Statistics 7 Review Problem 2 and 21

Sampling Distributions of Means and Medians & Central Limit Effect

Lucas Liona

Table of contents

Problem 2: Die Tossing Experiment

Consider the experiment of throwing an ideal die three times (i.e., all outcomes are equally likely). Let $\bar{X}$ and M denote the mean and median, respectively, of the three scores obtained.

Part (a): Theoretical Expected Value

Find $\mathrm{E}(\bar{X})$.

```
# For a fair die, the expected value of a single roll is
E_X <- (1 + 2 + 3 + 4 + 5 + 6)/6
E_X
```

```
[1] 3.5
```

```r
# The expected value of the mean of 3 rolls is the same
E_X_bar <- E_X
E_X_bar
```

```
[1] 3.5
```

The expected value of the mean of three die rolls is 3.5, which makes sense because each individual roll has an expected value of 3.5, and the mean of three rolls would maintain this same expected value.

Part (b): Experimental Results

The results of 10 repetitions of the preceding experiment are given in Table 1.

```r
# Create a matrix to represent the die tosses
die_tosses <- matrix(c(
  4, 2, 4, 2, 6, 2, 3, 5, 3, 2,
  3, 6, 6, 3, 1, 5, 1, 5, 4, 1,
  5, 5, 2, 5, 5, 2, 4, 2, 3, 6
), nrow = 3, byrow = TRUE)

# Display the data as a table
experiment_numbers <- 1:10
colnames(die_tosses) <- paste("Exp", experiment_numbers)
rownames(die_tosses) <- paste("Toss", 1:3)
knitr::kable(die_tosses, caption = "Die Tossing Results")
```

Table 1: Die Tossing Results

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7	Exp 8	Exp 9	Exp 10
Toss 1	4	2	4	2	6	2	3	5	3	2
Toss 2	3	6	6	3	1	5	1	5	4	1
Toss 3	5	5	2	5	5	2	4	2	3	6

```r
# Calculate medians and means for each experiment
medians <- apply(die_tosses, 2, median)
means <- apply(die_tosses, 2, mean)

# Combine results into a data frame
```

```r
results <- rbind(medians, means)
knitr::kable(results, digits = 3,
             caption = "Medians and Means for Each Experiment")
```

Table 2: Medians and Means for Each Experiment

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7	Exp 8	Exp 9	Exp 10
medians	4	5.000	4	3.000	5	2	3.000	5	3.000	2
means	4	4.333	4	3.333	4	3	2.667	4	3.333	3

```r
# Calculate summary statistics
median_mean <- mean(medians)
median_sd <- sd(medians)
mean_mean <- mean(means)
mean_sd <- sd(means)

cat("Sample statistics for the medians:\n")
```

```
Sample statistics for the medians:
```

```r
cat("Mean =", median_mean, ", Standard deviation =", median_sd, "\n\n")
```

```
Mean = 3.6 , Standard deviation = 1.173788
```

```r
cat("Sample statistics for the means:\n")
```

```
Sample statistics for the means:
```

```r
cat("Mean =", mean_mean, ", Standard deviation =", mean_sd, "\n")
```

```
Mean = 3.566667 , Standard deviation = 0.5676462
```

The sample mean of the 10 medians is 3.6 with a standard deviation of 1.174. The sample mean of the 10 means is 3.567 with a standard deviation of 0.568.

Notice that both estimators (the mean and median) have similar averages (close to the theoretical 3.5), but the standard deviation of the means is much smaller than the standard deviation of the medians.

Part (c): Histogram Comparison

Using the four histogram intervals [1.5, 2.5], [2.5, 3.5], [3.5, 4.5], and [4.5, 5.5], draw histograms of the two frequency distributions obtained in (b).

```
# Create frequency tables
breaks <- c(1.5, 2.5, 3.5, 4.5, 5.5)
median_freq <- table(cut(medians, breaks))
mean_freq <- table(cut(means, breaks))

# Display frequency tables
median_table <- data.frame(
  Interval = c("1.5-2.5", "2.5-3.5", "3.5-4.5", "4.5-5.5"),
  Frequency = as.vector(median_freq)
)
mean_table <- data.frame(
  Interval = c("1.5-2.5", "2.5-3.5", "3.5-4.5", "4.5-5.5"),
  Frequency = as.vector(mean_freq)
)

knitr::kable(median_table, caption = "Frequency Distribution of Medians")
```

