Inferential Statistics

1.Hypothesis Testing

2.P value

3.Confidence Interval

4.Significance Value

@datascience_2.0

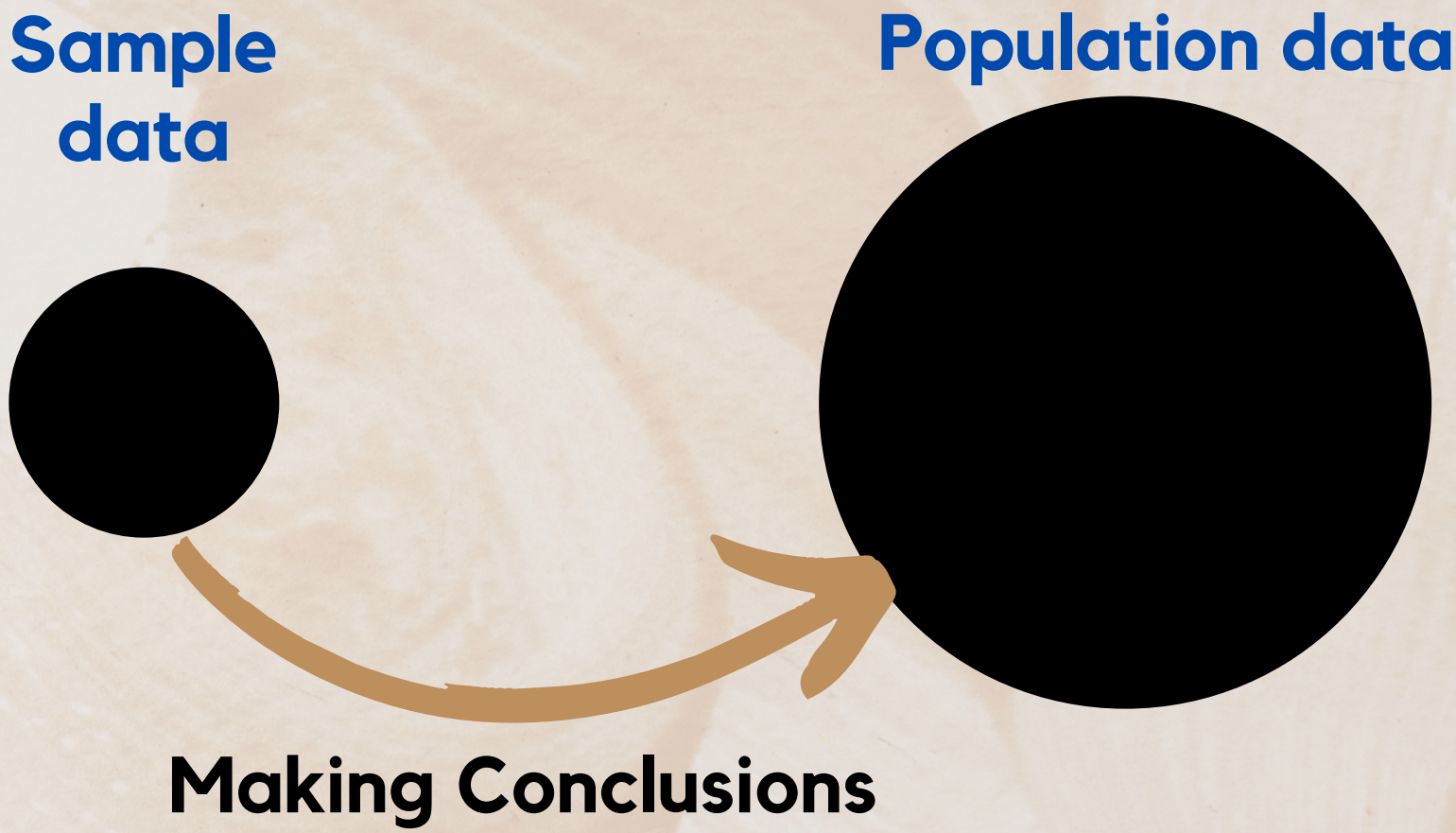
what is inferential statistics?

It consists of collecting sample data and making conclusions about population data using some experiments.

Making conclusions can be done by **Hypothesis testing**

ex: In a university there are 500 students.from it we are selecting a class of 60 students.

with the help of 60 students we are finding the average of them and making the conclusions about 500 students.



In order to validate and support these assumptions and conclusions we use **Hypothesis testing**

what is Hypothesis testing?

Investigating our ideas about the world using statistics.

Steps of hypothesis testing

Step1: State null and alternate hypothesis.

Step2:Collect data

Step3: Perform a statistical test

Step4: Decide whether to reject or fail to reject your null hypothesis

Step5: Conclusions

@datascience_2.0

Step 1:

Null hypothesis:

Default conclusions goes on null hypothesis.

(or)

A null hypothesis is a statement, in which there is no relationship between two variables.

ex: by default person is not a criminal untill unless he proved.

ex: by default coin is fair.

Alternate hypothesis:

Opposite of null hypothesis

(or)

An alternative hypothesis is statement in which there is some statistical significance between two measured phenomenon.

ex: Person had comitted crime

ex: coin is not fair.

@datascience_2.0

Step 2:

Collect data:

Collect evidences like DNA, fingerprint, weapons, eye witness, footage

Step 3:

Perform a statistical test:

There are a variety of statistical tests available, but they are all based on the comparison of within-group variance (how spread out the data is within a category) versus between-group variance (how different the categories are from one another).

Step 4:

Decide whether to reject or fail to reject your null hypothesis:

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

Step 5:

Conclusions:

we can make conclusions.

example:how are we going to conduct the experiment or test:

Toss a coin for 100 times

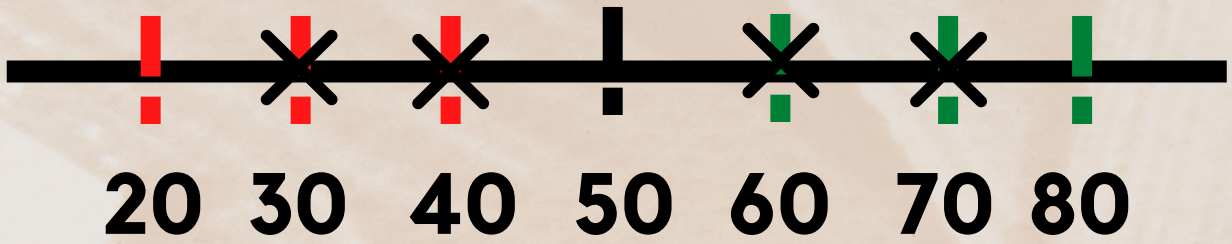
for example we got 50 times head and 50 times tail for the first experiment

for second experiment we got 60 times head and 40 times tail

Then we got 70 times head and 30 times tail

but 80 times head and 20 times tail vary a lot to say its a fair coin

so we are defining a range that it should be in between the range not less than 20 greater than 80.This range is known as **Confidence Interval**

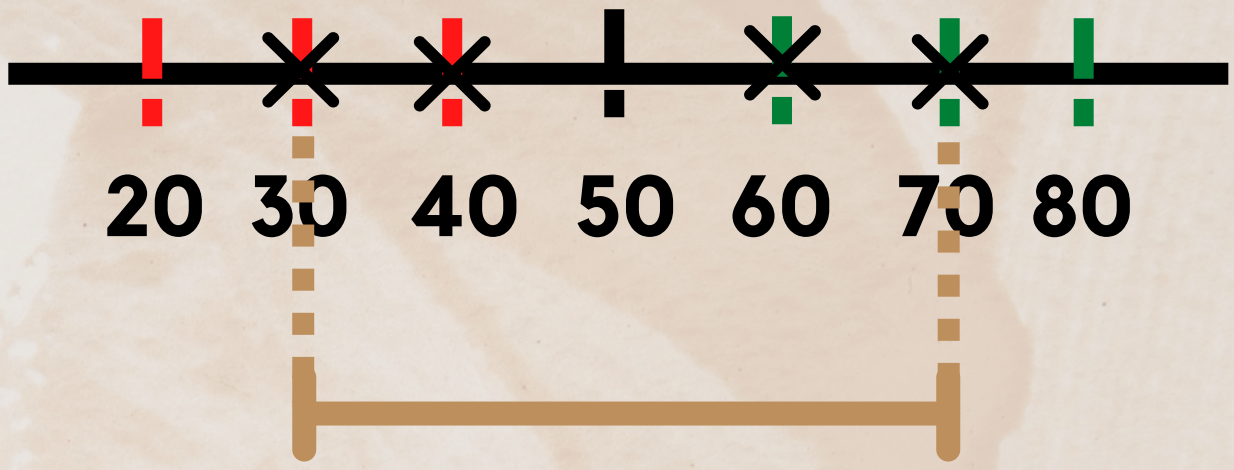


for example we got 50 heads out of hundred experiments we think that it's a fair coin.

for example we got 60 heads out of hundred experiments its ok we can consider that it as fair coin.

what if get we got 70 heads out of hundred experiments then we can ask for domain expert for a range known as confidence interval.

then domain expert say's that consider only if it is between 30 and 70.

