Final Exam

Han Nguyen - TXN200004

12/15/2024

Problem 1

Question 1.1

Suppose X has a Poisson distribution with an expected value of 8.9. What is the probability its value is at least as big as 7 but not bigger than 9?

```
# Calculate P(7 <= X <= 9) for Poisson distribution with lambda = 8.9
lambda <- 8.9

# P(7 <= X <= 9) = P(X <= 9) - P(X <= 6)
prob_1.1 <- ppois(9, lambda) - ppois(6, lambda)
prob_1.1
```

```
## [1] 0.3845391
```

The probability is 0.3845391.

Question 1.2

Calculate the 95% confidence interval of the mean for the COVID-19 daily deaths data.

```
# Data: 35 daily deaths
covid_deaths <- c(8, 12, 6, 12, 8, 8, 12, 5, 9, 10, 6, 8, 6, 16, 7, 13, 5, 10,
                  5, 7, 9, 7, 8, 8, 9, 10, 8, 9, 9, 8, 8, 6, 10, 9, 17)

# Sample statistics
n <- length(covid_deaths)
xbar <- mean(covid_deaths)
s <- sd(covid_deaths)

# Standard error
SE <- s / sqrt(n)

# Critical value for 95% CI (two-tailed)
# Using t-distribution with n-1 degrees of freedom
alpha <- 0.05
t_critical <- qt(1 - alpha/2, df = n - 1)

# Confidence interval
lower <- xbar - t_critical * SE
upper <- xbar + t_critical * SE

cat("Sample mean:", xbar, "\n")
```

```
## Sample mean: 8.8
```

```
cat("Sample standard deviation:", s, "\n")
```

```
## Sample standard deviation: 2.78441
```

```
cat("Standard error:", SE, "\n")
```

```
## Standard error: 0.4706513
```

```
cat("95% Confidence Interval: [", lower, ",", upper, "]\n")
```

```
## 95% Confidence Interval: [ 7.843522 , 9.756478 ]
```

The 95% confidence interval for the mean is [7.844, 9.756].

Question 1.3

Perform a Z-test at the 0.05 significance level with H_a: lambda > 8.1.

Why we can use the Z-test: We can use the Z-test here because the sample size (n = 35) is large enough for the Central Limit Theorem to apply. For a Poisson distribution, the sample mean is approximately normally distributed with mean lambda and variance lambda/n when n is large.

Step 1: Specify model for the data

Model: $X_i \sim \text{Poisson}(\lambda)$, $i = 1, 2, ..., 35$, where observations are independent and identically distributed.

Step 2: State Null and Alternative Hypothesis

- $H_0 : \lambda = 8.1$
- $H_a : \lambda > 8.1$ (right-tailed test)

Step 3: Specify Test Statistic

For Poisson distribution with large n, by the Central Limit Theorem:

$$Z = \frac{\bar{X} - \lambda_0}{\sqrt{\lambda_0/n}}$$

where $\lambda_0 = 8.1$ is the value under the null hypothesis.

Step 4: State α level

$\alpha = 0.05$

```
alpha <- 0.05
```

Step 5: Calculate test statistic

```
# Data from Question 1.2 (covid_deaths is already defined above)
lambda_0 <- 8.1
n <- length(covid_deaths)
xbar <- mean(covid_deaths)

# For Poisson: mean = variance = lambda
# Standard error under H_0: sqrt(lambda_0/n)
SE_null <- sqrt(lambda_0 / n)
z_stat <- (xbar - lambda_0) / SE_null
```

- Sample mean: $\bar{x} = 8.8$
- $\lambda_0 = 8.1$

- $n = 35$
- Standard error under H_0: $SE = \sqrt{\lambda_0/n} = 0.4811$
- Z-statistic: $Z = \frac{8.8-8.1}{0.4811} = 1.4551$

Step 6: Compute p-value

```
# For right-tailed test: P(Z > z_stat)
p_value <- 1 - pnorm(z_stat)
```

For a right-tailed test: $P(Z > 1.4551) = 0.0728$

Step 7: Interpret Results

Since p-value $= 0.0728 \geq \alpha = 0.05$, we fail to reject H_0.

