$X \approx$ Log Normal Distribution (μ, σ)

$\Downarrow$

$Y \approx \ln(X)$ = Normal Distribution

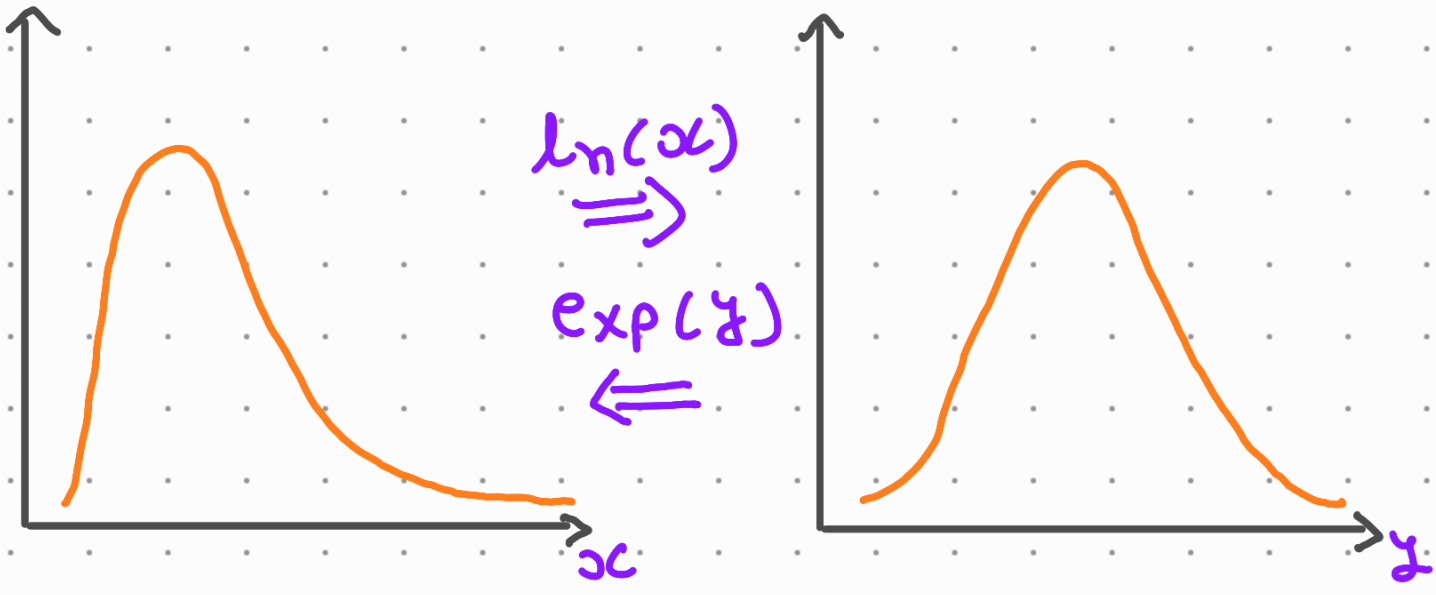
$\Downarrow$

Natural Log

$[\log_e]$

$\Downarrow$

$X \approx \exp(Y)$ = Log Normal Distribution



Log Normal Distribution

Normal Distribution

$\Downarrow$

QQ plot

e.g. ① The wealth distribution of the world

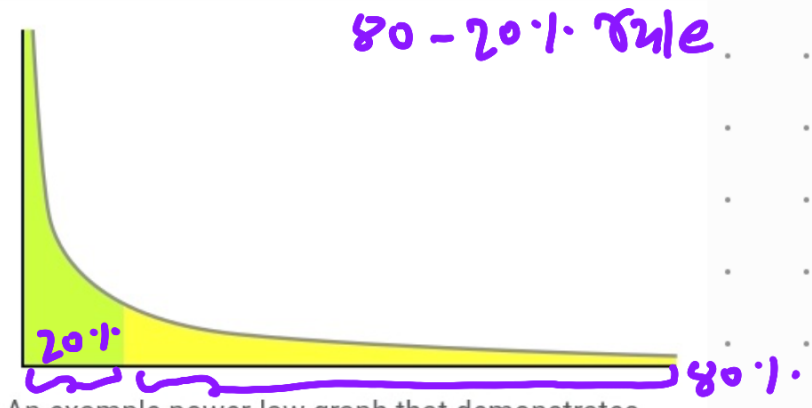
② Discussion Forum → length of the coments

③ Length of Chess Game

④ Dwell time on online articles (joke, news)

⑧ Power Law Distribution

In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a relative change in the other quantity proportional to a power of the change, independent of the initial size of those quantities: one quantity varies as a power of another.



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

e.g.

① -> 20% of Team is responsible for winning 80% of matches.

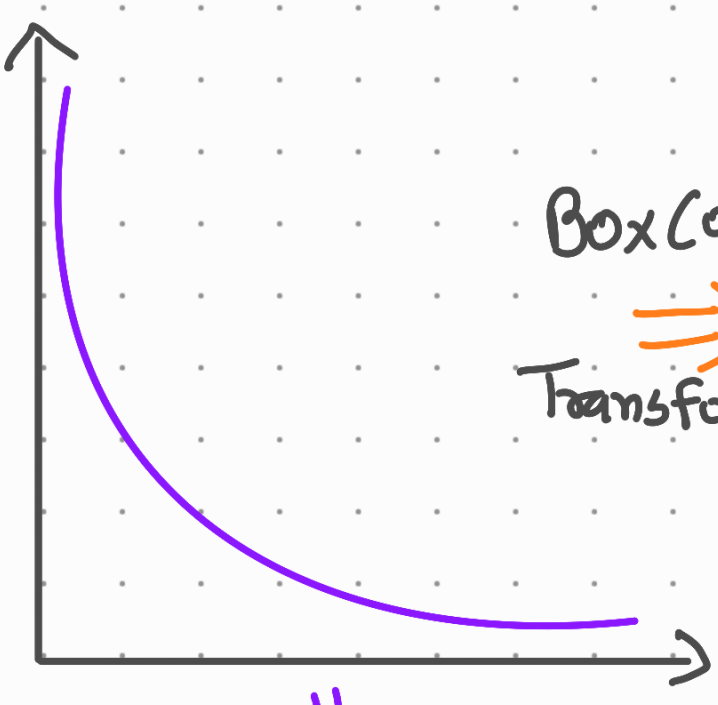
② -> 80% of wealth are distributed with 20% of the total population.

③ 80% of the total oil is with 20% of the nation

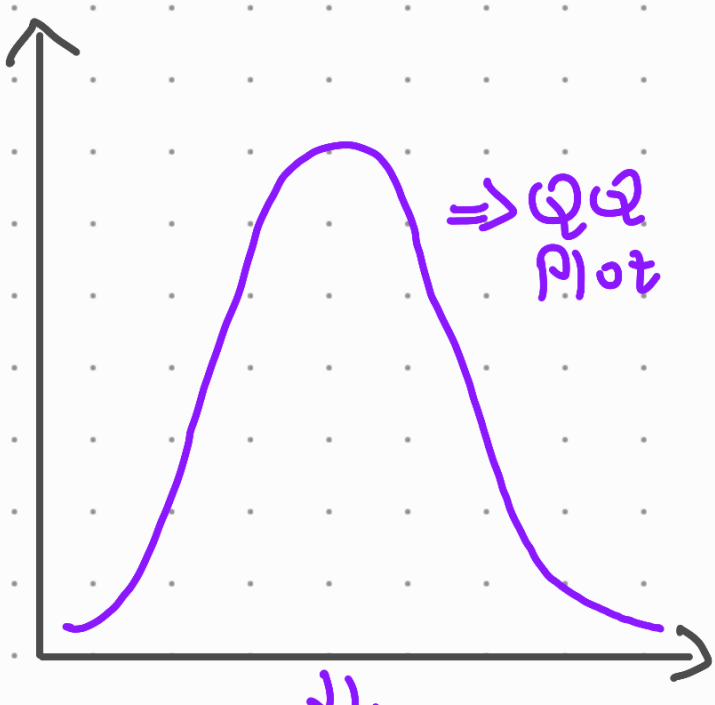
④ Frequencies of words in most languages.

