Stat 525 Homework 9

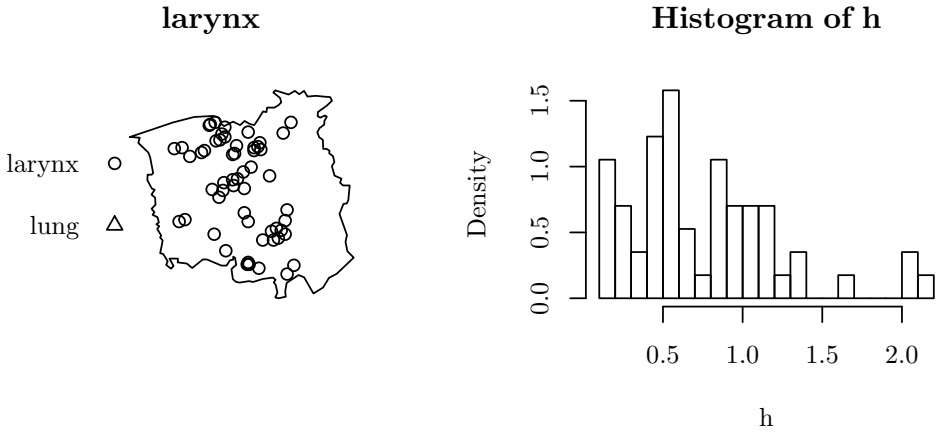
Kenny Flagg

November 14, 2016

1. *We looked at the Chorley-Ribble cancer data in class. For this problem we focus on the locations of the 57 larynx cancer cases.*

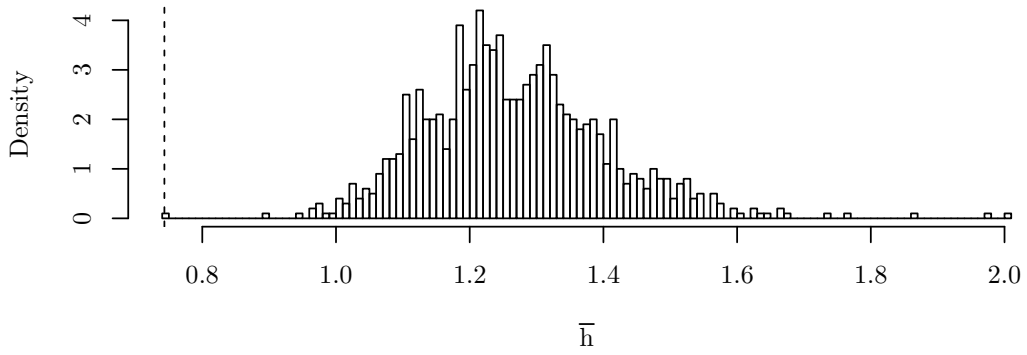
   (a) *Compute the mean nearest neighbor distance. Carry out a randomization test to see if this mean distance is different from that expected under CSR. Provide me with a histogram of the nearest neighbor distances and a randomization p-value based on the alternative of clustering. What is your conclusion and why? I provide you with some of the R code but want you to come up with the rest on your own. See page 11 in the spatial notes for an example involving the entire Chorley data set. You should be able to modify this for the larynx cases only. There is an error in that code. The last line of it should be* `sum(hbar.vec<=hbar)/1000`*. Provide me with a copy of the code you used.*

```
library(spatstat)
larynx <- chorley[chorley$marks == 'larynx']
larynx <- unique.ppp(larynx)

h <- nndist(larynx)
hbar <- mean(h)
lambdahat <- intensity(larynx)['larynx']

par(mfrow = c(1, 2))
plot(larynx)
hist(h, breaks = 20, freq = FALSE)
```



