Cover Sheet

By including this statement, we, all the students listed in the table below, declare that:

- We hold a copy of this assignment if the original is lost or damaged.
- We hereby certify that no part of this assignment has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of the assignment has been written for us by any other person except where collaboration has been authorised by the unit coordinator.
- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism; this software may retain a copy on its database for future plagiarism checking.
- We hereby certify that no part of this assignment or product has been submitted by any of us in another (previous or current) assessment, except where appropriately referenced, and with prior permission from the unit coordinator for this unit.
- We hereby certify that we have read and understand what the University considers to be academic misconduct, and that we are aware of the penalties that may be imposed for academic misconduct.

Name	Student Number	Contribution (%)
Heja Bibani	16301173	33.33
Tao Wei	18416536	33.33
Vinith Gunawardena	19620392	33.33

Part 1

Part 1(A)

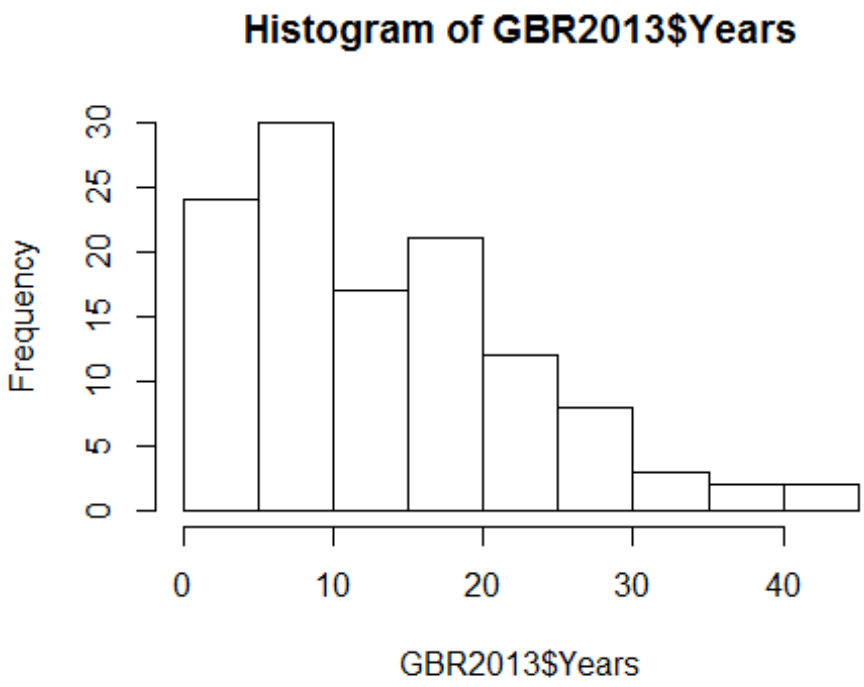
```
GBR2013 = read.csv(file.choose())
GBR2017 = read.csv(file.choose())

mean.Years2013 = mean(GBR2013$Years)
summary(GBR2013$Years, na.rm=TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00   13.00   14.43   20.00   44.00
```

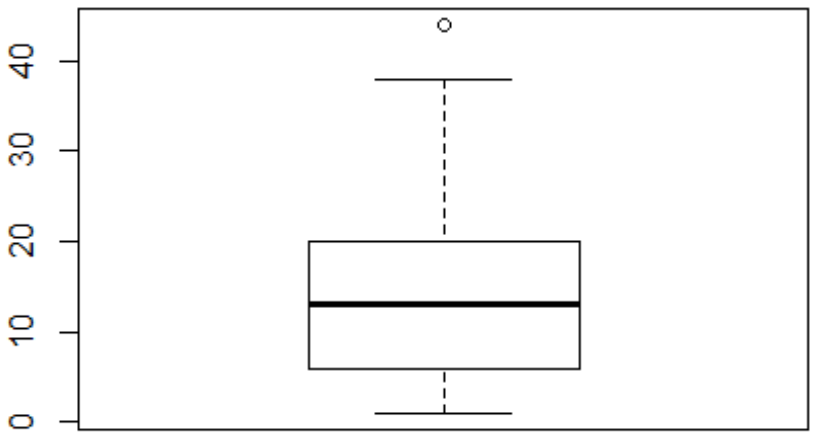
```
sd(GBR2013$Years, na.rm=TRUE)
```

```
## [1] 9.670479
```

```
hist(GBR2013$Years)
```




```
boxplot(GBR2013$Years)
```



