**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.**

**1. Produce a plot of the exponential probability density function (1) for the values *y ∈* (0*,*10), for *v* = 1, *v* = 0*.*5 and *v* = 2. Ensure the graph is readable, the axis are labelled appropriately and a legend is included.**

***Answer:***

***Interpretation:***

***Approach:***

# Install and load ggplot2 if you haven't already

install.packages("ggplot2")

library(ggplot2)

# Define the range of values (y) and values of v

y <- seq(0, 10, by = 0.1)

v\_values <- c(1, 0.5, 2)

# Create a data frame for plotting

df <- data.frame(y = rep(y, times = length(v\_values)), v = rep(v\_values, each = length(y)))

# Calculate the probability density function

df$probability <- dexp(df$y, rate = 1 / df$v)

# Create the plot

ggplot(df, aes(x = y, y = probability, color = factor(v))) +

geom\_line() +

labs(x = "y", y = "Probability Density", color = "v") +

scale\_color\_discrete(name = "v") +

ggtitle("Exponential Probability Density Function for Different v Values") +

theme\_minimal()



**Question 2.**

2. Imagine we are given a sample of *n* observations **y** = (*y*1*,...,yn*). Write down the joint probability of this sample of data, under the assumption that it came from an exponential distribution with log-scale parameter *v* (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. *(hint: remember that these samples are independent and identically distributed.)*

***Answer:***

***Interpretation:***

***Approach:***

Certainly, let's break down the derivation of the joint probability (likelihood) of a sample of n observations from an exponential distribution in more detail.

Given a sample of n independent and identically distributed (i.i.d) observations, denoted as \(y = (y\_1, y\_2, \ldots, y\_n)\), where each \(y\_i\) is assumed to follow an exponential distribution with log-scale parameter v, i.e., \(y\_i \sim \text{Exp}(v)\), we want to find the likelihood function \(L(v|y)\).

1. \*\*Individual Probability Density Function (PDF):\*\*

The probability density function (PDF) for a single observation \(y\_i\) from an exponential distribution with log-scale parameter v is given by:

\[p(y\_i|v) = \exp(-e^{-vy\_i} - v)\]

2. \*\*Joint Probability (Likelihood):\*\*

Since the observations are independent, the joint probability (likelihood) of the entire sample y can be expressed as the product of the individual PDFs:

\[L(v|y) = \prod\_{i=1}^{n} p(y\_i|v)\]

3. \*\*Taking the Logarithm:\*\*

It's common to work with the log-likelihood because it simplifies calculations and numerical stability. Therefore, we take the natural logarithm (ln) of the likelihood:

\[\ln(L(v|y)) = \ln\left(\prod\_{i=1}^{n} p(y\_i|v)\right)\]

4. \*\*Using Logarithmic Properties:\*\*

Applying the properties of logarithms, we can rewrite the product of probabilities as the sum of logarithms:

\[\ln(L(v|y)) = \sum\_{i=1}^{n} \ln(p(y\_i|v))\]

This step allows us to work with the sum of logarithms instead of the product of probabilities, which can simplify calculations.

Now, you have the log-likelihood function \(\ln(L(v|y))\) for the sample y under the exponential distribution assumption with the log-scale parameter v. This log-likelihood is used in maximum likelihood estimation (MLE) and other statistical analyses. To find the MLE for v, you typically differentiate this log-likelihood with respect to v and set it equal to zero, solving for v.

##############

Certainly, here's the derivation of the joint probability (likelihood) of a sample of n observations from an exponential distribution in full detail:

\*\*Step 1: Individual Probability Density Function (PDF):\*\*

The probability density function (PDF) for a single observation \(y\_i\) from an exponential distribution with log-scale parameter v is given by:

\[p(y\_i|v) = \exp(-e^{-vy\_i} - v)\]

This PDF represents the probability of observing the value \(y\_i\) under the exponential distribution assumption with the parameter v.

\*\*Step 2: Joint Probability (Likelihood):\*\*

Now, we want to find the joint probability (likelihood) of a sample of n observations, denoted as \(y = (y\_1, y\_2, \ldots, y\_n)\), given the parameter v. Since these observations are assumed to be independent, the joint probability is the product of the individual probabilities:

\[L(v|y) = p(y\_1|v) \cdot p(y\_2|v) \cdot \ldots \cdot p(y\_n|v)\]

This equation represents the likelihood of observing the entire dataset y under the assumption that each observation \(y\_i\) follows an exponential distribution with log-scale parameter v.

\*\*Step 3: Taking the Logarithm:\*\*

To simplify calculations and take advantage of properties of logarithms, we take the natural logarithm (ln) of the likelihood:

\[\ln(L(v|y)) = \ln\left(p(y\_1|v) \cdot p(y\_2|v) \cdot \ldots \cdot p(y\_n|v)\right)\]

\*\*Step 4: Using Logarithmic Properties:\*\*

Using the properties of logarithms, we can simplify the expression. When you take the logarithm of a product, it becomes the sum of the logarithms:

\[\ln(L(v|y)) = \ln(p(y\_1|v)) + \ln(p(y\_2|v)) + \ldots + \ln(p(y\_n|v))\]

This step allows us to work with the sum of logarithms instead of the product of probabilities, making it easier to calculate and manipulate.

So, the final result is the log-likelihood function \(\ln(L(v|y))\) for the sample y under the exponential distribution assumption with the log-scale parameter v. This function is used in statistical analysis, especially for maximum likelihood estimation (MLE), where you find the value of v that maximizes this log-likelihood to estimate the parameter.

