1. *Our text implies and others state outright that the BB, BW , and WW statistics reveal pretty much the same thing about spatial correlation. The joincount.mc function will carry out Monte Carlo tests based on the BB and WW statistics. We do not have an R formula for computing the BW statistic but it is possible to carry out a BW joincount test of spatial autocorrelation (or clustering) using Gearys c.*

   (a) *Show the relationship between Gearys c and BW.*

   (b) *Carry out a test based on the BW statistics using geary.mc. The data file atrplx.dat will be emailed to you at your math department email addresses. The first 2 columns contain the spatial coordinates and the fourth column contains the Z values you need. Use the R handout to generate the necessary neighbors and list objects.*

   All tests are done using the one sided alternative.

   Let c = Geary's C statistic, which is a measure of similarity in responses, accounting for spatial proximity

   H_0: c = E[c] = no spatial autocorrelation (random pattern)

   H_A: c < E[c] = positive spatial autocorrelation (positive non-random pattern, positive systematic pattern)

   Note the alternative makes sense because in (a) we showed that Geary's c is equivalent to computing a BW statistic, which measures the number of BW connections. The fewer BW connections, the more BB/WW connections, indicating more similar responses at closer locations and is why the sign is opposite.

   1000 Monte Carlo simulations led to a p-value of 0.127 using an observed Geary's C of 0.95. There is no evidence of a positive spatial autocorrelation in the responses.

   ```
   Monte-Carlo simulation of Geary C

   data:  atrplx$z
   ```

```
weights: atrplx_w
number of simulations + 1: 1000

statistic = 0.94756, observed rank = 131.5, p-value = 0.1315
alternative hypothesis: greater
```

MC Density of 1000 Permutations

Geary's C
Monte–Carlo simulation of Geary C
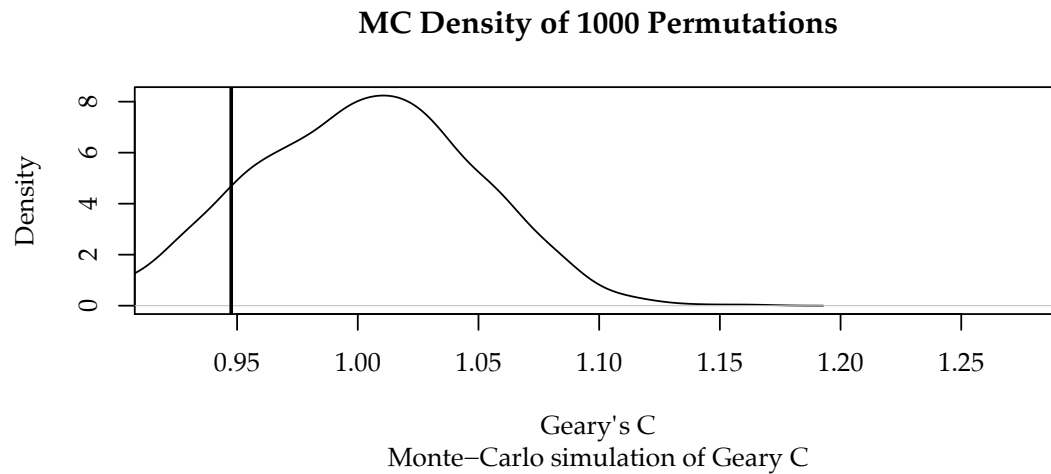
(c) *Using the output from geary.mc compute BW and E[BW]. Do you expect BW $<$E[BW] or BW $>$E[BW] in the presence of positive spatial clustering of the plants? Why or why not?*

In the prescense of positive spatial clustering of plants I would expect BW $<$E[BW]. More B-W joins suggests fewer similar plants (meaning B-B or W-W) clustered. Positive spatial clustering of plants occurs when similar plants (B-B or W-W) are located closer together, thereby fewer dis-similar plants (B-W) being located closer together.

The output from geary.mc resulted in a test statistic of 0.948, leading to a BW test statistic of 173.

$E[BW] = \pi(1 - \pi) * \Sigma_i \Sigma_j w_{ij}$ where $w_{ij} = w_{..}$

Using non-free sampling, E[BW] = 0.254*0.746*960 = 181.86.

BW is less than E[BW] under no spatial autocorrelation and aligns with what was proposed above given there was no evidence of spatial autocorrelation.

(d) *Reproduce the analysis I presented in class using the Atriplex data. Compare the results of the BW test to those of the BB and WW test that were discussed in class. Do these statistics all seem to indicate the same thing about spatial clustering of the plants.*

The three tests do not all result in the same conclusion. Using the BW (c) statistic and the WW statistic, there is no evidence of positive spatial clustering (p-values of 0.125 and 0.414, respectively), though the magnitudes of the p-values differ quite a bit. Using the BB statistic, with a p-value of 0.02 there is strong evidence of positive spatial clustering.

(e) *For grins compute Morans I and compare that result to those above.*

```
Monte-Carlo simulation of Moran I

data:  atrplx$z
weights: atrplx_w
number of simulations + 1: 1000

statistic = 0.068342, observed rank = 943, p-value = 0.057
alternative hypothesis: greater
```

Using Moran's I, there is some evidence of positive spatial autocorrelation with a p-value of 0.057, which falls between the amount of evidence for spatial autocorrelation found using Geary's C (BW) and the BB tests.

