STATISTICS

- DESCRIPTIVE
- INFERENTIAL

1 (#) DESCRIPTIVE

Central Tendency or Average.

- Mean (Arithmatic Mean) — Normal (Average)
- median
- mode

$\Rightarrow$

Most common Number

1 1 2 3 4 → (1)

1 1 2 3 4 4 → ?

median

1 1 2 3 4

(after sorting)

1 1 2 3 4 4

Arithmetic mean of these number $\Rightarrow$ **2.5**

Mean

$\frac{1+1+2+3+4}{5}$

ex:- 3 3 3 3 3 100

→ mean: $\frac{115}{6} = 19\frac{1}{6}$ ← central Tendency

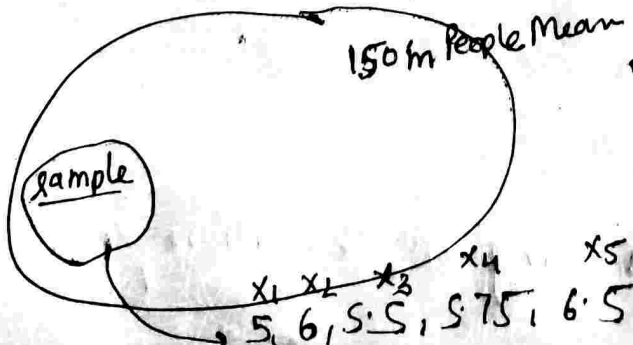
→ median: $\frac{3+3}{2} = 3$

→ Mode : 3.

(#) ⇒ **Sample Population :-**

μ = Population mean

$\bar{x}$ = sample mean.



$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n$$ ✓

$$\bar{x} = \sum_{n=1}^{n} x_m \Big/ n$$ ✓

(#) Measures of Dispersion:-

ex: 2 2 3 3

$\mu = \frac{2+2+3+3}{4} = 2.5$

close to 2.5

ex: 0 0 5 5

$\mu = \frac{0+0+5+5}{4} = 2.5$

away from 2.5

more dispersion

How to measure this ?

⇒ Variance : (σ^2)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

ex: 2 2 3 3 ⟶

i	x_i	μ	$x_i-\mu$	$(x_i-\mu)^2$
1	2	2.5	-0.5	0.25
2	2	2.5	-0.5	0.25
3	3	2.5	.5	0.25
4	3	2.5	.5	0.25

$\sum_{i=1}^{4}(x_i-\mu)^2 = 1$

$\sigma^2 = 1/4 = 0.25$ (Average squared distance from mean)

distance from mean

ex: 0, 0, 5, 5

↳ $\sigma^2 = 6.25$

Note: Everything we are doing in Population.

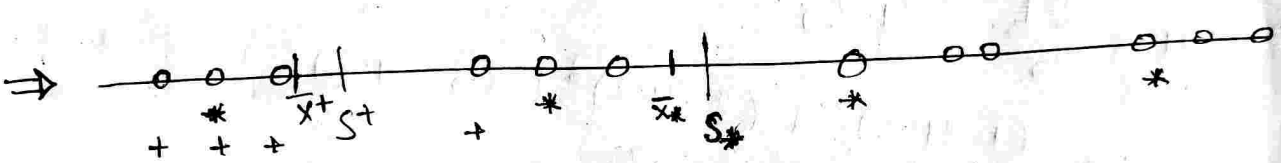
⇒ σ^2 = variance = $\dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$; $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$

⇒ Very hard to calculate for population./

⇒ Calculate parameters in Sample & make Inferences.

⇒ $\boxed{\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}}$ ← Sample mean

$\boxed{S_n^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$ ← Sample variance

⇒ 

⇒ * → sample 1- items;

$\bar{x}_*$ → sample1 mean ; S_* → sample1 variance

⇒ + → sample 2 : items ~ (Skewed Sample)

$\bar{x}_+$ → sample 2 mean ; S_+ → sample 2 variance

⇒ S_+ → underestimating the population variance.

⇒ Unbiased Sample Variance ~

$$\boxed{S^2 = S_{n-1}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

⊛ Standard deviation $\sigma = \sqrt{\sigma^2}$; $S = \sqrt{S^2}$

⊛ Why standard deviation is required:-

Suppose x is in meter; $x \rightarrow$ i/p data.

variance will be 2.5 (suppose) meter^2

So, here variance is in square meter.

Now, to get the value in appropriate unit we use standard deviation.

---

examples 1 2 3 8 7

$$\mu = \frac{21}{5} = 4.2;$$

$$\sigma^2 = \frac{(1-4.2)^2 + (2-4.2)^2 + (3-4.2)^2 + (8-4.2)^2 + (7-4.2)^2}{5}$$

$$\sigma^2 = \frac{38.8}{5} = 7.76 \qquad \sigma = 2.79$$

$$S^2 = 9.7 \quad ; \quad S = 3.11$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \rightarrow \frac{\sum_{i=1}^{N} (x_i^2 - 2x_i\mu + \mu^2)}{N}$$

$$\rightarrow \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 \sum_{i=1}^{N} 1}{N}$$

$$\rightarrow \frac{\sum_{i=1}^{N} x_i^2 - 2\mu \sum_{i=1}^{N} x_i + \mu^2 N}{N}$$

(#) $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - 2\mu\sum_{i=1}^{N} x_i + \mu^2 N}{N}$$ ($\sum_{i=1}^{N} x_i \to N\mu$)

$$\Rightarrow \frac{\sum_{i=1}^{N} x_i^2}{N} - 2\mu^2 + \mu^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$$

faster way of calculating variance

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$$ ↳ raw score method

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \frac{\left[\sum_{i=1}^{N} x_i\right]^2}{N^2}$$ ← We don't have to calculate mean ahead of time.

Normal Distributed : (Gaussian Distribution)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

↳ standard Z-score

$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

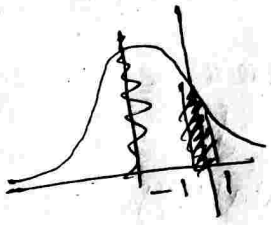
how much standard deviation we are away from mean.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2 e^{z^2}}}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \left(e^{\left(\frac{x-\mu}{\sigma}\right)^2}\right)^{-1/2}$$

(#) To know the probability from probability Density graph, We need to take small area (around that point) under the curve.

So: $$\int_{-1}^{1} \frac{1}{10\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x+5}{100}\right)^2} \cdot dx.$$

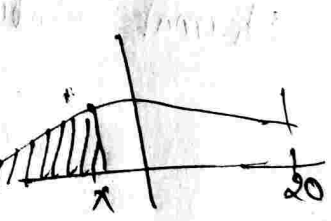


But; not an easy integral to evaluate. So can be done numerically.

A function is defined called Cumulative distribution function, that is useful tool to figure out this area

~~CMD~~ $$\boxed{CDF(X) = \int_{-\infty}^{X} P(x)\,dx}$$

↳ it tells you area under the curve upto X.



Helps you when you wants to know, ~~that~~ the probability to get the value less than 20,.

↓ to calculate are b/w -1 to 1.

↳ $\int_{-\infty}^{1} \ldots - \int_{-\infty}^{-1} \ldots$ ⇒ desired Area.

~~CDF(1) - CDF(-1)~~ // CDF(1) − CDF(−1)

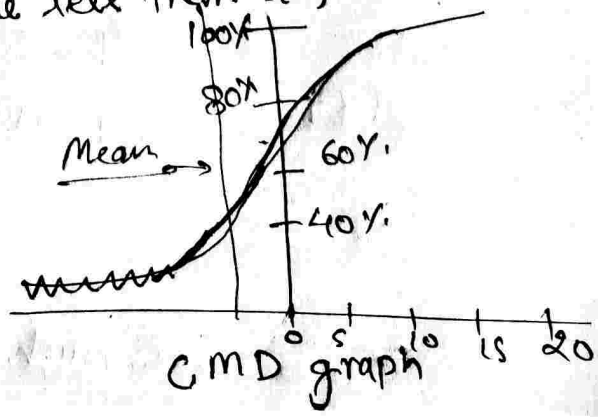


Table:

X	(X−μ)	(X−μ)/σ	CDF(X)	P(X)

↑ distance from

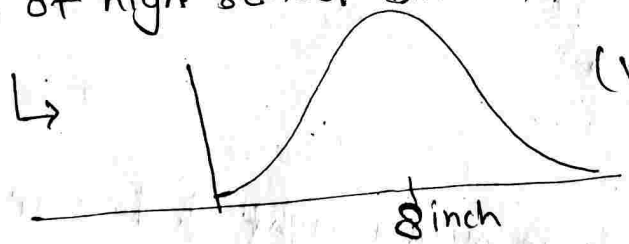
↑ Z-score How many standard deviation away from mean

↑ Cumulative Distribution function

---

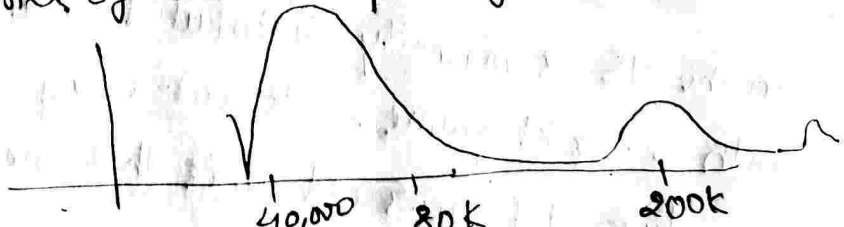
examples:- Which of the following data sets is most likely to be normally distributed? ~~For~~

(a) The hand span (measured from the trip of the ~~bomb~~ thumb to the tip of the 5th finger) of a random sample of high school seniors.



(✓).

(b) The annual salaries of all employees of a large shipping company.



(X) ✓

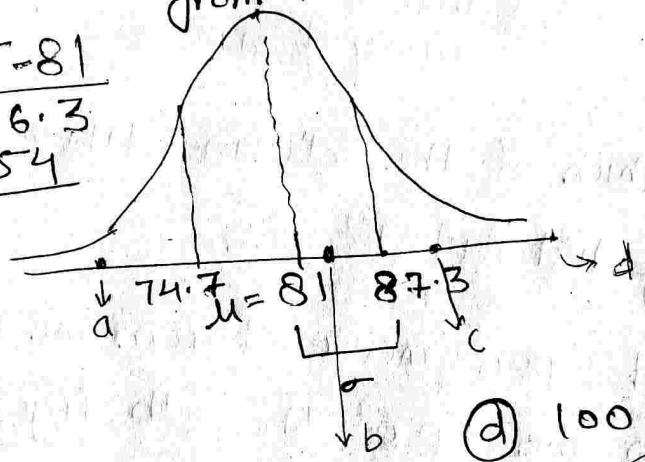
right-skewed distribution (+ve skewed distribution)

(mean is toward right, Hence it is called right-skewed

example 2:

The grades on a statistics mid-term for a high-school are normally distributed with $\mu = 81$, $\sigma = 6.3$. Calculate Z-score ? for each of the following exam grades.