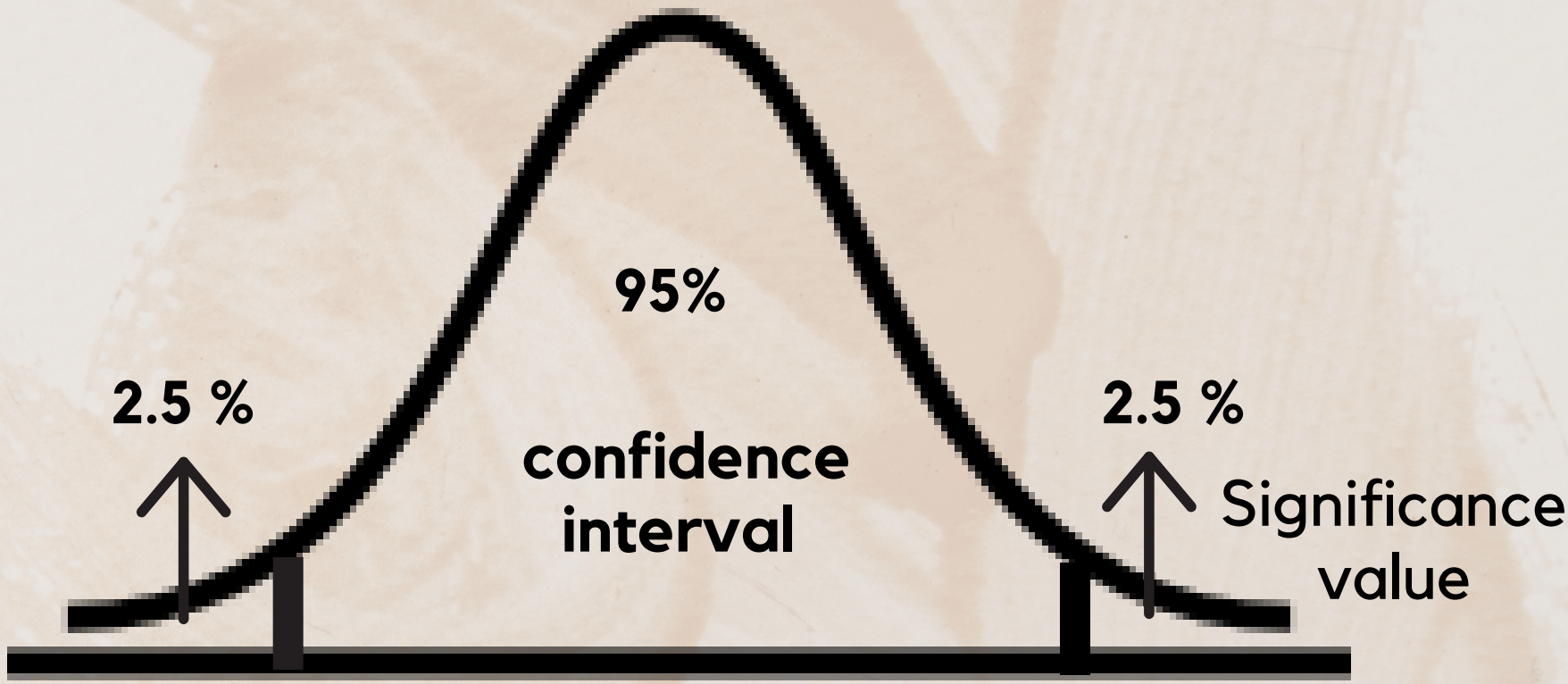


C.I = [30 - 70]

if we get the value in between 30 and 70 then we can consider the coin is fair we can accept null hypothesis.

if we get the value in less than 30 and greater than 70 then we can consider the coin is unfair we reject null hypothesis and accept alternate hypothesis.

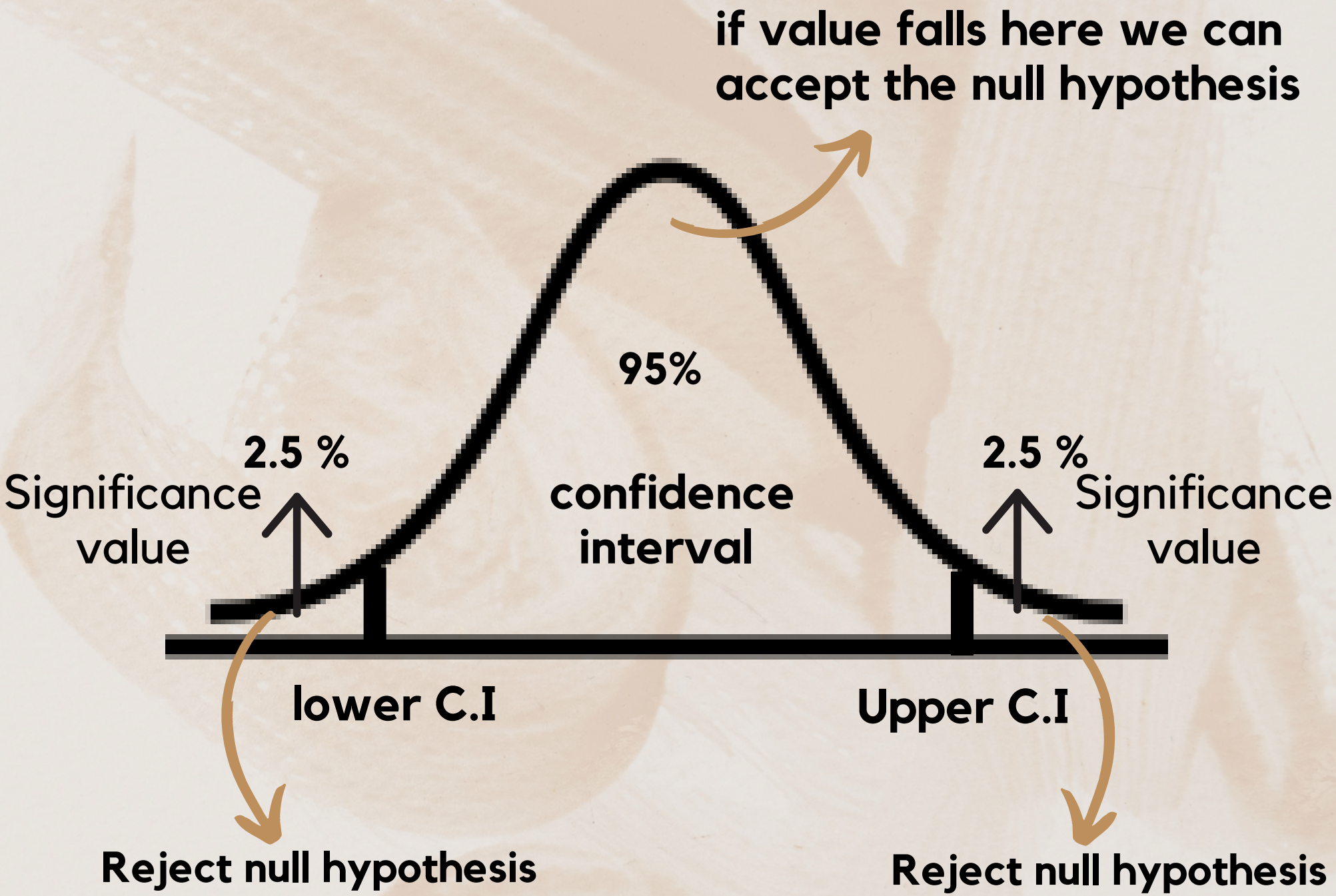
Confidence interval: (defined by domain expert)