Table 3: Frequency Distribution of Medians

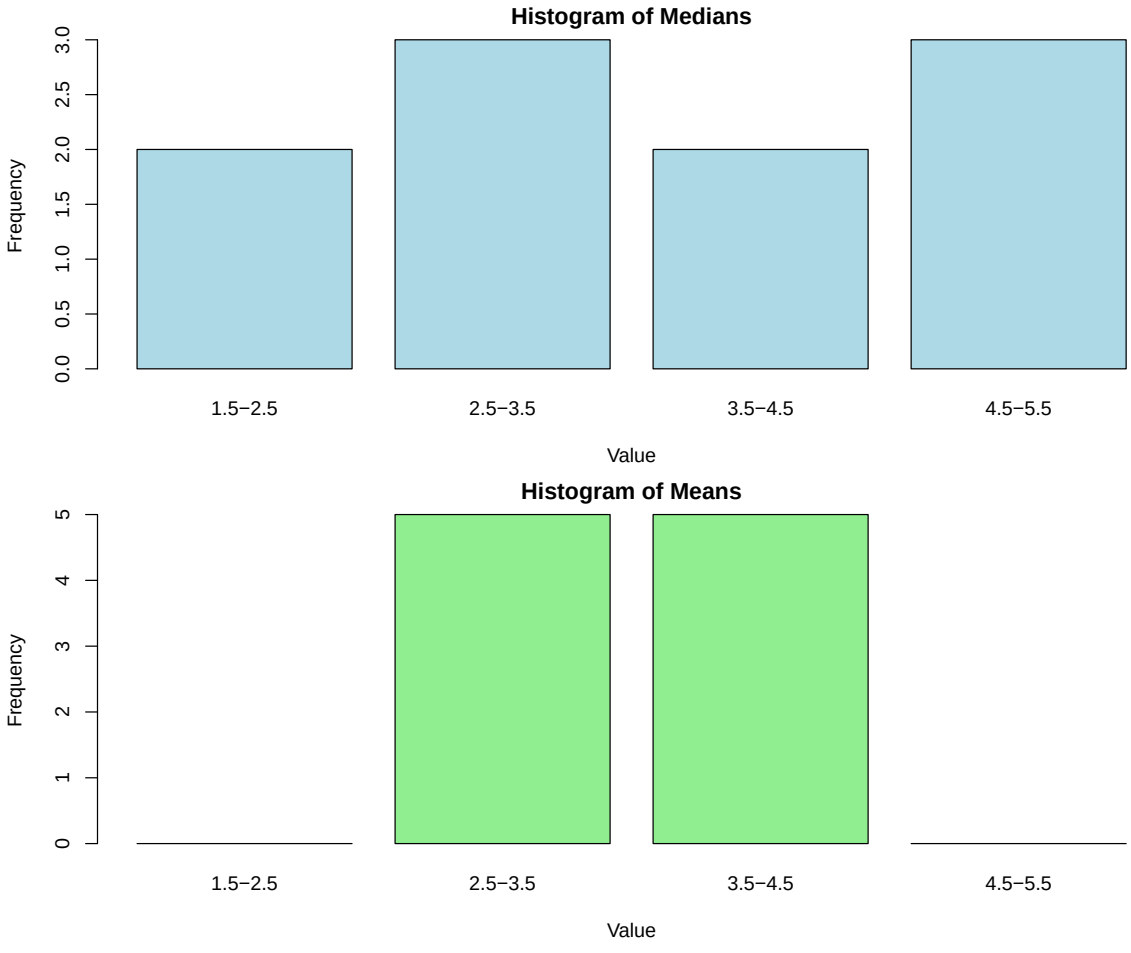
Interval	Frequency
1.5-2.5	2
2.5-3.5	3
3.5-4.5	2
4.5-5.5	3

```
knitr::kable(mean_table, caption = "Frequency Distribution of Means")
```

Table 4: Frequency Distribution of Means

Interval	Frequency
1.5-2.5	0
2.5-3.5	5
3.5-4.5	5
4.5-5.5	0

```r
# Create histograms
par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))

# Histogram for medians
barplot(median_freq, main = "Histogram of Medians",
        xlab = "Value", ylab = "Frequency",
        names.arg = c("1.5-2.5", "2.5-3.5", "3.5-4.5", "4.5-5.5"),
        col = "lightblue")

# Histogram for means
barplot(mean_freq, main = "Histogram of Means",
        xlab = "Value", ylab = "Frequency",
        names.arg = c("1.5-2.5", "2.5-3.5", "3.5-4.5", "4.5-5.5"),
        col = "lightgreen")
```



Part (d): Empirical Sampling Distributions

Compare the empirical sampling distributions of $\bar{X}$ and M.