(5) 20% of the major defects fixes the 80% of the upcoming defects in a software product.

→ Distribution Graph



Box Cox
⇒
Transform



⇒ QQ Plot

⇓

Power Law Distribution

⇓

Pareto Distribution

⇓

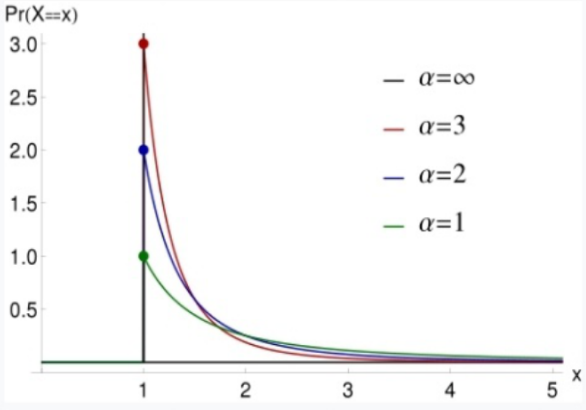
Normal Distribution

⑨ Pareto Distribution

The Pareto principle or "80-20 rule" stating that 80% of outcomes are due to 20% of causes was named in honour of Pareto, but the concepts are distinct, and only Pareto distributions with shape value (α) of log45 ≈ 1.16 precisely reflect it. Empirical observation has shown that this 80-20 distribution fits a wide range of cases, including natural phenomena and human activities.

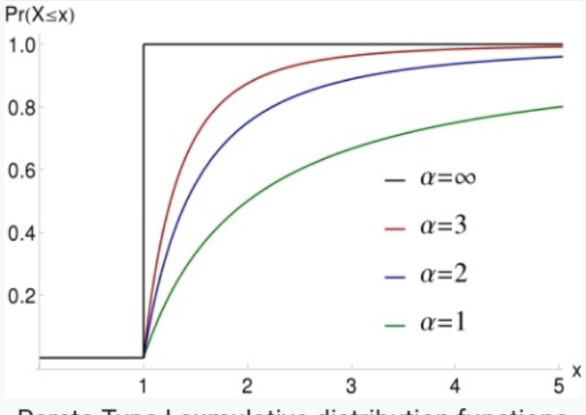
Pareto Type I

Probability density function



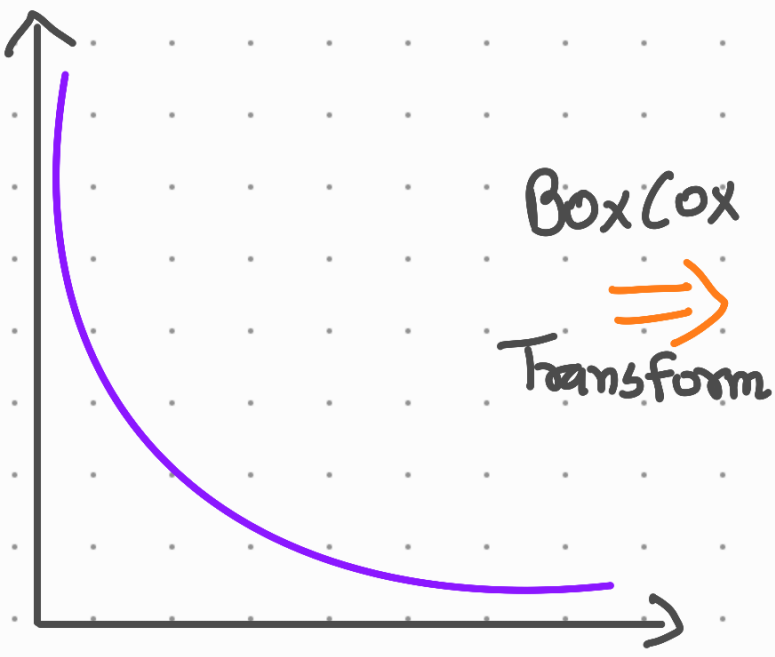
Pareto Type I probability density functions for various α with $x_\mathrm{m} = 1$. As $\alpha \to \infty$, the distribution approaches $\delta(x - x_\mathrm{m})$, where δ is the Dirac delta function.

Cumulative distribution function



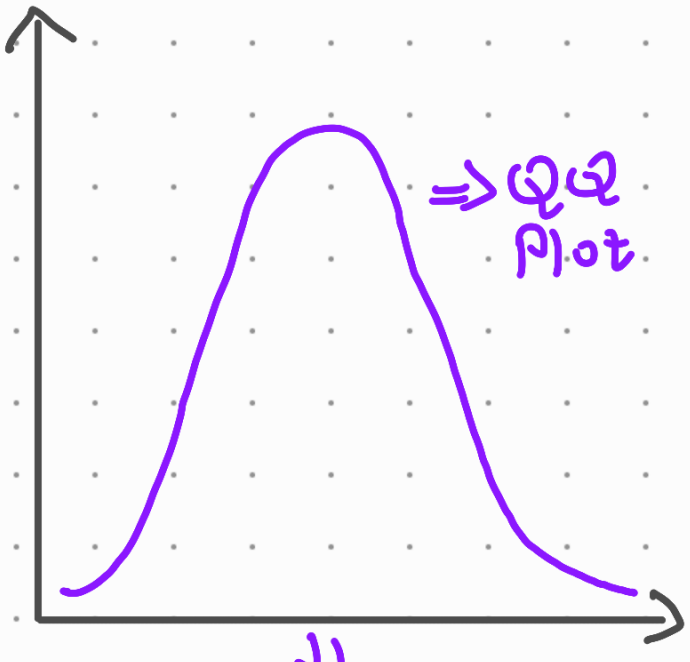
Pareto Type I cumulative distribution functions for various α with $x_\mathrm{m} = 1$.

Parameters	$x_\mathrm{m} > 0$ scale (real) $\alpha > 0$ shape (real)
Support	$x \in [x_\mathrm{m}, \infty)$
PDF	$\dfrac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}}$
CDF	$1 - \left(\dfrac{x_\mathrm{m}}{x}\right)^\alpha$
Quantile	$x_\mathrm{m}(1-p)^{-\frac{1}{\alpha}}$
Mean	$\begin{cases} \infty & \text{for } \alpha \le 1 \\ \dfrac{\alpha x_\mathrm{m}}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$
Median	$x_\mathrm{m}\sqrt[\alpha]{2}$
Mode	x_m
Variance	$\begin{cases} \infty & \text{for } \alpha \le 2 \\ \dfrac{x_\mathrm{m}^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \text{for } \alpha > 2 \end{cases}$
Skewness	$\dfrac{2(1+\alpha)}{\alpha-3}\sqrt{\dfrac{\alpha-2}{\alpha}}$ for $\alpha > 3$



Box Cox

⇒ Transform



⇒ QQ Plot

⇓

Pareto Distribution

⇓

Normal Distribution

e.g.

