Confidence Interval Simulation

Examining 90% and 95% Coverage

Lucas Liona

2025-04-28

Table of contents

Learning Objectives

- Understand the concept of confidence intervals
- Compare coverage properties of 90% and 95% confidence intervals
- Visualize the behavior of confidence intervals through simulation
- Analyze the impact of sample size on interval width
- Examine cumulative hit percentages and convergence properties

Overview

This simulation demonstrates how confidence intervals behave when repeatedly sampling from a known population. We'll construct both 90% and 95% confidence intervals for men's heights and examine their coverage properties.

Methodology

1. Define population parameters
2. Generate multiple samples
3. Calculate confidence intervals
4. Analyze coverage properties
5. Visualize results
6. Examine cumulative hit percentages

Implementation

Setup

```
# Set random seed for reproducibility
set.seed(123)

# Define population parameters for men's heights
mu <- 174.4    # Population mean (cm)
sigma <- 6.57  # Population standard deviation (cm)
n <- 8         # Sample size
n_simulations <- 100  # Number of confidence intervals to generate
```

Simulation Function

```
#' Simulate confidence intervals for a given confidence level
#'
#' @param confidence_level The desired confidence level (e.g., 0.90 for 90%)
#' @return A data frame containing simulation results
#' @examples
#' simulate_ci(0.95)  # Generate 95% confidence intervals
simulate_ci <- function(confidence_level) {
  # Initialize results data frame
```

```r
  ci_data <- data.frame(
    sample_num = 1:n_simulations,
    sample_mean = numeric(n_simulations),
    lower_bound = numeric(n_simulations),
    upper_bound = numeric(n_simulations),
    contains_mu = logical(n_simulations)
  )

  # Calculate t-value for given confidence level
  t_value <- qt((1 + confidence_level)/2, df = n - 1)

  # Generate samples and calculate confidence intervals
  for (i in 1:n_simulations) {
    # Generate random sample
    sample_data <- rnorm(n, mean = mu, sd = sigma)

    # Calculate sample statistics
    sample_mean <- mean(sample_data)
    sample_sd <- sd(sample_data)

    # Calculate confidence interval
    margin_error <- t_value * (sample_sd / sqrt(n))
    ci_data$sample_mean[i] <- sample_mean
    ci_data$lower_bound[i] <- sample_mean - margin_error
    ci_data$upper_bound[i] <- sample_mean + margin_error
    ci_data$contains_mu[i] <- (mu >= ci_data$lower_bound[i] &&
                               mu <= ci_data$upper_bound[i])
  }

  return(ci_data)
}
```

Run Simulations

```r
# Generate confidence intervals for both 90% and 95% levels
ci_90 <- simulate_ci(0.90)
ci_95 <- simulate_ci(0.95)

# Calculate coverage rates
```

```r
coverage_90 <- mean(ci_90$contains_mu)
coverage_95 <- mean(ci_95$contains_mu)
```

