homework 4

Gregory Matesi

Problem 1

The *Urkiola* data set in the spatstat package contains locations of birch and oak trees in secondary wood in Urkiola Natural Park. They are part of a ore extensive dataset collected and analyzed by Laskurain (2008). They coordinates of the trees are given in metersf. Let the "oak" trees be the cases and "birch" trees be the controls.

```r
rm(list = ls())
my.path <- "~/../Desktop/SpatialStatisticsClass/chapter6_casecontrol/homework_4/"
set.seed(1)
library(spatstat)
```

```
## Warning: package 'spatstat' was built under R version 4.0.5

## Loading required package: spatstat.data

## Warning: package 'spatstat.data' was built under R version 4.0.5

## Loading required package: spatstat.geom

## Warning: package 'spatstat.geom' was built under R version 4.0.5

## spatstat.geom 2.2-2

## Loading required package: spatstat.core

## Warning: package 'spatstat.core' was built under R version 4.0.5

## Loading required package: nlme

## Loading required package: rpart

## spatstat.core 2.3-0

## Loading required package: spatstat.linnet

## Warning: package 'spatstat.linnet' was built under R version 4.0.5

## spatstat.linnet 2.3-0

##
## spatstat 2.2-0       (nickname: 'That's not important right now')
## For an introduction to spatstat, type 'beginner'
```

```r
library(smacpod)

data("urkiola")
str(urkiola)
```

```
## List of 6
##  $ window    :List of 5
##   ..$ type  : chr "polygonal"
##   ..$ xrange: num [1:2] 0.05 219.95
##   ..$ yrange: num [1:2] 0.05 149.95
##   ..$ bdry  :List of 1
##   .. ..$ :List of 2
##   .. .. ..$ x: num [1:44] 210 220 220 210 210 ...
##   .. .. ..$ y: num [1:44] 10 10 30 30 60 ...
##   ..$ units :List of 3
##   .. ..$ singular  : chr "metre"
##   .. ..$ plural    : chr "metres"
##   .. ..$ multiplier: num 1
##   .. ..- attr(*, "class")= chr "unitname"
##   ..- attr(*, "class")= chr "owin"
##  $ n         : int 1245
##  $ x         : num [1:1245] 6.1 6.6 8.6 3.9 2.7 10 3.8 5.6 6.1 2.1 ...
##  $ y         : num [1:1245] 146 144 148 143 141 ...
##  $ markformat: chr "vector"
##  $ marks     : Factor w/ 2 levels "birch","oak": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "class")= chr "ppp"
```

```r
class(urkiola)
```

```
## [1] "ppp"
```

```r
# oak <- which(urkiola$marks == "oak")
```

Part a

Perform a test to determine whether the most unusual window of case/control event locations in the study area can be considered a cluster using the spatial scan statistic under the random labeling hypothesis. Use $n_{sim} = 499$ randomly labeled data sets and $\alpha = 0.10$. Make sure to clearly describe your null and alternative hypothesis. Make your conclusion in the context of the problem.

```r
n.sim <- 499

# urkiola_scan = spscan.test(urkiola, nsim = n.sim, case = "oak")
# save(urkiola_scan, file = paste0(my.path, "urkiola_scan.rda"))

load(paste0(my.path, "urkiola_scan.rda"))

summary(urkiola_scan, digits = 3)
```

```
##   centroid_x centroid_y radius events cases      ex    rr   stat     p
## 1      144.7       69.6 77.447    875   313 252.309 2.877 38.337 0.002
```

