Statistics 6 Review Exercises 20 and 21

Sampling Distributions and Simulation

Lucas Liona

Table of contents

Problem 20: Vehicle Fleet Analysis

Several years ago, a New Zealand government ministry wanted to reduce costs by cutting down on the number of vehicles in its fleet. The ministry estimated the number of idle days by sampling 60 days from the approximately 240 working days of the previous year. For 17 "medium cars" in a particular district, they found 824 idle vehicle-days when scaled up to the full year.

We'll analyze how accurate this estimate is by modeling the situation.

Part (a): Distribution of Idle Days

Let X_i be the number of days (out of the 60 sampled days) that the ith car was idle.

```
# Under the given assumptions, Xi follows a binomial distribution
# We'll define this for later use
car_idle_distribution <- "Binomial(n = 60, p)"
cat("Distribution of Xi:", car_idle_distribution, "\n")
```

```
Distribution of Xi: Binomial(n = 60, p)
```

Under the assumptions that the probability of a car being idle on any given day is p and that days and cars act independently, the variable X_i follows a Binomial(n = 60, p) distribution.

Part (b): Expected Value and Standard Deviation

For a binomial random variable, we have standard formulas for the expected value and standard deviation:

```
# For a binomial distribution with parameters n and p
# E(X) = np
# sd(X) = sqrt(np(1-p))

cat("E(Xi) = 60p\n")
```

```
E(Xi) = 60p
```

```
cat("sd(Xi) = sqrt(60p(1-p))\n")
```

```
sd(Xi) = sqrt(60p(1-p))
```

Part (c): Expected Value of the Total Estimate

The estimate was $Y = \frac{240}{60} \sum_{i=1}^{N} X_i = 4 \sum_{i=1}^{17} X_i$

```
# Calculate the expected value of Y
E_Y <- "4 × 17 × 60p = 4080p"
cat("E(Y) =", E_Y, "\n")
```

```
E(Y) = 4 × 17 × 60p = 4080p
```

This is a reasonable measure of the underlying idleness of the cars because the expected value represents the true number of idle car-days we would expect in a year if the probability of a car being idle on any day is p.

Part (d): Standard Deviation Formula

To find the standard deviation of Y, we use the properties of variance for independent variables:

```
# For Y = 4 * sum(Xi), where Xi are independent
# sd(Y) = 4 * sqrt(sum(Var(Xi)))
# sd(Y) = 4 * sqrt(17 * 60p(1-p))

sd_Y <- "4 × sqrt(17 × 60p(1-p)) = 127.75 × sqrt(p(1-p))"
cat("sd(Y) =", sd_Y, "\n")
```

```
sd(Y) = 4 × sqrt(17 × 60p(1-p)) = 127.75 × sqrt(p(1-p))
```

Part (e): Estimating p

We can estimate p using the observed data:

```
# Total car-days in the year: 17 × 240 = 4080
# Estimated idle car-days: 824
# Estimated p = 824/(17×240)

p_hat <- 824/(17*240)
cat("Estimated p =", p_hat, "\n")
```

```
Estimated p = 0.2019608
```

Part (f): Estimated Standard Deviation

Using our estimate of p, we can now estimate the standard deviation of Y:

```r
# Using the formula from part (d) with our estimate of p
p <- p_hat
sd_Y_estimated <- 127.75 * sqrt(p*(1-p))
cat("Estimated sd(Y) =", sd_Y_estimated, "\n")
```

```
Estimated sd(Y) = 51.28691
```

The standard deviation of our estimate is approximately 51.29 idle car-days.

