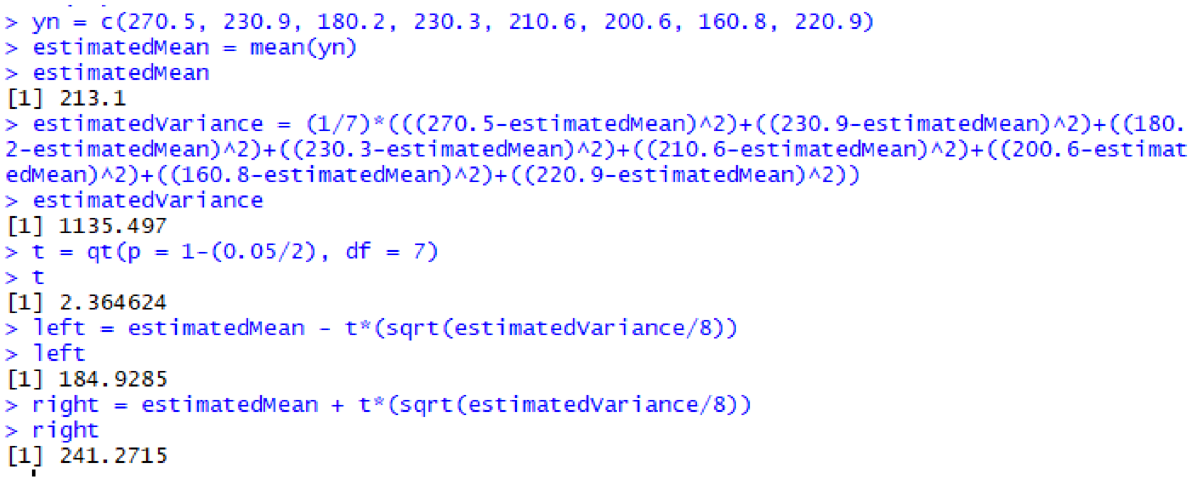
Question 1

The median house prices collected on those suburbs which do not have houses along the Charles River were:

$$y_n = (270.5, 230.9, 180.2, 230.3, 210.6, 200.6, 160.8, 220.9)$$

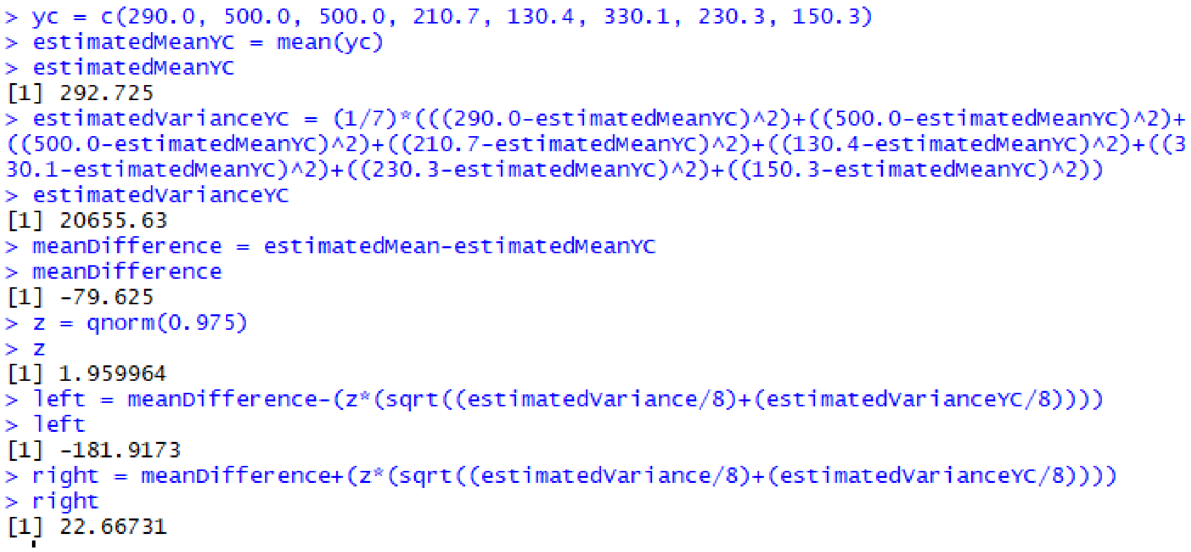
1. Calculate an estimate of the average median house price for suburbs of Boston which do not have houses on the Charles River. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. [4 marks]



```
> yn = c(270.5, 230.9, 180.2, 230.3, 210.6, 200.6, 160.8, 220.9)
> estimatedMean = mean(yn)
> estimatedMean
[1] 213.1
> estimatedVariance = (1/7)*(((270.5-estimatedMean)^2)+((230.9-estimatedMean)^2)+((180.
2-estimatedMean)^2)+((230.3-estimatedMean)^2)+((210.6-estimatedMean)^2)+((200.6-estimat
edMean)^2)+((160.8-estimatedMean)^2)+((220.9-estimatedMean)^2))
> estimatedVariance
[1] 1135.497
> t = qt(p = 1-(0.05/2), df = 7)
> t
[1] 2.364624
> left = estimatedMean - t*(sqrt(estimatedVariance/8))
> left
[1] 184.9285
> right = estimatedMean + t*(sqrt(estimatedVariance/8))
> right
[1] 241.2715
```