Sol: Z-score :- How many standard deviation away from the mean



(a) 65 $\Rightarrow \frac{65-81}{6.3} = -2.54$

(b) 83 $\Rightarrow \frac{83-81}{6.3} = 0.32$

(c) 93 $\Rightarrow \frac{93-81}{6.3} \quad \frac{x-\mu}{\sigma}$

$\Rightarrow 1.9$

(d) 100 ← distance from mean

$\Rightarrow \frac{100-81}{6.3}$ ← magnitude (scale)

$\Rightarrow 3.02$

---

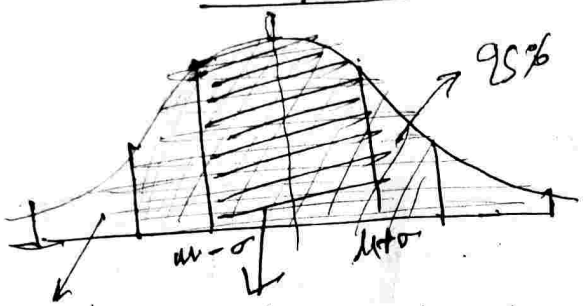
example: (3) Assume that the mean weight of 1-year-old girls in the US is normally distributed with a mean of about 9.5 kg with $\sigma \approx 1.1$ grams. Without using calculator, estimate the percentage of 1-year old girls that meet the following conditions. Draw a sketch & shade the proper region for each problem.

(a) less than 8.4 kg (b) b/w 7.3 kg & 11.7 kg

(c) more than 12.8 kg.

⑤

⇒ Sole Empiricale Rule:



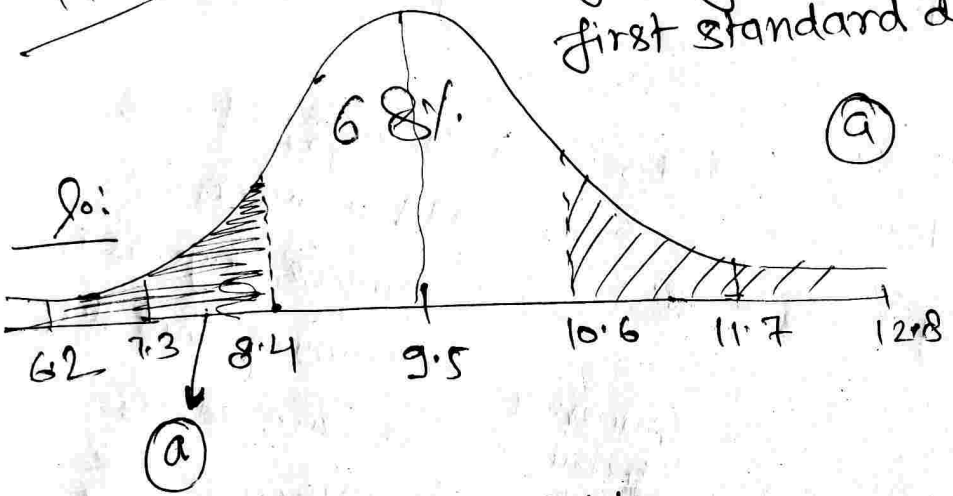
99.7% 68%

68-95-99.7

68% of chance finding result b/w first standard deviation

95% chance b/w second deviation

99.7% b/w third deviation



Ans:

ⓐ

ⓐ total area under curve

100 - 68% = 32%

left & right teg.

ⓑ: b/w 2 standard deviation & 95%

⇒ 32/2 = 16% ✓

ⓒ b/w 3 standard deviation = 99.7%

100 - 99.7% = 0.3%/2 = 0.15% ✓

Note: for standard Normal distribution &

Mean = 0

standard deviation = 1

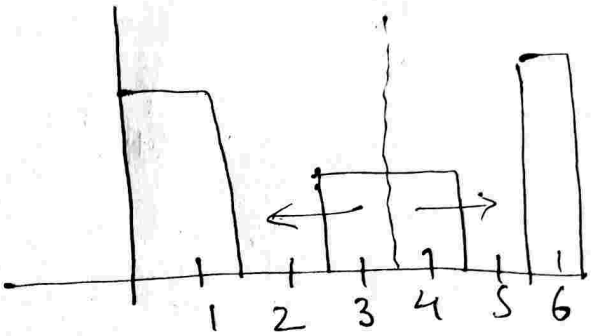
---

* empiricale Rule

68 - 95 - 99.7

Note! z-score can apply to non-normal distribution.

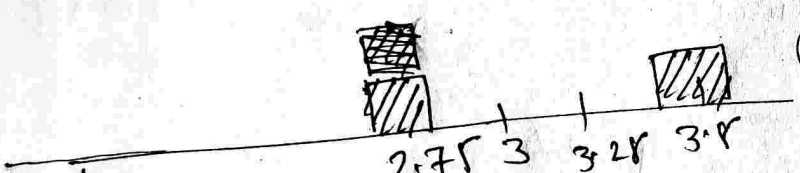
(#) Central Limit Theorem



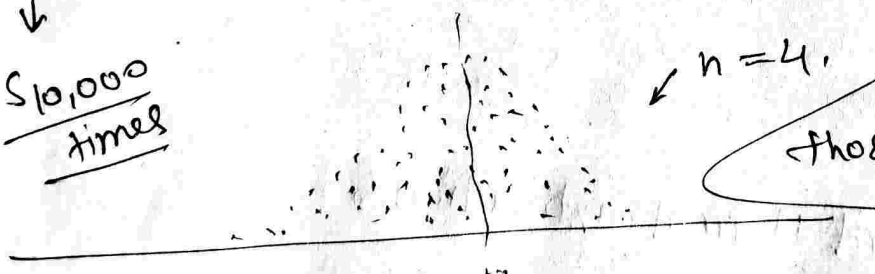
Sample size of $n = 4$

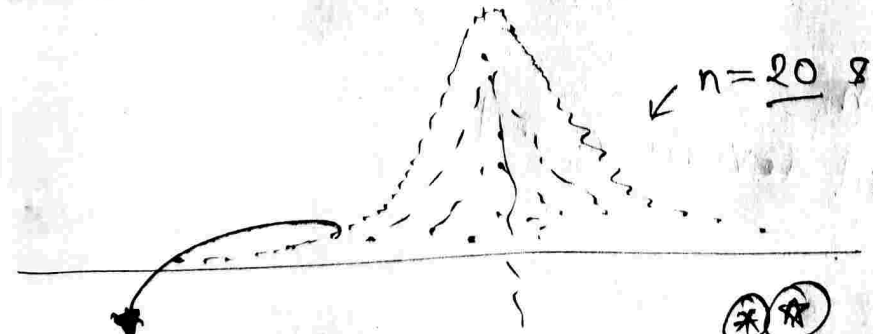
- $S_1 = [1, 1, 3, 6]$
  $\overline{x_1} = 2.75$
- $S_2 = [3, 4, 3, 1]$; $\overline{x_2} = 2.75$
- $S_3 = [1, 1, 6, 6]$; $\overline{x_3} = 3.5$

↙ frequency distributions



↓ 10,000 times





Sampling Distribution of the sample mean

It says; if we add a bunch of ~~samples~~ actions together, assuming that they all have same distribution, or we take mean of all of those actions, and if we were to plot the frequency of those means, we will get Normal distribution

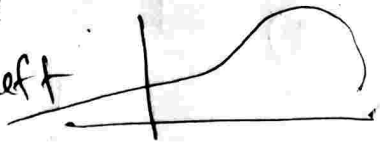
sample size → ∞

↓ 10,000 times

Normal distribution

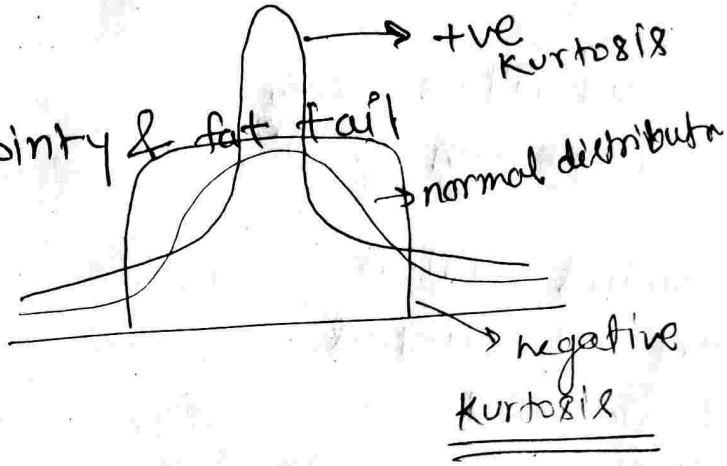
(#) positive skewed :- long tail toward right

Negative skewed : long tail toward left



(#) Kurtosis

- +ve kurtosis is pointy & fat tail



(#) Observation:

When sample size was

N=5 — skew = 0.12, kurtosis = -0.04 / -0.26

10,000 times

N=25 (✓) Normal dist. tighter fit to mean — skew = 0.03, kurtosis = -0.04

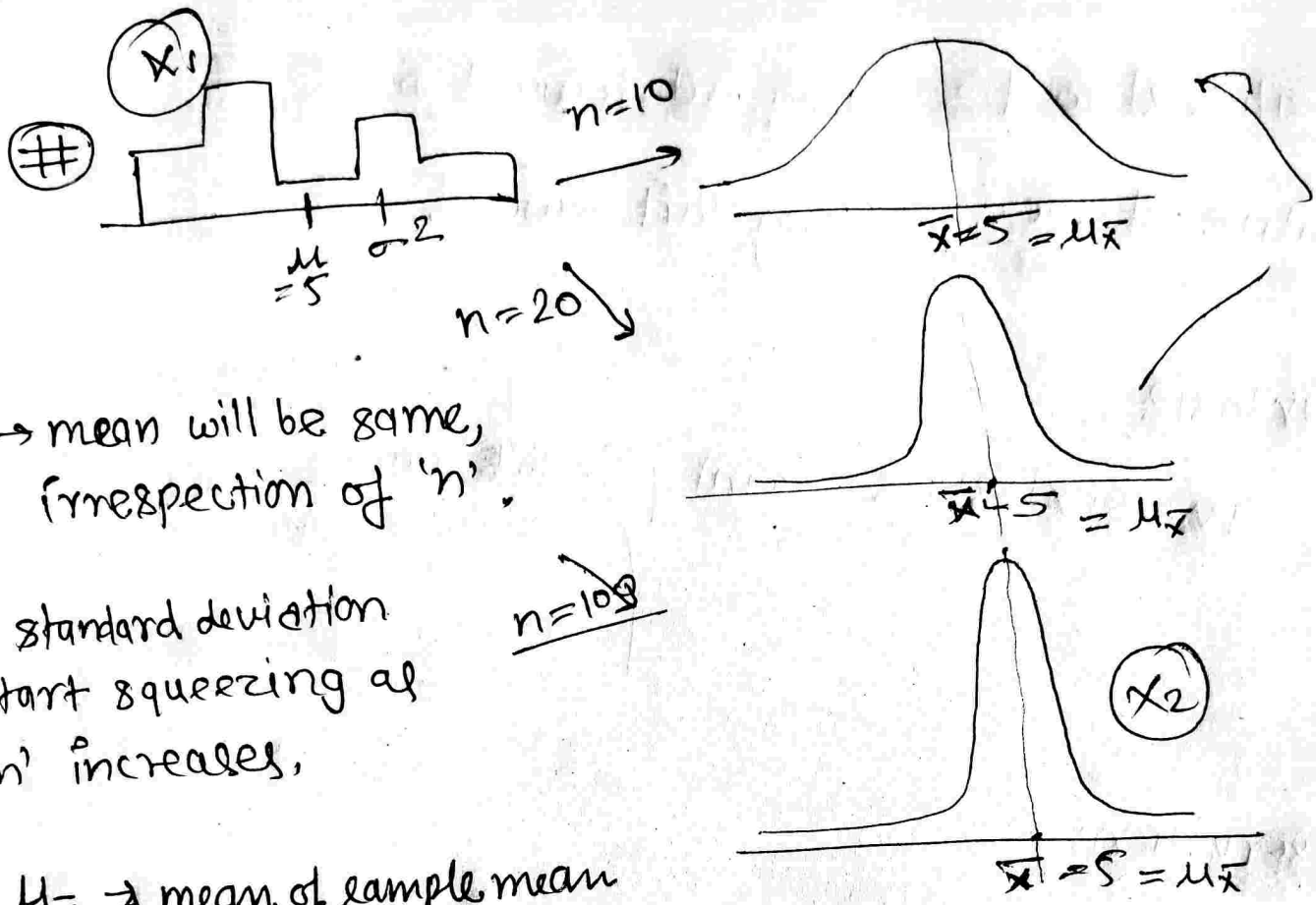
Conclusion: Central limit theorem says as you increase 'n' (sample size), you will lead to normal distribution

$n \to \infty \implies$ normal distribution

(sample size 4)

So: Sampling Distribution of Sample mean } will → normal distribution.

