Question 1

1. Using the following R code to calculate the $\hat{\mu}_{ML}$

```r
dogbite <- read.csv("dogbites.fullmoon.csv", header = TRUE)
dogbite_fullmoon <- dogbite[dogbite$is.full.moon == 1,]

my_estimates <- function(X) filter out data which is not full moon
{
  n = length(X)

  retval = list()
  # Calculate the sample mean
  retval$mu_ml = sum(X)/n

  # Calculate the squared deviations around the mean
  e2 = (X - retval$m)^2

  # Calculate the two estimates of variance
  retval$var_ml = sum(e2)/n
  retval$var_u  = sum(e2)/(n-1)

  return(retval)
}

my_estimates(dogbite_fullmoon$daily.dogbites)
```

The code returns the following output.

```
> my_estimates(dogbite_fullmoon$daily.dogbites)
$mu_ml
[1] 4.230769

$var_ml
[1] 6.023669

$var_u
[1] 6.525641
```

∴ The estimate of the average number of daily dog-bites for days on which there was a full moon is 4.23

Using the following R code to calculate the 95% confidence interval using t-distribution

```r
calcCI <- function(y, alpha)
{
  n = length(y)

  retval = list()

  # Simple error checking
  if (alpha <= 0 || alpha >= 1)
  {
    stop("Alpha must be a value greater than 0 and less than 1")
  }

  # Calculate the sample mean and (unbiased) estimate of variance
  retval$mu.hat = mean(y)
```

```r
calcCI <- function(y, alpha)
{
  n = length(y)

  retval = list()

  # Simple error checking
  if (alpha <= 0 || alpha >= 1)
  {
    stop("Alpha must be a value greater than 0 and less than 1")
  }

  # Calculate the sample mean and (unbiased) estimate of variance
  retval$mu.hat = mean(y)

  # the "var()" function returns the unbiased estimate of variance
  retval$sigma2.hat = var(y)

  # Calculate the multiplier for our CI based on t-distribution (un
  t = qt(1-alpha/2, n-1)

  # return the interval
  retval$CI = retval$mu.hat + c(-t * sqrt(retval$sigma2.hat/n),
                                t * sqrt(retval$sigma2.hat/n))
  return(retval)
}

est = calcCI(dogbite_fullmoon$daily.dogbites, 0.05)

est$mu.hat  # Estimated mean
est$CI
```

The code returns the following output

```
> est$mu.hat  # Estimated mean
[1] 4.230769
> est$CI
[1] 2.687080 5.774458
```

∴ The estimated mean daily dogbites when full moon is 4.23 times. We are 95% confident the population mean daily dogbites when full moon is between 2.69 and 5.77

2. Using the following code to calculate the difference in mean of dogbites in full moon and not full moon days. And the 95% confidence interval for the difference:

```r
### 2 ###
dogbite_notfullmoon <- dogbite[dogbite$is.full.moon == 0,]

est_fullmoon = calcCI(dogbite_fullmoon$daily.dogbites, alpha=0.05)
est_notfullmoon = calcCI(dogbite_notfullmoon$daily.dogbites, alpha=0.05)

n1 = length(dogbite_fullmoon)
n2 = length(dogbite_notfullmoon)

# get the difference
diff = est_fullmoon$mu.hat - est_notfullmoon$mu.hat

# calculate standard error
se.diff = sqrt(est_fullmoon$sigma2.hat/n1 + est_notfullmoon$sigma2.hat/n2)

# calculate the 95% CI
CI.diff = diff + c(-1.96*se.diff, 1.96*se.diff)
```

```r
### 2 ###
dogbite_notfullmoon <- dogbite[dogbite$is.full.moon == 0,]

est_fullmoon = calcCI(dogbite_fullmoon$daily.dogbites, alpha=0.05)
est_notfullmoon = calcCI(dogbite_notfullmoon$daily.dogbites, alpha=0.05)

n1 = length(dogbite_fullmoon)
n2 = length(dogbite_notfullmoon)

# get the difference
diff = est_fullmoon$mu.hat - est_notfullmoon$mu.hat

# calculate standard error
se.diff = sqrt(est_fullmoon$sigma2.hat/n1 + est_notfullmoon$sigma2.hat/n2)

# calculate the 95% CI
CI.diff = diff + c(-1.96*se.diff, 1.96*se.diff)

diff
CI.diff
```

The code returns the following output:

```
> diff
[1] -0.2842993
> CI.diff
[1] -6.364823  5.796224
```

∴ The estimated difference in mean dogbites between full moon days and not full moon days is -0.28. We are 95% confident the population mean difference in daily dogbites is between -6.36 up to 5.80. As the interval includes zero, we cannot rule out the possibility of there being no difference at a population level between full moon days and not full moon days.