Therefore, with the estimated mean median house price for suburbs of Boston which do not have houses on the Charles River, we are 95% confident the population mean median house price for this group is between 184.9285 and 241.2715. However, this result is derived from an estimated variance out of our small sample size which does not truly cover the true parameter value for 95% of possible samples, hence the result brings with it a degree of uncertainty.

2. The same real estate agency has also collected median houses prices for eight suburbs that do have houses on the Charles River. These are:
$$y_c = (290.0, 500.0, 500.0, 210.7, 130.4, 330.1, 230.3, 150.3)$$
The real estate agents want to know if there is a difference, at the population level, between average median houses prices in suburbs that do, and do not, have houses on the Charles River. Use this sample to answer this question. Using the approximate method for difference in means with unknown variances presented in Lecture 4, calculate the estimated mean difference in median house price between the suburbs with and without houses on the Charles River, and a 95% confidence interval for this difference. Summarise/describe your results appropriately. Show working as required. [4 marks]



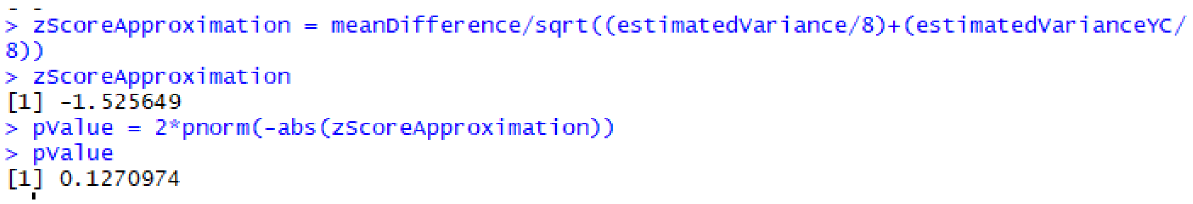
```
> yc = c(290.0, 500.0, 500.0, 210.7, 130.4, 330.1, 230.3, 150.3)
> estimatedMeanYC = mean(yc)
> estimatedMeanYC
[1] 292.725
> estimatedVarianceYC = (1/7)*(((290.0-estimatedMeanYC)^2)+((500.0-estimatedMeanYC)^2)+
((500.0-estimatedMeanYC)^2)+((210.7-estimatedMeanYC)^2)+((130.4-estimatedMeanYC)^2)+((3
30.1-estimatedMeanYC)^2)+((230.3-estimatedMeanYC)^2)+((150.3-estimatedMeanYC)^2))
> estimatedVarianceYC
[1] 20655.63
> meanDifference = estimatedMean-estimatedMeanYC
> meanDifference
[1] -79.625
> z = qnorm(0.975)
> z
[1] 1.959964
> left = meanDifference-(z*(sqrt((estimatedVariance/8)+(estimatedVarianceYC/8))))
> left
[1] -181.9173
> right = meanDifference+(z*(sqrt((estimatedVariance/8)+(estimatedVarianceYC/8))))
> right
[1] 22.66731
```

Therefore, with the estimated difference in mean between median houses prices in suburbs that do (sample size, n = 8) and do not (sample size, n = 8) have houses on the Charles River -79.625, we are 95% confident the population mean difference in median houses prices is between -181.9173 (median houses prices is lower in suburbs that do not have houses on the Charles River) up to 22.66731 (median houses prices is higher in suburbs that do not have houses on the Charles River). As the interval includes zero, we cannot rule out the probability of there being no difference at a population level between median houses prices in suburbs that do and do not have houses on the Charles River.

