STATISTICS FOR DATA SCIENCE DAY 5

* Inferential Statistics

(i) Hypothesis Testing
(ii) p-value
(iii) Confidence Interval
(iv) Significance Value

→ z-test
→ t-test
→ Chi square test
→ ANOVA test (F-test)

3- Distributions :- (i) Bernoulli
(ii) Binomial
(iii) Power law
} TRANSFORMATION

HYPOTHESIS TESTING

Steps of hypothesis testing

● **Null hypothesis** [the default one will be in null hypothesis]

e.g. (i) Person is not a criminal until and unless he is not proven we will say he is not criminal.

(ii) Let, we have tossed a coin, now we check whether coin is fair or not

So, By default we consider "Coin is fair" → null hypothesis

(●) **Alternate hypothesis** [opposite of Null hypothesis]

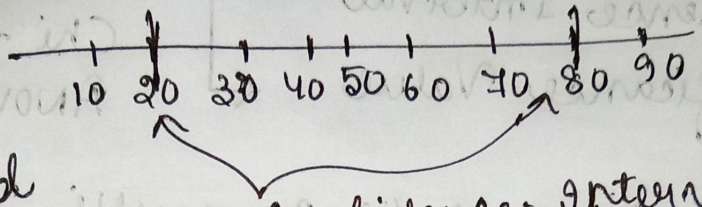
e.g. (i) Coin is not fair
(ii) Person has committed crime

• Perform Experiments

Let, say I have tossed 100 times coin

mean = 50 , Standar deviation = 10

50 times head
⇓
Fair



60 times head
⇓
Fair

70 times head
⇓ we go to
Domain expert to get a range
and that range is called confidence Interval

C.I = [20 - 80]
⇓
coin is fair

If we get 75 heads then null hypothesis is accepted

If we get 10 heads then null hypothesis is rejected and alternate hypothesis is accepted

* We reject the null hypothesis [outside C.I.]

* We fail to reject the null hypothesis [within C.I.]

Eg; whether the person is criminals or not [Murder case]

(i) Null hypothesis : Person is not criminal

(ii) Alternate hypothesis : Person is criminal

(iii) Experiment / Proof : DNA, fingerprint, weapons, eye, witness, footage

⇓

Judge [take decision]

Conclusion :- Based on Judge decision

- Confidence Internal (C.I.)

⇓

95%

$S.I = 1 - C.I.$

C.I. = 95%

Significance value = 1 - 0.95
= 0.05

* L.F. = Lower fence
* H.F = Higher fence

* In S.I. region we reject the null hypothesis
* In C.I. region we fail to reject the null hypothesis.

- Point Estimate :- The value of any statistics that estimates the value of a parameter is called Point Estimate.

e.g. mean of sample is used to estimate the mean of parameter

$\bar{x}$ ⟶ estimate ⟶ μ ⟶ parameter

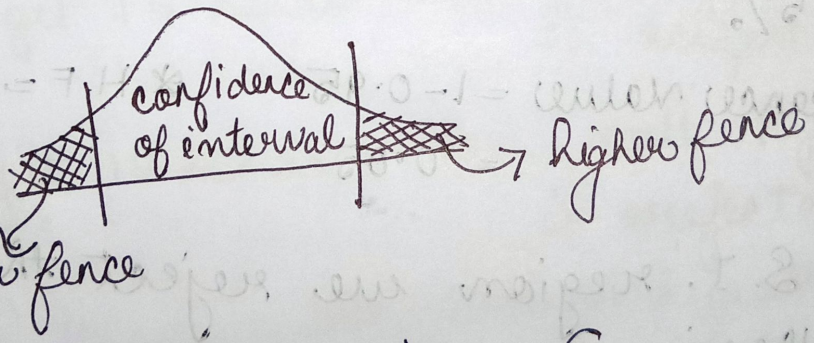
↖ statistic

Parameter ⇒ Population Mean

Point estimate ($\bar{x}$)	±	Margin of error	=	Parameter (μ)