3. Let μ_1 be the mean of dog bites on full moon days.
And μ_2 be the mean of dog bites on not full moon days.
We want to test:

$$H_0 : \mu_1 = \mu_2$$

vs

$$H_A : \mu_1 > \mu_2$$

Use the following R code to perform the hypothesis test

```r
dogbite_1<-dogbite[dogbite[,2]==1,1]
dogbite_0<-dogbite[dogbite[,2]==0,1]

t.test(dogbite_1,dogbite_0,var.equal = F)
```

The code have following output:

```
> dogbite_1<-dogbite[dogbite[,2]==1,1]
> dogbite_0<-dogbite[dogbite[,2]==0,1]
> t.test(dogbite_1,dogbite_0,var.equal = F)

        Welch Two Sample t-test

data:  dogbite_1 and dogbite_0
t = -0.38802, df = 13.722, p-value = 0.704
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.858745  1.290147
sample estimates:
mean of x mean of y
 4.230769  4.515068
```

∵ This is a two-sided test

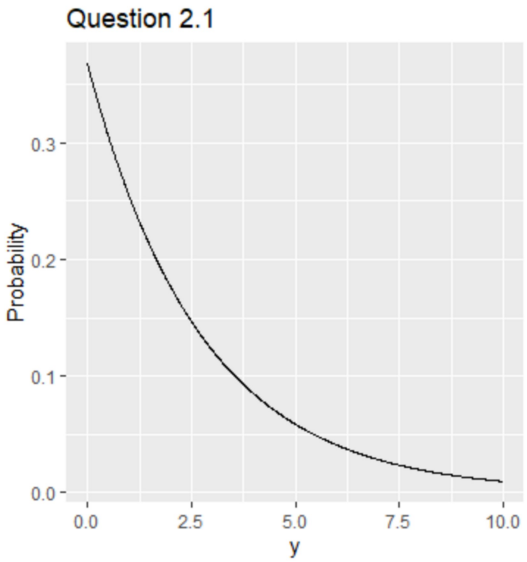
∴ p-value = 0.704 / 2 = 0.302

Question 2

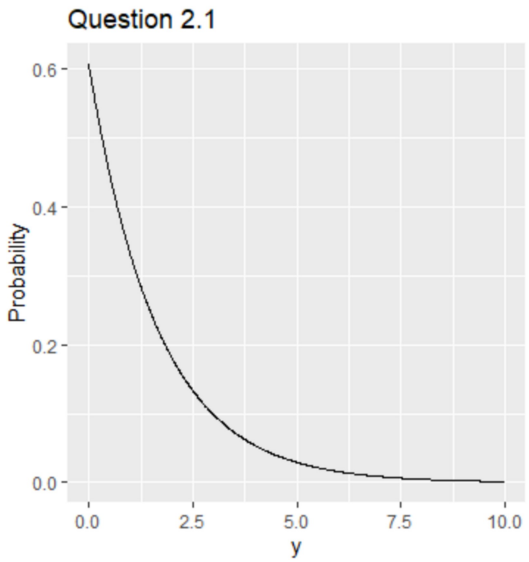
1. Use the following R code to plot the exponential pdf.

```r
y <- seq(0, 10, length.out = 1000)

v = 1

prob = exp(-exp(-v) * y - v)

dat <- data.frame(y = y, prob =prob)

library(ggplot2)

ggplot(dat, aes(x = y, y = prob)) + geom_line() + xlab("y") + ylab("Probability") + ggtitle("Question 2.1")
```

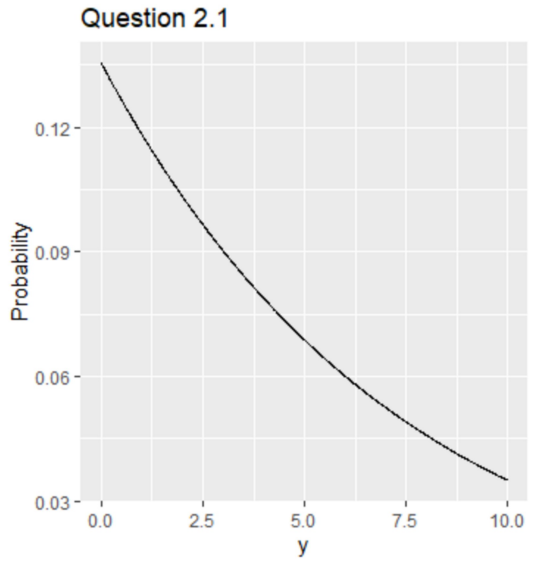
The plot for v=1:



The plot for v= 0.5:



The plot for $v = 2$:



2. $p(y|v) = \prod_{i=1}^{n} p(y_i|v)$

$= (\exp(-e^{-v}y_1 - v)) \cdot (\exp(-e^{-v}y_2 - v)) \cdot \ldots \cdot (\exp(-e^{-v}y_n - v))$

$= \exp[(-e^{-v}y_1 - v) + (-e^{-v}y_2 - v) + \ldots + (-e^{-v}y_n - v)]$

$= \exp[(-e^{-v}y_1 + (-e^{-v}y_2) + \ldots + (-e^{-v}y_n) - nv]$

$= \exp[-e^{-v}(y_1 + y_2 + \ldots + y_n) - nv]$

$= \exp(-e^{-v}\sum_{i=1}^{n} y_i - nv)$

3. Applying a negative logarithm of the likelihood:

$-\log(\exp(-e^{-v}\sum_{i=1}^{n} y_i - nv))$

$= -(-e^{-v}\sum_{i=1}^{n} y_i - nv)$

$= e^{-v}\sum_{i=1}^{n} y_i + nv$

4. To derive the maximum likelihood estimator $\hat{v}$, we minimise the equation by applying derivative to it

4. To derive the maximum likelihood estimator $\hat{v}$, we minimise the equation by applying derivative to it

$$\frac{dL(y|v)}{dv} = -ve^{-v}\sum_{i=1}^{n} y_i + n$$

Let the equation equal to zero to minimise v

$$-ve^{-v}\sum_{i=1}^{n} y_i + n = 0$$

$$-ve^{-v}\sum_{i=1}^{n} y_i = -n$$

$$ve^{-v}\sum_{i=1}^{n} y_i = n$$

$$ve^{-v} = \frac{n}{\sum_{i=1}^{n} y_i}$$

$$\frac{v}{e^{v}} = \frac{n}{\sum_{i=1}^{n} y_i}$$

$$\frac{e^{v}}{v} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$\because E[Y] = e^{v} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$\therefore v$ must be 1

$\therefore \hat{v} = 1$

5. $b_{\hat{v}} = 1 - e^{v}$

$$V[\bar{v}] = V\left[\frac{y_1 + y_2 + \cdots + y_n}{n}\right]$$

$$= \frac{V[Y]}{n}$$

$$= \frac{e^{2v}}{n}$$

Question 3

1. $\because$ The sample size $n = 124$

80 turns their head to the right.

Using the following R code:

```
binconf(80, 124,alpha = 0.05)
```

The output is:

```
> binconf(80, 124,alpha = 0.05)
  PointEst     Lower     Upper
0.6451613 0.5577452 0.7238536
```

∴ The estimate of perference for human turning their head to the right when kissing is 0.645.
We are 95% confident the population mean is between 0.558 and 0.724

2. According to the question, there doesn't exist a preference for human turning their head when kissing can be expressed as $\theta_R = \theta_L = 0.5$, so we can write our hypothesis as

$H_0 : \theta_R = 0.5$

vs

$H_A : \theta_R \neq 0.5$

From question 1 we know that $\hat{\theta}_R = 0.645$

Using this result, we can calculate the $Z_{\hat{\theta}}$

$$Z_{\hat{\theta}} = \frac{\hat{\theta}_R - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} = \frac{0.645 - 0.5}{\sqrt{0.5(0.5)/124}} = \frac{0.145}{0.045} = 3.22$$

Due to $H_0 : \theta_R = \theta_0$ vs $H_A : \theta \neq \theta_0$ from our hypothesis

$p \approx 2P(Z < -|z_{\hat{\theta}}|)$

≈ 0.00124

∴ In conclusion, there's enough evidance to reject the null hypothesis which says that the chance of human turning head to right over other as strong as we seen. From this, we could conclude that there is a preference of turning heads to right while human is kissing.

while human is kissing.

