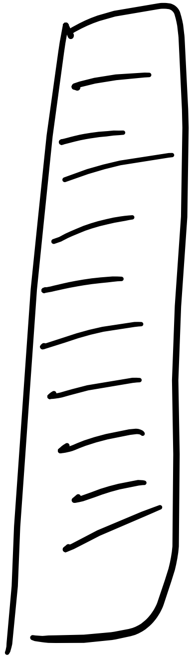
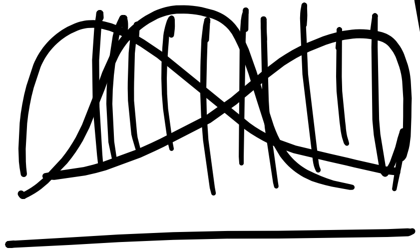
Central Limit Theorem:

Variable:



plot



size > 25

Enough samples

9/10

Samples — $S_1 \rightarrow$ Mean (S_1)

$S_2 \rightarrow$ Mean (S_2)

S_2

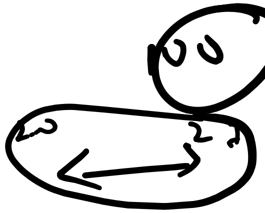
. . .

S_n

(sample mean)

plot the averages:

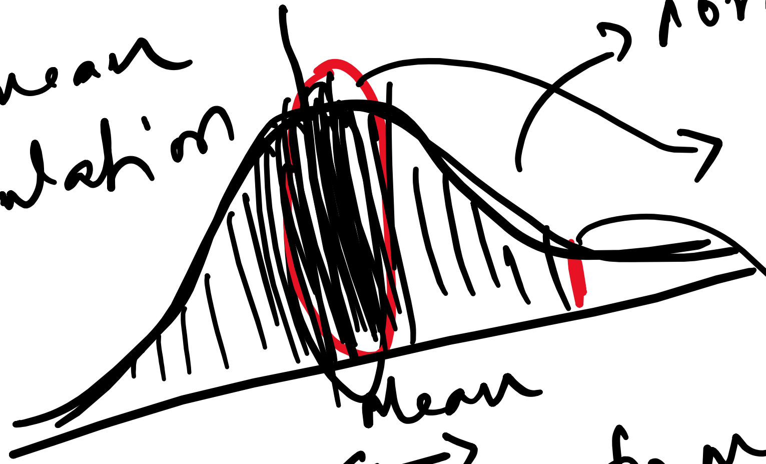
2x



$\rightarrow$ normal

Sample mean $\rightarrow$ pop. mean

Variation $\rightarrow$ population



Mean

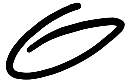
closer to mean

Statistic

⇕

pop.

M



Sampling:- Taking a few samples out of population

Sampling error: -> Different results

-> sampling error / variation due to sampling

- 'Always'

Population

↑

Statics - sample

Next month -> \$4.5

(\$4.2 - \$4.7)

→ Estimate the population parameter

↳ Good practice to give in a confidence Interval

Confidence Interval : Range ⟷ μ (lie)

What affects the width of C.I ?

1. The variation within the population ✓
2. The sample size ✓

1 Variation ⇒ ↓
Sample ⇒ ↓

$\boxed{C.I}$ → narrow

2 Sample size → CLT simulation

1000 = pop

N = 100

Heights

3.5

5.8

6.5

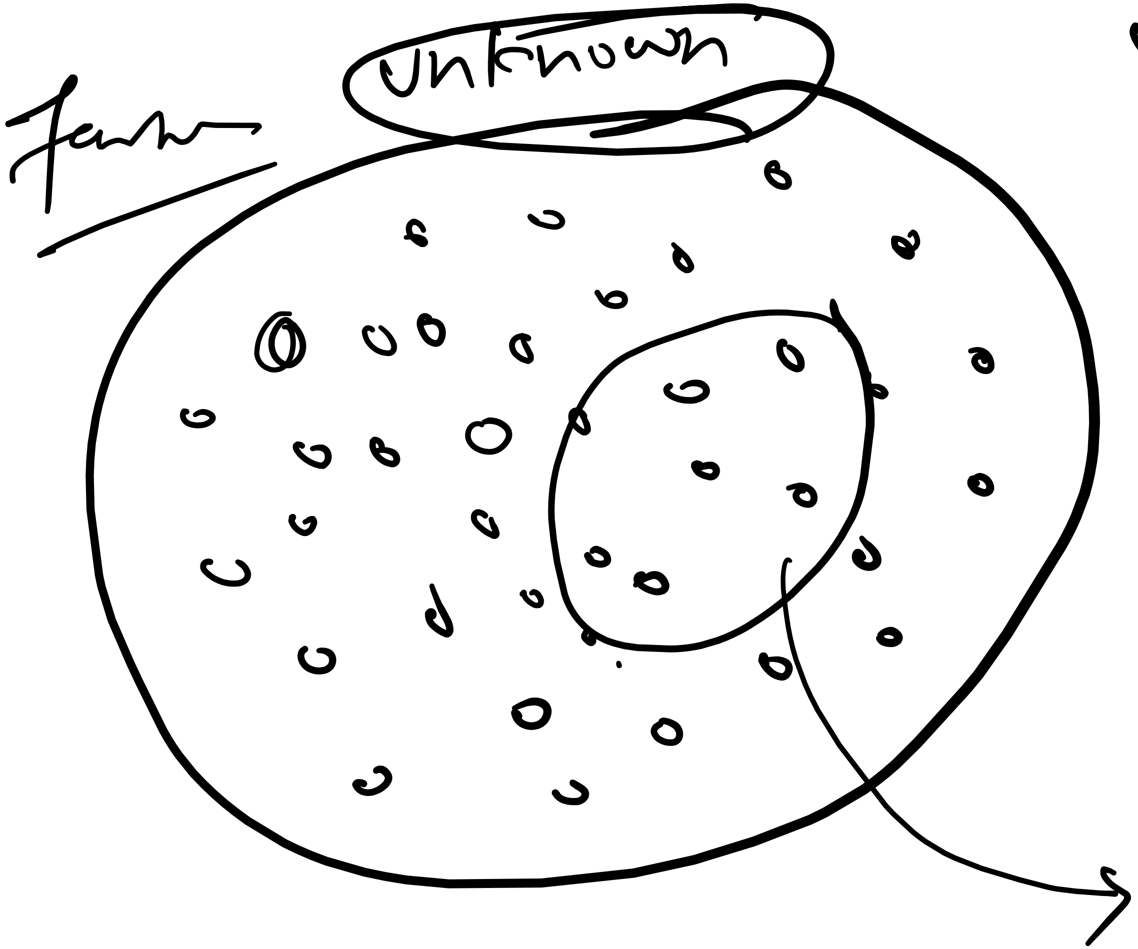
S1
S2
S4
S3
S..

↑ Sample

Similar

3.5
7.5

Average out



'Size of an apple'

sample variation can be an indicator of my pop var?

variation

Weight of samples

Avg? ↳ estimate

from CLT:

C.I

Confidence Interval of a mean $= \boxed{\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)}$

$\bar{x} \rightarrow$ sample mean

$t \rightarrow$ t-value

$s \rightarrow$ Std of sample

$n \rightarrow$ sample size

Level of confidence

sample size number

$n = 15$; $\boxed{\bar{x} = 149.3}$; $\boxed{\sigma = 4.758}$

$\Rightarrow \left(s/\sqrt{n} \right) = 4.758/\sqrt{15} = 1.22$

(std error)

90% → level

df = sample size - 1 = 14

Margin of error = 2.145 * 1.22 ⇒ 2.62

'spread' means to be'

$\bar{X} + 2.62$ & $\bar{X} - 2.62$

Range $\Rightarrow$ $\boxed{146.67}$ to $\boxed{151.9}$

95 % confident that the population parameter (μ) lies b/w above range,

Misconception:—

(Apple eg) (150 — μ — 170)

- The C.I is NOT the interval that will hold the weight of 95% of apples from the population