3. Test the hypothesis that the two groups are the same. Using the approximate hypothesis test for difference in means with unknown variances presented in Lecture 5, calculate an appropriate p-value under the null hypothesis that the average median house price for suburbs with and without houses on the Charles River are the same, showing working as required. Interpret this p-value; do you think the two groups of suburbs (with and without houses on the Charles River) have the different average median house price at the population level? [2 marks]

$$\begin{aligned} H_0 &: median\ house\ price\ for\ suburbs\ without\ houses\ on\ the\ Charles\ River \\ &= median\ house\ price\ for\ suburbs\ with\ houses\ on\ the\ Charles\ River \end{aligned}$$

$$\begin{aligned} H_1 &: median\ house\ price\ for\ suburbs\ without\ houses\ on\ the\ Charles\ River \\ &\neq median\ house\ price\ for\ suburbs\ with\ houses\ on\ the\ Charles\ River \end{aligned}$$

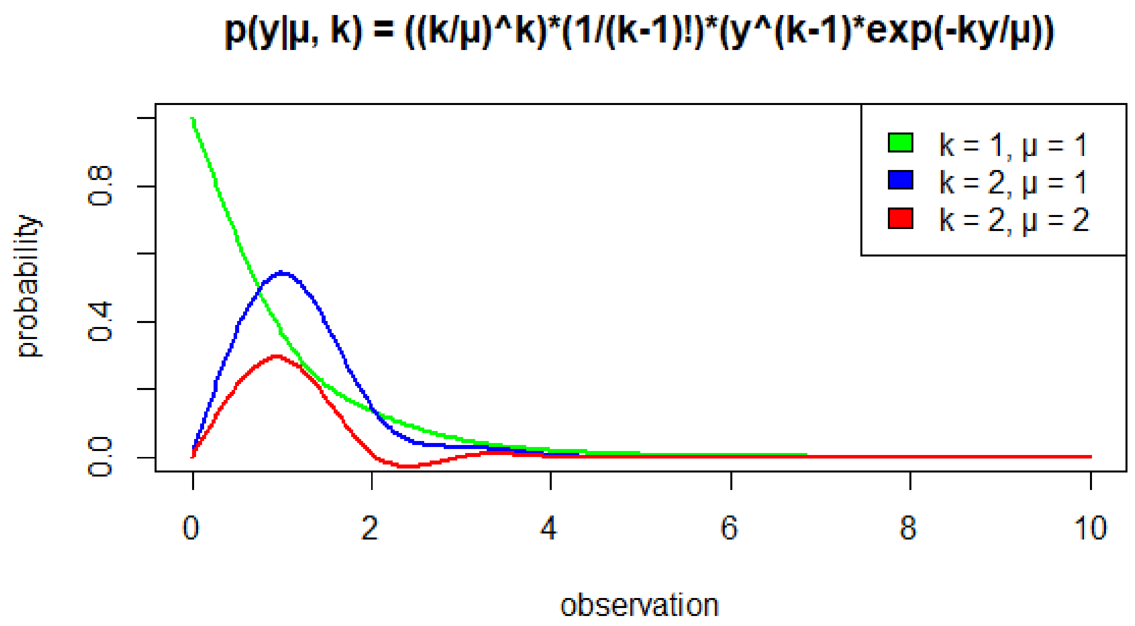


```
> zScoreApproximation = meanDifference/sqrt((estimatedVariance/8)+(estimatedVarianceYC/
8))
> zScoreApproximation
[1] -1.525649
> pValue = 2*pnorm(-abs(zScoreApproximation))
> pValue
[1] 0.1270974
```

The p-value (pValue) states that if there was no difference at the population level in median houses prices between suburbs with and without houses on the Charles River to expect a difference as large as the one we have observed: 12.7% ($> \alpha = 0.05/2$) at the time we drew two samples of size n = 8 from these populations. So the p-value (pValue) suggests that we do not have enough evidence to reject the null hypothesis that the median house price for suburbs without houses on the Charles River is the same as the median house price for suburbs with houses on the Charles River.