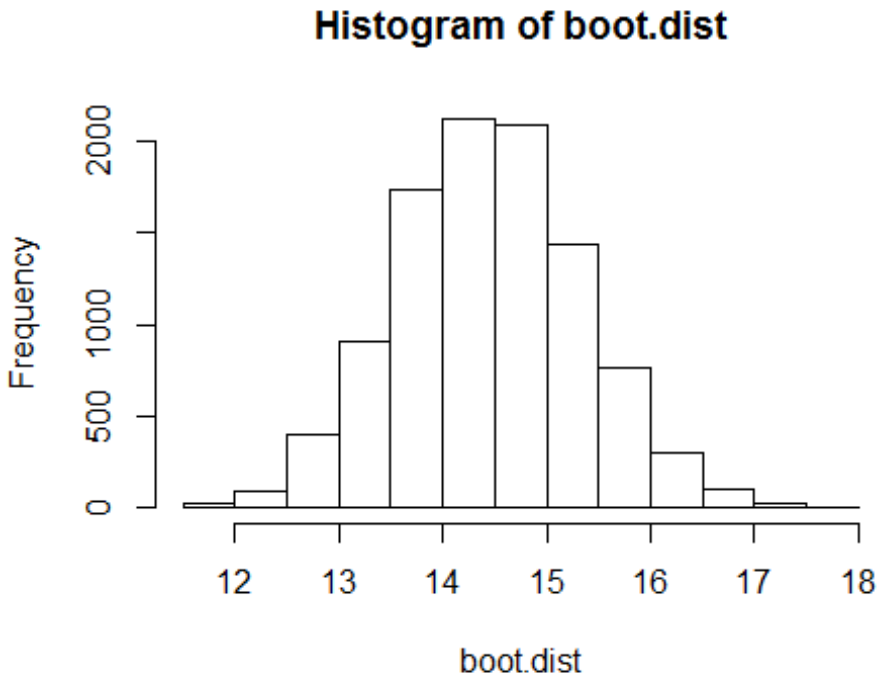
```
#The data is right-skewed when we observe this when we examine this in the hi
stogram. We can also examine this type of spread when we look at the box plot
. We have also identified one outlier; this and the spread has had some effec
t on the calculation of the mean (13, 14.43). The standard deviation is relat
ively large(9.67), and this can also be seen with the histogram and the boxpl
ot. We can see from the box plot, that the data is also spread a little uneve
nly, more so towards the end of the plot [which confirms the spread we see in
the histogram].
```

Part 1(B)

```
mean.Years2013 = mean(GBR2013$Years)
mean.Years2013
```

```
## [1] 14.42857
```

```
b = 10000
boot.dist = rep(NA, b)
for (i in 1:b){
boot.sample = sample(119, replace=TRUE)
boot.dist[i] = mean(GBR2013$Years[boot.sample])
}
hist(boot.dist)
```



```
SE.year2013 = sd(boot.dist)
SE.year2013
```

```
## [1] 0.8871243
```

```
CI.99.Bootstrap.Upper = mean.Years2013 + 2.576*(SE.year2013)
CI.99.Bootstrap.Upper
```

```
## [1] 16.7138
```

```
CI.99.Bootstrap.Lower = mean.Years2013 - 2.576*(SE.year2013)
CI.99.Bootstrap.Lower
```

```
## [1] 12.14334
```

```
#The 99% CI is between 12.14 and 16.71. We are 99% sure that population param
eter for the mean years is between this interval. [Note: I know that each boo
tstrap distribution will change the result of the Confidence interval, the on
e uses the standard error from one of the simulations]
```

Part 1C

```
GBR2013 = read.csv(file.choose())
GBR2017 = read.csv(file.choose())

#CLT states that SE using the formula SE = s / sqrt(n) applies when the size
is greater or equal to 30. This would suggest we can apply CLT to this case ,
since n = 119.

mean.Years2013 = mean(GBR2013$Years)
mean.Years2013
```

```
## [1] 14.42857
```

```
sd.Years2013 = sd(GBR2013$Years)
sd.Years2013
```

```
## [1] 9.670479
```

```
SE.CLT.Years2013 = sd.Years2013 / sqrt(119)
SE.CLT.Years2013
```

```
## [1] 0.8864913
```

```
tvalue.Years2013 = qt(0.995, 119)
tvalue.Years2013
```

```
## [1] 2.617776
```

```
CI.99t.Upper = mean.Years2013 + tvalue.Years2013*SE.CLT.Years2013
CI.99t.Upper
```

```
## [1] 16.74921
```

```
CI.99t.Lower =mean.Years2013 - tvalue.Years2013*SE.CLT.Years2013
CI.99t.Lower

## [1] 12.10794

#We notice that our SE is almost identical to using the bootstrap method; We
can infer that because the SE is almost identical, the CI is going to be very
similar. We know that when we are dealing with means, it follows a t-distribu
tion; so our value for z-score would be replaced by our t-value. The 99% CI f
or both of them are almost identical. CI using Bootstrapping method = 12.14 a
nd 16.71. CI using CLT method was between 12.10762 and 16.74953. This gives u
s an indication that the formula methodology that we are using is a valid rep
resentation of a simulation; and that it can be used as a suitable alternativ
e, if it meets the conditions for CLT.
```