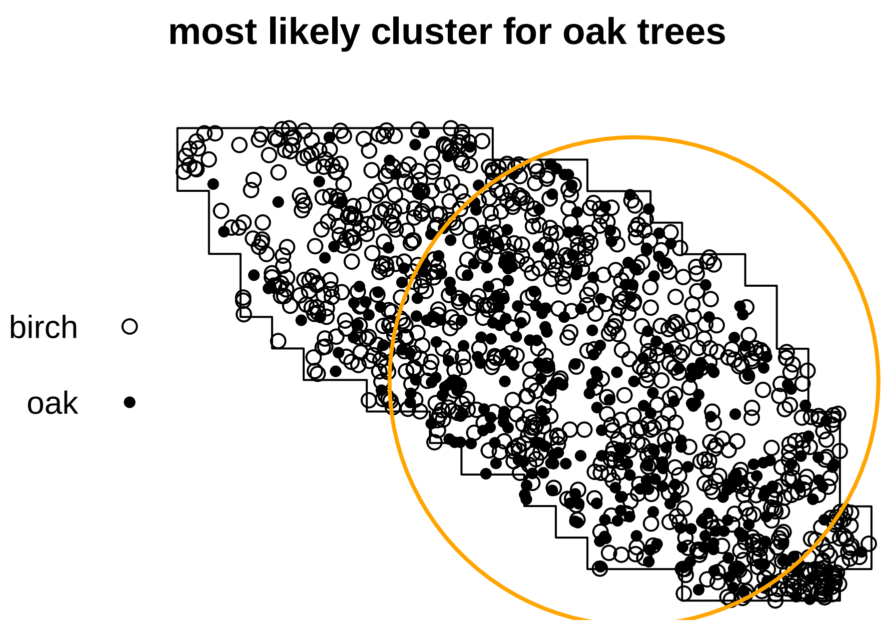
```r
# clusters(urkiola_scan)
```

The window that is least consistent with the null hypothesis that there is no significant clustering of oak trees in the study area yields a Monte Carlo p-vlue of 0.002. based on this p-value, there is significant evidence to that this cluster (the most unlikely cluster under the null hypothesis) has more clusters of oak trees (cases) than we expect under the null hyothesis. Our null hypothesis in this case was that there is at least one winder where the most unlikely cluster is more unusual than what we expect under the null hypothesis.

Part b

Using your analysis from the previous problem, create a plot of the case/control event locations, the associated study area boundary, and a legend indicating the cases/controls. Add the window identifying the most unusual window of case/control event locations (according to the spatial scan statistic) and any potential secondary clusters. Comment on the results.

```r
plot(urkiola_scan, chars = c(1, 20), main = "most likely cluster for oak trees",
     border = "orange")
```



There are no secondary unusual clusters. The most unusual (under the null hypothesis) cluster is rather large. It covers most of the bottom 3 quarters of the study area.

Part c

Perform a test for clustering using the q nearest neighbors method. Use $q = 3, 5, ..., 19$ and $n_{sim} = 499$ randomly labeled data sets. For which q are there more cases than we would expect

under random labeling in the q locations nearest each case? At what scale does this clustering appear to occur (use the contrasts)?

```
urkiola_qnn <- qnn.test(urkiola, q = c(3, 5, 7, 9, 11, 13, 15, 17, 19), nsim = n.sim, case = "oak")
```

```
## oak has been selected as the case group

## Q nearest neighbors test
##
## case label:  oak
## control label:  birch
##
## Summary of observed test statistics
##
##    q   Tq pvalue
##    3  377  0.002
##    5  637  0.002
##    7  887  0.002
##    9 1125  0.002
##   11 1374  0.002
##   13 1607  0.002
##   15 1855  0.002
##   17 2084  0.002
##   19 2285  0.002
##
## Summary of observed contrasts between test statistics
##
##    contrast Tcontrast pvalue
##     T5 - T3       260  0.002
##     T7 - T3       510  0.002
##     T9 - T3       748  0.002
##    T11 - T3       997  0.002
##    T13 - T3      1230  0.002
##    T15 - T3      1478  0.002
##    T17 - T3      1707  0.002
##    T19 - T3      1908  0.002
##     T7 - T5       250  0.002
##     T9 - T5       488  0.002
##    T11 - T5       737  0.002
##    T13 - T5       970  0.002
##    T15 - T5      1218  0.002
##    T17 - T5      1447  0.002
##    T19 - T5      1648  0.002
##     T9 - T7       238  0.014
##    T11 - T7       487  0.002
##    T13 - T7       720  0.002
##    T15 - T7       968  0.002
##    T17 - T7      1197  0.002
##    T19 - T7      1398  0.002
##    T11 - T9       249  0.002
##    T13 - T9       482  0.002
##    T15 - T9       730  0.002
##    T17 - T9       959  0.002
```

```
##   T19 - T9      1160  0.002
##  T13 - T11       233  0.024
##  T15 - T11       481  0.002
##  T17 - T11       710  0.002
##  T19 - T11       911  0.002
##  T15 - T13       248  0.002
##  T17 - T13       477  0.002
##  T19 - T13       678  0.010
##  T17 - T15       229  0.058
##  T19 - T15       430  0.192
##  T19 - T17       201  0.700
```

For each q equals 3 through 19, there is sufficient evidence to conclude that there are more cases among the q nearest neighbors for each case compared to the random labeling hypothesis. Across the board, the p-values for each q are 0002.

