QPM II

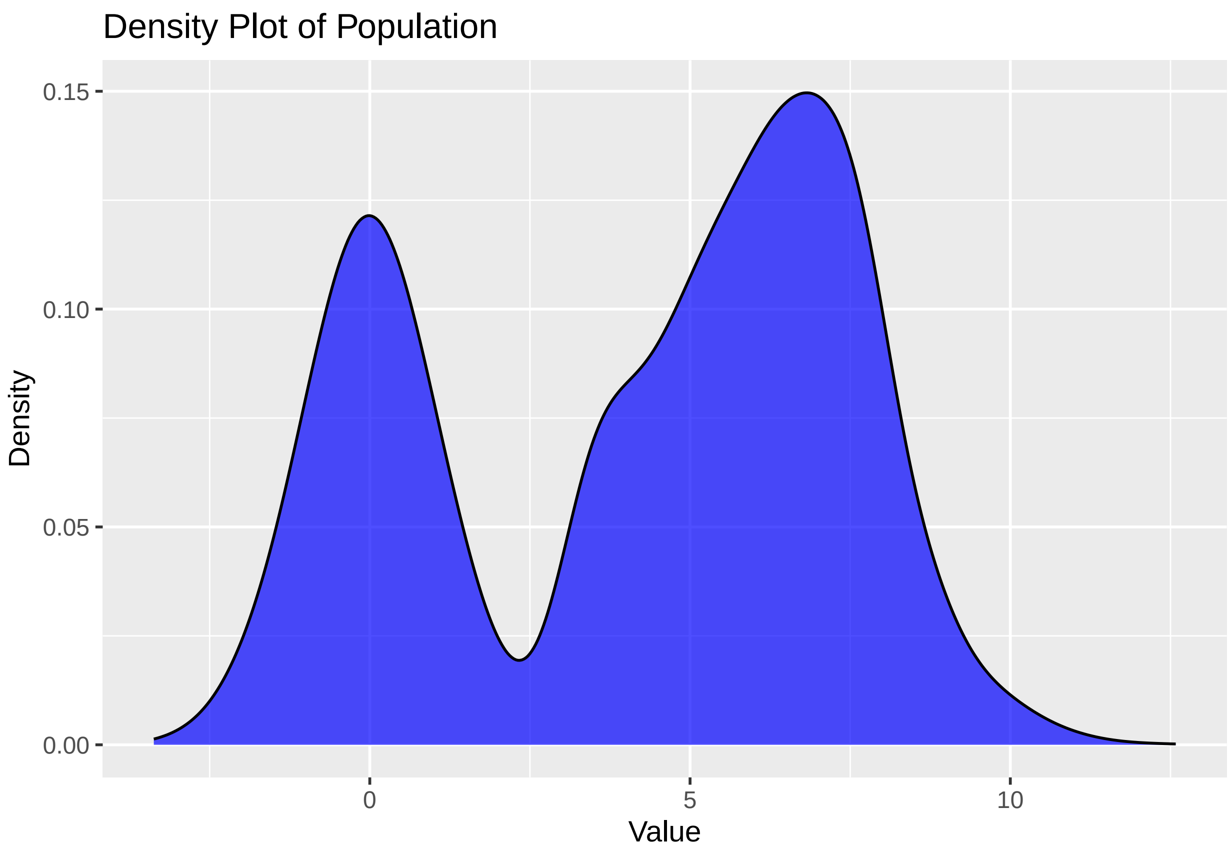
Problem Set 2

Daniel Cher

2024-08-08

Problem 12

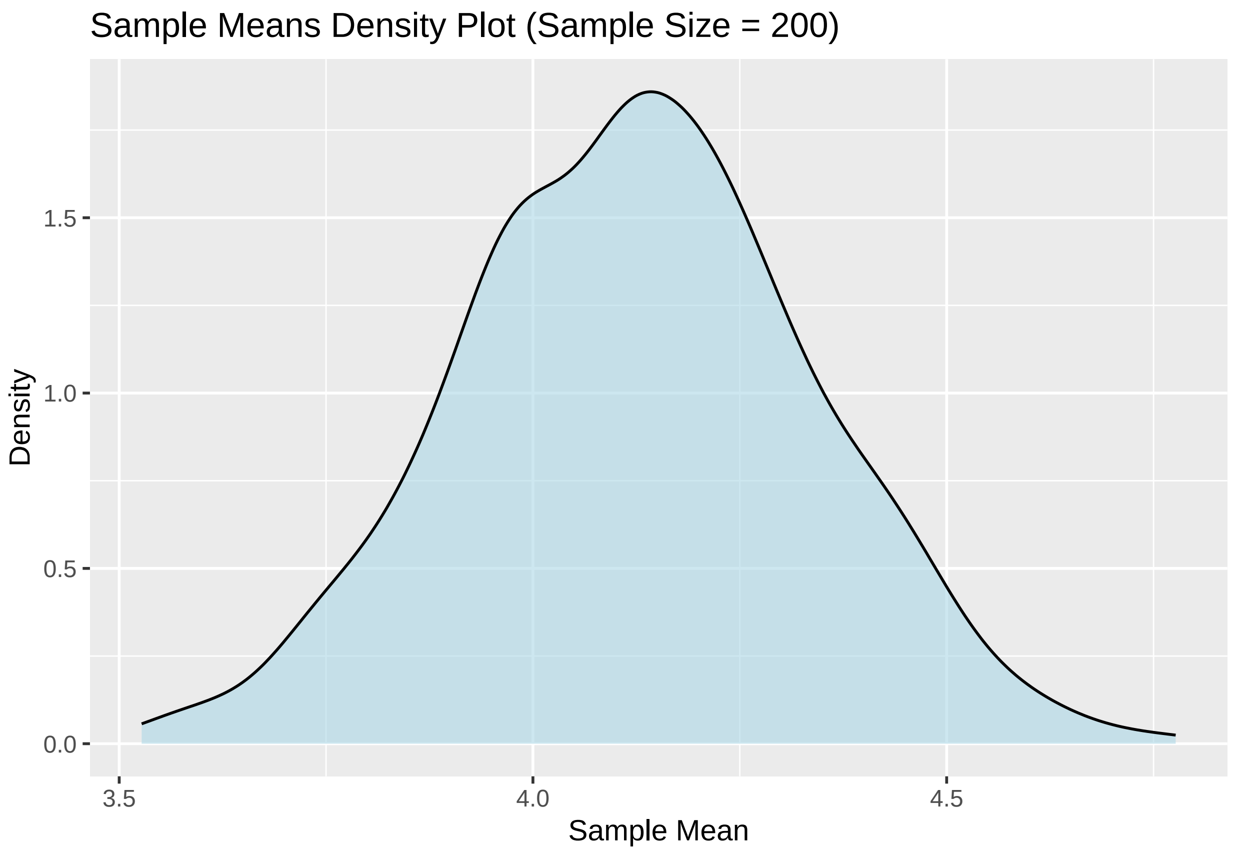
```r
n <- 10000

# Crazy distribution
data <- c(
  rnorm(n/3, mean = 0, sd = 1),
  runif(n/3, min = 3, max = 8),
  rnorm(n/3, mean = 6.9, sd = 1.5)
)

df <- data.frame(value = data)

# Density plot
ggplot(df, aes(x = value)) +
  geom_density(fill = "blue", alpha = 0.7) +
  labs(title = "Density Plot of Population",
       x = "Value",
       y = "Density")
```



