POL501 - Problem Set 3

Tutorial and instructions to get started

2024-11-05

Table of Contents

# Preliminary Setup

## Confidence Intervals and Hypothesis Testing Explained Intuitively

### Confidence Interval

A **confidence interval** is a range of values used to estimate an unknown population parameter, such as the mean. This interval is constructed around a sample estimate, allowing us to understand the potential range in which the true population parameter lies with a certain level of confidence.

For example, if we say that we have a 95% confidence interval for the mean age of a group, we are essentially saying that if we were to take many samples and construct confidence intervals for each, approximately 95% of these intervals would contain the true mean age of the population.

Mathematically, the confidence interval for a mean is given by:

Where:

* : Sample mean
* : Critical value corresponding to the confidence level
* : Population standard deviation (or sample standard deviation if unknown)
* : Sample size

The critical value changes depending on the confidence level, which is often denoted as **1 -** . For a 95% confidence level, , and the critical value can be found using a Z or t-distribution.

### Hypothesis Test

A **hypothesis test** is a formal statistical method used to evaluate if there is enough evidence to support a specific claim about a population parameter. The test involves the following steps:

1. **Formulate Hypotheses**:

* **Null Hypothesis ()**: The statement of no effect or no difference. It represents the status quo or the hypothesis to be tested.
* **Alternative Hypothesis ()**: The statement you want to provide evidence for. It usually represents a difference or an effect.

1. **Select a Significance Level ()**: The probability of rejecting the null hypothesis when it is actually true. Common choices are 0.05 or 0.01.
2. **Calculate the Test Statistic**: This could be a z-statistic, t-statistic, etc., depending on the test.
3. **Determine the Rejection Region**: Compare the test statistic to a critical value to determine if we should reject .
4. **Make a Decision**: Based on the p-value and the significance level, determine if we reject or fail to reject .

### Rejection Region

The **rejection region** is the range of values for the test statistic that leads us to reject the null hypothesis. For example, if we set our significance level to 0.05 in a two-tailed test, the rejection region would be in both tails of the distribution, covering the most extreme 5% of values.

### Z-Statistic and Critical Value

The **z-statistic** is used to measure how many standard deviations an element is from the mean. It is given by the formula:

Where:

* : Sample mean
* : Population mean under the null hypothesis
* : Population standard deviation under the null hypothesis
* : Sample size

The **critical value** is the value that separates the rejection region from the acceptance region. If the z-statistic falls beyond this value, we reject the null hypothesis.

### P-Value

The **p-value** is the probability of obtaining a test statistic at least as extreme as the one observed, given that the null hypothesis is true. It tells us how likely the observed data is under the null hypothesis: - A small p-value (typically less than 0.05) indicates strong evidence against the null hypothesis, so we reject . - A large p-value indicates weak evidence against , so we fail to reject .

In summary, the p-value helps us determine whether our results are statistically significant or if they could have occurred by random chance.

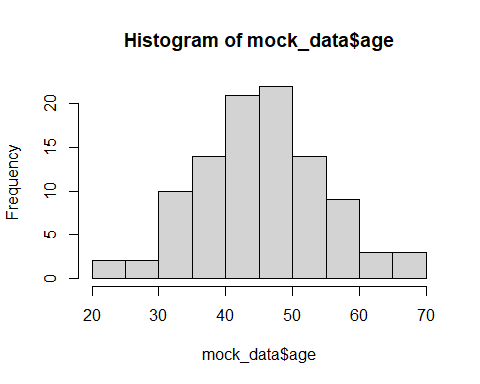
## Demonstration of the confidence\_interval Function to Compute Confidence Intervals

First, we define a user-defined function to compute confidence intervals.

