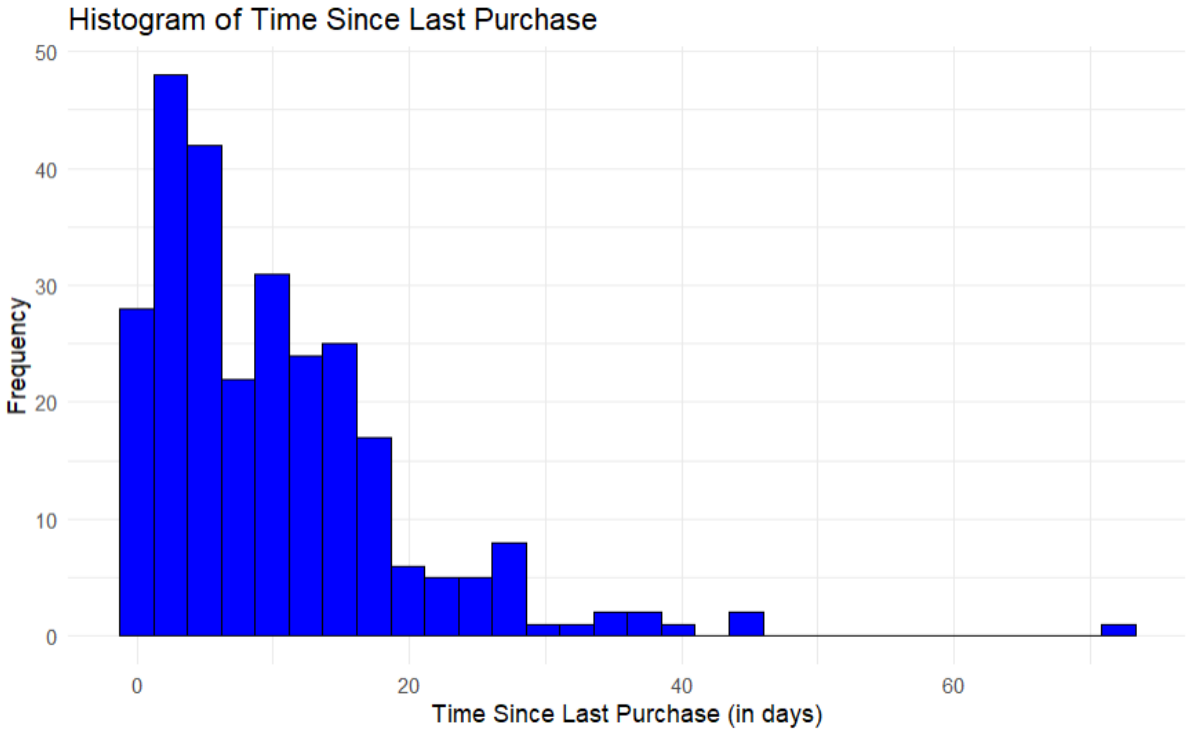
WALMART DATA ANALYSIS

Q1 GAMMA DISTRIBUTION

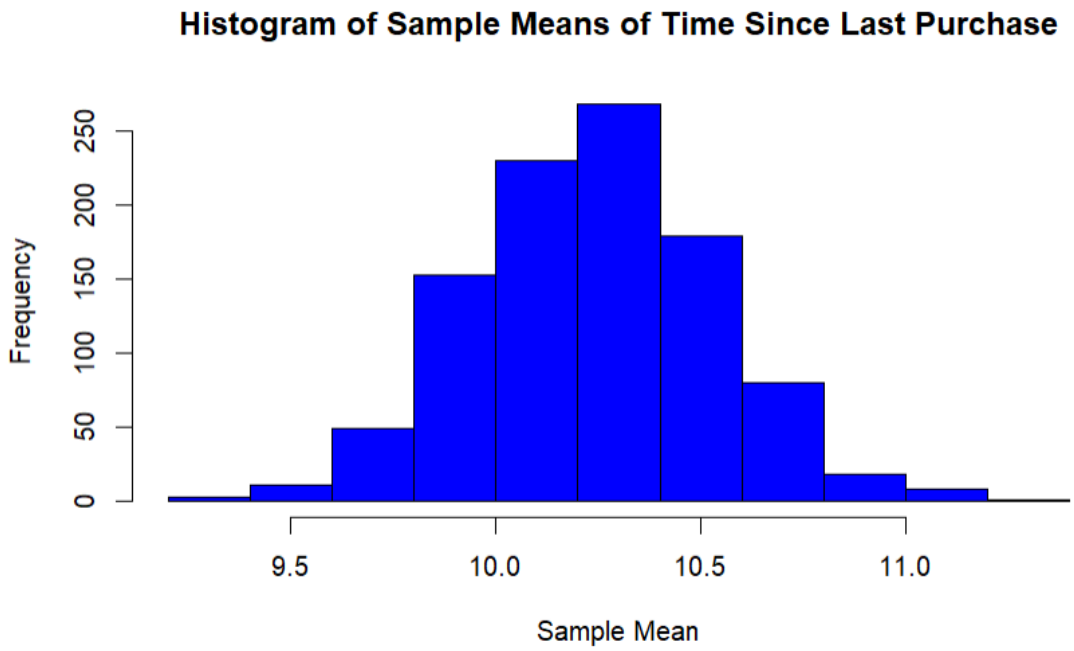


The "**Time_Since_Last_Purchase**" variable from the Walmart data shows a distribution that is right-skewed, with a peak near the origin and a long tail extending towards the right. This suggests that a large number of customers have made purchases relatively recently, with the frequency of customers who made their last purchase decreases as the time since the last purchase increases.

Implications for Customer Behavior: The implication of such a distribution in the context of customer purchases is that there's a significant number of customers who make frequent purchases, and as more time passes, the likelihood of a repeat purchase decreases. This can inform marketing strategies, inventory management, and customer retention efforts. Businesses might