Based on the results of the contrast statistics, the clustering of oak trees observed for q equal to 17 and 19 are caused by the clustering of the 15 nearest neighbors for each case.

Problem 2

Answer the same questions as problem 1 for the *paracou* data set in the **spatstat** packagef. Let the juveniles be the controls and adults be the cases.

Part a

Perform the test to determine whether the most unusual window of case/control event locations in the study area can be considered a cluster using the spatial scan statistic under the random labeling hypothesis. Use $N_{sim} = 499$ randomly labeled data sets and $\alpha = 0.10$. Make sure to clearly describe your null and alternative hypotheses. Make your conclusion in the context of the problem.

```r
rm(list = ls())
my.path <- "~/../Desktop/SpatialStatisticsClass/chapter6_casecontrol/homework_4/"
set.seed(1)
library(spatstat)
library(smacpod)

data("paracou")
str(paracou)
```

```
## List of 6
##  $ window    :List of 4
##   ..$ type  : chr "rectangle"
##   ..$ xrange: num [1:2] 0 401
##   ..$ yrange: num [1:2] 0 524
##   ..$ units :List of 3
##   .. ..$ singular  : chr "metre"
##   .. ..$ plural    : chr "metres"
##   .. ..$ multiplier: num 1
##   .. ..- attr(*, "class")= chr "unitname"
##   ..- attr(*, "class")= chr "owin"
```

```
##  $ n         : int 884
##  $ x         : num [1:884] 45.5 65.6 88 203.5 219.6 ...
##  $ y         : num [1:884] 457 484 436 452 521 ...
##  $ markformat: chr "vector"
##  $ marks     : Factor w/ 2 levels "adult","juvenile": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "class")= chr "ppp"
```

```
class(paracou)
```

```
## [1] "ppp"
```

```
adult <- which(paracou$marks == "adult")
```

```
rm(class.paracou)
```

```
## Warning in rm(class.paracou): object 'class.paracou' not found
```

```
n.sim <- 499

# paracou_scan = spscan.test(paracou, nsim = n.sim, case = "adult")
# save(paracou_scan, file = paste0(my.path, "paracou_scan.rda"))

load(paste0(my.path, "paracou_scan.rda"))

summary(paracou_scan, digits = 3)
```

```
##   centroid_x centroid_y radius events cases    ex     rr  stat     p
## 1   46.82281          2   30.5      6     4 0.312 13.937 8.272 0.152
```

```
length(paracou_scan)
```

```
## [1] 8
```

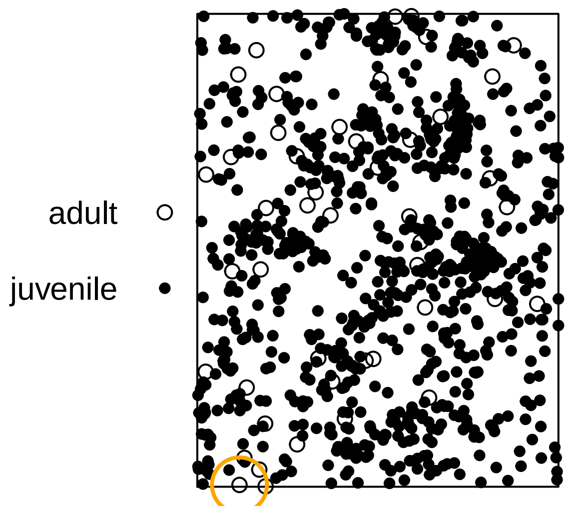
The null hypothesis is that there are no clusters of adult trees in the study area. The alternative hypothesis being that there is at least one cluster of adult trees in the study area. In order to reject the null hypothesis, we look at the window with the most unlikely cluster under the null hypothesis. This cluster provides a p-value based on the t scan statistic of 0.15. Based on this p-value, there is insuffiecient evidence to reject the null hypothesis that there are no windows that are more unusual than we expect under the null hypothesis.

