Frequently Asked Questions

1. State Central Limit Theorem?

   Answer:
   Suppose X1 ,X2,...Xn, be n independent random variables having the same probability density function each with E(Xi)=μ and V(Xi) = σ^2, for i=1,2 ,...,n then Sn = X1+X2+....+Xn is approximately Normal with mean nμ and variance nσ^2. Also

   $Z = \frac{\overline{X} - n\mu}{\sqrt{n}\sigma}$ is asymptotically N(0,1)

2. Why do we need the sample mean to be normally distributed?

   Answer:
   The sample mean is preferred to be Normally distributed because we want to use Z scores to analyze sample means. But to use Z scores, the data must be normally distributed.

3. What are the general assumptions of Central Limit Theorem

   Answer:
   Assumptions of CLT

   - The variables are independent
   - All the variables follow same distribution
   - The mean and variance of all the variables exists
   - The mean and variance of all the variables are same
   - Number of variables ate very large (that is n tends to infinity)

4. State Normal Theorem

   Answer:
   If the R.Vs X1,X2,...,Xn are independent and identically distributed as N(μ,σ^2) then

   $\bar{x} = \frac{\sum xi}{n}$ is distributed as N(μ,σ^2/n)

5. Suppose X1 ,X2,...Xn, be n independent random variables having the same probability density function each with E(Xi)=μ and V(Xi) = σ^2, for i=1,2 ,...,n then prove that Sn = X1+X2+....+Xn is approximately Normal with mean nμ and variance nσ^2. Also prove that

   $Z = \frac{\overline{X} - n\mu}{\sqrt{n}\sigma}$ is asymptotically N (0, 1?)

Answer:

Given that Sn= X1+X2+------+Xn

$$E(\frac{Sn}{n}) = E(\frac{X1 + X2 + ... + Xn}{n})$$

$$= \frac{1}{n}[E(X_1) + E(X_2) + ... + E(Xn)]$$

$$= \frac{1}{n}[\mu + \mu + ... + \mu] = \frac{n\mu}{n} = \mu$$

$$V(\frac{Sn}{n}) = V(\frac{X1 + X2 + ... + Xn}{n})$$

$$= \frac{1}{n^2}[V(X_1) + V(X_2) + ... + V(Xn)]$$

$$= \frac{1}{n}[\sigma^2 + \sigma^2 + ... + \sigma^2] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Then $Z = \dfrac{\frac{Sn}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$M_Z(t) = M_{\frac{\frac{Sn}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}}(t) = e^{\frac{-t\mu\sqrt{n}}{\sigma}} M_{\frac{Sn}{n}}(\frac{t\sqrt{n}}{\sigma})$

$$M_Z(t) = e^{\frac{-t\mu\sqrt{n}}{\sigma}}[M_X(\frac{\sqrt{n}t}{\sigma\sqrt{n}})]^n$$

Since the M.G.F of the sum of n independent variables is the product of the M.G.F's of the variables. Also M.G.F's of all variables are the same

$$M_Z(t) = e^{\frac{-t\mu\sqrt{n}}{\sigma}}[M_X(\frac{t}{\sigma\sqrt{n}})]^n$$

Taking Logarithm on both the sides

$$\text{Log } M_Z(t) = \frac{-t\mu\sqrt{n}}{\sigma} + n\log M_X(\frac{t}{\sigma\sqrt{n}})$$

Expanding the logarithmic function

Log (1+y) = y- $y^2/2+y^3/3-\ldots\ldots$

We get

$$\lim_{n\to\infty} \log M_Z(t) = \frac{t^2}{2}$$

$\therefore M_Z(t) = e^{\frac{t^2}{2}}$ is the M.G.F of Standard Normal Distribution

Therefore Z~ N(0,1) as n tends to infinity

Therefore $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is asymptotically Normally distributed with mean 0 and S.D 1

6. Explain the importance of CLT.

Answer:

The CLT plays an important role in Statistical Theory. It is one of the usual assumptions in Statistics to assume that the underlying observations to follow Normal distribution at least approximately. The theory of error used by physisits, or astromers can be justified on the basis of CLT. Thus investigating the most general set of conditions under which the CLT can hold is of theoretical and practical interest.

The significance of the central limit theorem lies in the fact that it permits us to use sample estimators to make inference about the population parameters without knowing

anything about the shape of the frequency distribution of that population other than what we can get from the sample.

The practical utility of the CLT is inherent in its approximation. The important contribution of CLT is in Statistical Inference. Many estimations that are used to make inferences about population parameters are sums or averages of the sample measurements. When the sample size is sufficiently large, you can expect these estimators to have sampling distribution that are approximately Normal. We can then use the Normal distribution to describe the behaviour of these estimators in repeated sampling and evaluate the probability of observing certain sample results. These probabilities are calculated using the Standard Normal R.V

Z= Estimator – mean /S.D

7. State the corollary for CLT.

Answer:

If X1, X2, Xn, are independent identically distributed R.Vs with common mean μ and variance σ^2 then

$$P\left[a < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < b \right] = \phi(b) - \phi(a)$$

This equation gives an approximate value of the probabilities of certain events described in terms of averages and sums.

8. What are the Similarities and dissimilarities between CLT and Normal Theorem?

Answer:

Similarities and dissimilarities between CLT and Normal Theorem are

Normal Theorem	CLT
Population Mean -=μ S.D =σ	Population Mean -=μ S.D =σ
Shape of the population is known to be a N(μ,σ^2) curve	Shape of the population is either unknown or not Normal
Sample average $\bar{x}$ is said to have N(μ,σ^2/n) for any n	Sample average $\bar{x}$ is said to have N(μ, σ^2/n) Approximately for only large n

9. How do you check for Normality?

Answer:
CHECK FOR NORMALITY:

- Use descriptive statistics. Construct stem-and-leaf plots for small or moderate-sized data sets and frequency distributions and histograms for large data sets

- Compute measures of central tendency (mean and median) and compare with the theoretical and practical properties of the normal distribution. Compute the inter quartile range. Does it approximate the 1.33 times the standard deviation?

