Homework 3 - Stat 534
Due Tuesday, February 8

1. On the last homework there was some confusion about two problems. I took off some points for one of those but am now giving you a chance to get them back.

   (a) It was pointed out in class that, conditional on n events, event locations are uniformly distributed for a homogeneous Poisson process. Show this result for a $1-d$ process. Hint: 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 $s < L$? Show that this is the cdf of a $Unif(0, L)$ distribution.

   (b) Suppose we have a realization of a spatial point process consisting of N event locations $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N\}$. 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). Derive the G function if the point process is CSR; i.e. what is $G(h) = P(H \leq h)$.

2. We looked at one simple method of using nearest neighbor distances to assess a null hypothesis of CSR. The method was based on using Monte Carlo tests to evaluate the deviation of the mean distance from that expected under CSR. We will look at another possible approach in this problem, one that theoretically would allow us to use a test based on normal theory. A question on Homework 2 asked you to find the probability density function of H, the distance between an event and the nearest neighboring event. If you worked this problem correctly you got

$$g(h) = 2\lambda\pi h \exp\left(-\lambda\pi h^2\right)$$

   where $\lambda > 0$. This is a Weibull distribution parametrized as

$$g(h) = \frac{\beta}{\theta^\beta} h^{\beta-1} \exp\left(-\frac{h}{\theta}\right)^\beta$$

   and with parameters $\beta = 2$ and $\theta = \left(\sqrt{\lambda\pi}\right)^{-1/2}$. We will be working with a homogeneous Poisson process with intensity $\lambda = 30$.

   (a) What are the mean and variance of $\overline{H} = (1/30)\sum H_i$ when $\lambda = 30$, i.e. both the sample size and the intensity equal 30?

   (b) What is the approximate sampling distribution of

$$\frac{\overline{H} - E[\overline{H}]}{\sqrt{\text{Var}[\overline{H}]}}$$

   under CSR and how do you know this?

   (c) Simulate 1000 realizations of complete spatial randomness in the unit square with 30 events in each realization. For each realization

      i. Calculate the distance between each event and its nearest-neighboring event (H_i for the ith event in the realization)

ii. Calculate and store the mean distance.

iii. Calculate and store the values of

$$Z = \frac{\overline{H} - E[\overline{H}]}{\sqrt{\text{Var}[\overline{H}]}}$$

using the mean and variance from part (a) above.

(d) Compute the mean and standard deviation of the 1000 simulated $\overline{H}$ values and compare them to what would be expected under CSR. Are they higher or lower than expected? What could explain this result?

(e) Produce a qqplot of the z scores. Comment.

(f) Use the following formulas for the expected value and variance of $\overline{H}$:

$$E\left[\overline{H}\right] = 0.5\sqrt{A/n} + 0.051P/n + 0.041P/n^{3/2}$$

$$\text{Var}\left[\overline{H}\right] = 0.0703A/n^2 + 0.037P\sqrt{A/n^5}$$

where A is the area and P is the perimeter of the spatial domain (the unit square). Compare the mean and standard deviation from these formulas to those you computed from the simulations above. Does this modification seem to help?

(g) The above procedure is called the Clark-Evans test. Use it to test the null hypothesis of CSR for the cells and redwood data sets. Interpret the results of each test. Also, compute approximate large sample 95% confidence intervals for the mean distance and interpret.

3. I am sending you a copy of a paper by Peter Diggle on the use of K and cross K functions in the analysis of spatial point patterns. The data he is referring to are in the amacrine data set in the spatstat library in R. Read the paper and reproduce the analysis. The data are in `spatstat` (use the command `data(amacrine)`. You do not have to carry out the significance tests he refers to but I would like for you to take the same approach I took on the analysis of the `betacells` data set we discussed in class. Write up a summary of your analysis. Pay attention to the distinction between the independence and random labelling hypotheses.