Part (g): Recommendation for Car Reduction

To determine how many cars to sell, we need to analyze how many cars are typically idle on any given day:

```r
# Let U be the number of cars idle on a given day
# U ~ Binomial(n = 17, p = 0.202)

# Probability that at least 1 car is idle
prob_at_least_1 <- 1 - dbinom(0, 17, p)
cat("Probability at least 1 car is idle:", prob_at_least_1, "\n")
```

```
Probability at least 1 car is idle: 0.9784021
```

```r
# Probability that at least 2 cars are idle
prob_at_least_2 <- 1 - pbinom(1, 17, p)
cat("Probability at least 2 cars are idle:", prob_at_least_2, "\n")
```

```
Probability at least 2 cars are idle: 0.8854833
```

Based on these calculations, on a typical day there are at least 2 cars idle with a probability of about 0.89 (89%). Since this probability is high, it would be reasonable to reduce the fleet by 1-2 cars. This would likely have minimal impact on operations while reducing costs.

Problem 21: Simulation of Women's Heights

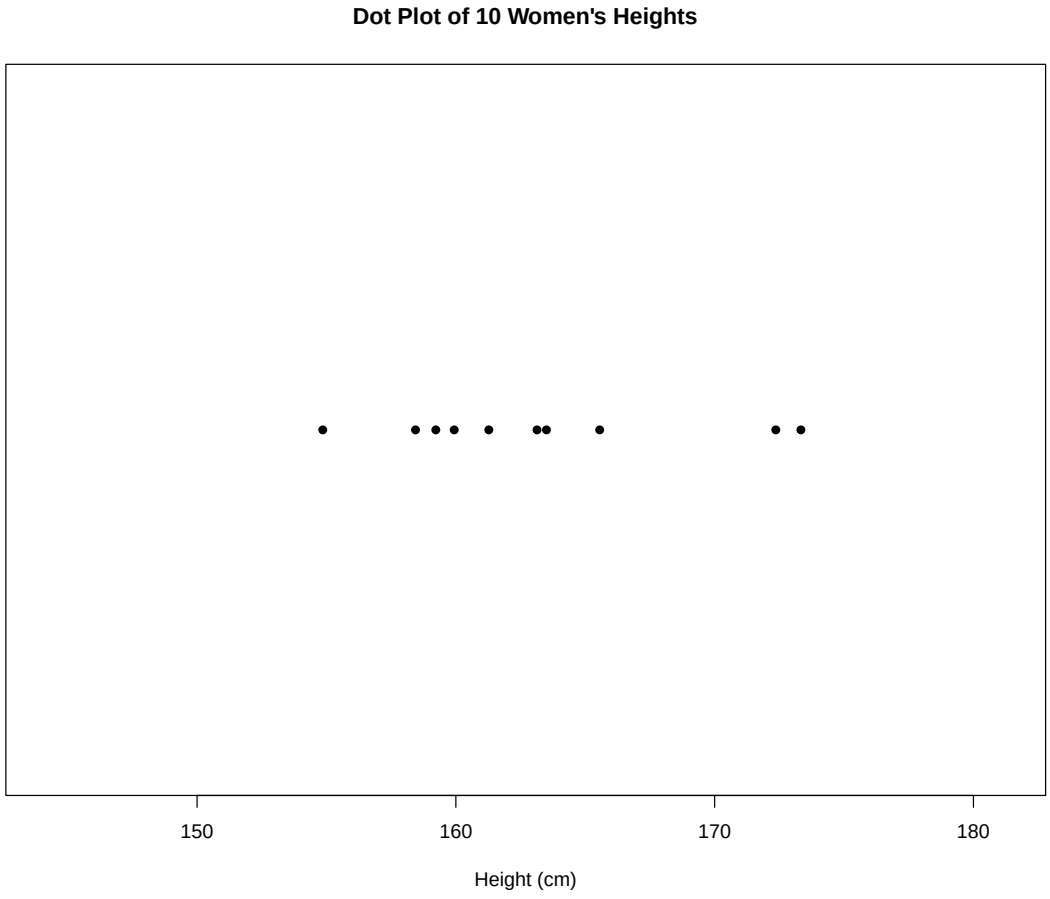
For this problem, we'll simulate samples of women's heights from a Normal distribution with mean 162.7 cm and standard deviation 6.2 cm.