Looking at both the numerical summaries and the histograms, we can make several observations:

1. The means are less variable than the medians - the standard deviation of the means (0.568) is less than half that of the medians (1.174).
2. The distribution of the means is more concentrated - all values fall within the middle two intervals (2.5-4.5), while the medians are spread across all four intervals.
3. Both distributions have similar centers (around 3.5-3.6), which agrees with the theoretical expected value of 3.5.
4. The median values show a more uniform distribution pattern, while the mean values show a more bell-shaped distribution.

This matches statistical theory - sample means tend to have less variability than other statistics (like medians) and tend to distribute more normally. Even with just 10 samples, we can see this difference in behavior starting to emerge.

The practical implication is that if we had to choose between using the mean or median as an estimator in this situation, the mean would generally provide more consistent results from sample to sample.

Problem 21: Central Limit Effect

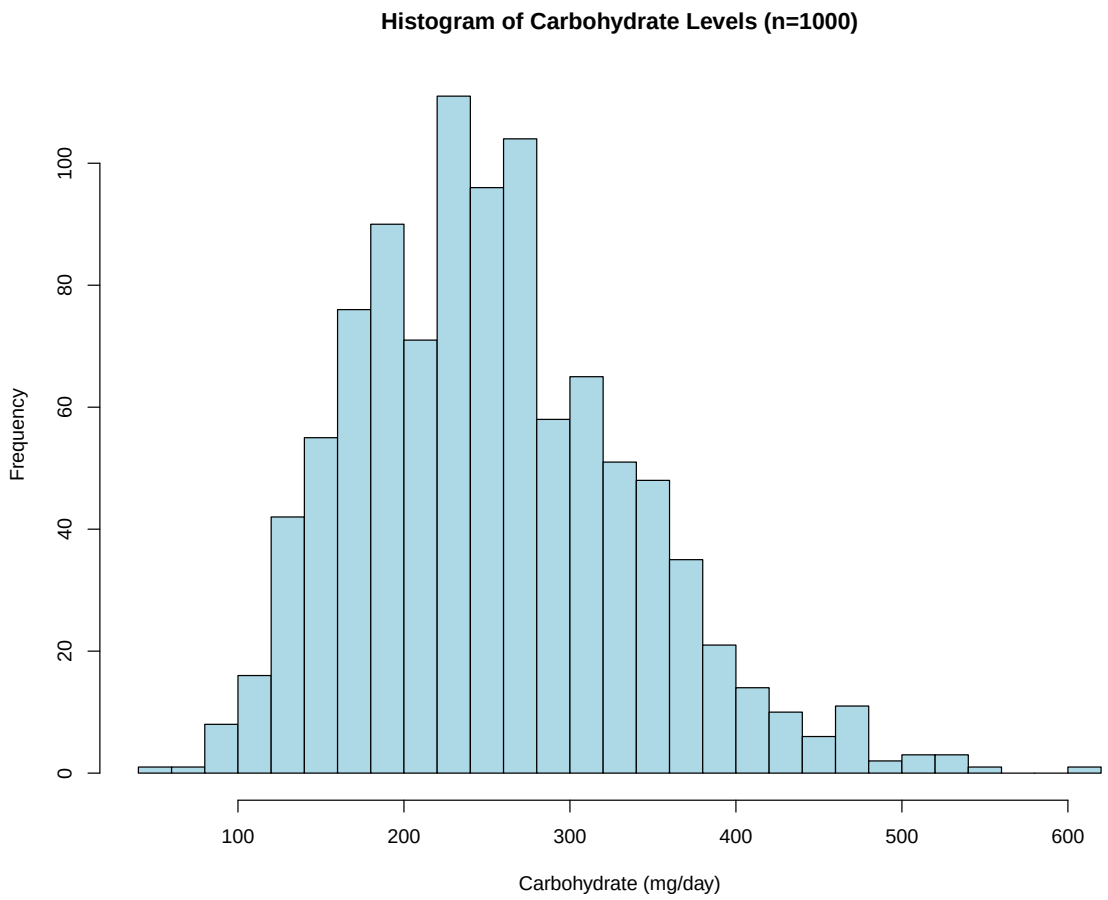
The central limit theorem is one of the most important results in statistics. It states that regardless of the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution as the sample size increases. Let's explore this through simulation.

