

**Batch: E2**

**Roll No.: 16010123325**

**Experiment No. 5**

**Title: Hypothesis Testing**

Aim : To perform hypothesis testing using the z-test and t-test in RStudio, and draw statistical inference from the results

Course Outcome:

**CO2 : Gain necessary mathematical skills in probability, statistics, and basic linear algebra.**

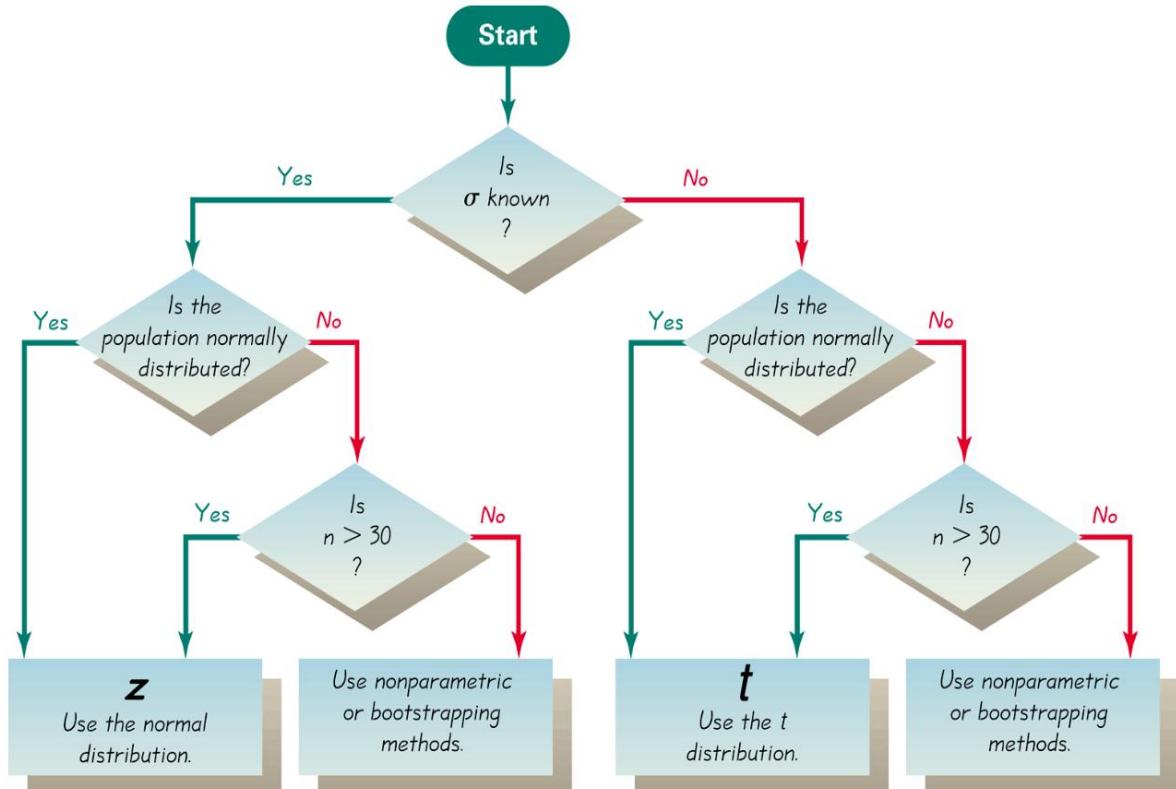
Books/ Journals/ Websites referred:

1. [The Comprehensive R Archive Network](#)
2. [Posit](#)

Resources used:

(Students should write the datasets used)

Theory:



Implementation:

### 1. One-Sample Z-Test (One-Tailed)

The population of all verbal GRE scores are known to have a standard deviation of 8.5. The UW Psychology department hopes to receive applicants with a verbal GRE scores over 210. This year, the mean verbal GRE scores for the 42 applicants was 212.79. Using a value of  $\alpha = 0.05$  is this new mean significantly greater than the desired mean of 210?

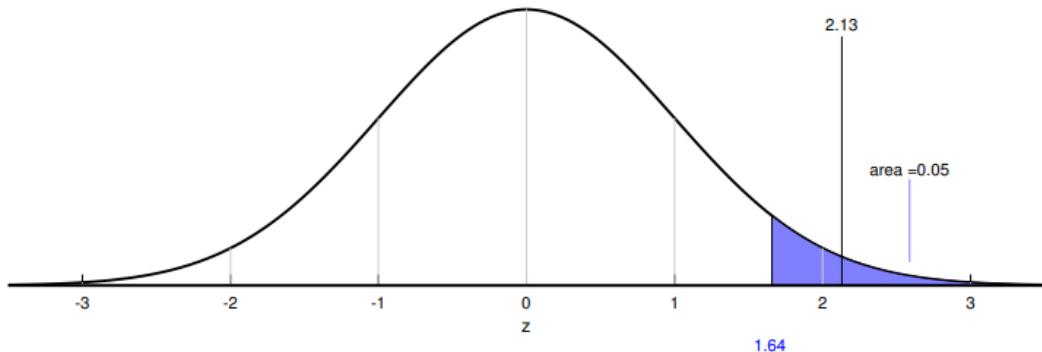
The standard error of the mean is therefore:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{8.5}{\sqrt{42}} = 1.31$$