Question 2

1. Produce a plot of the gamma probability density function (1) for the values $y \in (0, 10)$, for $(k = 1, \mu = 1), (k = 2, \mu = 1)$ and $(k = 2, \mu = 2)$. Ensure the graph is readable, the axis are labelled appropriately and a legend is included. [2 marks]



```
> x = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
> first = dgamma(x = x, shape = 1, scale = 1)
> plot(spline(x, first, method = "n", n = 250), type = "l", main = "p(y|μ,
k) = ((k/μ)^k)*(1/(k-1)!)*(y^(k-1)*exp(-ky/μ))", xlab = "observation", ylab
 = "probability", col = "green", lwd = 2)
> second = dgamma(x = x, shape = 2, scale = 1/2)
> lines(spline(x, second, method = "n", n = 250), type = "l", col = "blue",
 lwd = 2)
> third = dgamma(x = x, shape = 2, scale = 1/4)
> lines(spline(x, third, method = "n", n = 250), type = "l", col = "red", l
wd = 2)
> legend("topright", c("k = 1, μ = 1", "k = 2, μ = 1", "k = 2, μ = 2"), fil
l = c("green", "blue", "red"))
```

2. Imagine we are given a sample of n observations $y = (y_1, \ldots, y_n)$. Write down the joint probability of this sample of data, under the assumption that it came from a gamma distribution with mean parameter μ and shape parameter k (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.) [1 mark]

$$p(y|\mu,k) = \prod_{i=1}^{n} p(y_i|\mu,k\ )$$

$$= \prod_{i=1}^{n} \left(\frac{k}{\mu}\right)^{k} \left[\frac{1}{(k-1)!}\right] {y_i}^{k-1} e^{\frac{-ky_i}{\mu}}$$

$$= \left(\frac{k}{\mu}\right)^{nk} \left[\frac{1}{(k-1)!}\right]^{n} \prod_{i=1}^{n} {y_i}^{k-1} e^{\frac{-ky_i}{\mu}}$$

$$= \frac{k^{nk}}{\mu^{nk}((k-1)!)^{n}} \prod_{i=1}^{n} {y_i}^{k-1} e^{\frac{-ky_i}{\mu}}$$

3. Take the negative logarithm of your likelihood expression and write down the negative log-likelihood of the data y under the gamma model with mean parameter μ and shape parameter k. Simplify this expression. [1 mark]

$$L(y|\mu,k) = -\sum_{i=1}^{n} log\big(p(y_i|\mu,k\ )\big)$$

$$= -\sum_{i=1}^{n} log\left[\left(\frac{k}{\mu}\right)^{k} \left(\frac{1}{(k-1)!}\right) {y_i}^{k-1} e^{\frac{-ky_i}{\mu}}\right]$$

$$= -\sum_{i=1}^{n} \left(log\left[\left(\frac{k}{\mu}\right)^{k} \left(\frac{1}{(k-1)!}\right)\right] + log\left({y_i}^{k-1} e^{\frac{-ky_i}{\mu}}\right)\right)$$

$$= -\sum_{i=1}^{n} \left(log\left(\frac{k}{\mu}\right)^{k} + log\left[\frac{1}{(k-1)!}\right] + log({y_i}^{k-1}) + log\left(e^{\frac{-ky_i}{\mu}}\right)\right)$$

$$= -\sum_{i=1}^{n} \left(k\ log\left(\frac{k}{\mu}\right) + log\left[\frac{1}{(k-1)!}\right] + (k-1)\ log(y_i) - \frac{ky_i}{\mu}\right)$$

$$= -nk\ log(k) + nk\ log(\mu) + n\ log[(k-1)!] - (k-1)\sum_{i=1}^{n} log(y_i)$$

$$+ \frac{k}{\mu}\sum_{i=1}^{n} y_i$$

4. Derive the maximum likelihood estimator $\hat{\mu}$ for μ, under the assumption that k is known. That is, find the value of μ that minimises the negative log-likelihood with k assumed to be an arbitrary, known constant. You must provide working. [2 marks]

$$\frac{d}{d\mu}\big(L(y|\mu,k)\big)$$

$$= -\frac{d}{d\mu}\big(nk\,log(k)\big) + \frac{d}{d\mu}\big(nk\,log(\mu)\big) + \frac{d}{d\mu}(n\,log[(k-1)!])$$

$$-\frac{d}{d\mu}\left((k-1)\sum_{i=1}^{n} log(y_i)\right) + \frac{d}{d\mu}\left(\frac{k}{\mu}\sum_{i=1}^{n} y_i\right)$$

$$= -nk\frac{d}{d\mu}\big(log(k)\big) + nk\,\frac{d}{d\mu}\big(log(\mu)\big) + k\sum_{i=1}^{n} y_i\frac{d}{d\mu}\left(\frac{1}{\mu}\right)$$