if domain expert say's that 95% confidence interval

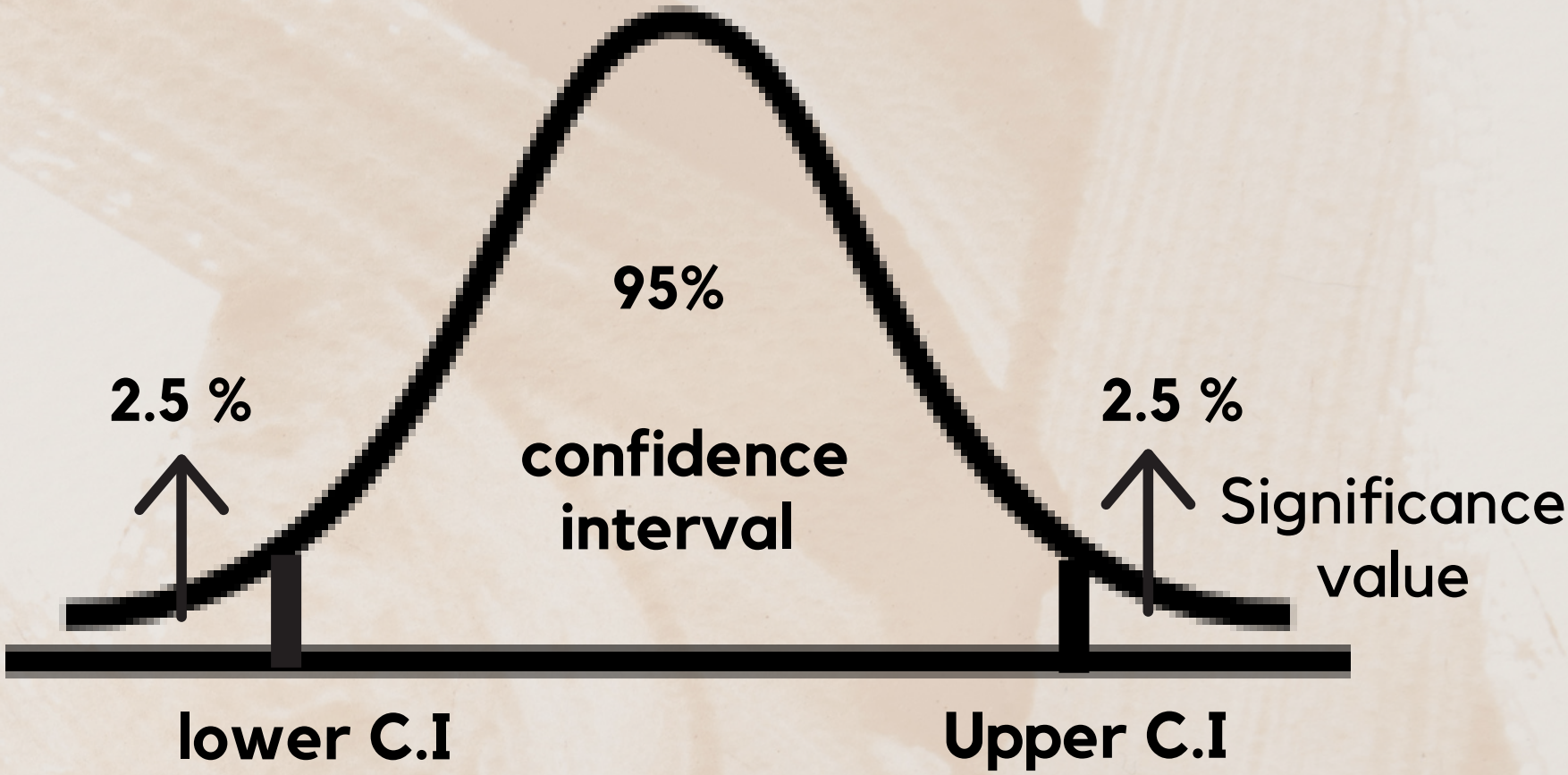
then significance value is **S.V = 1 - 0.95 = 0.05**

In medical field this confidence interval is very less



In medical field this confidence interval is very less

@datascience_2.0



How can we find out this lower C.I and upper C.I?

Point Estimate:

The value of any **statistic** that estimates the value of a **parameter** is called point estimate.

In inferential statistics we make conclusion on **population mean** with the help of **sample mean**

$\bar{x}$ = **sample mean**

μ = **Population mean**

statistic $\bar{x}$ **estimating** μ **parameter**

According to the definition **$\bar{x}$ is static and μ is parameter.**

Here we can consider **$\bar{x}$ as point estimate** because it is trying to estimate **μ (parameter).**

$$\text{statistic}\quad \bar{x} \xrightarrow{\text{estimating}} \mu \quad \text{parameter}$$

we can also observe that sample mean can be higher or lower than the population mean.

$$\bar{x} >= \mu \qquad\qquad \bar{x} <= \mu$$

so we can write

$$\text{Point estimate} \pm \text{Margin of error} = \mu$$

Lower C.I = Point estimate - Margin of error

Upper C.I = Point estimate + Margin of error

@datascience_2.0

Point estimate $\pm$ Margin of error $= \mu$

$$\text{Margin of error} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

α = Significance value

σ = Population standard deviation

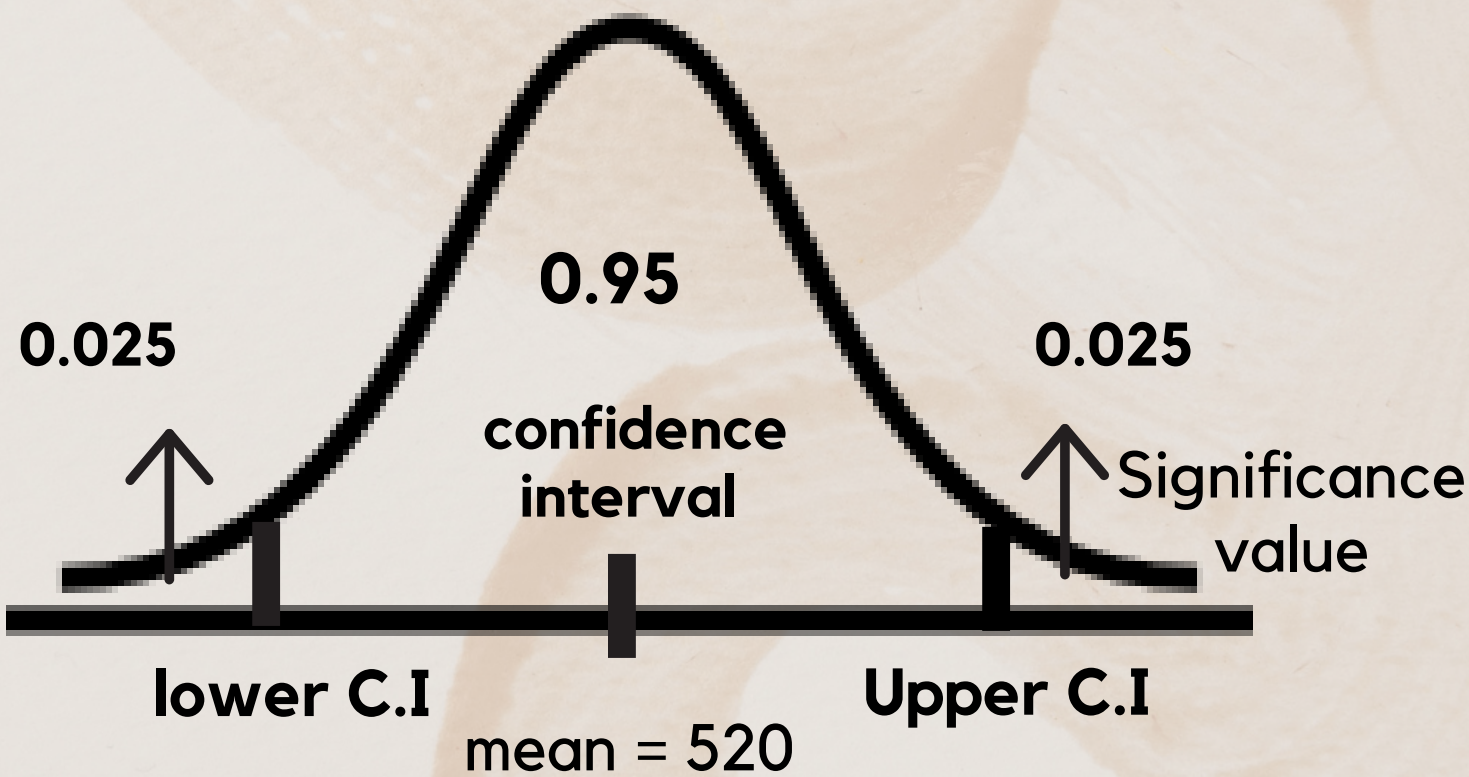
@datascience_2.0

Problem - 1

On the quant test of cat, a sample of 25 test takes has a mean of 520 with a population standard deviation of 100.construct a 95% of C.I about the mean?