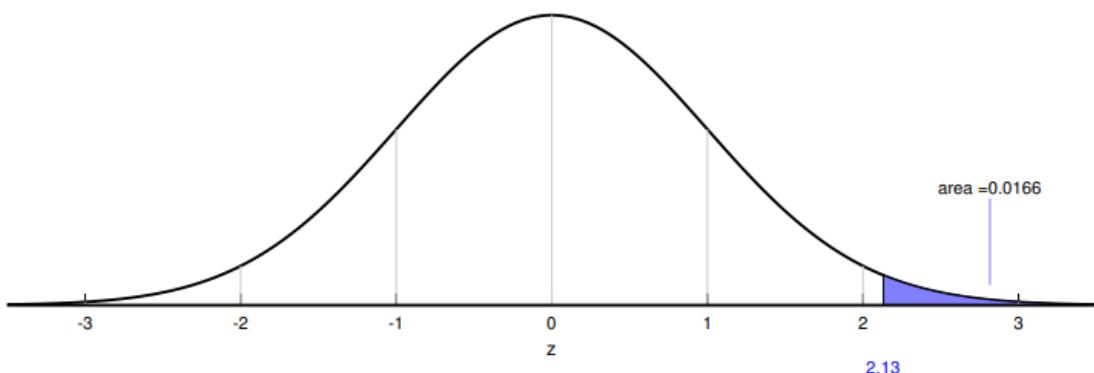
To find the probability of finding a mean above 212.79 we convert our observed mean,  $\bar{x}$ , to a z-score:

$$z = \frac{(\bar{x} - \mu_{hyp})}{\sigma_{\bar{x}}} = \frac{(212.79 - 210)}{1.31} = 2.13$$

```
> # Given data
> sigma <- 8.5
> n <- 42
> sample_mean <- 212.79
> pop_mean <- 210
> alpha <- 0.05
> # Compute Z-score
> z_score <- (sample_mean - pop_mean) / (sigma / sqrt(n))
> print(z_score)
[1] 2.127208
> # Compute p-value (one-tailed test)
> p_value <- 1 - pnorm(z_score)
> print(p_value)
[1] 0.01670141
> # Decision Rule
> if (p_value < alpha) {
+   cat("Reject H0: The new mean is significantly greater than 210.\n")
+ } else {
+   cat("Fail to Reject H0: The new mean is NOT significantly greater than 210.\n")
+ }
Reject H0: The new mean is significantly greater than 210.
```



Our observed value of  $z$  is 2.13 which is greater than the critical value of 1.64. We therefore reject  $H_0$ .

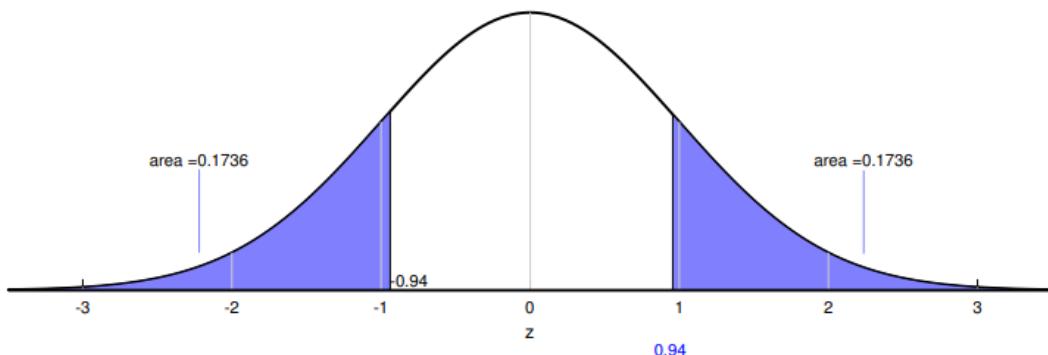


Our p-value is less than alpha (0.05). If the null hypothesis is true, then the probability of obtaining our observed mean or greater is less than 0.05. We therefore reject  $H_0$  and state that (in APA format):

The verbal GRE scores of applicants ( $M = 212.79$ ) is significantly greater than 210,  $z=2.13$ ,  $p = 0.0166$ .

## 2. One-Sample Z-Test (Two-Tailed)

Suppose you start up a company that has developed a drug that is supposed to increase IQ. You know that the standard deviation of IQ in the general population is 15. You test your drug on 36 patients and obtain a mean IQ of 97.65. Using an alpha value of 0.05, is this IQ significantly different than the population mean of 100?



To solve this, we first calculate the standard error of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5$$

and then convert our observed mean to a z-score:

$$z = \frac{(\bar{x} - \mu_{hyp})}{\sigma_{\bar{x}}} = \frac{(97.65 - 100)}{2.5} = -0.94$$

The rejection region contains values of z less than -1.96 and greater than 1.96. Our observed value of z falls outside the rejection region, so we fail to reject  $H_0$  and conclude that our drug did not have a significant effect on IQ.

```

> # Given values
> sigma <- 15
> n <- 36
> sample_mean <- 97.65
> pop_mean <- 100
> alpha <- 0.05
>
> # calculate the z-score
> z_score <- (sample_mean - pop_mean) / (sigma / sqrt(n))
>
> # calculate the p-value for a two-tailed test
> p_value <- 2 * pnorm(-abs(z_score))
>
> # Print the results
> cat("z-score:", z_score, "\n")
z-score: -0.94
> cat("p-value:", p_value, "\n")
p-value: 0.3472176
>
> # Decision rule
> if (p_value < alpha) {
+   cat("Reject the null hypothesis: There is a significant difference.\n")
+ } else {
+   cat("Fail to reject the null hypothesis: There is no significant difference.\n")
+ }
Fail to reject the null hypothesis: There is no significant difference.