$$= \frac{nk}{\mu} - \frac{k}{\mu^2}\sum_{i=1}^{n} y_i$$

$$0 = \frac{nk}{\mu} - \frac{k}{\mu^2}\sum_{i=1}^{n} y_i$$

$$= \frac{k}{\mu}\left(n - \frac{1}{\mu}\sum_{i=1}^{n} y_i\right)$$

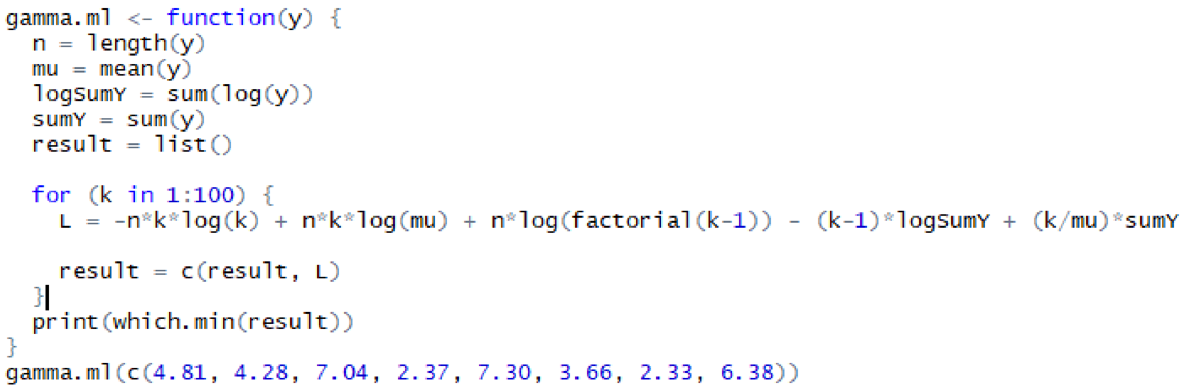
$$= \left(n - \frac{1}{\mu}\sum_{i=1}^{n} y_i\right)$$

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

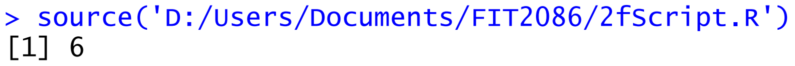
$$\therefore\ \widehat{\mu_{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

5. What is the bias and variance of the maximum likelihood estimator $\hat{\mu}$ of μ for the gamma distribution, assuming that the population value of k is known? Explain how you obtained your answer. [1 mark]

6. So far we have treated k as known. We can also estimate this by using the method of maximum likelihood, although we cannot do it by the usual procedure as no closed form solution exists. However, as k is an integer in our setting, we can instead use R to search for it by trying all values of $k = 1, \ldots, 100$ and searching for the one that minimises the negative log-likelihood of the data. Implement this idea in a function gamma.ml(y) that takes a data vector y and returns the maximum likelihood estimates $\hat{\mu}$ and $\hat{k}$. Once you have coded this, run your code on the data
$$y = (4.81, 4.28, 7.04, 2.37, 7.30, 3.66, 2.33, 6.38)$$
and report the values of the maximum likelihood estimates for μ and k. [2 marks]



```
gamma.ml <- function(y) {
  n = length(y)
  mu = mean(y)
  logSumY = sum(log(y))
  sumY = sum(y)
  result = list()

  for (k in 1:100) {
    L = -n*k*log(k) + n*k*log(mu) + n*log(factorial(k-1)) - (k-1)*logSumY + (k/mu)*sumY

    result = c(result, L)
  }
  print(which.min(result))
}
gamma.ml(c(4.81, 4.28, 7.04, 2.37, 7.30, 3.66, 2.33, 6.38))
```

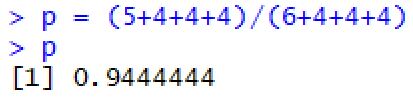


```
> source('D:/Users/Documents/FIT2086/2fScript.R')
[1] 6
```

$\therefore k = 6$

Question 3

1. Calculate an estimate of the German national team's success rate at converting penalties in World Cup penalty shoot-outs. [1 mark]