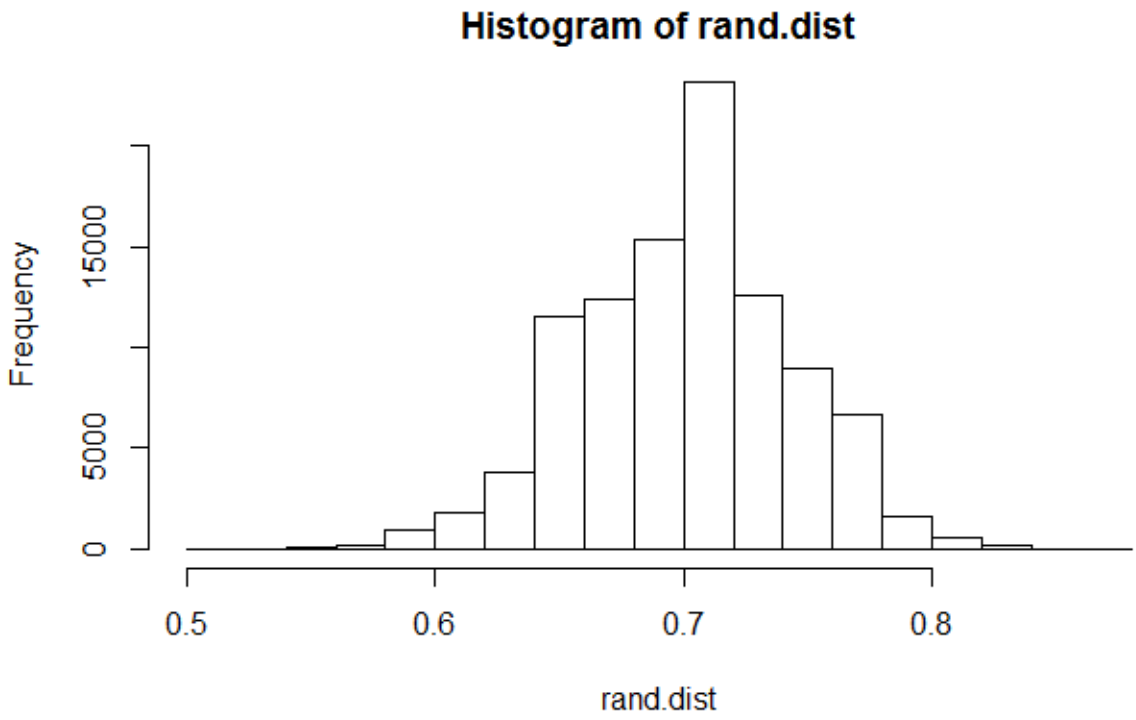
Part 2

Part 2(A)

```
--------------------
#(a) Hypothesis Test
--------------------

#H0: p = 0.7
#HA: p < 0.7

#Note: P is representing the population parameter(for proportions).
#Left-tailed test.

----
#Part (i)
----

#two values are missing and therefore we are using n = 117

----------------------------------------
#With and without Continuity Correction.
----------------------------------------

prop.test(72, 117, conf.level=0.95, correct = FALSE, alternative = "less", 0.
7)
prop.test(72, 117, conf.level=0.95, correct = TRUE, alternative = "less",  0.
7)
```

