The Central Limit Theorem (CLT)

- CLT is a fundamental principle in probability theory and statistics. It states that the sampling distribution of the mean of any independent, random variable will be approximately normally distributed, regardless of the shape of the original population distribution, as long as the sample size is large enough.

  $Central\ Limit\ Theorem$: $Let\ x_1, x_2, x_3, \ldots .\ be\ \ r.v\ which\ are\ idendently\ and\ identically\ distributed\ (iid)\ with\ mean\ \mu\ and\ variance\ \sigma^2.$

  $If\ \overline{x_n} = (x_1 + x_2 + x_3 + \cdots \ldots + x_n)/n$
  Where, 'n' is the number of samples

  CLT says:
  $Z_n = \frac{(\overline{x_n} - \mu)}{\sigma/\sqrt{n}}$, follows Standard Normal Distribution

- In simpler terms, it means that if you take multiple random samples from any population (even if that population is not normally distributed), and calculate the mean of each sample, the distribution of those sample means will be approximately normal.

Key Points:

1. **Large Sample Size:**
   - The CLT works best when the sample size is large (usually, at least 30 observations). As the sample size increases, the sample mean distribution becomes closer to a normal distribution.
2. **Independence:**
   - The samples must be independent, meaning that the selection of one individual does not affect the selection of another.
3. **Approximation:**
   - The larger the sample size, the better the approximation to a normal distribution. For small sample sizes, the approximation might not be very accurate.
4. **Universality:**
   - The CLT applies to many different types of random variables. It's a powerful tool in statistics because it allows us to make inferences about a population based on a sample. CLT is incredibly useful for making statistical inferences.

PROBLEMS

1) Suppose you have a population with a highly skewed distribution of incomes. The mean income is $50,000 with a standard deviation of $20,000. You take a random sample of 100 individuals from this population. What is the probability that the sample mean income will be between $48,000 and $52,000?

Ans:
CLT says:
$Z_n = \frac{(\overline{x_n} - \mu)}{\sigma/\sqrt{n}}$, follows Standard Normal Distribution

Given,

$$Population\ Mean\ (\mu) = \$50000$$
$$Population\ s.d\ (\sigma) = \$20000$$
$$Sample\ Size\ (n) = 100$$

When the Sample mean = $ 52000

$$Z_1 = \frac{52000 - 50000}{\frac{20000}{10}} = \frac{2000\ X\ 10}{20000} = 1$$

When, the Sample Mean = $ 48000
$Z_2 = (48000 - 50000)/2000$ = - 2000/2000= -1

So, the Z-scores will have following range:

$$-1 \leq Z \leq 1$$

From Standard Normal Distribution Table:

P(z < -1) ≈ 0.1587 and P(z < 1) ≈ 0.8413

P (48000< Z < 52000) = 0.8413 – 0.1587 = 0.6826

Ans. The probability that the sample mean income will be between $48,000 and $52,000 is 0.6826

2) You're conducting a study on the average height of students in a school. The population has a mean height of 65 inches with a standard deviation of 3 inches. You take a sample of 36 students. What is the probability that the sample mean height is greater than 64 inches?

Ans.
Given

$$Population\ Mean\ (\mu) = 65\ inches$$

$$Population\ s.d\ (\sigma) = 3\ inches$$
$$Sample\ Size\ (n) = 36$$

P (Z > 64) =?

When the Sample Mean = 64

$$Z_1 = \frac{64 - 65}{\frac{3}{6}} = -2$$

From Standard Normal Distribution Table:

P(z < -2) ≈ 0.0228

So,

P (x > 64)= 1 – P(x< 64) = 1 – P(z < -2)=1 – 0.0228 = 0.9772

Ans: The probability that the sample mean height is greater than 64 inches is 0.9772

3) Suppose you are conducting a survey to estimate the average number of hours students at a university spend on extracurricular activities per week. You collect data from a random sample of 50 students and find that the sample mean is 6 hours with a sample standard deviation of 1.5 hours. Calculate a 95% confidence interval for the true population mean.

Ans:

Given

$$Sample\ Mean\ (\overline{x_n}) = 6\ hours$$
$$Sample\ s.d\ (s) = 1.5\ hours$$
$$Sample\ Size\ (n) = 50$$

$$Value\ of\ \mu\ at\ 95\%\ Confidence\ interval?$$

Confidence Level (1 - α): 95% or α = (1-0.95) = 0.05

From Standard Normal Distribution Table, we can calculate the critical z-value is (i.e when α/2= 0.05/2 = 0.025) is approximately 1.96.

The margin of error (E) is calculated using the formula:

E = z * (s/√n) = $1.96 * (1.5/\sqrt{50}$ = 0.415

Confidence Interval = (x̄ - E, x̄ + E) = (6-0.415; 6+0.415) = (5.585, 6.415)

Ans. At 95% confidence interval the true population mean will lie in the range (5.585, 6.415)

Try at home:

4) You are studying the time taken by customers to complete a specific task on a website. From past data, you know that the population standard deviation is 4 minutes. You take a random sample of 25 customers and find that the sample mean time is 22 minutes. Can you determine whether this sample mean is significantly different from the population mean of 20 minutes at a 5% level of significance?