Part (a): Generating Carbohydrate Levels and Initial Histogram

We'll generate 1000 "carbohydrate levels" by taking random numbers from the Chi-square distribution with 17 degrees of freedom, multiplying by 15, and adding 4.5.

```
set.seed(123) # For reproducibility

# Generate 1000 carbohydrate levels
n_obs <- 1000
df <- 17
```

```r
carb_levels <- rchisq(n_obs, df) * 15 + 4.5

# Create histogram of the data
hist(carb_levels,
     breaks = 30,
     main = "Histogram of Carbohydrate Levels (n=1000)",
     xlab = "Carbohydrate (mg/day)",
     ylab = "Frequency",
     col = "lightblue")
```




```r
# Summary statistics
cat("Summary statistics for carbohydrate levels:\n")
```

```
Summary statistics for carbohydrate levels:
```

```r
cat("Mean =", mean(carb_levels), "\n")
```

```
Mean = 253.6062
```

```r
cat("Standard deviation =", sd(carb_levels), "\n")
```

```
Standard deviation = 83.88046
```

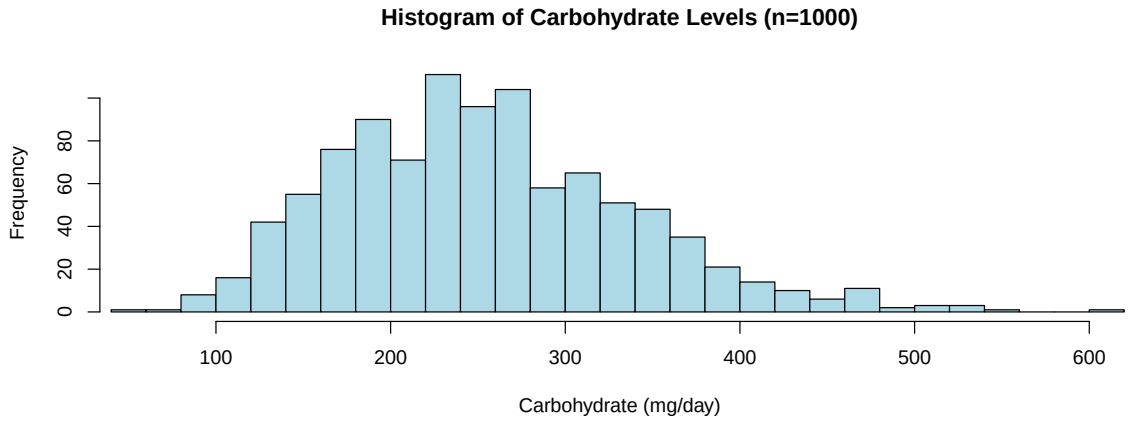
The histogram shows a right-skewed distribution, similar to Figure 6.1.1 from the textbook. This is characteristic of carbohydrate intake levels, where most people consume moderate amounts, but some individuals have much higher intake, creating a positive skew.

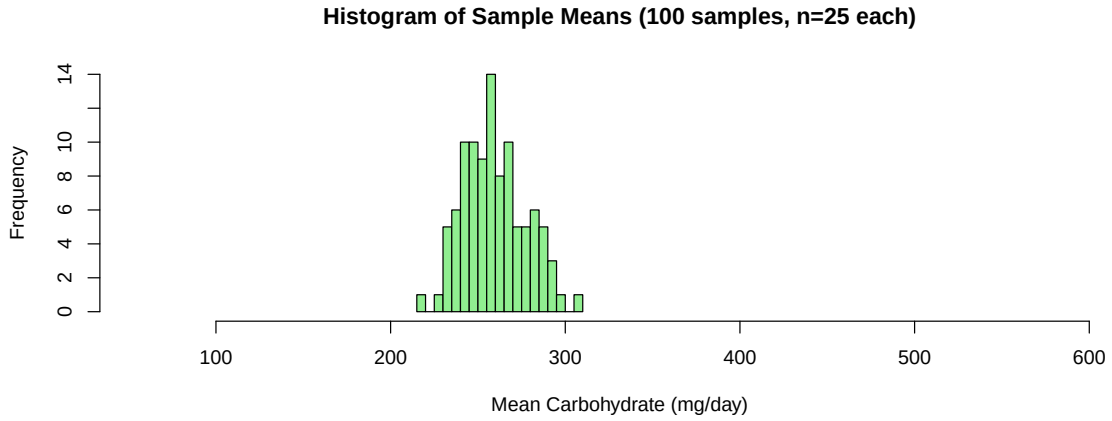
Part (b): Sampling Distribution of Means

