Stat 534 Homework 1

Kenny Flagg

January 18, 2017

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 Geary's c.*

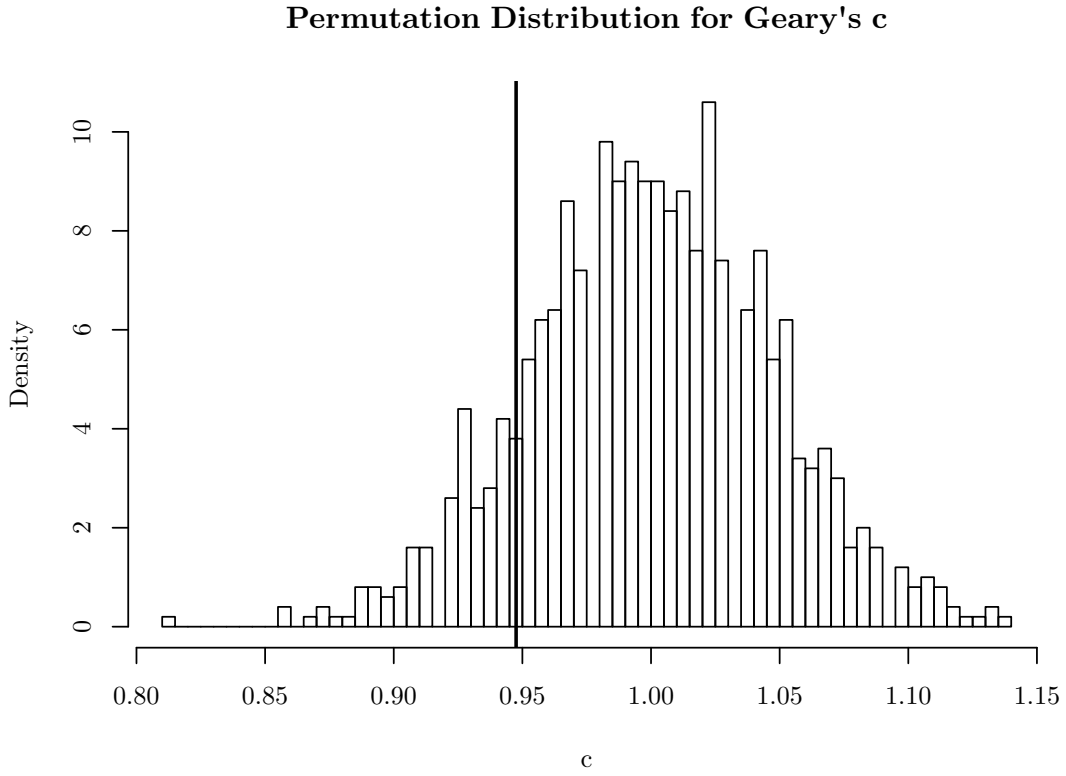
   (a) *Show the relationship between Geary's c and BW.*

$$\begin{aligned} BW &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \\ &= S^2 w_{\cdot\cdot}\left(\frac{1}{2S^2 w_{\cdot\cdot}}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\right) \\ &= S^2 w_{\cdot\cdot} c \end{aligned}$$

   (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.*

```
# Read data.
atrplx <- read.table('atrplx.dat', col.names = c('X', 'Y', 'ignore', 'Z'))

# Get rook adjacency list.
atrplx.nb <- dnearneigh(as.matrix(atrplx[, c('X', 'Y')]), d1 = 0, d2 = 1)

# Augment with binary-style w[ij].
atrplx.lw <- nb2listw(atrplx.nb, style = 'B')

# Now do the test.
atrplx.c <- geary.mc(atrplx$Z, atrplx.lw, 999)
atrplx.c


Monte-Carlo simulation of Geary C

data:  atrplx$Z
```

```
weights: atrplx.lw
number of simulations + 1: 1000

statistic = 0.94756, observed rank = 131, p-value = 0.131
alternative hypothesis: greater

# Make a plot.
hist(atrplx.c$res, breaks = 100, freq = FALSE,
     xlab = 'c', main = 'Permutation Distribution for Geary\'s c')
abline(v = atrplx.c$statistic, lwd = 2)
```



A permutation test with 1,000 permutations results in a p-value of 0.131, which is not convincing evidence the the plants are spatially clustered.