Part b

Using your analysis from the previous problem, create a plot of the case/control event locations, the associated study area boundary, and a legen indicating the cases/controls. Add a window identifying the most unusual collection of case/control event locations (according to the spatial scan statistic) and any potential secondary clusters. Comment on the results.

```
plot(paracou_scan, chars = c(1, 20), main = "most likely cluster for adult trees",
     border = "orange")
```

most likely cluster for adult trees



There are no secondary clusters in this analysis. The most unlikely cluster is located in the bottom left corner of the study area.

Part c

Perform a test for clustering using the q nearest neighbors method. Use $q = 3, 5, ..., 19$ and $n_{sim} = 499$ randomly labeled data sets. For which q are there more cases than we would expect under random labeling in the q locations nearest each case? At what scale does this clustering appear to occur (Use the contrasts)?

```
paracou_qnn <- qnn.test(paracou, q = c(3, 5, 7, 9, 11, 13, 15, 17, 19), nsim = n.sim, case = "adult")
```

```
## adult has been selected as the case group
```

```
## Q nearest neighbors test
##
## case label:  adult
## control label:  juvenile
##
## Summary of observed test statistics
##
##    q Tq pvalue
##    3  9  0.286
##    5 14  0.286
##    7 18  0.392
##    9 25  0.264
```

```
##   11 32  0.184
##   13 36  0.212
##   15 42  0.188
##   17 49  0.144
##   19 54  0.152
##
## Summary of observed contrasts between test statistics
##
##    contrast Tcontrast pvalue
##     T5 - T3         5  0.486
##     T7 - T3         9  0.598
##     T9 - T3        16  0.348
##    T11 - T3        23  0.238
##    T13 - T3        27  0.290
##    T15 - T3        33  0.250
##    T17 - T3        40  0.178
##    T19 - T3        45  0.186
##     T7 - T5         4  0.674
##     T9 - T5        11  0.358
##    T11 - T5        18  0.230
##    T13 - T5        22  0.304
##    T15 - T5        28  0.246
##    T17 - T5        35  0.178
##    T19 - T5        40  0.186
##     T9 - T7         7  0.208
##    T11 - T7        14  0.116
##    T13 - T7        18  0.218
##    T15 - T7        24  0.172
##    T17 - T7        31  0.118
##    T19 - T7        36  0.134
##    T11 - T9         7  0.224
##    T13 - T9        11  0.344
##    T15 - T9        17  0.284
##    T17 - T9        24  0.172
##    T19 - T9        29  0.204
##   T13 - T11         4  0.678
##   T15 - T11        10  0.448
##   T17 - T11        17  0.268
##   T19 - T11        22  0.282
##   T15 - T13         6  0.318
##   T17 - T13        13  0.160
##   T19 - T13        18  0.194
##   T17 - T15         7  0.180
##   T19 - T15        12  0.250
##   T19 - T17         5  0.512
```

The q nearest neighbors test for clustering provides insufficient evidence of clustering among any of q from 3 to 19 for any case.

Problem 3

Write your own function from scratch to implement the q nearest neighbors method, indluding performing a Monte Carlo simulation to assess significance of you results. You may not use any function from spatstat or smacpod.

