DAY 02 INTERVIEW QUESTIONS - Confidence Interval, level of significance, p-value

Q1) What is Confidence Interval?

Ans - Confidence Interval tells how much confident we can be the results from a poll or a survey reflect what we would expect to find if it were possible to survey entire population.

Confidence Interval can be intrepeted as Confidence levels. For example, 95% confidence level. It means if we repeat an experiment/survey over and over again, 95 percent of the time our result will match the result we will get from the population.

Example → US Bureau routinely uses confidence levels of 90% in survey. One survey of the number of people in poverty in 1995 stated confidence level of 90%, "Number of people in poverty in United State is 35 lacs to 37 lacs". Means if they repeated the survey using same technique, 90 percent of the time result will fall between 35 lacs to 37 lacs in poverty.

Q2) So confidence Interval give data width with Confidence levels? So what are the factors affect width of Confidence intervals.

Ans - i) Variation and ii) Sample size

i) Variation → Variation within population of interest.
- If all the values are in the sample are same/almost similar, then the data have low variation.
- Population with low variation leads to similar samples with low variation leads to narrow confidence level.
- Population with lots of variation leads to varied samples with high variation leads to wider confidence interval.

ii) Sample size → Sample size also affects the width of confidence interval.
- If we take small sample size, more samples will vary from each other, there will be more variation due to sampling or sampling error.
- Large samples are more similar to each other and have more information, which leads to narrow width of confidence intervals.

Q3) Any formula for Confidence Interval?

Confidence Interval, $CI = \bar{x} \pm t \frac{s}{\sqrt{n}}$

$\bar{x}$ → Sample mean, t → t distribution
$\sqrt{n}$ → square root of sample size
s → sample standard deviation
$s/\sqrt{n}$ → standard error
$\pm$ → + will give higher range, - lower range

Q4) What is Significance level (Alpha)?

- Significance level, also denoted as alpha or α, is the probability of rejecting the NULL HYPOTHESIS when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding of rejecting the null hypothesis.
- Significance levels shows how likely a pattern in data is due to chances. Mostly, no statistical package shows confidence Interval (suppose 0.95 means 95% chance of being true). Instead it will show "0.05", meaning that finding that five percent (0.05) chances of not being true, which is converse of a 95% of being true.

Example – Do you buy Bajaj Fixed Deposit?

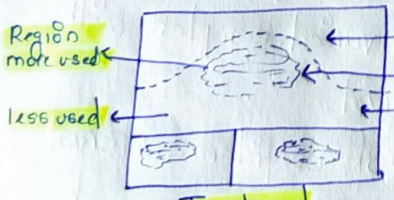
	AREA			Type of Vehicle			
	Total	State	City	Car	Bike	Bus	Van
Base	713	361	352	247	150	44	150
Yes	428 (60%)	215 (60%)	213 (60%)	131 (53%)	74 (49%)	29 (66%)	131 (73%)
No	285 (40%)	146 (40%)	139 (40%)	116 (47%)	76 (51%)	15 (34%)	49 (27%)
Chi Square	.910			.001			

There is no difference (probably no difference) in purchases of FD in city and state (as city is part of states mostly, other than Union territories) because probability is .91 (i.e, 9% chances difference is true).

In contrast, high significance level of type of vehicle (.001 or 99%) indicates there is almost certainly a true difference in purchase of FD by vehicle owners.

Q5) What is Pvalue?

Ans – P value is the "probability" for the "NULL HYPOTHESIS" to be "True".



Normalbell shaped curve

p value = 0.8, this means, region is touched more 80 times per 100 times.

Suppose here pvalue = 0.01 this means, this region is touched 1 times per 100 times.

Example, Create 2 groups of 4 plants each: Group 1 → 4 plants → Fertilizers A
Group 2 → 4 plants → Fertilizers B

Null hypothesis there is no difference in effect of fertilizers A and fertilizers B.

So in this we will apply t test (comparison of mean).

Pvalue can take value range from 0 to 1. p = 0.01, suggest 01 out of 100 times, 1 time null hypothesis is true. p = 0.10, suggest 10 out of 100 times, 10 times NULL is true.

By t test, p = 0.03, then 3 out of 100 times, null is true which suggest there will be no difference in effects of fertilizers A and B.

So, what is optimal P-value? Ans is level of significance.