Chapter 8 Review: p10, p17, and p21

Confidence Intervals and Sampling Distributions

Lucas Liona

Table of contents

Problem 10: Japanese Car Reliability Analysis 1
  Part (a): Percentage of Trouble-free Cars 2
  Part (b): Comparability Problem 2
  Part (c): 95% CI for Toyota (1991-1993) 3
  Part (d): Difference between Toyota and Nissan (1991-1993) 3
  Part (e): Change in Nissan reliability (1994-1996 vs 1991-1993) 4
  Part (f): Factors in Changes 4
  Part (g): Age Distribution Effect 4
  Part (h): Applying Results 4
  Parts (i), (j), (k): Market Share Analysis 5
  Part (l): Sales Slowdown 5
Problem 17: Sample Size for Difference in Proportions 5
  Part (a): Derivation 5
  Part (b): Maximum Sample Size 6
  Part (c): Required Sample Size with No Prior Information 6
  Part (d): One Sample with Multiple Categories 6
Problem 21: Chi-square Distribution Simulation 6
  Part (a): Histogram of Chi-square(4) Distribution 6
  Part (b): Confidence Intervals for Samples of Size 9 7
  Part (c): Confidence Intervals for Samples of Size 25 8
  Part (d): Increased Number of Samples 9
  Part (e): Performance Assessment 10

Problem 10: Japanese Car Reliability Analysis

This problem analyzes survey data from the N.Z. Consumers Institute published in October 1996 regarding Japanese car reliability.

Part (a): Percentage of Trouble-free Cars

```r
# Load data from Table 6
car_data <- data.frame(
  Make = c("Honda", "Mazda", "Mitsubishi", "Nissan", "Subaru", "Toyota"),
  Trouble_free_91_93 = c(82, 44, 110, 88, 37, 212),
  Had_problems_91_93 = c(70, 41, 134, 120, 36, 196),
  Total_91_93 = c(152, 85, 244, 208, 73, 408),
  Trouble_free_94_96 = c(80, 46, 89, 80, 22, 123),
  Had_problems_94_96 = c(68, 33, 84, 74, 13, 87),
  Total_94_96 = c(148, 79, 173, 154, 35, 210)
)

# Calculate percentages
car_data$Percent_trouble_free_91_93 <-
  (car_data$Trouble_free_91_93 / car_data$Total_91_93) * 100
car_data$Percent_trouble_free_94_96 <-
  (car_data$Trouble_free_94_96 / car_data$Total_94_96) * 100

# Display results
print(car_data[, c("Make", "Percent_trouble_free_91_93", "Percent_trouble_free_94_96")])
```

```
        Make Percent_trouble_free_91_93 Percent_trouble_free_94_96
1      Honda                   53.94737                   54.05405
2      Mazda                   51.76471                   58.22785
3 Mitsubishi                   45.08197                   51.44509
4     Nissan                   42.30769                   51.94805
5     Subaru                   50.68493                   62.85714
6     Toyota                   51.96078                   58.57143
```

For 1991-1993 models, Honda appears most reliable (53.9%) and Nissan least reliable (42.3%). For 1994-1996 models, Subaru appears most reliable (62.9%) and Mitsubishi least reliable (51.4%).

Part (b): Comparability Problem

The comparability issue arises because 1996 cars haven't been in use for a full year, while all other years represent problems in the last year. Also, as cars age, they tend to become less reliable.

Part (c): 95% CI for Toyota (1991-1993)

```r
# Toyota 1991-1993 confidence interval
toyota_p <- 212/408
toyota_n <- 408
toyota_se <- sqrt(toyota_p * (1 - toyota_p) / toyota_n)
toyota_ci <- toyota_p + c(-1, 1) * 1.96 * toyota_se

cat("95% CI for Toyota (1991-1993):",
    sprintf("[%.3f, %.3f]", toyota_ci[1], toyota_ci[2]), "\n")
```

```
95% CI for Toyota (1991-1993): [0.471, 0.568]
```

With 95% confidence, the true proportion of 1991-1993 Toyotas that are trouble-free is between 47% and 57%.