Lower Fence = point estimate − Margin of error

Higher Fence = point estimate + Margin of error



* Margin of error = $Z_{\alpha/2} \frac{6}{\sqrt{n}}$

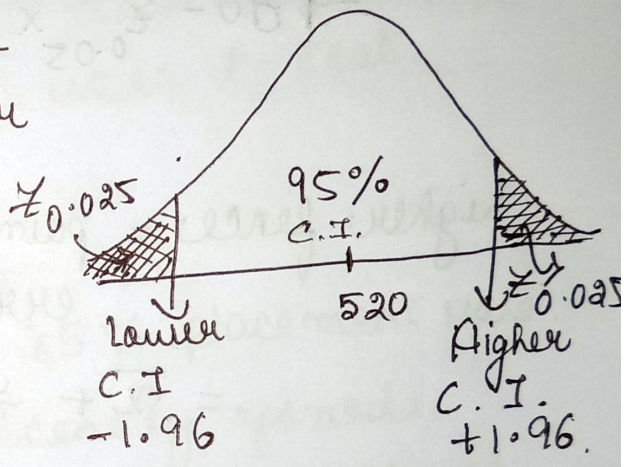
where, α = significance value

$\frac{6}{\sqrt{n}}$ = standard error

Problem: On the quant test of CAT Exam, a sample of 25 test takers has a mean of 520 with a ~~sample~~ population standard deviation of 100. Construct a 95% C.I. about the mean?

Sol:n Given, $n = 25$, $\bar{x} = 520$, $\sigma = 100$

$C.I = 95\%$, $S.V. = 1 - C.I = 1 - 0.95 = 0.05$



Lower C.I = Point estimate − Margin of error

$$= 520 - Z_{0.05/2} \frac{100}{\sqrt{25}}$$

$$= 520 - Z_{0.025}\, 20$$

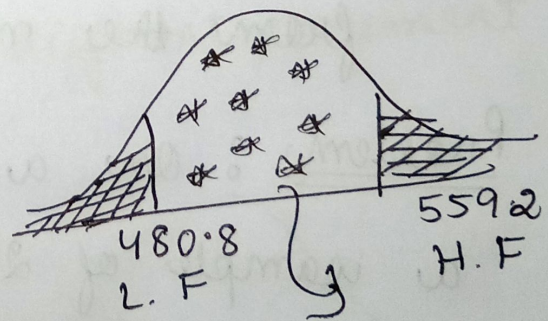
$1 - 0.025 = 0.975$

$= 520 - 1.96 \times 20$

$= 520 - 39.2 = 480.8$

Higher C.I = $520 + 39.2$

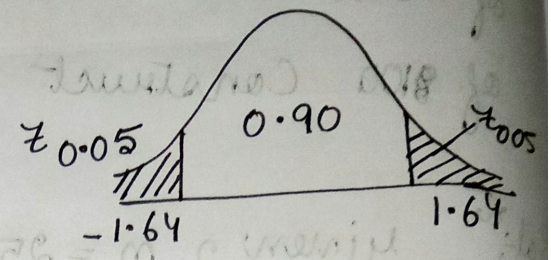
$= 559.2$



Value come under this fail to reject null hypothesis.

Problem $\bar{x} = 480$, $6 = 85$, $n = 25$, C.I = 90%

S.V. = 1 − C.I = 1 − 0.90 = 0.10



Lower fence = $\bar{x}$ − Margin of error

$= 480 - z_{0.05} \frac{85}{\sqrt{25}}$

$= 480 - z_{0.05} \times 17 = 480 - 1.64 \times 17$

$= 452.12$

∴ higher fence = point estimate + Margin of error

$= \bar{x} + z_{\alpha/2} \frac{6}{\sqrt{n}} = 480 + 1.64 \times 17$

$= 507.8$

* here ±1.64 is the standard deviation from the mean

Problem: On a quant test of CAT exam a sample of 25 test takers has a mean of 520, with a sample standard deviation of 80. Construct 95% C.I about the mean?