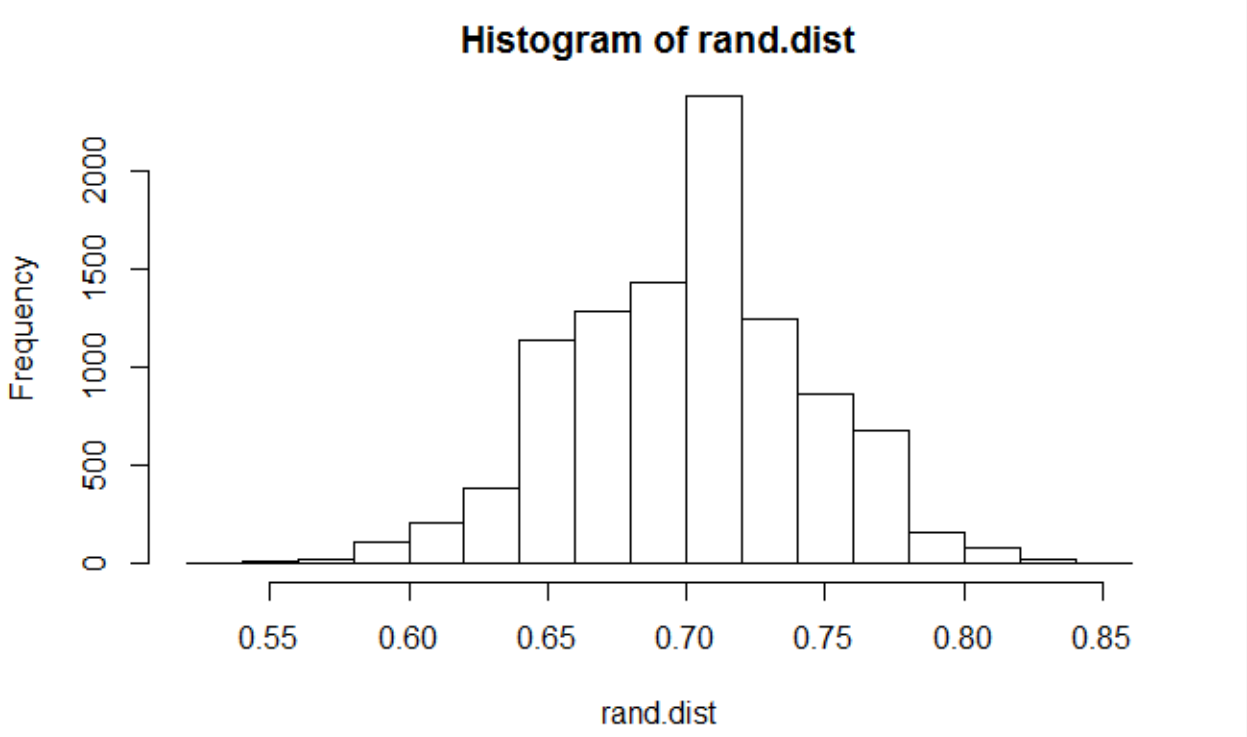
```
## 1-sample proportions test with continuity correction
##
## data:  72 out of 117, null probability 0.7
## X-squared = 3.5963, df = 1, p-value = 0.02895
## alternative hypothesis: true p is less than 0.7
## 95 percent confidence interval:
##  0.000000  0.1668754
## sample estimates:
##    p
## 0.6153846



#p-value without correction = 0.0229
#p-value with correction = 0.02895

#Doing some extra-tests to see if the prop.test is consistent with CLT method
.
#The test is also done with the intention of confirming our SE through random
ization.

se.ztest = sqrt(0.7*0.3/117)
se.ztest

prop.Optimistic = 72/117

ztest = (prop.Optimistic - 0.7) / sqrt(0.7*0.3/117)
ztest
pnorm(ztest, 0 , 1)

#p-value Z-test = 0.0229

-----
#Part (ii)
-----

mean(GBR2013$Optimistic >=6, na.rm=TRUE)

-----------------------------------------
#Method 1 [Centered at the null hypothesis]
-----------------------------------------
r = 100000
n = 117
rand.dist = rep(0,r)
for (a in 1:r) {
```

```
s = sample(c(TRUE,FALSE), n, replace = TRUE, prob = c(0.7,0.3))
rand.dist[a] = mean(s == TRUE)
}
hist(rand.dist)
sd(rand.dist)
pvalue.randomization1 = mean(rand.dist <= prop.Optimistic )
pvalue.randomization1
```




```
#pvalue1 = 0.031

----------------------------------------
#Method 2 [Done for extra Confirmation]
----------------------------------------

b = 10000
rand.dist = rep(NA, b)
for (i in 1:b) {
rand.dist[i] = (rbinom(1, size = 117, prob = 0.7)/117)
}
hist(rand.dist)
sd(rand.dist)
pvalue.randomization2 = mean(rand.dist <= prop.Optimistic )
pvalue.randomization2
```

```
#pvalue2 = 0.0302

#Two randomization methods with the intention of observing consistency. Stand
ard error is similar to the one produced through the formula.
```




```
-------
#Part (iii)
-------

#All three tests give p-values which are around 0.03 which is lower than the
alpha = 0.05. We notice some difference when we use the CLT with the z-score;
but this is corrected when continuity is on. We would reject the Null hypothe
sis with all of the tests [randomization and prop.test]. There is evidence th
at the mean proportion for those optimistic is less than 0.7.
```

Part 2(B)

```r
mean.2013.6 = mean(GBR2013$Optimistic >= 6, na.rm=TRUE)
mean.2017.6 = mean(GBR2017$Optimistic >= 6, na.rm=TRUE)

difference.mean = mean.2013.6 - mean.2017.6
difference.mean
```

```
## [1] 0.02398677
```