Part a

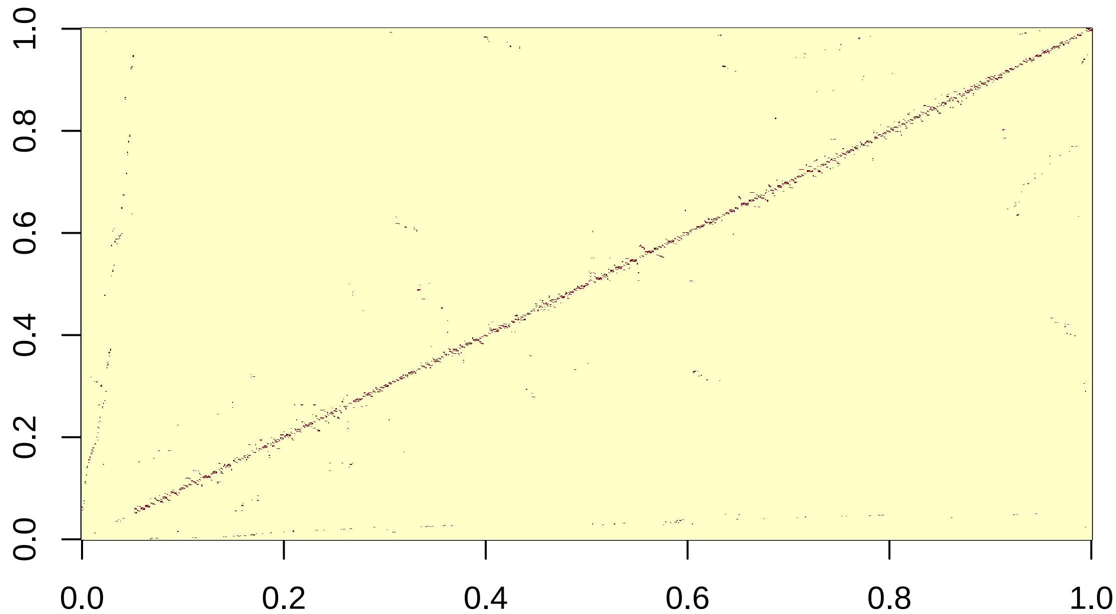
Create a function, W, that takes the event locations and q, the number of nearest neighbors, and, returns the W matrix from the book. Apply this function to the paracou data with q=3, then us the image function to plot the W matrix. Make sure to include your code here.

```
# rm(list = ls())
# data("paracou")
# N <- paracou$n
# W <- as.matrix(dist(cbind(paracou$x, paracou$y)))
#
# dim(W)
# W[1:4, 1:4]
W <- function(dat = data("paracou"), q = 3){
  n <- dat$n
  x <- dat$x
  y <- dat$y
  W <- as.matrix( dist( cbind(x,y) ))
  third <- apply(W, 1, function(x) sort(x)[q+1])

  for(i in 1:n){
    W[i,][W[i,] > third[i]] <- 0
    W[i, which(W[i,] > 0)] <- 1
    # W[i, which(W[i,] > 0)]
  }

  return(W)
}
```

```
wmat <- W(dat = paracou, q = 3)
dim(wmat)
```

```
## [1] 884 884
```

```
image(wmat)
```



Part b

Determine the δ vector discussed in the book for the paracou data, using the adults as cases. Use the formula $\delta^T W \delta$ to determine T_q for each simulated data set for q=3.

```
delta <- paracou$marks
levels(delta)[levels(delta)=="adult"] <- 1
levels(delta)[levels(delta)=="juvenile"] <- 0

# levels(delta[delta=="adult"]) <- 1
# delta[delta=="juvenile"] <- 0

delta <- as.numeric(delta)

delta[delta==2] <- 0

# delta
observed <- as.numeric(t(delta) %*% wmat %*% delta)
observed
```

```
## [1] 9
```

The observed test statistic if $T_3=$9.

Part c

Generate 499 datasets under the random labeling hypothesis for the paracou data, using the adults as cases. Determine T_q for each simulated data set for q=3. Compute the sample mean and variance for the statistics coming from the NULL data (do not include the observed statistic). Compute the Monte Carlo p-value for this test using the observed statistic and the 499 statistics from the simulated data. Make sure to provide your code and clearly indicate the sample mean, sample variance, and Monte Carlo p-value.

```r
# Permute delta
nsim <- 499
sum(delta)
```

```
## [1] 46
```

```r
n1 <- length(delta[delta==1])
n0 <- length(delta[delta==0])

sim.delta <- numeric(length = n0+n1)
sim.delta[sample(n0, size = n1)] <- 1
sim.delta
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
##  [38] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
##  [75] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [112] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
## [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0
## [260] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
## [297] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [334] 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
## [371] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [445] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0
## [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
## [630] 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [667] 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [704] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [741] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
## [778] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
## [815] 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```r
sum(sim.delta)==n1
```

```
## [1] TRUE
```

```r
sampleT <- numeric(nsim)
for (i in 1:nsim) {
  sim.delta <- numeric(length = n0+n1)
  sim.delta[sample(n0, size = n1)] <- 1

  sampleT[i] <-   t(sim.delta) %*% wmat %*% sim.delta

}
sample.mean <- mean(sampleT)
sample.var <- var(sampleT)
MonteCarlo_pval <- (length(which(sampleT>observed)) + 1) / (n0+n1+1)
```

The test statistic sample mean is 7.4228457. The test statistic sample variance is 11.53771. The Monte Carlo p-value is 0.1446328