(1) 80.1. of the entire project is done by 20.1. of the team.

(2) 80.1. of defects can be solved by 20.1. of the main defects.

Central Limit Theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large number of samples taken from a population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

① $x \approx N(\mu, \sigma)$

n = Sample size ⇒ Any

$S_1 = \{x_1, x_2, x_3 \ldots x_n\} = \bar{x}_1$

$S_2 = \{x_1, x_3, \ldots x_n\} = \bar{x}_2$

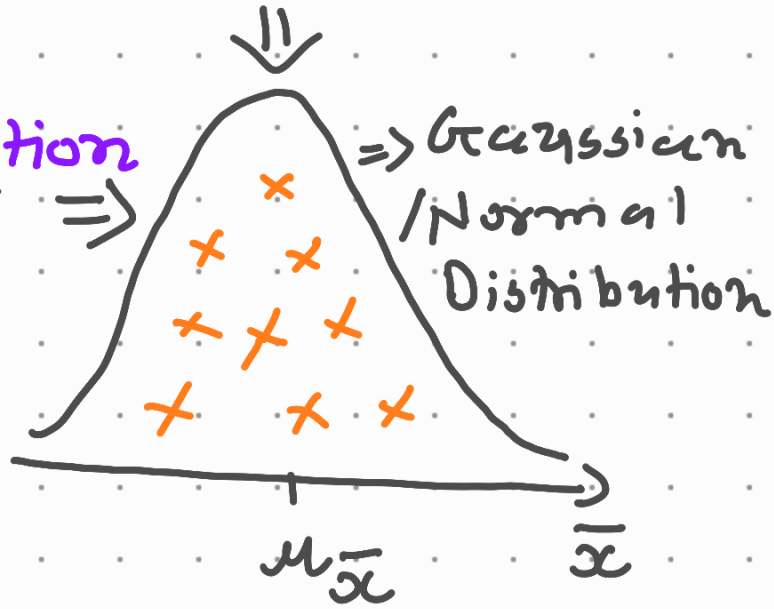
S_3 . . . . . .

S_4 . . . . . .

⋮

S_m . . . . . $\bar{x}_m$

Sampling Distribution of Mean ⇒



② $x \not\sim N(\mu, \sigma)$

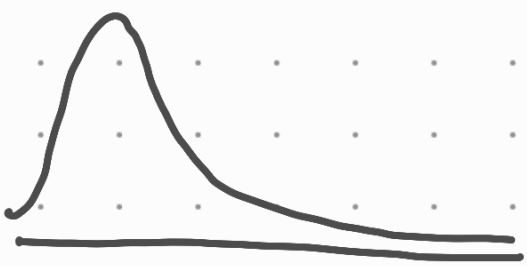
$\boxed{n \geq 30} \Rightarrow$ Sample size

