**Question 1.**

**1. Calculate an estimate of the average number of days to recovery using the provided data. Calculate a 95% confidence interval for this estimate using the *t*-distribution, and summarise/describe your results appropriately. Show working as required.**

***Answer:***

* the sample mean is ***14.25797***
* the sample standard deviation is ***6.64479***
* *95% Confidence Interval:* ***13.98935 to 14.52659***

***Interpretation:***

*The sample mean of 14.26 days suggests that, on average, COVID-19 patients in New South Wales take about 14.26 days to recover from the virus.*

*The 95% confidence interval (13.99 to 14.53 days) provides a range within which we can be reasonably confident (95% confidence level) that the true population average recovery time lies. This interval represents the uncertainty associated with our estimate.*

*In practical terms, this means that if we were to repeatedly sample from the same population and calculate confidence intervals, we would expect approximately 95% of these intervals to contain the true average recovery time.*

***Approach:***

# Load the data with the correct file path

> recovery\_times <- read.csv("C:/Users/admin/Desktop/Genilytics Solutions/Genilytics\_Solution-ML\_intern/4. Covid Analysis and Human Behavior/Instructions and Data/covid.19.ass2.2023.csv")

> # View the contents of the recovery\_times dataset

> str(recovery\_times)

***'data.frame': 2353 obs. of 1 variable:***

***$ Recovery.Time: int 14 11 22 19 14 24 18 11 14 12 ...***

> # Compute the sample mean

> sample\_mean <- mean(recovery\_times$Recovery.Time)

> # Print the sample mean

> cat("Sample Mean:", sample\_mean, "\n")

***Sample Mean: 14.25797***

> # Compute the sample standard deviation

> sample\_std\_deviation <- sd(recovery\_times$Recovery.Time)

> # Print the sample standard deviation

> cat("Sample Standard Deviation:", sample\_std\_deviation, "\n")

***Sample Standard Deviation: 6.64479***

> # Define the confidence level and degrees of freedom

> confidence\_level <- 0.95

> degrees\_of\_freedom <- length(recovery\_times$Recovery.Time) - 1

> degrees\_of\_freedom

***[1] 2352***

> # Find the critical t-value

> critical\_t\_value <- qt((1 + confidence\_level) / 2, df = degrees\_of\_freedom)

> critical\_t\_value

***[1] 1.960973***

> # Calculate the margin of error

> margin\_error <- (sample\_std\_deviation / sqrt(length(recovery\_times$Recovery.Time))) \* critical\_t\_value

> margin\_error

***[1] 0.2686222***

> # Calculate the confidence interval

> lower\_bound <- sample\_mean - margin\_error

> lower\_bound

***[1] 13.98935***

> upper\_bound <- sample\_mean + margin\_error

> upper\_bound

***[1] 14.52659***

> # Print the confidence interval

> cat("95% Confidence Interval:", lower\_bound, "to", upper\_bound, "\n")

***95% Confidence Interval: 13.98935 to 14.52659***

**Question 1.**

**2. Similar data was collected in 2020 by the Israeli Ministry of Health. While the specific data collected was not available, the summary statistics were provided, and from these I have simulated a dataset of n = 494 individuals from the Israeli study. The days to recovery in this group are provided in the file israeli.covid.19.ass2.2023.csv. Using the provided data and the approximate method for difference in means with (different) unknown variances, calculate the estimated mean difference in recovery times between the Israeli patients and the patients from NSW, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required.**

***Answer:***

Sample Mean (NSW): **14.25797**

Sample Standard Deviation (NSW): **6.64479**

Sample Mean (Israeli): **14.6498**

Sample Standard Deviation (Israeli): **5.520461**

Standard Error of Difference in Means: **0.2836475**

The margin of error for the Difference in Means: **0.5561756**

Estimated Mean Difference in Recovery Times: **-0.391829**

95% Confidence Interval for Mean Difference: **[ -0.9480046, 0.1643466]**

***Interpretation:***

1.Sample Mean (NSW): The average recovery time for patients in New South Wales (NSW) is approximately 14.26 days.

2. Sample Standard Deviation (NSW): The variation in recovery times among NSW patients is approximately 6.64 days, indicating some variability in the data.

3. Sample Mean (Israeli): The average recovery time for Israeli patients is approximately 14.65 days.

4. Sample Standard Deviation (Israeli): The variation in recovery times among Israeli patients is approximately 5.52 days, which suggests slightly less variability compared to NSW.