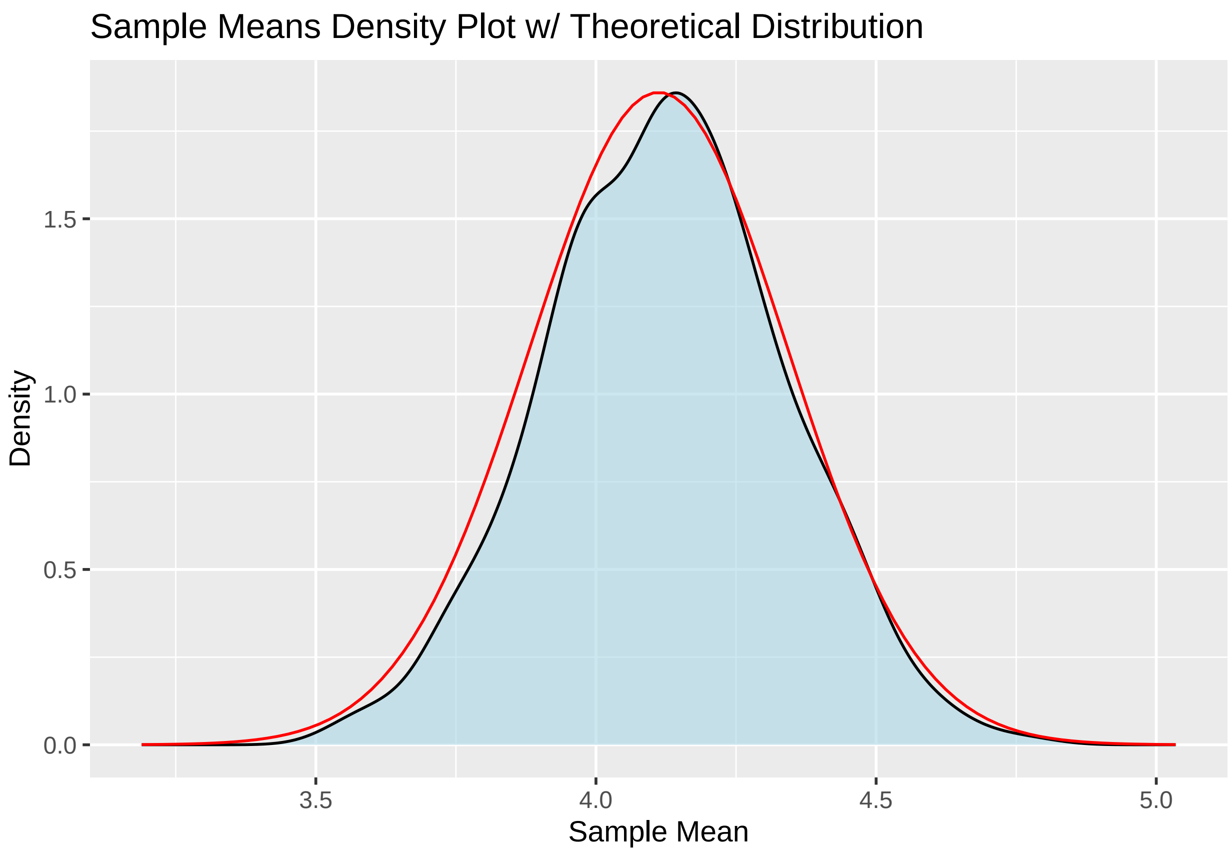
(a - e)

```r
sample <- sample(df$value, 200)
```

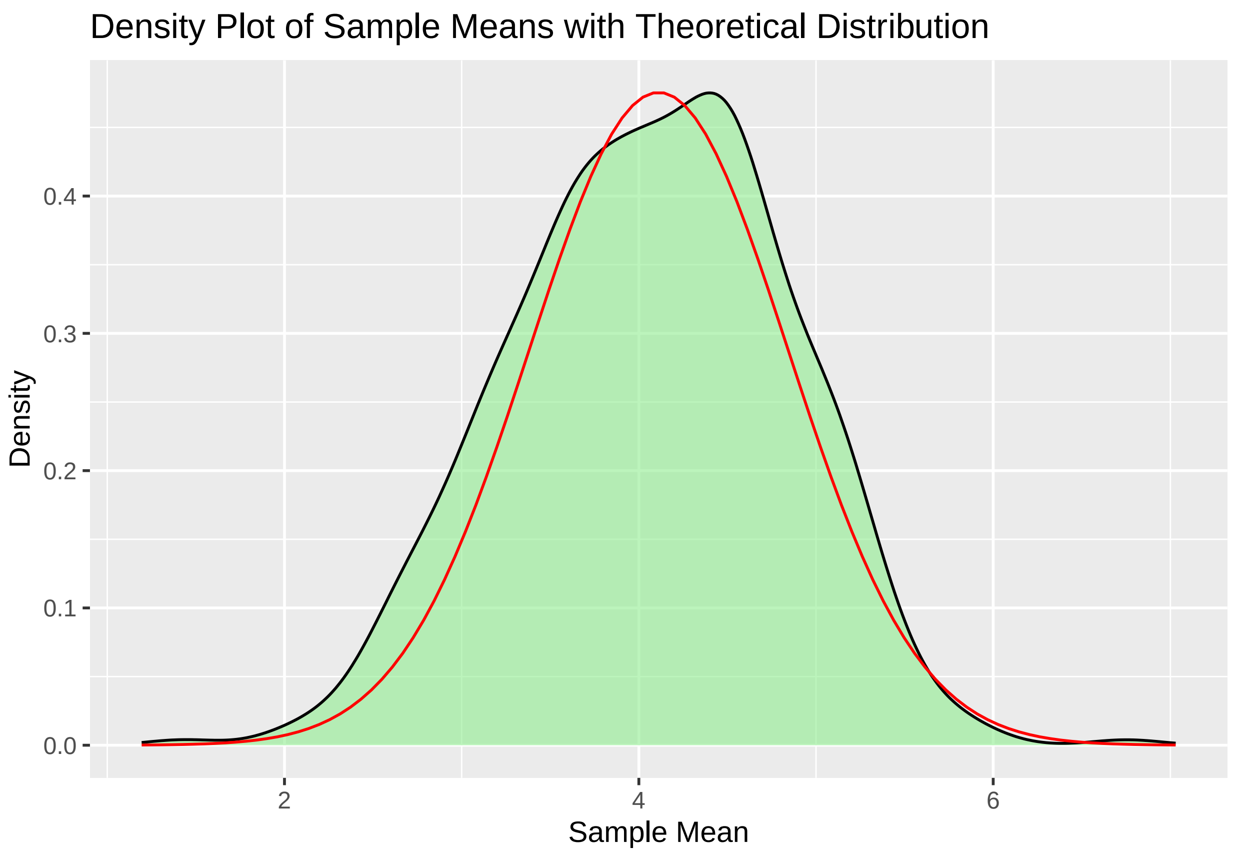
```r
# parameters
sample_200 <- 200
n_samples <- 500


# Get sample means
get_sample_means <- function(data, sample_size, n_samples) {
  means <- numeric(n_samples)
  for (i in 1:n_samples) {
    means[i] <- mean(sample(data, sample_size))
  }
  return(means)
}

sample_means_200 <- get_sample_means(df$value, sample_200, n_samples)

# Density Plot
plot_200 <- ggplot(data.frame(mean = sample_means_200), aes(x = mean)) +
  geom_density(fill = "lightblue", alpha = 0.6) +
  labs(title = "Sample Means Density Plot (Sample Size = 200)",
       x = "Sample Mean",
       y = "Density")

plot_200
```