$n \uparrow \rightarrow$ standard deviation squeezes.



→ mean will be same, irrespective of 'n'.

• standard deviation start squeezing as 'n' increases,

• $\mu_{\bar{x}}$ ⇒ mean of sample mean

(#) Relation b/w probability distributed graph (X_1) & normal function's standard deviation → 's standard deviation

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = s^2 \text{ (variance)}$$

$\sigma_{\bar{x}}$ ⇒ standard deviation of sampling distribution of sample mean

(standard deviation of the mean ☆☆☆

or

★ Standard Error of the mean ← $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ ←

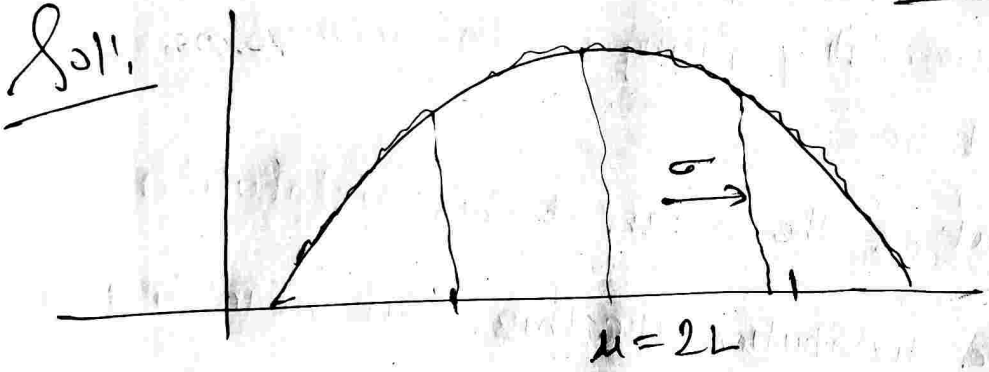
standard deviation

(#) example:

The average male drinks 2L of water when active outdoor (with standard deviation of 0.7L). You are planning a full day nature trip for 50 men & will bring 110 L of water. What is the probability that you will run out?

$\sigma = 0.7L$

Soln:

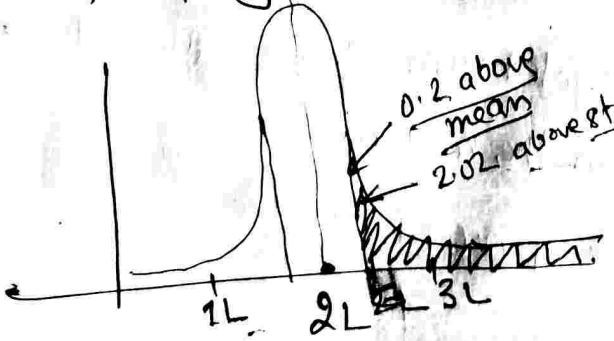


⇒ P(run out)

P(use more than 110L)

P(average water use per man) > 2.2L/m

⇒ Sampling distribution of the sample mean when n = 50



$\mu_{\bar{x}} = \mu = 2L$; $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \Rightarrow 0.099$

z-score :- $\frac{2.2-2}{0.099} = 2.020$

P($\bar{x}$ will be more than 2.020 std. deviation above the mean)

z(2.02) → 0.9783 → area till 2.020 ⇒ 1 − 0.9783 ⇒ 2.17%

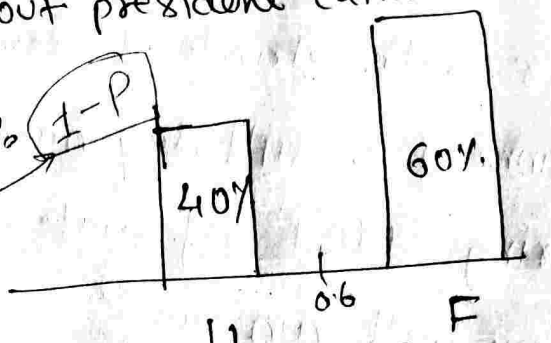
(Simplest Case of Binomial Distribution)

(#) ~~Bernat~~ Bernoulli Distribution

In population servey: about president candidate

Unfavorable view:- 40% (1-P)

favorable — 60% → P



pick a random member of that population & say what is the expected favorability rating of that member, ? what would it be?

or

what is the mean of this distribution?

(*) And for discrete distribution like this, your mean or expected value is just going to be

probability weighted sum of the different values

suppose U = 0, F = 1

$\mu = 0.4 * 0 + 0.6 * 1 = \underline{0.6}$

variance: σ^2 = probability weighted sum of the squared distances from the mean.

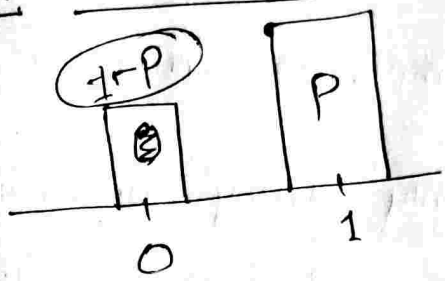
$= 0.4(0-0.6)^2 + 0.6(1-0.6)^2$

$\sigma^2 = \underline{0.24}$; $\sigma = \underline{0.49}$

(#) General formula for Bernoulli Distribution:

favorable → 60% — P

Unfavorable — 40% — 1-P



$\mu = (1-P)\cdot 0 + P(1)$

$\boxed{\mu = P}$ mean.

$\sigma^2 = (1-P)\cdot(0-\underset{=P}{\mu})^2 + P(1-P)^2$

$\sigma^2 = (1-P)(-P)^2 + P(1-P)^2$

$\sigma^2 = (1-P)(P)^2 + P(1-P)^2$

$\sigma^2 = P^2 - P^3 + P - 2P^2 + P^3$

$\boxed{\sigma^2 = P(1-P)}$ Variance

$\boxed{\sigma = \sqrt{P(1-P)}}$ → standard deviation.

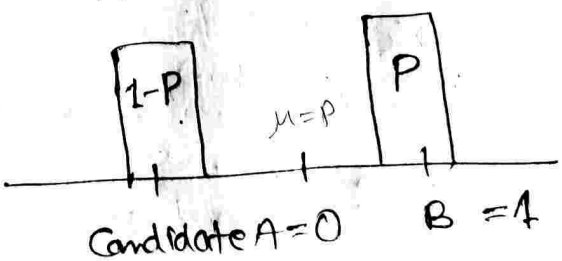
(#) Margin of Error:-

from Bernoulli Distribution:-

$\mu = P$

Random survey:

Presidential Election

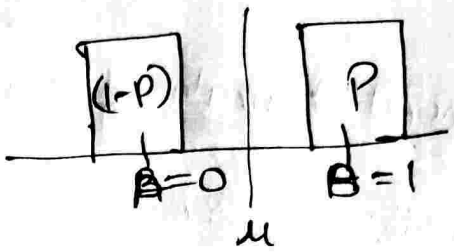


So; to make it manipulatable mathematically

A=0; B=1

→ Margin of Error

↳ Sample of population:



→ Randomly survey 100 people:

57 people → vote-for → A ⇒ 0

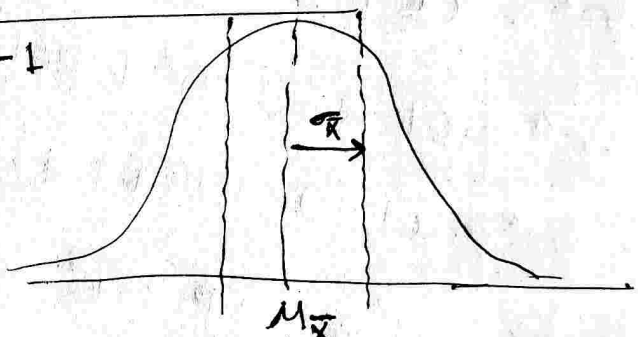
43 people → — — → B ⇒ 1

Sample Mean: $\frac{57*0+43*1}{100} = 0.43$

S^2: $\frac{57(0-0.43)^2 + 43(1-0.43)^2}{100-1} = 0.2475$

$S = 0.50$

sample standard deviation



sampling dist. of sample mean

↳ standard deviation of the sampling ~~mean~~ dist. of sampling mean

$\mu_{\bar{x}} = p$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ → sample size

→ population std. deviation

to know this

↓

we need to actually survey 100 million people,

↓

so, we will use our sampling std. deviation

$S = 0.50$

best estimate for σ

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}}$

$\sigma_{\bar{x}} \approx \frac{0.50}{10}$ → depend on sample

$\sigma_{\bar{x}} \approx 0.05$

↳ best estimation of $\sigma_{\bar{x}} \approx 0.05$ ⇒ 5%

⇒ Now, come up with an interval around the sample mean where I'am reasonalley confident using all my estimates that there's 95% chance that true mean is within two standard deviations

↳ $\mu = p$

↦ ~~[illegible]~~

$P(\bar{x}$ is within $2\sigma_{\bar{x}}$ of $\mu_{\bar{x}}) \approx 95\%$

$\bar{x}$ ↓ random sample mean

$P(\mu_x$ is within $2\sigma_{\bar{x}}$ of $\bar{x}) \approx 95.4$

μ_x / mean of sampling dist. ↓ p

$P(p$ is within ~~[illegible]~~ $2(0.05)$ of $\bar{x}) \approx 95\%$

proportion of the population saying 1

$P(p$ is within 0.10 of $\bar{x}) \approx 95\%$

$P(p$ is within $(0.43 \pm 0.10) \approx 95\%$

95% Confidence interval of 33% to 53%

↑ p range

43% will vote for candidate B

57% will vote for A

margin of error :— 10% ✓

example

Confidence Interval

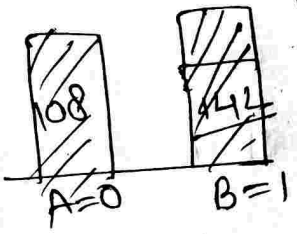
Ques:

In a local teaching district a technology grant is available. to teachers in order to install a cluster of four computers in their classrooms.