5. Standard Error of Difference in Means: The standard error of the difference in means is approximately 0.28 days. It represents the variability of the sample mean difference if we were to take multiple random samples.

6. Margin of Error for Difference in Means: The margin of error for the difference in means is approximately 0.56 days. This value indicates the range within which we can reasonably expect the true mean difference in recovery times to fall with 95% confidence.

7. Estimated Mean Difference in Recovery Times: The estimated mean difference in recovery times between Israeli and NSW patients is approximately -0.39 days. This negative value suggests that, on average, Israeli patients may recover slightly faster than NSW patients, although the difference is relatively small.

8. 95% Confidence Interval for Mean Difference: The 95% confidence interval for the mean difference ranges from approximately -0.95 days to 0.16 days. This interval tells us that we can be 95% confident that the true mean difference falls within this range. Since it includes zero, it suggests that there may not be a statistically significant difference in recovery times between the two groups.

In summary, while there is a small estimated mean difference in recovery times between Israeli and NSW patients, the confidence interval includes zero, indicating that this difference may not be statistically significant. Further analysis or a larger sample size may be needed to draw more conclusive results about any potential differences in recovery times between the two populations.

***Approach:***

>israeli\_data <- read.csv("C:/Users/admin/Desktop/Genilytics Solutions/Genilytics\_Solution-ML\_intern/4. Covid Analysis and Human Behavior/Instructions and Data/israeli.covid.19.ass2.2023.csv")

> str(israeli\_data)

***'data.frame': 494 obs. of 1 variable:***

***$ Recovery.Time: int 20 21 22 28 8 8 10 5 13 22 ...***

> # Calculate sample mean and sample standard deviation for both datasets

> mean\_nsw <- mean(recovery\_times$Recovery.Time)

> sd\_nsw <- sd(recovery\_times$Recovery.Time)

> mean\_israeli <- mean(israeli\_data$Recovery.Time)

> sd\_israeli <- sd(israeli\_data$Recovery.Time)

> # Calculate the sample sizes (n1 and n2)

> n1 <- length(recovery\_times$Recovery.Time)

> n2 <- length(israeli\_data$Recovery.Time)

> # Calculate the standard error of the difference in means

> standard\_error\_diff <- sqrt((sd\_nsw^2 / n1) + (sd\_israeli^2 / n2))

> # Calculate t-score for a 95% confidence interval

> alpha <- 0.05 # 1 - confidence level

> t\_score <- qt(1 - alpha / 2, df = n1 + n2 - 2) # t-score for a two-tailed test

> # Calculate margin of error

> margin\_of\_error\_diff <- t\_score \* standard\_error\_diff

> # Calculate the estimated mean difference

> mean\_difference <- mean\_nsw - mean\_israeli

> # Compute lower and upper bounds of the confidence interval

> lower\_bound\_diff <- mean\_difference - margin\_of\_error\_diff

> upper\_bound\_diff <- mean\_difference + margin\_of\_error\_diff

> # Summarize the results

> cat("Sample Mean (NSW):", mean\_nsw, "\n")

***Sample Mean (NSW): 14.25797***

> cat("Sample Standard Deviation (NSW):", sd\_nsw, "\n")

***Sample Standard Deviation (NSW): 6.64479***

> cat("Sample Mean (Israeli):", mean\_israeli, "\n")

***Sample Mean (Israeli): 14.6498***

> cat("Sample Standard Deviation (Israeli):", sd\_israeli, "\n")

***Sample Standard Deviation (Israeli): 5.520461***

> cat("Standard Error of Difference in Means:", standard\_error\_diff, "\n")

***Standard Error of Difference in Means: 0.2836475***

> cat("Margin of Error for Difference in Means:", margin\_of\_error\_diff, "\n")

***Margin of Error for Difference in Means: 0.5561756***

> cat("Estimated Mean Difference in Recovery Times:", mean\_difference, "\n")

***Estimated Mean Difference in Recovery Times: -0.391829***

> cat("95% Confidence Interval for Mean Difference: [", lower\_bound\_diff, ", ", upper\_bound\_diff, "]\n")

***95% Confidence Interval for Mean Difference: [ -0.9480046 , 0.1643466 ]***

**Question 1.**

**3. It is of interest to determine if there are any differences, at a population level, in recovery times for patients in different countries. Test the hypothesis that the population average time taken to recover for the Israeli cohort is the same as in the NSW cohort. Write down explicitly the hypothesis you are testing, and then calculate a *p*-value using the approximate hypothesis test for differences in means with (different) unknown variances. What does this *p*-value suggest about the difference in mean recovery time between the two cohorts of patients?**

***Answer:***

p-value: ***1.83283***

***Interpretation:***

Fail to reject the null hypothesis: There is no significant difference in mean recovery time between Israeli and NSW cohorts.