n = 25 C.I = 95% S.D = 100

$\bar{x}$ = 520 S.V = 1 - C.I = 1 - 0.95 = 0.05



Here we have to find the lower C.I and Upper C.I

@datascience_2.0

Formula to calculate lower C.I and Upper C.I

Lower C.I = Point estimate - Margin of error

Upper C.I = Point estimate + Margin of error

$$\text{Margin of error} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\text{Lower C.I} = 520 - Z_{\frac{0.05}{2}} \frac{\sigma}{\sqrt{n}}$$

$$520 - Z_{0.025} \frac{100}{5}$$

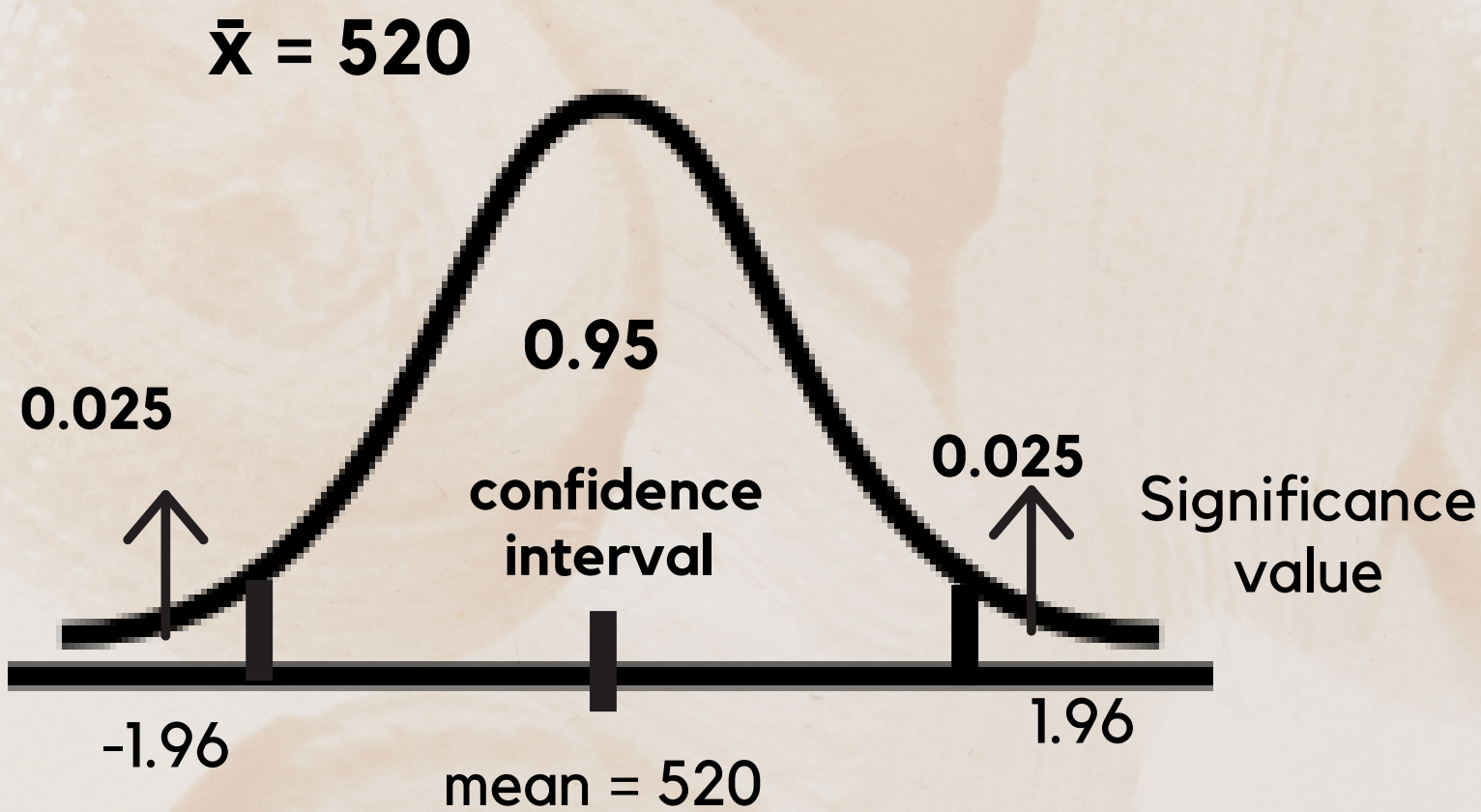
let's find the value of z(0.025) from z table

@datascience_2.0

If we consider the total grapth is 1 then 1-0.025 = 0.975

we can find the value of **0.975** from **z-table**

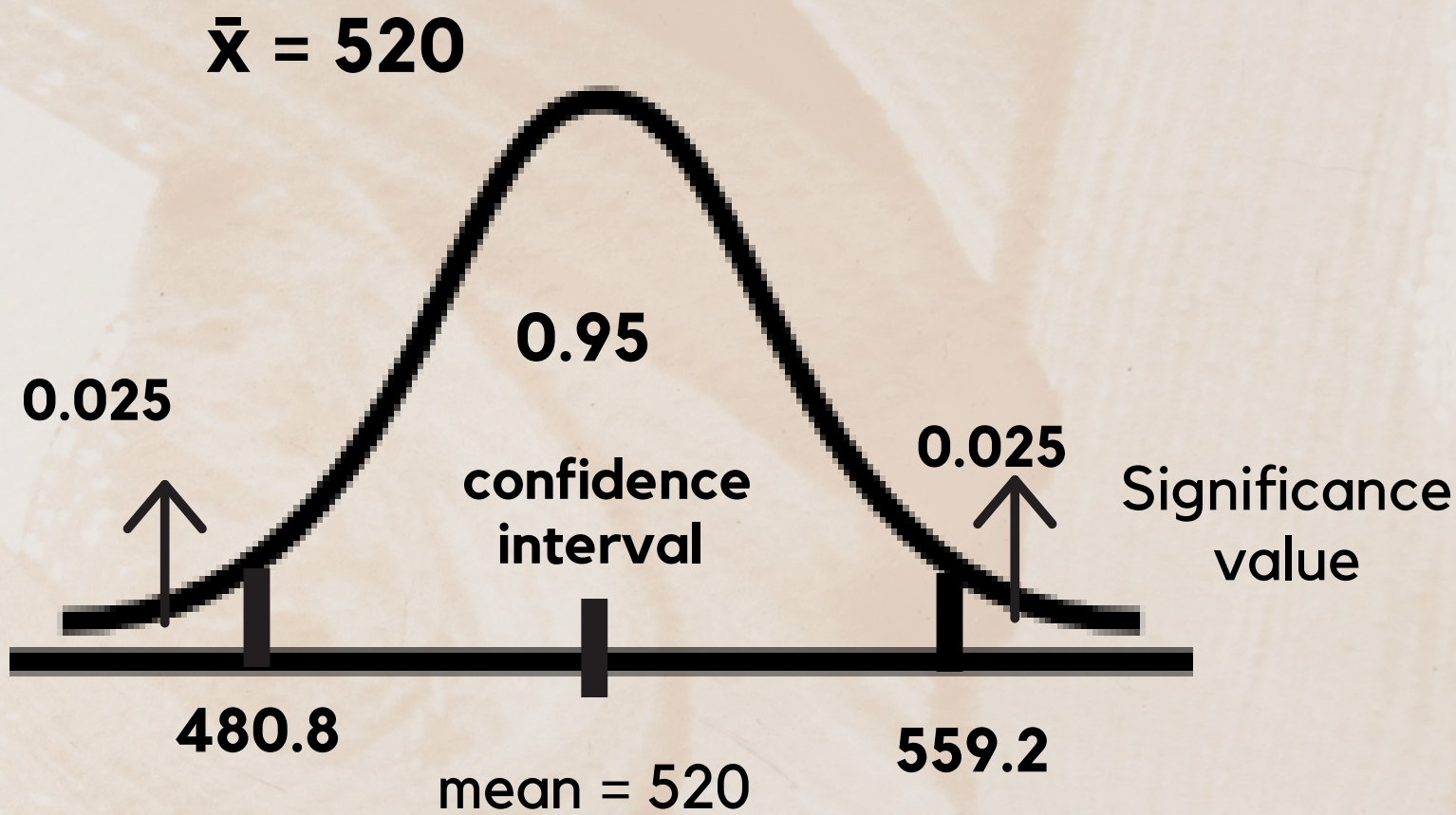
for 0.0025 we get the value as -1.96 and +1.96