From the 6250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. of those selected, 142 teachers felt that computers were an essential teaching tool.

Sol: sample size = 250,

$n = 250$, $N = 6250$



142 people say → B=1

108 people say A=0

$\bar{x}$

250
-142
108

Questions:

① Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential tool.

② How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?

Sol:

sample mean $\bar{x} = \frac{108*0 + 142*1}{250} = 0.568.$

sample variance $s^2 = \frac{108(0-0.568)^2 + 142(1-0.568)^2}{250-1}$

$= \frac{0.322624 + 26.50}{249} = \frac{0.10077}{0.246}$

Std. deviation $S = 0.3282$ ✓

Sample

Sampling distribution of sampling means

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ ← population std. deviation

← sample size.

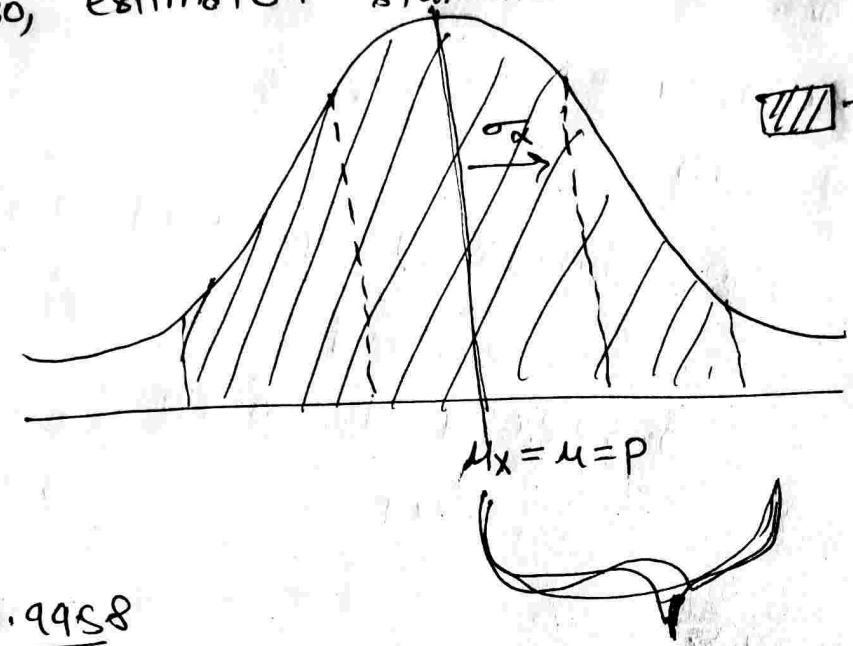
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{250}}$ $\quad \sigma_{\bar{x}} \approx \frac{0.50}{\sqrt{250}} = \frac{0.031}{15.81}$

Estimated Std. deviation of Sampling dist. of Sampling mean

$\sigma_{\bar{x}} \approx 0.031$ ↳

↳ How many standard deviations away from the mean do we have to be that we can be 99% confident that any sample from the sampling distribution will be in that interval?

So, estimated standard deviation $\sigma_{\bar{x}} \approx 0.031$



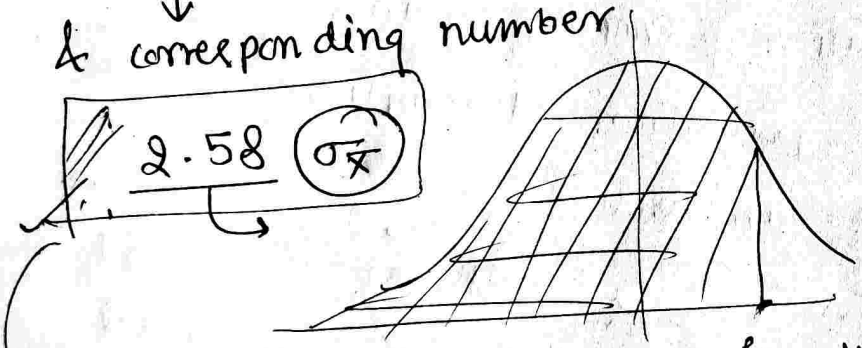
▨ → 99% confidence zone.
↳ so basically find its boundary in std. deviations terms.

Check. 0.9958

value in z-score table ← $\frac{99}{2} = 0.495$ as it symmetric around μ.

↓

& corresponding number



99% chance that a random $\bar{x}$ is with $2.58\ \sigma_{\bar{x}}$ of p

"Confident" 99% chance that ~~$\bar{x}$~~ p is within

~~0.08~~ 0.08 / ~~$\bar{x}$~~ of ~~$\bar{x}$~~ "

0.568

→ 0.568 ± 0.08 ✓ (0.488 to 0.648)

$\bar{x}$

↓

48.8 to 64.8 %

(6): take more samples.

Example: 7 patients blood pressures have been measured after having been given a new drug for 3 months. They had blood pressure increases of 1.5, 2.9, 0.9, 3.9, 3.2, 2.1 & 1.9. Construct a 95% confidence interval for the true expected blood pressure increase for all patient in a population

Sol: sample size = 7

$$\bar{x} = \frac{1.5+2.9+0.9+3.9+3.2+2.1+1.9}{7} = \frac{16.4}{7} = 2.342$$

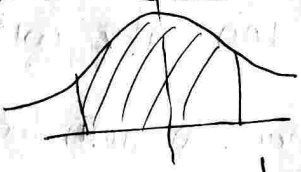
$$S^2 = 1.04$$

So, we estimate the true standard-deviation of the population with our sample standard deviation.

$$\sigma \approx S = 1.04$$

But here only 7-sample so, not so good estimates.

because n is small. ($n<30$)

t-distribution → 

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{7}} \approx \frac{S}{\sqrt{7}} \Rightarrow$$

sample size is small

↓

normal ~~distribution~~

t-distribution ✓

for t-distribution

use t-table!

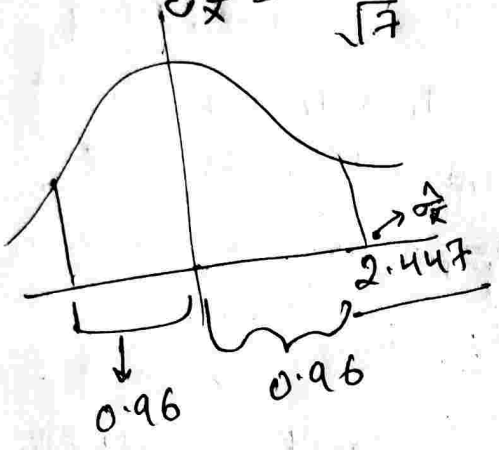
$n=7$; so $(n-1)$ degree of freedom

Two-sided row → 95%, 6 degree

t table → 2.447 approximated standard deviation

↳ $\hat{\sigma}_{\bar{x}}$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{7}} \approx \frac{(s)}{\sqrt{7}} = \hat{\sigma}_{\bar{x}} = \underline{0.39}$



$\rightarrow 2.447\ (\sigma_{\bar{x}})$

$\rightarrow 2.447(0.39) = 0.96$

Confident 95% chance, $\bar{x}$ 2.34 is within 0.96 of true population mean. (μ).

$\Rightarrow$ 95% chance μ is within 0.96 of 2.34

(#) Hypothesis testing & P-values

example: A neurologist is testing a effect of a drug on response time by injecting 100 rats with a unit dose of the the drug, subjecting each to neurological stimulus, and recording its response time.

mean response time for rats not injected with the drugs is 1.2 seconds.

mean of 100 injected rats' response times is 1.05 sec with sample standard deviation of 0.5 seconds.

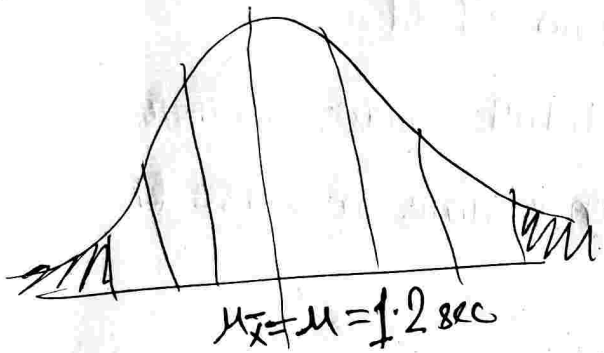
Do you think that the drug has an effect on response time?

Sol:

H_0 : Drug has no effect $\Rightarrow \mu = 1.2$ sec (even w/o drug)

H_1 :- Drug has an effect $\Rightarrow \mu \neq 1.25$ when the drug is given.

Assume H_0 is true:



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}} \Rightarrow \frac{0.5}{\sqrt{100}} = \frac{0.5}{10}$$

$$\sigma_{\bar{x}} = 0.05$$

What is the probability of getting 1.05 seconds?

Or another way is how many standard deviation away from this mean is 1.05 sec so that we can use empirical rule.

Z-score is

$$\Rightarrow \frac{(1.2 - 1.05)}{0.05}$$

$$Z = 3$$

3 standard deviation away from mean

probability of getting result (-ve direction + +ve direction

3-deviation area $\doteq$ 99.7%.

$1 - 99.7 = 0.3\%$ $\quad .003$

So, H_0 should be rejected as it has

value 0.3%, probability of getting a result more extreme than this given the null hypothesis is called p-value.

(*) p-value (probability value) = 0.003

(*) So, null hypothesis has no effect.