3. Let θ_L be the preference of human turning head to right while kissing, θ_R be the oppusite as the previous question.
∴ We can have our hypothesis:

$$H_0 : \theta_R = 0.5$$

vs

$$H_A : \theta_R \neq 0.5$$

Using the following R code to perform a binominal hypothesis test.

```
11  binom.test(80,124,p=0.5,alternative='two.sided')
```

The output is:

```
> binom.test(80,124,p=0.5,alternative='two.sided')

        Exact binomial test

data:  80 and 124
number of successes = 80, number of trials = 124, p-value =
0.001565
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5542296 0.7289832
sample estimates:
probability of success
             0.6451613
```

The p-value = 0.0016

4. According to the question, we are trying to test if right / left-handedness could affect the preference of head turning.
So let θ_1 be the estimation of preference of turning head to right, and θ_2 be the estimation of right-handedness.
We can test the hypothesis as

$$H_0 : \theta_1 = \theta_2$$

vs

$$H_0 : \theta_1 = \theta_2$$

vs

$$H_A : \theta_1 \neq \theta_2$$

$n_1 = 124 \qquad n_2 = 100$

$\hat{\theta}_1 = \frac{80}{124} \qquad \hat{\theta}_2 = \frac{83}{100}$

∴ pooled estimate of θ

$$\hat{\theta}_p = \frac{80 + 83}{124 + 100} = 0.728$$

$$Z(\hat{\theta}_1 - \hat{\theta}_2) = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\theta_p(1-\theta_p)(1/n_1 + 1/n_2)}}$$

$$= \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{0.728 \times 0.272 \times 0.018}}$$

$$= \frac{-0.185}{0.060}$$

$$= -3.083$$

Due to $H_0 : \theta_1 = \theta_2$ vs $H_A : \theta_1 \neq \theta_2$ from our hypothesis.

$p \approx 2P(Z < -|z(\hat{\theta}_1 - \hat{\theta}_2)|)$

≈ 0.002

∴ In conclusion, we have enough evidence to reject the null hypothesis which says that the chance of human turning head to one side is affected by the right/left handedness over other as strong as we seen. From this, we could conclude that there is no relation between human turning head to a specific side and right/left handedness.

5. The sample size could be bigger, and country selection could be broader instead of just US, Germany and Turkey.

5. The sample size could be bigger, and country selection could be broader instead of just US, Germany and Turkey.

Question 4.

1. Load the data into R, use the following code to fit a linear model to find the predictors associated with fuel efficiency.

```
getwd()
setwd("C:/Users/yjb13/OneDrive/Monash Uni/Year 2/FIT2086 Modelling for Data Analysis/Assignment2")

fuel <- read.csv("fuel2017-20.csv", header = TRUE)

summary(fuel)

fit<-lm(Comb.FE ~.,data=fuel)

summary(fit)
```

The output is:

```
> fit<-lm(Comb.FE ~.,data=fuel)
> summary(fit)

Call:
lm(formula = Comb.FE ~ ., data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2229 -0.9985 -0.0975  0.7149 11.4355

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.003e+02  7.241e+01  -2.766  0.00573 **
Model.Year               1.074e-01  3.588e-02   2.993  0.00279 **
Eng.Displacement        -1.287e+00  8.674e-02 -14.832  < 2e-16 ***
No.Cylinders             2.569e-03  5.767e-02   0.045  0.96447
AspirationOT            -2.471e-01  6.343e-01  -0.390  0.69692
AspirationSC            -1.015e+00  1.995e-01  -5.089 3.94e-07 ***
AspirationTC            -1.268e+00  1.085e-01 -11.685  < 2e-16 ***
AspirationTS            -1.183e+00  4.215e-01  -2.807  0.00506 **
No.Gears                -1.745e-01  2.534e-02  -6.888 7.58e-12 ***
Lockup.Torque.ConverterY -7.859e-01  9.506e-02  -8.267 2.48e-16 ***
Drive.SysA              -3.829e-02  1.294e-01  -0.296  0.76725
Drive.SysF               1.512e+00  1.438e-01  10.511  < 2e-16 ***
Drive.SysP              -4.435e-01  2.427e-01  -1.827  0.06781 .
Drive.SysR               9.319e-02  1.243e-01   0.750  0.45349
Max.Ethanol             -6.993e-03  2.490e-03  -2.808  0.00503 **
Fuel.TypeGM              5.696e-01  3.752e-01   1.518  0.12913
Fuel.TypeGP              5.024e-01  1.163e-01   4.321 1.63e-05 ***
Fuel.TypeGPR             2.066e-01  1.199e-01   1.723  0.08500 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.619 on 1982 degrees of freedom
Multiple R-squared:  0.6639,    Adjusted R-squared:  0.661
F-statistic: 230.3 on 17 and 1982 DF,  p-value: < 2.2e-16
```

According to the table, predictors with p-value < 0.5 could be considered as associated with fuel efficiency.