Lower C.I = 500 - 1.96 * 20 = 480.8

Upper C.I = 500 + 1.96 * 20 = 559.2

@datascience_2.0



C.I ranges in between 480.8 and 559.2

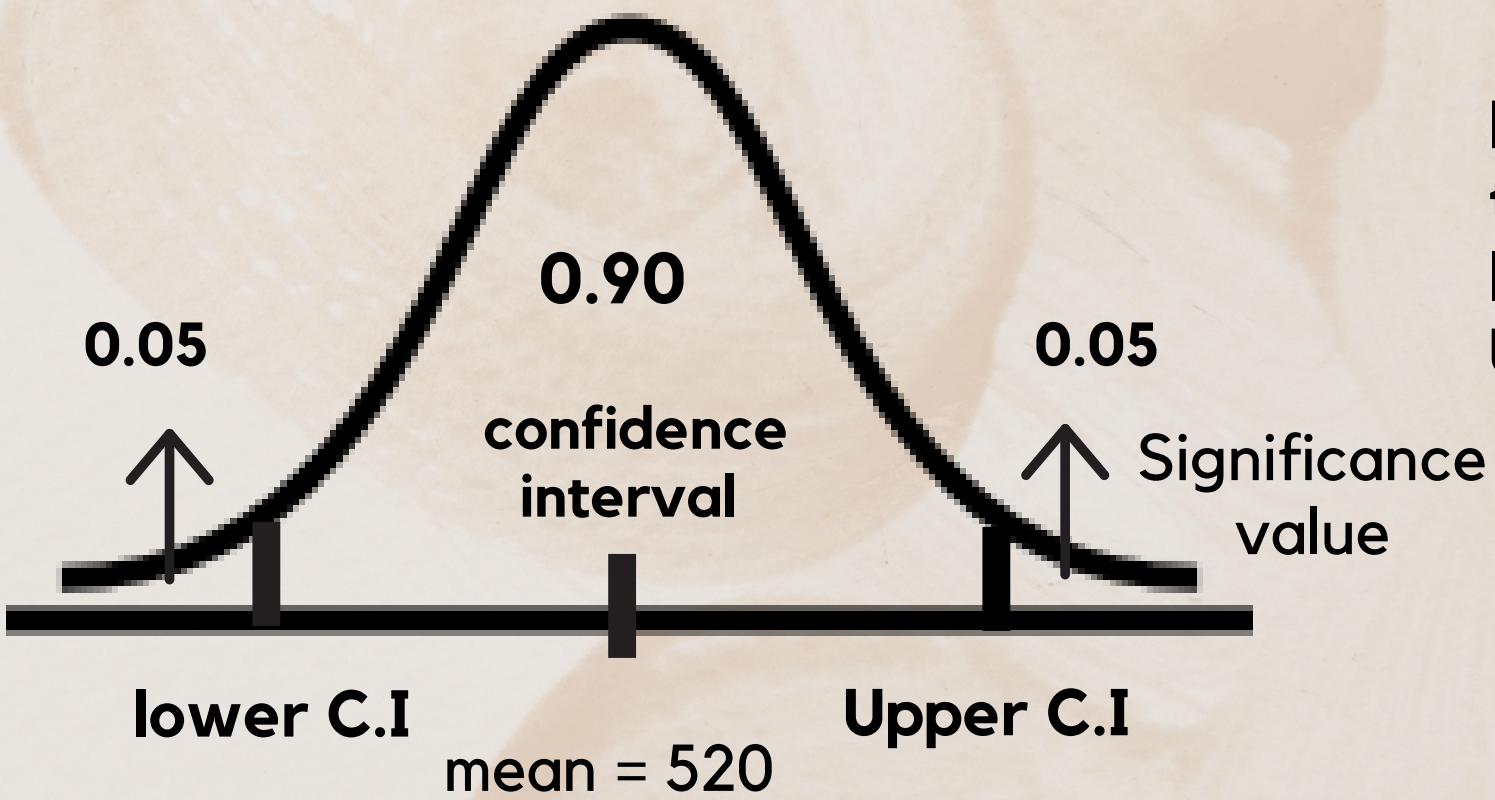
if the value is falling between C.I range then we accept null hypothesis and reject alternate hypothesis if not vise versa

@datascience_2.0

Problem - 2

n = 25 C.I = 90% S.D = 85

$\bar{x}$ = 480 S.V = 1 - C.I = 1 - 0.90 = 0.10



Here we have to find the lower C.I and Upper C.I

Formula to calculate lower C.I and Upper C.I

Lower C.I = Point estimate - Margin of error

Upper C.I = Point estimate + Margin of error

$$\text{Margin of error} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\text{Lower C.I} = 480 - Z_{0.05} \frac{\sigma}{\sqrt{n}}$$

$$= 480 - Z_{0.05} \frac{85}{5}$$

$$= 480 - 1.64 * 17$$

let's find the value of z(0.05) from z table

$$= 480 - 27.8 = 452.12$$

@datascience_2.0

$$\text{Margin of error} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\text{Upper C.I} = 480 + Z_{0.05} \frac{\sigma}{\sqrt{n}}$$

$$= 480 + Z_{0.05} \frac{85}{5}$$

$$= 480 + 1.64 * 17$$

$$= 480 + 27.8 = 507.8$$

C.I range = [452.12 - 507.8]