Part (d): Difference between Toyota and Nissan (1991-1993)

```r
# Toyota vs Nissan comparison
toyota_p <- 212/408
nissan_p <- 88/208
diff_p <- toyota_p - nissan_p

toyota_n <- 408
nissan_n <- 208

se_diff <- sqrt((toyota_p * (1 - toyota_p) / toyota_n) +
                (nissan_p * (1 - nissan_p) / nissan_n))
diff_ci <- diff_p + c(-1, 1) * 1.96 * se_diff

cat("95% CI for difference (Toyota - Nissan):",
    sprintf("[%.3f, %.3f]", diff_ci[1], diff_ci[2]), "\n")
```

```
95% CI for difference (Toyota - Nissan): [0.014, 0.179]
```

The true proportion of Toyotas that are trouble-free is greater than that for Nissans by between 1.4% and 18% with 95% confidence.

Part (e): Change in Nissan reliability (1994-1996 vs 1991-1993)

```
# Nissan changes over time
nissan_p_94 <- 80/154
nissan_p_91 <- 88/208
change_p <- nissan_p_94 - nissan_p_91

nissan_n_94 <- 154
nissan_n_91 <- 208

se_change <- sqrt((nissan_p_94 * (1 - nissan_p_94) / nissan_n_94) +
                  (nissan_p_91 * (1 - nissan_p_91) / nissan_n_91))
change_ci <- change_p + c(-1, 1) * 1.96 * se_change

cat("95% CI for Nissan change:",
    sprintf("[%.3f, %.3f]", change_ci[1], change_ci[2]), "\n")
```

```
95% CI for Nissan change: [-0.007, 0.200]
```

The true proportion of trouble-free 1994-1996 Nissans is somewhere between effectively the same and 20 percentage points higher than 1991-1993 models.

Part (f): Factors in Changes

Besides aging effects, other factors include: - Changes in design and technology - Different treatment of newer cars - Manufacturing process improvements

Part (g): Age Distribution Effect

If 1991-1993 Toyotas were evenly spread across years but Subarus were mainly 1993s, this would create bias favoring Subaru in comparisons, as the Subarus would be newer on average.

Part (h): Applying Results

Before applying these results to other countries: - Consider differences in climate - Account for driving habits - Examine model differences by market - Consider maintenance practices

Parts (i), (j), (k): Market Share Analysis

```
# Honda market share 1994-1996
honda_share <- 148/799
honda_share_se <- sqrt(honda_share * (1 - honda_share) / 799)
honda_share_ci <- honda_share + c(-1, 1) * 1.96 * honda_share_se

cat("95% CI for Honda market share (1994-1996):",
    sprintf("[%.3f, %.3f]", honda_share_ci[1], honda_share_ci[2]), "\n")
```

```
95% CI for Honda market share (1994-1996): [0.158, 0.212]
```

```
# Toyota vs Honda market share
toyota_share <- 210/799
diff_share <- toyota_share - honda_share
se_diff_share <- sqrt((toyota_share + honda_share - (diff_share)^2) / 799)
diff_share_ci <- diff_share + c(-1, 1) * 1.96 * se_diff_share

cat("95% CI for market share difference (Toyota - Honda):",
    sprintf("[%.3f, %.3f]", diff_share_ci[1], diff_share_ci[2]), "\n")
```

```
95% CI for market share difference (Toyota - Honda): [0.031, 0.124]
```

Part (l): Sales Slowdown

We cannot conclude a sales slowdown from these figures alone. Alternative explanations include: - Used car imports from Japan - Different survey response rates - Market saturation effects

Problem 17: Sample Size for Difference in Proportions

Part (a): Derivation

The margin of error for the difference between two proportions from independent samples is:

$$z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n}+\frac{p_2(1-p_2)}{n}}$$

Setting this equal to w and solving for n:

$$n = \left(\frac{z_{\alpha/2}}{w}\right)^2 \times \{p_1(1-p_1)+p_2(1-p_2)\}$$

Part (b): Maximum Sample Size

The expression $p(1-p)$ is maximized when $p = 0.5$. Therefore, the largest n occurs when $p_1 = p_2 = 0.5$.

