Workshop 08

Digvijay Jondhale

Problem Statement

The GitHub repository for the Tidy Tuesday project contains a dataset for each Tuesday from the past ~6 years; i.e., around 300 datasets. A random sample of 40 dataset names was taken from the full list and the number of characters in each title was found. These values are contained in the file, ws08-exercise-title_length.rds.

```
title_data <- readRDS("ws08-exercise-title_length.rds")

summary(title_data)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4.00    8.00   12.00   12.65   15.25   27.00 
```

```
length(title_data)
```

```
[1] 40
```

To check for normality conditions of clt , the following conditions should be satisfied :

- Random sampling
- Sample size (n>=30)
- Independence
- Plotting rhe distribution and checking its shape

Random sampling :

This condition is met , as it is already stated in the problem statement that the sample of 40 datasets was randomly sampled.

Sample size

The sample size for clt to apply should be minimum 30, as the sample size is 40 >30 , this condition has been met.

```
if (length(title_data) >= 30) {
  cat("Sample size is large enough (n >= 30) for CLT to apply.\n")
} else {
  cat("Sample size is not large enough (n < 30)for clt to apply .\n")
      }
```

```
Sample size is large enough (n >= 30) for CLT to apply.
```

Independence

For independence , we need to make sure that the sample size n = 40 should be less than 10% of the population size (300).

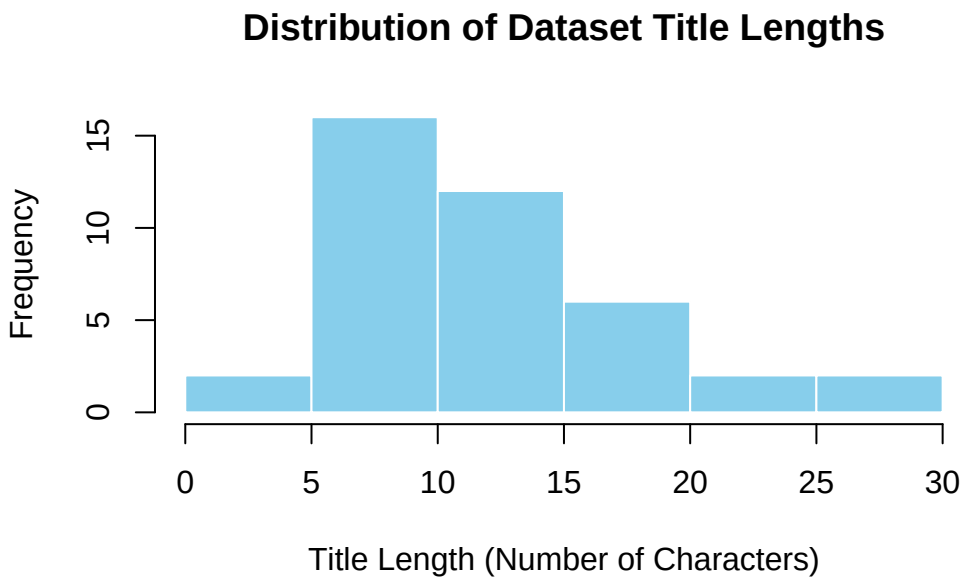
```
population <- 300
independence <- length(title_data) < (0.1 * population)
cat("Independence Condition Met:", independence, "\n")
```

```
Independence Condition Met: FALSE
```

Form the above results ,the conditions for independence are not met.

Checking distribution shape

```
hist(title_data, main = "Distribution of Dataset Title Lengths",
     xlab = "Title Length (Number of Characters)", col = "skyblue", border = "white")
```



From the histogram above , it is clear that the shape is unimodal, but it is right skewed and is not normally distributed. Hence CLT does not applies here as independence condition was not met.

Results can be verified by Shapiro-Wilk test

```
shapiro_test <- shapiro.test(title_data)
cat("Shapiro-Wilk p-value:", shapiro_test$p.value, "\n")
```

```
Shapiro-Wilk p-value: 0.02279463
```

```
if (shapiro_test$p.value < 0.05) {
  cat("Not Normally Distributed (p < 0.05).\n")
} else {
  cat("Approximately Normally Distributed (p >= 0.05).\n")
}
```

```
Not Normally Distributed (p < 0.05).
```

Results

The above test results confirm our analysis, that Central Limit theorem does not apply as the independence condition was violated, indicating that the sample means are **Not approximately Normally** distributed.

Note : *As the sample size = 40 was > 30 ,If we overlook the 10% rule of sample size then we can get a approximately Normal distribution of means using CLT*