# Function to calculate confidence interval  
confidence\_interval <- function(data, var\_name, confidence\_level = 0.95, method = "t") {  
  
 # Extract the data column  
 data\_column <- data[[var\_name]]  
   
 # Validate inputs (method, var\_name)  
 if (!(method %in% c("t", "z"))) {  
 stop("Method must be either 't' or 'z'.")  
 }  
 if (!(var\_name %in% colnames(data))) {  
 stop("var\_name must be the name of a column in the data frame.")  
 }  
   
 # Calculate sample statistics  
 sample\_mean <- mean(data\_column, na.rm = TRUE)  
 sample\_sd <- sd(data\_column, na.rm = TRUE)  
 n <- length(na.omit(data\_column))  
 standard\_error <- sample\_sd / sqrt(n)  
   
 # Set alpha for confidence level  
 alpha <- 1 - confidence\_level  
   
 # Calculate the critical value based on the selected method  
 if (method == "t") {  
 critical\_value <- qt(1 - alpha / 2, df = n - 1) # t-distribution critical value  
 } else {  
 critical\_value <- qnorm(1 - alpha / 2) # z-distribution critical value  
 }  
   
 # Calculate margin of error (MOE)  
 margin\_of\_error <- critical\_value \* standard\_error  
   
 # Calculate confidence interval bounds  
 lower\_bound <- sample\_mean - margin\_of\_error  
 upper\_bound <- sample\_mean + margin\_of\_error  
   
 # Prepare output as a named vector  
 output <- c(  
 `Sample Mean of` = var\_name,  
 Estimate = round(sample\_mean, 3),  
 MOE = round(margin\_of\_error, 3),  
 `Lower CI Bound` = round(lower\_bound, 3),  
 `Upper CI Bound` = round(upper\_bound, 3)  
 )  
  
 # Return the result  
 return(output)  
}

Now we simulate some data and demonstrate how to use this function:

# Example usage with mock data  
# Create a mock data frame with a column of ages (normal distribution with mean 45 and SD 10)  
set.seed(123)  
mock\_data <- data.frame(age = abs(round(rnorm(100, mean = 45, sd = 10), 0)))  
  
# Visualize the simulated data  
hist(mock\_data$age)



Using this simulated data, we illustrate how to use the function to compute confidence intervals. Since we simulated the data under the assumption of a normal distribution, we use the z critical values to calculate the margin of error (MOE).

# Calculate the confidence interval for the 'age' column with 95% confidence using the z-distribution  
result\_CI\_z\_95 <- confidence\_interval(data = mock\_data, var\_name = 'age', confidence\_level = 0.95, method = "z")  
  
# Print the result  
print(result\_CI\_z\_95)

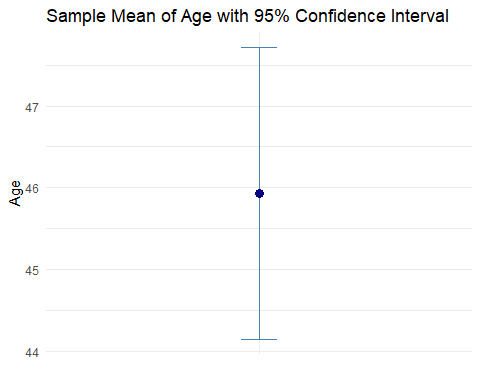
## Sample Mean of Estimate MOE Lower CI Bound Upper CI Bound   
## "age" "45.93" "1.792" "44.138" "47.722"

# Display the result in a table  
kable(data.frame(result\_CI\_z\_95))

|  | result\_CI\_z\_95 |
| --- | --- |
| Sample Mean of | age |
| Estimate | 45.93 |
| MOE | 1.792 |
| Lower CI Bound | 44.138 |
| Upper CI Bound | 47.722 |

To better present this result, we can use ggplot to plot it with a bar graph and whiskers representing the lower and upper bounds of the confidence interval.