(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?*

```
# Sample variance.
var(atrplx$Z)

[1] 0.1901808

# Neighbor information.
atrplx.lw
```

```
Characteristics of weights list object:
Neighbour list object:
Number of regions: 256
Number of nonzero links: 960
Percentage nonzero weights: 1.464844
Average number of links: 3.75

Weights style: B
Weights constants summary:
    n    nn  S0   S1    S2
B 256 65536 960 1920 14624
```

```
# Mean of the permutation distribution.
mean(atrplx.c$res)
```

```
[1] 1.000238
```

We have $S^2 = 0.19018$, and the number of nonzero links tells us that $w_{\cdot\cdot} = 960$. So

$$BW = 0.19018 \times 960 \times 0.94756 = 173.$$

Then the expected value is

$$BW = 0.19018 \times 960 \times 1.0002 = 182.62.$$

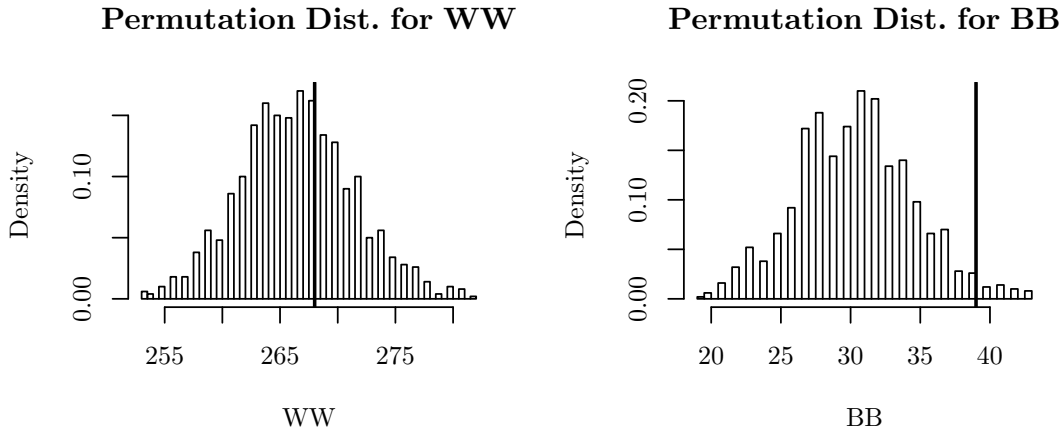
Under positive spatial clustering I would expect $BW < E[BW]$ because clustering will place quadrats with plants together, leading to more black-black and white-white joins, and fewer black-white joins than would occur under if plants were placed completely randomly.

(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?*

```
atrplx.join <- joincount.mc(factor(atrplx$Z), atrplx.lw, 999)
atrplx.join
```

```
Monte-Carlo simulation of join-count statistic

data:  factor(atrplx$Z)
weights: atrplx.lw
number of simulations + 1: 1000

Join-count statistic for 0 = 268, rank of observed statistic = 618,
p-value = 0.382
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
              266.67467               24.79085


Monte-Carlo simulation of join-count statistic
```

```
data:  factor(atrplx$Z)
weights: atrplx.lw
number of simulations + 1: 1000

Join-count statistic for 1 = 39, rank of observed statistic = 972,
p-value = 0.028
alternative hypothesis: greater
sample estimates:
    mean of simulation variance of simulation
              30.54154               17.64732

par(mfrow = c(1, 2))
hist(atrplx.join[[1]]$res, breaks = 50, freq = FALSE,
     xlab = 'WW', main = 'Permutation Dist. for WW')
abline(v = atrplx.join[[1]]$statistic, lwd = 2)
hist(atrplx.join[[2]]$res, breaks = 50, freq = FALSE,
     xlab = 'BB', main = 'Permutation Dist. for BB')
abline(v = atrplx.join[[2]]$statistic, lwd = 2)
```