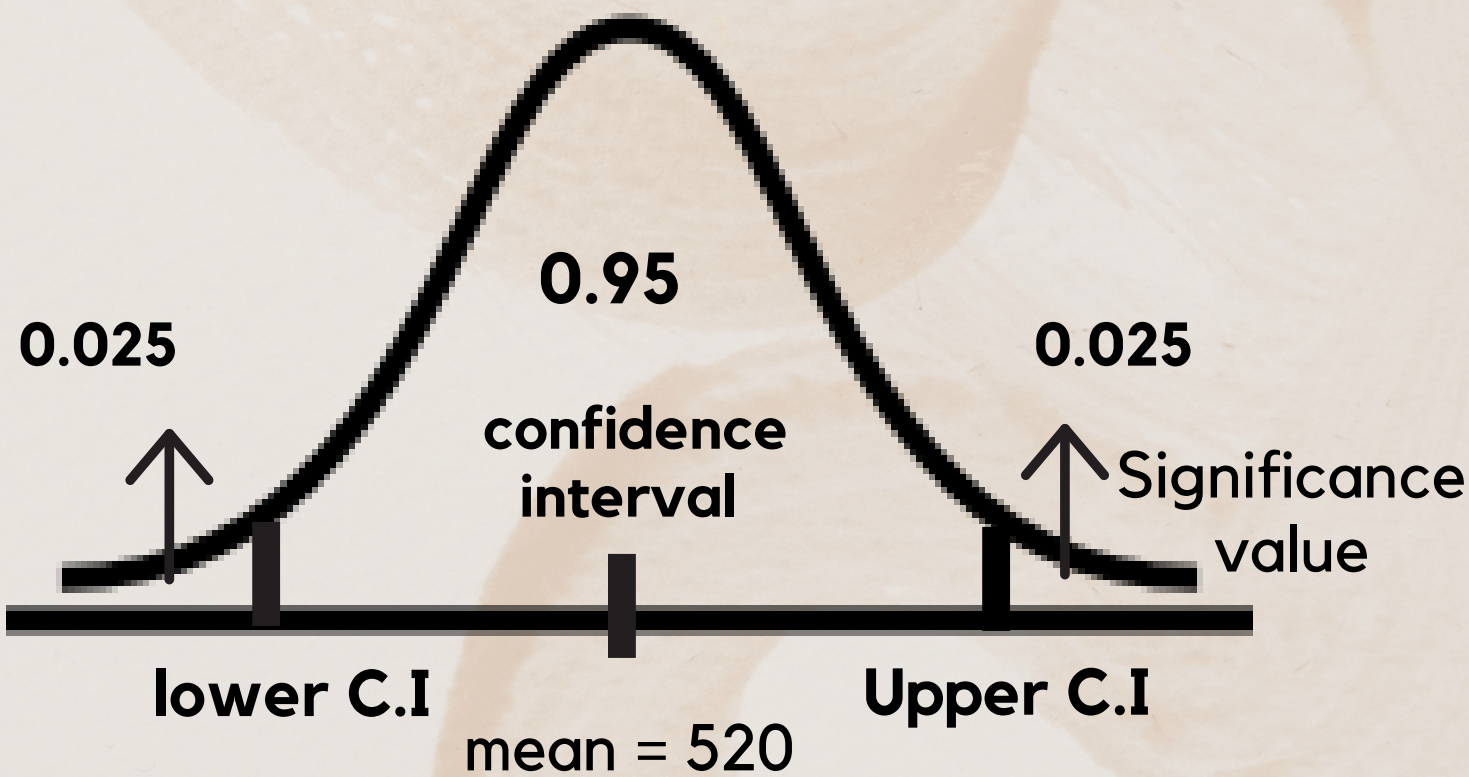
@datascience_2.0

Problem - 3

On the quant test of cat, a sample of 25 test takes has a mean of 520 with a sample standard deviation of 80.construct a 95% of C.I about the mean?

n = 25 C.I = 95% sample sd = 80

$\bar{x}$ = 520 S.V = 1 - C.I = 1 - 0.95 = 0.05



Here we have to find the lower C.I and Upper C.I

@datascience_2.0

Formula to calculate lower C.I and Upper C.I

Lower C.I = Point estimate - Margin of error

Upper C.I = Point estimate + Margin of error

here sample standard deviation is given so we are using t table

$$\text{Margin of error} = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$\text{Lower C.I} = 520 - t_{\frac{0.05}{2}} \frac{s}{\sqrt{n}}$$

$$520 - t_{0.025} \frac{80}{5}$$

let's find the value of t(0.025) from t table

@datascience_2.0

$$520 - t_{0.025} \frac{80}{5}$$

let's find the value of t(0.025) from t table

we have to find the degree of freedom for t test = n - 1 = 25 - 1 = 24

Lower C.I = 520 - 2.064 * 16

Lower C.I = 486.976

Upper C.I = 520 + 2.064 * 16

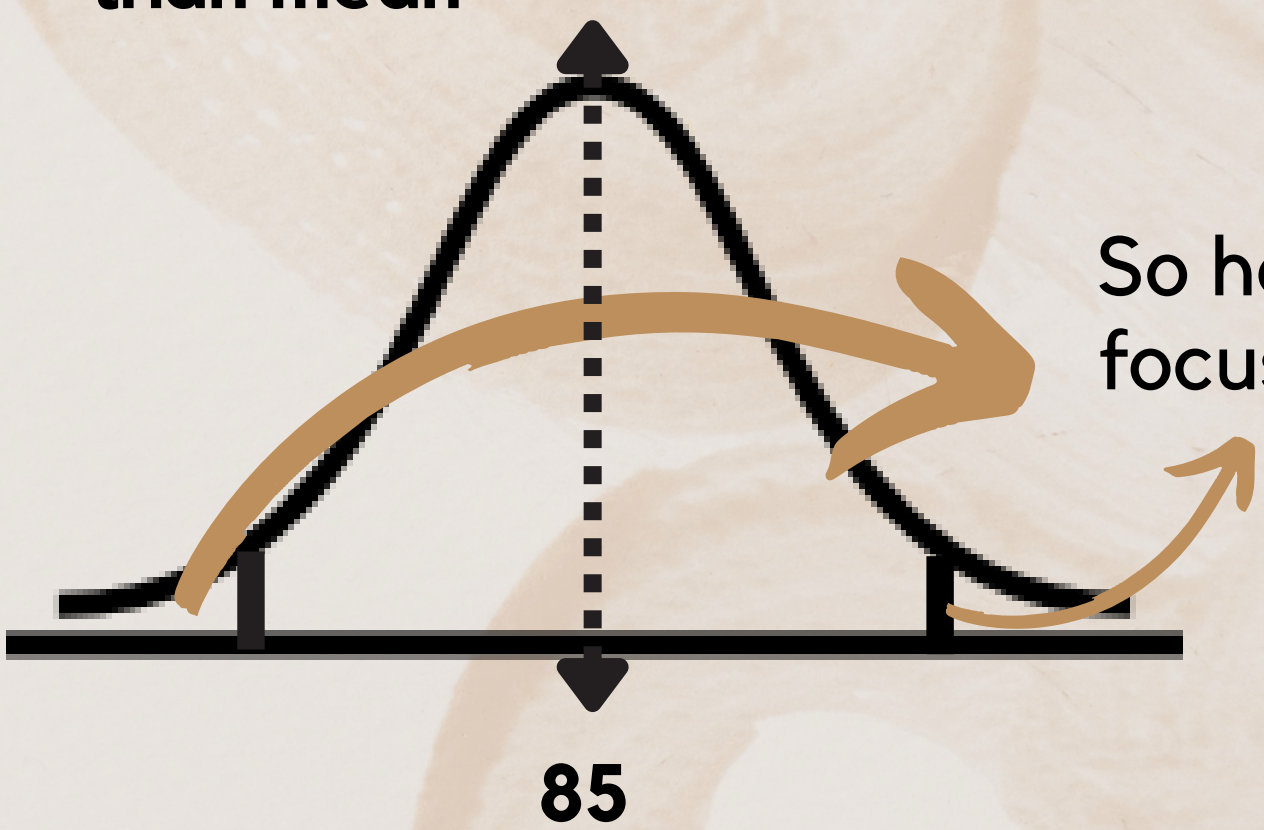
Upper C.I = 553.024

@datascience_2.0

1 - Tail and 2 - Tail Test

Colleges in town A has 85% placement rate.A new college was recently found that a sample of 150 had a placement rate of 88% with a standard deviation of 4% does this college has a **different placement** rate with 95% C.I?

from the question we can observe that **different placement** then it can be **greater than mean** or **less than mean**