***Approach:***

Step 1: Null and Alternative Hypotheses:

: (Population average recovery times are equal for the Israeli and NSW cohorts.)

: (Population average recovery times are not equal for the Israeli and NSW cohorts)

2. Calculate the t-statistic:

The t-statistic is calculated as follows:

Where:

* are the sample means for the Israeli and NSW cohorts, respectively.
* and are the sample standard deviations for the Israeli and NSW cohorts, respectively.
* and are the sample sizes for the Israeli and NSW cohorts, respectively.

3. Find the degrees of freedom:

The degrees of freedom df for the t-test can be calculated using the formula:

4. Calculate the p-value:

We can use the t-distribution to calculate the p-value associated with the calculated t-statistic. We're interested in a two-tailed test because the alternative hypothesis is two-sided (). we will use the `pt` function in R to find the p-value.

5. Interpret the p-value:

* If the p-value is less than your chosen significance level (0.05), we would reject the null hypothesis.
* If the p-value is greater than your significance level, we would fail to reject the null hypothesis.

**R-Code:**

> recovery\_times <- read.csv("C:/Users/admin/Desktop/Genilytics Solutions/Genilytics\_Solution-ML\_intern/4. Covid Analysis and Human Behavior/Instructions and Data/covid.19.ass2.2023.csv")

> str(recovery\_times)

***'data.frame': 2353 obs. of 1 variable:***

***$ Recovery.Time: int 14 11 22 19 14 24 18 11 14 12 ...***

> israeli\_data <- read.csv("C:/Users/admin/Desktop/Genilytics Solutions/Genilytics\_Solution-ML\_intern/4. Covid Analysis and Human Behavior/Instructions and Data/israeli.covid.19.ass2.2023.csv")

> str(israeli\_data)

***'data.frame': 494 obs. of 1 variable:***

***$ Recovery.Time: int 20 21 22 28 8 8 10 5 13 22 ...***

> # Calculate sample means and standard deviations for both cohorts

> mean\_nsw <- mean(recovery\_times$Recovery.Time)

> sd\_nsw <- sd(recovery\_times$Recovery.Time)

> n\_nsw <- length(recovery\_times$Recovery.Time)

> mean\_israeli <- mean(israeli\_data$Recovery.Time)

> sd\_israeli <- sd(israeli\_data$Recovery.Time)

> n\_israeli <- length(israeli\_data$Recovery.Time)

> # Calculate the t-statistic

> t\_statistic <- (mean\_israeli - mean\_nsw) / sqrt((sd\_israeli^2 / n\_israeli) + (sd\_nsw^2 / n\_nsw))

> t\_statistic

***[1] 1.381394***

>

> # Calculate degrees of freedom

> df <- ((sd\_israeli^2 / n\_israeli + sd\_nsw^2 / n\_nsw)^2) / ((sd\_israeli^2 / n\_israeli^2) / (n\_israeli - 1) + (sd\_nsw^2 / n\_nsw^2) / (n\_nsw - 1))

> df

***[1] 25216.89***

>

> # Calculate the p-value

> p\_value <- 2 \* pt(abs(t\_statistic), df = df)

> p\_value

***[1] 1.83283***

>

> # Interpret the p-value

> alpha <- 0.05

> if (p\_value < alpha) {

+ cat("Reject the null hypothesis: There is a significant difference in mean recovery time between Israeli and NSW cohorts.\n")

+ } else {

+ cat("Fail to reject the null hypothesis: There is no significant difference in mean recovery time between Israeli and NSW cohorts.\n")

+ }

***Fail to reject the null hypothesis: There is no significant difference in mean recovery time between Israeli and NSW cohorts.***

> # Include the t-statistic and p-value in your report

> cat("t-statistic:", t\_statistic, "\n")

***t-statistic: 1.381394***

> cat("Degrees of freedom:", df, "\n")

***Degrees of freedom: 25216.89***

> cat("p-value:", p\_value, "\n")

***p-value: 1.83283***

**Question 2.**