Part (a): Generating 10 Random Heights

```r
# Set a seed for reproducibility
set.seed(123)

# Generate 10 random heights
mu <- 162.7
sigma <- 6.2
n <- 10

heights_10 <- rnorm(n, mean = mu, sd = sigma)
heights_10
```

```
 [1] 159.2251 161.2729 172.3640 163.1372 163.5016 173.3334 165.5577 154.8566
 [9] 158.4415 159.9369
```

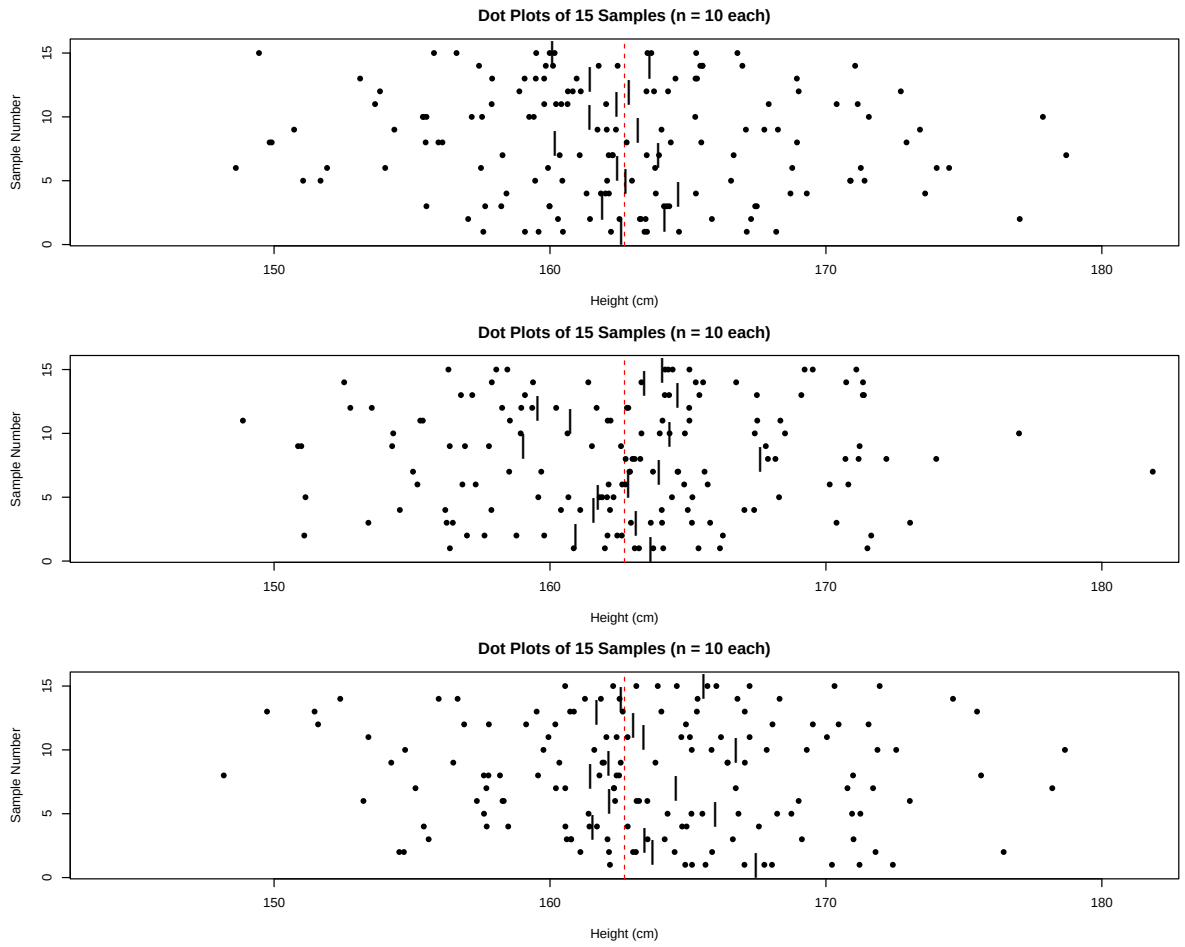
```r
# Create a dot plot
plot(heights_10, rep(1, n),
     pch = 16,
     xlim = c(mu - 3*sigma, mu + 3*sigma),
     ylim = c(0.5, 1.5),
     xlab = "Height (cm)",
     ylab = "",
     yaxt = "n",
     main = "Dot Plot of 10 Women's Heights")
```