Certainly, let's continue with the steps:

\*\*Step 5: Simplifying the Log-Likelihood Function:\*\*

Now, let's express each term \(\ln(p(y\_i|v))\) in more detail. Recall the original PDF:

\[p(y\_i|v) = \exp(-e^{-vy\_i} - v)\]

Taking the natural logarithm of this PDF:

\[\ln(p(y\_i|v)) = -e^{-vy\_i} - v\]

\*\*Step 6: Substituting Back into the Log-Likelihood:\*\*

Now, substitute this expression for each term into the log-likelihood:

\[\ln(L(v|y)) = (-e^{-v y\_1} - v) + (-e^{-v y\_2} - v) + \ldots + (-e^{-v y\_n} - v)\]

\*\*Step 7: Combining Like Terms:\*\*

Next, combine the like terms. You have n identical terms, each with a negative exponential and a negative v:

\[\ln(L(v|y)) = -\left(e^{-vy\_1} + e^{-vy\_2} + \ldots + e^{-vy\_n}\right) - nv\]

\*\*Step 8: Simplifying Further:\*\*

Now, you can simplify the expression further. Note that the term inside the parentheses is the sum of exponential functions, which can be expressed as a sum:

\[\ln(L(v|y)) = -\left(\sum\_{i=1}^{n} e^{-vy\_i}\right) - nv\]

This is the simplified log-likelihood function for the sample y under the exponential distribution assumption with the log-scale parameter v.

\*\*Step 9: Maximizing the Log-Likelihood:\*\*

In statistical analysis, you would typically find the value of v that maximizes this log-likelihood function. This value of v is the maximum likelihood estimator (MLE) for the parameter v based on your observed data y. To do this, you would differentiate the log-likelihood with respect to v, set the derivative equal to zero, and solve for v. The MLE for v will be the value that maximizes this log-likelihood given your data.

Certainly, let's continue with the next steps:

\*\*Step 10: Maximizing the Log-Likelihood (Continued):\*\*

To find the maximum likelihood estimator (MLE) for the parameter v, we differentiate the log-likelihood with respect to v, set the derivative equal to zero, and solve for v:

\[\frac{d}{dv}\ln(L(v|y)) = -\frac{d}{dv}\left(\sum\_{i=1}^{n} e^{-vy\_i}\right) - n = 0\]

\*\*Step 11: Solving for the MLE v^:\*\*

Solving for v in the above equation will give us the MLE v^:

\[-\frac{d}{dv}\left(\sum\_{i=1}^{n} e^{-vy\_i}\right) - n = 0\]

This step involves calculus, and you would typically perform the differentiation and algebraic manipulation to solve for v^.

\*\*Step 12: Determining Bias and Variance:\*\*

To determine the approximate bias and variance of the MLE v^ for the exponential distribution, you would need to calculate the bias and variance of the estimator. The bias is the difference between the expected value of the estimator and the true parameter v. The variance measures the spread or uncertainty in the estimator.

The bias of the MLE is often difficult to find analytically and may require simulation or additional mathematical techniques. The variance can often be calculated analytically.

The bias and variance of an estimator provide insights into its performance. A lower bias and lower variance are generally desirable qualities for an estimator.

Please note that the exact calculations for bias and variance can be quite complex and may depend on the specific form of the likelihood function and the sample size. These calculations are typically covered in advanced statistical courses or require specialized statistical software.

Certainly, let's continue:

\*\*Step 13: Bias and Variance Calculation (Continued):\*\*

To further discuss the approximate bias and variance of the maximum likelihood estimator (MLE) \(v^\) for the exponential distribution, we can provide a general idea of how these concepts are typically assessed.

\*\*Bias:\*\*

The bias of an estimator is defined as the difference between the expected value (mean) of the estimator and the true parameter value. In mathematical terms:

\[ \text{Bias}(v^) = \mathbb{E}(v^) - v \]

To find the bias, you would need to calculate the expected value of \(v^\) using the distribution of the estimator. The bias gives you an idea of whether the estimator tends to overestimate or underestimate the true parameter.

\*\*Variance:\*\*

The variance of an estimator measures its spread or variability. In mathematical terms:

\[ \text{Var}(v^) = \mathbb{E}((v^ - \mathbb{E}(v^))^2) \]

To calculate the variance, you would need to compute the expected value of the squared difference between the estimator and its expected value. The variance provides insight into how much the estimator's values tend to deviate from their mean.

\*\*Bias-Variance Tradeoff:\*\*

In general, estimators aim for a balance between bias and variance. A low bias is desirable because it means the estimator is unbiased and, on average, provides estimates close to the true parameter. A low variance is also desirable because it means the estimator's values are tightly clustered around their mean.

However, there is often a tradeoff between bias and variance. Estimators with lower bias tend to have higher variance, and vice versa. The choice of estimator depends on the specific application and the tradeoff you are willing to accept.

Calculating the exact bias and variance for the MLE of the exponential distribution may involve more detailed mathematical analysis, and it can depend on the specific sample size and data characteristics. You might need to use specialized statistical software or conduct simulation studies to estimate bias and variance in practice.

In a report or analysis, you would typically summarize your findings regarding bias and variance, and these metrics help assess the performance of your estimator.

**Question 2.**

3. Take the negative logarithm of your likelihood expression and write down the negative loglikelihood of the data **y** under the exponential model with log-scale *v*. Simplify this expression.

***Answer:***

***Interpretation:***

***Approach:***

**Question 2.**

1. Derive the maximum likelihood estimator *v*ˆ for *v*. That is, find the value of *v* that minimises the negative log-likelihood. You must provide working.

***Answer:***

***Interpretation:***

***Approach:***