focus on strategies to encourage repeat purchases sooner or to re-engage customers who are moving towards the long tail of the distribution.

Understanding the distribution of time since the last purchase can help Walmart in planning promotions, loyalty programs, and other customer engagement strategies to possibly shift the distribution towards more frequent purchases, thereby increasing the frequency of customer visits to the store.

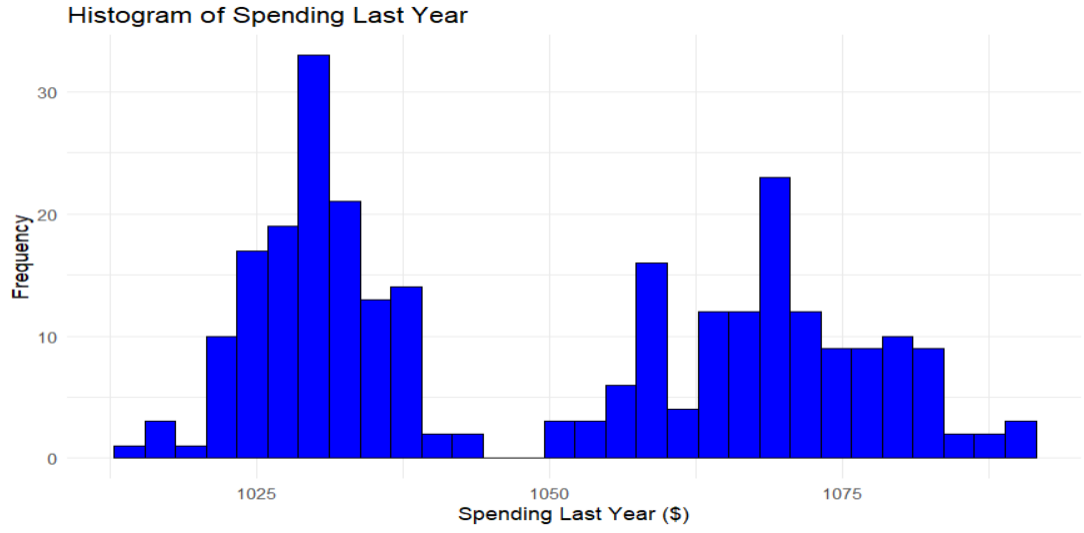
Q2 NORMAL DISTRIBUTION



The histogram of the sample means for "**Time_Since_Last_Purchase**" from the Walmart data, as you've confirmed to follow a Normal Distribution, reflects the principles of the Central Limit Theorem (CLT). This theorem states that the distribution of the sample means will tend to be normal if the sample size is large enough, regardless of the distribution of the underlying data.