larynx **Histogram of h**

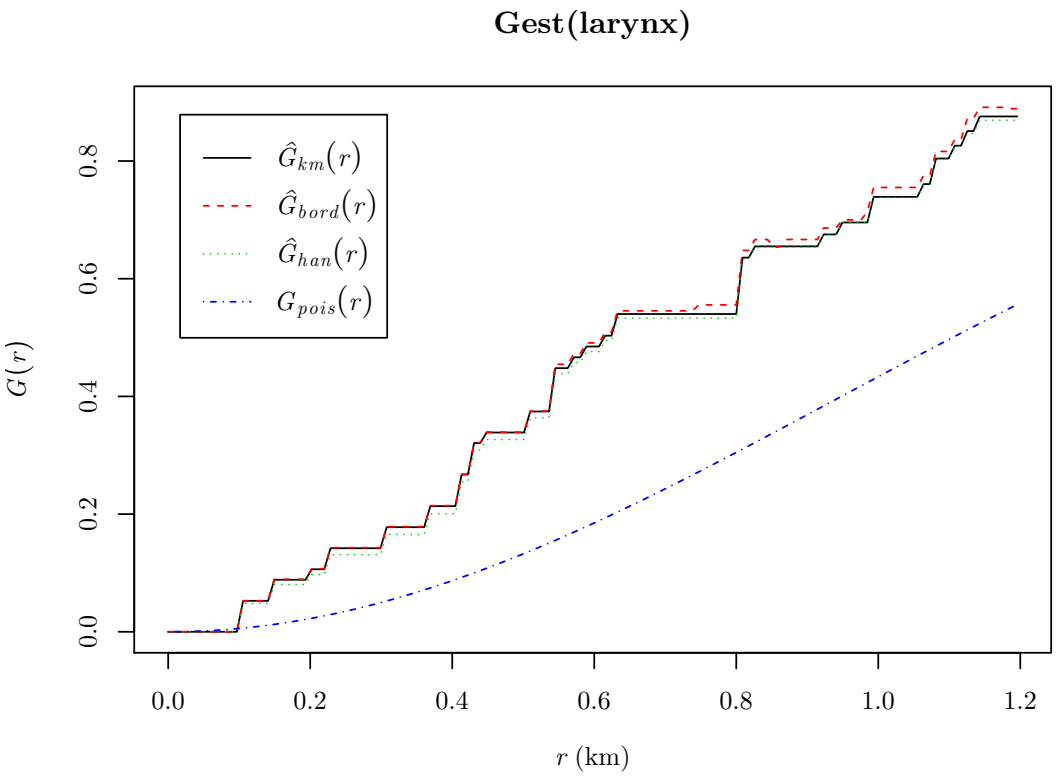
```
set.seed(782315)
hbar.sim <- c(hbar, replicate(999, mean(nndist(rpoispp(lambdahat, win = Window(larynx))))))
pval <- mean(hbar.sim <= hbar)

hbar

[1] 0.7432806

lambdahat

   larynx
0.1808632

pval

[1] 0.001

hist(hbar.sim, breaks = 100, freq = FALSE, xlab = expression(bar(h)),
     main = 'Histogram of 1,000 Simulated Mean Nearest Neighbor Distances')
abline(v = hbar, lty = 2)
```



We observed a mean nearest-neighbor distance of 0.743 km and estimated the intensity to be 0.181 cases per km^2. Under the null hypothesis of complete spatial randomness, we have very strong evidence (p-value = 0.001) that the true mean nearest-neighbor distance is smaller than expected. This implies that the cases tend to be located close together and are therefore clustered.

(b) *Plot the theoretical and empirical G functions. What do they suggest about the spatial distribution of the points and why?*