Now, let's generate 100 samples, each of size 25, calculate the sample mean for each sample, and plot the distribution of these means.

```r
# Generate 100 samples of size 25
n_samples <- 100
sample_size <- 25
sample_means <- numeric(n_samples)

for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df) * 15 + 4.5
  sample_means[i] <- mean(sample_data)
}

# Create histogram of sample means using same x-axis range as original data
hist_range <- range(carb_levels)

par(mfrow = c(2, 1))

# Original data histogram
hist(carb_levels,
     breaks = 30,
     main = "Histogram of Carbohydrate Levels (n=1000)",
     xlab = "Carbohydrate (mg/day)",
     ylab = "Frequency",
     col = "lightblue",
     xlim = hist_range)
```

```r
# Sample means histogram
hist(sample_means,
     breaks = 15,
     main = "Histogram of Sample Means (100 samples, n=25 each)",
     xlab = "Mean Carbohydrate (mg/day)",
     ylab = "Frequency",
     col = "lightgreen",
     xlim = hist_range)
```






```r
# Summary statistics for sample means
cat("\nSummary statistics for sample means:\n")
```

```
Summary statistics for sample means:
```

```r
cat("Mean of sample means =", mean(sample_means), "\n")
```

```
Mean of sample means = 259.5856
```

```r
cat("Standard deviation of sample means =", sd(sample_means), "\n")
```

```
Standard deviation of sample means = 17.7676
```

```r
cat("Theoretical standard error =", sd(carb_levels)/sqrt(sample_size), "\n")
```

```
Theoretical standard error = 16.77609
```

Comparing the two histograms, we can see:

1. The distribution of sample means is much more concentrated around the center compared to the original data.
2. The distribution of sample means appears more symmetric and bell-shaped (closer to a normal distribution), even though the original data is right-skewed.
3. The standard deviation of the sample means is approximately 1/5 of the standard deviation of the original data, which is consistent with theory (SE = /√n).

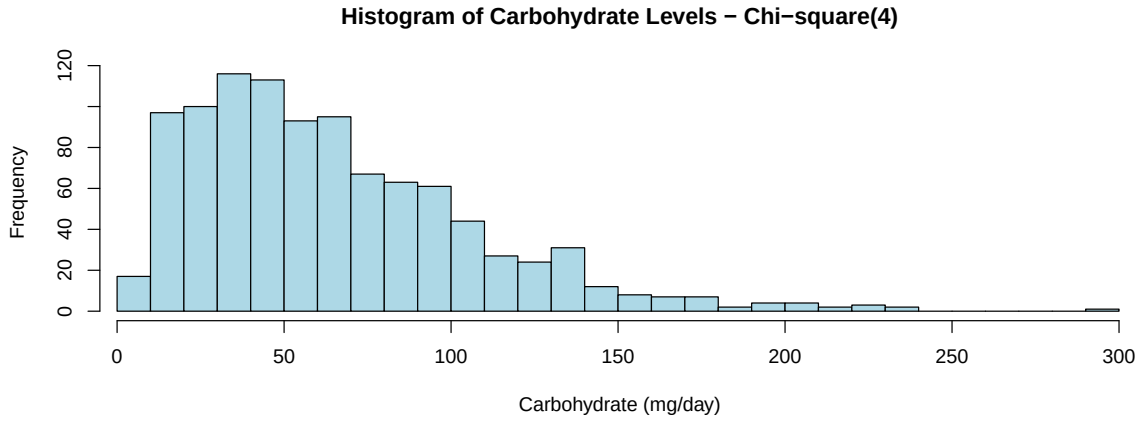
These observations confirm the central limit theorem in action - as we take means of samples, the distribution becomes more normal, regardless of the shape of the original distribution.