2. *Categorize the following examples of spatial data as to their data type:*

   (a) *Elevations in the foothills of the Allegheny mountains.*

   Geostatistical

   (b) *Highest elevation within each state in the United States.*

   Lattice

   (c) *Concentration of a mineral in soil.*

   Geostatistical

   (d) *Plot yields in a uniformity trial.*

   Lattice

   (e) *Crime statistics giving names of subdivisions where break-ins occurred in the previous year and property loss values.*

   Point process

   (f) *Same as previous, but instead of the subdivisions, the individual dwelling, is identified.*

   Point process

   (g) *Distribution of oaks and pines in a forest stand.*

   Point process

3. *Show that Morans I is a scale-free statistic, i.e. Z(s) and λ Z(s) yield the same value for any constant $\lambda \neq 0$.*

4. (a) *Show the $Var(\bar{Y}) =$*

(b) *Let n = 10 and = 0.26. Compare and contrast a 95% confidence interval for μ computed using the true standard deviation of Y and one computed assuming independence.*

With n = 10 and $\rho = 0.26$, a 95% confidence interval accounting for the spatial correlation structure will be about twice the width of that found assuming independence. However, both have the same center and the same multiplier.

Assuming Independence:

$\bar{y} \pm 1.96(\frac{\sigma^2}{10})^{1/2}$

$= \bar{y} \pm 0.62\sigma$

Assuming Correlated:

$\bar{y} \pm 1.96(\frac{\sigma^2}{n}[1 + (n-1)\rho])^{1/2})$

$= \bar{y} \pm 1.96(\frac{\sigma^2}{10}[1 + (10-1)*0.26])^{1/2})$

$= \bar{y} \pm 1.13*\sigma$

(c) *Given independence, we know that Y is the "best" estimator of μ. One nice property it has is that it is a consistent estimator of the mean. Is Y a consistent estimator of the mean given the correlation structure above? Justify your answer.*

(d) *Recall that effective sample size is a measure of the effect of correlation on inference. An equation for the effective sample size under the equicorrelation model is:*

$$n' = \frac{n}{1+(n-1)\rho}$$

The effective sample size is defined to be the sample size n' of uncorrelated observations that provide the same information (in a sense) as a sample of n correlated observations.

  i. *Compute the effective sample size when n = 10, 100, and 1000 and ρ = 0.05, 0.1, 0.25, and 0.5.*

	rho0.05	rho0.1	rho0.25	rho0.5
n10	6.90	5.26	3.08	1.82
n100	16.81	9.17	3.88	1.98
n1000	19.63	9.91	3.99	2.00

 ii. *Find lim n' as $n \to$ infinite.*

 iii. *The effect is extreme here but we would not expect to see this type of correlation structure in a spatial setting. Why not?*

     We would not expect the equicorrelation structure in a spatial setting because the equicorrelation structure would assume that points are equally correlated, regardless of the distance between them. One of the major aspects of spatial data is that observations in closer proximity to each other tend to have more similar responses, which is not accounted for in the equicorrelation structure.

R Code

```r
#first compute number of "distance based" neighbors (vs. color based)
#using lower and upper distance bounds from notes
atrplx_nn <- dnearneigh(as.matrix(atrplx[,c(1,2)]), d1 = 0, d2 = 1)
card(atrplx_nn)

#calculate spatial weights -> "B" = binary
atrplx_w <- nb2listw(atrplx_nn, style = "B")

#do monte carlo simulation based test
gc.out <- geary.mc(atrplx$z,atrplx_w, nsim = 999)


gc.out

# randomly shuffles response and re-assigns to atrplx_w 1000 times
# calculate geary's c for each shuffle
# plot the shuffled geary's c's and find the rank of that observed

plot(gc.out, main = "MC Density of 1000 Permutations", xlab = "Geary's C")

# computing the BW statistic using (a) and (b)

#note third column has BW coding

# the output of the dnearneigh() object gives the total number
# of non-zero joints in the distance matrix, which is w_dot,dot

w2dot <- 960
z_mean <- mean(atrplx[,4])

s2 <- (sum((atrplx[,4] - z_mean)^2))/(dim(atrplx)[1] - 1)

# from (a), get to BW from c by multiplying c by s2*w2dot

bw <- gc.out$statistic*s2*w2dot

# use the observed proportion of blacks = n1 -> non-free sampling
pi <- sum(atrplx[,4])/dim(atrplx)[1]

e_bw <- pi*(1-pi)*w2dot

gc.out_mean <- mean(gc.out$res)
gc.out_obs <- gc.out$statistic

# first output is WW joins, second is BB
atrplx_jc <- joincount.mc(as.factor(atrplx$z), atrplx_w, nsim = 999)

atrplx_mi <- moran.mc(atrplx$z, atrplx_w, nsim = 999)
atrplx_mi

n_prime <- function(t,p){
  t/(1+(t-1)*p)
}

n <- c(10,100,1000)
rho <- c(0.05, 0.1, 0.25, 0.5)

n_prime.out <- matrix(0,nrow=length(n), ncol=length(rho))

for(i in 1:length(n)){
  for(j in 1:length(rho)){
    n_prime.out[i,j] <- n_prime(n[i],rho[j])
```

```
  }
}

n_prime.out <- data.frame(n_prime.out)
rownames(n_prime.out) <- c("n10", "n100", "n1000")
colnames(n_prime.out) <- c("rho0.05", "rho0.1",
                          "rho0.25", "rho0.5")

print(xtable(n_prime.out, align = "||l|l|l|l|l||"))
```