H_1 has an effect; but didn't know whether the drug would lower the response time or raise the response time

In this situation, where we're really just testing to see if it had an effect, whether an extreme positive effect, or an extreme negative effect, both considered as an effect. called a two-tailed test (*)(*)

If we want to do one-tailed test (*)

H_0 → Drug has no effect $\mu_{drug} = 1.2$ sec → does not lower

H_1 → Drug lowers response $\mu_{drug} < 1.2$ sec

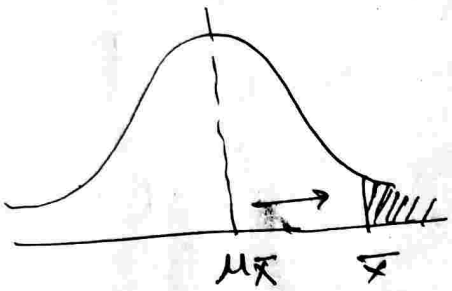
→ Care about one direction

H_0 → 0.15% → 0.0015

(*)

(#) Z-statistic

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



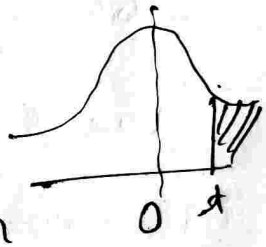
$$\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \Rightarrow \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

$$\Rightarrow Z \approx \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} \quad ; \quad \text{'ok' if } n > 30$$

↳ Z table

normal distributed

t-statistic : $\frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$ if n is small → t-distribution



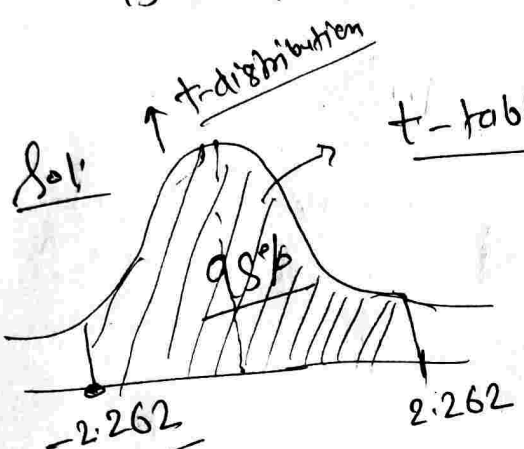
↳ t-table

(#) Type-I error :- Rejecting null hypothesis H_0 even though it is true

(#) example: find 95% confidence interval $n = 10$

15.6 16.2 22.5 20.8 16.4 19.4 16.6 17.9 12.7 13.9

Sol:



t-table: two-sided row $n = 10$ degree $= 9$

t-value = 2.262

there is 95% chance, that if you pick random T-statistic is going to be less than 2.262

$$-2.262 < t < 2.262$$

sample mean: $17.17 \Rightarrow \bar{x}$

$S = 2.98$

$$t\text{-statistic} \Rightarrow \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} \Rightarrow < 2.267$$

$$-2.267 < \frac{17.17 - \mu_{\bar{x}}}{\frac{2.98}{\sqrt{10}} \rightarrow 0.9423} < 2.268$$

$$-2.267 * 0.9423 < 17.17 - \mu < 2.268 * 0.942$$

$$-2.13 < 17.17 - \mu < 2.13$$

multiple (-1)

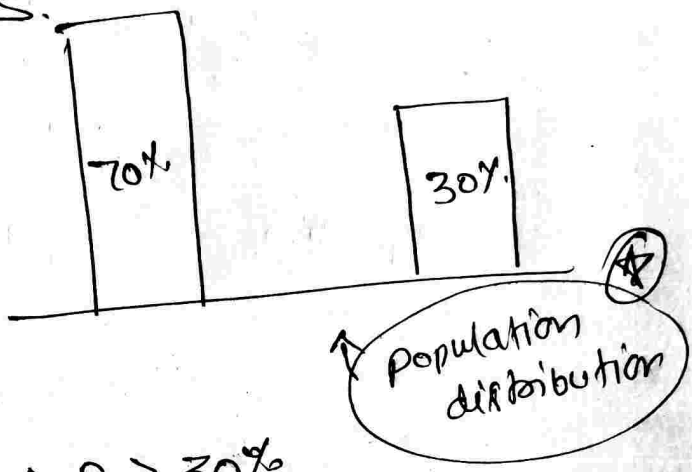
$$2.13 < \mu - 17.17 < -2.13$$

$$2.13 + 17.17 < \mu > -2.13 + 17.17$$

$$19.3 > \mu > 15.04$$

↑ 95% chance

(#) We want to test the hypothesis that more than 30% of U.S. households have internet access (with a significance of level 5%). We collect a sample of 150 household & find that 57 have access.



Sol

Bernoulli Distribution

Variance: $\sigma^2 = P(1-P)$

$\mu = P$

$H_0 : P \leq 30\%$ $\quad H_1 \Rightarrow P > 30\%$

Assume P-value based on null hypothesis (portion based on null hypothesis of population & given that assumption what is the probability that ≤ 57 out of 150 of our samples actually have internet. And if that probability is less than 5% → reject H_0

Assume H_0 true

sample proportion $\Rightarrow \frac{57}{150}$

$\bar{P} = 0.38$

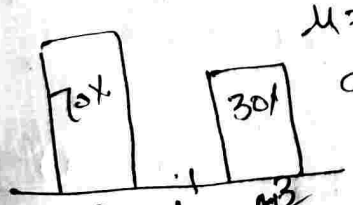
$P_{H_0} = 0.30$

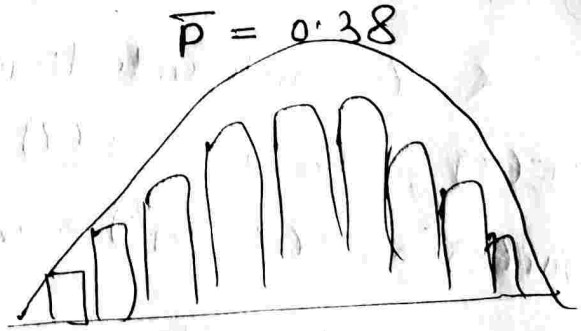
$\mu = P$

$\sigma_{H_0} = \sqrt{P(1-P)}$

$= \sqrt{(0.3)(0.7)}$

$= \sqrt{.21}$

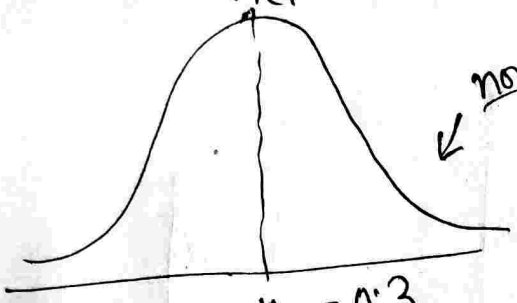




distribution of sample proportion

Binomial distribution

if $np > 5$ → $n(1-p) > 5$ } we can assume, distribution is normal.



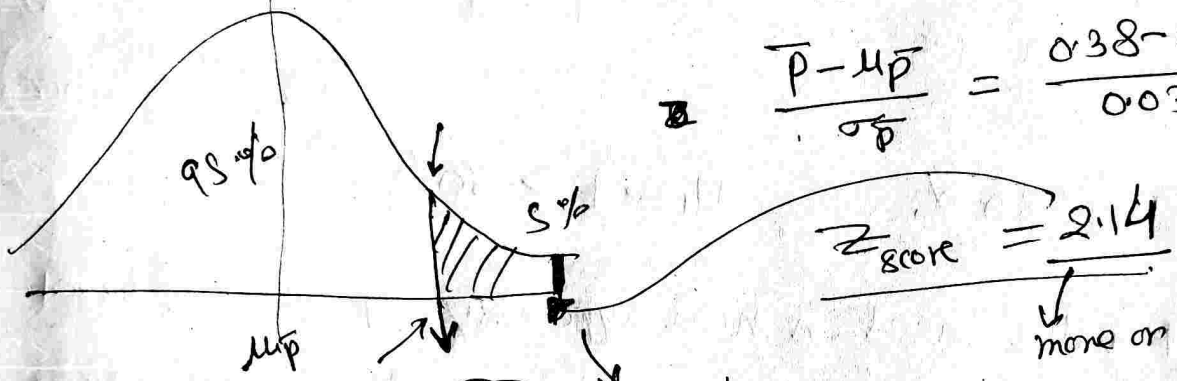
$\sigma_{\bar{p}} = \frac{\sigma_{H_0}}{\sqrt{n}} = \frac{\sqrt{.21}}{\sqrt{150}} = 0.037$

$\sigma_{\bar{p}}$ → population

$\mu_{\bar{p}} = 0.3$

how much away from std. dev

$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{0.38 - 0.3}{0.037} = \frac{0.08}{0.037}$



$z_{score} = 2.14$ → more or less 5%

what z value which has 95%

probability here → p-value

Variance of differences of random variables ☆

X Y independent random variables.

$E(X) = \mu_X$; $E(Y) = \mu_Y$;

$Var(X) = E((X-\mu_X)^2) = \sigma_X^2$; $Var(Y) = E((Y-\mu_Y)^2) = \sigma_Y^2$

third random variable

$Z = X + Y$ so $E(Z) = ? = E(X \pm Y) = E(X) \pm E(Y)$

$\mu_Z = \mu_X \pm \mu_Y$ ——— ☆//

variance (Z) = Var(X) + Var(Y)

$$\sigma_Z^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \quad \& \quad \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

$$\sigma_A^2 = \sigma_{X-Y}^2 = \sigma_{X+(-Y)}^2 = \sigma_X^2 + \sigma_{-Y}^2$$

same as it is square

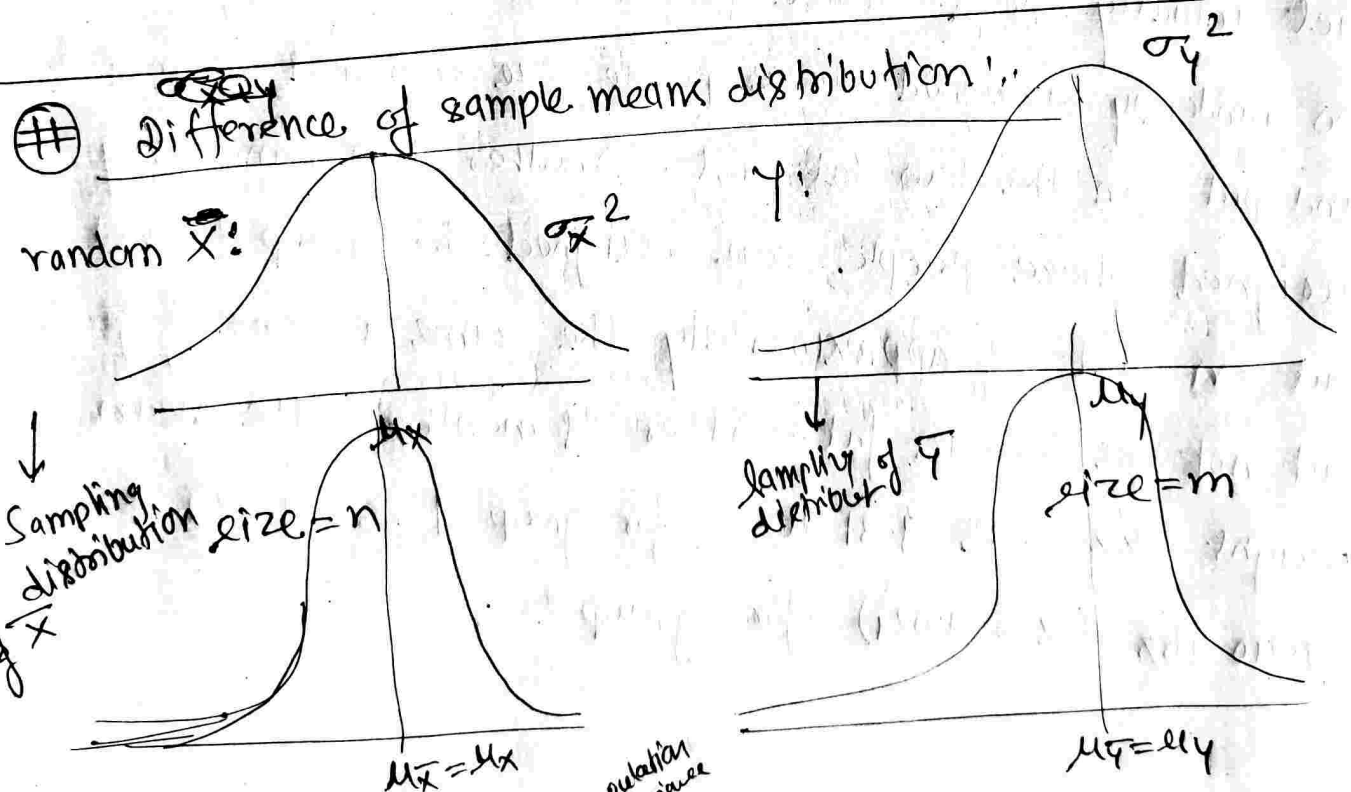
$$\sigma_{-Y}^2 = \text{Var}(-Y) = E((-Y - E(-Y))^2$$

$$= E((-1)^2 (Y + E(-Y)^2))$$

$$\therefore E(-Y) = -E(Y)$$

$$E((Y - E(Y)^2) = \sigma_Y^2$$

Difference of sample means distribution :-



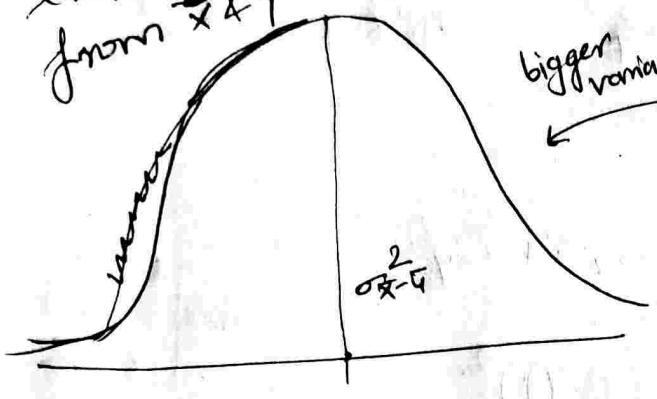
$$\sigma_{\bar{X}}^2 = \frac{\sigma_{\text{pop } X}^2}{n} \quad ; \quad \sigma_{\bar{Y}}^2 = \frac{\sigma_{\text{pop } Y}^2}{m}$$