Part (b): Generating 15 Sets of 10 Observations

```r
# Function to generate and plot multiple samples
generate_sample_plots <- function(n_samples, sample_size, mu, sigma) {
  # Create a plot area
  plot(0, 0, type = "n",
       xlim = c(mu - 3*sigma, mu + 3*sigma),
       ylim = c(0.5, n_samples + 0.5),
       xlab = "Height (cm)",
       ylab = "Sample Number",
       main = paste("Dot Plots of", n_samples, "Samples (n =", sample_size, "each)"))

  # Add a vertical line at the population mean
```

```r
  abline(v = mu, lty = 2, col = "red")

  # Generate and plot each sample
  for (i in 1:n_samples) {
    sample_heights <- rnorm(sample_size, mean = mu, sd = sigma)
    points(sample_heights, rep(i, sample_size), pch = 16)

    # Add a small vertical bar to indicate the sample mean
    points(mean(sample_heights), i, pch = "|", cex = 2)
  }
}

# Generate three sets of plots
par(mfrow = c(3, 1), mar = c(4, 4, 3, 1))
for (j in 1:3) {
  set.seed(j * 100)  # Different seed for each panel
  generate_sample_plots(15, 10, mu, sigma)
}
```



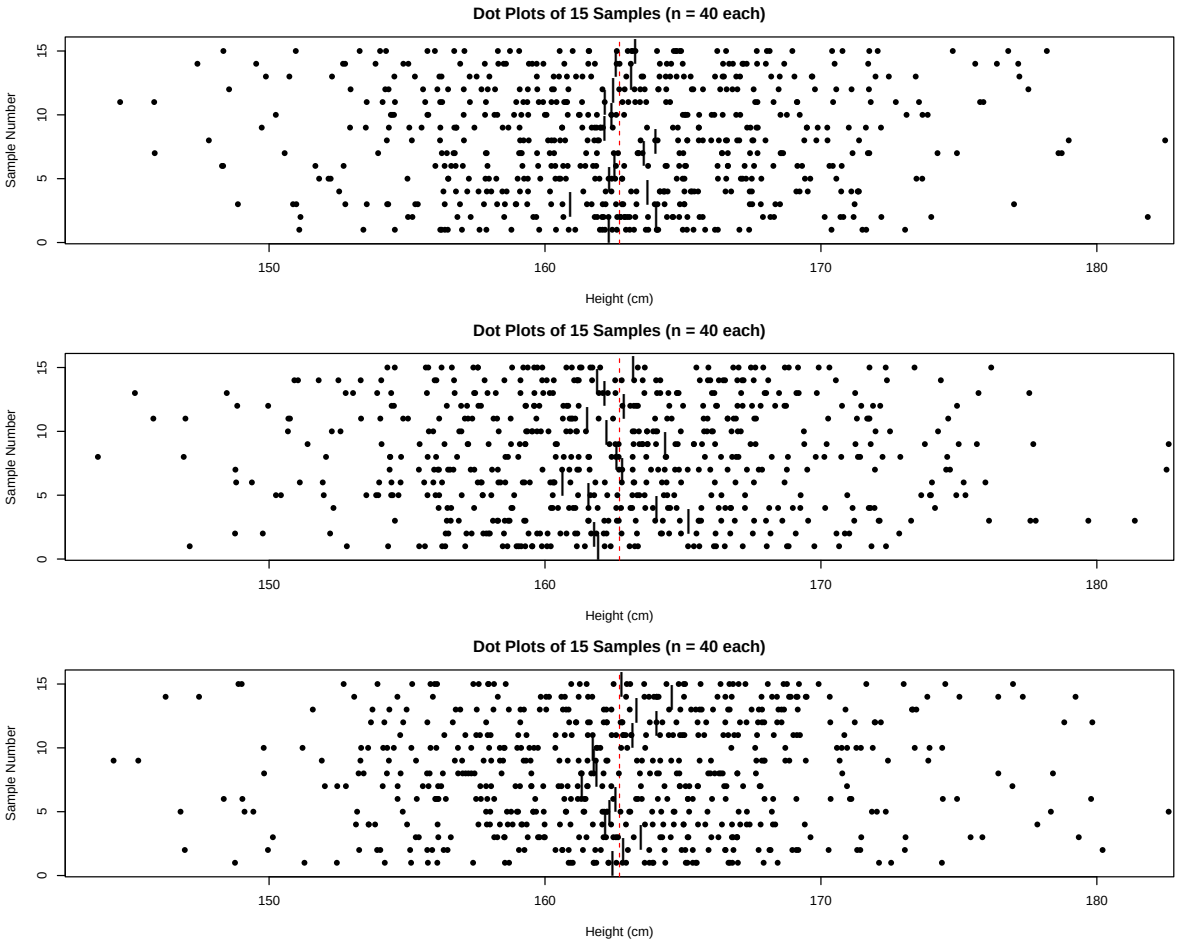
Looking at these plots, we can observe significant variation between samples:

1. Some samples appear to have outliers (points far from the rest of the sample)
2. Some samples seem to have gaps or clusters
3. Some samples appear slightly skewed (more spread out on one side)
4. The sample means (marked by vertical bars) vary considerably around the true population mean

All of these patterns occur naturally by chance, even though the data comes from a perfectly normal distribution. This illustrates how random sampling can produce apparent patterns that might be misinterpreted as true features of the population.

Part (c): Samples of Size 40

```r
# Generate three panels of 15 samples with n=40 each
par(mfrow = c(3, 1), mar = c(4, 4, 3, 1))
for (j in 1:3) {
  set.seed(j * 200)  # Different seed for each panel
  generate_sample_plots(15, 40, mu, sigma)
}
```



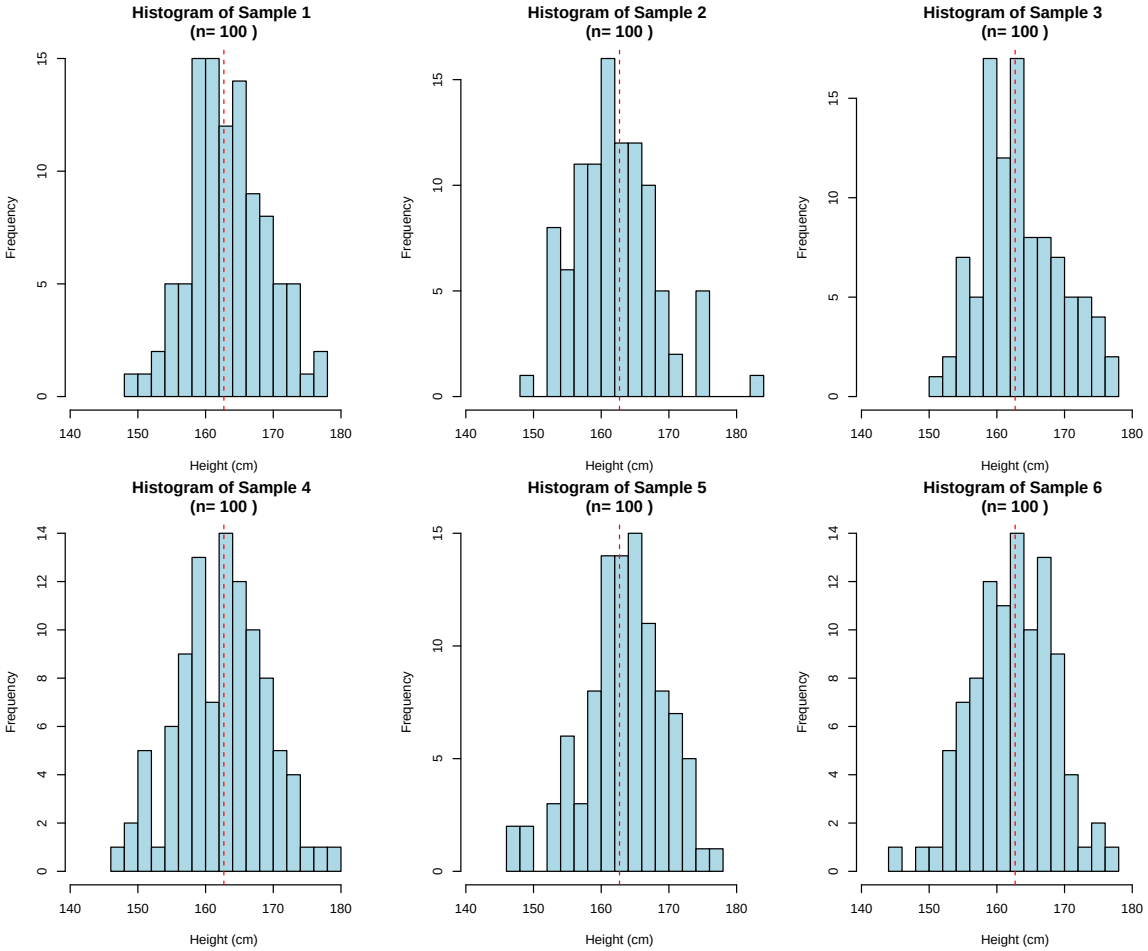
With the larger sample size of 40:

1. The samples appear more consistent and more normally distributed
2. There are fewer extreme outliers
3. The sample means (vertical bars) vary less around the true population mean
4. There is less apparent skewness and fewer gaps or clusters

This demonstrates the law of large numbers: as sample size increases, the sample characteristics tend to more closely match the true population characteristics.

Part (d): Histograms of Samples of Size 100

```r
# Function to generate histogram for a sample
generate_histograms <- function(n_samples, sample_size, mu, sigma) {
  # Set up a grid of plots
  par(mfrow = c(2, 3), mar = c(4, 4, 3, 1))

  # Common x-axis limits for comparison
  x_lim <- c(mu - 3.5*sigma, mu + 3.5*sigma)

  # Generate and plot histograms for each sample
  for (i in 1:n_samples) {
    sample_heights <- rnorm(sample_size, mean = mu, sd = sigma)
    hist(sample_heights,
         breaks = 12,
         xlim = x_lim,
         main = paste("Histogram of Sample", i, "\n(n=", sample_size, ")"),
         xlab = "Height (cm)",
         col = "lightblue")
    abline(v = mu, col = "red", lty = 2)
  }
}

# Generate histograms for 6 samples of size 100
set.seed(123)
generate_histograms(6, 100, mu, sigma)
```

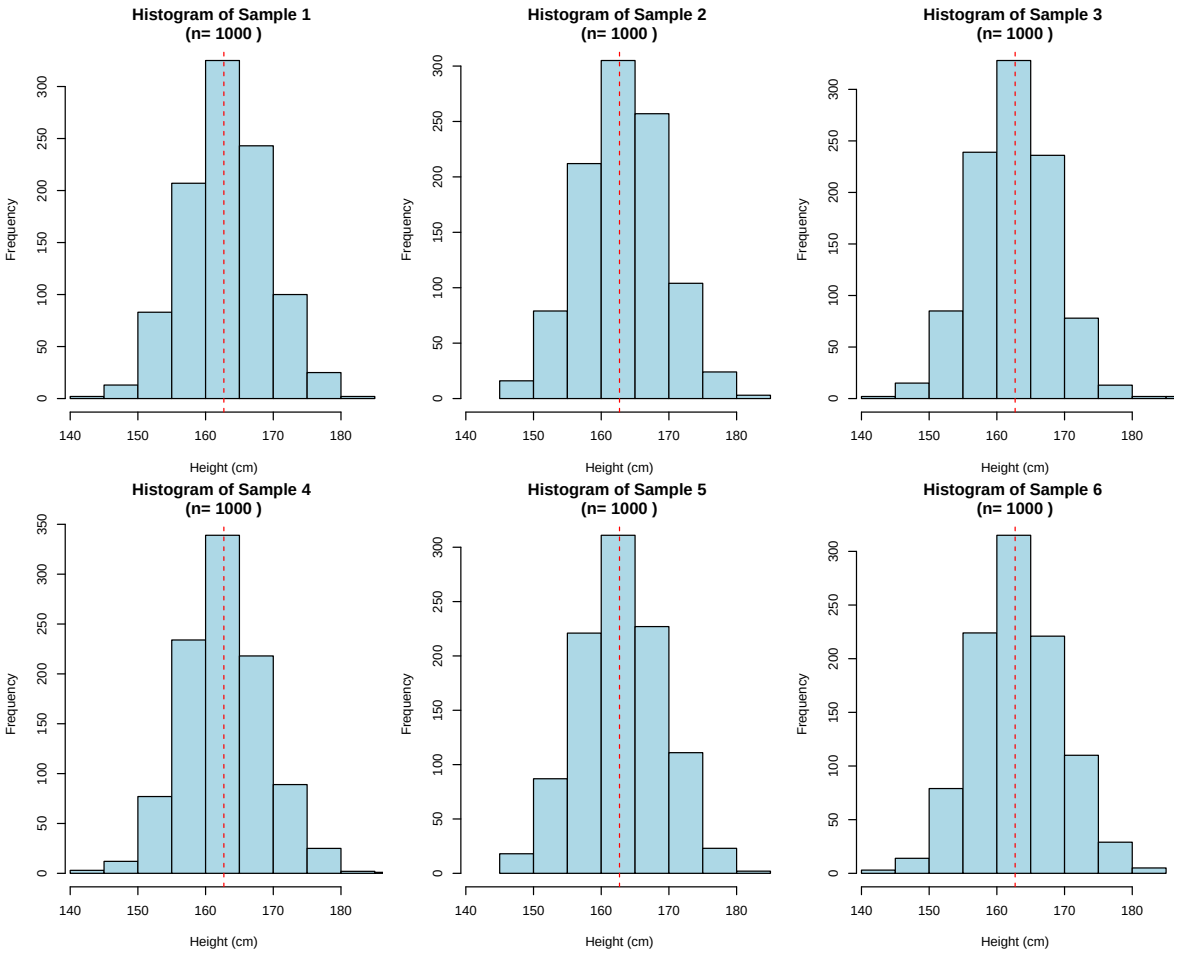


With samples of size 100, we observe:

1. The histograms begin to show a bell-shaped pattern resembling the normal distribution
2. There is still considerable variation in the exact shape from sample to sample
3. Some histograms appear slightly asymmetric despite coming from a symmetric normal distribution
4. The centers of the distributions tend to be close to the true population mean

Part (e): Histograms of Samples of Size 1000

```
# Generate histograms for 6 samples of size 1000
set.seed(456)
generate_histograms(6, 1000, mu, sigma)
```



With samples of size 1000:

1. All histograms closely approximate the bell-shaped normal curve
2. There is much less variation in shape between samples
3. The centers of all distributions are very close to the true population mean
4. The spread of each distribution is more consistent
5. Any asymmetry or irregular features are much less pronounced

The key difference between parts (d) and (e) is that with larger samples, the empirical distributions much more closely and consistently approximate the theoretical normal distribution. This is a demonstration of the central limit theorem, which states that as sample size increases, the sampling distribution approaches normality, regardless of the shape of the original population distribution (though in this case, the original distribution was already normal).

This exercise illustrates why we need to be cautious when interpreting patterns in small samples, as random variability can produce apparent features that aren't representative of the true population.