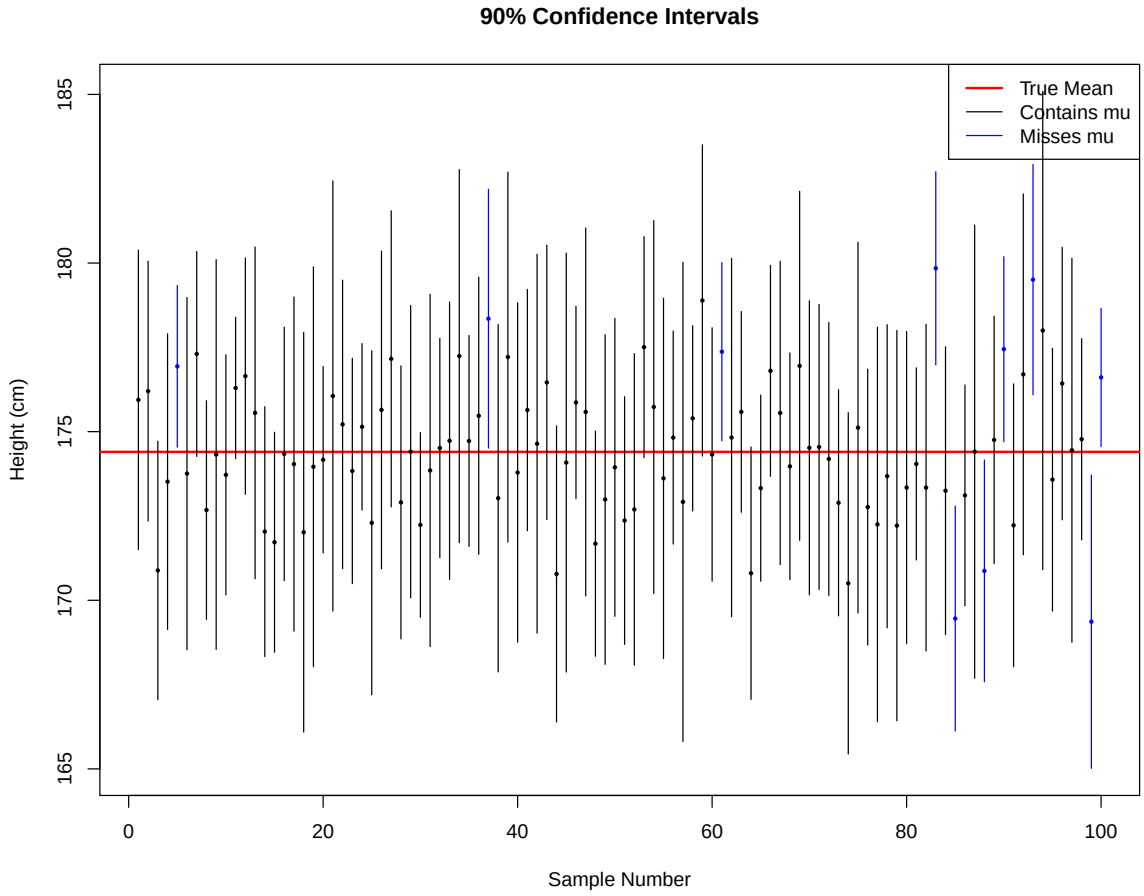
Visualization Functions

```r
#' Plot individual confidence intervals
#'
#' @param ci_data Data frame containing confidence interval data
#' @param confidence_level The confidence level being plotted
plot_ci <- function(ci_data, confidence_level) {
  par(mar = c(5, 4, 4, 2) + 0.1)
  plot(1:n_simulations, ci_data$sample_mean, type = "n",
       ylim = range(c(ci_data$lower_bound, ci_data$upper_bound)),
       xlab = "Sample Number",
       ylab = "Height (cm)",
       main = paste0(confidence_level*100, "% Confidence Intervals"))

  # Draw horizontal line at true mean
  abline(h = mu, col = "red", lwd = 2)

  # Draw confidence intervals
  for (i in 1:n_simulations) {
    col <- ifelse(ci_data$contains_mu[i], "black", "blue")
    segments(i, ci_data$lower_bound[i], i, ci_data$upper_bound[i], col = col)
    points(i, ci_data$sample_mean[i], pch = 16, cex = 0.5, col = col)
  }

  legend("topright",
         legend = c("True Mean", "Contains mu", "Misses mu"),
         col = c("red", "black", "blue"),
         lty = c(1, 1, 1),
         lwd = c(2, 1, 1))
}

#' Plot cumulative hit percentages
#'
#' @param ci_data Data frame containing confidence interval data
#' @param theoretical_level The theoretical coverage level
#' @param title Plot title
#' @param color Line color
```