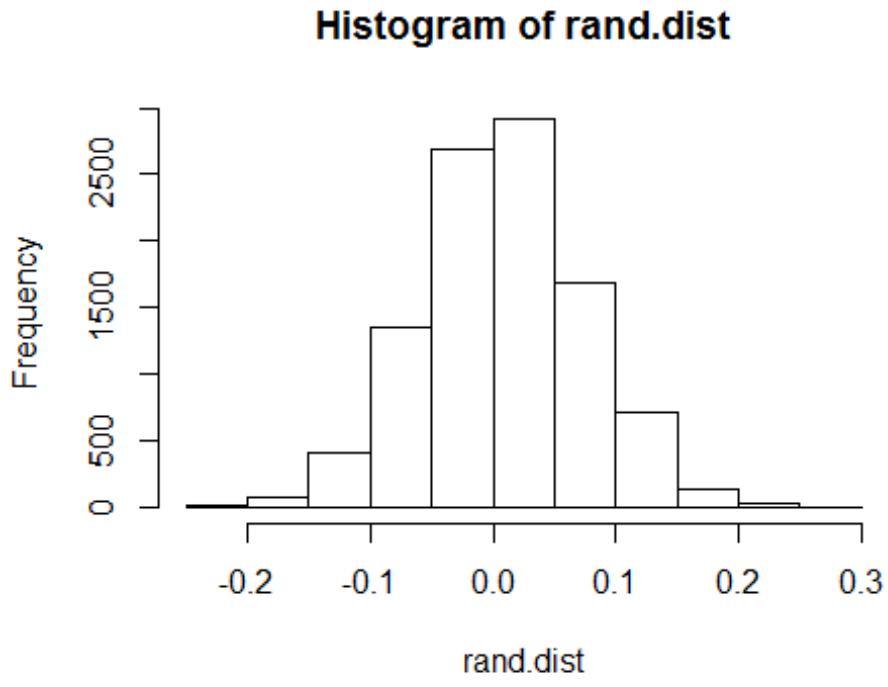
```r
----
#Part (i)
----

#Note: There are two missing values for both 2013 and 2017, and we are using
n(2013) = 117 and n(2017) = 93

#H0 : p(2013) - p(2017) = 0
#HA : p(2013) - p(2017) != 0

#p is representative of the population parameter [proportions]

-----------------------------------------
#Prop.test [With Correct False and True]
-----------------------------------------
prop.test(c(72, 55), c(117, 93), correct = FALSE)
prop.test(c(72, 55), c(117, 93), correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(72, 55) out of c(117, 93)
## X-squared = 0.044557, df = 1, p-value = 0.8328
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1189019  0.1668754
## sample estimates:
##    prop 1    prop 2
## 0.6153846 0.5913978
```

```r
#pvalue[correction] = 0.8328
#pvalue[without correction] = 0.724

#We notice that the correction for continuity has had a big influence on the
p-value; change it from 0.724 to 0.8328.

#Prop-test has its own unique method, and I had thought to add a third(z-test
) to make some clarifications. The test is valuable when we want to examine w
hether or not our SE in our randomization is the same value as the formula. T
he formula was very useful in our investigation, because it has given us some
```

```
hints about how to produce our randomization. The formula gave us some indica
tion that it was imperative that we need to split the values in an uneven fas
hion (117, 93).

table(GBR2017$Optimistic)

##
##  1  2  3  4  5  6  7  8  9 10
## 11  3  3  6 15  9 18 13  3 12

table(GBR2013$Optimistic)

##
##  1  2  3  4  5  6  7  8  9 10
##  6  3  9  9 18 10 12 20  2 28

pooledtable = table(GBR2017$Optimistic) + table(GBR2013$Optimistic)
pooledtable

##
##  1  2  3  4  5  6  7  8  9 10
## 17  6 12 15 33 19 30 33  5 40

#Total n = 210 (missing two values in each year)

pooled.proportion = 127/210
pooled.proportion

## [1] 0.6047619

p.1minusp =  (pooled.proportion)*(1-pooled.proportion)
p.1minusp

## [1] 0.2390249

se.pooled.proprotion = sqrt(p.1minusp/119  + p.1minusp/ 95 )
se.pooled.proprotion

## [1] 0.06726563

zscore.pooled = (mean.2013.6 - mean.2017.6) / se.pooled.proprotion
zscore.pooled

## [1] 0.3565977

pvalue = 2*(1 - pnorm(zscore.pooled, 0, 1))
pvalue

## [1] 0.721393

-----
#Part (ii)
-----
```

```
------------------------------------------------------------
# Randomization
------------------------------------------------------------

#The formula for the SE splits the sizes of the sample into two separate calc
ulations I would expect that splitting the samples in the same manner accordi
ng to the formula will give a similar Standard error. We have two methods tha
t are shown below which will demonstrate this phenomenon. Replacement is fals
e, and the values has been successfully split.


b = 10000
rand.dist = rep(NA, b)
for (i in 1:b) {
unifiedSample=c(GBR2013$Optimistic,GBR2017$Optimistic)
rand.samp=(sample(210, 117, replace = FALSE))
group.size117 = unifiedSample[rand.samp]
group.size93 = unifiedSample[-rand.samp]
rand.dist[i] = mean(group.size117  >= 6, na.rm=TRUE) - mean(group.size93 >=6,
na.rm=TRUE)
}
hist(rand.dist)
```




```
sd(rand.dist)

## [1] 0.06700783
```

```
prop.difference = mean.2013.6 - mean.2017.6

pvaluernd1 = 2*mean(rand.dist >= prop.difference)
pvaluernd1

## [1] 0.8266

#We notice that the prop.test function has slightly different p-value when co
ntinuity is on. Our p-value using randomization is around 0.82. Continuity is
more similar to the randomization, and it seems as if the way it calculates t
he value is a better representation than the general CLT method. This is very
similar to our p-value when we randomized 0.8266. In either case; when signif
icance level alpha is set to 0.05, since 0.82 > 0.05, we do not reject the nu
ll hypothesis, and there is no evidence that the optimism was different betwe
en the years 2013-2017.
```