```

### 3. Two-Sample Z-Test (Independent)

#### Example: Male vs Female Heights

A researcher wants to test whether male and female average heights are significantly different. The population standard deviations for both groups are known:

Male height:  $\sigma = 7$  cm

Female height:  $\sigma = 6$  cm

A random sample of 50 males and 50 females is taken:

Sample mean height for males: 175 cm

Sample mean height for females: 170 cm

Using  $\alpha = 0.05$ , determine whether the difference in average heights is statistically significant.

```

> # Given data
> x1 <- 175 # Sample mean of males
> x2 <- 170 # Sample mean of females
> sigma1 <- 7 # Population SD of males
> sigma2 <- 6 # Population SD of females
> n1 <- 50 # sample size of males
> n2 <- 50 # sample size of females
> alpha <- 0.05 # Significance level
>
> # Calculate z-score
> z_value <- (x1 - x2) / sqrt((sigma1^2 / n1) + (sigma2^2 / n2))
>
> # Compute two-tailed p-value
> p_value <- 2 * (1 - pnorm(abs(z_value)))
>
> # Critical z-value at 95% confidence
> z_critical <- qnorm(1 - alpha/2)
>
> # Decision based on z-value
> if (abs(z_value) > z_critical) {
+   z_decision <- "Reject H0 (Significant Difference in Heights)"
+ } else {
+   z_decision <- "Fail to Reject H0 (No Significant Difference)"
+ }

> # Decision based on p-value
> if (p_value < alpha) {
+   p_decision <- "Reject H0 (significant difference in Heights)"
+ } else {
+   p_decision <- "Fail to Reject H0 (No significant difference)"
+ }
>
> # Print results
> cat("Z-value:", z_value, "\n")
Z-value: 3.834825
> cat("Critical Z-value:", z_critical, "\n")
Critical Z-value: 1.959964
> cat("P-value:", p_value, "\n")
P-value: 0.0001256538
> cat("Decision based on Z-value:", z_decision, "\n")
Decision based on Z-value: Reject H0 (Significant Difference in Heights)
> cat("Decision based on P-value:", p_decision, "\n")
Decision based on P-value: Reject H0 (Significant Difference in Heights)

```

#### 4. Two-Sample Z-Test (Left-Tailed)

A company claims that **employees in the Marketing department** have significantly lower stress levels (measured on a 1-100 scale) than employees in the Sales department.

- The **population standard deviation** for stress levels in Sales is **10**.
- The **population standard deviation** for stress levels in Marketing is **8**.
- A random sample of **50 Sales employees** has a mean stress level of **70**.
- A random sample of **45 Marketing employees** has a mean stress level of **65**.
- Using  $\alpha = 0.05$ , test whether Marketing employees have significantly **lower stress levels** than Sales employees.

```
> # Given data
> x1 <- 70 # Sample mean of Sales employees
> x2 <- 65 # Sample mean of Marketing employees
> sigma1 <- 10 # Population SD of Sales employees
> sigma2 <- 8 # Population SD of Marketing employees
> n1 <- 50 # Sample size of Sales employees
> n2 <- 45 # Sample size of Marketing employees
> alpha <- 0.05 # significance level
>
> # Calculate Z-score
> z_value <- (x1 - x2) / sqrt((sigma1^2 / n1) + (sigma2^2 / n2))
>
> # Compute one-tailed p-value (left-tailed test)
> p_value <- pnorm(z_value)
>
> # critical z-value for left-tailed test
> z_critical <- qnorm(alpha)

> # Decision based on z-value
> if (z_value < z_critical) {
+   z_decision <- "Reject H0 (Marketing employees have significantly lower stress levels)"
+ } else {
+   z_decision <- "Fail to Reject H0 (No significant difference)"
+ }
>
> # Decision based on p-value
> if (p_value < alpha) {
+   p_decision <- "Reject H0 (Marketing employees have significantly lower stress levels)"
+ } else {
+   p_decision <- "Fail to Reject H0 (No significant difference)"
+ }
>
> # Print results
> cat("Z-value:", z_value, "\n")
Z-value: 2.702812
> cat("Critical Z-value:", z_critical, "\n")
Critical Z-value: -1.644854
> cat("P-value:", p_value, "\n")
P-value: 0.9965622
> cat("Decision based on Z-value:", z_decision, "\n")
Decision based on Z-value: Fail to Reject H0 (No significant difference)
> cat("Decision based on P-value:", p_decision, "\n")
Decision based on P-value: Fail to Reject H0 (No significant difference)
```

## 5. Paired z-test

A researcher wants to determine whether a new diet plan is effective in reducing weight. To test this, they measure the weight (in kg) of 10 participants before and after following the diet plan for 6 weeks. The data is as follows:

Participant	Before Diet	After Diet
1	70	68
2	75	74
3	80	78
4	85	83
5	90	88
6	72	70
7	78	76
8	82	80
9	88	86
10	95	92

```
> library(BSDA)
>
> # Data: Before and After weights
> before <- c(70, 75, 80, 85, 90, 72, 78, 82, 88, 95)
> after <- c(68, 74, 78, 83, 88, 70, 76, 80, 86, 92)
>
> # Calculate differences
> differences <- after - before
>
> # Perform paired z-test
> z_test_result <- z.test(x = differences,
+                           sigma.x = sd(differences),
+                           mu = 0,
+                           alternative = "two.sided")
>
> # Print results
> print(z_test_result)
```

One-sample z-Test

```
data: differences
z = -13.416, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-2.292174 -1.707826
sample estimates:
mean of x
-2
```

### t-Tests:

#### 1. One-Sample T-Test (Two Tailed)

A normal cholesterol level is considered **less than 200 mg/dL**.

