Session 12: Central Limit Theorem and Confidence Interval

Agenda

Sl No.	Agenda Topics
1.	Central Limit Theorem
2.	R Code For Understanding Central Limit Theorem
3.	How does Central Limit Theorem work
4.	Mechanism
5.	Plotting Now
6.	CLT Facts
7.	Practical Application 1 Of CLT
8.	Practical Application 2 of CLT
9.	Confidence Interval & Probability
10.	(Mis)interpreting The Confidence

Sl No.	Agenda Topics
11.	How To Interpret Confidence Interval
12.	P-value , Z-score
13.	Confidence Intervals For Unknown Mean And Known Standard Deviation
14.	t-Distribution
15.	t-Distribution Using R
16.	Poison Distribution
17.	Simulation Of Poisson Random Variables
18.	Question And Answer on Poison Distribution
19.	Exercises

Central Limit Theorem

- Given a distribution with a mean μ and variance σ2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ2/n as n, the sample size, increases
- The amazing and counter-intuitive thing about the central limit theorem is that no matter what the shape of the original (parent) distribution, the sampling distribution of the mean approaches a normal distribution
- Three different components of the central limit theorem
  - (1) successive sampling from a population
  - (2) increasing sample size
  - (3) population distribution
- Remember that this theorem applies only to the mean and not to other statistics

R Code For Understanding Central Limit Theorem

- Preliminaries in terms of reading the dataset

```
mba.data<-read.csv("mba.csv",header=T)
attach(mba.data)
library(car)
library(moments)
```

How Does Central Limit Theorem Work?

- A graph depicting that the distribution of sample means tends towards normality

- Multiple graphs depicting how the histogram becomes thinner, as the sample size (n) increases

- Values of E(sample means) and (mean of the original variable=mu)

- Values of Var (sample means) and (standard error of the original variable = (sigma^2/n))

Mechanism

1) The histogram of actual values of the variable

2) A histogram of 1000 sample means, where each sample is of size sample size1; drawn from the variable of interest

3) A histogram of 1000 sample means, where each sample is of size sample size2; drawn from the variable of interest

4) Compare the values mentioned above

Mechanism (Contd.)

- Next three lines are your input

```
x = workex #Define the variable that you want to use here

sample.size1 = 64 #Define the smaller sample size here

sample.size2 = 256 #Define the larger sample size here
```

- The main code

```
par(mfrow=c(3,1)) #This adjusts the chart area to fit 3 graphs in a
3rows X 1column formation

hist(x,prob=TRUE,col='dodgerblue4') #Histogram of the original
variable
```

Mechanism (Contd.)

- sample1.mean <- replicate(mean(sample(x,size=sample.size1,replace=TRUE)),n=10000) #This calculates the mean of a random sample of size sample.size1, drawn from the variable x. And the 'replicate' command replicates this and calculates means for n (10000 here) such samples
- sample2.mean <- replicate(mean(sample(x,size=sample.size2,replace=TRUE)),n=10000) #This calculates the mean of a random sample of size sample.size2, drawn from the variable x. And the 'replicate' command replicates this and calculates means for n (10000 here) such samples
- hist(sample1.mean,prob=TRUE,ylim=c(0,0.2),col="dodgerblue4",xlim=c(min(min(sample1.mean),min(sample2.mean)),max(max(sample1.mean),max(sample2.mean)))) #Histogram of the means from the first sample of sample.size1. The limits on the X-axis are set in this manner to ensure that the histograms of means from two sample sizes can be plotted on the same range. This will make them comparable.
- mean1.clt = mean(sample1.mean) #Mean of sample means from sample.size2

Mechanism (contd.)

- sd1.clt = sd(sample1.mean) #Standard error of the sample means from sample.size1
- curve(dnorm(x,mean1.clt,sd1.clt),col="red",lwd=5,add=TRUE,ylim=c(0,1)) #Gives the CLT approximation to the normal density with mean=mean(variable of interest) and sd=standard error of the variable of interest
- hist(sample2.mean,prob=TRUE,ylim=c(0,0.4),col="dodgerblue4",xlim=c(min(min(sample1.mean),min(sample2.mean)),max(max(sample1.mean),max(sample2.mean)))) #Histogram of the means from the first sample of sample.size1.
- mean2.clt = mean(sample2.mean) #Mean of sample means from sample.size1
- sd2.clt = sd(sample2.mean) #Standard error of the sample means from sample.size2

