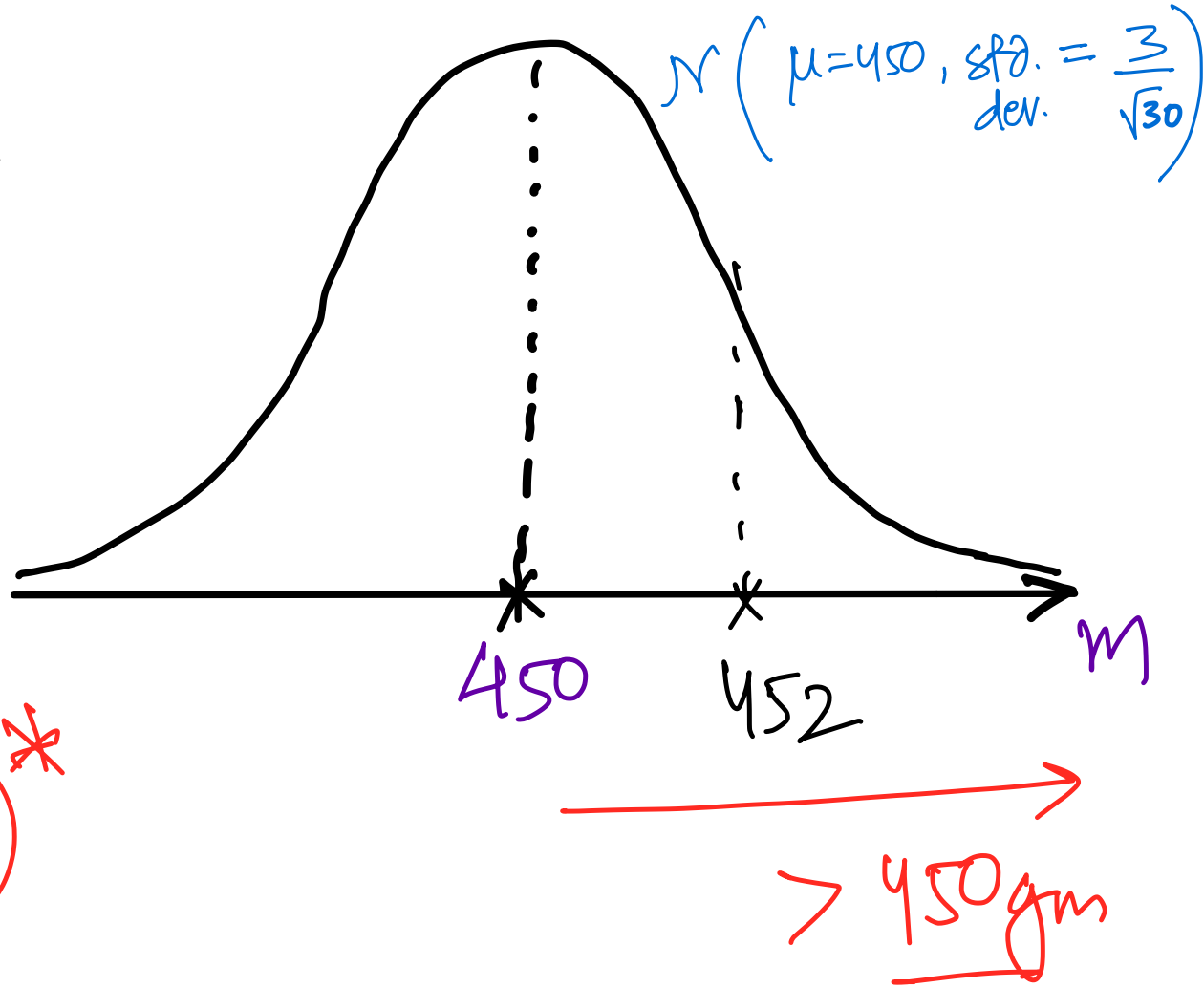
Z score : Test Statistic

$$Z = \frac{452 - 450}{(3/\sqrt{30})}$$

$$\boxed{Z = 3.65}$$ ✓

iii) Test setup:

H_A: $\mu > 450$ gm. ✓



Right tailed. *

iv) Compute p-value :

→ Test statistic

p-value : Prob of. observing data as extreme as what was observed under the assumption H_0 is True

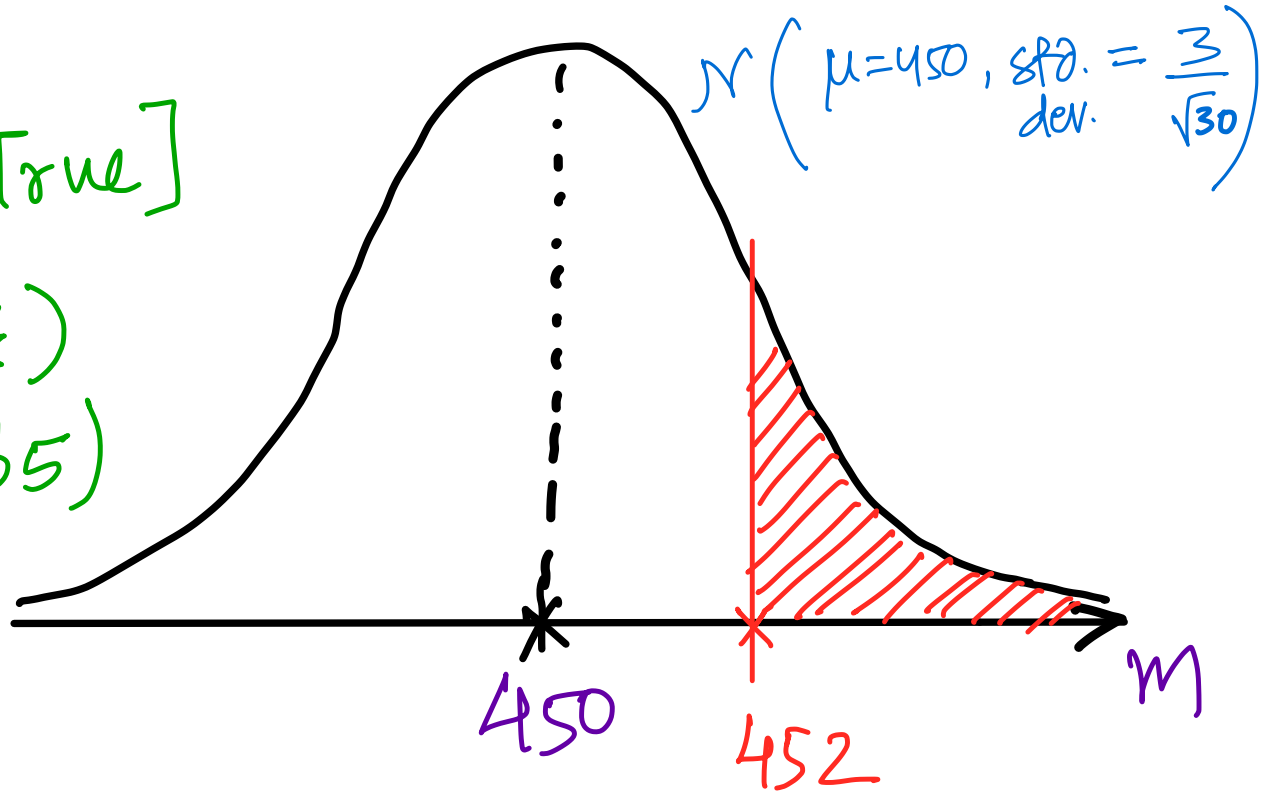
p-value

$= P[m > 452 \mid H_0 \text{ is True}]$

$= 1 - \text{norm.cdf}(Z)$

$= 1 - \text{norm.cdf}(3.65)$

p-value = 0.00013



V) Compare p-value with α:

Confidence required = 99%

→ Significance level

$\therefore \alpha = 1 - 0.99 \Rightarrow \alpha = 0.01$

p-value = 0.00013

- Clearly p-value $< \alpha$

i.e.) Reject H_0

→ We can conclude with 99% confidence that machine is OVERFILLING.