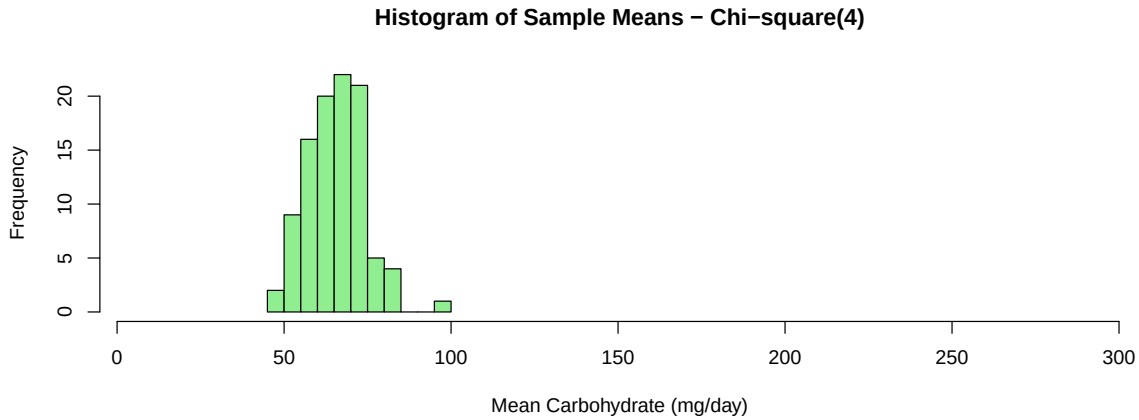
Part (c): Repeating with Chi-square(4) Distribution

Now let's repeat the process using a Chi-square distribution with 4 degrees of freedom, which will be even more skewed than the previous distribution.

```r
set.seed(456) # Different seed for reproducibility

# Generate 1000 carbohydrate levels with Chi-square(4)
df2 <- 4
carb_levels2 <- rchisq(n_obs, df2) * 15 + 4.5

# Generate 100 samples of size 25
sample_means2 <- numeric(n_samples)
```

```r
for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df2) * 15 + 4.5
  sample_means2[i] <- mean(sample_data)
}

# Create histograms
hist_range2 <- range(carb_levels2)

par(mfrow = c(2, 1))

# Original data histogram
hist(carb_levels2,
     breaks = 30,
     main = "Histogram of Carbohydrate Levels - Chi-square(4)",
     xlab = "Carbohydrate (mg/day)",
     ylab = "Frequency",
     col = "lightblue",
     xlim = hist_range2)

# Sample means histogram
hist(sample_means2,
     breaks = 15,
     main = "Histogram of Sample Means - Chi-square(4)",
     xlab = "Mean Carbohydrate (mg/day)",
     ylab = "Frequency",
     col = "lightgreen",
     xlim = hist_range2)
```






```
# Summary statistics
cat("\nSummary statistics for Chi-square(4) data:\n")
```

```
Summary statistics for Chi-square(4) data:
```

```
cat("Mean of original data =", mean(carb_levels2), "\n")
```

```
Mean of original data = 64.81651
```

```
cat("Standard deviation of original data =", sd(carb_levels2), "\n")
```

```
Standard deviation of original data = 42.24794
```

```r
cat("\nSummary statistics for sample means:\n")
```

```
Summary statistics for sample means:
```

```r
cat("Mean of sample means =", mean(sample_means2), "\n")
```

```
Mean of sample means = 65.67284
```

```r
cat("Standard deviation of sample means =", sd(sample_means2), "\n")
```

```
Standard deviation of sample means = 8.576327
```

```r
cat("Theoretical standard error =", sd(carb_levels2)/sqrt(sample_size), "\n")
```

```
Theoretical standard error = 8.449588
```

With the Chi-square(4) distribution, which is even more skewed than Chi-square(17), we observe:

1. The original distribution is highly skewed to the right, with a long tail.
2. Despite the severe skewness of the original distribution, the distribution of sample means still appears more symmetric and bell-shaped.
3. The standard deviation of the sample means is again approximately what we would expect from theory (/√n).

This further confirms the power of the central limit theorem. Even with a more extremely skewed original distribution, the sampling distribution of the mean still approaches normality as sample size increases.

The transformation to normality is somewhat less complete than with the Chi-square(17) distribution, which is expected since the original distribution is more severely non-normal. This suggests that for more skewed distributions, larger sample sizes may be needed to achieve approximate normality in the sampling distribution.