Plotting Now

- curve(dnorm(x,mean2.clt,sd2.clt),col="red",lwd=5,add=TRUE,ylim=c(0,1)) #Gives the CLT approximation to the normal density with mean=mean(variable of interest) and sd=standard error of the variable of interest
- par(mfrow=c(1,1))
- qqPlot(sample1.mean) #Normal quantile plot, to check if the sample mean is normally distributed
- qqPlot(sample2.mean) #Normal quantile plot, to check if the sample mean is normally distributed

CLT Facts

- If you draw samples from a normal distribution, then the distribution of sample means is also normal
- The mean of the distribution of sample means is identical to the mean of the "parent population“, the population from which the samples are drawn
- The higher the sample size that is drawn, the "narrower" will be the spread of the distribution of sample means

Practical Application 1 Of CLT

- The mean salary of the 9,000 employees at Holley.com is $\mu = 26{,}000$ with a standard deviation of $\sigma = 2420$. A pollster samples 400 randomly selected employees and finds that the mean salary of the sample is 26 650. Is it likely that the pollster would get these results by chance, or does the discrepancy suggest that the pollster's results are fake?

Practical Application 1 of CLT (Contd.)

- The question deals with the mean of a group of 400 individuals, which is a case for the Central Limit Theorem. The theorem tells us that if we select many groups of 400 individuals and compute the mean of each group, the distribution of means will be close to normal with a mean of $\mu = 26,400$ and a standard deviation of = $\sigma/\sqrt{n} = 2400 / \sqrt{400} = 121$. Within the distribution of means, a mean salary of 26,650 has a z-score of

- Z = data value – mean / standard deviation = 26,650-26400 / 121 = 2.07

Practical Application 1 of CLT (Contd.)

- In other words, if we assume that the sample is randomly selected, its mean salary is more than 2 standard deviations above the mean salary of the entire company. According to the z-score chart, a z-score of 2.07 lies near the 98th percentile. Thus, the mean salary of this sample is greater than the mean salary we would find in 98% of the possible samples of 400 workers. That is, the likelihood of selecting a group of 400 workers with a mean salary above 26,650 is about 2 % or 0.02. The mean salary of the sample is surprisingly high; perhaps the survey was flawed.

Practical Application 2 of CLT

- Engineers must consider the breadths of male heads when designing motorcycle helmets for men. Men have head breadths that are normally distributed with a mean of 6.0 inches and a standard deviation of 1.0 inch

a. If one male is randomly selected, what is the likelihood that his head breadth is less than 6.2 inches?

- z = 6.2-6/ 1
- z = .2
- 7.93 is the area between z and mean.

Practical Application 2 of CLT (Contd.)

- We add 50 to this number because we know where it lies on the chart.

The answer: 57.93 % have head sizes less than 6.2

b. The Safeguard Helmet company plans an initial production run of 100 helmets. How likely is it that 100 randomly selected men have a mean head breath of less than 6.2 inches?

1/ sq. root of 100 = st. error

0.1 = st. error

6.2 – 6.0 / .1

z-score = 2

- Area between mean and z is 47.72. We add 50%. 97. 72 % of the sample means will be less than 6.2

Practical Application 2 of CLT (Contd.)

- The production manager sees the result in part b and reasons that all helmets should be made for men with head breadths of less than 6.2 inches, because they would fit all but a few men. What is wrong with that reasoning?

Practical Application 2 of CLT (Contd.)

- The answer in part A tells us that the reasoning in part b is wrong.

Confidence Interval & Probability

- A confidence interval is expressed in terms of a range of values and a probability (e.g. my lectures are between 60 and 70 minutes long 95% of the time)
- For this example, the confidence level that is used is the 95% level, which is the most commonly used confidence level
- Other commonly selected confidence levels are 90% and 99%, and the choice of confidence level to be used when constructing an interval often depends on the application

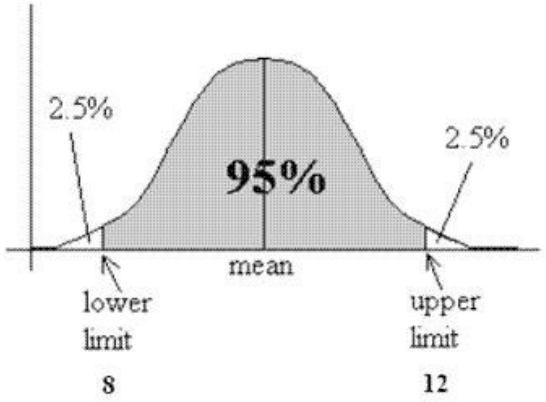
(Mis)interpreting The Confidence Interval