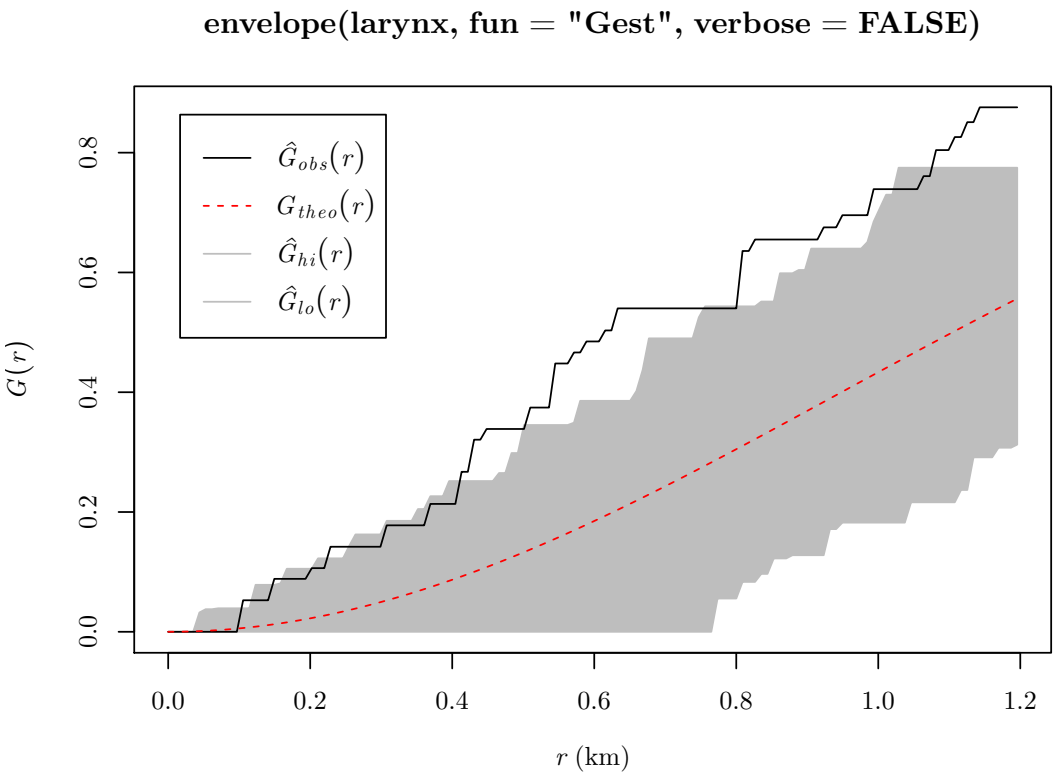
```
plot(Gest(larynx))
```



The empirical G functions are all larger than the theoretical G function under the complete spatial randomness assumption for distances above 0.1 km. This suggests that the observed point pattern has more small nearest-neighbor distances than expected under CSR, suggesting that larynx cancer cases are clustered close together.

(c) *Plot the simulation envelope. Does it provide addtional evidence to back up your conclusion in part (a)? Why or why not.*

```
plot(envelope(larynx, fun = 'Gest', verbose = FALSE))
```

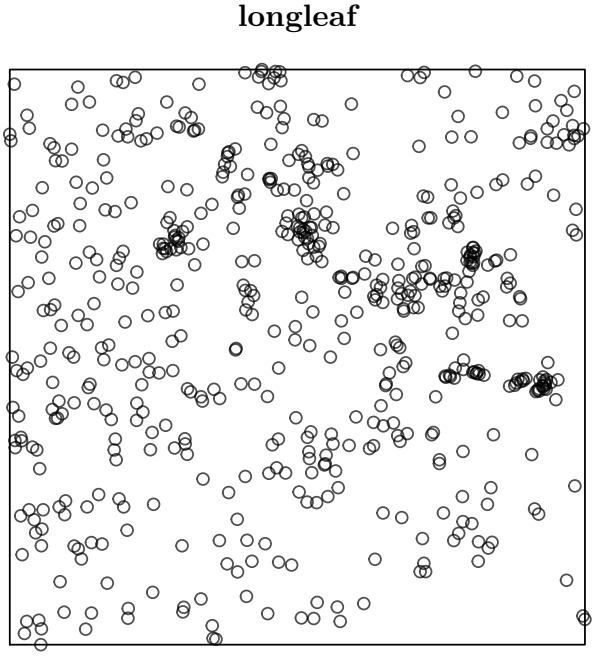


The empirical G function stays near the upper edge of the simulation envelope, leaving the envelope at a few distances. This provides weak evidence that larynx cancer is not a homogeneous process.