(population variance)

taking sample from $\bar{x}$ & $\bar{y}$

$Z = \bar{x} - \bar{y}$ → useful to get confidence level

$\mu_{\bar{x}-\bar{y}} = \mu_Z = \mu_{\bar{x}} - \mu_{\bar{y}}$



bigger variance

$\sigma^2_{\bar{x}-\bar{y}} = \sigma^2_{\bar{x}} + \sigma^2_{\bar{y}}$

↓ for $\bar{x}$ ↓ for $\bar{y}$

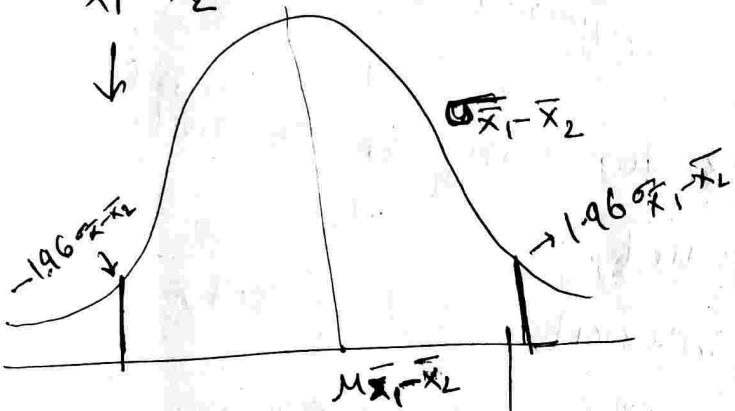
$\sigma^2_{\bar{x}-\bar{y}} = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$ → population for y

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

example: We're trying to ~~get~~ test whether a new, low fat diet actually helps obese people ~~less~~ lose weight. 100 randomly assigned obese people are assigned to group 1 and put on the low fat diet. Another 100 randomly assigned obese people are assigned to group 2 and put on diet of approximately the same amount of food, but not as low in fat. After 4 months, the mean weight loss was 9.31 lbs. for group 1 ($s = 4.67$) & 7.40 lbs. ($s = 4.04$) for group 2.

Low-fat: $\bar{x}_1 = 9.31$ lbs, $s_1 = 4.67$ } given Control: $\bar{x}_2 = 7.40$, $s_2 = 4.04$

$\bar{x}_1 - \bar{x}_2 = 9.31 - 7.40 = 1.91$ → differences of sample



→ maybe you lose an incremental 1.91 pounds every 4 month if you are on this diet.

$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$

create 95% interval

↓ how many std. deviation

↓ look z table

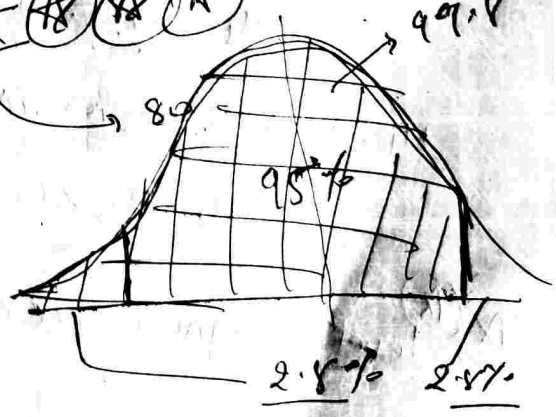
↓ gives cumulative value

1.96

95% chance that 1.91 is within $1.96\sigma_{\bar{x}_1-\bar{x}_2}$ of the $\mu_{\bar{x}_1-\bar{x}_2}$

So get a 95% confidence interval around this number (★)

do we lose weight?



$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{S_1^2}{100} + \frac{S_2^2}{100}}$

$= \sqrt{\frac{(4.67)^2}{100} + \frac{(4.04)^2}{100}} = 0.617$

So $1.96 * 0.617 = 1.21$

so 95% level 1.91 ± 1.21 ✓

↓ so find z for 95 + 2.5% = 97.5%

95% confidence interval for $\mu_{\bar{x}_1 - \bar{x}_2}$

1.91 ± 1.21

$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$ ← 95% confidence level of this

μ_1 & μ_2 ↓ population

The true diff. in weight lose b/w going on the low-fat diet & not going on low-fat diet.

95% confidence interval 0.7 — 3.12

means 95% confidence that you will lose some weight %

Hypothesis test for above example:

Ho :- low-fat does nothing

$\mu_1 - \mu_2 = 0 \Rightarrow \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = 0 \Rightarrow \mu_{\bar{x}_1 - \bar{x}_2} = 0$

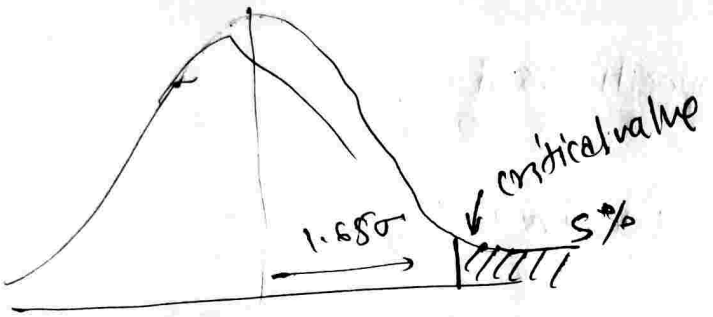
H1 :- improvement

$\mu_1 - \mu_2 > 0 \Rightarrow \mu_{\bar{x}_1 - \bar{x}_2} > 0$

Level of significance: (threshold level or alpha) ⇒ 5%

significance level of 95% $\alpha =$ 5%

⇒ 5%



$\mu_{\bar{x}_1 - \bar{x}_2} = 0$

step (1) critical z-value for 95% ⟶ critical value.

from z table ⟹ 1.65

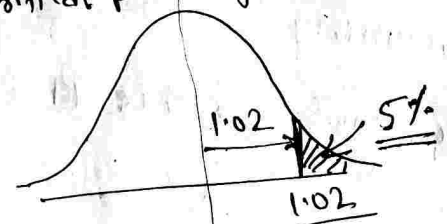
step (2) $\sigma_{\bar{x}_1 - \bar{x}_2} = ?$

$\sigma_{\bar{x}_1 - \bar{x}_2} = 0.617$ (already calculated)

step (3) so; z value * $\sigma_{\bar{x}_1 - \bar{x}_2}$

$1.65 * (0.617) = 1.02$ } → critical point on graph

critical point from mean is 1.02 far.



step (4) $\bar{x}_1 - \bar{x}_2 = 1.91$; where it lies; 95% or 5% interval

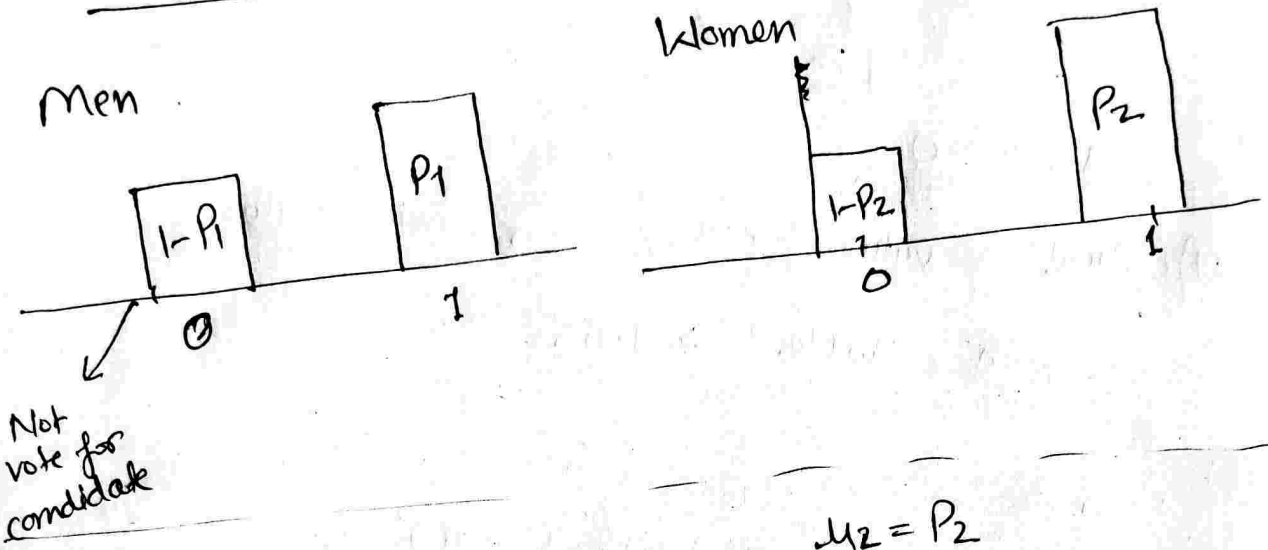
its in 5% interval

so Ho has less than 5% probability.

less than α.

Ho (rejected)

(#) Comparing population proportions 1 :-



$\mu_1 = P_1$

$\sigma_1^2 = P_1(1-P_1) =$

$\mu_2 = P_2$

$\sigma_2^2 = P_2(1-P_2)$

? Get the meaningful difference b/w the way that the men will vote & women will vote.