Df

100

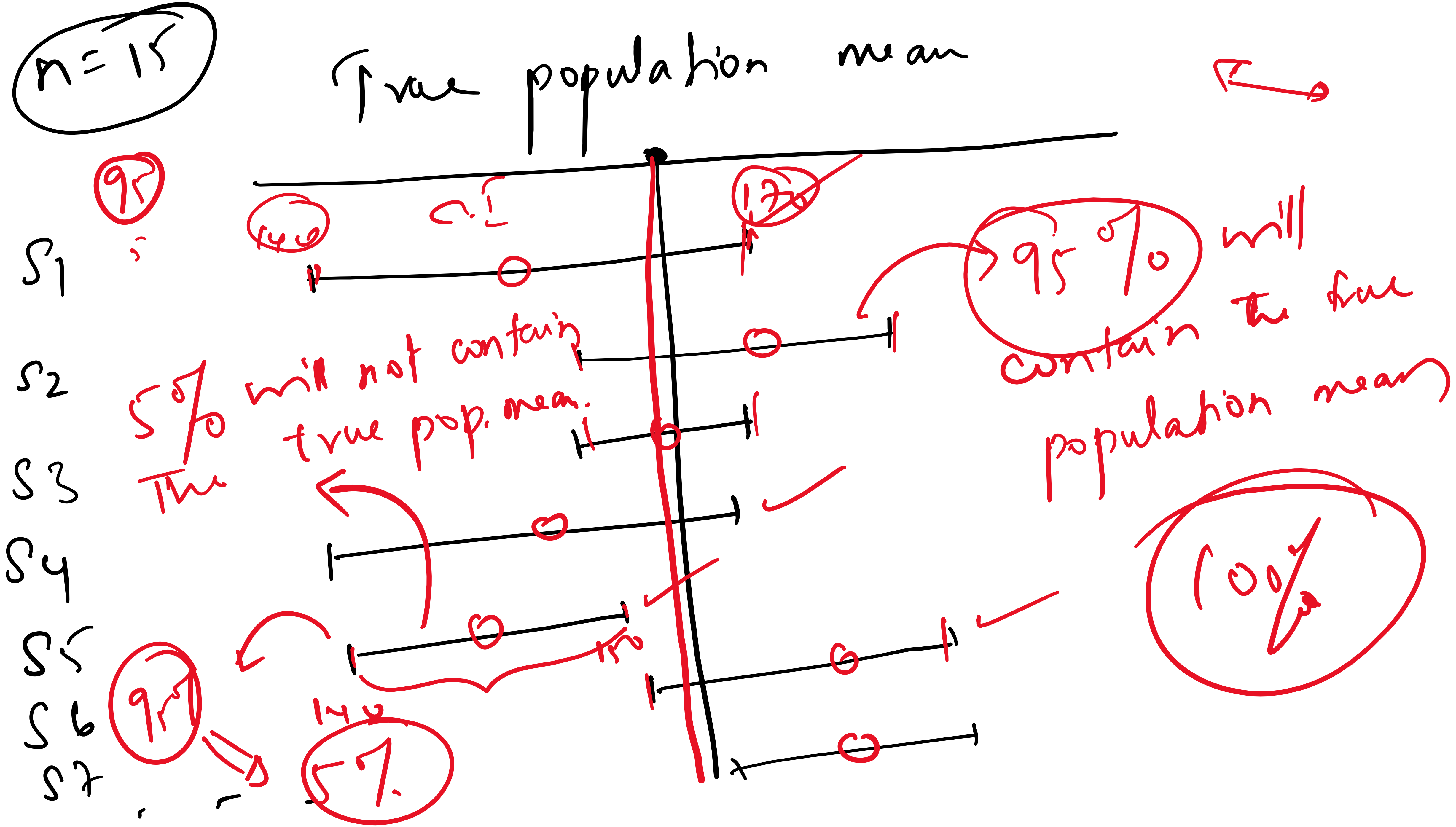
95

n = 15

Eg S1, S2, S3, S4, ... Sn

95% C.I

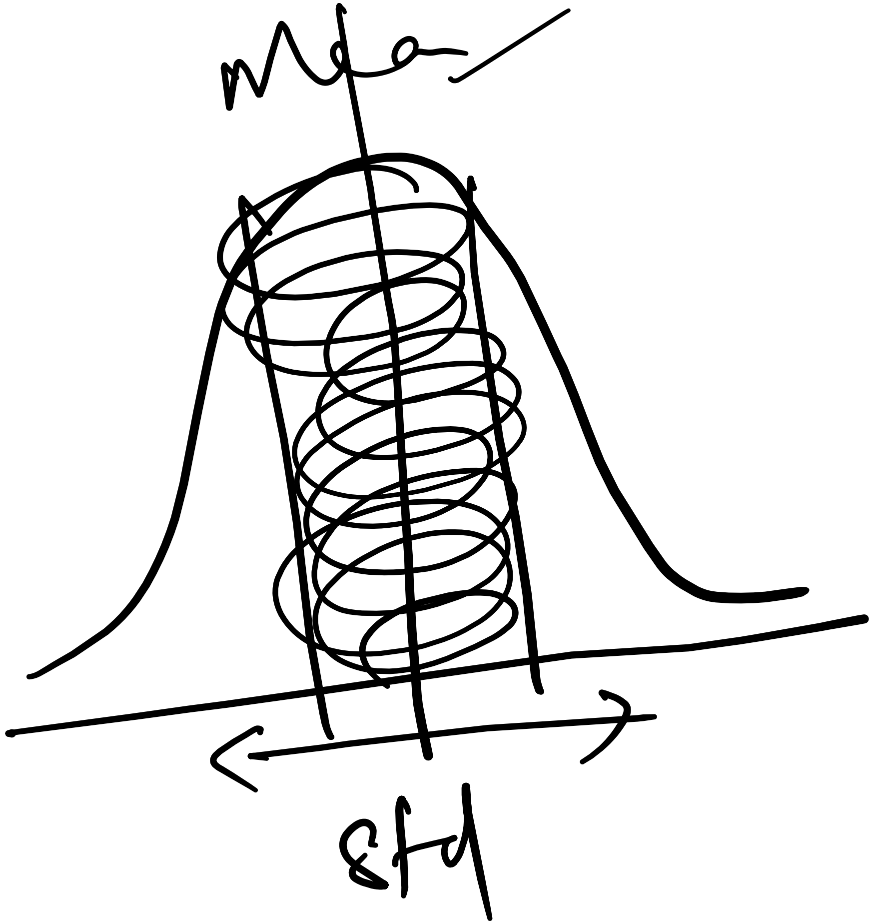
true population mean

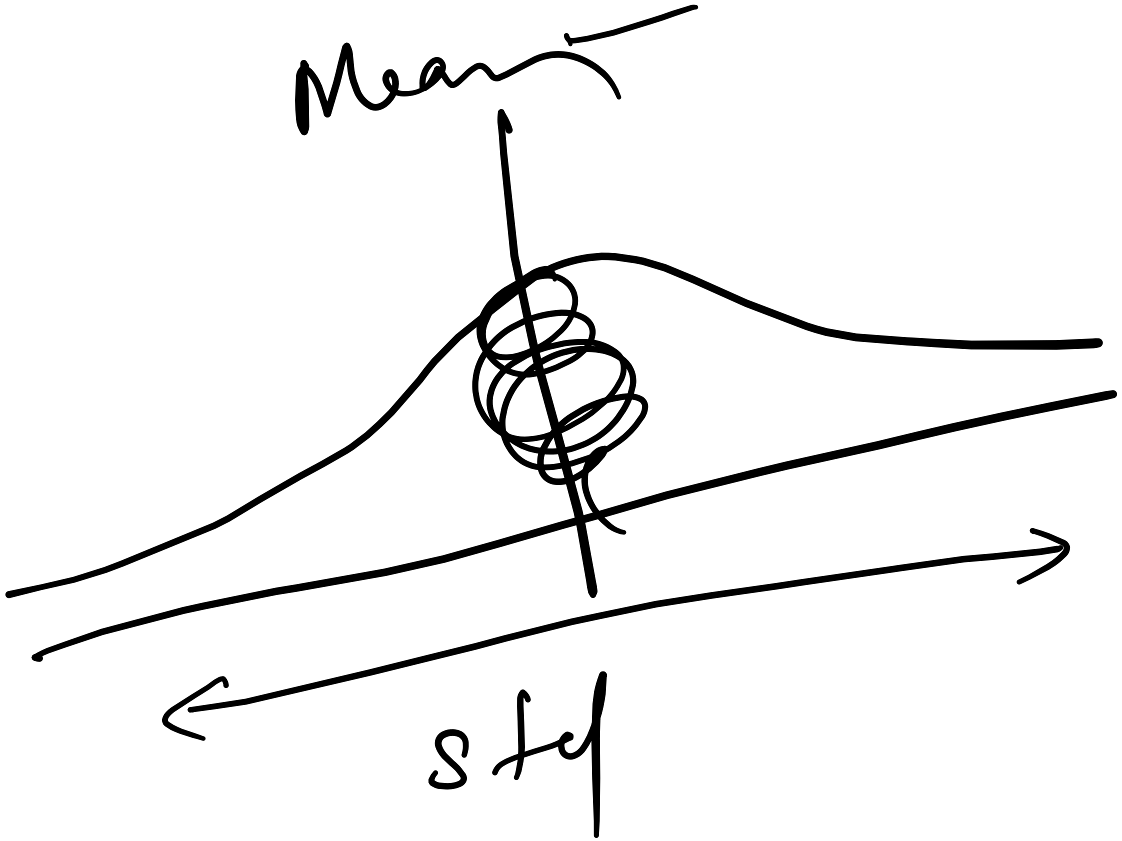


n=15
True population mean
95
140
C.I
170
S1
S2
S3
S4
S5
S6
S7
95% will contain the true population mean
5% will not contain the true pop. mean.
The
100%
150
95
140
5%

Normal Distribution:

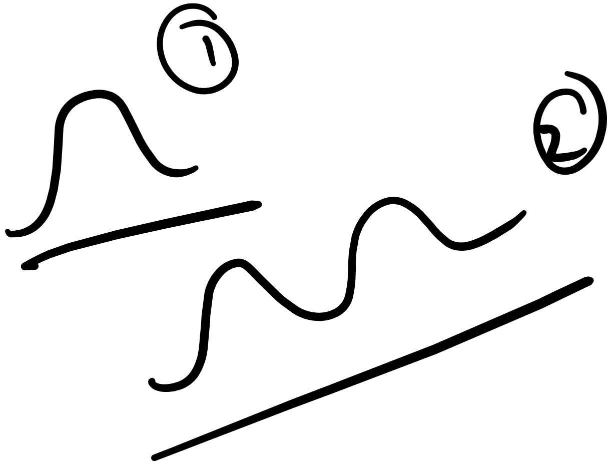
Parameter		Statistics
A number that describes the data from a population.	Mean	A number that describes the data from a sample
μ / σ	Std	$\bar{x}$ / s





