Stat 534 Homework 2

Kenny Flagg

January 26, 2017

1. *It was pointed out in class that, conditional on n events, event locations are uniformly distributed for a homogeneous Poisson process. We will consider the simplest example of this. Consider a one-dimensional process on a transect of length L, $(0, L]$. Given that one event has occurred on the interval $(0, L]$ what is the probability that it occurred in the subinterval $(0, s]$ for $s < L$?*

   Let X be the location of the event. Then $X|N((0, L]) = 1 \sim \text{Unif}(0, L)$. So

   $$P(0 < X \leq s|N((0, L]) = 1) = \int_0^s \frac{1}{L} ds = \frac{s}{L}.$$

2. *We looked at an example of* ***quadrat.test*** *on the amacrine data set in class. We will use it to analyze another data set, called* ***redwood****. You can read about this in the* ***spatstat*** *help material. You will be using the* ***quadrat.test*** *function. You can also read about this function in the help material.*

   (a) *Read the help pages on the* ***quadrat.test*** *function. What null hypothesis do they claim to be testing?*

   ```
   require(spatstat)
   data(redwood)
   help(redwood) # optional information on the data set
   help(quadrat.test)
   ```

   The help file says pretty clearly that "we test the null hypothesis that the data pattern is a realisation of Complete Spatial Randomness (the uniform Poisson point process)."

   (b) *Use* ***quadrat.test*** *on the redwood data set.*

   ```
   redwood.fit <- quadrat.test(redwood)
   ```

   ```
   Warning:  Some expected counts are small; chi^2 approximation may be inaccurate
   ```

   ```
   redwood.fit
   ```

   ```
   Chi-squared test of CSR using quadrat counts
   Pearson X2 statistic

   data:  redwood
   ```

```
X2 = 64.613, df = 24, p-value = 0.00002774
alternative hypothesis: two.sided

Quadrats: 5 by 5 grid of tiles
```

The quadrat test on a 5×5 grid yields a test statistic of $\chi^2_{24} = 64.613$ with a p-value of < 0.0001, very strong evidence that the locations of the redwood trees do not follow complete spatial randomness.

The default partitioning of the grid is 5 × 5. Does that appear appropriate here? Justify your answer.

R gave a warning that some of the grid cells had small expected counts, so the 5×5 grid is too fine for these data. Bigger grid cells would be better.

(c) *Redo the analysis using a 3 × 3 grid.*

```
redwood.fit <- quadrat.test(redwood, nx = 3, ny = 3)
redwood.fit
```

```
Chi-squared test of CSR using quadrat counts
Pearson X2 statistic

data:  redwood
X2 = 22.774, df = 8, p-value = 0.007333
alternative hypothesis: two.sided

Quadrats: 3 by 3 grid of tiles
```

This time we don't get a warning message, so the expected cell counts in the larger grid cells are big enough to use the χ^2 distribution. The quadrat test on a 3×3 grid yields a test statistic of $\chi^2_8 = 22.774$ with a p-value of 0.0073, still very strong evidence that the locations of the redwood trees do not follow complete spatial randomness.

(d) *Is the value of the test statistic X^2 indicative of clustering, CSR, or a regular pattern? Justify your answer. Note that I am only asking you to compare the observed value of X^2 to what you would expect under each of these three patterns. You do not need to calculate a P-value just yet.*

Under the null hypothesis, the test statistic follows a χ^2_8 with mean 8. The observed value of 22.774 is larger, indicating greater than excepted variability in quadrat counts which suggests clustering.