We randomly sample **100 individuals** from the NHANES dataset and test:

- **H<sub>0</sub> (Null Hypothesis):** The mean cholesterol level in our sample is **200 mg/dL** (population mean).
- **H<sub>1</sub> (Alternative Hypothesis):** The mean cholesterol level **is different** from 200 mg/dL.
- We use  $\alpha = 0.05$  for significance.

```
> library(NHANES)
library(ggplot2)

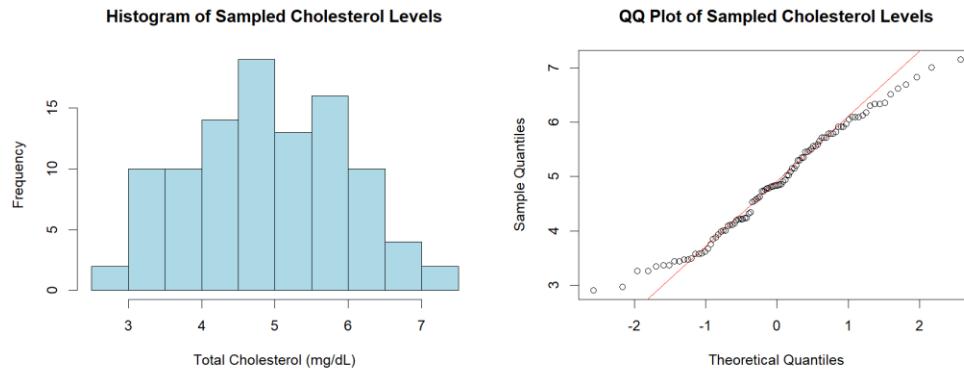
# Load and Clean Data
set.seed(123) # Ensuring reproducibility
chol_data <- na.omit(NHANES$TotChol) # Remove NA values

# Randomly select 100 Rows
sample_data <- sample(chol_data, 100, replace = FALSE)

# Visualizing the Sample Distribution
hist(sample_data, breaks = 15, col="lightblue", border="black",
     main="Histogram of Sampled Cholesterol Levels",
     xlab="Total Cholesterol (mg/dL)", ylab="Frequency")

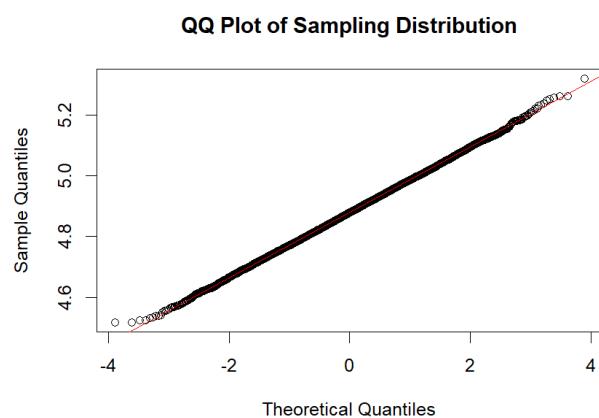
# QQ Plot to Check Normality
qqnorm(sample_data, main="QQ Plot of Sampled Cholesterol Levels")
qqline(sample_data, col="red")
```

Checking the distribution of underlying data



```

> library(NHANES)
> library(ggplot2)
>
> # Remove NAs and Extract Total Cholesterol Data
> chol_data <- na.omit(NHANES$TotChol)
>
> # Set Parameters for Sampling
> sample_size <- 100 # Choose a large enough sample size
> num_samples <- 10000 # Number of samples to take
>
> # Generate Sampling Distribution
> sample_means <- replicate(num_samples, mean(sample(chol_data, sample_size, replace=TRUE)))
>
> # Plot Histogram of Sample Means
> ggplot(data.frame(sample_means), aes(x=sample_means)) +
+   geom_histogram(binwidth=1, fill="lightblue", color="black", alpha=0.7) +
+   ggtitle("Sampling Distribution of Sample Means") +
+   xlab("Sample Mean of Total Cholesterol") +
+   ylab("Frequency") +
+   theme_minimal()
>
> # QQ Plot to Check Normality
> qqnorm(sample_means, main="QQ Plot of Sampling Distribution")
> qqline(sample_means, col="red")
  
```



```
# One-Sample T-Test
t_test_result <- t.test(sample_data, mu=200, alternative="two.sided")

# Print Test Results
print(t_test_result)

# Decision Rule
if (t_test_result$p.value < 0.05) {
  cat("Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.\n")
} else {
  cat("Fail to Reject H0: No significant difference from 200 mg/dL.\n")
}

One Sample t-test

data: sample_data
t = -1891.3, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 4.68681 5.09619
sample estimates:
mean of x
 4.8915

>
> # Decision Rule
> if (t_test_result$p.value < 0.05) {
+   cat("Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.\n")
+ } else {
+   cat("Fail to Reject H0: No significant difference from 200 mg/dL.\n")
+ }
Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.
```

**Students have to explore the datasets within R libraries, Kaggle, UCI or any other freely available data repositories, frame the null and alternate hypothesis on suitable attributes and perform all the z-tests and t-tests.**

**Students have to clearly mention the data sources, paste their code, and screenshots of output in the writeup.**

**Conclusion: Students should write in their own words.**

A factory produces light bulbs with a known population mean lifespan of 1,000 hours and a standard deviation of 50 hours. A quality control team tests  $n = 36$  randomly selected bulbs and finds that their mean lifespan is 1,020 hours. The team wants to determine whether the new batch of bulbs has a significantly higher average lifespan at a 5% significance level.

```
> # 1. One-Sample z-Test (One-Tailed)
> # Given data
> sigma <- 50      # Population standard deviation
> n <- 36          # Sample size
> sample_mean <- 1020 # Sample mean
> pop_mean <- 1000  # Population mean
> alpha <- 0.05    # Significance level
>
> # Compute z-score
> z_score <- (sample_mean - pop_mean) / (sigma / sqrt(n))
> print(z_score)
[1] 2.4
>
> # Compute p-value (one-tailed test)
> p_value <- 1 - pnorm(z_score)
> print(p_value)
[1] 0.008197536
>
> # Decision Rule
> if (p_value < alpha) {
+   cat("Reject H0: The new batch has a significantly higher lifespan than 1,000 hours.\n")
+ } else {
+   cat("Fail to Reject H0: There is no significant increase in the lifespan.\n")
+ }
Reject H0: The new batch has a significantly higher lifespan than 1,000 hours.
>
```