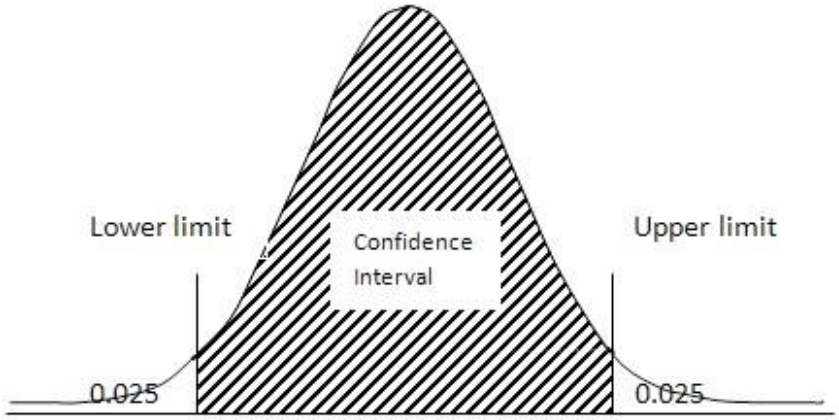
- 90% Confidence Interval for the mean score of the students is [75-80]

- Does this mean that
  - 90% of the students have a score in this range?
  - The mean score of the class lies in this range?
    – The mean score is in this range, 90% of the time?

How To Interpret Confidence Interval

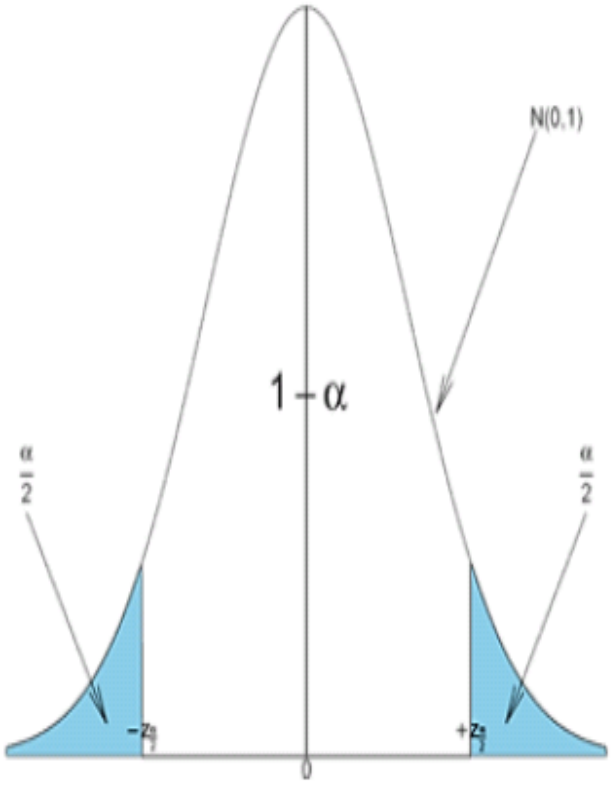
- A 95% confidence interval is a range of values that you can be 95% certain and contains the true mean of the population. This is not the same as a range that contains 95% of the values.





P-value , Z-score

- The central region on this graph is the acceptance area and the tail is the rejection region, or regions. In this particular graph of a two tailed test, the rejection region is shaded blue. The tail is referred to as “alpha“, or p-value (probability value). The area in the tail can be described with z-scores. For example, if the area of the tails was 5% (2.5% each side), the z-score would be 1.96 (from the z-table), which represents 1.96 standard deviations from the mean. The null hypothesis will be rejected if z is less than -1.96 or greater than 1.96.



Confidence Intervals For Unknown Mean And Known Standard Deviation

- For a population with unknown mean(μ) and known standard deviation(σ), a confidence interval for the population mean, based on a simple random sample (SRS) of size n is:

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}},$$

where z* is the upper (1-C)/2 critical value for the standard normal distribution.

- An increase in sample size will decrease the length of the confidence interval without reducing the level of confidence. This is because the standard deviation decreases as *n* increases.
- As the level of confidence decreases, the size of the corresponding interval will decrease.