(e) *The investigator suspected a clustered pattern and the plot would seem to be consistent with this. Rerun the test with `alternative="c"` in the argument list. Give the p-value for the test and interpret the results. Does the test provide evidence against CSR and for clustering? Justify your answer.*

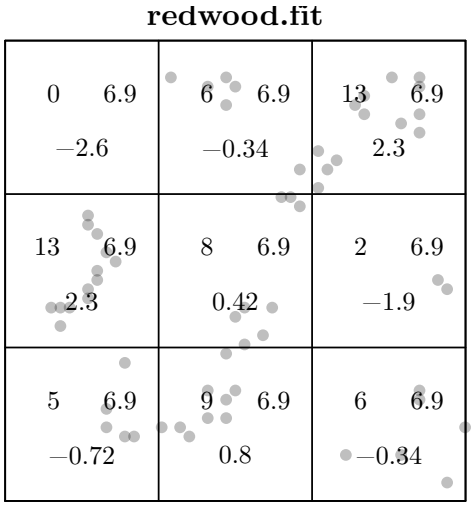
```
redwood.fit <- quadrat.test(redwood, nx = 3, ny = 3, alternative = 'c')
redwood.fit
```

```
Chi-squared test of CSR using quadrat counts
Pearson X2 statistic

data:  redwood
X2 = 22.774, df = 8, p-value = 0.003667
alternative hypothesis: clustered

Quadrats: 3 by 3 grid of tiles
```

The quadrat test on a 3×3 grid with a test statistic of $\chi^2_8 = 22.774$ has a one-sided p-value of 0.0037, very strong evidence that the locations of the redwood trees are clustered.

(f) *We can plot the results of the fit.*

```
par(mar = c(0, 0, 0.5, 0))
plot(redwood.fit)
points(redwood, pch = 16, col = '#00000040') # Add semitransparent points for fun!
```



You will see a plot of the 3 × 3 grid. There are 3 numbers in each cell: the observed count (upper left), expected count under CSR (upper right), and a scaled residual (lower number). The sum of the scaled residuals is the X^2 statistic. Give the results of the test and using the plot indicate where CSR seems to break down, if it does.

The one-sided p-value of 0.0037 provides very strong evidence that trees are clustered. The upper-left and center-right cells have fewer trees than expected under CSR, with large-magnitude negative residuals. The center-left and upper-right cells have large positive residuals, indicating that these cells contain relatively dense clusters.

(g) *Quadrat size can be important. Repeat the analysis using a 2 × 2 grid. Give the results and compare to what we saw with the 3 × 3 grid.*

```
redwood.fit <- quadrat.test(redwood, nx = 2, ny = 2)
redwood.fit
```

```
Chi-squared test of CSR using quadrat counts
Pearson X2 statistic

data:  redwood
X2 = 6.5161, df = 3, p-value = 0.1781
alternative hypothesis: two.sided

Quadrats: 2 by 2 grid of tiles
```

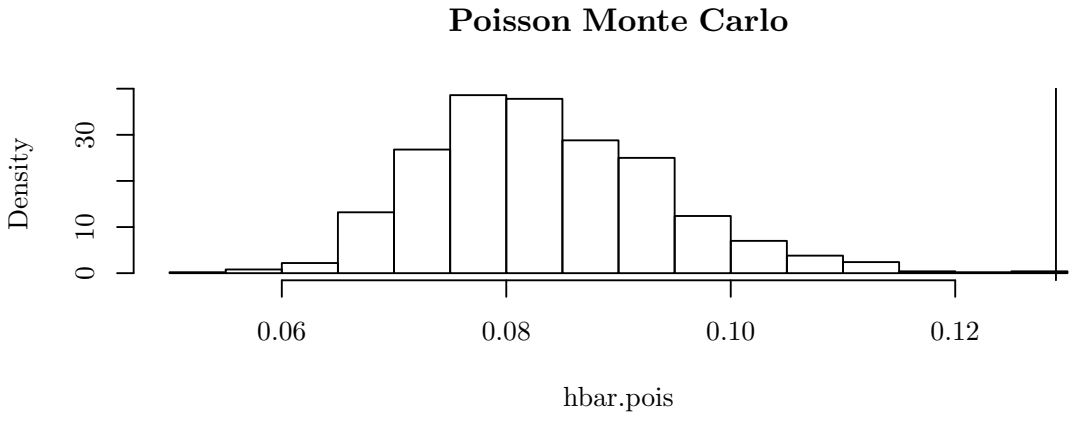
The quadrat test on a 2×2 grid yields a test statistic of $\chi^2_3 = 6.516$ with a two-sided p-value of 0.1781, which is no evidence against complete spatial randomness and contradicts the results of the test on the 3×3 grid.

