

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  $s < L$ ?
  
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?

The help page for `quadrat.test` suggest  $H_0$  is complete spatial randomness for a given point pattern.

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

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

## redwood.fit

0 2.5 -1.6	1 2.5 -0.94	4 2.5 0.97	3 2.5 0.33	5 2.5 1.6
1 2.5 -0.94	0 2.5 -1.6	1 2.5 -0.94	7 2.5 2.9	1 2.5 -0.94
9 2.5 4.1	2 2.5 -0.3	2 2.5 -0.3	0 2.5 -1.6	2 2.5 -0.3
1 2.5 -0.94	1 2.5 -0.94	7 2.5 2.9	0 2.5 -1.6	2 2.5 -0.3
0 2.5 -1.6	7 2.5 2.9	2 2.5 -0.3	1 2.5 -0.94	3 2.5 0.33

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

data: redwood
X2 = 64.613, df = 24, p-value = 2.774e-05
alternative hypothesis: two.sided

Quadrats: 5 by 5 grid of tiles
```

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

It was recommended to have at least 5 observations in each grid cell when using the  $\chi^2$  distribution to approximate the p-value. Several cells have less than two seedlings or saplings and so the 5X5 grid is not appropriate. A warning message also indicates the low cell count is a problem.

- (c) Redo the analysis using a 3 3 grid.

Using the 3X3 grid cell counts seem to be adequate to use the the  $\chi^2$  distribution.

