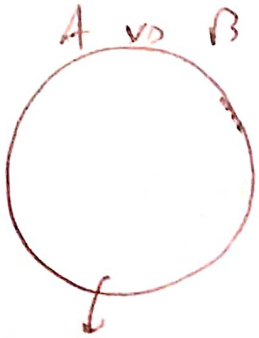
4-12-2020

Unit - II

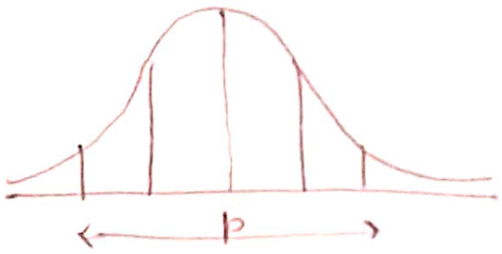
Confidence intervals

- Let's suppose there is an election going on b/w A and B. Let p be the prob that A will win. Out of the ~~two~~ total population, proportion that voted for A is p.



p that support A.

- Now since the population is very large, it is impossible (nearly) to calculate p.
- What we will instead do is infer the value of p (population) from samples along with its confidence.
- If we take a sample and calculate the proportion of people that support A. (sample is of size $n = 100$)
- We know from last chapter that the distribution of these proportions of samples will be normal.



→ mean (p = population parameter)

std = $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- Now for a sample if $\hat{p} = 0.54$.
- Can we infer the value of p (population) from $\hat{p}$ (sample).
- There's 95% chance that $\hat{p}$ will lie within $2\,\sigma_{\hat{p}}$ from p.
- Or we can say that we are 95% sure that p will lie b/w $(\hat{p} - 2.\sigma_{\hat{p}}$ to $p + 2.\sigma_{\hat{p}})$.

∴ With 95% confidence, p will lie betwe

$0.54 - 2.\sigma_{\hat{p}}$, $0.54 + 2.\sigma_{\hat{p}}$

- How to calculate $\sigma_{\hat{p}}$. (since it needs p) $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- Calculate std dev of sample.

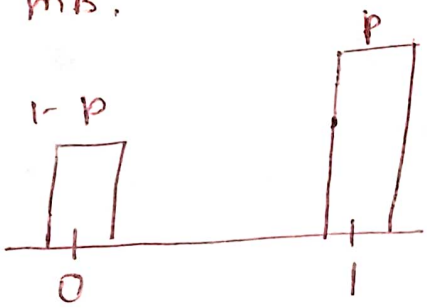
  Now we are estimating that std dev of population = std dev of sample.

- from population, we can calculate std dev for sampling distribution.

$$= \frac{\sigma}{\sqrt{n}}$$

⇒ Summary so far.

- East Bengal and Mohan Bagan.
- For whole of Bengal, p percent of population supports EB whereas (1-p)% of population support MB.



For this population, $\mu = p$, and Std dev.

$$\sigma = p(1-p)$$

- We want to find out the value of p.
- But since the population Bengal is very high, it is impossible to calculate p.
- hence we will estimate the value of p through sampling.

- Let's suppose that I took a sample of 100 Bengali's,

  57 → EB → 0

  43 → mB → 1

For this sample,

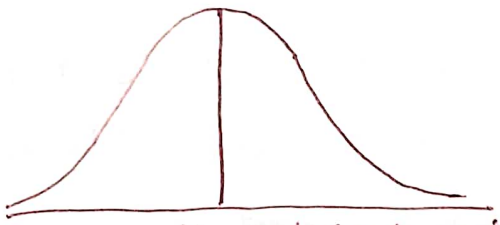
$$\bar{x} = \frac{57(0) + 43(1)}{100} = 0.43 \quad \text{(sample mean)}$$

$$s^2 = \frac{57(1-0.43)^2 + 43(1-0.43)^2}{(100-1)}$$

$$s^2 = 0.2475$$

$$s = 0.50 \quad \text{(sample std dev)}$$

- Sampling dist. of sample mean



$\mu_{\bar{x}} = p$ (population mean)

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

$$= \frac{\sigma}{10} \quad (n = 100)$$

- The sample mean ($\bar{x}$) will also be from this distribution.
- Also since it is a normal dist, $\bar{x}$ will lie within 2 std dev from p, with prob of 0.95.
- We can also say that p will be within 2 std dev of $\bar{x}$, with a prob of 0.95.