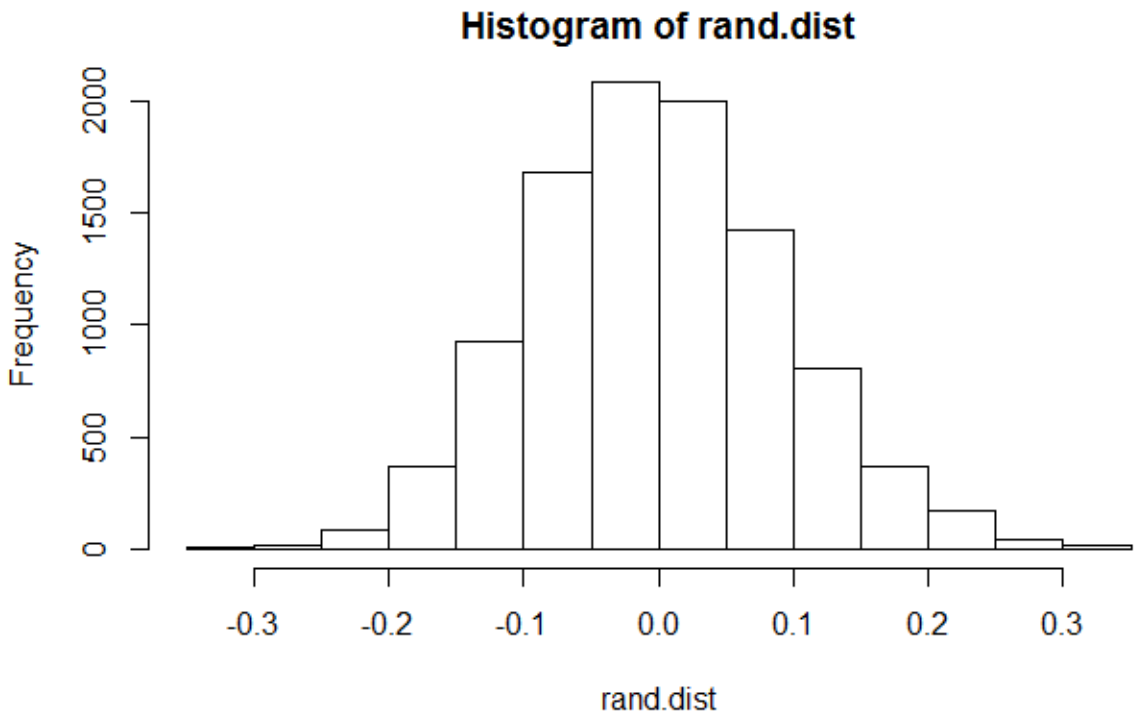
Part 3

Part 3(a)

```
GBR2013$QualityOfLife
GBR2013$WantToProtect

H0 : p(correlation) = 0
HA : p (correlation) > 0

#Note: Both values are representative of the population parameter[correlation
].
```

Part 3(b)

```
cor.Value.2013 = cor(GBR2013$QualityOfLife,GBR2013$WantToProtect, use="comple
te.obs")
cor.Value.2013

#We use in this section; replacement = false.

b = 10000
rand.dist = rep(NA, b)
for (i in 1:b) {
  shuffle = sample(GBR2013$QualityOfLife)
```




```
  rand.dist[i] = cor(GBR2013$WantToProtect,
                     shuffle, use="complete.obs" )
}
hist(rand.dist)
sd(rand.dist)

pvalue = mean(rand.dist >= cor.Value.2013 )
pvalue
```

Part 3(c)

Alpha = 0.10

#Our p-value is 0.185, and this is larger than alpha = 0.10. We do not reject null hypothesis. This suggests that there is no evidence to support the claim that that the "The GBR contributes to my quality of life" and "I would like t o do more help to protect the GBR" are positively correlated.

