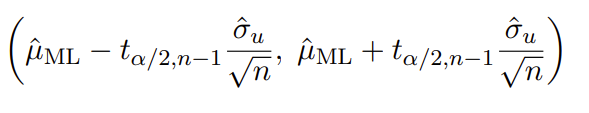
# Assignment 2 Rui Qin 30874157

## Question 1

### Question 1A

Below is the formula of the estimate using the t-distribution



Based on the formula we can calculate the result in RStudio

# Load the data

df <- read.csv("covid.19.ass2.2023.csv")

# Question 1A

df\_mean <- mean(df$Recovery.Time)

df\_sd <- sd(df$Recovery.Time)

df\_var <- var(df$Recovery.Time)

se <- df\_sd / sqrt(length(df$Recovery.Time))

t\_critical <- qt(1 - 0.05/2, df = length(df$Recovery.Time) - 1)

margin\_of\_error <- t\_critical \* se

upper\_limit <- df\_mean + margin\_of\_error

lower\_limit <- df\_mean - margin\_of\_error

df\_mean

cat("[", lower\_limit, ",", upper\_limit, "]\n")

> df\_mean

[1] 14.25797

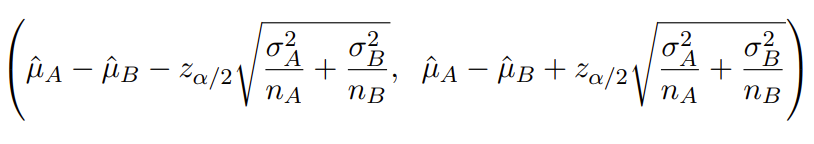
> cat("[", lower\_limit, ",", upper\_limit, "]\n")

[ 13.98935 , 14.52659 ]

The average duration for Covid-19 patients in New South Wales to recover is about 14.25797 days, supported by a 95% confidence interval spanning from 13.98935 to 14.52659 days.

### Question 1B

Below is the formula of confidence interval with difference of means



Based on the formula we can calculate the result in RStudio

# Question 1B

israeli\_df <- read.csv("israeli.covid.19.ass2.2023.csv")

israeli\_mean <- mean(israeli\_df$Recovery.Time)

israeli\_var <- var(israeli\_df$Recovery.Time)

israeli\_se <- israeli\_sd / sqrt(length(israeli\_df$Recovery.Time))

mean\_difference <- israeli\_mean - df\_mean

se\_difference <- sqrt((df\_var / length(df$Recovery.Time))

+ (israeli\_var / length(israeli\_df$Recovery.Time)))

margin\_of\_error <- t\_critical \* se\_difference

diff\_upper\_limit <- mean\_difference + margin\_of\_error

diff\_lower\_limit <- mean\_difference - margin\_of\_error

mean\_difference

cat("[", diff\_lower\_limit, ",", diff\_upper\_limit, "]\n")

> mean\_difference

[1] 0.391829

> cat("[", diff\_lower\_limit, ",", diff\_upper\_limit, "]\n")

[ -0.1643962 , 0.9480542 ]

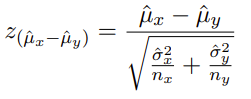
The mean difference is 0.391829. And we hold a 95% confidence that the variance in mean recovery times between the Israeli patients and those in New South Wales lies within the interval from -0. 1643962 days to 0.9480542 days.

### Question 1C

Based on the question we are working on testing the difference of means with unknown variances

Hypotheses:

* Null Hypothesis (H0): The average recovery time for the Israeli cohort matches that of the NSW cohort at the population level.
  + H0: μ1 = μ2
* Alternative Hypothesis (H1): The average recovery time for the Israeli cohort does not match that of the NSW cohort at the population level.
  + H1: μ1 ≠ μ2



# Question 1C

test\_statistic <- (df\_mean - israeli\_mean) / se\_difference

2 \* pnorm(-abs(test\_statistic))

> 2 \* pnorm(-abs(test\_statistic))

[1] 0.1671578

The p-value at 0.1675329, provides evidence against the Alternative hypothesis. This supports the assertion that there is no distinction in the average recovery times between Israeli and NSW patients.

# Question 2

### Question 2A

For each y value, we calculate with its v value put it into the same data set, and finally print it out with ggplot.