size ↑

Characteristics of N.D:-

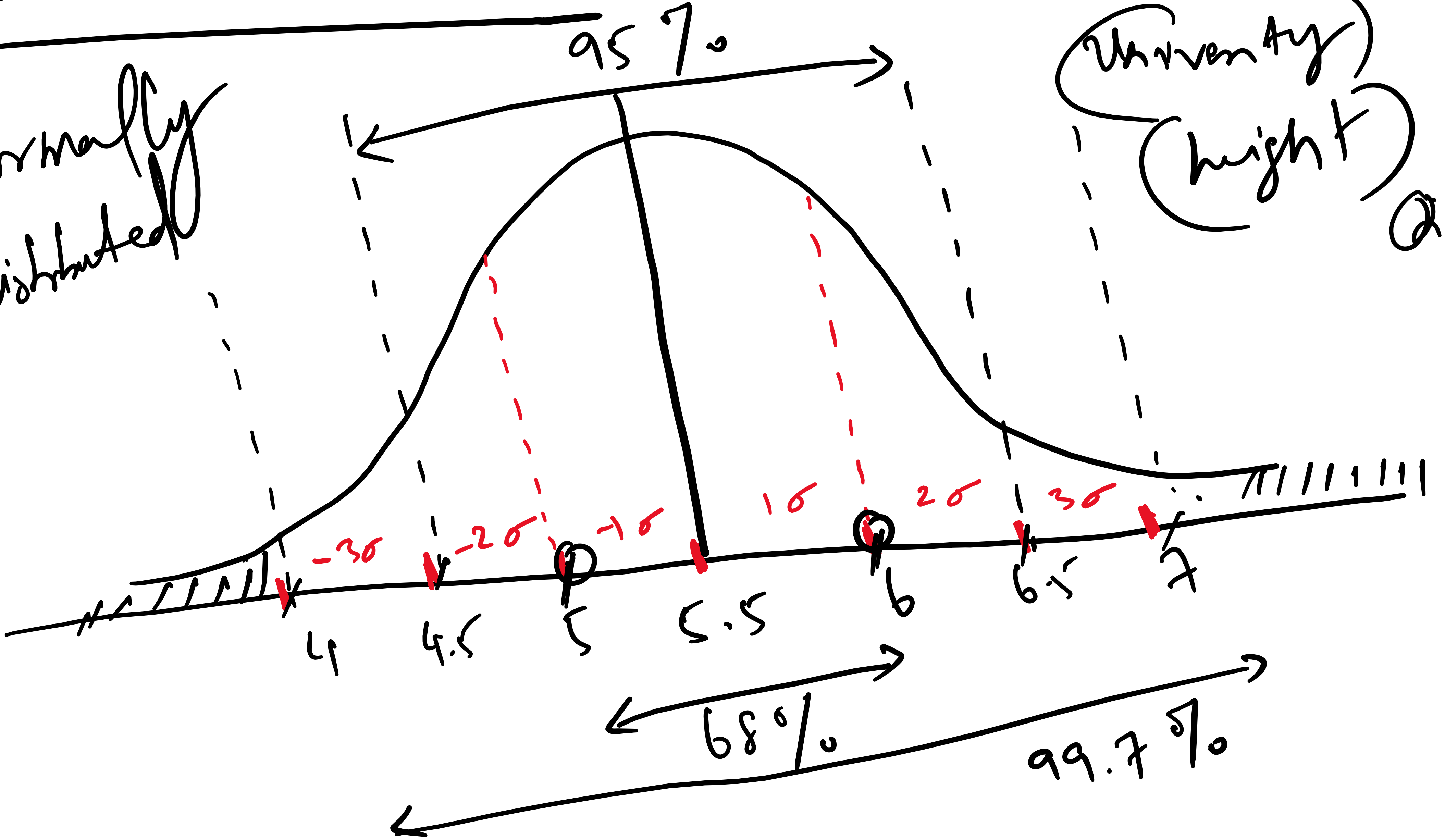


- It's a unimodal
- Symmetric about its mean
- The μ & σ normal distribution

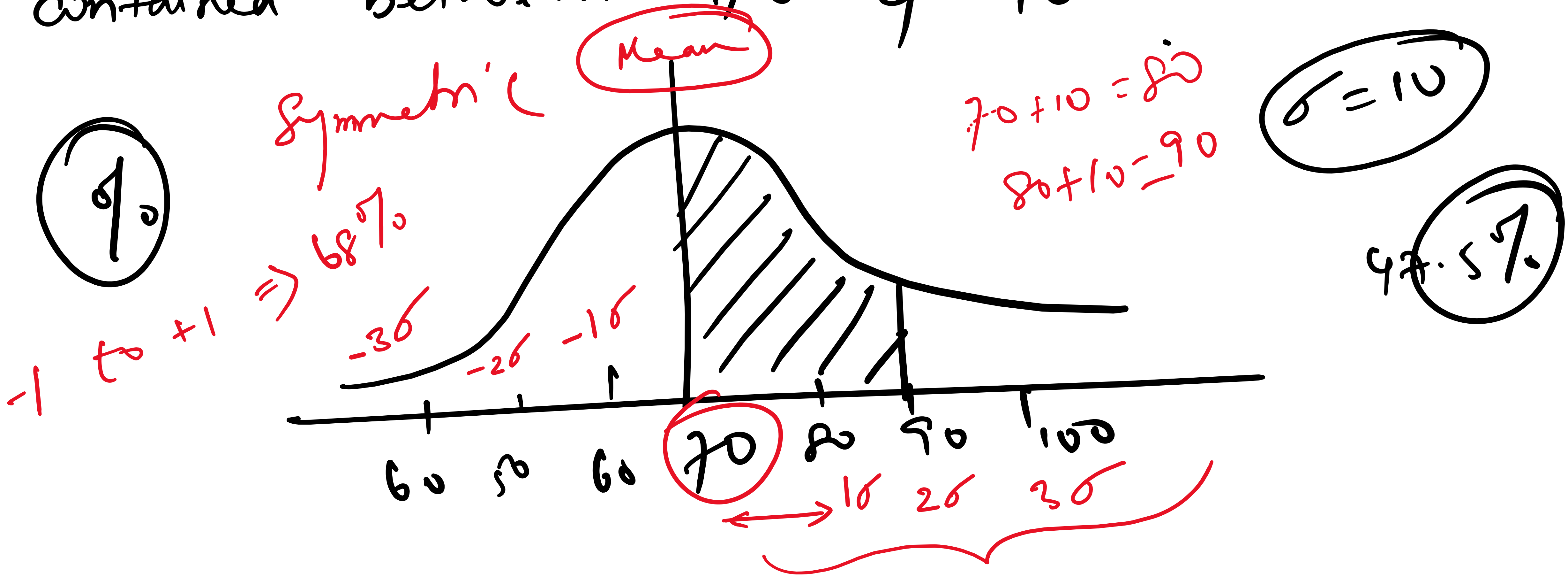
68 - 95 - 99.7 Rule :- $\mu = 5.5$ feet ; $\sigma = 0.5$ feet

Normally distributed

(University height) Q.3



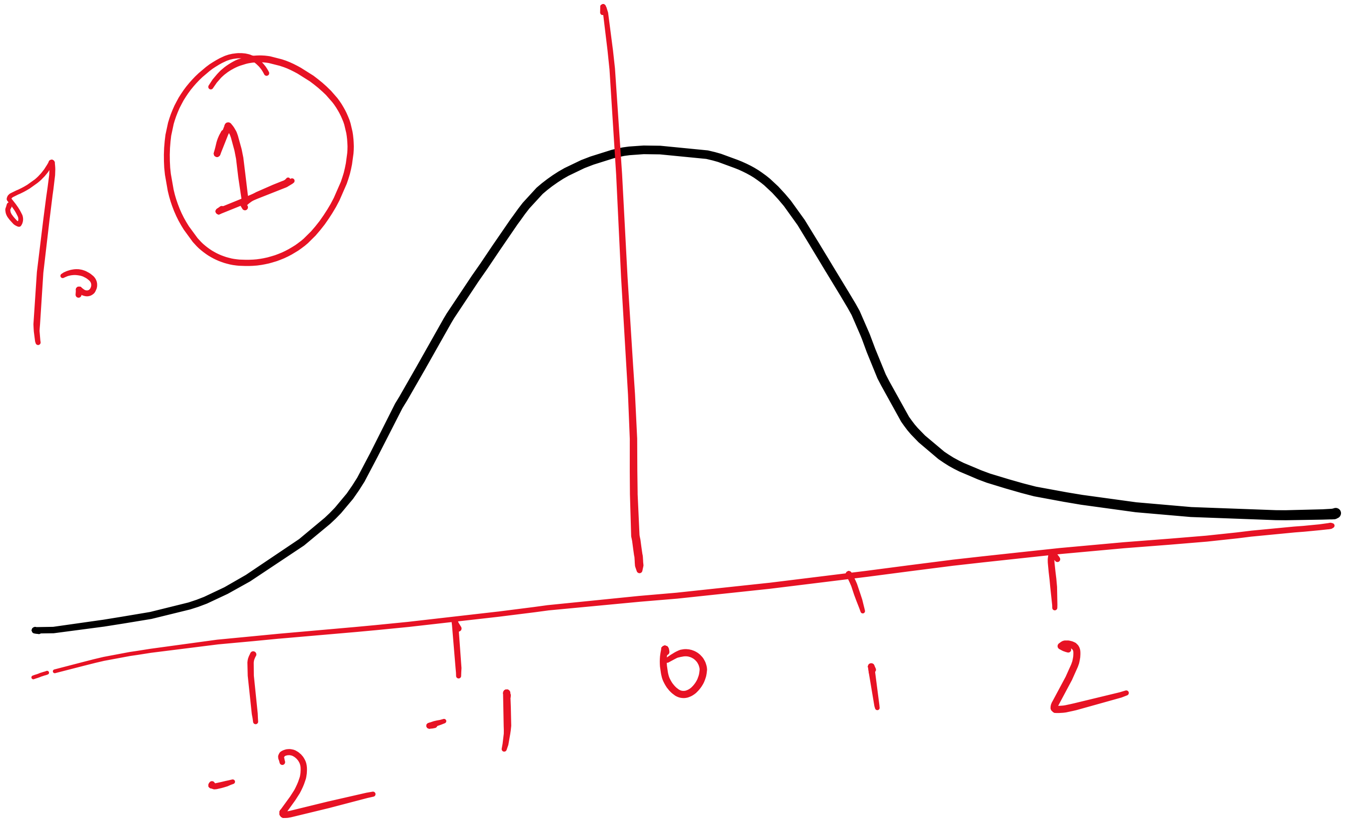
① The normal distribution has a $\sigma = 10$; $\boxed{\mu = 70}$; approximately what area is contained between 70 & 90?