$H_0: \mu = 450$, $H_A: \mu > 450$ gm

CASE II : · QA team wants to perform test

↓

Is machine UNDERFILLING. ?

Mean wt. of 30 Bags = 448.7 gm

- H_0 : $\mu = 450$

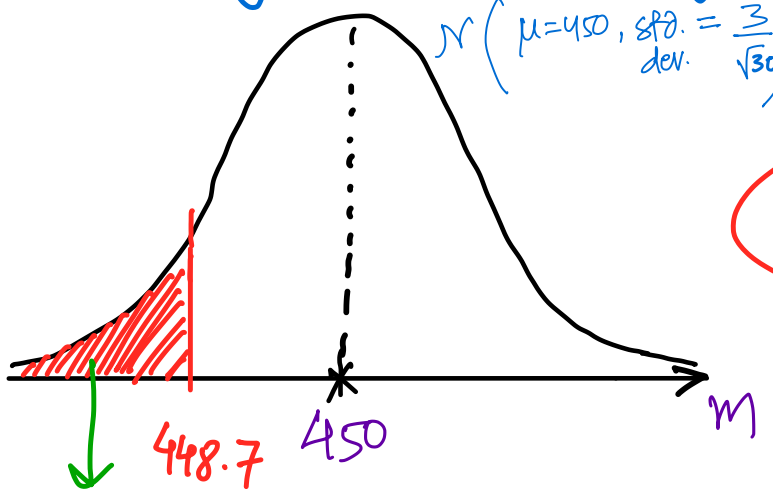
  H_A : $\mu < 450$

- Test statistic →

  → Z score of 448.7

$$Z = \frac{448.7 - 450}{3/\sqrt{30}}$$

$$Z = -2.37$$



Left tailed *

- p-value = norm.cdf(−2.37)
- p-value = 0.008
  $\alpha = 0.01$

$p < \alpha$

∴ Reject H_0

CASE III: · QA team wants to perform test

↓

Is machine UNDERFILLING/OVERFILLING or Working correctly or NOT

Mean wt. of 30 Bags = 449 gm

- H_0: $\mu = 450$
  H_A: $\mu \neq 450$

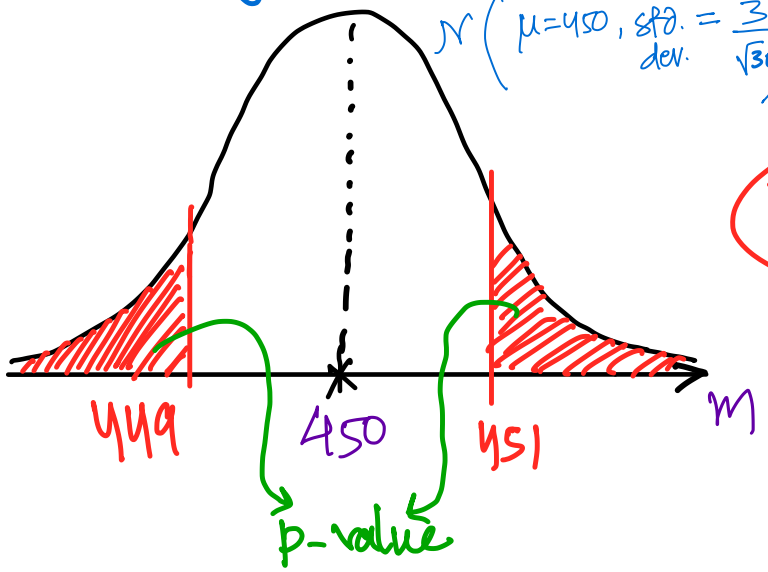
- Test statistic →

→ Z score of 449

$$Z = \frac{449 - 450}{3/\sqrt{30}}$$

$$Z = -1.826$$

- Z of 451 = +1.826



TWO tailed *

- p-value = norm.cdf(−1.826) + [1 − norm.cdf(+1.826)]
- p-value = 0.068

$\alpha = 0.01$ ✓

Clearly, $p\text{-value} > \alpha$

H_0: working correctly

H_A: not working correctly

- failed to Reject H_0
- i.e. we don't have enough evidence to conclude that machine is not working correctly.