Confidence Intervals for Unknown Mean and Unknown Standard Deviation (Contd.)

- Mostly, the standard deviation for the population of interest is not known. In this case, the standard deviation σ is replaced by the estimated standard deviation s, also known as the standard error
- Since the standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean is no longer normal with mean μ and standard deviation σ/√n. Instead, the sample mean follows the t distribution with mean and standard deviation. Instead, the sample mean follows the ***t distribution*** with mean μ, and standard deviation s/√n

t-Distribution

- The use of a t-distribution is precluded by the standard deviation of the population parameter being unknown and allows the analyst to approximate probabilities, based on the mean of the sample, the population, the standard deviation of the sample and the sample's degrees of freedom.
- The t-distribution is also described by its degrees of freedom. For a sample of size n, the t distribution will have n-1 degrees of freedom.
- As the sample size n increases, the t-distribution becomes closer to the normal distribution, since the standard error σ approaches the true standard deviation for large n.
- For a population with unknown mean μ and unknown standard deviation, a confidence interval for the population mean, based on a simple random sample (SRS) of size n, is $\overline{x} \pm t^* \frac{s}{\sqrt{n}}$, , where t* is the upper (1-C)/2 critical value for the t-distribution with n-1 degrees of freedom, t(n-1).

t-Distribution Using R

- There are four functions that can be used to generate the values associated with the t-distribution. A full list of them and their options can be obtained using the help command:

  >help(Tdist)

The four functions are :

- Distribution function
- Cumulative probability distribution function
- Inverse cumulative probability distribution function
- Random numbers

Poison Distribution

- Suppose an event has a small probability of occurring and a large number of independent trials take place. Suppose further that you know the average number of occurrences μ over a period of time. Then the Poisson random variable, denoted X ~ Poi(μ), counts the total number of occurrences during a given time period.

- The **probability** of having exactly k occurrences during this time, for k >= 0, is given by:

  P(X = k) = e ^(- μ) * μ^k / k! .

- The **average** (or **mean**) number of occurrences is given by E[X] = μ.

- The **variance** of the number of occurrences is given by Var(X) = μ.

Simulation Of Poisson Random Variables

- rpois() function can be used to simulate N independent Poisson random variables
- we can generate 12 Poisson random numbers with parameter $\lambda = 3$ as follows:

```
> rpois(12,3)
 [1] 6 4 0 3 2 3 3 7 4 5 6 2
```

Simulation Of Poisson Random Variables (Contd.)

- The mean and variance are equal for Poisson random variables.
- The standard deviation is $\sqrt{\lambda}$.
- For the Poisson random variable with $\lambda = 2.25$, we can calculate these quantities by using R:

```
> lambda <- 2.25
> lambda # Expected Value, and Variance
[1] 2.25
> sqrt(lambda) # Standard Deviation
[1] 1.5
```

Simulation Of Poisson Random Variables (Contd.)

- We can compare these with what we would obtain from a simulated sample of 10000 binomial random variables

```
> P <- rpois(10000, 2.25)
> mean(P) # sample mean
[1] 2.246
> var(P) # sample variance
[1] 2.220906
> sd(P) # sample standard deviation
[1] 1.49027
```

Question on Poison Distribution

- On an average, if there are 8 cows crossing a road per minute, find the probability of having 12 or more cows crossing the road in a particular minute.

Answer

- In R we have function ppois for poison distribution
- This would be lower tail
- Therefore, for 12 or more cows to cross the road, we have upper tail of probability density function

```
> ppois(11,lambda=8)
[1] 0.888076
```

```
> ppois(11,lambda=8,lower=FALSE)
[1] 0.111924
```

Next Class: Hypothesis Testing

- Define Hypothesis Testing
- Null Hypothesis Vs Alternative Hypothesis
- Type I Vs Type II Errors
- P-value And Significance Level
- α Vs β
- Steps Of Hypothesis Testing
- Determine A P-value When Testing A Null Hypothesis
- Question And Answer
- Upper-Tailed, Lower-Tailed, Two-Tailed Tests
- Problems And Answer
- Types Of Hypothesis Testing

Contact Info:

- **Website** : http://www.datasciencenirvana.com/
- **LinkedIn** : https://www.linkedin.com/in/gautham111/
- **Email** : egautham@gmail.com