Z-Test

⇒ Quick Recap → Hypothesis
→ CLT

→ Z test

→ Critical val

→ Confidence Invertals

⇒ **Hypothesis Testing Framework** →

→ Assumption → Collect data → Showcase some Result
⇓
was the assumption Correct

Detailed Framework →

⇒ 1) Setup Null Hypothesis (H_0) & Alternate Hypo (H_a)

2) Choose a distribution →
↳ Normal, uniform, binomial

3) Choose which tailed test to use →

decreasing ↓ left tailed

increasing ↓ right tailed

changing ↓ two-tailed

4) Calculate P-value ?? [Based on tail & distribution]

Prob. of (data | H_0 is true)

5) Compare P value to α [Significance level]

p value $< \alpha$ → reject H_0

p value $> \alpha$ → fail to reject H_0.

---

ex → Identify if a Coin is fair or not?

1) H_0 → It is a fair Coin

H_a → Unfair Coin.

2) Binomial dist^n → $p = 0.5$, $n = 100$

3) Two tailed

4) Calculate p_vale.

→ You have to perform the experiment

100 Tosses → 71 H, 29 T } → real experiment data

Calculate → Chance of seeing this or more extreme than this data | H_0 is true.

→ Binomial dist^n $n=100$, $p=0.5$ → Prob $(H > 71) = ??$

Prob (H > 71 given Binomial dist^n)

$P(H > 71)$

$n = 100$
$p = 0.5$

$\Rightarrow \underline{[1 - cdf(70)]}$ p value $\rightarrow$ 0.00001 ✓

5) Compare p value < α $\rightarrow$ reject the H_0

$\Rightarrow$ Coin is unfair.

CLT $\rightarrow$ $\mu = 65$ inch
$\sigma = 2.5$ inch

The average height is 65 inches with std dev 2.5. We take a sample of 50 people.
Let "m" represent the sample mean. What distribution does "m" follow?

$\Rightarrow$ Sample mean's distribution $\rightarrow$ Normal dist^n

Mean of SM dist $\equiv$ Population Mean

Std dev of SM dist $\equiv \dfrac{\sigma_{popl}}{\sqrt{n}}$

The average height is 65 inches with std dev 2.5. We take a sample of 50 people.
Let "m" represent the sample mean. What is the average or expected value of “m”

Mean of SM dist
$E(m)$ $\equiv \mu_{popula}$ $\rightarrow$ 65 inch

Average height is 65 inches with std dev 2.5. We take a sample of 50 people.
Let "m" represent the sample mean. What is the standard deviation of “m”

Std dev of SM

$$\underbrace{\text{Std deve of } m}_{SE} \equiv \frac{\sigma_{pop}}{\sqrt{n}} = \frac{2.5}{\sqrt{50}}$$

⇒ 5 people Sample vs 50 people Sample

SE for 5 people > SE for 50 people

Mean value for 5 people m ≈ mean value of 50 people m.

$$\mu_{n=50} = \mu_{n=5} = 65$$

$$\sigma_{n=50} = \frac{2.5}{\sqrt{50}} < \sigma_{n=5} = \frac{2.5}{\sqrt{5}}$$

⇒ Mama Earth → Shampoos

→ 2000 Stores across India

→ Historical weekly sales avg = 1800

σ = 100

Sales/store/week

⇒ We want to increase Sales

↓

do some promotion

⇒ We hire some agency →

⇒ Firm A → 50 Stores

→ avg 1850 bottles/week/store

⇒ Firm B → 5 Stores

→ avg 1900 bottles/week/store

⇒ Results we are seeing → Are they statistically significant or not.

$\Rightarrow$ Firm A $\rightarrow$ $\mu_{pop} = 1800$ $\mu_{sample} = 1850$

$\sigma_{pop} = 100$ $n = 50$

1) $H_0 \rightarrow \mu$ is Constant $\mu = 1800$

$H_a \rightarrow \mu > 1800$

2) Dist^n of SM $(n=50) \rightarrow$ normal dist^n

3) Right tailed test

4) P value $\rightarrow$

Chances that you will see 1850 or more given your data follows a normal dist^n

with $\mu = 1800$

$$\sigma = \frac{\sigma_{pop}}{\sqrt{50}} = \frac{100}{\sqrt{50}}$$

$P(X > 1850)$ given Data follow $N\left(1800, \frac{100}{\sqrt{50}}\right)$

1 - norm.cdf(z)

$$z = \frac{X - \mu}{\sigma}$$

$$= \frac{1850 - 1800}{100/\sqrt{50}}$$

⇒ Chances to see the $X > 1850$
given H_0 is true

$\boxed{0.002}$ p value

⇒ Reject the null hypothesis as p-value < α

$\alpha = 0.01$ (99% surity or Confidence)

⇒ Marketing Firm A → increased the sales.