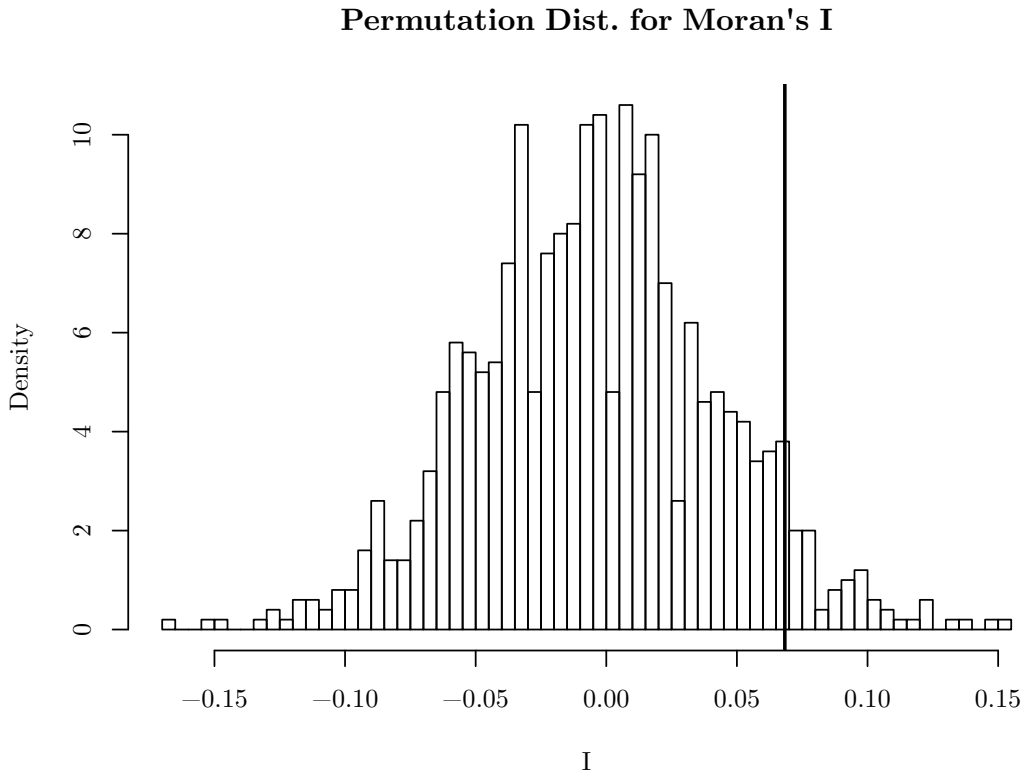


The WW test results in a p-value of 0.382, no evidence of clustering. However, the BB test gives a p-value of 0.382, strong evidence of clustering. The p-value of the BW test is in between these values. These tests do not indicate the same thing about the distribution of the plants because the BB statistic suggests they are clustered, by the WW and BW statistics do not.

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

```
atrplx.i <- moran.mc(atrplx$Z, atrplx.lw, 999)
atrplx.i


Monte-Carlo simulation of Moran I

data:  atrplx$Z
weights: atrplx.lw
number of simulations + 1: 1000
```

```
statistic = 0.068342, observed rank = 938.5, p-value = 0.0615
alternative hypothesis: greater

hist(atrplx.i$res, breaks = 50, freq = FALSE,
     xlab = 'I', main = 'Permutation Dist. for Moran\'s I')
abline(v = atrplx.i$statistic, lwd = 2)
```



With 1,000 permutations, we get a p-value of 0.0615, which gives weak evidence of spatial clustering. This is not as convincing as the result of the test based on BB, and definitely inconsisitent with the tests based on WW and BW.

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

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

   These are geostatistical data.

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

   These are lattice data.

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

   These are geostatistical data.

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

   These are lattice data.

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

   If the list only includes the subdivisions where crimes occurred then the list itself is random, so **these are marked point process data.**

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

   These are marked point process data.

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

   These are marked point process data.

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

   First of all, the mean of $\lambda Z(\mathbf{s})$ is

$$\frac{\sum_{i=1}^n \lambda Z(\mathbf{s}_i)}{n} = \lambda \overline{Z}$$

   and the sample variance of $\lambda Z(\mathbf{s})$ is

$$\frac{\sum_{i=1}^n (\lambda Z(\mathbf{s}_i) - \lambda\overline{Z})^2}{n-1} = \lambda^2 \frac{\sum_{i=1}^n (Z(\mathbf{s}_i) - \overline{Z})^2}{n-1} = \lambda^2 S^2.$$