s_1 $\quad$ $\bar{x}_1$

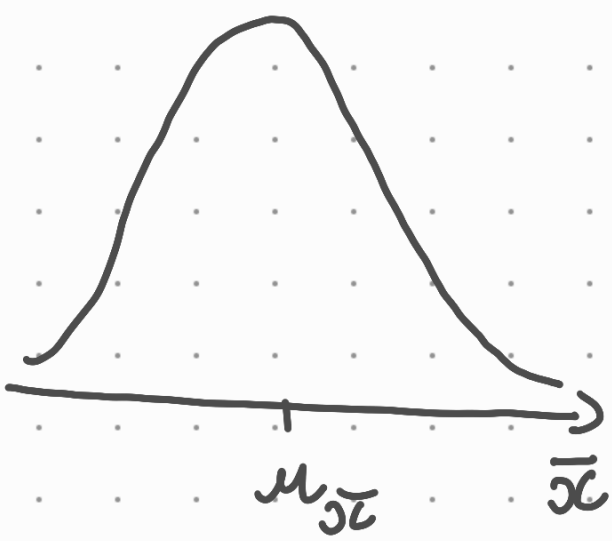
s_2 $\quad$ $\bar{x}_2$

$\vdots$ $\quad$ $\vdots$

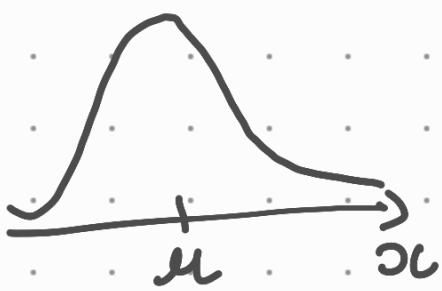
s_m $\quad$ $\bar{x}_m$



$\Downarrow$ CLT



① Normal Distribution

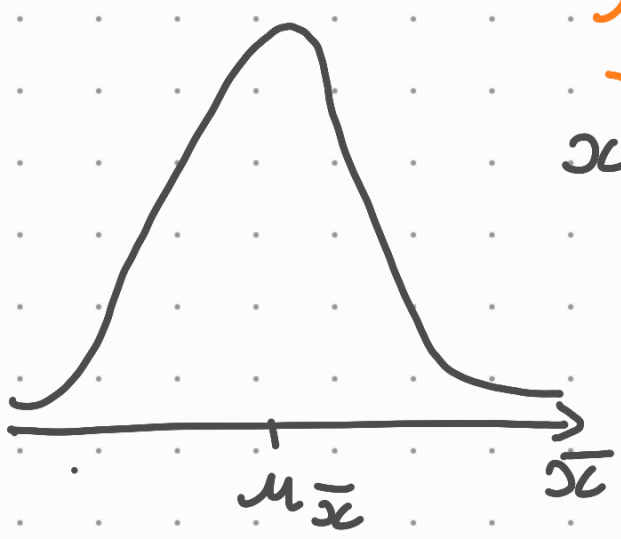


$x \sim N(\mu, \sigma)$

Sampling Distribution of Mean

σ = Population S.D.
μ = Population Mean
n = Sample Size

$$\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

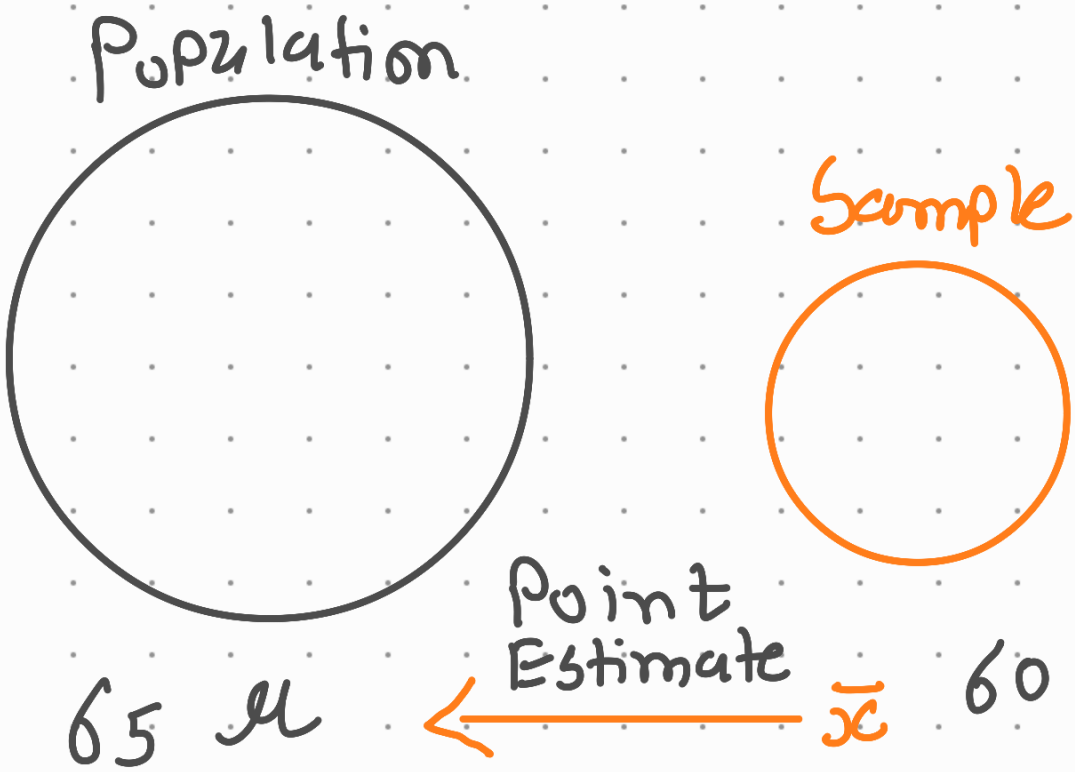


Estimates

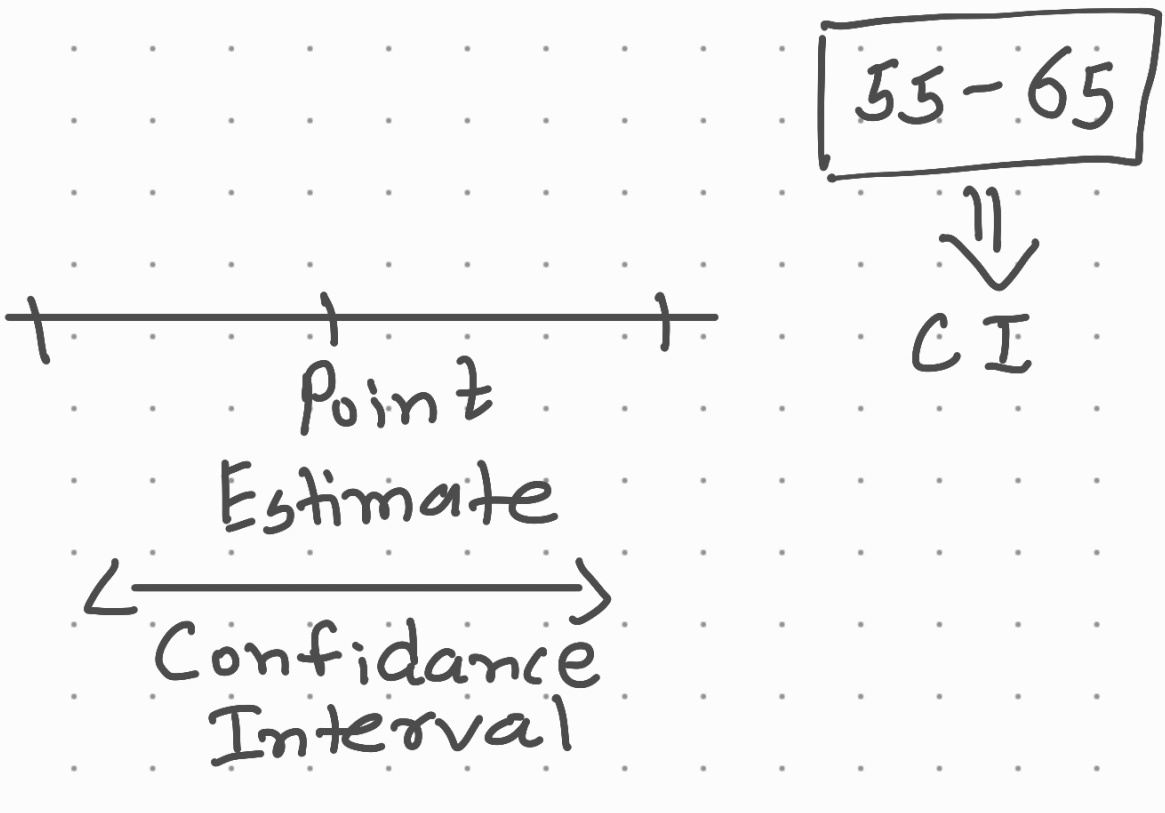
It is specified observed numerical value used to estimate an unknown population parameter.

① **Point Estimate**: Single numerical value used to estimate an unknown population parameter.

e.g. Sample mean is a point estimate of population mean.

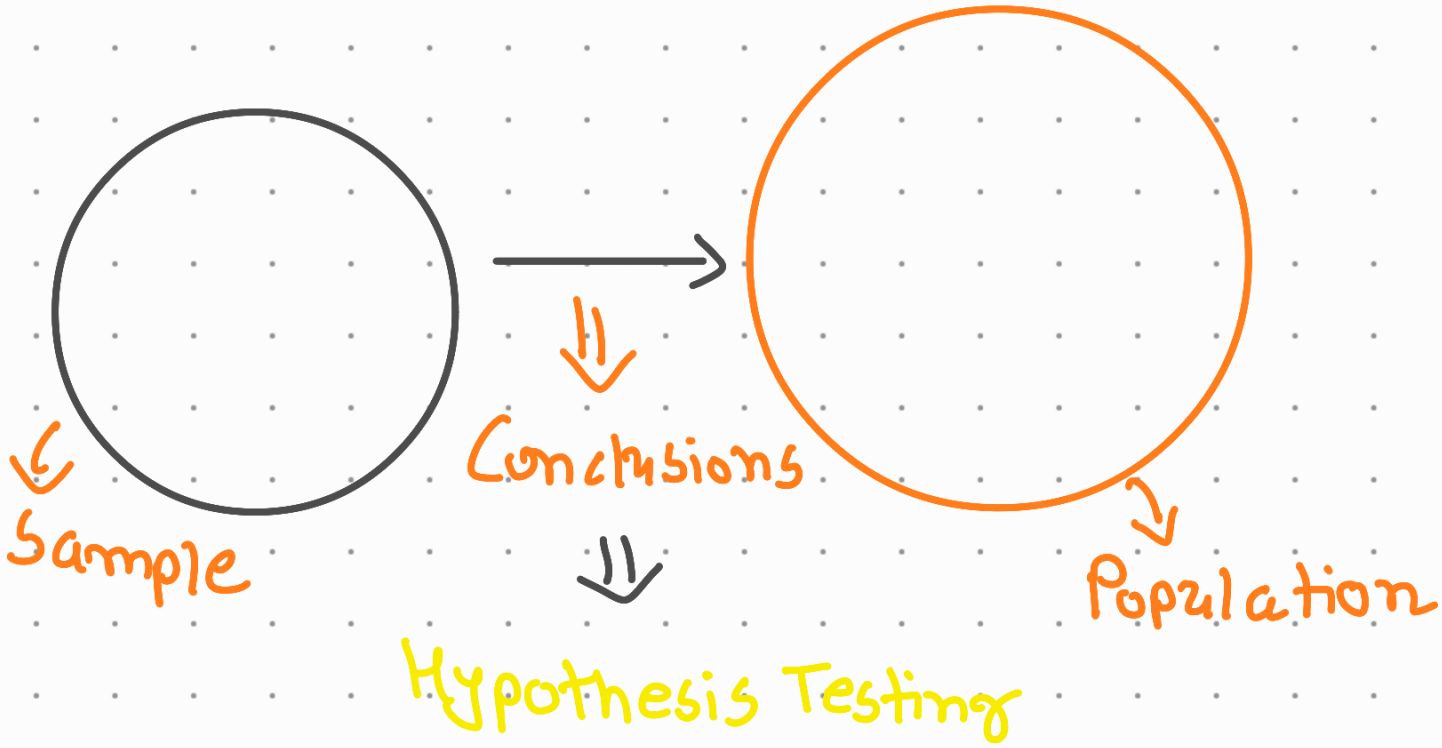


② Interval Estimate: Range of values is used to estimate unknown population parameter.