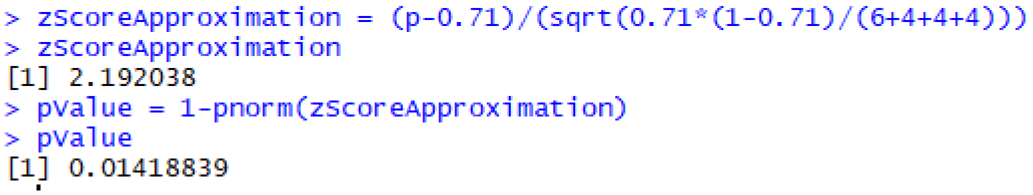


```
> p = (5+4+4+4)/(6+4+4+4)
> p
[1] 0.9444444
```

2. The average rate of penalty conversion across all games at the world cup is71%. Using hypothesis testing, test the hypothesis that the German national team has a penalty conversion rate that is better than the world cup average. Write down explicitly the hypothesis that you are testing, and then calculate a p-value using the approximate approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? [2 marks]

$$H_0: p_{GER} \leq p_{WC}$$

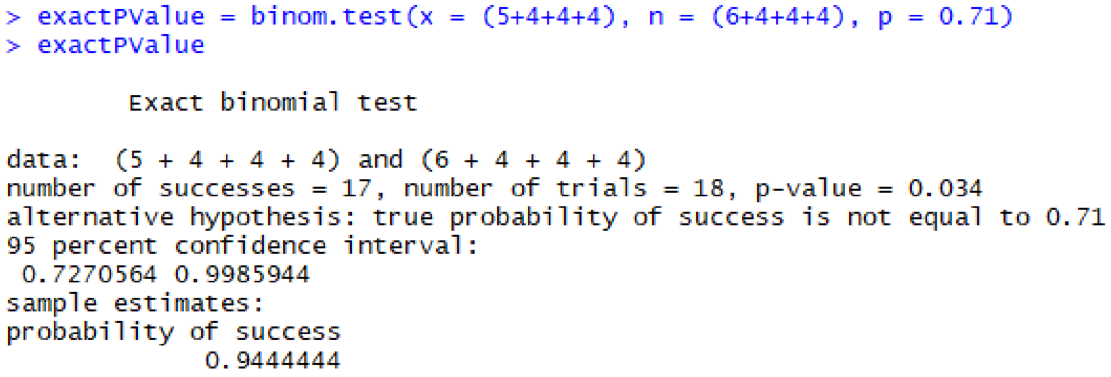
$$H_1: p_{GER} > p_{WC}$$



```
> zScoreApproximation = (p-0.71)/(sqrt(0.71*(1-0.71)/(6+4+4+4)))
> zScoreApproximation
[1] 2.192038
> pValue = 1-pnorm(zScoreApproximation)
> pValue
[1] 0.01418839
```

Using this technique, we can say that if the penalty conversion rate of the German national team is less than or equal to the average rate of penalty conversion across all games at the world cup, then the chance of the German national team having 94% success rate at converting penalties in the World Cup is around 1.42%. This is not particularly likely, and therefore offers a strong evidence against the null hypothesis that the penalty conversion rate of the German national team may actually be more than the average rate of penalty conversion across all games at the world cup.

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. [1 mark]



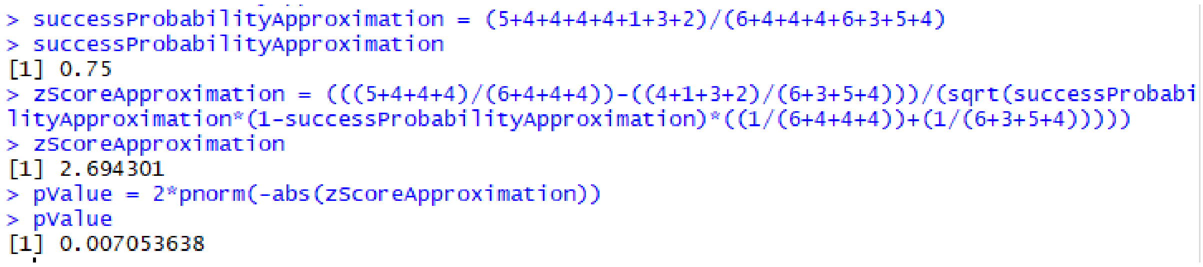
```
> exactPValue = binom.test(x = (5+4+4+4), n = (6+4+4+4), p = 0.71)
> exactPValue

        Exact binomial test

data:  (5 + 4 + 4 + 4) and (6 + 4 + 4 + 4)
number of successes = 17, number of trials = 18, p-value = 0.034
alternative hypothesis: true probability of success is not equal to 0.71
95 percent confidence interval:
 0.7270564 0.9985944
sample estimates:
probability of success
             0.9444444
```

This p-value says that the chance of the German national team having 94% success rate at converting penalties in the World Cup is around 3.4%. This is as weak an evidence as suggested by our approximate test; from this, we could conclude that it is likely that the penalty conversion rate of the German national team may actually be more than the average rate of penalty conversion across all games at the world cup. Our stand do not change.