Sol^n Given, $n = 25$, $\bar{x} = 520$, $s = 80$

S.V. = 1 − 0.95 = 0.05 C.I = 95%

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

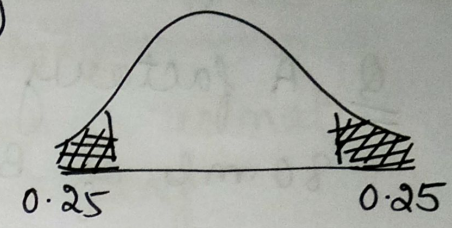
t-test

Degree of freedom
$= n-1 = 25-1 = 24$

Lower C.I. $= 520 - t_{0.05/2}\left(\frac{80}{5}\right)$

$= 520 - 2.064 \times 16$

$= 486.976$



Higher C.I $= 520 + 2.064 \times 16 = 553.024$

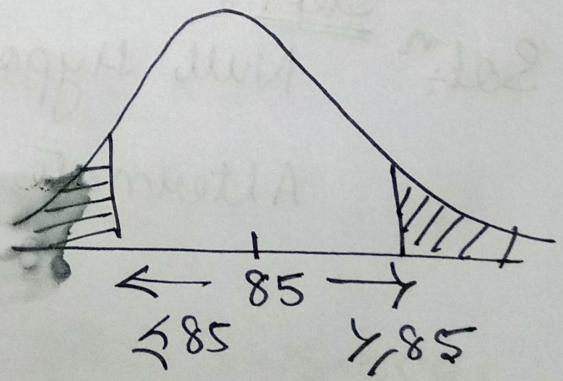
* Since sample standard deviation is given therefore we use t-test.

1 Tail and 2 Tail Test

(*) Colleges in Town A has 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88% with a standandard deviation of 4%. Does this college has a different placement rate with 95% C.I. ?

* greater than 85%
  ⇓
  1 tail test Right side

* lesser than 85%
  ⇓
  1 tail test Left side



① z - test

② t - test

Hypothesis Testing Problem

Q 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 80ml. Using [Two tailed Test] 40 samples, he measures the average amount dispersed by the machine is to be 78 ml with a standard deviation of 2.5.

State (a) Null & Hypothesis

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

given : $\mu = 80ml$, $n = 40$, $\bar{x} = 78$, $s = 2.5$

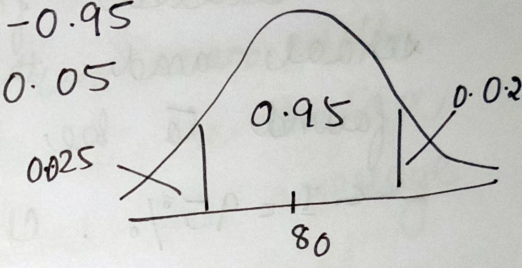
C.I. = 95% = 0.95

~~$\therefore \alpha = 1 - 0.95 = 0.05$~~

Step1: Null Hypothesis : $\mu = 80$ i.e. machine is working

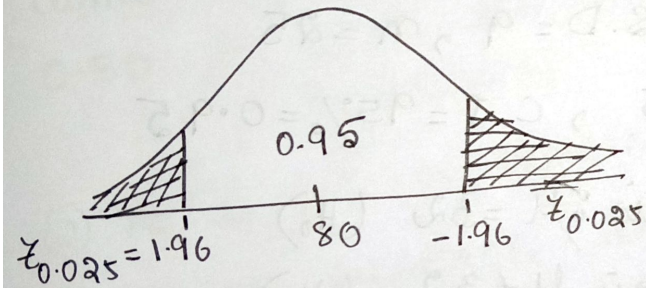
Alt. hypothesis : $\mu \neq 80$ i.e. machine is not working

Step2 : S.V (α) = 1 - C.I = 1 - 0.95 = 0.05



Step3: z-test

$z_{0.025} = 1.96$



* when $n \geq 30$ or pop s.d is given ↳ used z-test

* when $n < 30$ and sample s.d is given used t-test

Decision Boundary

1.96 & -1.96 is the boundary.