Firm B

$\mu_{pop} = 1800$ $\mu_{sample} = 1900$

$\sigma_{pop} = 100$ $n = 5$

1) $H_0 \rightarrow \mu = 1800$

$H_a \rightarrow \mu > 1800$

2) Normal distn → $N\left(1800, \frac{100}{\sqrt{5}}\right)$

3) Right tailed test

4) pvalue → $Z = \dfrac{X - 1800}{100/\sqrt{5}} = \dfrac{1900 - 1800}{100/\sqrt{5}}$

p_value ⇒ 1 − norm.cdf (z) =

p_value = $\underline{0.012}$

5). 99% surity / confidence → $\alpha = 0.01$ → accept H_0

↳ Firm has no impact

• 95% surity / confidence → $\alpha = 0.05$ → reject H_0

↳ Firm has increased the impact.

⇒

Case I → Buisness Owner want 99% guarantee.

↳ A has increased → ✓

B has not

Case II → BO wants 95% guarantee

↳ A has increased

↳ B has increased

$\boxed{\text{One sample Z- Test}}$

- Compare Increase / decrease from original

- $n = 50 \rightarrow 1850 \leftarrow 1800$ (50 increase was okay)

  $n = 5 \rightarrow 1900 \leftarrow 1800$ (100 increase was not okay)

A fitness App claims that its users walk an average of 8,000 steps per day.
A random sample of 30 users showed an average of 7,600 steps per day with a standard deviation of 1,200 steps.
Conduct a right-tailed Z-test at a 5% significance level to determine if the App's claim is supported.
What is the p-value?

Right tailed test

$H_0 \rightarrow \mu = x$

$H_a \rightarrow \mu > x$

$\Rightarrow$ $\boxed{\begin{array}{l} \mu_{pop} \rightarrow 8000 \\ \sigma_{pop} \rightarrow 1200 \end{array}}$

If σ_{pop} is unknown $\rightarrow$ assume it same as sample σ (s)

$\boxed{\begin{array}{l} \mu_{sample} = 7600 \\ \sigma_{sample} = 1200 \end{array}}$

SM should follow a norm dist^n

with $\mu = \mu_{pop} = 8000$

$\sigma = \frac{\sigma_{pop}}{\sqrt{n}} = \frac{1200}{\sqrt{30}}$

p-value $\Rightarrow$ 1 − norm.cdf (z)

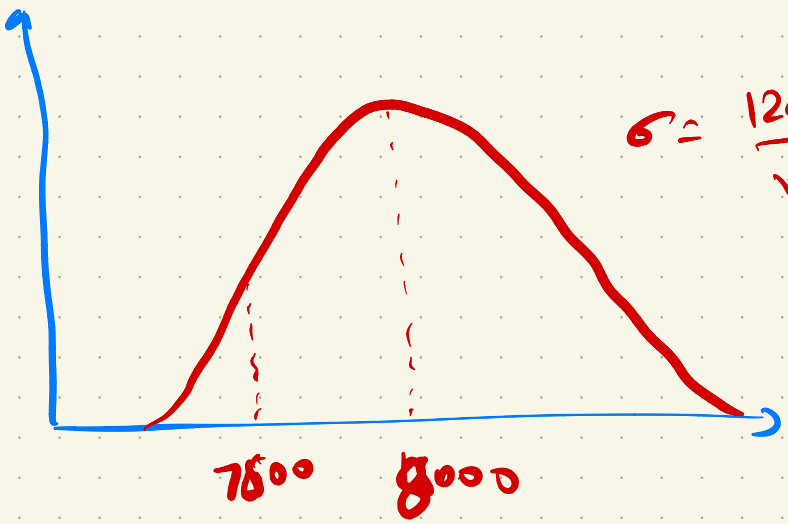
$z = \frac{7600 - 8000}{1200/\sqrt{30}}$