A company claims that the average weight of a certain type of packaged cereal is 500g. A quality control team selects a random sample of 40 packages and finds that the sample has an average weight of 495g with a known population standard deviation of 12g. The team wants to test whether the actual mean weight differs significantly from 500g at a 5% significance level.

```
> # 2. One-Sample Z-Test (Two-Tailed)
> # Given values
> sigma <- 12      # Population standard deviation
> n <- 40          # Sample size
> sample_mean <- 495 # Sample mean
> pop_mean <- 500  # Population mean
> alpha <- 0.05    # Significance level
>
> # calculate the z-score
> z_score <- (sample_mean - pop_mean) / (sigma / sqrt(n))
>
> # calculate the p-value for a two-tailed test
> p_value <- 2 * pnorm(-abs(z_score))
>
> # Print the results
> cat("z-score:", z_score, "\n")
z-score: -2.635231
> cat("p-value:", p_value, "\n")
p-value: 0.008407995
>
> # Decision rule
> if (p_value < alpha) {
+   cat("Reject the null hypothesis: There is a significant difference in weight.\n")
+ } else {
+   cat("Fail to reject the null hypothesis: There is no significant difference in weight.\n")
+ }
Reject the null hypothesis: There is a significant difference in weight.
> |
```

A researcher wants to compare the average test scores of two groups of students who used different study techniques.

- Group 1 (Technique A) has a sample mean score of 82, a population standard deviation of 10, and a sample size of 60.
- Group 2 (Technique B) has a sample mean score of 78, a population standard deviation of 12, and a sample size of 65.
- The researcher tests at a 5% significance level to determine if there is a significant difference between the two groups' average test scores.

```

> # 3. Two-Sample Z-Test (Independent)
> # Given data
> x1 <- 82 # Sample mean of group 1 (Technique A)
> x2 <- 78 # Sample mean of group 2 (Technique B)
> sigma1 <- 10 # Population SD of group 1
> sigma2 <- 12 # Population SD of group 2
> n1 <- 60 # Sample size of group 1
> n2 <- 65 # Sample size of group 2
> alpha <- 0.05 # Significance level
>
> # calculate z-score
> z_value <- (x1 - x2) / sqrt((sigma1^2 / n1) + (sigma2^2 / n2))
>
> # Compute two-tailed p-value
> p_value <- 2 * (1 - pnorm(abs(z_value)))
>
> # Critical z-value at 95% confidence
> z_critical <- qnorm(1 - alpha/2)
>
> # Decision based on z-value
> if (abs(z_value) > z_critical) {
+   z_decision <- "Reject H0 (significant difference in test scores)"
+ } else {
+   z_decision <- "Fail to Reject H0 (No significant difference in test scores)"
+ }
>
> # Decision based on p-value
> if (p_value < alpha) {
+   p_decision <- "Reject H0 (significant difference in test scores)"
+ } else {
+   p_decision <- "Fail to Reject H0 (No significant difference in test scores)"
+ }
>
> # Print results
> cat("Z-value:", z_value, "\n")
Z-value: 2.030156
> cat("Critical Z-value:", z_critical, "\n")
Critical z-value: 1.959964
> cat("P-value:", p_value, "\n")
P-value: 0.04234071
> cat("Decision based on Z-value:", z_decision, "\n")
Decision based on Z-value: Reject H0 (significant difference in test scores)
> cat("Decision based on P-value:", p_decision, "\n")
Decision based on P-value: Reject H0 (significant difference in test scores)
> |

```

A researcher wants to compare the **reaction times** (in milliseconds) of two groups of drivers:

- **Young drivers (Group 1)** have a **sample mean reaction time of 280 ms**, a **population standard deviation of 50 ms**, and a **sample size of 55**.
- **Older drivers (Group 2)** have a **sample mean reaction time of 300 ms**, a **population standard deviation of 60 ms**, and a **sample size of 50**.
- The researcher wants to test at a **5% significance level** whether **young drivers have significantly lower reaction times** (i.e., faster reactions) than older drivers.

```
> # 4. Two-Sample z-Test (Left-Tailed)
> # Given data
> x1 <- 280 # Sample mean of young drivers
> x2 <- 300 # Sample mean of older drivers
> sigma1 <- 50 # Population SD of young drivers
> sigma2 <- 60 # Population SD of older drivers
> n1 <- 55 # Sample size of young drivers
> n2 <- 50 # sample size of older drivers
> alpha <- 0.05 # Significance level
>
> # calculate z-score
> z_value <- (x1 - x2) / sqrt((sigma1^2 / n1) + (sigma2^2 / n2))
>
> # Compute one-tailed p-value (left-tailed test)
> p_value <- pnorm(z_value)
>
> # Critical z-value for left-tailed test
> z_critical <- qnorm(alpha)
>
> # Decision based on z-value
> if (z_value < z_critical) {
+   z_decision <- "Reject H0 (Young drivers have significantly lower reaction times)"
+ } else {
+   z_decision <- "Fail to Reject H0 (No significant difference)"
+ }
>
> # Decision based on p-value
> if (p_value < alpha) {
+   p_decision <- "Reject H0 (Young drivers have significantly lower reaction times)"
+ } else {
+   p_decision <- "Fail to Reject H0 (No significant difference)"
+ }
>
> # Print results
> cat("Z-value:", z_value, "\n")
Z-value: -1.845419
> cat("Critical Z-value:", z_critical, "\n")
Critical Z-value: -1.644854

> cat("Decision based on Z-value:", z_decision, "\n")
Decision based on Z-value: Reject H0 (Young drivers have significantly lower reaction times)
> cat("Decision based on P-value:", p_decision, "\n")
Decision based on P-value: Reject H0 (Young drivers have significantly lower reaction times)
> |

-- library (lattice)
Loading required package: lattice

Attaching package: 'BSDA'

The following object is masked from 'package:datasets':
  orange

>
> # Data: Before and After weights
> before <- c(70, 75, 80, 85, 90, 72, 78, 82, 88, 95)
> after <- c(68, 74, 78, 83, 88, 70, 76, 80, 86, 92)
>
> # Calculate differences
> differences <- after - before
>
> # Perform paired Z-test
> z_test_result <- z.test(x = differences,
+                           sigma.x = sd(differences),
+                           mu = 0,
+                           alternative = "two.sided")
>
> # Print results
> print(z_test_result)

  One-sample z-Test

data: differences
z = -13.416, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-2.292174 -1.707826
sample estimates:
mean of x
-2

>
```