A Hypothesis And Hypothesis Testing Mechanism

→ Inferential Stats : Conclusion or Inference



→ Hypothesis Testing Mechanism

① Null Hypothesis (H_0) - The person is not guilty

→ The assumption you are begining with

② Alternate Hypothesis (H_1) - The person is guilty

→ Opposite of Null Hypothesis

(3) Experiments - Statistical Analysis

→ Collect Proof (e.g. DNA, Finger Print)

(4) Accept the null Hypothesis or
Reject the null Hypothesis

Eg: Colleges or District A states its average passed percentages of Students are 85%. A new college opened in the district and it was found that a sample of student 100 have a pass percentage of 90%. with a standard deviation of 4%.

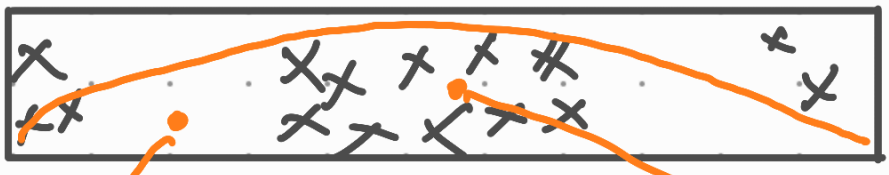
Does this school have a different passed percentage.

Ans Null Hypothesis (H_0) = $\mu = 85.1$.

Alternate Hypothesis (H_1) = $\mu \neq 85.1$.

P-Value

A p-value, or probability value, is a number describing how likely it is that your data would have occurred under the null hypothesis of your statistical test



Space Bar → p-value = 0.8

p-value = 0.2 ⇒ Out of 100 touches on space bar, we touch around 20 times in this region.

→ Hypothesis Testing

e.g. Coin is Fair or Not {100 Tosses}

P(H) = 0.5 P(T) = 0.5

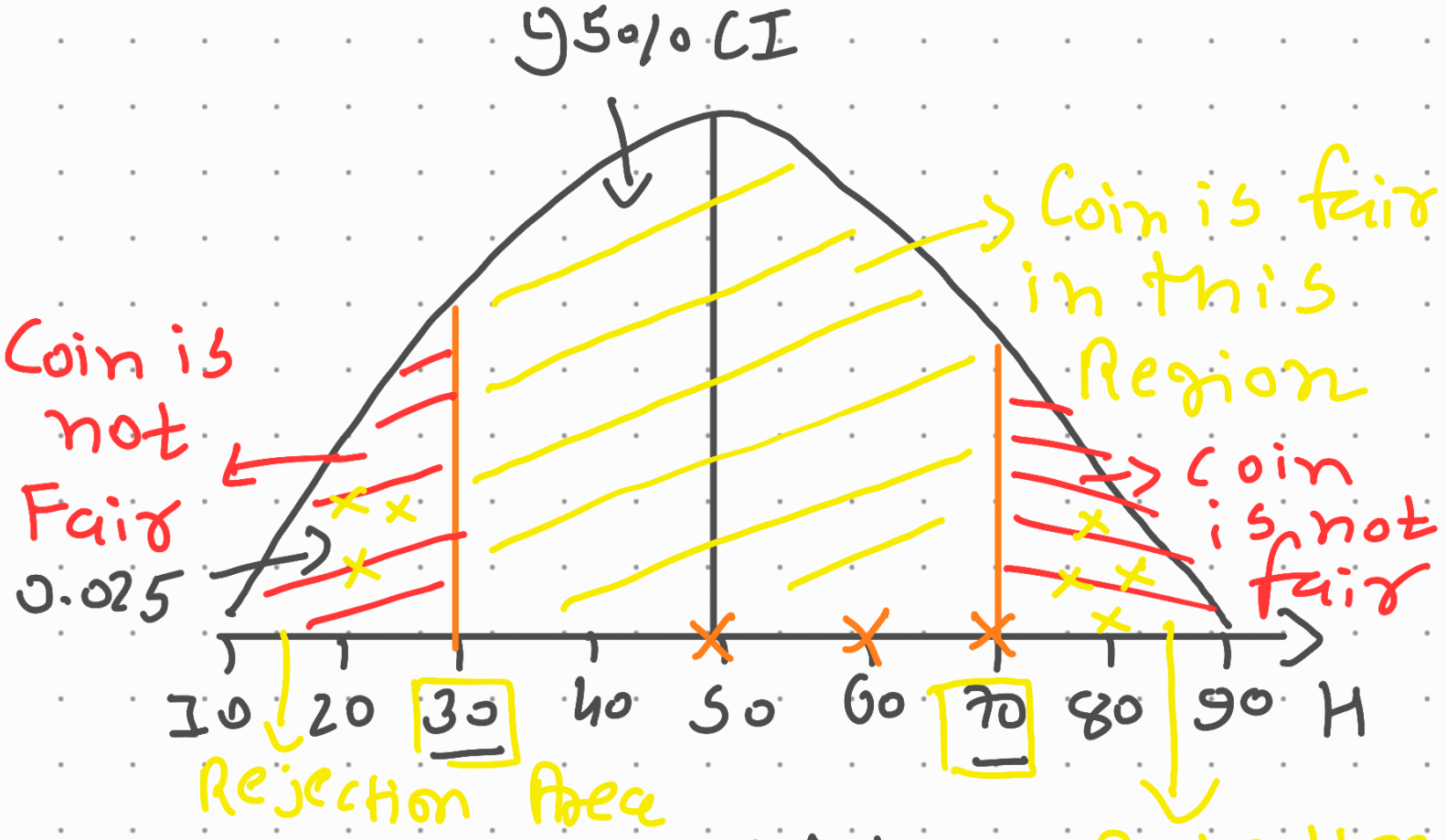
if P(H) = 0.6 P(T) = 0.4

if P(H) = 0.7 P(T) = 0.3

① Null Hypothesis: H_0
↳ The Coin is Fair

② Alternate Hypothesis: H_1
↳ The Coin is Not Fair

③ Experiment : 100 times



④ Significance Value

→ $\alpha = 0.05$

→ $CI = 1 - 0.05 = 0.95$

→ P < Significance Value

Reject the Null Hypothesis

→ else

Fail to Reject the Null Hypothesis

Statistical Analysis

① Z-Test

② T-Test

} ⇒ Average

⇓

Z-Table → Z-score & P-Value

T-Table

③ CHI SQUARE Test ⇒ Categorical Data

④ ANNOVA Test ⇒ Variance of Data

(1) Z-Test

Conditions To Use

(i) Population Std

(ii) $n \geq 30$

with a $\sigma = 3.9$

1) The average heights of all residents in a city is 168cm. A doctor believes the mean to be diffuent. He measured the height of 36 individuals and found the average height to be 169.5 cm.