$\frac{8000 - 7600}{1200/\sqrt{30}}$

$\Rightarrow$ p_value = 0.966 $\rightarrow$

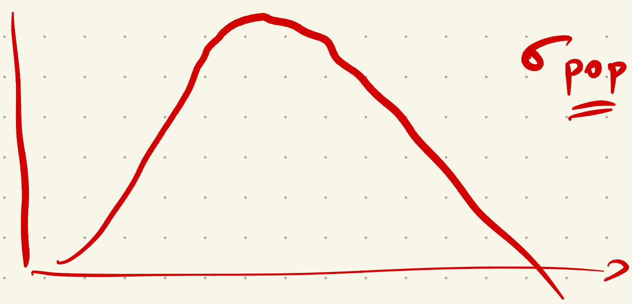
$\alpha = 0.05$

p_value $> \alpha$ $\rightarrow$ fail to reject.



$\sigma = \frac{1200}{\sqrt{30}}$

$\Rightarrow$ Population $\rightarrow$ 1 M people

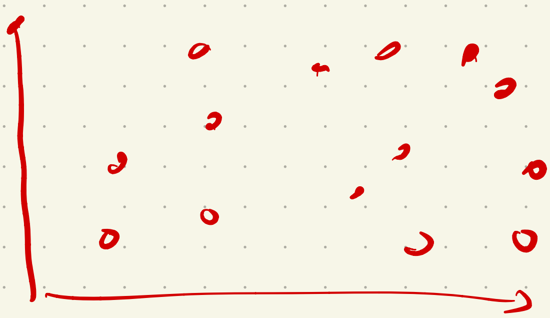


σ_{pop}

$\Rightarrow$ Sample $\rightarrow$ 30 people $\swarrow$

$\sigma_{sample} \leftarrow 1200$

30 values $\rightarrow$



Std dev of these 30 people

Sample mean = 9600

⇒ Dist^n of SM →

1000's of exp with 30 people at a time.

→ μ_1 σ_1

→ μ_2 σ_2

→ ⋮

μ_{100} σ_2

$\mu_1, \mu_2, \ldots \mu_{1000}$

↳ Std dev of this SM dist $\equiv SE = \dfrac{\sigma_{pop}}{\sqrt{n}}$

---

① Why is it Right tailed (7600 < 8000)

→ Assumption → H_0 H_a

→ Distribution of H_0 →

→ Choose what test I want to perform

Conduct the exp → give you some data
↓
Calculate Pvalue

→ Case I → 30 people → 8900 avg step

II 30 people → 8200 avg

III 30 people → 7600 avg

IV 30 people → 1500 Steps avg

$\sigma = \frac{1200}{\sqrt{30}}$

$P_{8400} < P_{8200} < P_{7600}$

⇒ • $\mu_{pop} = \underline{8000}$ ← $\underline{H_0}$

⇒ $\underline{\sigma_{pop}}$ ⇒ $S = \sigma_{sample}$

⇒ Can we calculate a value after which we will say that p_value $< \alpha$.

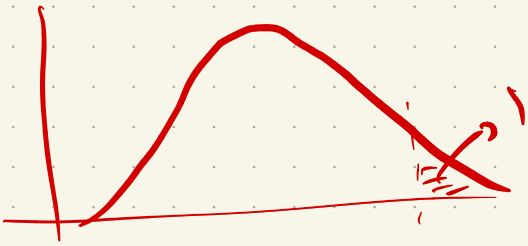
⇒ find a value x such that if experiment result are better than x then we can say that H_0 was rejected always.