A fitness coach wants to analyze whether a **new diet plan** significantly reduces **blood pressure levels**.

- **Blood pressure readings were recorded before and after** following the diet for 8 weeks.

- We assume the **population standard deviation** of the differences is known.
- The coach conducts a **paired Z-test** to determine if there is a significant reduction.

```

> # 5. Paired z-test
> library(BSDA)
>
> # Data: Before and After Blood Pressure Levels
> before <- c(135, 140, 150, 145, 155, 138, 142, 148, 152, 160)
> after <- c(130, 136, 145, 140, 150, 134, 138, 144, 148, 155)
>
> # calculate the differences (Before - After)
> differences <- before - after
>
> # Perform Paired Z-Test
> z_test_result <- z.test(x = differences,
+                           sigma.x = sd(differences), # Population SD of differences
+                           mu = 0, # Null hypothesis: No mean difference
+                           alternative = "greater") # One-tailed test (Checking if Before > After)
>
> # Print results
> print(z_test_result)

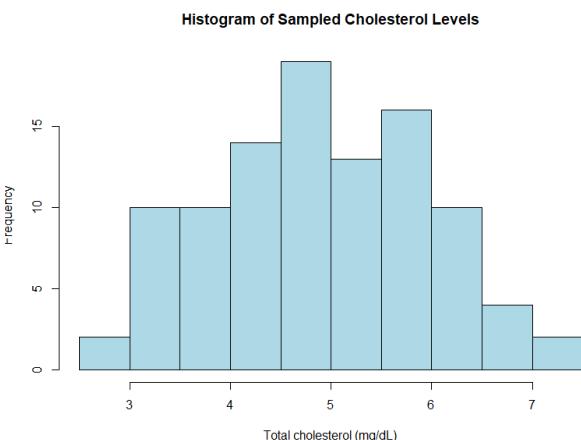
One-sample z-Test

data: differences
z = 27, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 4.225858      NA
sample estimates:
mean of x
        4.5

>
> # Decision Rule
> alpha <- 0.05 # Significance level
> if (z_test_result$p.value < alpha) {
+   cat("Reject H0: The new diet plan significantly reduces blood pressure. \n")
+ } else {
+   cat("Fail to Reject H0: No significant reduction in blood pressure. \n")
+ }
Reject H0: The new diet plan significantly reduces blood pressure.
>

> library (NHANES)
> library(ggplot2)
> # Load and Clean Data
> set.seed(123) # Ensuring reproducibility
> chol_data <- na.omit (NHANES$TotChol) # Remove NA values
> # Randomly select 100 Rows
> sample_data <- sample(chol_data, 100, replace = FALSE)
> # Visualizing the Sample Distribution
> hist(sample_data, breaks = 15, col="lightblue", border="black",
+       main="Histogram of Sampled Cholesterol Levels",
+       xlab="Total cholesterol (mg/dL)", ylab="Frequency")
>

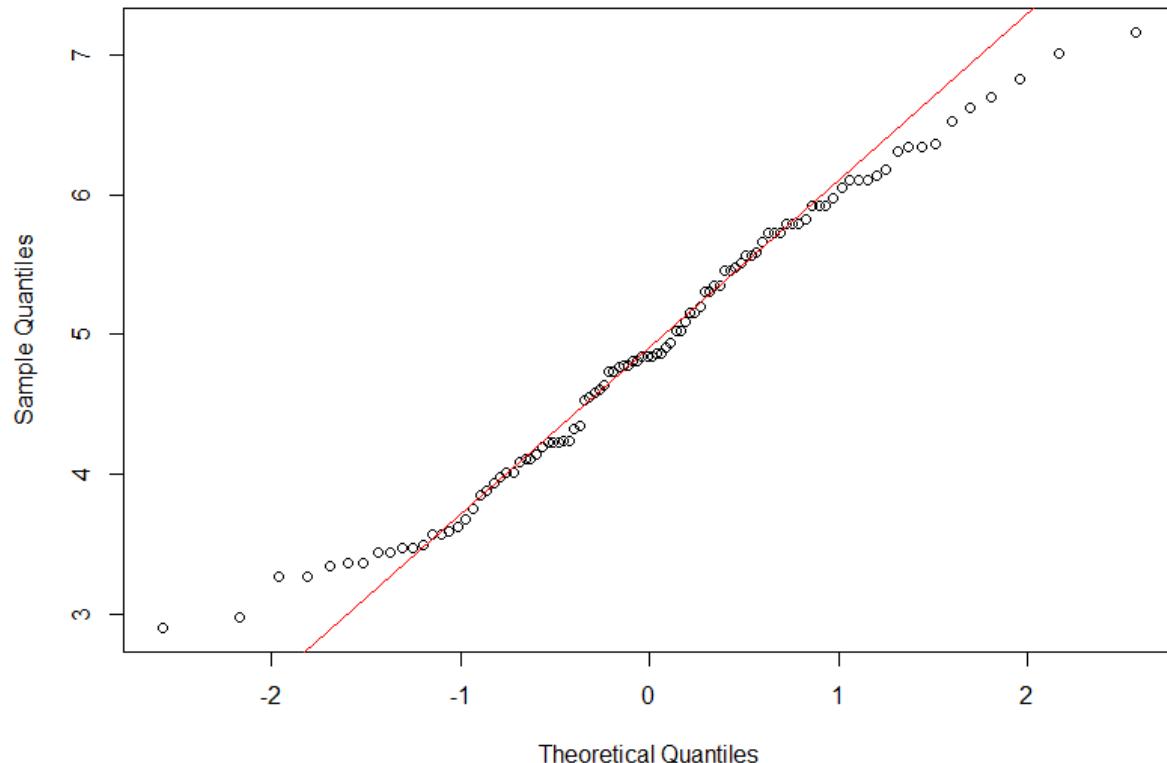
```



```

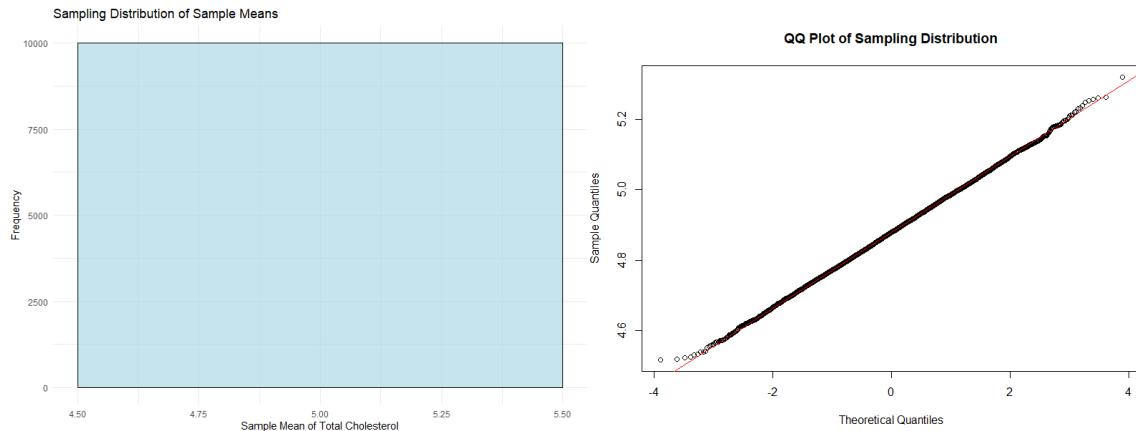
> # QQ Plot to Check Normality
> qqnorm(sample_data, main="QQ Plot of sampled cholesterol Levels")
> qqline (sample_data, col="red")
>
  