81.5 % ①

$\mu = 0$

$\sigma = 1$

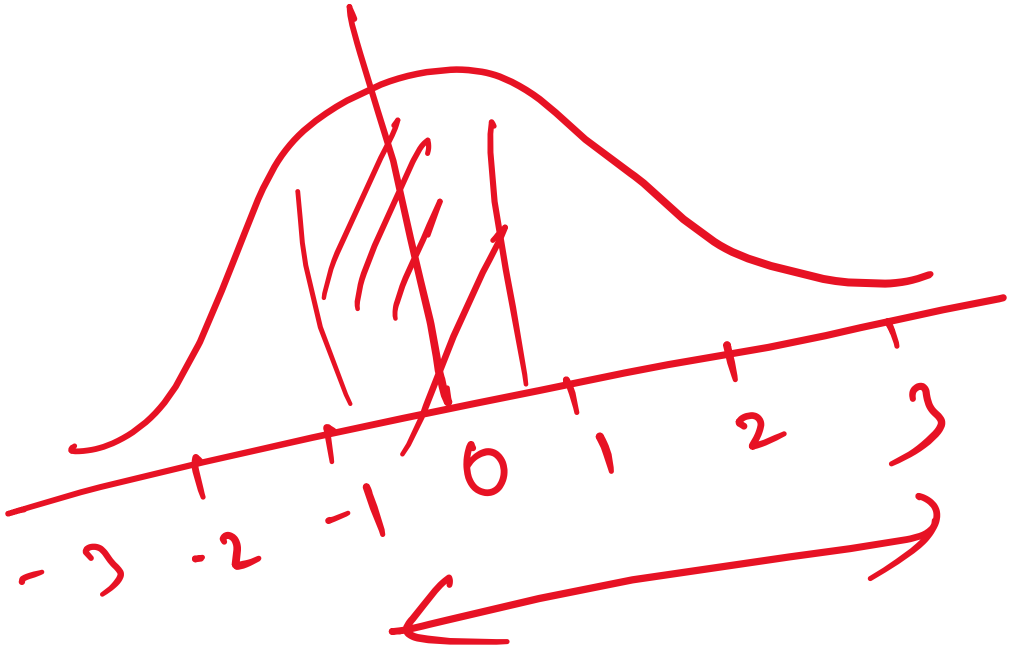


② Find the area b/w $-2 + 1$?

Z-score & Standardization:- (ML preprocessing)

Standard normal distribution?

$\mu = 0$; $\sigma = 1$



Z-score → -2

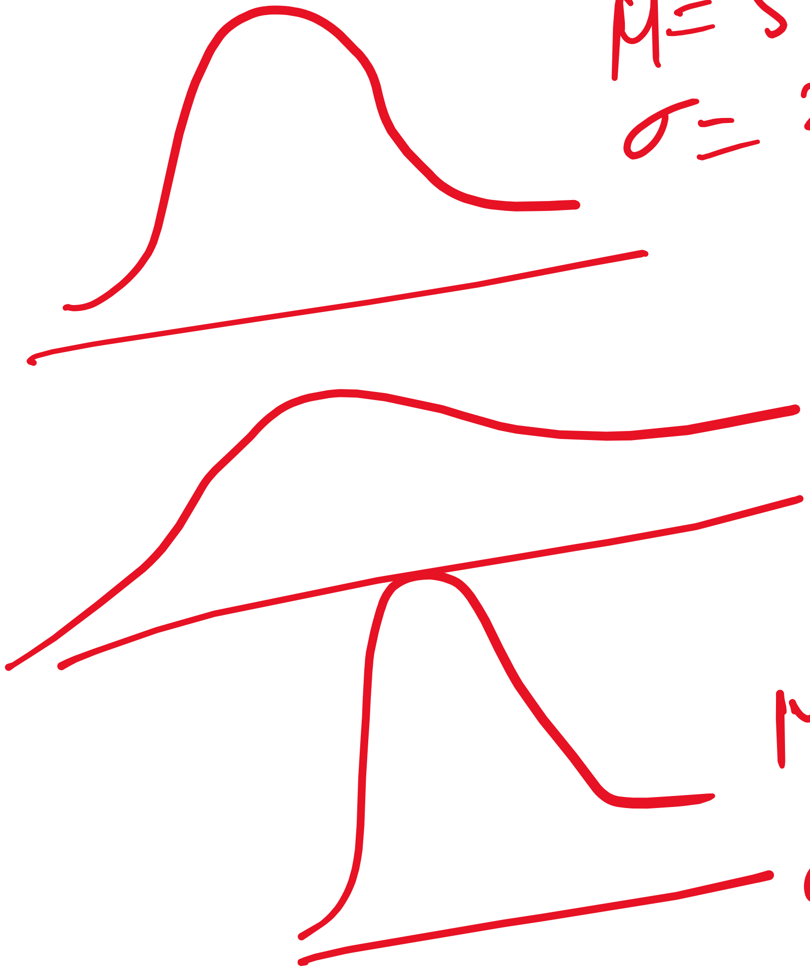
2 std

- Each number on the x-axis corresponds to 'z-score'

- z-score tells us how many 'σ' an observations is away from the mean

→ Calculate the area of distribution with the help of a z-table

Normal distribution



$\mu = 5$

$\sigma = 2$

$\mu = 16$

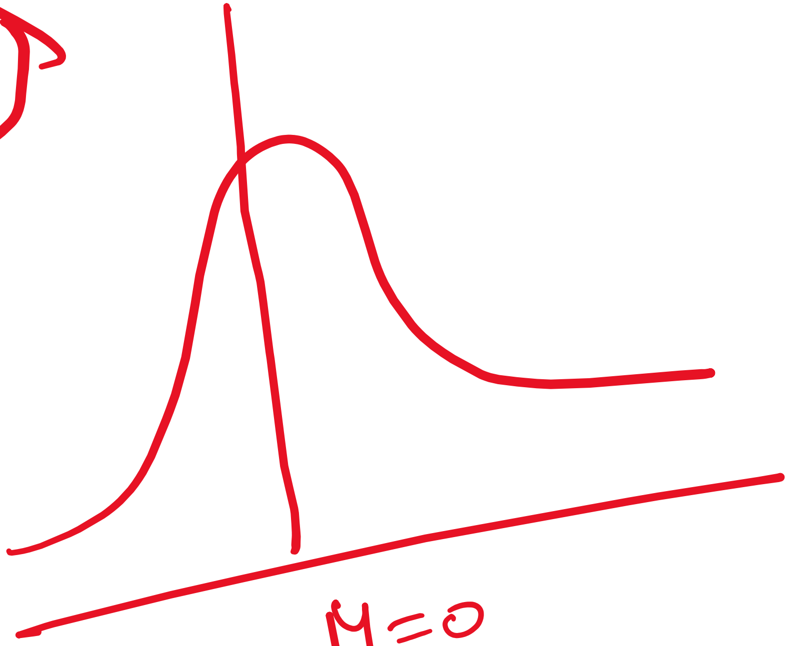
$\sigma = 3$

$\mu = 15$

$\sigma = 1$

Standardization

Std normal dist



$\mu = 0$

$\sigma = 1$

F Y

Formula: $\dfrac{x_i - \mu}{\sigma}$

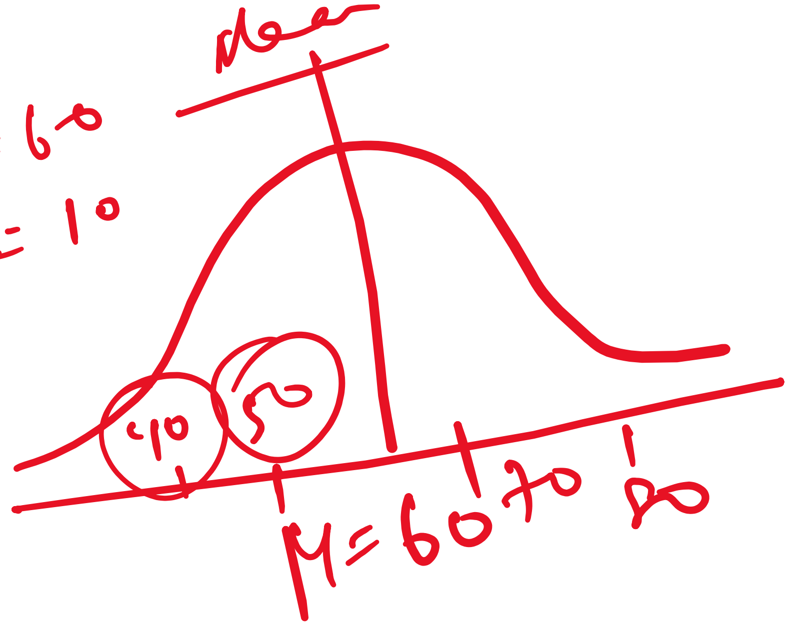
$x_i \rightarrow$ Observation

$\mu \rightarrow$ Population mean

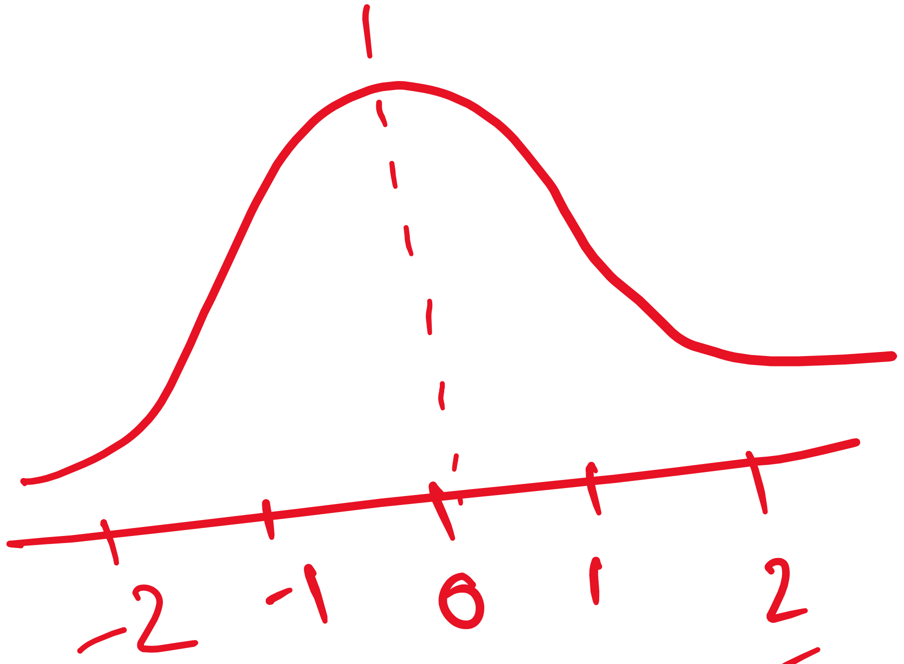
$\sigma \rightarrow$ " std

Eg: 1 : Class X ; scores of marks of students follow a normal distribution $\mu = 60$; $\sigma = 10$

$\mu = 60$
$\sigma = 10$



$$\boxed{Z = \frac{X_i - \mu}{\sigma}}$$



$$\frac{40-60}{10} = \frac{-20}{10} = -2$$

$$\frac{50-60}{10} = -1$$

Std norml dist

$\mu = 0 \quad \sigma = 1$

Q: What population of students scored less than 49 on the exam?