But we don't know the std dev of sampling dist.

$$p \approx (\bar{x} - 2 \times \sigma_{\bar{x}}, \bar{x} + 2 \times \sigma_{\bar{x}})$$ 95% chance.

- We can calculate sample std dev.

from that we can estimate population std dev = sample std dev. = 0.50

- Now, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{100}} = \frac{0.5}{10} = 0.05$

So, now we have an estimated value of $\sigma_{\bar{x}}$.

We can now say that

$p \approx (\bar{x} - 2 \times \sigma_{\bar{x}}, \bar{x} + 2 \times \sigma_{\bar{x}})$ with a 95% chance

$p \approx (0.43 - 0.1, 0.43 + 0.1)$

$p \approx (0.33, 0.53)$ with 95% chance.

⇒ Condition for confidence intervals

1. Random sample
2. Normal condition ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$)
3. Independence condition. sample is less than 10% of population.

Generally

$$p \approx \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

z^* critical value (dependent on what confidance we want)

for 95%, $z^* = 2 \approx 1.96$

22-12-2020

T distributions

- So far we are trying to estimate a population parameter by using a sample parameter.

% we are concerned about mean, then,

$\mu \in \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ → σ: st. dev. of population

$\mu \in \bar{x} \pm z^* \frac{s}{\sqrt{n}}$ → σ is estimated by s

- But it turns out that there is a better representation for this and this underfits the true values (when we are using s) not σ.

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

(where t comes from t distribution)

⇒ Condition for Inference on a mean :-

i) Random ii) Normal iii) Independent.

Random :-

Selection of sample should be random

Normal :-

Sampling dist should be normal

either $n \geq 30$, or original population

Independent :-

Independency is assumed if sample contains $\leq 10\%$ of population.

Note :

While looking for value of t for some confidence, we also have to find degree of freedom (df)

$$df = n - 1$$

- If we know σ, (population std dev), we will use

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$$

If we only know s (sample std dev), we will use,

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

⇒ Making t interval for paired data

- In paired data we have two observations on the same individual.

For eg. student's pre test & post test score,

- Example for making t interval

Let's say we have 2 watches A and B and we want that are used to record the distances travelled.

- We want to find out whether there is a plausible difference in the distances recorded in both watches.

We take a sample of 5 people to run 10 km wearing both the watches and calculate the readings.

	1	2	3	4	5
A					
B					

Steps:

1. Find the difference:
When dealing with paired data, we're interested in dist. of differences

	1	2	3	4	5
A					
B					
(B-A)					

2. Check conditions :

- Random
- Normal
- Independent

3. Construct the interval

Find mean and std dev of difference $\bar{x}_{diff}$ and s_{diff}.

and find the interval. by :

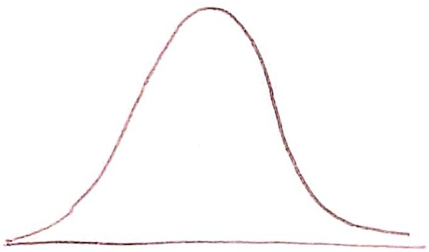
$$\bar{x}_{diff} \pm t^* \frac{s_{diff}}{\sqrt{n}}$$

Interval $\approx$ (lower, higher)

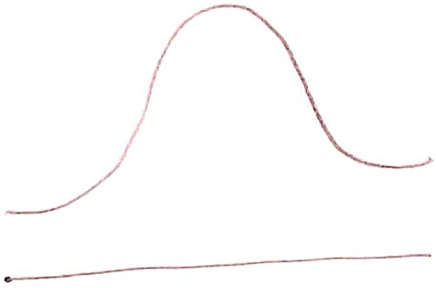
4. Interpret the interval

check whether 0 is contained in the interval.

Normal dist.



T distribution



Sampling dist. is normal when sample size is large enough.

When sample size is small, it is t distribution.