```

**QQ Plot of Sampled Cholesterol Levels**



```

> library (NHANES)
> library(ggplot2)
>
> # Remove NAs and Extract Total Cholesterol Data
> chol_data <- na.omit (NHANES$TotChol)
>
> # Set Parameters for Sampling
> sample_size <- 100 # Choose a large enough sample size
> num_samples <- 10000 # Number of samples to take
> # Generate Sampling Distribution
> sample_means <- replicate (num_samples, mean (sample(chol_data, sample_size, replace=TRUE) ))
>
> # Plot Histogram of Sample Means
> ggplot(data.frame (sample_means), aes(x=sample_means)) +
+   geom_histogram(binwidth=1, fill="lightblue", color="black", alpha=0.7) +
+   ggtitle("Sampling Distribution of Sample Means") +
+   xlab ("Sample Mean of Total Cholesterol") +
+   ylab("Frequency") +
+   theme_minimal()
>
> # QQ Plot to Check Normality
> qqnorm(sample_means, main="QQ Plot of sampling Distribution")
> qqline (sample_means, col="red")
>
> # One-Sample T-Test
> t_test_result <- t.test(sample_data, mu=200, alternative="two. sided")
  
```



```
# One-Sample T-Test
t_test_result <- t.test(sample_data, mu=200, alternative="two.sided")

# Print Test Results
print(t_test_result)

# Decision Rule
if (t_test_result$p.value < 0.05) {
  cat("Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.\n")
} else {
  cat("Fail to Reject H0: No significant difference from 200 mg/dL.\n")
}

One Sample t-test

data: sample_data
t = -1891.3, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
 4.68681 5.09619
sample estimates:
mean of x
 4.8915

>
> # Decision Rule
> if (t_test_result$p.value < 0.05) {
+   cat("Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.\n")
+ } else {
+   cat("Fail to Reject H0: No significant difference from 200 mg/dL.\n")
+ }
Reject H0: The mean cholesterol level is significantly different from 200 mg/dL.
```

### Post Lab questions:

1. A university requires applicants to have a **minimum GRE score of 320**. You sample **36 applicants**, and their **mean GRE score is 322.5**, with a known **population standard deviation of 8**. At  $\alpha = 0.05$ , can we conclude that the average GRE score of applicants is significantly greater than 320?

```

> #Q1
> # Input data
> sample_mean <- 322.5 # Sample mean
> pop_mean <- 320 # Hypothesized population mean
> pop_sd <- 8 # Population standard deviation
> n <- 36 # Sample size
> alpha <- 0.05 # Significance level
>
> # Calculate the z-test statistic
> z <- (sample_mean - pop_mean) / (pop_sd / sqrt(n))
> z # Print the z-value
[1] 1.875
>
> # Calculate the critical z-value for a one-tailed test
> critical_z <- qnorm(1 - alpha)
> critical_z # Print the critical z-value
[1] 1.644854
>
> # Compare the z-value to the critical z-value
> if (z > critical_z) {
+   print("Reject the null hypothesis: The average GRE score is significantly greater than 320.")
+ } else {
+   print("Fail to reject the null hypothesis: The average GRE score is not significantly greater than 320.")
+ }
[1] "Reject the null hypothesis: The average GRE score is significantly greater than 320."
>
> # Calculate the p-value
> p_value <- 1 - pnorm(z)
> p_value # Print the p-value
[1] 0.03039636
>
> # Compare the p-value to alpha
> if (p_value < alpha) {
+   print("Reject the null hypothesis: The average GRE score is significantly greater than 320.")
+ } else {
+   print("Fail to reject the null hypothesis: The average GRE score is not significantly greater than 320.")
+ }
[1] "Reject the null hypothesis: The average GRE score is significantly greater than 320."
>