Std

-3.49

$.09$

$P(x < 49)$?

$$\boxed{Z = \frac{x_i - 60}{10}} \Rightarrow \frac{49-60}{10} = -1.1$$

$P(Z < -1.1) \Rightarrow$ Look up in the z table

$= 0.1357$

→ How much area is associated with the z-score → calculated.

formula ?
Std.

→ z value of -1.1 is '0.13'

$\mu = 0$

$\sigma = 1$

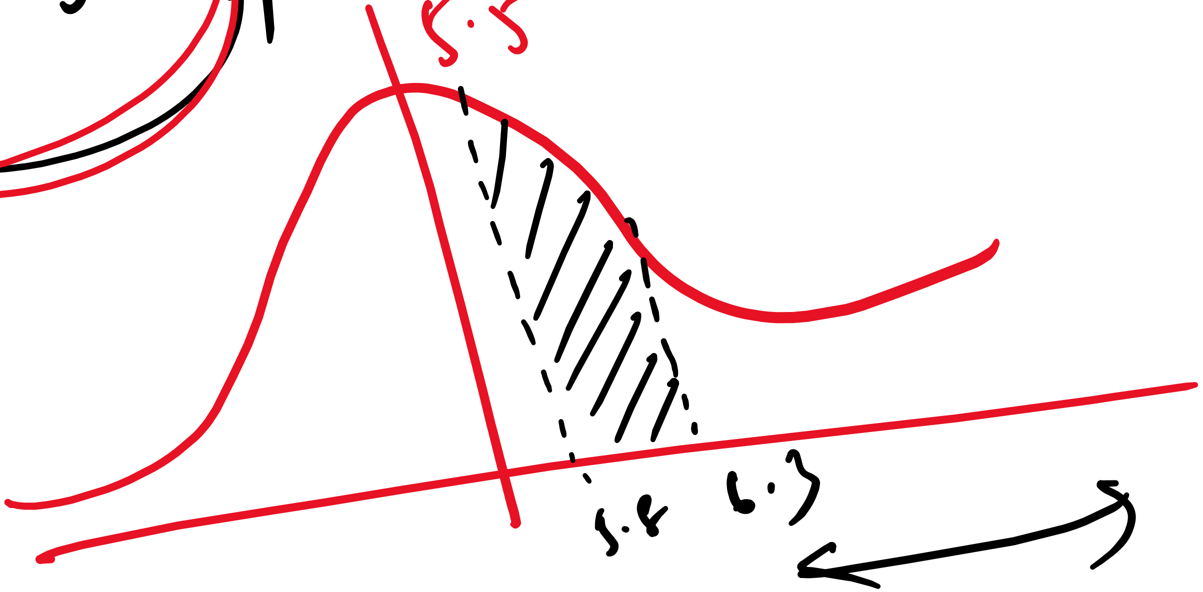
0.13

population

Eg: 2 Heights of students

$\mu = 5.5$ feet ; $\sigma = 0.5$ feet

Q: What proportion of students b/w 5.81 and 6.3 feet tall



1. Standardization
2. Z value.

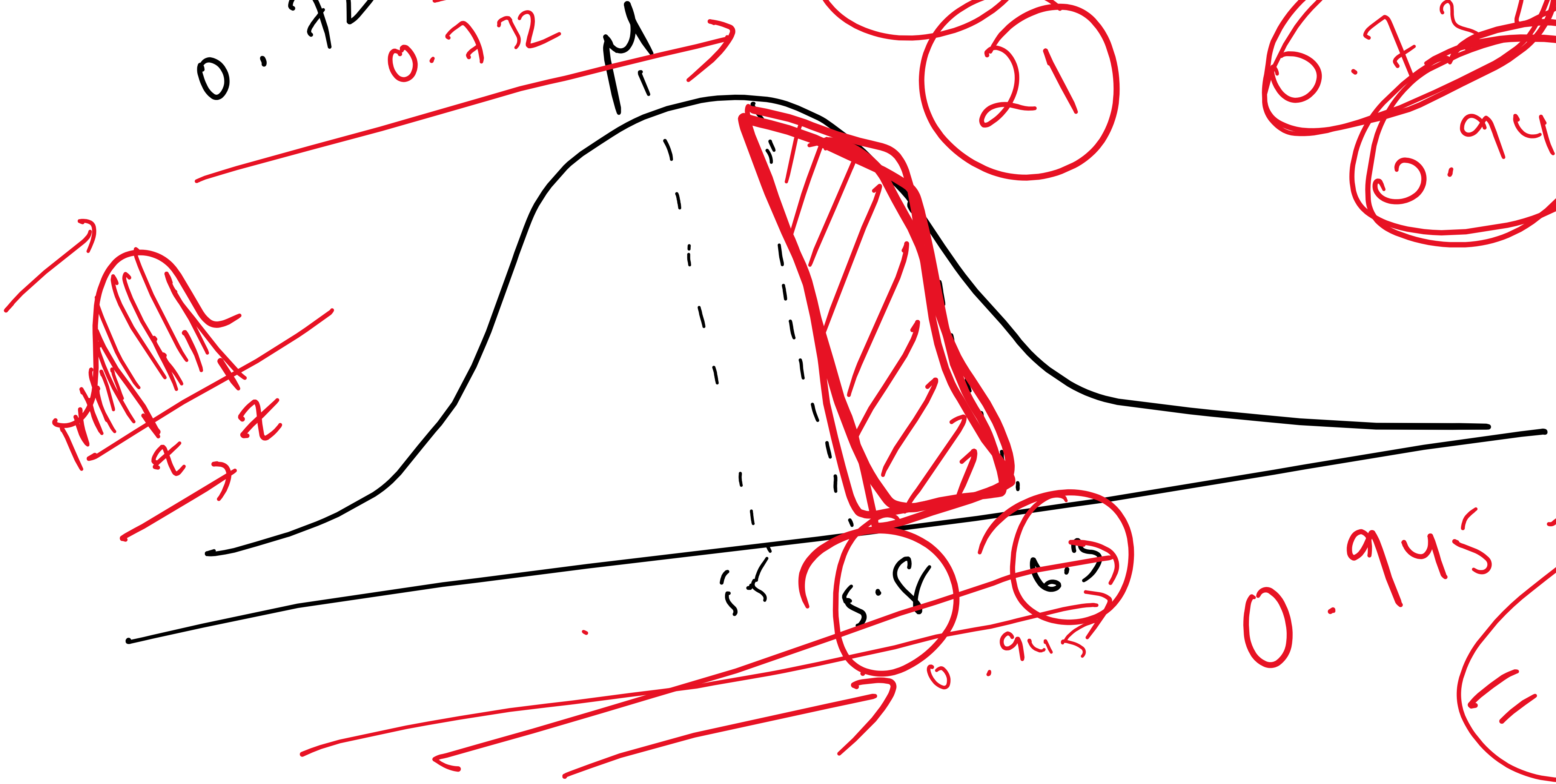
0.945 to 0.7257

Z-value

100

21

0.62 to 1.6



0.732

0.945

low

high

0.945 − 0.732

= 0.213

X = Age 25, 35, 45

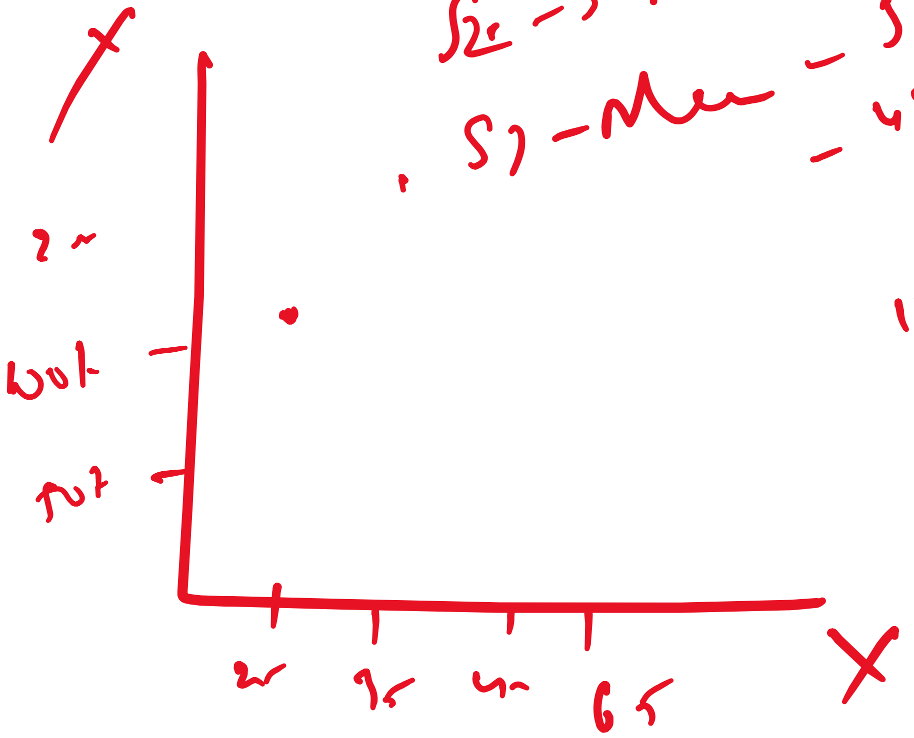
Y = Salary 100k, 200k, 300k

S1 — Mean — 50

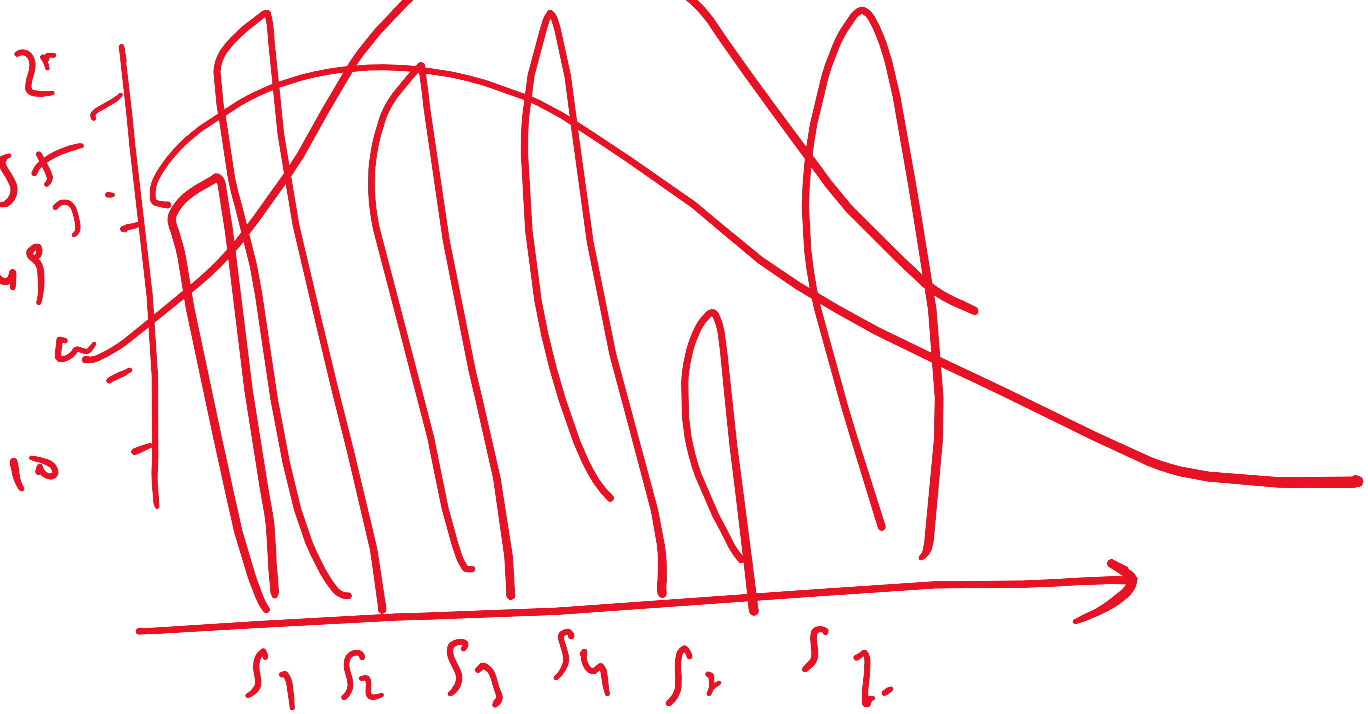
S2 —> Mean — 52

S3 — Mean — 55

— 49



Density / Distribution



Hypothesis Testing :- (Assumption)

Covid

D D

1. Person 1 — 3
2. Person 2 — 4

Med 1

X

α

Med 2

1 / 2

65 35

60 40

55 45

Sklearn

H_2O.ai

fast.ai

P value

n = 200

X Med 2

n = 100

Med 1

Cure	Not
97	3

S1

Med 2

Cure	Not
3	97

S2

Med 1

C	N
73	125

S3 37%

C	N
59	131

S4 31%

Eg:2 'Snickers'

— 200g ; 70g of nuts

$H_0 \Rightarrow 70g$

$H_a < 70g$

H_1

S1 S2 S3 S4

70.1 69.2 69 71.1

$\overset{V_1}{D} = \overset{V_2}{D}$ ✗ Ha ✗

α

$\alpha >$

T

$V_1 \neq V_2$

$V_1 < V_2$

$V_1 > V_2$

$\frac{200}{100}$