In essence, the histogram suggests that with a large number of samples, Walmart can expect a stable and reliable average of the time since last purchase, which is crucial for various strategic and operational decisions.

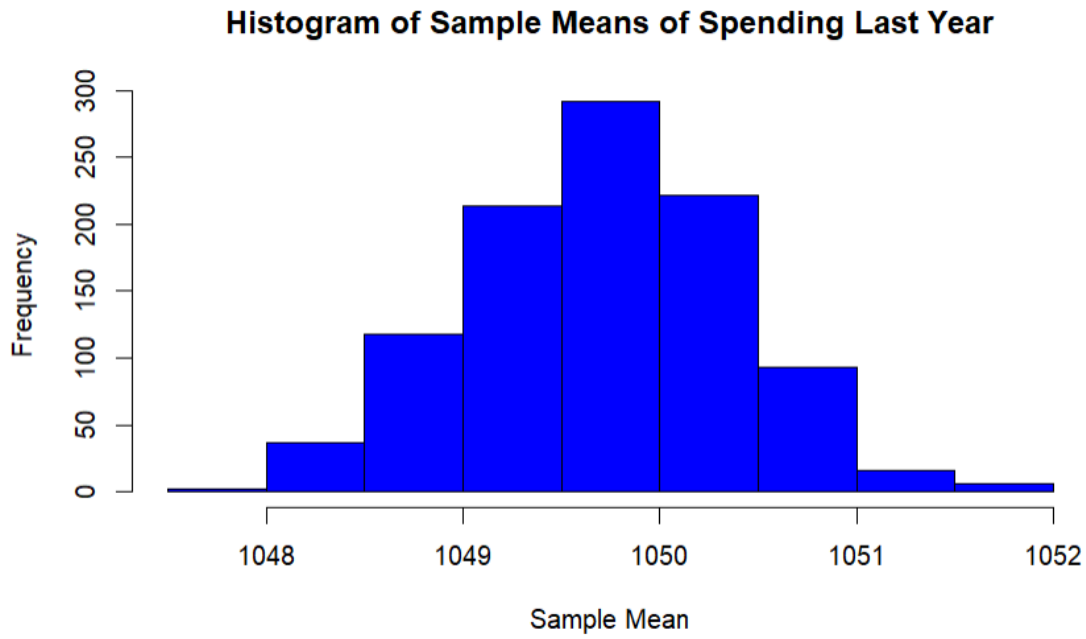
Q3 BIMODAL (2 PEAKS)



The histogram of the "**Spending_Last_Year**" variable from the Walmart data, which you've confirmed as having a bimodal distribution, shows two distinct peaks. A bimodal distribution implies that there are two different common spending amounts among the customers surveyed, suggesting the presence of two subgroups within the customer base with different spending behaviors.

The presence of a bimodal distribution is an insightful discovery that can lead to a deeper understanding of consumer behavior and help inform a range of business decisions.

Q4 NORMAL DISTRIBUTION



The histogram for the "**Spending_Last_Year**" variable's sample means shows what appears to be a normal distribution, which aligns with the expectations set by the Central Limit Theorem. This theorem states that as the sample size gets larger, the distribution of the sample means will tend to be approximately normal, regardless of the shape of the original data distribution.

The normally distributed sample means provide a solid foundation for various statistical analyses and can be a significant asset for Walmart's business strategies.

Q5

The resulting output relate to performing a t-test on the **'Time_Since_Last_Purchase'** data from the Walmart dataset.

The output of this test:

- Lower Bound of Confidence Interval: 9.11

- Upper Bound of Confidence Interval: 11.37

This means that based on the sample data from Walmart, we can be 95% confident that the true mean of the 'Time_Since_Last_Purchase' for the entire customer population lies between 9.11 and 11.37 days. In other words, the average number of days since the last purchase for all Walmart customers is estimated to fall within this range with a 95% level of confidence.