There is insufficient evidence to conclude that $\lambda > 8.1$.

Problem 2

Email data for student and instructor from Sunday to Saturday:

```
# Create the data table
email_data <- matrix(c(9, 19, 12, 18, 11, 15, 10,
                       20, 17, 23, 14, 23, 20, 20),
                     nrow = 2, byrow = TRUE,
                     dimnames = list(c("Student", "Instructor"),
                                     c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")))

# Display the table
print(email_data)
```

```
##            Sun Mon Tue Wed Thu Fri Sat
## Student      9  19  12  18  11  15  10
## Instructor  20  17  23  14  23  20  20
```

```
# Calculate totals
row_totals <- rowSums(email_data)
col_totals <- colSums(email_data)
grand_total <- sum(email_data)

cat("\nRow totals:", row_totals)
```

```
##
## Row totals: 94 137
```

```
cat("\nColumn totals:", col_totals)
```

```
##
## Column totals: 29 36 35 32 34 35 30
```

```
cat("\nGrand total:", grand_total, "\n")
```

```
##
## Grand total: 231
```

Question 2.1

Calculate the expected counts matrix.

```r
# Calculate expected counts
# Expected Count = (Row Total * Column Total) / Grand Total

expected_counts <- matrix(0, nrow = 2, ncol = 7,
                          dimnames = list(c("Student", "Instructor"),
                                          c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")))

for (i in 1:2) {
  for (j in 1:7) {
    expected_counts[i, j] <- (row_totals[i] * col_totals[j]) / grand_total
  }
}

print(expected_counts)
```

```
##                 Sun      Mon      Tue      Wed     Thu      Fri      Sat
## Student    11.80087 14.64935 14.24242 13.02165 13.8355 14.24242 12.20779
## Instructor 17.19913 21.35065 20.75758 18.97835 20.1645 20.75758 17.79221
```

Question 2.2

Using a chi-squared test of independence at alpha = 0.05, test whether the number of emails received is independent of whether it's a professor or a student.

Step 1: Specify model for the data

The data follows a contingency table (2x7) with counts from a multinomial distribution. Row and column totals are fixed.

Step 2: State Null and Alternative Hypothesis

- H_0: The number of emails received is independent of person type (student vs instructor)
- H_a: The number of emails received is NOT independent of person type

Step 3: Specify Test Statistic

Chi-squared test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} = observed count in cell (i, j) and E_{ij} = expected count in cell (i, j).

Step 4: State α level

$\alpha = 0.05$

```r
alpha <- 0.05
```

Step 5: Calculate test statistic

```r
observed <- email_data
expected <- expected_counts

# Calculate chi-squared statistic
chi_sq_stat <- sum((observed - expected)^2 / expected)

# Degrees of freedom: (r - 1) * (c - 1)
df <- (nrow(observed) - 1) * (ncol(observed) - 1)
```

- Chi-squared statistic: $\chi^2 = 8.825$
- Degrees of freedom: $df = (2-1) \times (7-1) = 6$

Step 6: Compute p-value

```
p_value <- 1 - pchisq(chi_sq_stat, df)
```

$P(\chi^2_6 > 8.825) = 0.1837$

Step 7: Interpret Results

Since p-value $= 0.1837 \geq \alpha = 0.05$, we fail to reject H_0.

There is insufficient evidence to reject independence. We do not have enough evidence to conclude that the number of emails received depends on whether the person is a student or an instructor.

Problem 3

Question 3.1

What are the null hypotheses corresponding to the t-statistics (t-value) shown in the regression output?

Answer:

For the linear regression model: divorce $= \beta_0 + \beta_1$ * femlab $+ \epsilon$

The null hypotheses corresponding to the t-statistics are:

- For the intercept (β_0): $H_0 : \beta_0 = 0$
- For the slope (β_1): $H_0 : \beta_1 = 0$

These test whether each coefficient is significantly different from zero.