⇒ $\mu_{pop} = 1800$ $n = 50$

$\sigma_{pop} = 100$ $\underline{\mu_{sample}} = ??$

Such that p_value $< \alpha$ →

$\alpha = 0.01$ → Can I Calculate a Corresponding z value



→ $p = 1 - \text{norm.cdf}(z)$

$0.01 = 1 - \text{norm.cdf}(z)$

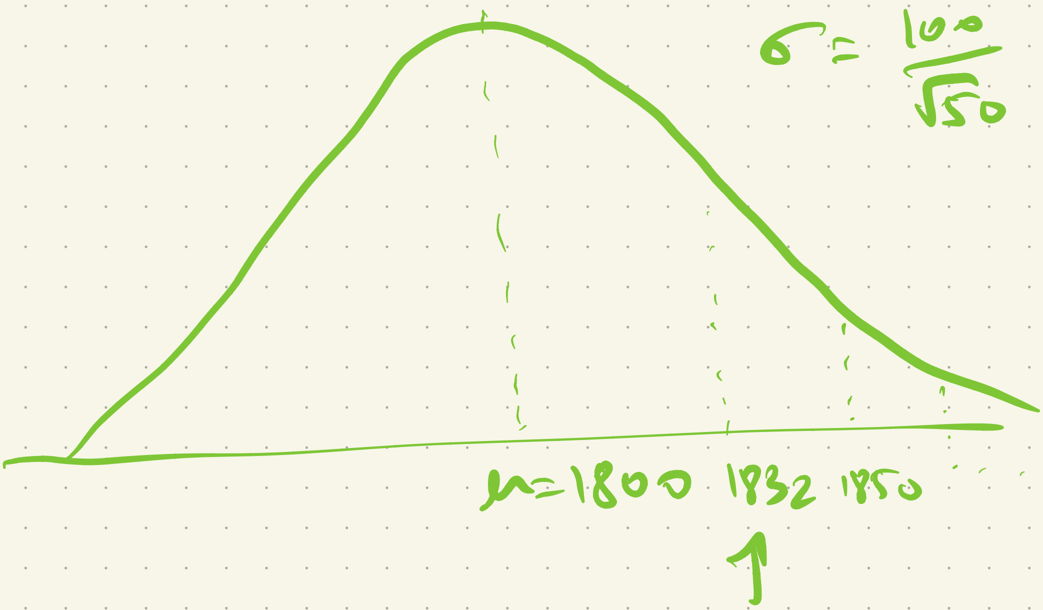
$\text{norm.cdf}(z) = 0.99$

→ z → $\text{norm.ppf}(0.99)$

$\frac{x - \mu}{\sigma}$ ⇒ $z = \frac{x - \mu_{pop}}{\sigma_{pop}/\sqrt{n}}$

$x = \mu_{pop} + \text{norm.ppf}(1 - \alpha) \frac{\sigma_{pop}}{\sqrt{n}}$

$\sigma = \frac{100}{\sqrt{50}}$



1832 → Critical value

the region after 1832 → Critical region.

→ 50 sample item → $\mu = 1800$ →

Test

If Firm A > 1832

Reject H0

have increased the sale

$\bar{x} = 1870$ ✓

D = 1840 ✓

E = 1820 ×× → random chance

F = 1200 ×× → unable to reject

⇒ 5 Store Sample.

$\alpha = 0.01$

$1-\alpha = 0.99$

$$\text{norm.ppf}(0.99) = \frac{X - 1800}{100/\sqrt{5}}$$

$$X = 1800 + \frac{100}{\sqrt{5}} \times \text{norm.ppf}(0.99)$$

$$\boxed{X = 1904}$$

If firm B would have done more than 1904 sales → reject H_0

But as they did 1900 → fail to reject H_0.

In a dataset of exam scores with a mean of 60 and a standard deviation of 15,
What is the critical value for the corresponding Z-score at a 95% confidence level?

$\mu = 60$

$\sigma = 15$

$\alpha = 0.05$

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

$= \text{norm.ppf}(0.95)$

$Z = 1.644$

$\frac{x-\mu}{\sigma} = \text{norm.ppf}(0.95)$

$\boxed{x = 84.67}$

→ Someone claims that they have better result than the avg →

⇒ Confidence Interval →

→ Sample data is known → predict a range for $\mu_{pop.}$

$$\mu_{sample} \pm SE * Z$$

$$CI = \mu_{sample} \pm \frac{\sigma_{pop}}{\sqrt{n}} Z$$

⇒ $\mu = 1800$ $\qquad \sigma = 100$

Sample = 1850 → predicted a CI.

$\sigma_{pop} = 100$

$n = 50$

⇒ ⇒ $\mu_{sample} \pm \frac{\sigma_{pop}}{\sqrt{n}} z$

$$1850 \pm \frac{100}{\sqrt{50}} z$$

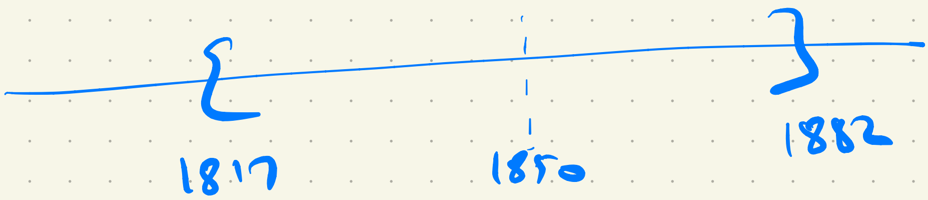
99% sure $\qquad 1850 \pm \frac{100}{\sqrt{50}}$ norm.ppf(0.99)

1850 $\xrightarrow[\text{99\% Confidence}]{}$ 1817 – 1882

If actual population mean is outside this range → reject H_0.