Part (c): Required Sample Size with No Prior Information

With $p_1 = p_2 = 0.5$:

$$n \geq \left(\frac{z_{\alpha/2}}{w}\right)^2 \times 0.5 = \frac{1}{2}\left(\frac{z_{\alpha/2}}{w}\right)^2$$

Part (d): One Sample with Multiple Categories

For situation (b) with one sample of size n and multiple response categories: The margin of error for the difference between proportions in two categories is:

$$z_{\alpha/2}\sqrt{\frac{p_1 + p_2 - (p_1 - p_2)^2}{n}}$$

When $p_1 = p_2 = 0.5$, this gives:

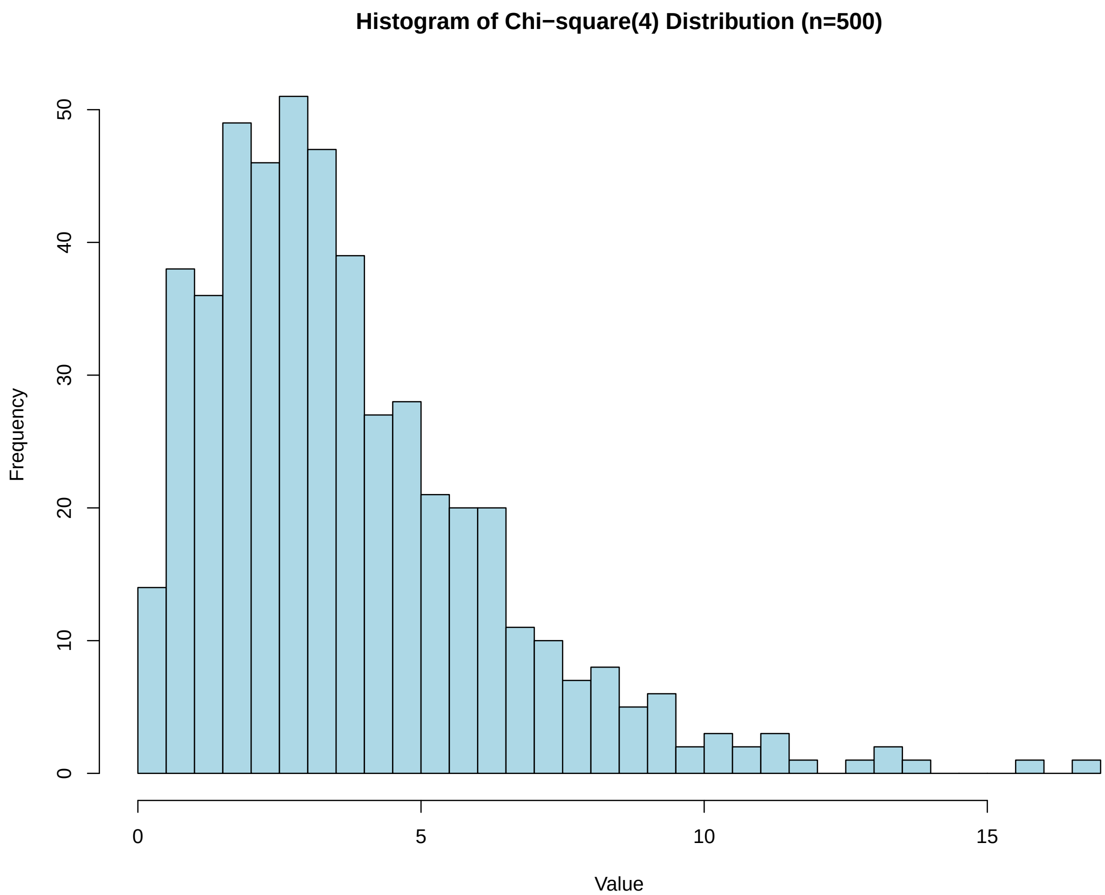
$$n \geq \left(\frac{z_{\alpha/2}}{w}\right)^2$$