$M = 68.7g$

'p'value = 0.18

18%

69.9 1m

69.4 70

68.2 70.1

69.4 2.

68.9

Hypothesis Testing: $H_0 \rightarrow$ Null ; $H_a \rightarrow$ Alternate

- Frame a hypothesis - (about a population)
- Set a significance level (α)
- Sampling
- Calculate 'p' value
- Decision (Accept / Reject the H_0)

Z-test :-

Scenario :- A complaint that the students are malnutritioned

statistical data chart

Age = 10 ; $n = 25$; $\mu = 32$ kg ; $\sigma = 9$ kg

Sampling :- $\boxed{\bar{x} = 29.5 \text{ kg}}$

A vs. B

32 ≠ 29.5 / 31.

Proof

1. $H_0 \Rightarrow \mu = 32$ (No significant diff)

$H_a \Rightarrow \mu < 32$ (significant differen)

$\neq$, $<$, $>$

(2) $\boxed{\alpha = 0.05}$ *   0.025   0.10

95%   97.5   90%

(3) Samplin $\mu \leq 32$

(4) Calculate the z value

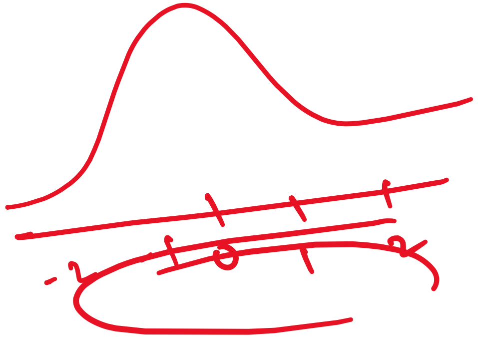
$\bar{x} \rightarrow$ sample mean

$\mu \rightarrow$ pop. mean

$\sigma \rightarrow$ " std

$n \rightarrow$ sample size

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



$$= \frac{29.5 - 32}{9/\sqrt{25}} = \frac{-2.5}{9/5} = \frac{-2.5}{1.8} = -1.39$$

$\boxed{Z = -1.39}$

1 → −1.3

② .09 → Y

→ find the p-value (from the Z table)

$\boxed{\alpha = 0.05}$

5%

3%

2%

$\alpha = 0.02$ Reject

$p(z = -1.39) = 0.0823$

$\boxed{p = 0.0823} > 0.05$ Accept

'If p value is low; Null must go'

$p(z) < 0.05$;

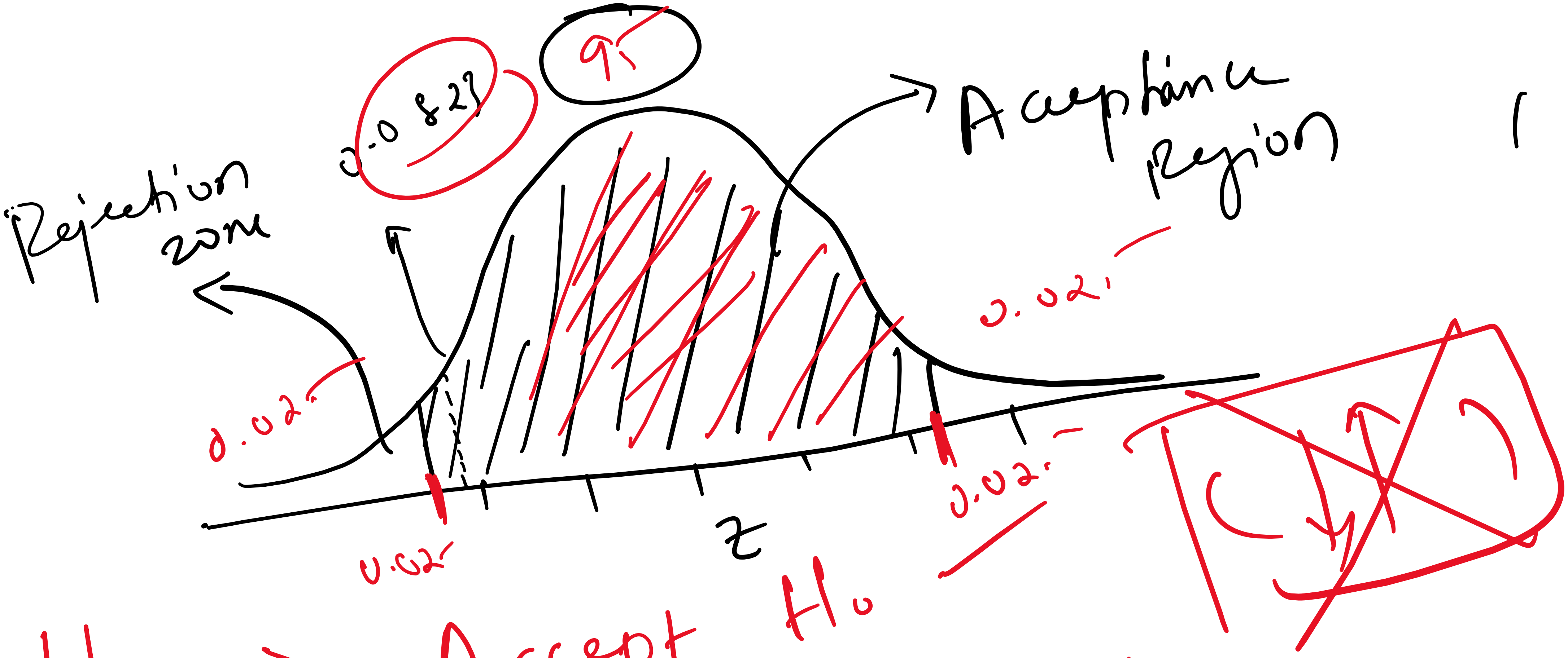
$p(z) > 0.05$;