According to the table, predictors with p-value < 0.5 could be considered as associated with fuel efficiency.
Which are all predictors except number of cylinders, Aspiration (Other) and Drive System (All-wheel).

Engine Displacement, Aspiration (Turbocharged) and Drive system (Front-wheel) appears to be the strongest predictors to fuel efficiency due to they have smallest p-value.

2. ∵ There are 17 predictors and $\alpha = 0.05$
   $0.05 / 17 = 0.003$
   ∴ Yes, we should filter out predictors with p-value < 0.03 in stead of 0.05

3. According to the table, with the model year grow by 1, fuel efficiency increase by 0.1074.
With number of Gear increase by 1, fuel efficiency will decrease by 0.1745.

4. Use the following R code to fit a new model:

```
fit.sw.bic = step(fit, k = log(length(fuel$Model.Year)))
fit.sw.bic%>%summary()
```

The output is:

```
> fit.sw.bic%>%summary()

Call:
lm(formula = Comb.FE ~ Model.Year + Eng.Displacement + Aspiration + 
    No.Gears + Lockup.Torque.Converter + Drive.Sys + Max.Ethanol, 
    data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1799 -1.0033 -0.0835  0.6849 11.4237
```

```
> fit.sw.bic%>%summary()

Call:
lm(formula = Comb.FE ~ Model.Year + Eng.Displacement + Aspiration +
    No.Gears + Lockup.Torque.Converter + Drive.Sys + Max.Ethanol,
    data = fuel)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1799 -1.0033 -0.0835  0.6849 11.4237

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -2.097e+02  7.261e+01  -2.887 0.003927
Model.Year                1.120e-01  3.598e-02   3.113 0.001881
Eng.Displacement         -1.253e+00  3.698e-02 -33.897  < 2e-16
AspirationOT             -1.014e-01  6.294e-01  -0.161 0.872034
AspirationSC             -7.208e-01  1.866e-01  -3.863 0.000116
AspirationTC             -1.093e+00  9.018e-02 -12.116  < 2e-16
AspirationTS             -1.100e+00  4.098e-01  -2.685 0.007309
No.Gears                 -1.606e-01  2.493e-02  -6.442 1.47e-10
Lockup.Torque.ConverterY -7.999e-01  9.341e-02  -8.563  < 2e-16
Drive.SysA                7.188e-02  1.242e-01   0.579 0.562843
Drive.SysF                1.545e+00  1.401e-01  11.027  < 2e-16
Drive.SysP               -5.454e-01  2.376e-01  -2.295 0.021813
Drive.SysR                1.689e-01  1.231e-01   1.372 0.170300
Max.Ethanol              -8.184e-03  2.460e-03  -3.327 0.000893

(Intercept)              **
Model.Year               **
Eng.Displacement         ***
AspirationOT
AspirationSC             ***
AspirationTC             ***
AspirationTS             **
No.Gears                 ***
Lockup.Torque.ConverterY ***
Drive.SysA
Drive.SysF               ***
Drive.SysP               *
Drive.SysR
Max.Ethanol              ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 1986 degrees of freedom
Multiple R-squared:  0.6603,    Adjusted R-squared:  0.6581
F-statistic:   297 on 13 and 1986 DF,  p-value: < 2.2e-16
```

5. The BIC model suggest we should choose cars with later year of sale, Drive system in all-wheel mode or front-wheel mode or rear-wheel mode to improve the fuel efficiency.

6.

a. This new car located in the first row

∴ Using the following R code.

```
23 fuel_test <- read.csv("fuel2017-20.test.csv", header = TRUE)
24
25 new_car <- predict(fit.sw.bic, newdata = fuel_test[1,], interval = "predict")
```

The ouwput is:

```
> new_car <- predict(fit.sw.bic, newdata = fuel_test[1,], interval = "predict")
> new_car
       fit      lwr      upr
1 8.467534 5.273121 11.66195
```

∴ The mean fuel efficiency for this new car is 8.47 km/l
We are 95% confident that the prediction mean is between 5.27 km/l and 11.66 km/l

b. The current car's fuel efficiency is 8.5 km/l which is greater than the prediction of this new car 8.47 km/l
∴ The current car has better fuel efficiency.