Question 3.2

Based on the output, interpret the result in terms of slope. Based on the p-values, would you conclude that the univariate relationship is statistically significant at alpha = 0.01?

Answer:

Slope interpretation: The estimated slope coefficient is $\hat{\beta}_1 = 0.43867$. This means that for each 1% increase in female labor force participation, the divorce rate increases by approximately 0.439 divorces per 1,000 women aged 15+, on average.

Statistical significance: The p-value for the slope is p < 2e-16, which is much less than $\alpha = 0.01$. Therefore, we conclude that the univariate relationship between female labor market participation and the divorce rate is statistically significant at the 0.01 significance level. We reject the null hypothesis that $\beta_1 = 0$ and conclude there is a significant positive relationship between female labor force participation and divorce rate.

Question 3.3

Test the hypothesis that if female labor force participation increases 5%, we would see over 2 more divorces per 1,000 women.

Step 1: Specify model for the data

Simple linear regression model: $\text{divorce}_i = \beta_0 + \beta_1 \times \text{femlab}_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Step 2: State Null and Alternative Hypothesis

The claim: A 5% increase in femlab leads to over 2 more divorces, meaning $5 \times \beta_1 > 2$, or $\beta_1 > 0.4$.

- $H_0 : \beta_1 = 0.4$ (5% increase leads to exactly 2 divorces)
- $H_a : \beta_1 > 0.4$ (5% increase leads to MORE than 2 divorces)

Step 3: Specify Test Statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

where $\beta_{1,0} = 0.4$ is the value under the null hypothesis.

Step 4: State α level

$\alpha = 0.01$

```
# Values from regression output
beta_1_hat <- 0.43867  # Estimated slope coefficient
se_beta_1 <- 0.02302   # Standard error of slope
df <- 75               # Degrees of freedom (from output)
alpha <- 0.01
```

Step 5: Calculate test statistic

```
beta_1_0 <- 0.4  # Value under null hypothesis
t_stat <- (beta_1_hat - beta_1_0) / se_beta_1
```

- $\hat{\beta}_1 = 0.43867$
- $\beta_{1,0} = 0.4$
- $SE(\hat{\beta}_1) = 0.02302$
- $t = \frac{0.43867 - 0.4}{0.02302} = 1.6798$
- Degrees of freedom: $df = 75$

Step 6: Compute p-value

```
# Right-tailed test
p_value <- 1 - pt(t_stat, df)
```

For a right-tailed test: $P(t_{75} > 1.6798) = 0.0486$

Step 7: Interpret Results

Since p-value $= 0.0486 \geq \alpha = 0.01$, we fail to reject H_0.

There is insufficient evidence that a 5% increase in female labor force participation leads to more than 2 additional divorces per 1,000 women.

Question 3.4

Explain the agreement or disagreement between Questions 3.3 and 3.2.

Answer:

Question 3.2 tests whether the slope is significantly different from zero ($H_0 : \beta_1 = 0$), which establishes whether there is any relationship at all between female labor force participation and divorce rate. In Question 3.2, we rejected H_0 because the p-value ($<$ 2e-16) was much less than 0.01, confirming that a significant positive relationship exists.

Question 3.3 tests a more specific claim: whether the slope is greater than 0.4 ($H_0 : \beta_1 = 0.4$ vs $H_a : \beta_1 > 0.4$), which tests whether the relationship is stronger than the newspaper's claim. In Question 3.3, we fail to reject H_0 because the p-value is greater than 0.01.

These results are not contradictory. The estimated slope $\hat{\beta}_1 = 0.43867$ is close to 0.4, so while we can confidently say that the slope is different from zero (Question 3.2), we cannot confidently say it is greater

than 0.4 at the 0.01 significance level (Question 3.3). This means there is strong evidence of a relationship between female labor force participation and divorce rate, but insufficient evidence that the effect is as large as the newspaper claims (more than 2 divorces per 1,000 women for a 5% increase).