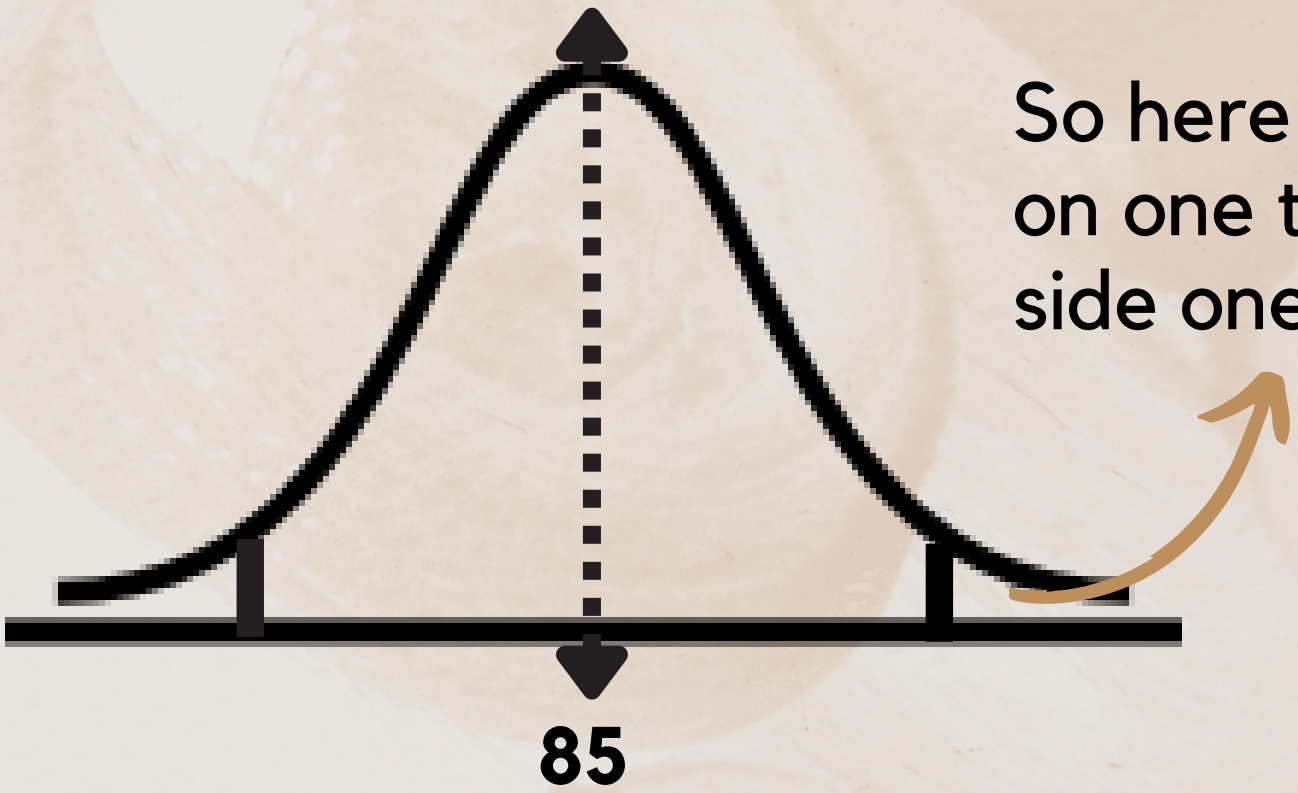


So here we are focusing on both tails

@datascience_2.0

1 - Tail and 2 - Tail Test

if the same question changed to **different placement** rate greater than 95% C.I?



So here we are focusing on one tail which is right side one.

if it is less than we will focus on left side tail.

Hypothesis testing Problem - 1

A factory has a machine that fills 80 ml of baby medicines in a bottle.An employee believes the average amount of baby medicine is not 80 ml.using 40 samples, he measures the average amount dispersed by the machine to be 78 ml with a standard deviation of 2.5.

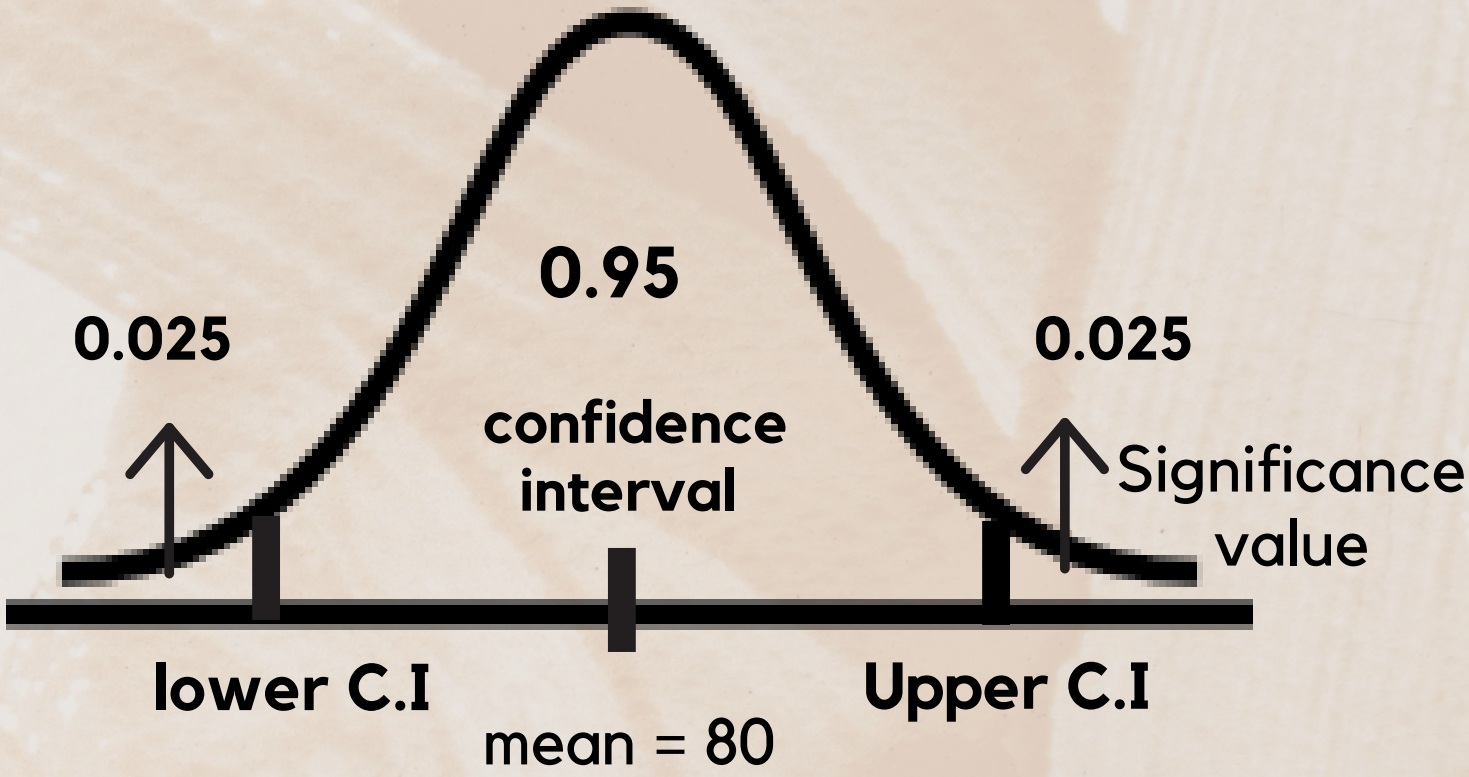
a. state null and alternate hypothesis.

b. At 95% C.I, Is there enough evidence to support machine is working properly or not.

$n = 40$ C.I = 95% $\mu = 80$ ml

$\bar{x} = 78$ S.V = 1 - C.I = 1 - 0.95 = 0.05

@datascience_2.0



step 1:

State Null and Alternate hypothesis

Null hypothesis $\mu = 80$ ml

Alternate hypothesis $\mu \neq 80$ ml

@datascience_2.0

step 2:

C.I = 95% **S.V = 1 - C.I = 1 - 0.95 = 0.05**

step 3:

n = 40 **s = 2.5**

Note:

when **n>=30 or** Population std given the we have to use **Z-Test**

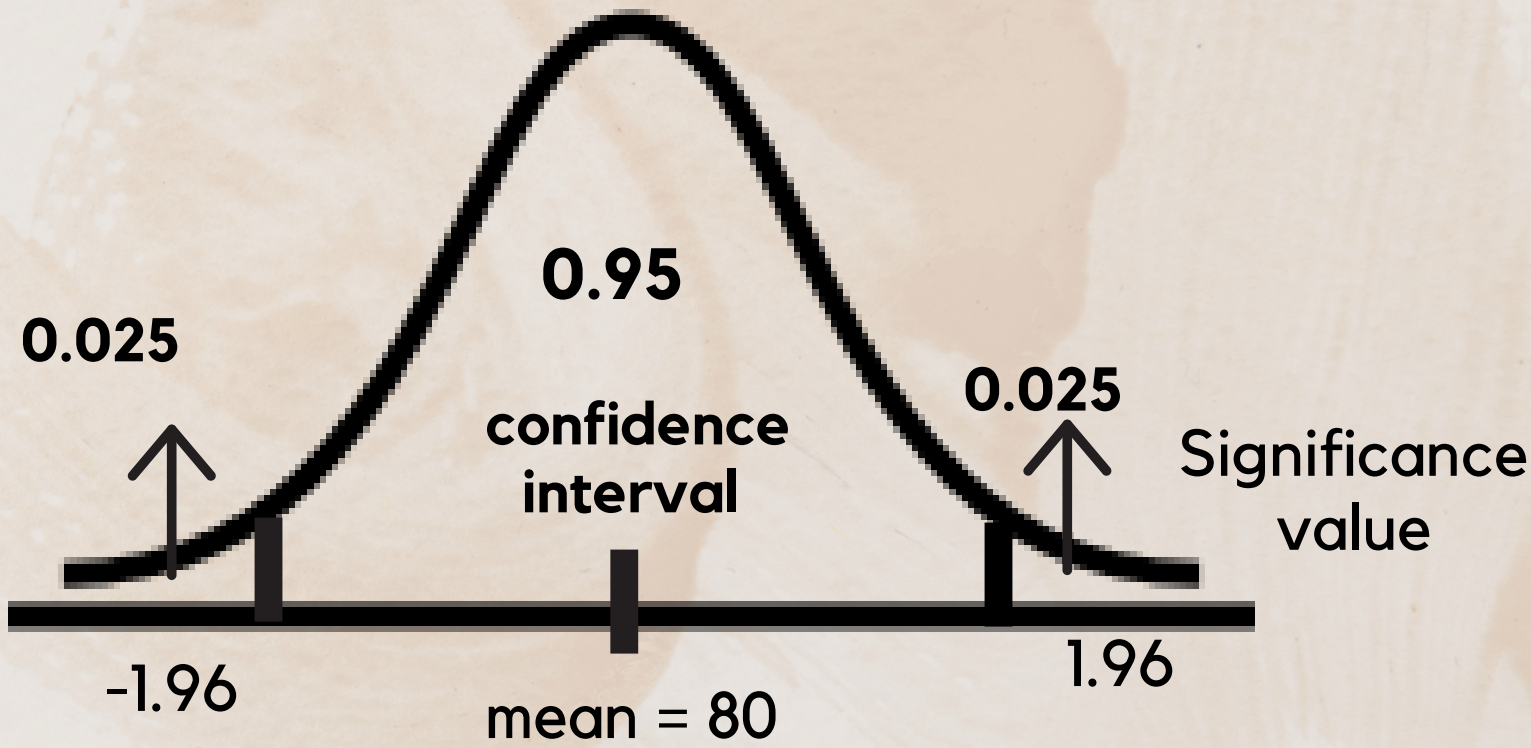
when **n<30** and Sample std given the we have to use **T-Test.**

@datascience_2.0

step 4:

let's perform the experiment.

here we are using z-test



If we consider the total grapth is 1 then 1-0.025 = 0.975

Calculate test statistics:(Z-test)

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

Standard error

$$Z = \frac{78 - 80}{\frac{2.5}{\sqrt{40}}} = -5.05$$

@datascience_2.0

Conclusions:

Decision rule :

Z = -5.05 is less than -1.96 so we are rejecting null hypothesis

so we are concluding that there is some fault in **machine**

Hypothesis testing Problem - 2

A complain was registered that boys in a government school was underfed.Average weight of the boys of age 10 is 32 kg with sd = 9 kgs. A sample of 25 boys were selected from the government school and the average weight was found to be 29.5 kgs. with C.I is 95%. check it is true or false.

$n = 25$ C.I = 95% $\mu = 32$ kgs $\sigma = 9$

$\bar{x} = 29.5$ S.V = 1 - C.I = 1 - 0.95 = 0.05

Conditions for Z-test

- we may know the population sd or we do not know the population sd but our sample is n>=30

Conditions for T - test

- we do not know the population std.
- sample size is $n<30$.
- sample std is given.

step 1:

State Null and Alternate hypothesis

Null hypothesis $\mu = 32$

Alternate hypothesis $\mu \neq 32$

@datascience_2.0

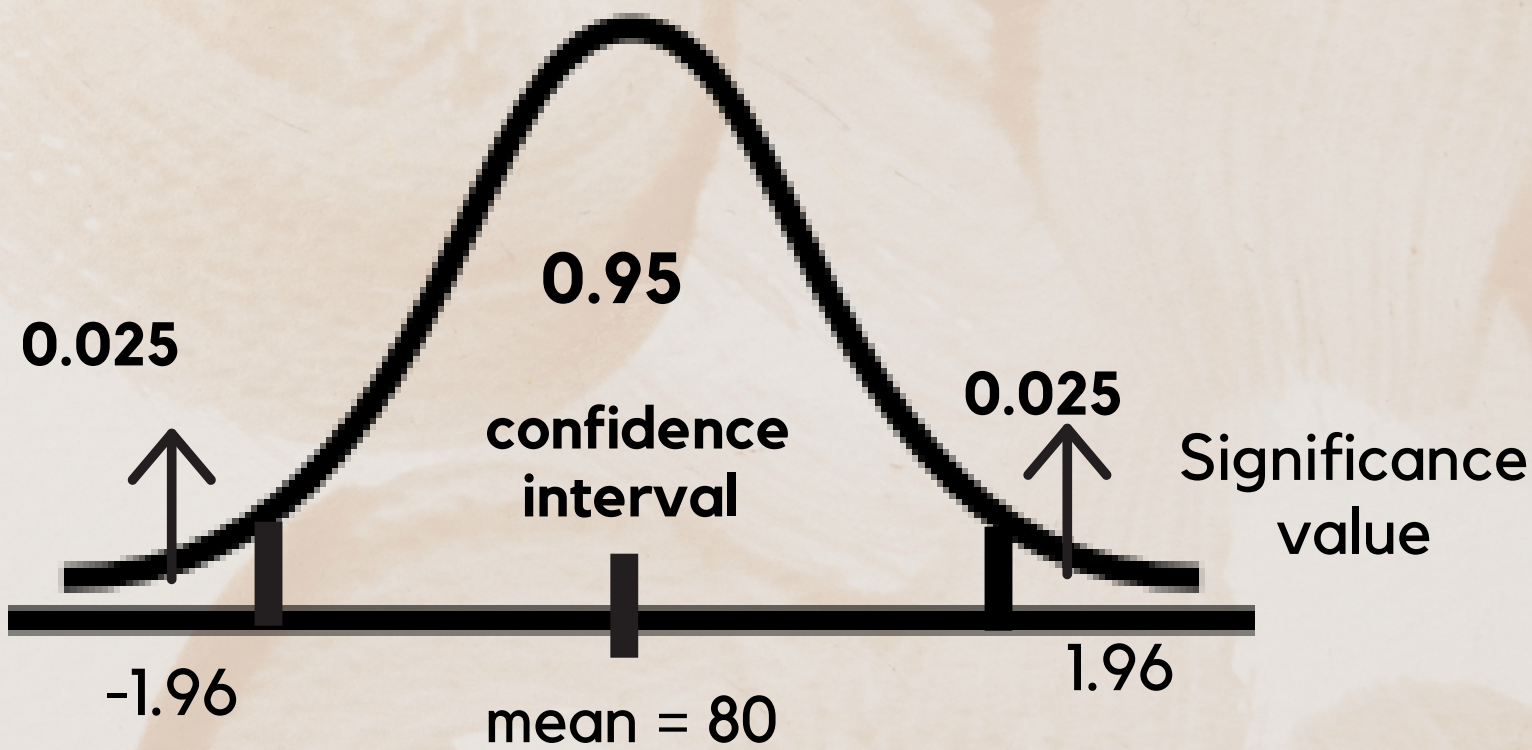
step 2:

C.I = 95% **S.V = 1 - C.I = 1 - 0.95 = 0.05**

step 3:

let's perform the experiment.

here we are using z-test



If we consider the total grapth is 1 then 1-0.025 = 0.975

@datascience_2.0

Calculate test statistics:(Z-test)

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

Standard error

$$Z = \frac{29.5 - 32}{\frac{9}{\sqrt{25}}} = -1.39$$

@datascience_2.0

Conclusions:

Decision rule :

Z = -1.39 is greater than -1.96 so we are accepting null hypothesis with C.I 95%

so we are concluding that boys are fed well.

Thank You

Follow me on instagram

@datascience_2.0

@urstruly_sandeeprayudu