⇒ Case 1

1800



⇒ Case 2 →

1800 — 1832 — 1864

⇒ Case 3 →

1797 — 1800 — 1820 — 1852

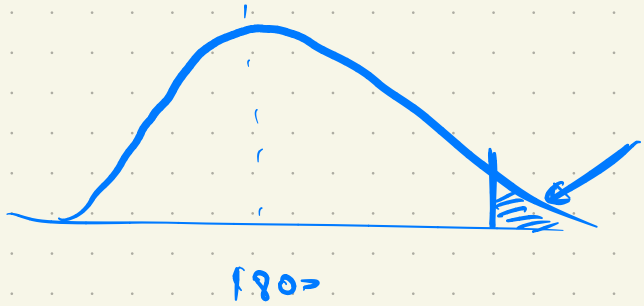
⇒

Avg Case

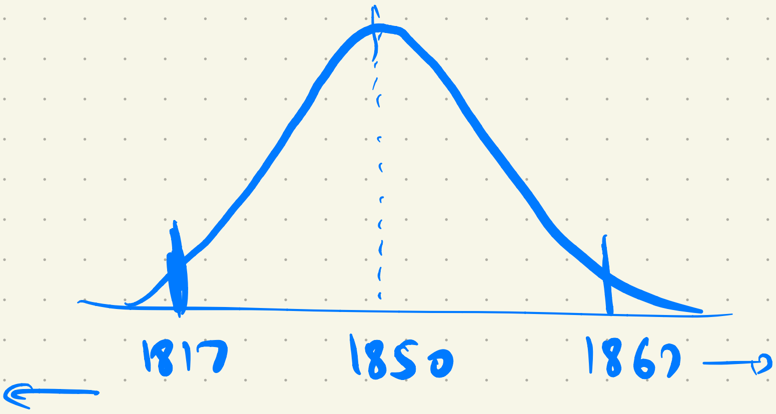
New Case

Significantly different than original

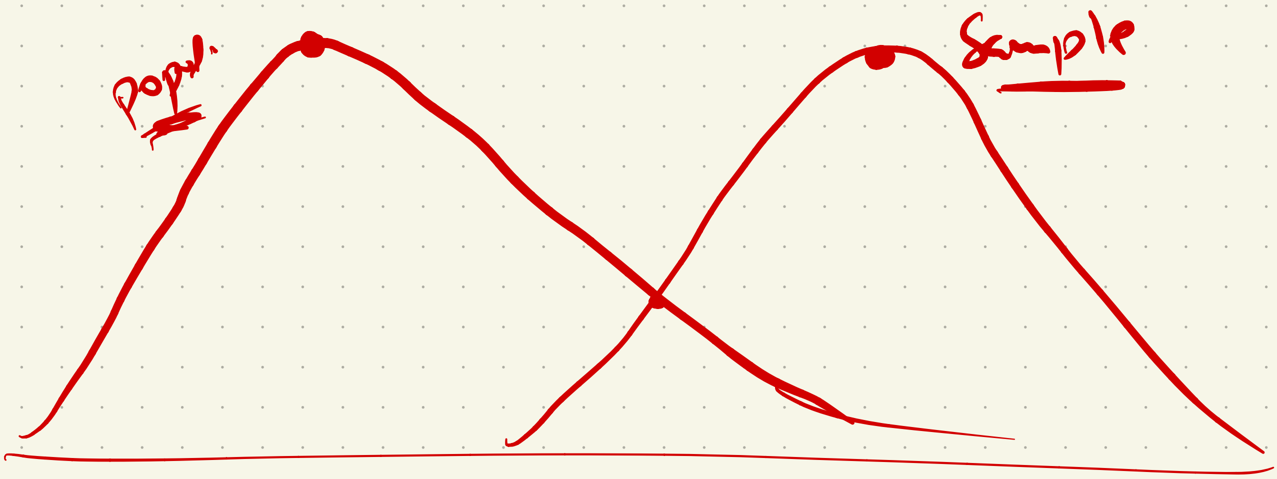
⇒ First Test → H_0 is distribute



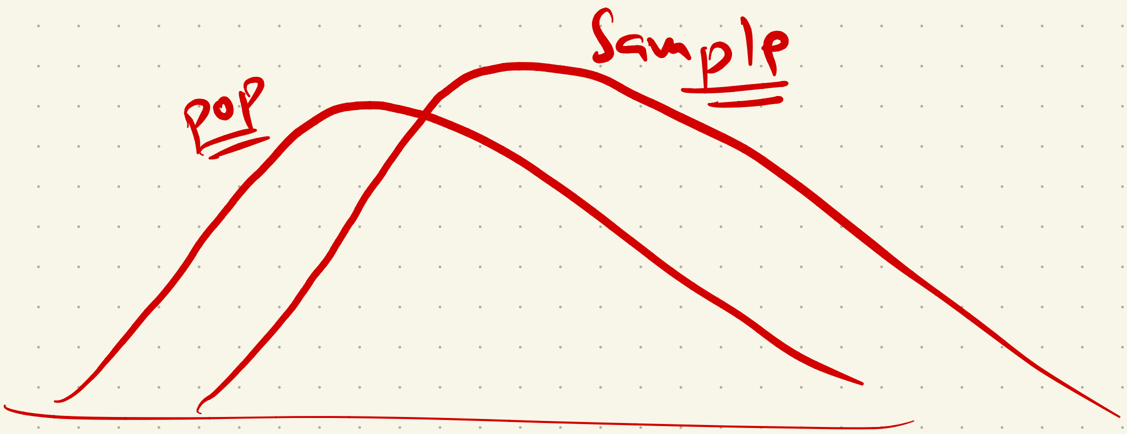
Second way → How the raw data will follow a distⁿ



⇒ ①



⇒ ②



Doubts

$\alpha = 0.01$ pvalue $= 0.009$

One tailed $\rightarrow$ reject H_0

two tailed $\frac{\alpha}{2} = 0.005$ $\rightarrow$ accept H_0

p-value

A data scientist is working on a credit scoring model for a bank.

The goal is to determine if an applicant is creditworthy (has a low credit risk) or not creditworthy (has a high credit risk) based on various financial factors.

The bank has set a credit score threshold, and applicants above this threshold are considered creditworthy, while those below it are considered not creditworthy.

Testing if a person is Credit worthy or Not

← Reality

Test outcome	Worthy	Not worthy
Worthy	A TP	B FP
Not worthy	C FN	D TN

	Reject	Not Reject
Reject	D TP	C
Not Reject	B	A TN

Test $\rightarrow$ Threshold

$X \rightarrow > X_T$ worthy

$X < X_T$ Not worthy

Reality: Applicant is not creditworthy.

Decision: The model correctly identifies the applicant as not creditworthy.

True -ve

Reality: Applicant is not creditworthy.

Decision: The model incorrectly classifies the applicant as creditworthy.

False +ve

↓

Type I

Reality: Applicant is creditworthy.

Decision: The model incorrectly classifies the applicant as not creditworthy.