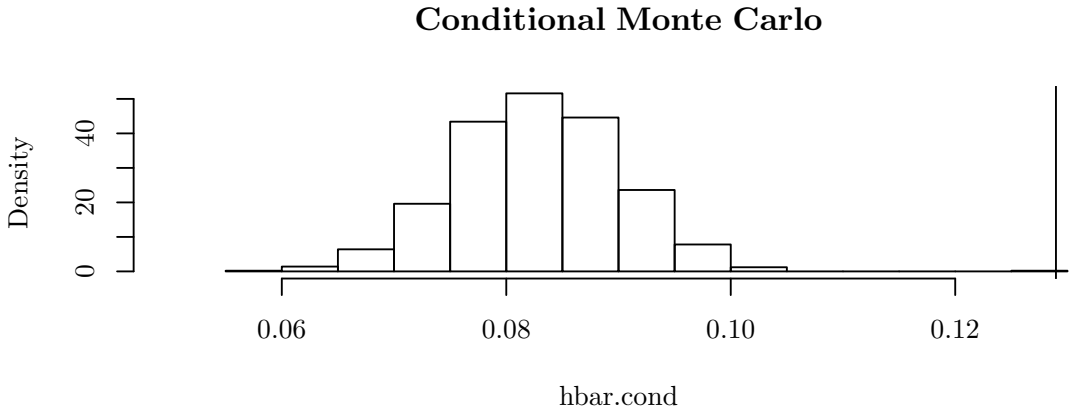
3. *We will compare results from Monte Carlo procedures based on Poisson sampling and based on conditioning on the number of observed points. We will use the* `cells` *data set. The R code to accomplish that is shown below. Compare the two procedures. What do they indicate about the spatial pattern and why? Which procedure do you like best for this data set and why?*

```
data(cells)
hbar <- mean(nndist(cells))
hbar
```

```
[1] 0.1289729
```

```
hbar.pois <- rep(0, 1000)
hbar.cond <- rep(0, 1000)
hbar.pois[1] <- hbar
hbar.cond[1] <- hbar
for(i in 2:1000){
  # Poisson Monte Carlo
  dat.pois <- rpoispp(42)
  hbar.pois[i] <- mean(nndist(dat.pois))
  # Conditional Monte Carlo
  dat.cond <- runifpoint(42)
  hbar.cond[i] <- mean(nndist(dat.cond))
}
par(mfrow = c(2, 1))
hist(hbar.pois, prob = TRUE, xlim = c(0.05, 0.13), main = 'Poisson Monte Carlo')
abline(v = hbar)
hist(hbar.cond, prob = TRUE, xlim = c(0.05, 0.13), main = 'Conditional Monte Carlo')
abline(v = hbar)
```






```
# Poisson P-value
2 * sum(hbar.pois >= hbar) / 1000
```

```
[1] 0.002
```

```
# Conditional P-value
2 * sum(hbar.cond >= hbar) / 1000
```

```
[1] 0.002
```

Both methods give similar p-values of about 0.002, very strong evidence against complete spatial randomness. (The observed average nearest neighbor distance is larger than expected under the null, implying regularity.) The null distribution for the Poison Monte Carlo method is has a larger spread than the null distribution for the conditional Monte Carlo method, reflecting variability in the number of events. For this dataset, the conditional method is more appropriate because the researcher chose the cells to look at and then rescaled the viewing window to the unit square. Therefore it is appropriate to consider permutations of these 42 cells rather than simulating new realizations with random numbers of cells.