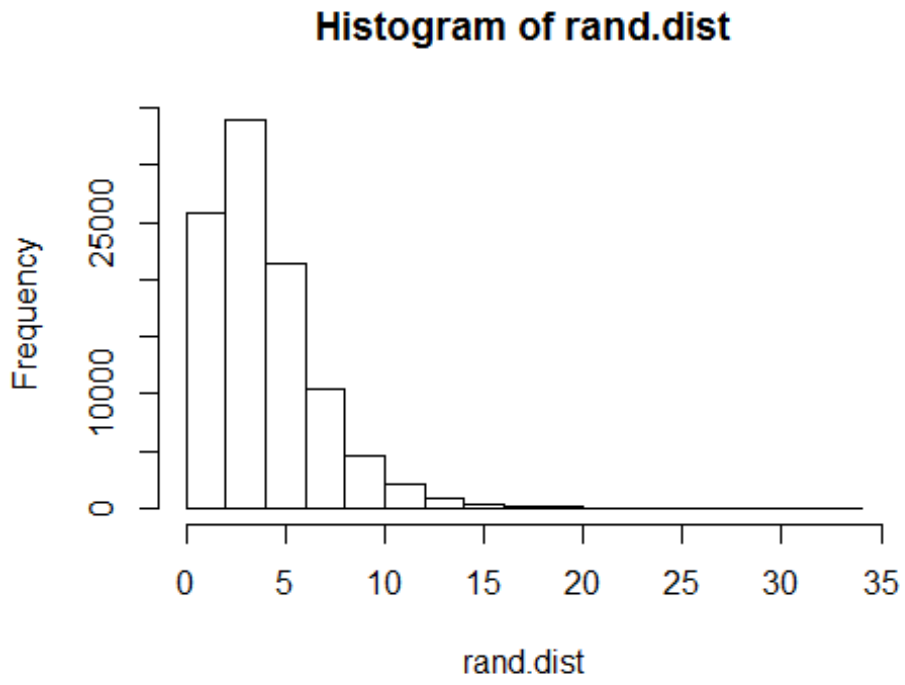
Part 4

Part 4 (a)

```
-------
#Table:
-------

#Note some values are missing n(2013) = 118 and n(2017) = 92

table.2013 = table(GBR2013$ClimateChangeView)
table.2013

##
##  1  2  3  4  5
## 59 21 27  5  6



table.2017 = table(GBR2017$ClimateChangeView)
table.2017

##
##  1  2  3  4  5
## 58  4 22  4  4

#Note: We must conduct randomization because three values are below 5.

-------------------------
#Prop Table[manual work]:
-------------------------

proportion.table2013 = prop.table(table(GBR2013$ClimateChangeView))
proportion.table2013

##
##          1          2          3          4          5
##   0.5000000 0.17796610 0.22881356 0.04237288   0.05084746



proportion.table2017 = prop.table(table(GBR2017$ClimateChangeView))
proportion.table2017

##
##          1          2          3          4          5
## 0.63043478 0.04347826 0.23913043 0.04347826 0.04347826

expectedCounts = proportion.table2013 * 92
```

```r
chisq.test = sum(((table.2017 - expectedCounts)^2 )/ expectedCounts)

chisq.test

# Chisq.test = 12.624

------------------------------
#Chi-squared goodness for fit
------------------------------


Note: In this situation we are testing if the proportions in 2017 are differe
nt from the year 2013's proportions.

H0: p1 = 0.5, p2=0.178, p3= 0.2288 p4=0.0424 p5=0.0508 2017 proportions are n
ot different from 2013's proportions [assume proportions will stay the same]

Ha : That there is some probability that is different.

#Note: We must conduct randomization because three values are below 5.

chisq.test( table(GBR2017$ClimateChangeView), p = c(0.5,0.17796610,0.22881356
, 0.04237288, 0.05084746))

#pvalue[without randomization(function)] = 0.01567. We did this just for some
additional insight into the change between the values(if there are any).


#Simulated Randomization using Chis-quared function
chisq.test( table(GBR2017$ClimateChangeView), p = c(0.5,0.17796610,0.22881356
, 0.04237288, 0.05084746), simulate.p.value = T, B=100000)


##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 1e+05 replicates)
##
## data:  table(GBR2017$ClimateChangeView)
## X-squared = 12.624, df = NA, p-value = 0.01492

##pvalue[with randomization] = 0.01327

#Simulated Randomization; manual[three values are less than 5]

r = table(GBR2017$ClimateChangeView)
e = c(47.5, 16.906780,21.737288,4.025424,4.8300508)

b = 100000
rand.dist = rep(NA, b)
for (i in 1:b) {
```