```
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
```

- (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.*

The value of the  $\chi^2$  test statistic is 22.8 which suggests spatial clustering. A test statistic of 8 is indicative of CSR (though that cannot be proved) and a test statistic below 8 is indicative of regularity.

*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.*

The default alternative hypothesis is a two sided. Based on a p-value of 0.00367, there is strong evidence against CSR and for spatial clustering.

- (e) *We can plot the results of the fit.*

*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  $\chi^2$  statistic. Give the results of the test and using the plot indicate where CSR seems to break down, if it does.*

### Redwood 3X3 – Ha: Clustered

0 6.9	6 6.9	13 6.9
-2.6	-0.34	2.3
13 6.9	8 6.9	2 6.9
2.3	0.42	-1.9
5 6.9	9 6.9	6 6.9
-0.72	0.8	-0.34

The observed  $\chi^2$  statistic was 22.774 which resulted in a p-value of 0.00367 when compared to a  $\chi^2_8$  distribution. There is strong evidence of spatial clustering. CSR appears to break down in cells (1,1); (2,1); (1,3); (2,3) as there is a large discrepancy between the expected and observed cell counts. The scaled residuals are also larger in these cells. Note that under CSR, the expected cell counts are all the same. and spatial clustering is suggested (vs. regularity) because the cells have large discrepancies. Under regularity, there would be very little if no difference between the observed and expected counts.

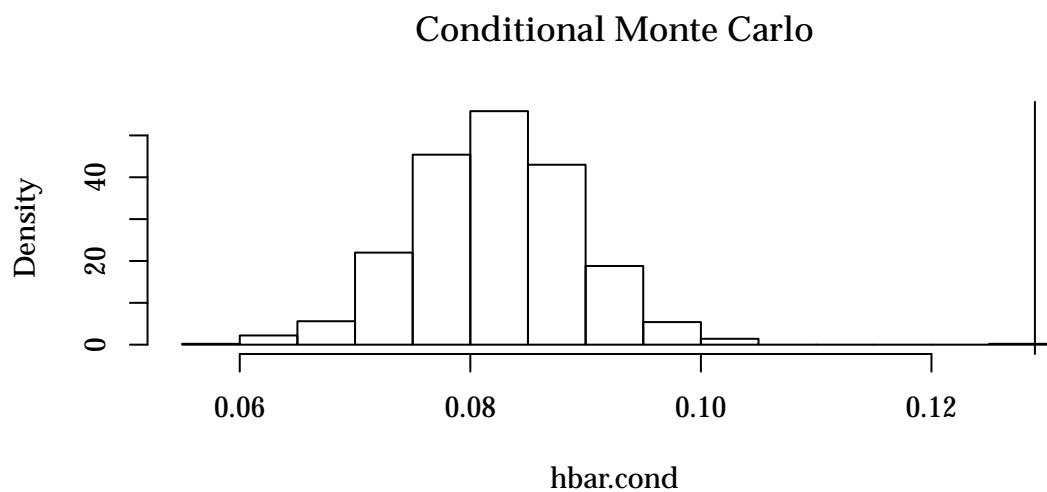
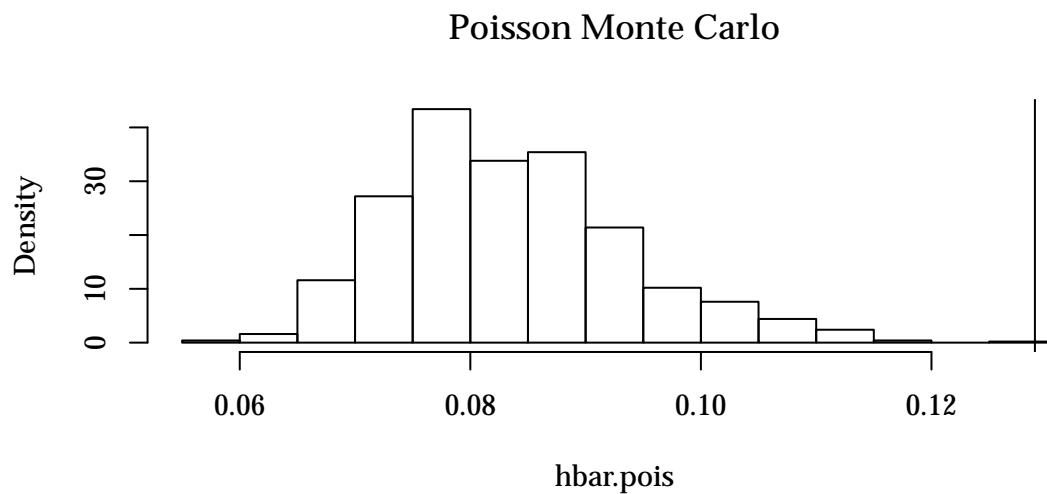
- (f) *Quadrat size can be important. Repeat the analysis using a 2X2 grid. Give the results and compare to what we saw with the 3 3 grid.*

Using a 2X2 grid, a  $\chi^2_3$  test statistic of 6.52 resulted in a p-value of 0.089. There is some evidence against CSR and for spatial clustering of the seedlings.

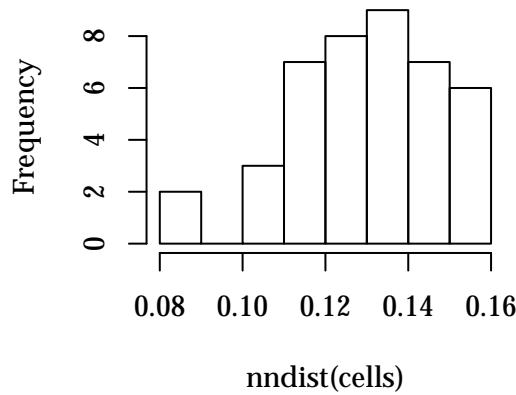
There is less evidence for spatial clustering when using the 2X2 grid than using the 3X3 grid. As the number of cells decrease, it is less likely to see more discrepancies.

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?*

Both procedures indicate the observed average nearest neighbor distance is much farther than would be expected under CSR and is indicative of spatial regularity.



### Observed Nearest Neighbor Distance Distribution



I prefer the Conditional Monte Carlo procedure here because the simulated distribution is symmetric.

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

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.

A Poisson goodness of fit test was carried to assess evidence against a CSR process modeled by a  $\text{Poisson}(1.43)$  distribution. The formal hypothesis are below:

$H_0$ : The distribution of counts is  $\text{Poisson}(1.43)$

$H_a$ : The distribution of counts is not  $\text{Poisson}(1.43)$

	0	1	2	3	4	at least 5
Expected	23.93	34.22	24.47	11.66	4.17	1.55
Observed	34.00	33.00	17.00	7.00	3.00	6.00

The observed data resulted in a  $\chi^2$  test statistic of 21.569. Comparing the test statistic to a  $\chi^2_4$  distribution resulted in a p-value of  $6 \times 10^{-4}$ . There is some evidence against the distribution of seedlings being  $\text{Poisson}(1.43)$ .

Clustering vs. regularity must be established by looking at the data. The expected counts are more distant than the observed counts which suggests clustering.

5. Suppose we have a realization of a spatial point process consisting of  $N$  event locations  $s_1, s_2, \dots, s_N$ . Let  $H_i$  denote the distance between the  $i$ th 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  $G(h) = P(H \leq h)$  ?*

(b) *Find the pdf of H.*

(c) *Find  $E[H]$  and  $\text{Var}(H)$ . 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.*