4. *Below is the frequency distribution of the number of trees per quadrat in a sample of 100 quadrats each of radius 6 m.*

Trees per quadrat	0	1	2	3	4	≥ 5
Count	34	33	17	7	3	6

The data were pooled for counts ≥ 5 to meet the assumptions of the method. Carry out a Poisson goodness-of-fit test based on an assumption of CSR. Discuss the results. The sample mean of the observed counts was 1.43.

```
# Observed number of quadrats in each bin.
Observed <- c(`0` = 34, `1` = 33, `2` = 17, `3` = 7, `4` = 3, `>=5` = 6)

# Expected number of quadrats in each bin.
Expected <- 100 * c(dpois(0:4, 1.43), ppois(4, 1.43, lower.tail = FALSE))

rbind(Observed, Expected)

                0        1        2        3        4      >=5
Observed 34.00000 33.00000 17.00000  7.00000 3.000000 6.000000
Expected 23.93089 34.22118 24.46814 11.66315 4.169575 1.547069

X2 <- sum((Observed - Expected)^2 / Expected)
X2

[1] 21.56901

pchisq(X2, 4, lower.tail = FALSE)

[1] 0.0002441518
```

With a test statistic of $\chi^2_4 = 21.569$ and a p-value of 0.0002 there is very strong evidence that the quadrat counts do not come from a Poisson(1.43) distribution. The test result does not indicate what type of spatial pattern the data follow, but we observed more counts of zero and more counts ≥ 5 than expected under CSR, which suggests clustering.

5. *Suppose we have a realization of a spatial point process consisting of N event locations $\{s_1, s_2, \ldots, s_N\}$. Let H_i denote the distance between the ith event and the nearest neighboring event. The cumulative distribution function of H (the nearest event-event distance) is the G function. (This problem will be continued on the next homework assignment).*

   (a) *What is the G function if the point process is CSR; i.e. what is $G(h) = P(H \leq h)$?*

$$\begin{aligned} G(h) &= P(H \leq h) \\ &= P(\text{at least 1 event in a circle of radius } h) \\ &= 1 - P(0 \text{ events in a circle of radius } h) \\ &= 1 - \frac{e^{-\lambda\pi h^2}\lambda^0}{0!} \\ &= 1 - e^{-\lambda\pi h^2}, \quad h > 0 \end{aligned}$$

   (b) *Find the pdf of H.*

$$f_H(h) = \frac{dG(h)}{dh} = 2\lambda\pi h e^{-\lambda\pi h^2}, \quad h > 0$$

   (c) *Find $E(H)$ and $Var(H)$. Hint: you found the pdf but before you start evaluating a gnarly integral take a close look at that pdf and see if you cannot identify the family of distributions it belongs to. If you can do that then you can use that knowledge to find the mean and variance.*

   H follows a Weibull distribution with (in the parameterization of Casella and Berger) $\gamma = 2$ and $\beta = \frac{1}{\lambda\pi}$. This distribution has mean

$$E(H) = \beta^{\frac{1}{\gamma}}\Gamma\left(1 + \frac{1}{\gamma}\right) = \frac{1}{\sqrt{\lambda\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{1}{\sqrt{\lambda\pi}}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{\sqrt{\lambda\pi}}\frac{1}{2}\sqrt{\pi} = \frac{1}{2\sqrt{\lambda}}$$

   and variance

$$\begin{aligned} Var(H) &= \beta^{\frac{2}{\gamma}}\left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right] \\ &= \frac{1}{\lambda\pi}\left[\Gamma(2) - \Gamma^2\left(\frac{3}{2}\right)\right] \\ &= \frac{1}{\lambda\pi}\left[1 - \left(\frac{\sqrt{\pi}}{2}\right)^2\right] \\ &= \frac{1}{\lambda\pi}\left[1 - \frac{\pi}{4}\right]. \end{aligned}$$