$P_1 - P_2$ is meaning

& come up 95% confidence level for this $(P_1 - P_2)$

$n = 1000$ men

642 → vote → for 1

rest → 0

1000 women

591 → 1

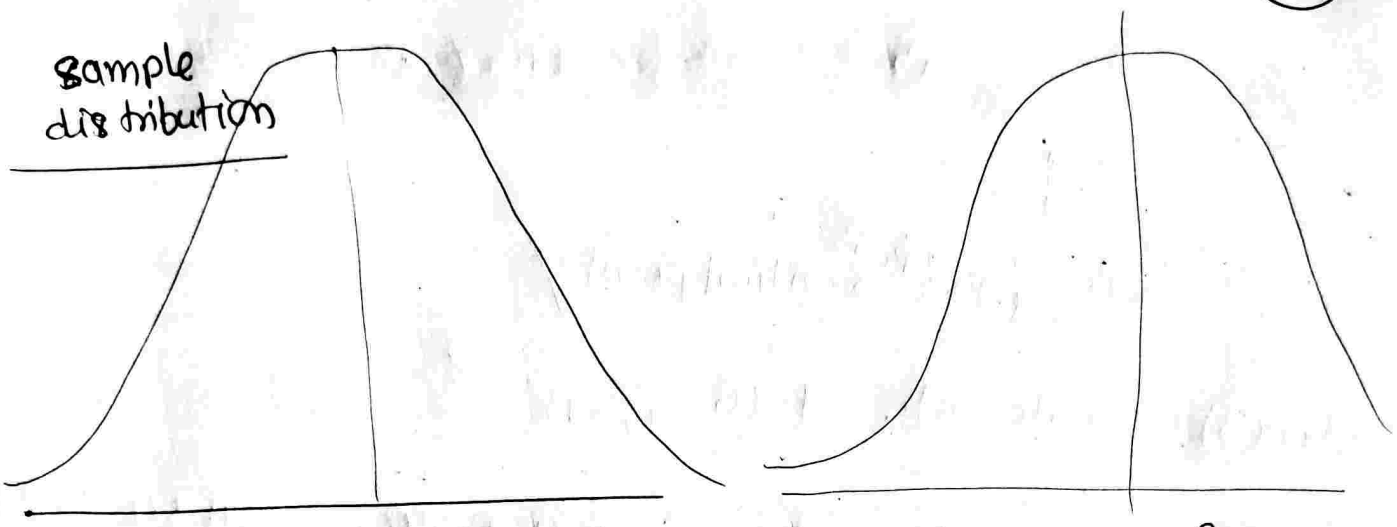
rest → 0

$\bar{P} = 0.642$

sample mean or proportion

$\bar{P} = 0.591$

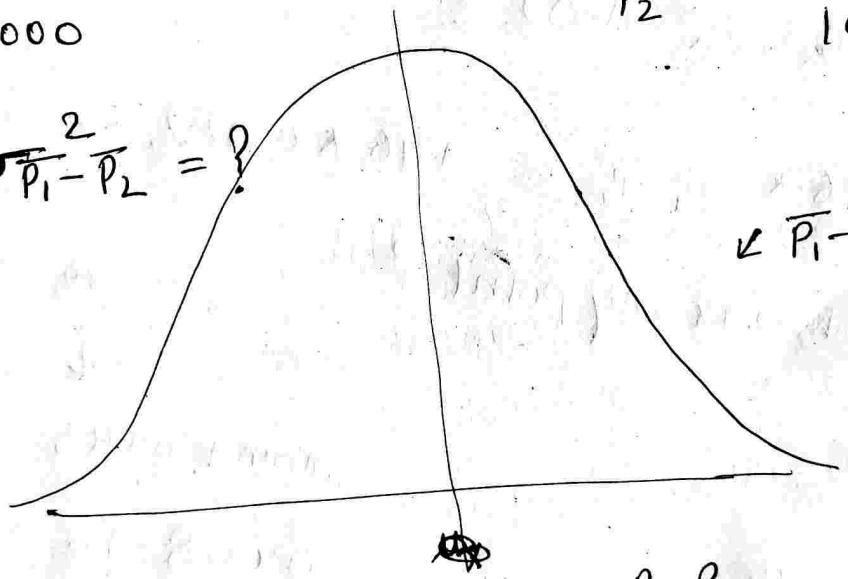
Sample distribution



$\mu_{\bar{p}_1} = p_1$ $\mu_{\bar{p}_2} = p_2$

$$\sigma^2_{\bar{p}_1} = \frac{p_1(1-p_1)}{1000} \qquad \sigma^2_{\bar{p}_2} = \frac{p_2(1-p_2)}{1000}$$

step (1) & $\sigma^2_{\bar{p}_1 - \bar{p}_2} = ?$



↙ $\bar{p}_1 - \bar{p}_2$

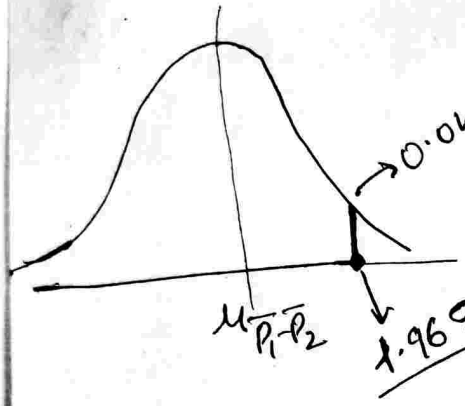
$\mu_{\bar{p}_1 - \bar{p}_2} = p_1 - p_2$

$$\sigma^2_{\bar{p}_1 - \bar{p}_2} = \sigma^2_{\bar{p}_1} + \sigma^2_{\bar{p}_2}$$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{1000} + \frac{p_2(1-p_2)}{1000}}$$

step (2) critical point for 95% interval

$\bar{p}_1 - \bar{p}_2 = 0.051$ ← (our point)



look for 97.5% in Z-table.

→ critical point

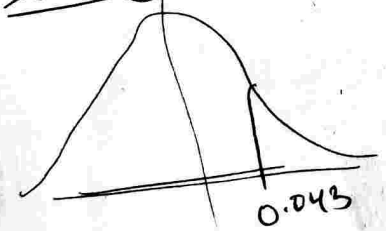
Step ③ Calculate $1.96\ \sigma_{\bar{P}_1-\bar{P}_2}$?

$$\therefore \sigma_{\bar{P}_1-\bar{P}_2} \approx \sqrt{\frac{0.642*(1-0.642)}{1000} + \frac{.591*(1-.591)}{1000}}$$

$$\sigma_{\bar{P}_1-\bar{P}_2} \approx 0.022$$

$$1.96 * \sigma_{\bar{P}_1-\bar{P}_2} = 1.96 * 0.022 = 0.043$$

Step ④ where is my point ?



so range is

↓

mean ± 0.043

0.008 to 0.094

↑

& 0.051 lies within

↓

so 95% confidence

95% chance men are more likely to vote for candidate than women.

Hypothesis test:

H_0 : No difference $P_1 - P_2 = 0$;

H_1 : $P_1 - P_2 > 0$; $P_1 \neq P_2$

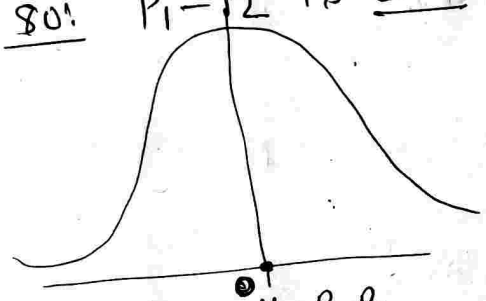
$\alpha = 5\%$

Assume H_0 is true; figure out probability for ★★

$P(\bar{P_1} - \bar{P_2} \mid H_0) < 5\%$

then reject H_0

So: $P_1 - P_2$ is zero



critical point

$\bar{P_1} = 0.642$ $\quad \bar{P_2} = 0.591$

$\bar{P_1} - \bar{P_2} = 0.051$ → my point on graph

probability to get this

$$Z\text{-score} = \frac{0.051 - 0}{\sigma_{\bar{P_1}+\bar{P_2}}}$$ → new

$$\sigma_{\bar{P_1}-\bar{P_2}} = \sqrt{\frac{P_1(1-P_1)}{1000} + \frac{P_2(1-P_2)}{100}}$$

in H_0 $P_1 = P_2$

$$= \sqrt{\frac{1}{1000}\left(P(1-P) + P(1-P)\right)}$$

$$\sigma_{\bar{P_1}-\bar{P_2}} = \sqrt{\frac{2P(1-P)}{1000}}$$

so our sample $1000 + 1000 = 2000$

$$\bar{P} = \frac{642 + 591}{2000} = 0.6165$$

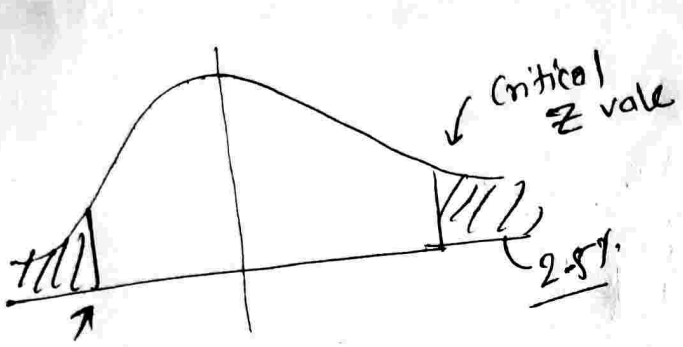
$$\sigma_{\bar{P_1}-\bar{P_2}} = \sqrt{\frac{2 * .6165(1-.6165)}{1000}}$$

$$= 0.0217$$

$$Z = \frac{0.051}{0.0217} \Rightarrow 2.38$$

↓ std. deviat away

my point in term of σ



97.5% → Z 1.96

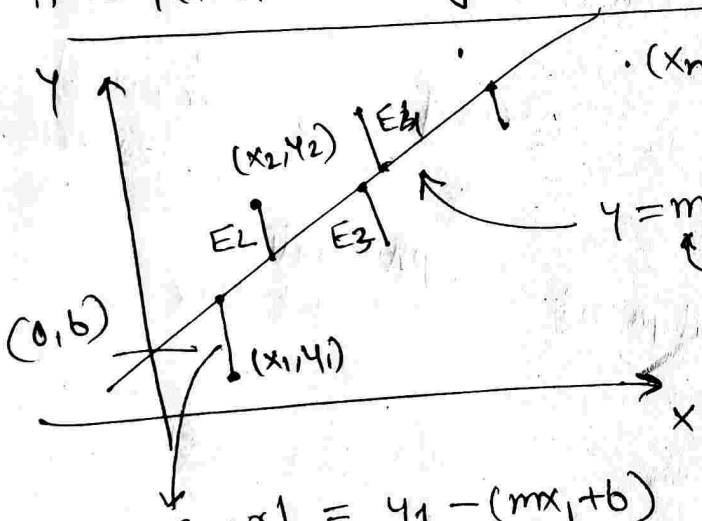
Z value in term of std. deviation

⇒ 2.35 is extreme right/left of 1.96

Hence probability is < 5%

⇒ Hence, reject Ho

Squared error of regression line :-



Error 1 = $y_1 - (mx_1 + b)$

$E2 = y_2 - (mx_2 + b)$

$En = y_n - (mx_n + b)$

$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \cdots + (y_n - (mx_n + b))^2$

↑ find m & b that minimize SE_{Line}

$$SE_{LINE} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \cdots + (y_n - (mx_n + b))^2$$

SE_{LINE}

$$= y_1^2 + 2y_1(mx_1 + b) + (mx_1 + b)^2 + y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 + \cdots + y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2$$

$$= y_1^2 - 2y_1mx_1 - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 + y_2^2 - 2y_2mx_2 - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 + \cdots + y_n^2 - 2y_nmx_n - 2y_nb + m^2x_n^2 + 2mx_nb + b^2$$

$$= (y_1^2 + y_2^2 + \cdots + y_n^2) - 2m(x_1y_1 + x_2y_2 + \cdots + x_ny_n) - 2b(y_1 + y_2 + \cdots + y_n) + m^2(x_1^2 + x_2^2 + \cdots + x_n^2) + 2mb(x_1 + x_2 + \cdots + x_n) + nb^2 \quad \text{——— (1)}$$

SE_{LINE} is a function of $\frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n} = \overline{y^2}$ ——— (a)

$\left(\Rightarrow y_1^2 + y_2^2 + \cdots + y_n^2 = n\overline{y^2}\right)$

$$\frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{n} = \overline{xy} \Rightarrow x_1y_1 + x_2y_2 + \cdots + x_ny_n = n\overline{xy} \quad \text{——— (b)}$$

similiarly we have $n\overline{y}$, nx^2, $n\overline{x}$ ——— (c)

Plug ⓐⓑⓒ in ①.