Q6

The provided R code and its output are about performing a t-test on the **'Spending_Last_Year'** data from the Walmart dataset, aiming to determine the 95% confidence interval for the mean spending by a customer last year.

The results of this test:

- Lower Bound of Confidence Interval for Spending Last Year: 1047.16
- Upper Bound of Confidence Interval for Spending Last Year: 1052.21

These results indicate that, based on the sampled data, we can be 95% confident that the true average spending of all Walmart customers last year falls between $1047.16 and $1052.21. This range represents our best estimate for the average annual spending of a Walmart customer, with a 95% level of confidence.

The relatively narrow range of this confidence interval suggests a high level of precision in our estimate of the average spending. It provides valuable insights into customer spending behavior, which can be utilized for business analysis, marketing strategies, and financial forecasting.

Q7

The provided output, "**Lower Bound of Confidence Interval: 0.21, Upper Bound of Confidence Interval: 0.38,"** represents the 95% confidence interval for a proportion derived from a statistical test (likely a two-sample z-test for proportions). This interval suggests that, with 95% confidence, the true proportion of the population being studied (in this case, customers who either shopped or did not shop last month at Walmart) lies between 21% and 38%.

In practical terms, if this confidence interval pertains to the proportion of customers who shopped last month, it means we can be 95% confident that between 21% to 38% of all customers made a purchase last month. This range provides a measure of the reliability and precision of the estimated proportion, indicating that the true proportion is likely within these bounds, based on the sample data analyzed.

Q8

The output "**Lower Bound of Confidence Interval: 0.18, Upper Bound of Confidence Interval: 0.4**" represents the 99% confidence interval for a proportion determined through a statistical analysis (likely a two-sample z-test for proportions).

In this context, the confidence interval indicates that, with 99% certainty, the true proportion of the population being examined (such as the proportion of Walmart customers who either shopped or did not shop last month) falls between 18% and 40%.

This wider interval, as compared to a 95% confidence interval, reflects a higher degree of certainty about where the true proportion lies. In practical terms, if this is related to the proportion of customers who shopped last month, it implies that we are 99% confident that the actual percentage of all customers who made a purchase last month is somewhere between 18% and 40%. The 99% confidence level provides a higher level of assurance about the estimate, but it also results in a broader range, reflecting increased uncertainty about the precise proportion.

Q9

The output "**Lower Bound of Confidence Interval: 0.22, Upper Bound of Confidence Interval: 0.36**" represents the 90% confidence interval for a proportion determined through a statistical test, likely a two-sample z-test for proportions.

This confidence interval suggests that, with 90% certainty, the true proportion of the population being analyzed (such as the proportion of Walmart customers who shopped or did not shop last month) lies between 22% and 36%.

Since this is a 90% confidence interval, it offers a lower degree of certainty compared to a 95% or 99% confidence interval, but as a trade-off, it provides a narrower range. In practical terms, if this interval pertains to the proportion of customers who shopped last month, it implies we can be 90% confident that the actual percentage of all customers who made a purchase last month is somewhere between 22% and 36%. The 90% confidence level indicates a moderate level of assurance about the estimate, with a relatively tighter range indicating less uncertainty about the precise proportion.

Q10

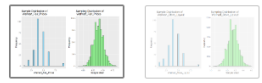
In statistics, the width of a confidence interval is directly related to the level of confidence chosen. A higher confidence level results in a wider confidence interval, while a lower confidence level leads to a narrower interval. This is because a higher confidence level requires more range to ensure that the true population parameter is captured within the interval.