False -ve

↓

Type II

Reality: Applicant is creditworthy.

Decision: The model correctly identifies the applicant as creditworthy.

True +ve

As a data scientist you are working for an e-commerce company, and you want to determine if the introduction of a new algorithm has led to an increase in the average order value of customer purchases.

After conducting an analysis of customer purchases before and after the introduction of the new algorithm, the data scientist obtains a p-value of 0.002.

The significance level (α) is set at 0.05.

What would be the appropriate hypotheses and conclusion to this situation?

$H_0 \rightarrow$ Nothing has change

$H_a \rightarrow$ increased.

reject H_0 as p-value $< \alpha$.

$\Rightarrow$ It has increased the

Which of the following would be the appropriate test to claim that the mean weight of airline passengers with carry-on baggage is at most 195 lb ?

$\rightarrow$ Left one tail.

The verbal reasoning in the GRE has an average score of 150 and a standard deviation of 8.5.

A coaching centre claims to improve these numbers for their students. How should the null and alternate hypotheses be set up?

$H_0 \rightarrow$ Does not improve

$H_a \rightarrow$ It improves the score.

It is known that students of a coaching centrehave better results then average. You claim that this is false. How should the null and alternate hypotheses be set up?

$H_0 \rightarrow$ the score is better

$H_a \rightarrow$ Score is same as mean $\leq$ mean

Collab Link : https://colab.research.google.com/drive/1HpDaz5Aia9HwmcE1bx5KXMtFQ35j6l0J?usp=sharing