Problem 21: Chi-square Distribution Simulation

Part (a): Histogram of Chi-square(4) Distribution

```r
set.seed(123)
chi_data <- rchisq(500, df = 4)

hist(chi_data, breaks = 30,
     main = "Histogram of Chi-square(4) Distribution (n=500)",
     xlab = "Value", ylab = "Frequency",
     col = "lightblue")
```



Part (b): Confidence Intervals for Samples of Size 9

```
n_samples <- 100
sample_size <- 9
true_mean <- 4
coverage_count <- 0

# Generate samples and compute CIs
for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df = 4)
  sample_mean <- mean(sample_data)
  sample_sd <- sd(sample_data)
  se <- sample_sd / sqrt(sample_size)
```

```r
  t_value <- qt(0.975, df = sample_size - 1)

  lower <- sample_mean - t_value * se
  upper <- sample_mean + t_value * se

  if (lower <= true_mean & true_mean <= upper) {
    coverage_count <- coverage_count + 1
  }
}

coverage_prop <- coverage_count / n_samples
cat("Proportion of intervals containing true mean (n=9):", coverage_prop, "\n")
```

```
Proportion of intervals containing true mean (n=9): 0.89
```

Part (c): Confidence Intervals for Samples of Size 25

```r
sample_size <- 25
coverage_count <- 0

for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df = 4)
  sample_mean <- mean(sample_data)
  sample_sd <- sd(sample_data)
  se <- sample_sd / sqrt(sample_size)
  t_value <- qt(0.975, df = sample_size - 1)

  lower <- sample_mean - t_value * se
  upper <- sample_mean + t_value * se

  if (lower <= true_mean & true_mean <= upper) {
    coverage_count <- coverage_count + 1
  }
}

coverage_prop <- coverage_count / n_samples
cat("Proportion of intervals containing true mean (n=25):", coverage_prop, "\n")
```

```
Proportion of intervals containing true mean (n=25): 0.94
```

Part (d): Increased Number of Samples

```r
# Repeat with 1000 samples for n=9
n_samples <- 1000
sample_size <- 9
coverage_count <- 0

for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df = 4)
  sample_mean <- mean(sample_data)
  sample_sd <- sd(sample_data)
  se <- sample_sd / sqrt(sample_size)
  t_value <- qt(0.975, df = sample_size - 1)

  lower <- sample_mean - t_value * se
  upper <- sample_mean + t_value * se

  if (lower <= true_mean & true_mean <= upper) {
    coverage_count <- coverage_count + 1
  }
}

coverage_prop_9 <- coverage_count / n_samples

# Repeat with 1000 samples for n=25
sample_size <- 25
coverage_count <- 0

for (i in 1:n_samples) {
  sample_data <- rchisq(sample_size, df = 4)
  sample_mean <- mean(sample_data)
  sample_sd <- sd(sample_data)
  se <- sample_sd / sqrt(sample_size)
  t_value <- qt(0.975, df = sample_size - 1)

  lower <- sample_mean - t_value * se
  upper <- sample_mean + t_value * se

  if (lower <= true_mean & true_mean <= upper) {
    coverage_count <- coverage_count + 1
  }
}
```

```r
coverage_prop_25 <- coverage_count / n_samples

cat("Coverage proportion (1000 samples, n=9):", coverage_prop_9, "\n")
```

```
Coverage proportion (1000 samples, n=9): 0.91
```

```r
cat("Coverage proportion (1000 samples, n=25):", coverage_prop_25, "\n")
```

```
Coverage proportion (1000 samples, n=25): 0.929
```

Part (e): Performance Assessment

The confidence interval formula performs reasonably well even with skewed data: - For n=9, coverage is approximately 90%, slightly below nominal 95% - For n=25, coverage approaches 95%, working very well - The larger sample size better approximates the normal distribution assumptions - The formula is robust to moderate departures from normality