# **TikTok Data Analysis: Sampling Distributions & Central Limit Theorem**

## **Lesson Plan for Web R Statistics**

### **Overview**

This lesson uses TikTok performance data to introduce students to sampling distributions and the Central Limit Theorem. By analyzing factors that contribute to TikTok video success, students will gain practical experience with statistical concepts while working with engaging, real-world data.

### **Learning Objectives**

By the end of this lesson, students will be able to:

1. Explain the concept of sampling distributions
2. Demonstrate the Central Limit Theorem using TikTok data
3. Calculate and interpret confidence intervals
4. Apply statistical methods to analyze factors affecting social media engagement
5. Use R to perform data analysis and visualization

### **Materials**

* Web R environment (accessible via browser)
* TikTok dataset CSV file (provided)
* R script with example code (provided)
* Student worksheet (optional)

### **Prerequisite Knowledge**

* Basic understanding of descriptive statistics (mean, median, standard deviation)
* Familiarity with basic R syntax
* Understanding of probability distributions

### **Time Required**

* 60-120 minutes depending on student statistics background and depth of discussion

## **Lesson Outline**

### **1. Introduction (10-15 minutes)**

* Engage students with a brief discussion about TikTok and what makes content go viral
* Present the research question: "What factors contribute to TikTok video success?"
* Explain that we'll use statistical methods to analyze real TikTok data
* Introduce the concept of sampling distributions as a foundation for statistical inference

### **2. Data Exploration (20-25 minutes)**

* Have students load the TikTok dataset into Web R
* Guide students through initial data exploration:
  + Examine structure and summary statistics
  + Create visualizations of key variables (views, engagement rate, duration)
  + Identify patterns in the data

#### **Sample Code for Initial Exploration:**

****# Load the data

tiktok\_data <- read.csv("tiktok\_data.csv", stringsAsFactors = TRUE)

# View structure and summary

str(tiktok\_data)

summary(tiktok\_data)

# Create visualizations

library(ggplot2)

ggplot(tiktok\_data, aes(x = Views)) +

geom\_histogram(bins = 20, fill = "skyblue", color = "black") +

labs(title = "Distribution of TikTok Video Views")

### **3. Sampling Distributions Concept (15-20 minutes)**

* Explain the concept of sampling distributions
* Discuss why we need sampling distributions in statistical inference
* Demonstrate how to create a sampling distribution using TikTok data:
  1. Take many random samples of a specific size from the data
  2. Calculate the mean of each sample
  3. Plot the distribution of these sample means

#### **Activity: Creating Sampling Distributions**

Have students:

1. Take 1000 random samples of size n=5 from the "Views" variable
2. Calculate the mean of each sample
3. Create a histogram of these sample means
4. Repeat with n=30
5. Compare the resulting distributions

### **4. Central Limit Theorem (20-25 minutes)**

* Introduce the Central Limit Theorem and its importance in statistics
* Key points to cover:
  + The sampling distribution of the mean approaches a normal distribution as sample size increases
  + The mean of the sampling distribution equals the population mean
  + The standard deviation of the sampling distribution decreases as sample size increases

#### **Demonstration:**

* Show how the sampling distributions created earlier demonstrate the CLT
* Calculate and compare the standard deviations of the original data and the sampling distributions
* Verify the relationship: σₓ = σ/√n

### **5. Application to TikTok Analysis (20-25 minutes)**

* Guide students in applying these concepts to analyze the TikTok data
* Activities:
  + Create confidence intervals for mean views by content category
  + Compare engagement rates across different time slots
  + Analyze factors associated with viral videos
  + Use bootstrap sampling to estimate parameters

#### **Sample Analysis Questions:**

1. Which content category has the highest average views?
2. Is there a significant difference in engagement rates between videos that use trending sounds and those that don't?
3. What time of day is associated with the highest average views?
4. Is there a relationship between video duration and number of views?

### **6. Student Challenge (Remaining time)**

Have students work individually or in small groups on one of the following challenges:

1. Create sampling distributions for other variables (likes, comments, shares)
2. Investigate how sample size affects the normality of sampling distributions
3. Create confidence intervals for the mean engagement rate of different categories
4. Develop a model to predict whether a TikTok video will go viral based on its characteristics

## **Assessment**

### **Formative Assessment**

* Participation in class discussions
* Completion of in-class activities
* Quality of data visualizations created

### **Summative Assessment**

Students submit a brief report (1-2 pages) that includes:

1. At least three visualizations created in R
2. An explanation of how the Central Limit Theorem applies to the TikTok data
3. Conclusions about what factors influence TikTok video success
4. A reflection on how sampling distributions help us understand variability in social media performance

## **Extensions and Differentiation**

### **For Advanced Students**

* Implement a multiple regression model to predict views
* Create interactive visualizations using Shiny
* Compare different sampling methods and their effects on the resulting sampling distributions

### **For Students Needing Support**

* Provide step-by-step instructions for R code
* Offer partially completed R scripts that students can modify
* Focus on core concepts with simplified data

## **Conclusion and Reflection**

End the lesson by having students reflect on:

* How statistical methods can provide insights into modern phenomena like social media
* How the Central Limit Theorem enables us to make inferences even when our data isn't normally distributed
* The relationship between sample size and the reliability of our statistical estimates

## **Additional Resources**

### **For Further Reading**

* "Statistical Inference" by Casella and Berger
* "R for Data Science" by Hadley Wickham
* Articles on TikTok analytics and social media statistics

### **Online Tools**

* RStudio Cloud for additional practice
* TikTok Creator Analytics documentation
* Kaggle datasets with social media data

### **Video Tutorials**

* Khan Academy videos on sampling distributions
* YouTube tutorials on data analysis with R