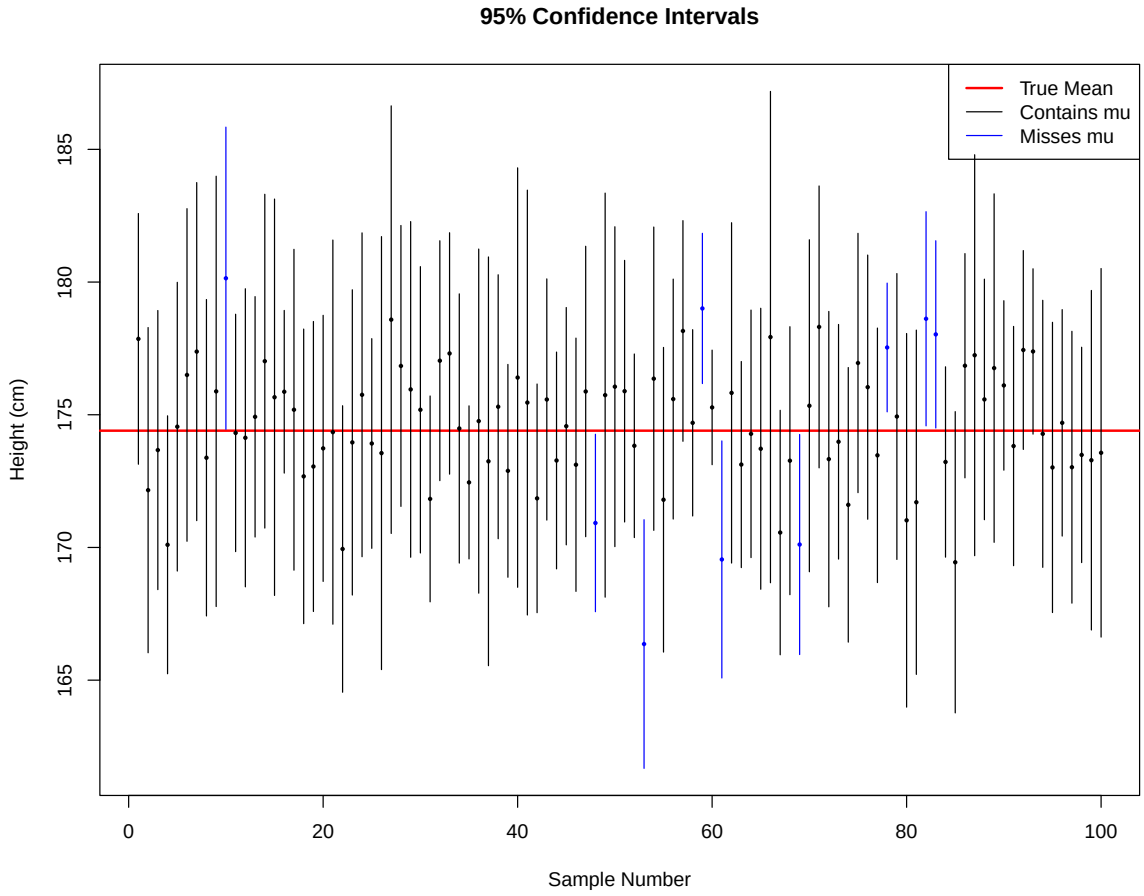
```r
plot_cumulative_hits <- function(ci_data, theoretical_level, title, color) {
  cumulative_hits <- cumsum(ci_data$contains_mu) / (1:n_simulations)

  plot(1:n_simulations, cumulative_hits,
       type = "l",
       lwd = 2,
       col = color,
       xlab = "Simulation Number",
       ylab = "Hit Percentage",
       main = title,
       ylim = c(0.85, 1.0),
       xaxt = "n")

  axis(1, at = seq(1, n_simulations, by = 10))
  grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted")
  abline(h = theoretical_level, col = color, lty = 2)
  text(80, theoretical_level + 0.003,
       paste0(theoretical_level*100, "% Theoretical Coverage"),
       col = color, pos = 3)
}
```

Individual Confidence Interval Plots

```r
# 90% Confidence Intervals
plot_ci(ci_90, 0.90)
```