```

- The average SAT Math score in a country is **500**, with a known standard deviation of **100**. You collect a **random sample of 50 students**, and their **mean score is 520**. At  **$\alpha = 0.05$** , can we conclude that the sample mean significantly differs from 500?

```

> #Q2
> # Input data
> sample_mean <- 520 # Sample mean
> pop_mean <- 500 # Hypothesized population mean
> pop_sd <- 100 # Population standard deviation
> n <- 50 # Sample size
> alpha <- 0.05 # Significance level
>
> # Calculate the z-test statistic
> z <- (sample_mean - pop_mean) / (pop_sd / sqrt(n))
> z # Print the z-value
[1] 1.414214
>
> # Calculate the critical z-value for a two-tailed test
> critical_z <- qnorm(1 - alpha / 2)
> critical_z # Print the critical z-value
[1] 1.959964
>
> # Compare the z-value to the critical z-value
> if (abs(z) > critical_z) {
+   print("Reject the null hypothesis: The sample mean significantly differs from 500.")
+ } else {
+   print("Fail to reject the null hypothesis: The sample mean does not significantly differ from 500.")
+ }
[1] "Fail to reject the null hypothesis: The sample mean does not significantly differ from 500."
>
> # Calculate the p-value
> p_value <- 2 * (1 - pnorm(abs(z)))
> p_value # Print the p-value
[1] 0.1572992
>
> # Compare the p-value to alpha
> if (p_value < alpha) {
+   print("Reject the null hypothesis: The sample mean significantly differs from 500.")
+ } else {
+   print("Fail to reject the null hypothesis: The sample mean does not significantly differ from 500.")
+ }
[1] "Fail to reject the null hypothesis: The sample mean does not significantly differ from 500."
>

```

3. A weight-loss program was tested on **20 individuals** by measuring their weights **before and after** the program. The **mean weight before** the program was **80 kg** ( $SD = 5$ ). The **mean weight after** the program was **78 kg** ( $SD = 4.5$ ). The **differences** are assumed to be normally distributed. At  $\alpha = 0.05$ , is the weight loss significant?

```
> #Q3
> # Given data
> n <- 20      # Sample size
> mean_before <- 80 # Mean weight before program
> mean_after <- 78 # Mean weight after program
> sd_before <- 5  # Standard deviation before program
> sd_after <- 4.5 # Standard deviation after program
> alpha <- 0.05  # Significance level
>
> # calculate the mean difference
> mean_diff <- mean_after - mean_before
>
> # Compute the standard deviation of differences (pooled standard deviation)
> sd_diff <- sqrt((sd_before^2 + sd_after^2) / 2)
>
> # Compute the z-score
> z_score <- mean_diff / (sd_diff / sqrt(n))
>
> # Compute two-tailed p-value
> p_value <- 2 * (1 - pnorm(abs(z_score)))
>
> # Critical z-value for two-tailed test at  $\alpha = 0.05$ 
> z_critical <- qnorm(1 - alpha/2)
>
> # Print results
> cat("Mean Difference:", mean_diff, "\n")
Mean Difference: -2
> cat("Standard Deviation of Differences:", sd_diff, "\n")
Standard Deviation of Differences: 4.756574
> cat("Z-score:", z_score, "\n")
Z-score: -1.880402
> cat("Critical Z-value:", z_critical, "\n")
Critical Z-value: 1.959964
> cat("P-value:", p_value, "\n")
P-value: 0.06005331
>
> # Decision Rule
> if (p_value < alpha) {
+   cat("Reject H0: The weight loss program has a significant effect. \n")
+ } else {
+   cat("Fail to Reject H0: No significant weight loss effect. \n")
+ }
Fail to Reject H0: No significant weight loss effect.
> |
```

4. When should you use a Z-test instead of a T-test? Explain with examples and discuss the role of sample size in determining which test to use.

#### -Test:

- Used when the **population standard deviation ( $\sigma$ ) is known**.
- Suitable for **large sample sizes ( $n \geq 30$ )** due to the **Central Limit Theorem**.
- Assumes data is **normally distributed**.

#### Example:

A university claims that the average GRE score of applicants is 320. If we sample 50 students and find a sample mean of 322 with a known **population standard deviation ( $\sigma = 8$ )**, we use a **one-sample Z-test** to check if the mean is significantly different.

### T-Test:

- Used when the **population standard deviation ( $\sigma$ ) is unknown.**
- Suitable for **small samples ( $n < 30$ ).**
- Uses **sample standard deviation (s) instead of population standard deviation.**
- More flexible for real-world scenarios where  $\sigma$  is often unknown.

### Example:

A weight-loss program is tested on 15 participants by measuring their weight before and after the program. Since  **$\sigma$  is unknown and  $n < 30$** , we use a **paired t-test** to analyze the difference.

---

### Role of Sample Size:

- **Large samples ( $n \geq 30$ ):** The **Z-test** is preferred because sample means approximate normality due to the **Central Limit Theorem (CLT)**.
- **Small samples ( $n < 30$ ):** The **T-test** is used as it accounts for **increased variability in small samples** using Student's **t-distribution**.