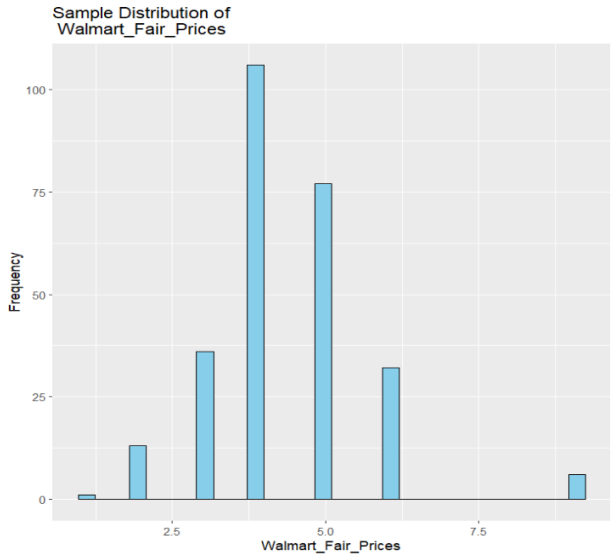
From the questions above, three different confidence levels were used: 90%, 95%, and 99%. Among these:

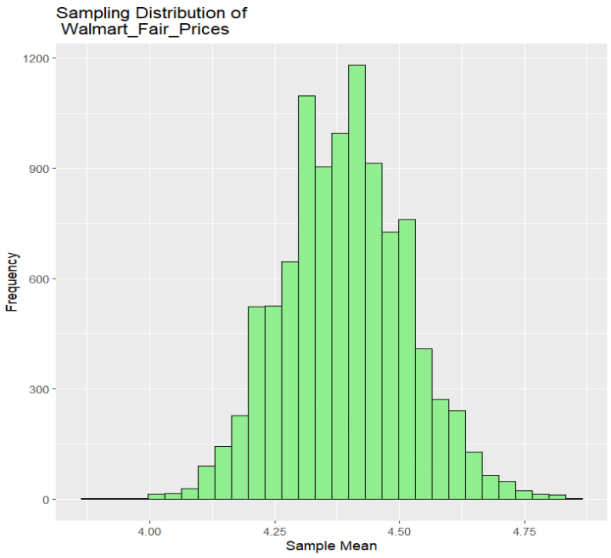
- A 90% confidence level will generate the narrowest confidence interval.
- A 95% confidence level will provide a wider interval than the 90% level.
- A 99% confidence level will generate the widest confidence interval of all.

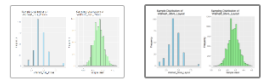
Therefore, the confidence level that generates the widest confidence interval is the 99% confidence level.

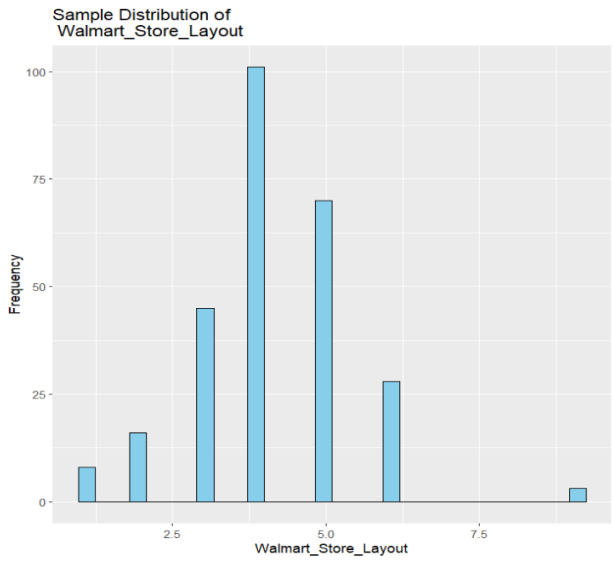
Q11

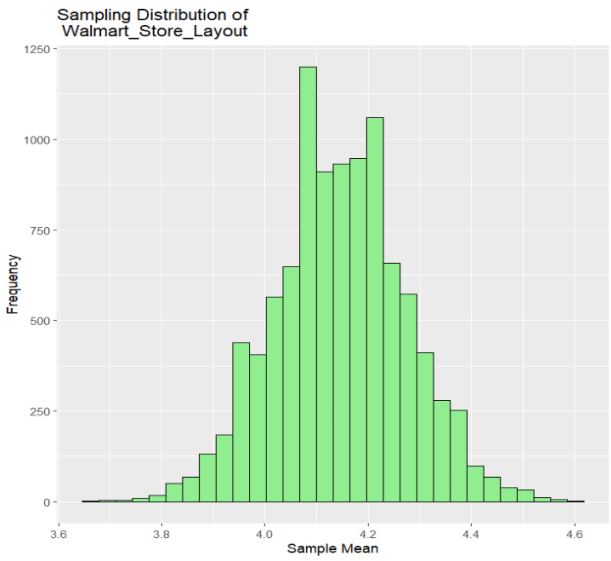












Analysis of Two Additional Variables from Walmart Data

Chosen Variables: Walmart_Fair_Prices & Walmart_Store_Layout

Plots and Distribution Shapes:

1. **Walmart_Fair_Prices:**

Sample Distribution: The sample distribution plot for **Walmart_Fair_Prices** shows how the respondents rated the fairness of prices at Walmart. The shape of this distribution can be inferred from the histogram.

Sampling Distribution: The sampling distribution, created using 10,000 drawn samples, demonstrates the behavior predicted by the **Central Limit Theorem**. It should approximate a normal distribution, with the sample means clustering around the true population mean.

2. Walmart_Store_Layout:

Sample Distribution: This plot reflects the distribution of respondents' ratings for Walmart's store layout. The distribution's shape (e.g., skewed, normal, bimodal) is visible in the histogram.

Sampling Distribution: Similarly, the sampling distribution for **Walmart_Store_Layout** is formed by 10,000 drawn samples. It should exhibit a normal distribution pattern as per the Central Limit Theorem.

Comparing Sample and Sampling Distributions:

Sample Distributions: These may vary in shape based on the actual data distribution (e.g., skewed, uniform, bimodal).

Sampling Distributions: Regardless of the shape of the sample distributions, the sampling distributions tend to be normal. This is a result of the **Central Limit Theorem**.

Central Limit Theorem (CLT):
The CLT is the theorem that predicts the sampling distribution will approach a normal distribution when a large number of samples are drawn. It states that as the sample size **becomes larger**, the distribution of the sample means will approximate a normal distribution, regardless of the population's original distribution. This theorem is foundational in statistics, enabling the use of normal distribution assumptions in various statistical methods, even when the underlying data does not strictly follow a normal distribution.

Conclusion:
The analysis of the two additional variables from the Walmart data illustrates the practical application of the Central Limit Theorem. By comparing the sample distributions with their corresponding sampling distributions, we observe how the latter conforms to a normal distribution as the number of drawn samples increases, providing a powerful tool for inferential statistics.

My Use of Confidence Intervals in Managing My Coffee Shop

As the owner of a neighborhood coffee shop, I'm constantly looking for ways to enhance our operations and profitability. Recently, I've been contemplating extending our shop's

operating hours. While this decision could potentially increase our sales and customer reach, I'm keenly aware that it also might lead to higher operating costs and could impact my team's work-life balance. To make an informed decision, I decided to use statistical methods, specifically confidence intervals, to analyze various aspects of our business.

Key Variables I Considered:

Sales Revenue: I wanted to understand the possible increase in sales if we extended our hours. By estimating a confidence interval for projected sales, I could gauge the range of potential revenue enhancements.

Customer Foot Traffic: It was crucial to determine whether the extended hours would actually attract more customers. Analyzing customer visits during our current and proposed extended hours would offer insights into this.

Employee Workload: My team's well-being is paramount. I needed to assess if the longer hours would require hiring more staff or lead to overwork.

Customer Satisfaction: Maintaining high customer satisfaction is at the core of our values. I planned to measure this before and after any change in operating hours.

Operating Costs: Lastly, I looked at the potential increase in costs such as utilities and wages. It was vital to ensure that any increase in revenue would not be negated by these additional expenses.

Stakeholders Involved:
My staff, who are concerned about their schedules and job satisfaction. Our loyal customers, who value both our service quality and availability. Local suppliers, who might need to adjust their delivery schedules in response to our changes. The local community, which benefits from a thriving local business.

By applying confidence intervals to these variables, I could estimate the potential impacts of extended hours with a certain level of statistical confidence. This approach offered a balanced view, helping me weigh the benefits against the risks from a data-driven perspective. It wasn't just about the numbers; it was about making a decision that aligned with our values and goals as a community-focused coffee shop.