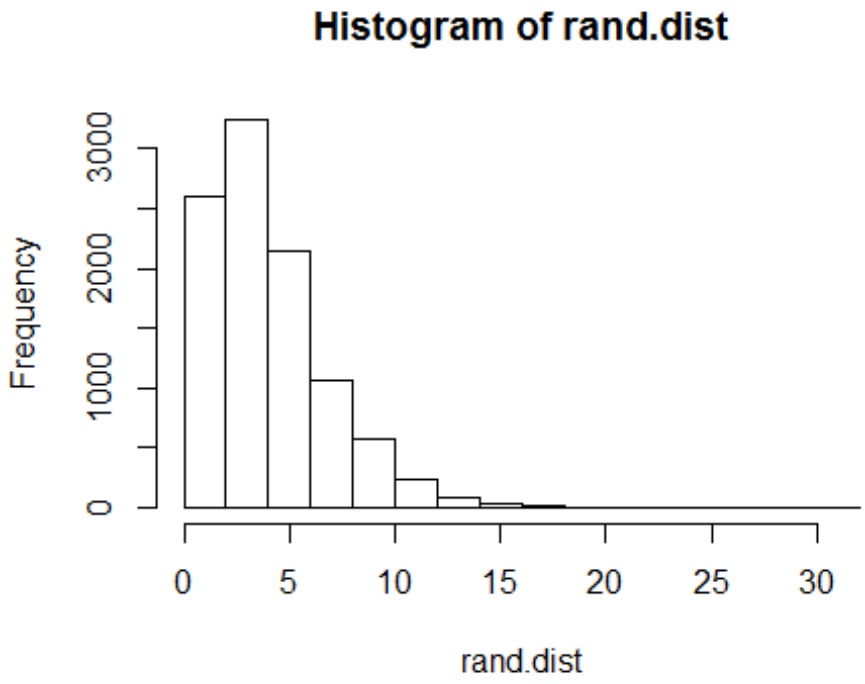
```
r = rmultinom(1,92, c(0.5,0.17796610,0.22881356, 0.04237288, 0.05084746))
rand.dist[i] = sum((r-e)^2/e)
}
hist(rand.dist)
```




```
mean(rand.dist >= 12.624)

## [1] 0.01293

#pvalue[without randomization(manual)] = 0.01308

#P-Values for both types of tests(randomization and without) are around 0.013
-0.015. It hadn't changed the p-value too greatly, but for formal reasons, we
will stick with the value that was given from randomization methods. There wa
sn't much difference between the simulations, in comparison to the general te
st; this is probably because the values are very close to 5. We need to use r
andomization in this case, and this has given us a value of p=0.01327, which
is lower than alpha = 0.05. This would suggest that there is evidence against
the null hypothesis, and for the alternative, however; we would not be able t
o test this without confirmation of the significance level.
```

Part 4(b)

```r
#H0 : Attitude on climate change is not associated with the year
#Ha : Attitude on climate change is associated with the year


degreeOfFreedom = (2-1)*(5-1)

F = matrix(c(59,21,27,5,6,58,4,22,4,4), nrow = 2, ncol=5, byrow=TRUE)
n = sum(F)

E = rowSums(F) %o% colSums(F) / n

#Randomiztion Simulation using ChiSquared function
chisq.test(F, p = E, simulate.p.value = T, B = 10000)

##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  F
## X-squared = 9.5167, df = NA, p-value = 0.0483

#pvalue(randomization[function]) = 0.0447

----------------------------------------
#Simulation using randomization manually
----------------------------------------
rowSums(F)

## [1] -118  -92

colSums(F)

## [1] 117  25  49   9  10

V = sum((F-E)^2/E)
V

## [1] 9.516694

pval.merge = pchisq(V, df=4, lower.tail=FALSE)

pf = c(rep(1,59+58), rep(2,21+4), rep(3,27+22), rep(4, 5+4,), rep(5,6+4))
gr = c(rep(1,59+21+27+5+6), rep(2,58+4+22+4+4))

b = 10000
rand.dist = rep(NA, b)
for (i in 1:b) {
r = table(sample(gr), pf)
rand.dist[i] = sum((r-E)^2/E)
}
hist(rand.dist)
```




```
mean(rand.dist >=V)
```

```
## [1] 0.0463
```

```
#pvalue(randomization[manual]) = 0.0463

#The p-value is 0.0447[chis-squared test function] and 0.0463 for simulation(
randomization). This is a relatively low p-value, and the outcome of the test
would be determined about our choice of Alpha.
```

Part 4(c)

The p-values are little bit different, part(a) = 0.013, and part(b) 0.045. The difference isn't massive because the tests are very similar. However, the difference is significant enough that could cause a different outcome of a hypothesis test depending on alpha (alpha = 0.01 for example).

The goodness of fit tests whether or not there is a difference in proportions from a set of values to the other. A goodness for association(independence) determines whether or not the responses are dependent on the year. The test for association will give us insight about whether there is some property about the year which causes a change in a response or value.

It is a little difficult to determine which test is more suitable, because the question can be interpreted in many ways. From our research into the question, we found that there are certain statisticians who prefer one of the methods over the other. We had identified a similar question from the text book [Question 7.45] where part B's method was employed. To determine which test to choose is dependent on the question being asked; the question itself can have a significant impact on the way we manipulate the data and what tests to employ.

This particular question can be interpreted in many ways, firstly, on one hand it can be interpreted to see if there was a difference that has occurred between the years, and if it is interpreted in this way, then part (a) is more suitable, since we don't require a second test in the opposite direction. However, if we are interpreting the question in a different way, and we are trying to determine if there is some attribute or property about the year (dependent on time or year) which is causing this change, then the methodology in part (b) is more suitable. The good thing about the test for independence is that it will give us an understanding of what to expect, if the responses are dependent upon time (or year) then we will know that the responses will vary between the years (and therefore expect a difference, because the response is dependent on time). This test gives us more insight into answering the question; because it has the capacity to answer both interpretations of the question.