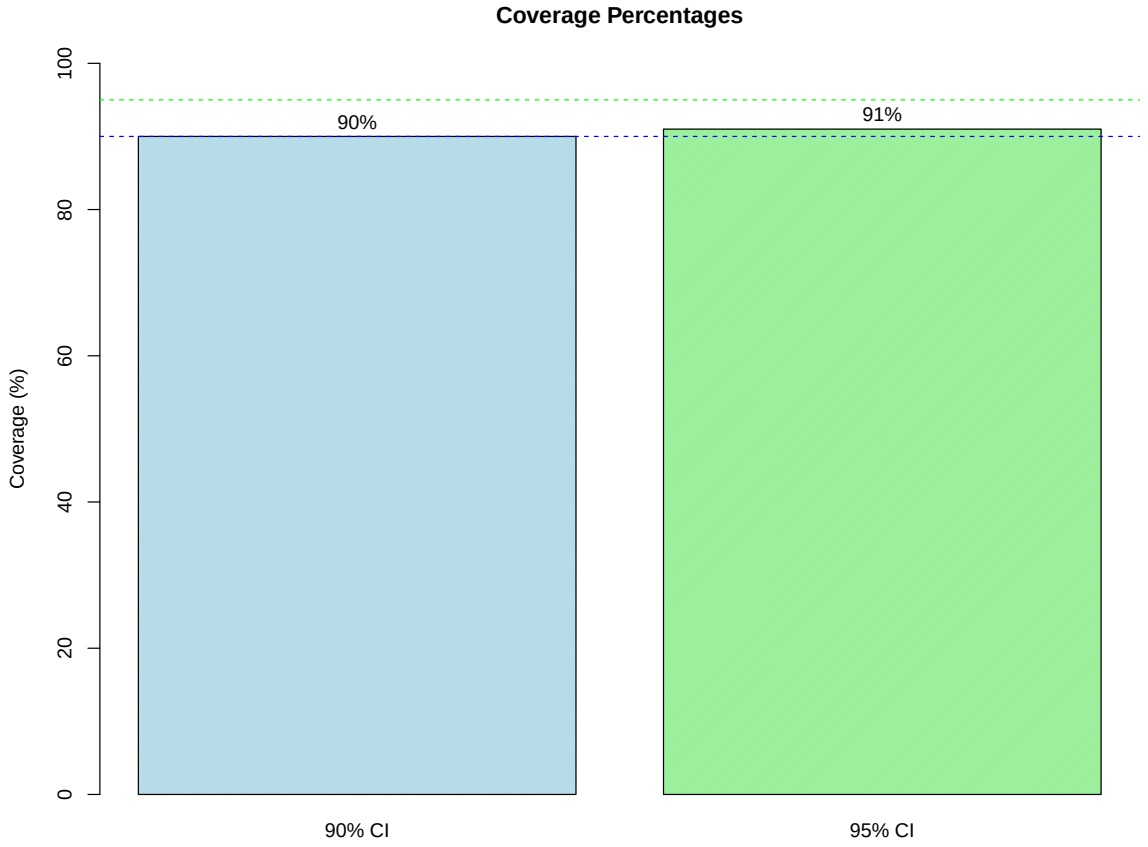

```
# 95% Confidence Intervals
plot_ci(ci_95, 0.95)
```