# Question 2A

library(ggplot2)

y <- seq(0, 10, by = 0.001)

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

data <- data.frame()

for (v in v\_values) {

density <- exp((-exp(-v) \* y) - v)

data <- rbind(data, data.frame(y = y, density = density, v = as.factor(v)))

}

ggplot(data, aes(x = y, y = density, color = v)) +

geom\_line() +

labs(

x = "y",

y = "Probability Density",

title = "Exponential Probability Density Function",

) +

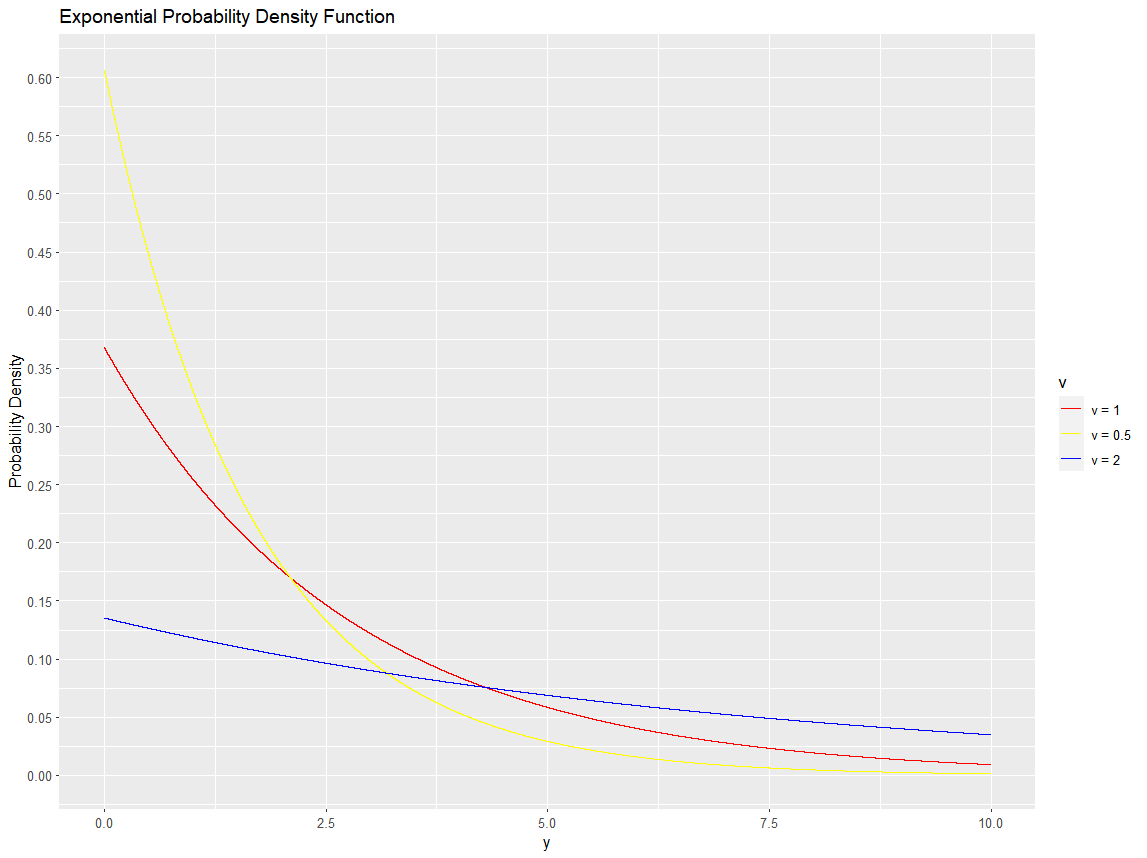
scale\_color\_manual(

values = c("1" = "red", "0.5" = "yellow", "2" = "blue"),

labels = c("v = 1", "v = 0.5", "v = 2")

)+

scale\_y\_continuous(breaks = seq(0, 0.7, by = 0.05))



### Question 2B

L(v ; y) = ∏( i=1, n) p(yi​∣v) = ∏( i=1, n) exp ((−e ^ −v) yi − v)

= exp [ ((− e ^ −v) y1 – v) + ((− e ^ −v) y2 – v) + … + ((− e ^ −v) yn – v) ]

= exp [ (− e ^ −v) y1 + (− e ^ −v) y2 + … + (− e ^ −v) yn – nv ]

= exp [ (− e ^ −v) (y1 + y2 + … + yn) – nv ]

= exp [ -∑(i=1, n) (e ^ −v) yi – nv ]

= exp [ - ((e ^ −v) ∑(i=1, n) yi + nv ) ]

### Question 2C

− logL(v ; y)

= - log { exp [ - ((e ^ −v) ∑(i=1, n) yi + nv ) ] }

= - (- ((e ^ −v) ∑(i=1, n) yi + nv ))

= (e ^ −v) ∑(i=1, n) yi + nv

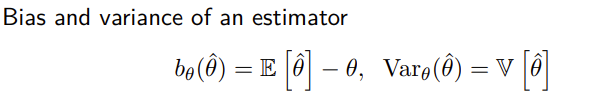
### Question 2D

NL (v ; y) = − logL(v ; y) = (e ^ −v) ∑(i=1, n) yi + nv

d/dv NL (v ; y) = - e ^ −v ∑(i=1, n) yi + n = 0

* e ^ −v ∑(i=1, n) yi = n
* e ^ −v = - n / ∑(i=1, n) yi
* - v = ln ( n / ∑(i=1, n) yi )
* V\_estimator = - ln ( n / ∑(i=1, n) yi )