- How are the observations in the data set distributed? Do approximately two thirds of the observations lie between the mean and plus or minus 1 standard deviation? Do approximately four-fifths of the observations lie between the mean and plus or minus 1.28 standard deviations? Do approximately 19 out of every 20 observations lie between the mean and plus or minus 2 standard deviations?

10. How do you decide when the sample size is large enough?

Answer:

- If the sampled population is normal, then the sampling distribution of x bar will also be Normal. No matter what sample size you choose. The result can be proven theoretically but it should not be too difficult for you to accept without proof
- When the sampled population is approximately systematic the sampling distribution of x bar becomes approximately Normal for relatively small values of n.
- When the sampled population is skewed the sample size n must be larger, with n at least 30, before the sampling distribution of X bar becomes approximately Normal.
- A sample size of 100 or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed, $n > 30$ may be sufficient to use the CLT.

11. How do you interpret CLT?

Answer:
If random samples of n observations are drawn from a Nonnormal population with finite mean μ and S.D σ, then when n is large the sampling distribution of the sample mean x bar is approximately Normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$

The approximation becomes more accurate as n becomes large

Regardless of its shape the sampling distribution of x bar always has a mean identical to the mean of the sampled population and a standard deviation equal to the population

S.D $\sigma/\sqrt{n}$. Consequently the spread of the distribution of the sample means is considerably less than a spread of the sampled population

12. Narrate with an example that the shape of the population varies according to the size of the sample

Answer:
The CLT states that under rather general conditions sums and means of random samples of measurements drawn from a population tend to have an approximately Normal distribution. Suppose you toss a balanced die n=1 time. The RV X is the number observed on the upper face. This familiar RV can take six values, each with probability 1/6 and its probability distribution. The shape of the distribution is flat or Uniform and systematic about the mean $\mu=3.5$ with a SD $\sigma=1.71$

Now take a sample of size n=2, from this population that is toss 2 dice and record the sums of numbers on the two upper faces $\sum xi=x1+x2$. There are 36 possible outcomes each with probability 1/36. The sampling distribution of x bar there is dramatic difference in the shape of the sampling distribution. It is now roughly mound shaped but still symmetric about the mean $\mu=3.5$

Similarly For n=3 and n=4 the sampling distribution clearly shows the mound shape of the Normal Probability distribution still centred at $\mu=3.5$. Notice also that the spread of the distribution is slowly decreasing as the sample size n increases. The distribution of x bar is approximately Normally distributed based on a sample as small as n=4. This phenomenon is the result of an important statistical theorem called the CLT.

13. An insurance company has 10,000 automobile policy holders. If the expected yearly claim per policy holder is $260 with an Standard deviation of $800 , approximate the probability that the total yearly claim exceeded $2.8 million

Answer:
Let Xi denote the yearly claim of policy holder i, i=1, 2... 10,000

By CLT $X= \sum Xi$ will have approximately Normal distribution with mean $10,000*260 = 2.6 * 10^6$ and Standard deviation $800 \sqrt{10^4} = 8*10^4$

Hence $P[X>2.8 * 10^{6]} = P [(X-2.6 * 10^6)/ 8*10^4 > (2.8 * 10^6 -2.6 * 10^6)/ 8*10^4$

$=P[Z>20/8]= P[Z>2.5]=0.0062$

That is there are only 6 chances out of 1000 that the total yearly claim will exceed $2.8 million

14. A random sample of size n was taken from a population with mean μ and Standard deviation σ. Find the distribution of sample mean $\bar{x}$, for large n

Answer:

Suppose X1, X2 ...Xn, be n independent random sample drawn from the population.

Given E (Xi) =μ and V (Xi) = σ^2, for i=1, 2··· n

Let X = X1+X2+....+Xn

$$E(X) = E(X1 + X2 + ... + Xn)$$

$$= \sum E(Xi) = \sum \mu = n\mu$$

$$V(X) = V(X1 + X2 + ... + Xn)$$

$$= \sum V(X_i) = n\sigma^2$$

Then $Z = \frac{X - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$

But $\frac{X}{n} = \frac{\sum x_i}{n} = \bar{x}$

$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ asymptotically

Therefore $\bar{x}$ follows Normal distribution with mean μ and variance σ/√n for large values of n

15. In a game involving repeated throws of a balanced die, a person receives Rs. 3 if the receiving number is greater than or equal to 3 and looses Rs. 3 otherwise. use CLT to find the probability that in 25 trials his total earnings exceeds Rs. 25

Answer:

Let Xi be the earnings in the ith game. Then,

X	Probability	Xp	X^2 p
3	2/3	2	6
3	1/3	-1	3

E (Xi) = ∑Xi p =1

E (Xi^2) = ∑Xi^2 p =9

V (Xi) = E (Xi^2)-[E (Xi)]^2 =8

Let X=X1+X2+...+X25

Then E(X) = n E(Xi) = 25X1= 25

V(X) = V(∑Xi) = ∑V(Xi)= 25 X 8= 120

By the CLT

$Z = \frac{X - 25}{\sqrt{120}} \sim N(0,1)$ asymptotically

Consider $P(X > 25) = P[\frac{X - 25}{\sqrt{120}} > 0] = P[Z > 0] = 0.5$

From the table of standard Normal probabilities

Therefore the probability that the total earnings exceeds Rs.25 is equal to 0.5