   Then Moran's I for $\lambda Z(\mathbf{s})$ is

$$\begin{aligned}
\frac{n\sum_{i=1}^n \sum_{j=1}^n w_{ij}(\lambda Z(\mathbf{s}_i) - \lambda\overline{Z})(\lambda Z(\mathbf{s}_j) - \lambda\overline{Z})}{(n-1)\lambda^2 S^2 w_{\cdot\cdot}} &= \frac{n\sum_{i=1}^n \sum_{j=1}^n w_{ij}\lambda(Z(\mathbf{s}_i) - \overline{Z})\lambda(Z(\mathbf{s}_j) - \overline{Z})}{(n-1)\lambda^2 S^2 w_{\cdot\cdot}} \\
&= \frac{n\lambda^2\sum_{i=1}^n \sum_{j=1}^n w_{ij}(Z(\mathbf{s}_i) - \overline{Z})(Z(\mathbf{s}_j) - \overline{Z})}{(n-1)\lambda^2 S^2 w_{\cdot\cdot}} \\
&= \frac{n\sum_{i=1}^n \sum_{j=1}^n w_{ij}(Z(\mathbf{s}_i) - \overline{Z})(Z(\mathbf{s}_j) - \overline{Z})}{(n-1)S^2 w_{\cdot\cdot}} \\
&= I
\end{aligned}$$

   which is Moran's I for $Z(\mathbf{s})$, so Moran's I is scale-free.

4. *Let $Y_1, \ldots, Y_n$ be normally distributed with unknown mean μ and known variance σ^2. Let $Cov(Y_i, Y_j) = \sigma^2\rho$ for $i \neq j$. We will further assume that > 0.*

   (a) *Show that* $Var\left(\overline{Y}\right) = \dfrac{\sigma^2}{n}\left[1 + (n-1)\rho\right]$.

   Note that there are $n^2 - n$ pairs (i, j) such that $i \neq j$. So,

$$\begin{aligned} Var\left(\overline{Y}\right) &= Var\left(\frac{\sum_{i=1}^n}{n}\right) \\ &= \frac{Var\left(\sum_{i=1}^n\right)}{n^2} \\ &= \frac{\sum_{i=1}^n \sum_{j=1}^n Cov(Y_i, Y_j)}{n^2} \\ &= \frac{\sum_{i=1}^n Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j)}{n^2} \\ &= \frac{n\sigma^2 + (n^2 - n)\sigma^2\rho}{n^2} \\ &= \frac{\sigma^2}{n}[1 + (n-1)\rho]. \end{aligned}$$

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

   Using the true standard deviation:

$$\overline{Y} \pm 1.96\frac{\sigma^2}{10}\left[1 + 9 \times 0.26\right] = \overline{Y} \pm 0.65464\sigma^2$$

   Assuming independence:

$$\overline{Y} \pm 1.96\frac{\sigma^2}{10} = \overline{Y} \pm 0.196\sigma^2$$

   The confidence interval computed using the true standard deviation is quite a bit wider than the confidence interval computed assuming independence. Ignoring the correlation in this case means we will think our estimates are more precise than they actually are.

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

   No, $\overline{Y}$ is not consistent in this situation. Consistency is convergence in probability to the quantity of interest, which, by definition 5.5.1 in Casella and Berger, is

$$\lim_{n \to \infty} P\left(|\overline{Y} - \mu| \geq \epsilon\right) = 0$$

   for all $\epsilon > 0$. This is not possible because

$$\lim_{n \to \infty} Var(\overline{Y}) = \lim_{n \to \infty} \frac{\sigma^2}{n}[1 + (n-1)\rho] = \frac{\sigma^2\rho}{n}$$

   is greater than zero if $\rho > 0$.

(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.*

```
n <- matrix(rep(c(10, 100, 1000), 4), ncol = 4,
            dimnames = list(paste('\\(n\\) =', c(10, 100, 1000)),
                            paste('\\(\\rho\\) =', c(0.05, 0.1, 0.25, 0.5))))
rho <- matrix(rep(c(0.05, 0.1, 0.25, 0.5), each = 3), ncol = 4,
              dimnames = list(paste('\\(n\\) =', c(10, 100, 1000)),
                              paste('\\(\\rho\\) =', c(0.05, 0.1, 0.25, 0.5))))
nprime <- n / (1 + (n - 1) * rho)
xtable(nprime, digits = 4, align = '|r|rrrr|')
```

	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.25$	$\rho = 0.5$
$n = 10$	6.8966	5.2632	3.0769	1.8182
$n = 100$	16.8067	9.1743	3.8835	1.9802
$n = 1000$	19.6271	9.9108	3.9880	1.9980

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

Solution:

$$\begin{aligned}\lim_{n\to\infty} n' &= \lim_{n\to\infty} \frac{n}{1 + (n-1)\rho} \\ &= \lim_{n\to\infty} \frac{n}{1 + n\rho - \rho} \\ &= \frac{1}{\rho}\end{aligned}$$

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

Tobler's First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." The setting in this problem ignores distance, making distant locations just as correlated as nearby locations. This is not how things usually work in reality.