4. Part of winning a penalty shoot-out is denying your opponent from scoring penalties. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the German penalty conversion rate is different to the penalty conversion rate of their opponents – at least in shoot-outs against Germany – using the data provided in Table 1. Summarise your findings. What does the p-value suggest? [2 marks]

$$H_0 : p_{GER} = p_{opponent(s)}$$

$$H_1 : p_{GER} \neq p_{opponent(s)}$$



```
> successProbabilityApproximation = (5+4+4+4+4+1+3+2)/(6+4+4+4+6+3+5+4)
> successProbabilityApproximation
[1] 0.75
> zScoreApproximation = (((5+4+4+4)/(6+4+4+4))-((4+1+3+2)/(6+3+5+4)))/(sqrt(successProbabi
lityApproximation*(1-successProbabilityApproximation)*((1/(6+4+4+4))+(1/(6+3+5+4)))))
> zScoreApproximation
[1] 2.694301
> pValue = 2*pnorm(-abs(zScoreApproximation))
> pValue
[1] 0.007053638
```

Therefore, with the assumption that there is no difference between the penalty conversion rate of the German national team and the penalty conversion rate of their opponents, then the chance of the German national team having 94% success rate at converting penalties in the World Cup is around 7.05%. This is not particularly likely, and therefore offers a strong evidence against the null hypothesis that the penalty conversion rate of the German national team may be different from the penalty conversion rate of their opponents.

5. Can you identify any possible problems with the way in which the penalty data is sampled that might introduce some biases into your analysis? [1 marks]

Data taken from years when rules are changed or new rules are introduced may affect the numbers.

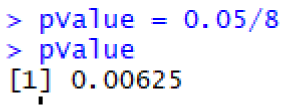
Question 4

1. Fit a multiple linear model to the concrete data using R. Using the results of fitting the linear model, which predictors do you think are possibly associated with compressive strength, and why? Which three variables appear to be the strongest predictors of compressive strength, and why? [2 marks]

```
> setwd("D:/Users/Documents/FIT2086")
> dataset = read.csv("concrete.csv", header = TRUE)
> fit = lm(Strength ~ ., dataset)
> summary(fit)

Call:
lm(formula = Strength ~ ., data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.817  -8.874   1.175   7.631  22.636 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1.175e+02  7.251e+01  -1.621   0.1063    
Cement              1.440e-01  2.827e-02   5.092 7.15e-07 ***
Blast.Furnace.Slag  1.243e-01  2.774e-02   4.480 1.16e-05 ***
Fly.Ash             7.170e-02  4.605e-02   1.557   0.1208    
Water              -9.918e-03  9.995e-02  -0.099   0.9210    
Superplasticizer    5.944e-02  2.292e-01   0.259   0.7956    
Coarse.Aggregate    5.778e-02  2.756e-02   2.096   0.0371 *  
Fine.Aggregate      5.335e-02  2.635e-02   2.025   0.0440 *  
Age                 8.790e-02  1.086e-02   8.093 2.86e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.26 on 241 degrees of freedom
Multiple R-squared:  0.5673,	Adjusted R-squared:  0.553 
F-statistic:  39.5 on 8 and 241 DF,  p-value: < 2.2e-16
```

Let null hypothesis being the coefficient $\beta_j = 0$ (that is, it is unassociated with the target). Looking at the p-values for each of the variables, we see that the three predictors *Cement, Blast.Furnace.Slag* and *Age* seem to deviate the most ($p < 0.1$) from our null hypothesis and are probably our best guesses of which variables might be associated with *Strength*. A small p-value will suggest that the data is at odds with the null hypothesis of no association, and we should potentially consider the predictor as being associated with the target. This is due to a p-value of 0.1 means that the chance of seeing an association as strong as the one we have observed by chance, even if there was no association at the population level, is 1 in 10 – which is neither highly likely, nor is it particularly unlikely – hence these may be highly associated data as the p-values of these variables are much smaller despite a small sample size. Thus, the result brings with it minimal uncertainty as the effect is very strong.