(a) State null and Alternate Hypothesis

(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis.

→ $\mu = 168$ cm

$\sigma = 3.9$

$n = 36 \geq 30$

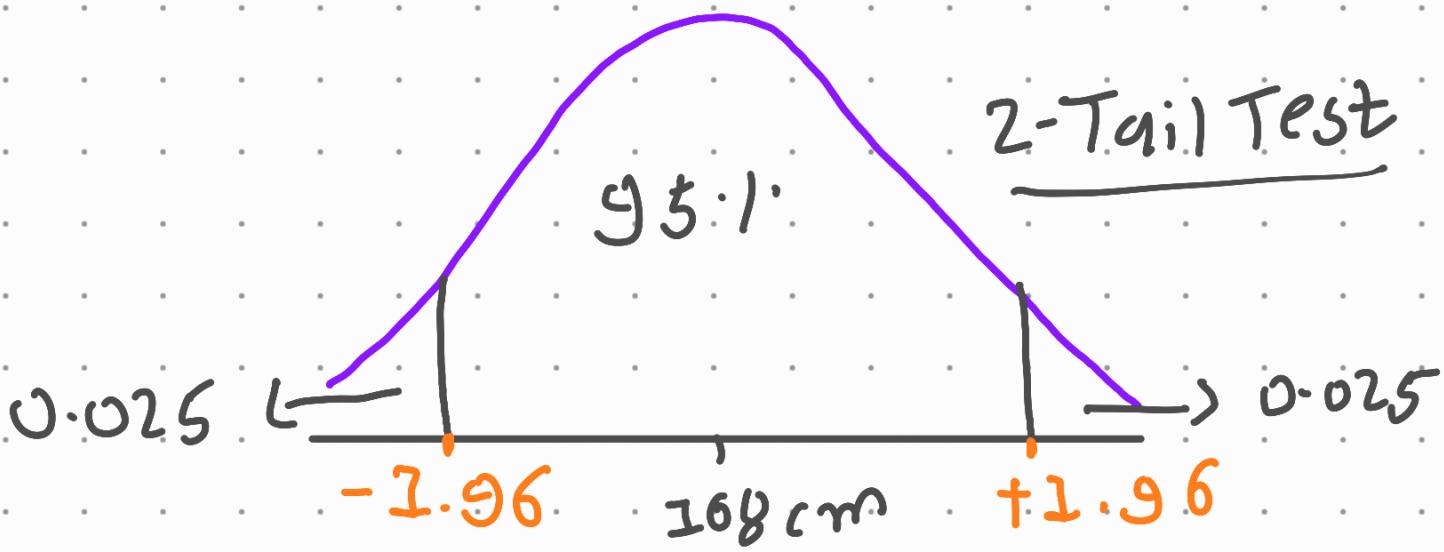
$\bar{x} = 169.5$ cm

$CI = 0.95$ $\alpha = 1 - 0.95 = 0.05$

(a) Null Hypothesis $H_0 = \mu = 168$ cms

(b) Alternate Hypothesis $H_1 = \mu \neq 168$ cms

(c) Based on C.I. Draw decision Boundry



$= 1 - 0.025$

$= 0.9750$ ⇓ Area

$\Rightarrow$ Z-Score $= 1.96$ (find from Z-table)

→ if Z is less than -1.96 or greater than or equal to $+1.96$, Reject the NULL HYPOTHESIS.

→ Z-Test

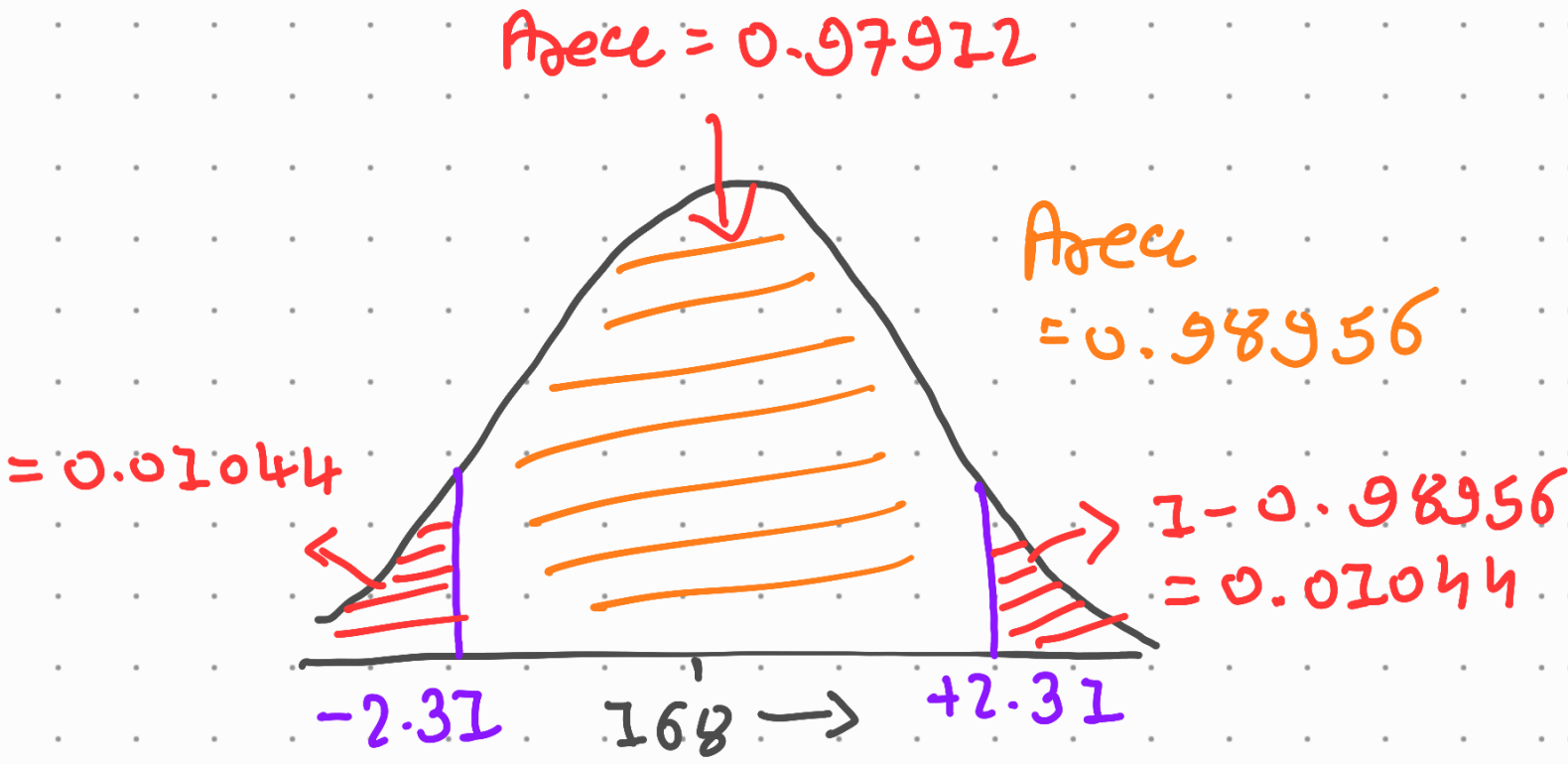
$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \text{Sample Data}$$

$$= \frac{169.5 - 168}{3.9/\sqrt{36}}$$

$$= \frac{1.5}{0.65}$$

$Z_d = 2.31 > 1.96$ so we reject the Null Hypothesis

$p < 0.05$



→ p-value = 0.02044 + 0.02044

= 0.02088

$p < 0.05$

→ So we reject the NULL Hypothesis.