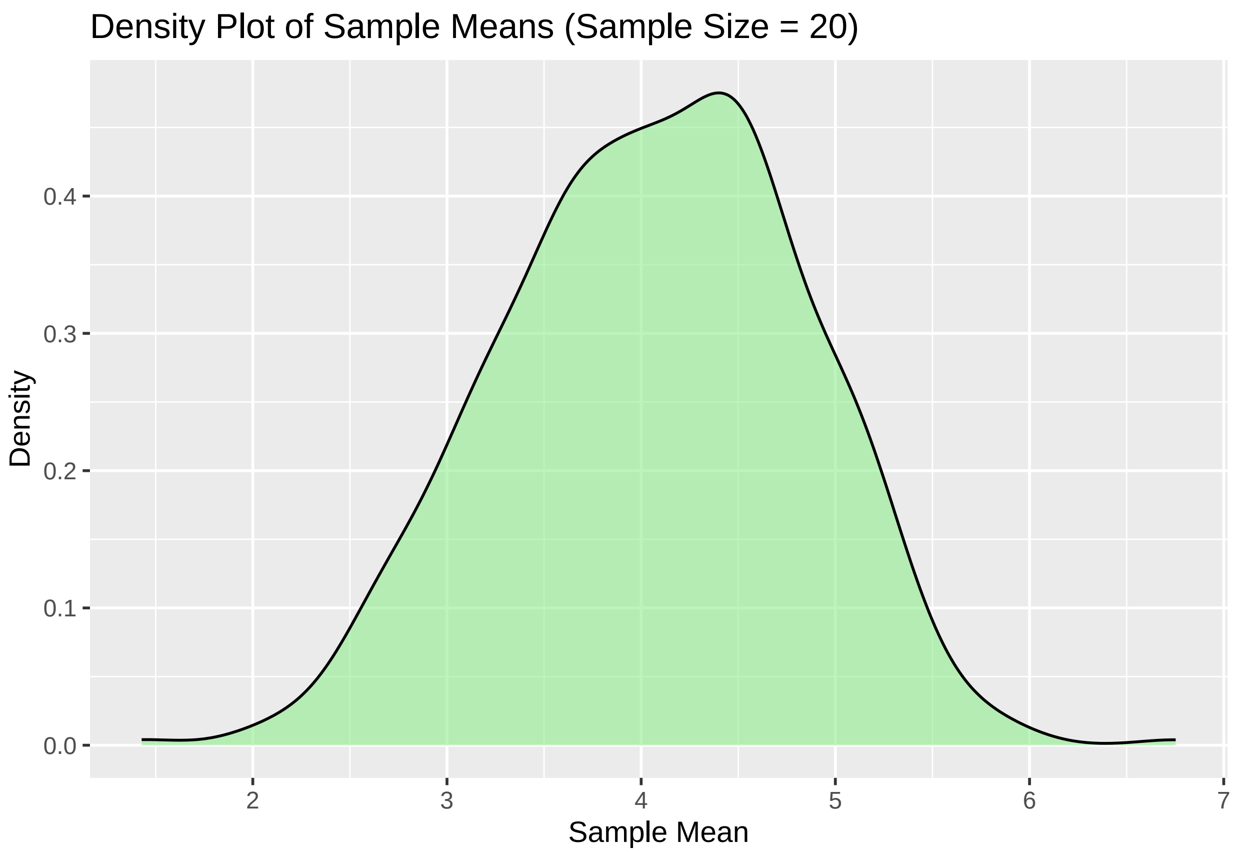
```r
# Calculate theoretical mean and standard deviation
theoretical_mean <- mean(df$value)
theoretical_sd <- sd(df$value) / sqrt(sample_200)

# Create range of values from theoretical calculated mean & sd -> density
x_values <- seq(theoretical_mean - 4 * theoretical_sd,
                 theoretical_mean + 4 * theoretical_sd,
                 length.out = 100)

theoretical_density <- dnorm(x_values, mean = theoretical_mean, sd = theoretical_sd)

# Scale so it's consistent with crazy distribution
scaled_theoretical_density <- theoretical_density * (max(density(sample_means_200)$y) / max(theoretical_

# Plot the sample means density and theoretical density
theoretical_data <- data.frame(x = x_values, y = scaled_theoretical_density)

plot_200 +
  geom_line(data = theoretical_data, aes(x = x, y = y), color = "red") +
  labs(title = "Sample Means Density Plot w/ Theoretical Distribution")
```



```
# Draw samples again with sample_size = 20
sample_20 <- 20
sample_means_20 <- get_sample_means(df$value, sample_20, n_samples)

# Remake density plot with 20
plot_20 <- ggplot(data.frame(mean = sample_means_20), aes(x = mean)) +
  geom_density(fill = "lightgreen", alpha = 0.6) +
  labs(title = "Density Plot of Sample Means (Sample Size = 20)",
       x = "Sample Mean",
       y = "Density")

# Do same thing as above
theoretical_mean <- mean(df$value)
theoretical_sd <- sd(df$value) / sqrt(sample_20)
x_values <- seq(theoretical_mean - 4 * theoretical_sd,
                 theoretical_mean + 4 * theoretical_sd,
                 length.out = 100)
theoretical_density <- dnorm(x_values, mean = theoretical_mean, sd = theoretical_sd)
scaled_theoretical_density <- theoretical_density * (max(density(sample_means_20)$y) / max(theoretical_
theoretical_data <- data.frame(x = x_values, y = scaled_theoretical_density)

# Plot the sample means density and theoretical density
plot_20 +
  geom_line(data = theoretical_data, aes(x = x, y = y), color = "red") +
  labs(title = "Density Plot of Sample Means with Theoretical Distribution")
```


Density Plot of Sample Means with Theoretical Distribution
Density
0.0
0.1
0.2
0.3
0.4
2
4
6
Sample Mean

```
plot_20
```



When using a smaller sample size the distribution of sample means will be wider, and more variable, because it is capturing less information about the population. Due to the Law of Large Numbers, as the sample size increases, the sample mean converges to the population mean, leading to a tighter distribution.

(e) If the population was normal then the distribution of sample means would also be normal. Due to the CLT, means of samples drawn from a normal distribution are also normally distributed. I would expect that the distribution of sample means would be less variable than the crazy distribution, since the samples will be more closely clustered to the true population mean.

13) The CLT is amazing because you can take pretty much any distribution (like the crazy one above) as long as it has finite mean and variance, and through a simple procedure of taking averages of n samples from the distribution, you can create a normal distribution centered at something that resembles the population mean. That means that if we want to understand some part of the world, we just need to sample from that population, and you'll most likely be able to understand a lot about the population just through those samples (with caveats of course). In a more philosophical sense, it uses randomness as a tool for knowledge, which is pretty amazing because randomness is naturally quite baffling. Finally, it also makes proofs a lot simpler, which is a HUGE plus, especially for graduate students.