# Create a data frame for ggplot  
plot\_data <- data.frame(  
 Estimate = as.numeric(result\_CI\_z\_95["Estimate"]),  
 Lower = as.numeric(result\_CI\_z\_95["Lower CI Bound"]),  
 Upper = as.numeric(result\_CI\_z\_95["Upper CI Bound"])  
)  
  
# Plot using ggplot2 with a dot and whiskers for confidence interval  
ggplot(plot\_data, aes(x = "Age", y = Estimate)) +  
 geom\_errorbar(aes(ymin = Lower, ymax = Upper), width = 0.1, color = "steelblue") + # Whiskers for CI  
 geom\_point(size = 3, color = "navyblue") + # Dot for the sample mean  
 labs(  
 title = "Sample Mean of Age with 95% Confidence Interval",  
 x = "",  
 y = "Age"  
 ) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_blank(), axis.ticks.x = element\_blank())



If we want a confidence interval with 99% confidence, we can adjust the function parameters accordingly:

# Calculate the confidence interval for the 'age' column with 99% confidence using the z-distribution  
result\_CI\_z\_99 <- confidence\_interval(data = mock\_data, var\_name = 'age', confidence\_level = 0.99, method = "z")  
  
# Print the result  
print(result\_CI\_z\_99)

## Sample Mean of Estimate MOE Lower CI Bound Upper CI Bound   
## "age" "45.93" "2.355" "43.575" "48.285"

# Display the result in a table  
kable(data.frame(result\_CI\_z\_99))

|  | result\_CI\_z\_99 |
| --- | --- |
| Sample Mean of | age |
| Estimate | 45.93 |
| MOE | 2.355 |
| Lower CI Bound | 43.575 |
| Upper CI Bound | 48.285 |

To compare multiple confidence intervals of different confidence levels or variables, we can combine the results using cbind:

# Combine results for different confidence intervals  
merge\_columns <- cbind(result\_CI\_z\_95, result\_CI\_z\_99)  
print(merge\_columns)

## result\_CI\_z\_95 result\_CI\_z\_99  
## Sample Mean of "age" "age"   
## Estimate "45.93" "45.93"   
## MOE "1.792" "2.355"   
## Lower CI Bound "44.138" "43.575"   
## Upper CI Bound "47.722" "48.285"

# Display the combined results in a table  
kable(data.frame(merge\_columns))

|  | result\_CI\_z\_95 | result\_CI\_z\_99 |
| --- | --- | --- |
| Sample Mean of | age | age |
| Estimate | 45.93 | 45.93 |
| MOE | 1.792 | 2.355 |
| Lower CI Bound | 44.138 | 43.575 |
| Upper CI Bound | 47.722 | 48.285 |

## Demonstration of the z.test Command for Hypothesis Testing

The z.test function tests the null hypothesis that the mean of a sample equals a hypothesized population mean. This test is suitable when the sample size is large (typically n > 30) or the population standard deviation is known and the underlying variable follows a normal distribution.

First, install the BSDA package if it is not already installed. This package includes useful statistical methods for data analysis.

# Install BSDA package if not already installed  
if (!requireNamespace("BSDA", quietly = TRUE)) {  
 install.packages("BSDA")  
}

Next, we use the z.test on our mock data to determine if the mean age differs from 45 years.

# Load BSDA library  
library(BSDA)  
  
# Run the z-test on the age data to test if the mean is significantly different from 45  
z\_test\_result <- z.test(  
 x = mock\_data$age,  
 mu = 45, # Hypothesized population mean  
 sigma.x = sd(mock\_data$age), # Population standard deviation (assuming known)  
 conf.level = 0.95  
)  
  
# Print the z-test result  
print(z\_test\_result)

##   
## One-sample z-Test  
##   
## data: mock\_data$age  
## z = 1.017, p-value = 0.3091  
## alternative hypothesis: true mean is not equal to 45  
## 95 percent confidence interval:  
## 44.13772 47.72228  
## sample estimates:  
## mean of x   
## 45.93

The output includes the z-value, p-value, and confidence interval of the sample mean. The p-value helps determine whether we can reject the null hypothesis. If the p-value is less than the significance level (typically 0.05), we reject the null hypothesis, concluding that there is sufficient evidence to suggest that the sample mean differs from the hypothesized population mean. In this example, we see that the p-value is much larger than 0.10, therefore, at 0.10, 0.05, and 0.01 significance levels we would not be able to reject the null hypothesis.