95%

95%

90%

$\alpha = 0.10$



$H_0 \Rightarrow$ Accept H_0

$H_a \Rightarrow$ Reject H_a

One tail vs two tail:

① $H_0 \Rightarrow \mu = 100g$ ✓ ×

② $H_a \Rightarrow \mu \neq 100g$

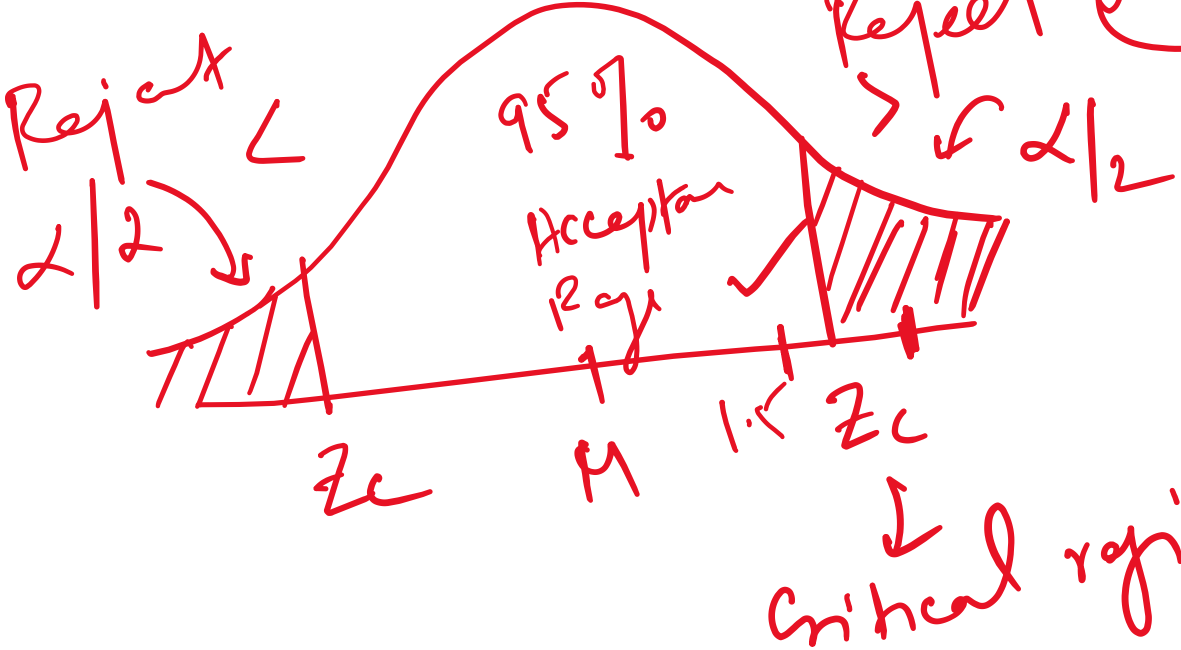
$\boxed{\alpha = 0.05}$

$0.025 < 0.05$ (α)

Sampling:

Z value

P value



✓

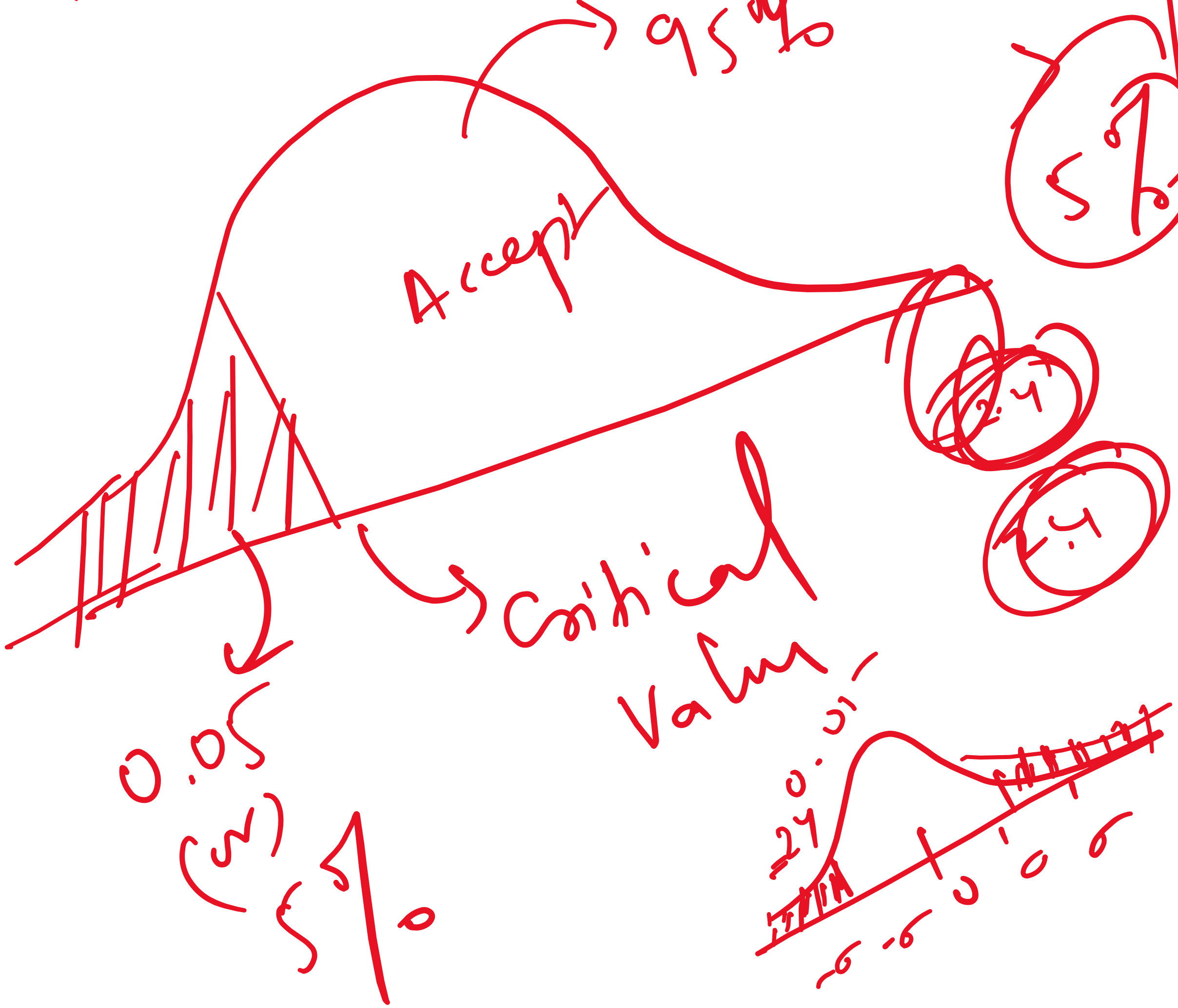
$\boxed{-0.05 < Z < +0.5}$

95%

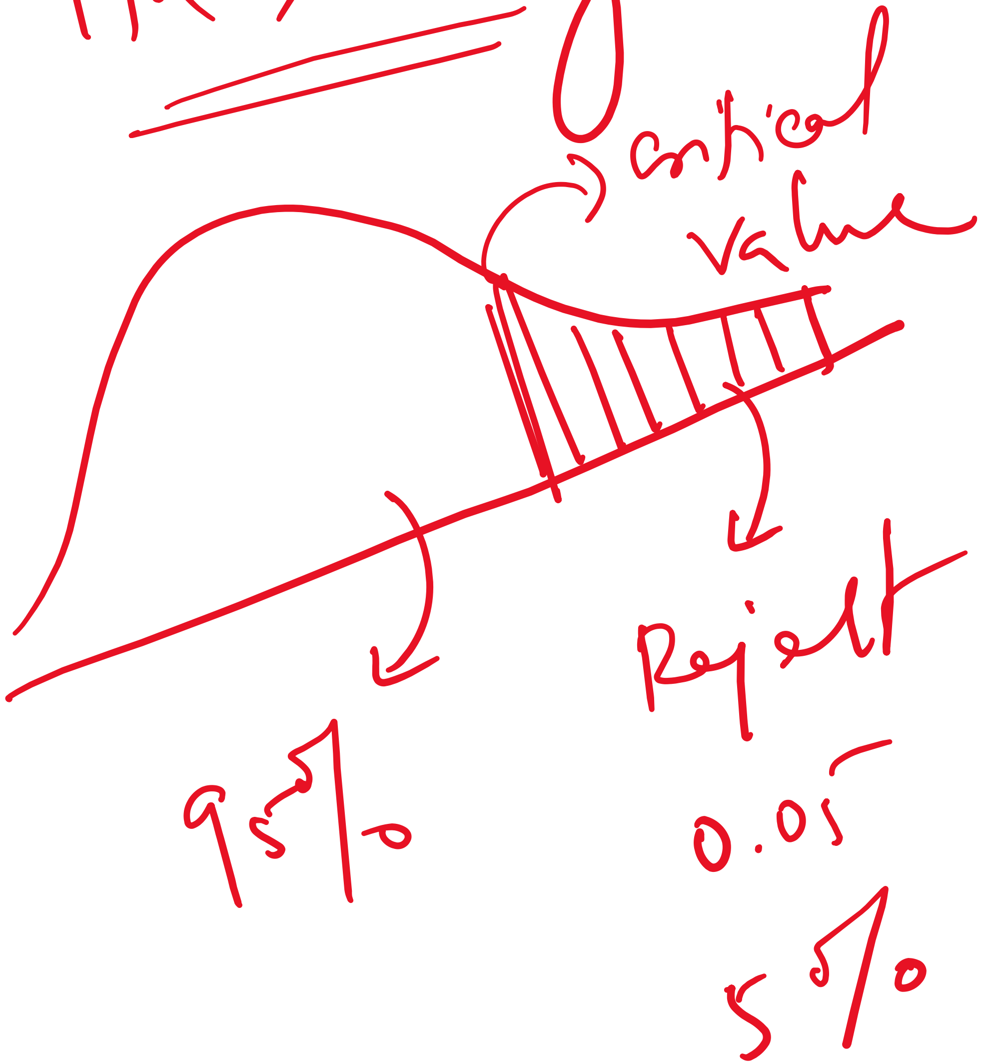
Ha < 100g

$\boxed{\alpha = 0.05}$

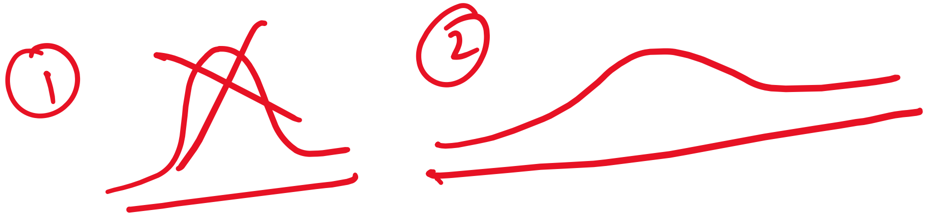
Ha > 100g



5%



T- Distribution:



Z	vs t

- Works with normal distribution
- **Small** samples
- For large samples, t-dist = z-score (normal distribution)
- T-Dist ⇒ high standard deviation.

→ if $n > 30$; 't' similar to 'z'.

Parameter → take a sample → Calculate statistic

Param | Stats

μ σ

For eg:- 'Mean' ($\bar{x}$)

→ Add a 'margin of error' to our statistic ($\bar{x}$)

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

σ → population s.d

s → sample std

⇒ Statics $\pm$ $t^* \frac{s}{\sqrt{n}}$

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

$$\boxed{\bar{x} \pm t^* \frac{s}{\sqrt{n}}}$$

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

→ 'σ' → 's' ; High variability

→ df ⇒ sample size - 1 ; $n \uparrow \Rightarrow$ t ⇒ 'z'

Constructing a C.I. (95%) for pop. mean :-

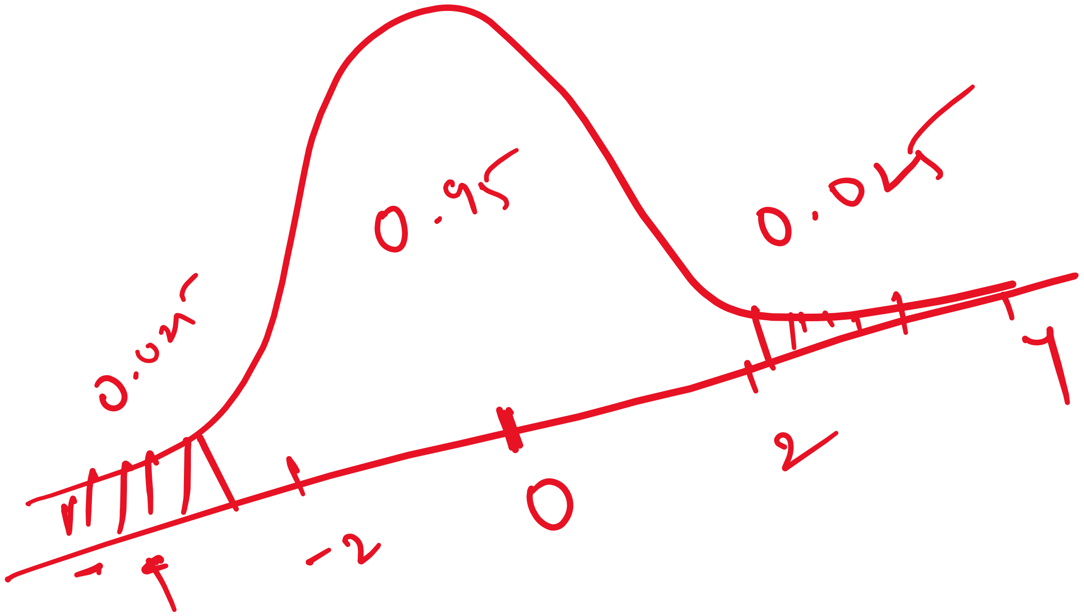
→ for 't' distribution ; 'σ' is unknown

If 'σ' is known

(Z)

$\bar{x} \pm Z * \sigma/\sqrt{n}$

$\boxed{\text{Regardless of 'n'}}$



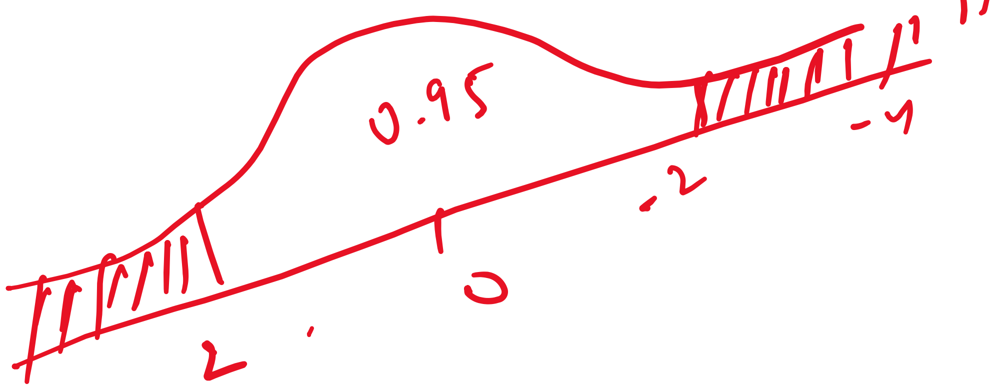
If 'σ' is Unknown

(t)

$\bar{x} \pm ? * S/\sqrt{n}$

$\bar{x} \pm t * S/\sqrt{n}$

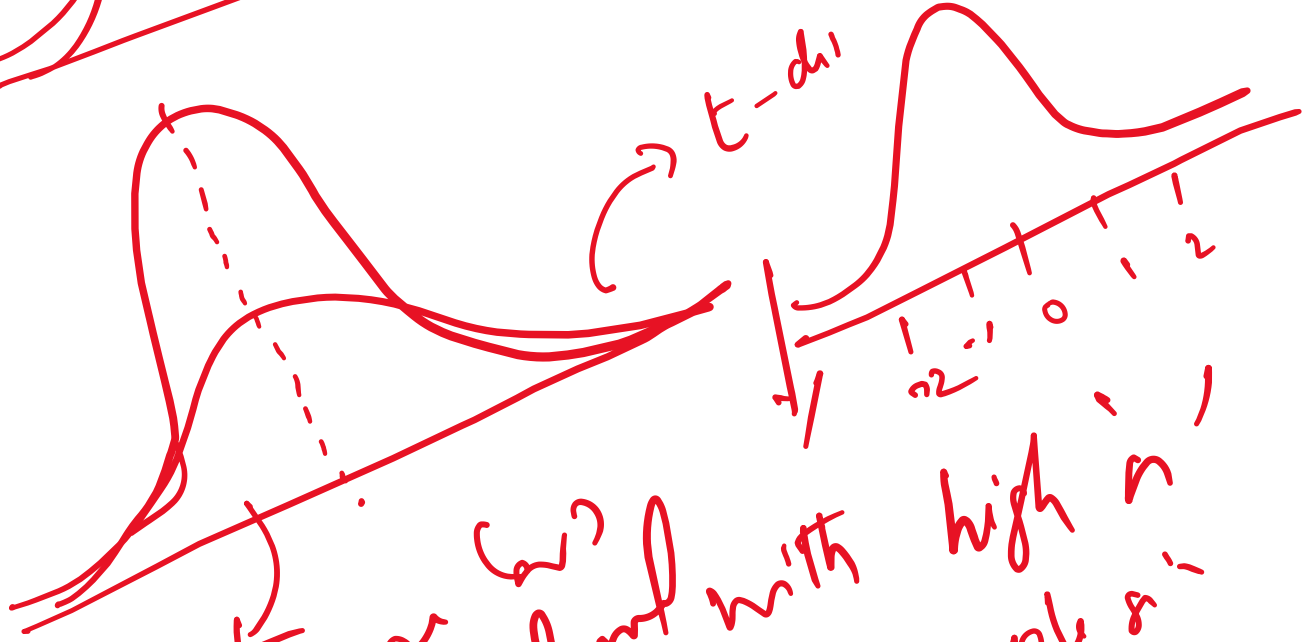
'less sample' if I have more 'n'



$t_{0.025}$ for (97.5% C.I)

n	df	$t_{0.025}$
6	5	2.571
11	10	2.228
31	30	2.042
61	60	2.000
81	80	1.990
101	100	1.984
∞	∞	1.960

C.I	z-score
90	1.65
95	1.96
99	2.5



z-score (or) standard dist with high 'n' sample's

Misconception:

'If $n > 30$; forget 't'; just use 'z' ~~X~~

$df > 30$; t-value $\Rightarrow$ 2.042; $z = 1.96$

$df = 100$; $t \Rightarrow$ 1.98; $z = 1.96$ (2)

'Margin of error'

diff (margin error)

$\boxed{\text{'t' test not 'z'}}$

C.I. $\bar{x} \pm z * \frac{s}{\sqrt{n}}$

(or)

$\bar{x} \pm t * \frac{s}{\sqrt{n}}$