Calculate test statistics (z-test)

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{78 - 80}{2.5/\sqrt{40}} = \frac{-2}{2.5/\sqrt{40}} = -5.0596$$

Conclusion : **Decision Rule** → If $z = -5.05$ is less than -1.96 or greater that +1.96 we reject null hypoth

Reject the Null hypothesis with 95% C.I
i.e. there is some faults in the machine

Q A complain was registered, the boys in a government school are underfed. Average weight of the boys of age 10 is 32kgs. with S.D = 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 = 95%. check it is True or False

Sol: n Z test $\Rightarrow n > 30$ or $n \leq 30$ or pop^n s.d.

Given, $\mu = 32$, S.D = 9, $n = 25$

$\bar{x} = 29.5$, C.I = 95% = 0.95

Step1: Null hypothesis: $\mu = 32$ (H_0)

Alt. hypothesis $\mu \neq 32$ (H_1)

Step2: S.V(x) = 1 − C.I = 1 − 0.95

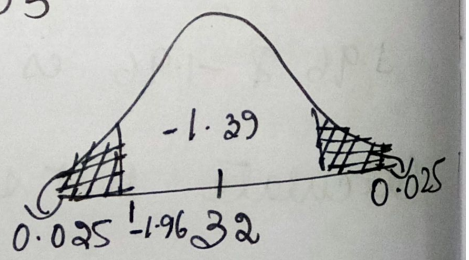
= 0.05

Step3: Z-test

$$Z_{0.025} = \frac{\bar{x} - \mu}{S.D/\sqrt{n}}$$

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

$Z_{(\alpha/2)}$ by Z-test table -1.96



Conclusion :- $-1.394 < -1.96$

$\therefore$ Accept the null hypothesis with 95% of C.I and we fail to reject null hypothesis. i.e. the boys are fed well.

Q. A factory manufactures cars with a warranty of 5 years or more on the engine and transmisson. An engineer believes that the engine or transmisson will malfunction is less than 5 years. He tests a sample of 40 cars and finds the average time to be 4.8 years with a S.D of 0.50 ① State the null and alternate hypothesis

② At a 2% significance level, is there enough evidence to support the idea that the warning should be received

Sol:- given, $\mu = 5$, $n = 40$, $\bar{x} = 4.8$, $S = 0.50$.

① Step 1: H_0 (Null hypothesis) $\geq \mu \leq 5$

H_1 (Alt. hypothesis) : $\mu < 5$

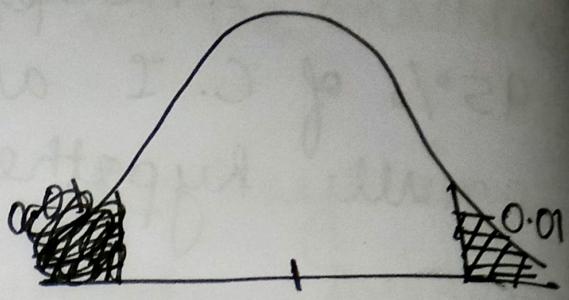
② Step 2: S.L. $= 0.02$, C.I $= 1 - 0.02 = 0.98$

Step 3 : Compute test-statistic

$$z_{0.01} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{4.8 - 5}{0.50/\sqrt{40}} = \frac{-0.2}{0.0790}$$

$$= -2.529 = -2.53$$



[Right Tailed Test]

$$z_{\alpha} = -2.33$$

$\because$ $-2.53 < -2.33$

$\therefore$ we accept the null hypothesis with 98% of C.I. and reject the alternate hypothesis i.e. yes the warning is received.