### Question 2E



In this case, v is the true parameter, and V\_estimator is the estimator

E[ V\_estimator ]

= E[ - ln ( n / ∑(i=1, n) yi ) ]

= - E[ ln ( n / ∑(i=1, n) yi ) ]

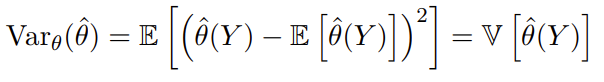
= - E[ ln(n) – ln (∑(i=1, n) yi) ]

= - ln(n) + E [ ln (∑(i=1, n) yi) ]

∵ E [ Y ] = e ^ v

∴ - E[ ln ( n / ∑(i=1, n) yi ) ] = - ln ( n/( e^v ) ) = - ln(n) + v

∴ Bias( V\_estimator ) = - ln(n) + v – v = - ln(n)



Var( V\_estimator ) = E[ ( V\_estimator - E[ V\_estimator ] ^ 2) ]

= E[ ( v – (- ln(n)))^2 ]

= E[ ( ln(n) - ln ( n / ∑(i=1, n) yi )))^2 ]

= E[ ( ln( ∑(i=1, n) yi ))^2 ]

= ln^2( E[∑(i=1, n) yi] )

= ln^2 (e^v)

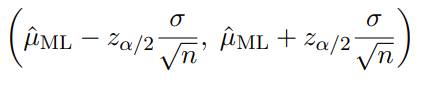
= (ln(e^v))^2

= v^2

## Question 3

### Question 3A

We can calculate based on the question so below is the formula we going to use.



# Question 3A

# 124 volunteer, 80 to the right

n <- 124

x <- 80

p <- x / n

# Calculate the standard error

se <- sqrt((p \* (1 - x / n)) / n)

# Calculate the confidence interval

lower <- p - 1.96 \* se

upper <- p + 1.96 \* se

p

cat("[", lower, ",", upper, "]\n")

> p

[1] 0.6451613

> cat("[", lower, ",", upper, "]\n")

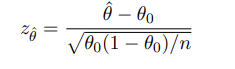
[ 0.5609452 , 0.7293773 ]

The estimated proportion of humans turning their heads to the right is 0.6451613. We are 95% confident that the true population mean within this group falls between 0.5609452 and 0.7293773.

### Question 3B

Hypotheses:

* Null Hypothesis (H0​): The proportion of couples turning their heads to the right when kissing is equal to 0.5.
* Alternative Hypothesis (H1​): The proportion of couples turning their heads to the right when kissing is not equal to 0.5.



Based on the z-score formula we can calculate the P value

# Question 3B

p\_0 <- 0.5

z <- (p - p\_0) / sqrt((p\_0 \* (1 - p\_0)) / n)

p\_value <- 2 \* (1 - pnorm(abs(z)))

z

p\_value

> z

[1] 3.232895

> p\_value

[1] 0.001225424

We accept the alternative hypothesis over the null hypothesis, indicating that humans do not turn their heads to a specific side when kissing.

### Question 3C

# Question 3C

binom.test(x, n, p = 0.5)$p.value

> binom.test(x, n, p = 0.5)$p.value

[1] 0.001564734

The p-value calculated for the preference for head tilting when kissing is approximately 0.0012 (approximate p-value) or 0.0016 (exact p-value). These p-values suggest that the chance of observing a preference stronger if the sample is larger, but it is still very low which gives the same conclusion.

### Question 3D

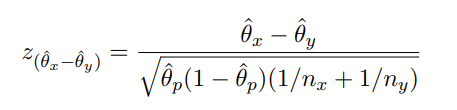
To test whether the proportion of people who prefer using their right-hand matches the proportion of people who tilt their head to the right when kissing, we consider individuals who are right-handed in equal proportion when tilting their head to the right during a kiss as the null hypothesis.

Hypotheses:

* Null Hypothesis (H0​): θ right\_hand = θ right\_head
* Alternative Hypothesis (H1​): θ right\_hand != θ right\_head

H0 : θright\_hand = θright\_head vs H1 : θA != θB

We can use the below formula to calculate the z-score



# Question 3D

p\_head <- x/n

p\_hand <- 83/(83+17)

p\_hat <- (x+83)/(n+83+17)

z <- (p\_hand - p\_head) / sqrt(p\_hat\*(1 - p\_hat)\*(1/n + 1/(83+17)))

z

2 \* (pnorm(-abs(z)))

> z

[1] 3.089364

> 2 \* (pnorm(-abs(z)))

[1] 0.002005856

This p-value (0.002) is smaller than 0.05. So, we would reject the null hypothesis positing that the rate of right-handedness in the population is equivalent to the preference for turning heads to the right during kissing. The results suggest that there is a relationship between right-handedness and the preference for head-turning direction when kissing.