Final Conclusion

→ The average height ≠ 168 cms.

The average height seems to be increasing based on the sample data.

② A factory manufactures bulbs with a average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years.

(a) State null and alternate hypothesis

(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised?

→ $\sigma = 0.50$ $\mu = 5$

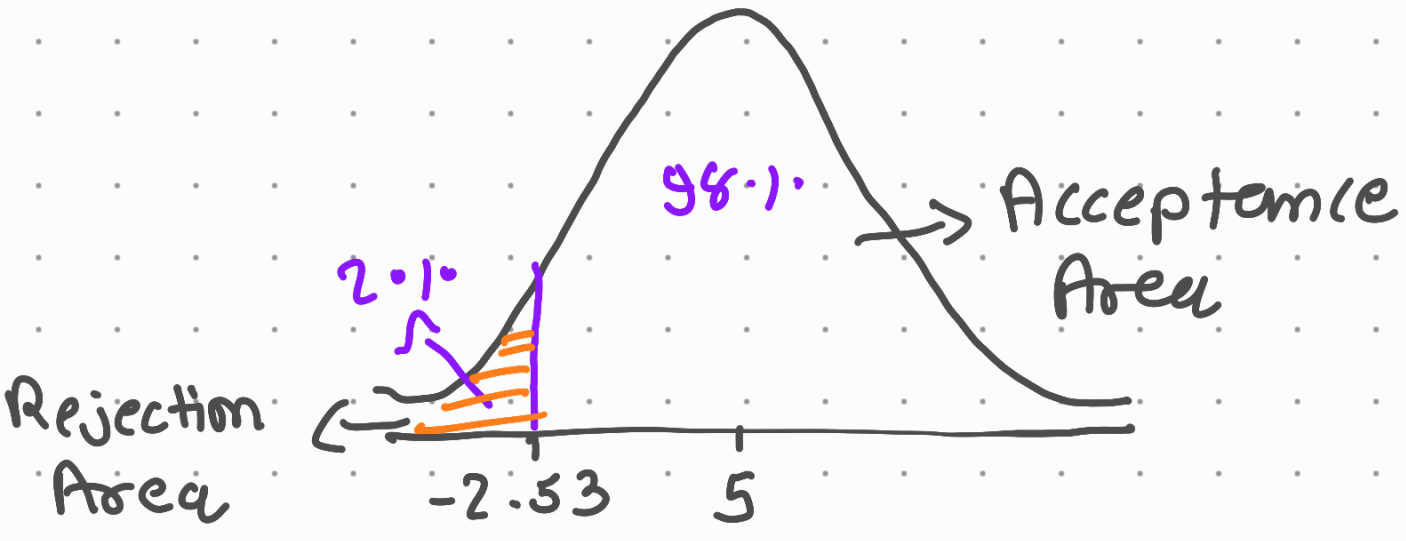
$n = 40$ $\bar{x} = 4.8$

(a) Null Hypothesis $H_0 = \mu = 5$

(b) Alternate Hypothesis $H_1 = \mu < 5$

{1-tail test}

(c) Decision Boundry



Z-Test

$$Z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{4.8 - 5}{0.50/\sqrt{40}}$$

$Z_d = -2.53$ => Area under the Curve
= 0.00570

P-Value = 0.00570

-> Compare P-Value with Significance Value

Significance Value $\alpha = 0.02$

$0.00570 < 0.02$ => True

-> We reject the Null Hypothesis.

Student T Distribution

→ In Z-stats when we perform any analysis using Z-score, we require σ (Population Standard Deviation) → is already known.

→ How do we perform analysis when we don't know the population standard deviation?

⇓

We use Student T Distribution

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \boxed{t = \frac{\bar{x} - \mu}{S/\sqrt{n}}}$$

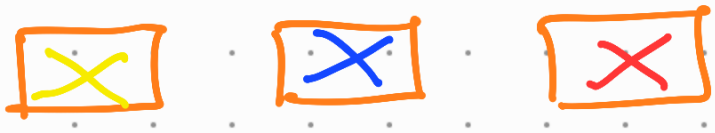
→ t-test or t-stats

S = Sample Standard Deviation

→ We will use t-table.

Degree of Freedom

$$dof = n - 1$$

3 people => 

→ First will have 3 choices where to seat

→ Second will have 2 choices where to seat → $dof = 3 - 1 = 2$

→ But First will have to seat on the spot remaining → $dof = 3 - 2 = 1$

T-Stats With T-test Hypothesis Testing

(One Sample T-test)

① In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication affect intelligence? [redacted]

→ CI = 95% n = 30

$\alpha = 0.05$ $\bar{x} = 140$

$\mu = 100$ $s = 20$

(a) Null Hypothesis $H_0 = \mu = 100$

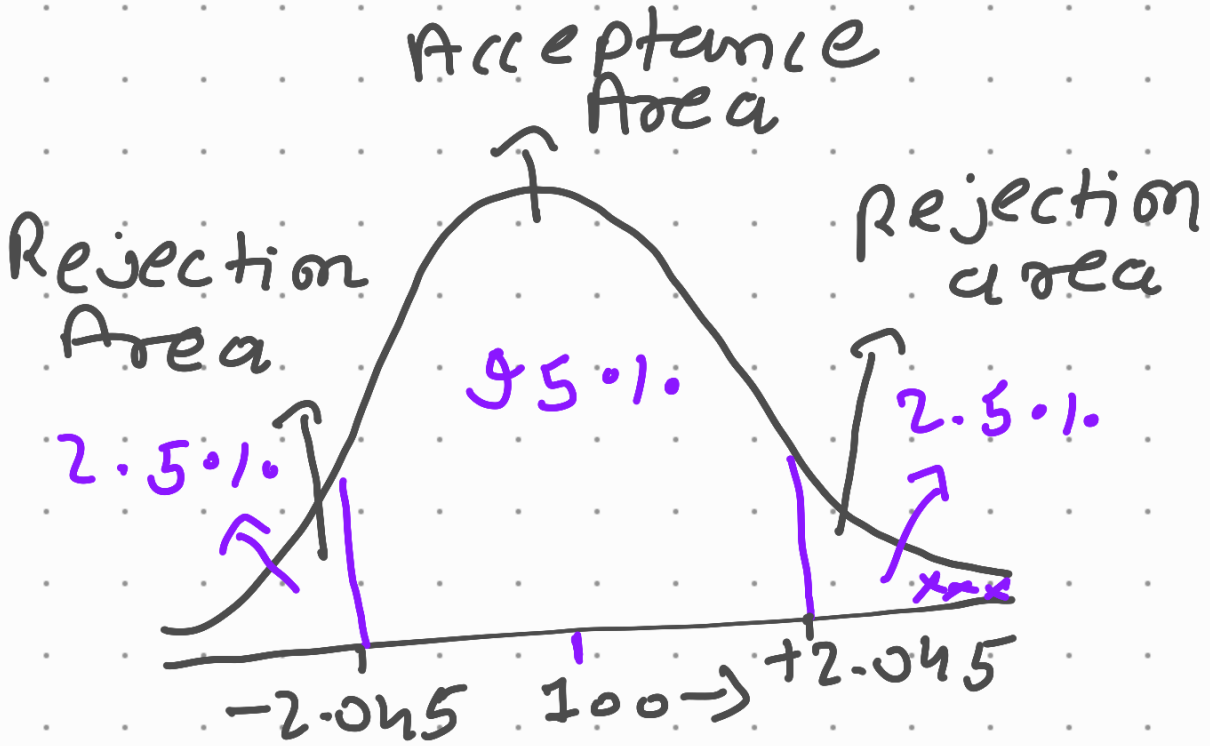
(b) Alternate Hypothesis $H_1 = \mu \neq 100$