2. Would your assessment of which predictors are associated change if you used the Bonferroni procedure with α = 0.05? [1 marks]



```
> pValue = 0.05/8
> pValue
[1] 0.00625
```

This p-value says that the chance of seeing an association as strong as the one we have observed by chance, even if there was no association at the population level, is around 6.25%. Our stand do not change. Variables mentioned may still be considered as highly associated data as the p-values of these variables are still much smaller compared to 6.25% despite a small sample size. Thus, the result brings with it minimal uncertainty as the effect is very strong.

3. Describe what effect water (Water) in the concrete mix appears to have on the mean compressive strength. Describe the effect that the Age variable has on the mean compressive strength of the concrete. [2 marks]

$$\mathrm{E}[Strength] = -117.5395 + (-0.00991795)Water + (0.08790315)Age$$

From this, we can see that the equation says that as Water increases, the predicted Strength decreases. For every unit increase Water, the Strength score decreases by approximately 0.01. So increased Water seems to make the concrete less sturdy.

On the other hand, the equation says that as Age increases, the predicted Strength increases. For every unit increase in Age, the Strength score goes up by approximately 0.09. So increased Age seems to make the concrete sturdier.

4. Use the stepwise selection procedure with the BIC penalty to prune out potentially unimportant variables. Write down the final regression equation obtained after pruning. [1 mark]

```
> fitStepwise = step(fit, direction = "backward", k = log(250))
> summary(fitStepwise)

Call:
lm(formula = Strength ~ Cement + Blast.Furnace.Slag + Fly.Ash + 
    Coarse.Aggregate + Fine.Aggregate + Age, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.9193  -8.7236   0.9863   7.7506  22.5486 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -129.92503   23.63939  -5.496 9.79e-08 ***
Cement                0.15096    0.01360  11.096  < 2e-16 ***
Blast.Furnace.Slag    0.13016    0.01503   8.659 6.70e-16 ***
Fly.Ash               0.08026    0.03156   2.543 0.011597 *  
Coarse.Aggregate      0.06149    0.01594   3.857 0.000147 ***
Fine.Aggregate        0.05914    0.01001   5.909 1.16e-08 ***
Age                   0.08818    0.01071   8.237 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.22 on 243 degrees of freedom
Multiple R-squared:  0.5671,	Adjusted R-squared:  0.5564 
F-statistic: 53.06 on 6 and 243 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\mathrm{E}[Strength] = &-129.925 + (0.1509552)Cement + (0.1301643)Blast.Furnace.Slag \\ &+ (0.08026474)Fly.Ash + (0.06149068)Coarse.Aggregate \\ &+ (0.05914188)Fine.Aggregate + (0.08818403)Age\end{aligned}$$

5. If we wanted to improve the strength of our concrete, what does this model suggest we could do? Can we use this model, as it stands, to find the “optimal” mixture? [2 marks]

From this, we can see that our model says that as Cement and Blast.Furnace.Slag increases, the predicted Strength increases. For every unit increase in Cement and Blast.Furnace.Slag content, the Strength score goes up by approximately 0.28. So increased Cement and Blast.Furnace.Slag content seems to make stronger concrete.

No, we cannot use this model, as it stands, to find the “optimal” mixture because coefficient does not mean causation. Determining the strength variable depending on the coefficients of each predictors does not provide us with enough reliability considering many assumptions were made before deciding on this model.

6. Imagine that a civil engineer proposes to use a new mix of concrete for a project with the mixture given in Table 3. The engineer asks you to predict the mean compressive strength of this new concrete mix after it has set for 28 days.
   a. Use your model to predict the mean compressive strength for this mix. Provide a 95% confidence interval for this prediction. [1 mark]

```
> predict(fitStepwise, data.frame(Cement = 491, Blast.Furnace.Slag = 26, F
ly.Ash = 123, Water = 210, Superplasticizer = 3.9, Coarse.Aggregate = 882,
 Fine.Aggregate = 699, Age = 28), interval = "confidence", level = 0.95)
       fit      lwr      upr
1 55.49492 47.38629 63.60355
```

   b. The current mix of concrete the engineer is using has a mean compressive strength of 52.35mPa after setting for 28 days. Does your model suggest that the newly proposed mix is better than the current mix? [1 mark]

Based on the model, the proposed mix mean compressive strength is better than current mix mean compressive strength, hence yes.