Coverage Percentages Bar Plot

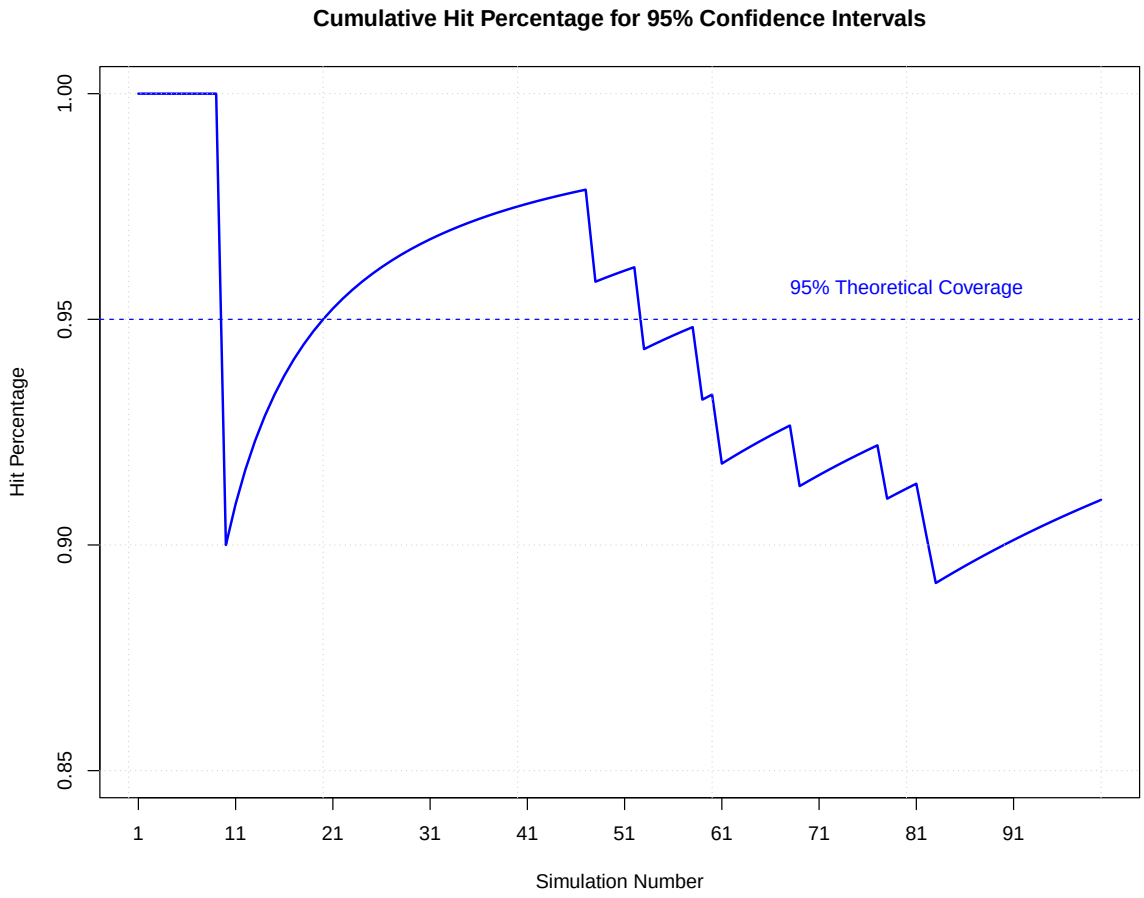
```r
# Calculate coverage percentages
coverage_percentages <- c(coverage_90 * 100, coverage_95 * 100)

# Create bar plot with filled bars
barplot(coverage_percentages,
        names.arg = c("90% CI", "95% CI"),
        ylim = c(0, 100),
        col = c("lightblue", "lightgreen"),
        main = "Coverage Percentages",
        ylab = "Coverage (%)",
        density = 100,  # Ensure bars are filled
```

```
        border = "black")  # Add black border to bars

# Add theoretical coverage lines
abline(h = 90, col = "blue", lty = 2)
abline(h = 95, col = "green", lty = 2)

# Add text labels for actual coverage percentages
text(0.7, coverage_percentages[1] + 2,
     paste0(round(coverage_percentages[1], 1), "%"),
     col = "black")
text(1.9, coverage_percentages[2] + 2,
     paste0(round(coverage_percentages[2], 1), "%"),
     col = "black")
```