{2-Tail Test}

(c) Degree of Freedom

$df = n - 1 = 30 - 1 = 29$

(d) Decision Rule



→ if t-test is less than -2.045 or greater than +2.045 we reject the Null Hypothesis.

(c) Calculate Test statistics

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

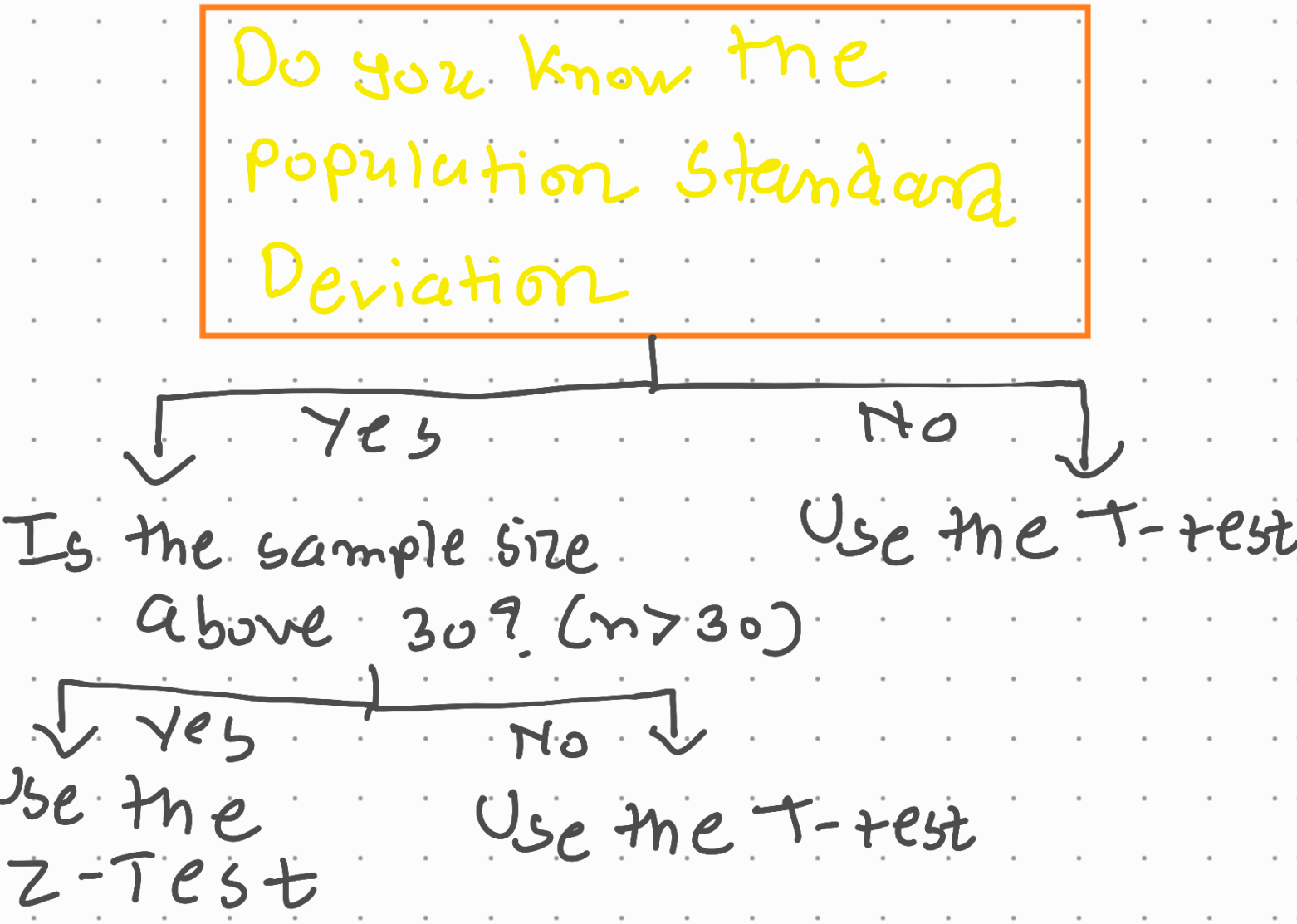
$$= \frac{140 - 100}{20/\sqrt{30}}$$

$$\boxed{t = 10.96}$$

→ Since $t = 10.96 > +2.045$, so & Reject The Null Hypothesis

→ Conclusion: Medication used has affected the intelligence. It has increased the IQ.

When To Use T-test Vs Z-Test



Type-1 and Type-2 Errors

Reality: Null Hypothesis is True or
Null Hypothesis is False

Decision: Null Hypothesis is True or
Null Hypothesis is False

Outcome-1 :- we reject the Null Hypothesis when in reality it is False → Good

Outcome-2 :- we reject the Null Hypothesis when in reality it is True → Type-1 Error

Outcome-3 :- we retain the Null Hypothesis, when in reality it is False → Type-2 Error

Outcome-4 :- We retain the Null Hypothesis, when in reality it is True → Good

Bayes Statistics (Bayes Theorem)

Bayesian statistics is an approach to data analysis and parameter estimation based on Bayes' theorem.

Bayes' Theorem

Probability → Independent Events
Probability → Dependent Events

① Independent Events

ex: Rolling a dice

$\{1, 2, 3, 4, 5, 6\}$

$P_3(1) = 1/6$

⋮ ⋮

$P_3(6) = 1/6$

② Tossing a coin

$P_B(H) = 0.5$

$P_B(T) = 0.5$

② Dependent Events

e.g. ① 3 yellow, 2 red marbles in a bag.

P_B(Get one red) = 2/5 after this

P_B(Yellow) = 3/4

→ P_B(Red and Yellow)

$= P_B(\text{Red}) * P_B(\frac{\text{Yellow}}{\text{Red}})$

↳ Conditional Probability

→ $P_B(A \text{ and } B) = P_B(B \text{ and } A)$

$= P_B(A) * P_B(B/A)$

$= P_B(B) * P_B(A/B)$

$$\therefore P_s(B/A) = \frac{P_s(B) \times P_s(A/B)}{P_s(A)}$$

↳ Bayes' Theorem

$$P_s(A/B) = \frac{P_s(A) \times P_s(B/A)}{P_s(B)}$$

A, B = Events

$P_s(A/B)$ = Probability of A given B is True

$P_s(B/A)$ = Probability of B given A is True

$P_s(A), P_s(B)$ = Independent Probabilities of A and B