Problem 4

Write a function to study the empirical coverage of a confidence interval procedure.

```r
ci_coverage <- function(population, alpha, n, iterations) {
  # Step 1: Compute the true parameter from the full population
  truth <- mean(population)

  # Step 2: Create a vector to store 0/1 indicators
  hits <- numeric(iterations)

  # Step 3: For each iteration, draw a sample and build a CI
  for (i in 1:iterations) {
    # 3a. Draw a random sample of size n from population
    # Note: Using replace = FALSE for finite population sampling
    # Each iteration draws a fresh sample without replacement
    sample_i <- sample(population, size = n, replace = FALSE)

    # 3b. Compute the sample mean and sample standard deviation
    xbar <- mean(sample_i)
    s <- sd(sample_i)

    # 3c. Compute the standard error of the mean
    SE <- s / sqrt(n)

    # 3d. Find the critical value for a (1 - alpha) CI
    # For large n, use the standard normal z_{1 - alpha/2}
    z_star <- qnorm(1 - alpha/2)

    # 3e. Construct the two-sided confidence interval
    lower <- xbar - z_star * SE
    upper <- xbar + z_star * SE

    # 3f. Check if the true mean is inside the interval
    if (truth >= lower && truth <= upper) {
      hits[i] <- 1
    } else {
      hits[i] <- 0
    }
  }

  # Step 4: After the loop, compute the empirical coverage
  coverage <- mean(hits)

  # Step 5: Return the coverage proportion
  return(coverage)
}
```

Question 4.1

Run the function with gamma distribution parameters and report results.

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

# Generate population
population_gamma <- rgamma(10000, shape = 24, rate = 5)

# Run the coverage function
coverage_4.1 <- ci_coverage(population = population_gamma,
                            alpha = 0.05,
                            n = 100,
                            iterations = 10000)

cat("Coverage proportion:", coverage_4.1, "\n")
```

```
## Coverage proportion: 0.948
```

```
cat("Expected coverage (1 - alpha):", 1 - 0.05, "\n")
```

```
## Expected coverage (1 - alpha): 0.95
```

Explanation:

The coverage proportion should be close to 0.95 (1 - alpha = 1 - 0.05).

The gamma distribution with shape = 24 has a relatively symmetric shape (as shape increases, gamma approaches normal). With n = 100 (large sample size), the Central Limit Theorem ensures that the sampling distribution of the mean is approximately normal. Therefore, using the z-critical value for constructing the confidence interval is appropriate, and we expect the empirical coverage to closely mirror the nominal confidence level (95%).

If the coverage is close to 0.95, this confirms that our CI procedure is working correctly for this population.

Question 4.2

Run the function with Poisson distribution parameters and report results.

```
# Set seed for reproducibility
set.seed(456)

# Generate population
population_pois <- rpois(100000, lambda = 0.05)

# Run the coverage function
coverage_4.2 <- ci_coverage(population = population_pois,
                            alpha = 0.05,
                            n = 11,
                            iterations = 10000)

cat("Coverage proportion:", coverage_4.2, "\n")
```

```
## Coverage proportion: 0.439
```

```
cat("Expected coverage (1 - alpha):", 1 - 0.05, "\n")
```

```
## Expected coverage (1 - alpha): 0.95
```

Explanation:

The coverage proportion is approximately 0.439, which is much lower than the expected 0.95 (1 - alpha). This demonstrates severe **undercoverage**.

This happens because:

1. **Extreme skewness**: The Poisson distribution with lambda = 0.05 is extremely sparse, with most values being 0 and occasional 1s. The population mean is only 0.05.
2. **Very small sample size**: With n = 11, the Central Limit Theorem does not apply well, especially for such a skewed distribution.
3. **Zero variance in many samples**: Many samples will consist entirely of zeros (or nearly all zeros), resulting in sample standard deviation s = 0. When s = 0, the standard error SE = 0, causing the confidence interval to collapse to a single point [xbar, xbar].
4. **Invalid normal approximation**: The large-sample normal-based CI procedure assumes the sampling distribution is approximately normal, which is severely violated here.

This demonstrates that the normal-based CI procedure can have severe undercoverage when the sample size is too small and the population distribution is far from normal, particularly when many samples have zero or near-zero variance.