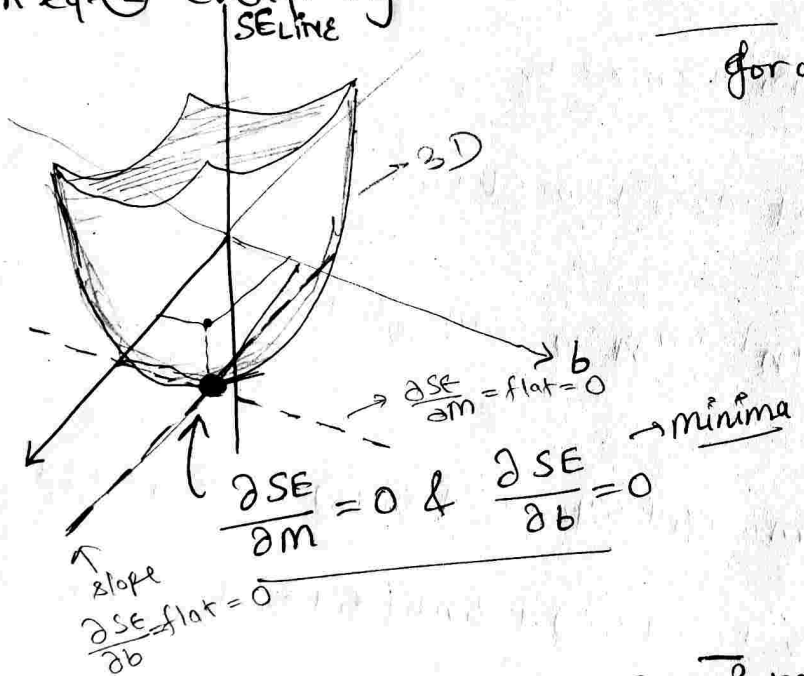
$$SE_{LINE} = n\bar{y}^2 - 2mn\overline{xy} - 2bn\bar{y} + m^2n\bar{x}^2 + 2mbn\bar{x} + nb^2 \quad \text{---- ②}$$

eq-② breaking into 3-D calculus

eq-② is a surface ;

In eq-② everything is a constant except m's & b's.

for all comb^n of m's & b's



$$\frac{\partial SE}{\partial m} = 0 \;\&\; \frac{\partial SE}{\partial b} = 0$$

$$SE_{LINE} \Rightarrow \frac{\partial SE}{\partial m} = -2n\overline{xy} + 2n\bar{x}^2 m + 2bn\bar{x} = 0 \;; \rightarrow /2$$

$$\frac{\partial SE}{\partial b} = -2n\bar{y} + 2mn\bar{x} + 2nb = 0 \;; \rightarrow /2n$$

$$-\overline{xy} + m\bar{x}^2 + b\bar{x} = 0 \quad \text{(1)}$$

$$-\bar{y} + m\bar{x} + b = 0 \quad \text{(2)} \quad \text{simplified}$$

add $\overline{xy}$ both side. at eq-ⓐ ⑪ & adding $\bar{y}$ both side

$$\overline{xy} - \overline{xy} + m\bar{x}^2 + b\bar{x} = 0 + \overline{xy}$$

$$\bar{y} - \bar{y} + m\bar{x} + b = 0 + \bar{y}$$

$$m\bar{x} + b = \bar{y}$$

$m\bar{x}^2 + b\bar{x} = \overline{xy}$

$m\bar{x} + b = \bar{y}$ → $(\bar{x}, \bar{y})$ lies on the line

$y = mx + b$

divide $\bar{x}$

$m\frac{\bar{x}^2}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$ → $\left(\frac{\bar{x}^2}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ on best fitting line

So $(\bar{x}, \bar{y})$ & $\left(\frac{\bar{x}^2}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ lies on $y = mx + b$

$m\bar{x} + b = \bar{y}$

$m\frac{\bar{x}^2}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$

$- \quad - \quad -$

$m\left(\bar{x} - \frac{\bar{x}^2}{\bar{x}}\right) = \bar{y} - \frac{\overline{xy}}{\bar{x}}$

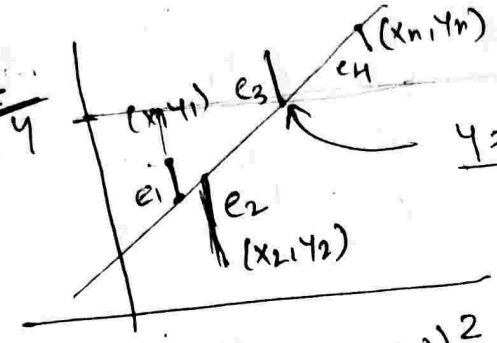
$m = \dfrac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\bar{x}^2}{\bar{x}}}$ $\dfrac{(\bar{x})}{(\bar{x})}$ ⇒ $\dfrac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \bar{x}^2}$

$b = \bar{y} - m\bar{x}$ ✓

$m = \dfrac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \bar{x}^2}$ ✓ $= \dfrac{\overline{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}$

x	y	x^2	xy
$\bar{x}$	$\bar{y}$	$\bar{x}^2$	$\overline{xy}$

ex1.



$y = mx + b$ → How good fit it is ??

$$SE_{LINE} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots + (y_n - (mx_n + b))^2 \quad \text{—(1)}$$

(actual − predicted)²

How much (what %) of the total variation in y is described by the variation in x.

total variation in y $\leftarrow \Sigma$(actual − mean)²

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2 \Rightarrow SE_{\bar{y}}$$

↑ squared error from $\bar{y}$

variance = $\frac{\text{total variation}}{n}$

$\frac{SE_{LINE}}{SE_{\bar{y}}}$ → tells what % of ~~var~~ total variation is described by variation in x (or by the regression line)

How much of the total variation is not described by the regression line :- ?

then remainder will be described by the line.

$1 - \frac{SE_{LINE}}{SE_{\bar{y}}}$ → $\frac{\text{b/w actual \& line}}{\text{squared error b/w } \bar{y}}$

↳ what % of total percentage variation described by line.

(#) Coefficient of determination

$$r^2 = \boxed{1 - \frac{SE_{LINE}}{SE_{\bar{y}}}}$$

if squared is really small :⟶ line is good fit

⟶ $\frac{SE_{LINE}}{SE_{\bar{y}}}$ will be small

⟶ r^2 will be close to 1 (*) (*)

⟶ if large ⟶ r^2 will close to 0.

example: for points $(-2,-3)$, $(-1,-1)$, $(1,2)$ & $(4,3)$

sol: best fit line :

by formula : $m = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$ & $b = \bar{y} - m\bar{x}$

Step (1)

$$y = \frac{41}{42}x - \frac{5}{21}.$$

Step (2) : sqrd from mean $\bar{y}$

total squared error from $\bar{y}$ ⇒

$SE_{\bar{y}} = 22.75$

total sqrd error ⇒

$SE_{LINE} = 2.74$

step (3) $\frac{2.74}{22.75}$ → % of total variation not explained by variation in x.

$\Rightarrow r^2 = 1 - 0.12 = 0.88$

88% of total variation is explained by variance in X.

(#) Covariance the regression line (#) (*)

$$cov(X,Y) = E[(X - E[X])(Y - E(Y))]$$

$$= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]$$

$$= E[XY] - E[XE[Y]] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]]$$

$$cov(X,Y) = E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]$$

$$\boxed{cov(X,Y) = E[XY] - E[Y]E[X]} \checkmark$$

$E[Y] \rightarrow \bar{Y}$, $E[X] \rightarrow \bar{X}$

$E[XY] \approx \overline{XY}$

$$cov(X,Y) = \overline{XY} - \bar{Y}\bar{X}$$ ← numerator of slope of line

$$\hat{m} = \frac{\overline{XY} - \bar{Y}\bar{X}}{\overline{X^2} - (\bar{X})^2}$$ ← $cov(X,Y)$ (#)

$$\overline{X \cdot X} - \bar{X} \cdot \bar{X}$$

$\rightarrow cov(X,X)$ (*)

$$\hat{m} = \frac{cov(X,Y)}{Var(X)} \checkmark$$

covariance means:- how changes in X are associated to Y

Chi-square distribution

$X \rightarrow$ random variable; normally distributed

$X \sim N(0, 1)$ $\quad E[X] = 0$; $\quad Var(X) = 1$.

(0 ↑ mean; 1 ↑ variance)

$Q_1 = X^2$

↳ the distribution for this random variable will be an example of chi-square distribution

$Q_1 \sim \chi^2_1$ ← degree of freedom

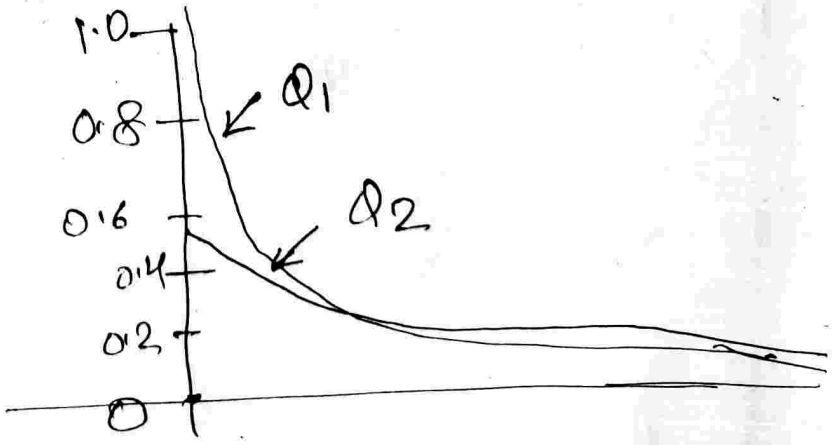
$Q_2 = X_1^2 + X_2^2$

↳ chi-square distribution.

$Q_2 \sim \chi^2_2$ ← degree of freedom

$Q_3 = X_1^2 + X_2^2 + X_3^2$

↳ $Q_3 \sim \chi^2_3$



What is the probability of $Q_2 > 2.41$?

$P(Q_2 > 2.41) \rightarrow$ look χ-square table against d (degree of freedom)

$= 0.30$

30%