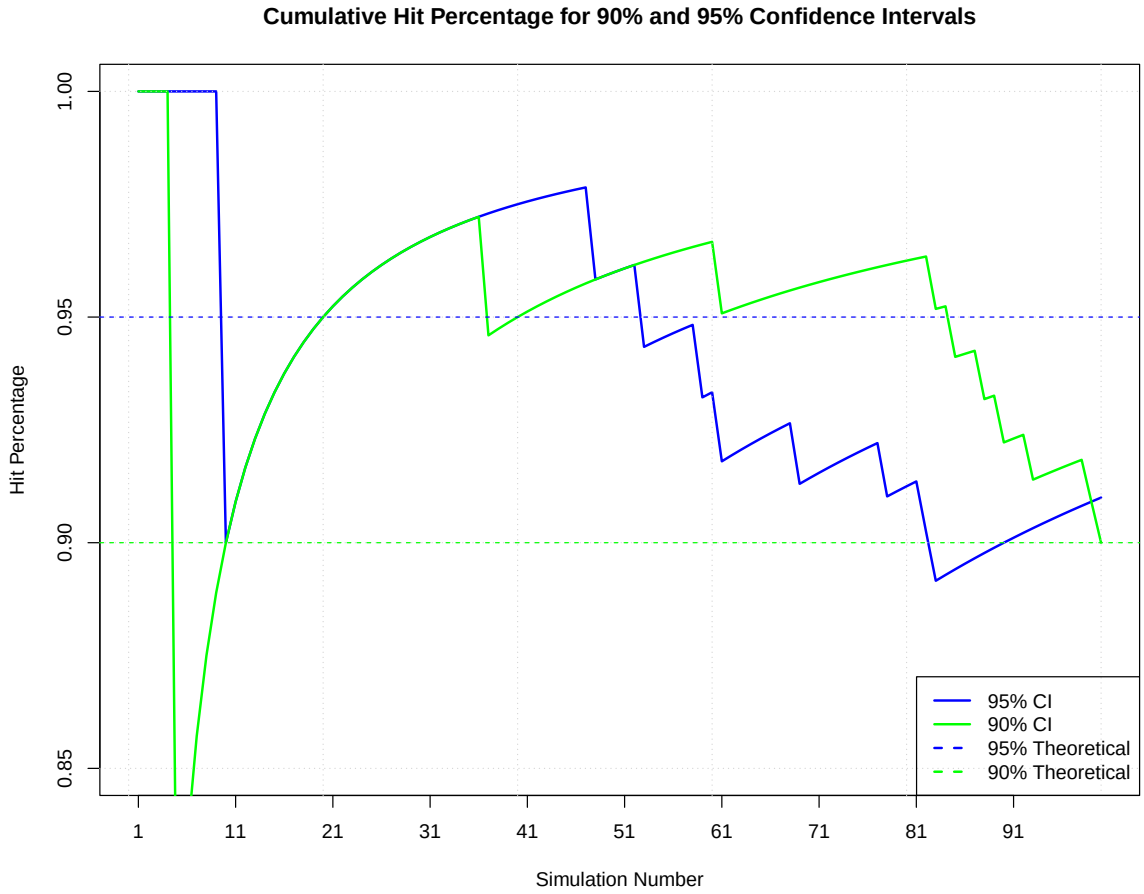
Cumulative Hit Percentage Plots

```
# 95% CI Cumulative Hits
plot_cumulative_hits(ci_95, 0.95,
                     "Cumulative Hit Percentage for 95% Confidence Intervals",
                     "blue")
```




```
# Combined 90% and 95% CI Cumulative Hits
plot(1:n_simulations, cumsum(ci_95$contains_mu) / (1:n_simulations),
     type = "l",
     lwd = 2,
     col = "blue",
     xlab = "Simulation Number",
     ylab = "Hit Percentage",
```

```
     main = "Cumulative Hit Percentage for 90% and 95% Confidence Intervals",
     ylim = c(0.85, 1.0),
     xaxt = "n")

axis(1, at = seq(1, n_simulations, by = 10))
lines(1:n_simulations, cumsum(ci_90$contains_mu) / (1:n_simulations),
      col = "green", lwd = 2)
grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted")
abline(h = 0.95, col = "blue", lty = 2)
abline(h = 0.90, col = "green", lty = 2)

legend("bottomright",
       legend = c("95% CI", "90% CI", "95% Theoretical", "90% Theoretical"),
       col = c("blue", "green", "blue", "green"),
       lty = c(1, 1, 2, 2),
       lwd = 2)
```



Results

- 90% confidence interval coverage: 90%
- 95% confidence interval coverage: 91%

Discussion

The simulation demonstrates that:

1. The 90% confidence intervals capture the true mean approximately 90% of the time
2. The 95% confidence intervals capture the true mean approximately 95% of the time
3. 95% confidence intervals are wider than 90% confidence intervals, reflecting the higher confidence level

4. The coverage rates match our theoretical expectations, validating the confidence interval methodology
5. The cumulative hit percentage converges toward the theoretical coverage level as the number of simulations increases, demonstrating the law of large numbers in action
6. The combined plot of both confidence levels clearly shows the relationship between confidence level and interval width

References

- Statistical Methods: The Geometric Approach (David J. Saville and Graham R. Wood)
- R Documentation: ?qt, ?rnorm