2. *There is a dataset in the* ***spatstat*** *package containing locations of 584 longleaf pines in a plot in Georgia. The investigators expected the distribution to be clustered.*

   (a) *Provide me with a plot of the pine locations. The plot shows circles of different diameters associated with marks based on tree size - ignore those.*

```
par(mar = c(0, 0, 2, 0))
plot(longleaf, use.marks = FALSE)
```



   (b) *Test the hypothesis of clustering against the null hypothesis of CSR using* ***quadrat.test****. Discuss the results.*

```
fit <- quadrat.test(longleaf, alternative = 'clustered')
fit
```

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

data:  longleaf
X2 = 152.64, df = 24, p-value < 2.2e-16
alternative hypothesis: clustered

Quadrats: 5 by 5 grid of tiles
```

   With $\chi^2_{24} = 152.64$ and p-value < 0.0001, there is very strong evidence that the counts of trees in the quadrats are more variable than expected under CSR, indicating that the trees are spatially clustered.

(c) *It is possible to "plot" the results. The R code will do this for you. You will see 3 numbers in each grid cell, the observed count, the expected count, and (below those two) the standardized residual. Based on this graphical summary where does it appear CSR breaks down and how?*

```
par(mar = c(0, 0, 2, 0))
plot(fit)
```

fit

20 23.4 −0.7	25 23.4 0.34	37 23.4 2.8	7 23.4 −3.4	26 23.4 0.55
25 23.4 0.34	34 23.4 2.2	50 23.4 5.5	51 23.4 5.7	27 23.4 0.75
29 23.4 1.2	22 23.4 −0.28	15 23.4 −1.7	31 23.4 1.6	37 23.4 2.8
26 23.4 0.55	12 23.4 −2.4	24 23.4 0.13	19 23.4 −0.9	8 23.4 −3.2
18 23.4 −1.1	14 23.4 −1.9	12 23.4 −2.4	8 23.4 −3.2	7 23.4 −3.4

There are some very large positive standardized residuals just above and to the right of the center of the region, meaning that part of the site has a cluster of trees at a higher spatial intensity than expected under CSR, and there are some large negative standardized residuals at the bottom of the region. This matches the plot in part (a) which shows many trees close together in the upper half of the region, and few trees in the bottom right corner.

(d) *Below is the frequency distribution of the number of trees in a sample of 100 quadrats each of radius 6 meters.*

Trees per quadrat	0	1	2	3	4	≥ 5
Observed Count	34	33	17	7	3	6
Expected Count	23.93	34.22	24.47	11.66	4.17	1.547

The data were pooled for counts ≥ 5 to meet the assumptions of the method. Carry out a goodness-of-fit test based on an assumption of CSR. Give the expected frequencies under CSR. Discuss the results, paying particular attention to where CSR breaks down (if it does) - be careful because CSR can break down under both clustering and regularity. You are expecting clustering - do the results support that expectation and why? The sample mean of the observed counts was 1.43.

```
# observed counts
long.obs <- c(34,33,17,7,3,6)
# expected counts for 0 to 4
long.exp <- 100 * dpois(0:4, 1.43)
# expected count for >=5
long.exp <- c(long.exp, 100 - cumsum(long.exp)[5])
# so the expected counts are
long.exp
```

```
[1] 23.930892 34.221176 24.468141 11.663147  4.169575  1.547069
```

```
# I will leave the rest up to you.

x2 <- sum(((long.obs - long.exp)^2)/long.exp)
x2
```

```
[1] 21.56901
```

```
pchisq(x2, 4, lower.tail = FALSE)
```

```
[1] 0.0002441518
```

With a test statistic of $\chi^2_4 = 21.57$, p-value $= 0.0002$, there is strong evidence that the quadrat counts do not come from a Poisson(1.43) distribution. We observed more quadrats containing zero or at least 5 trees than expected, which is consistent with clustering.

3. *STAT 525: Suppose we have a realization of a spatial point process of n event locations. The cumulative distribution function of H (the nearest event-event distance) is the G function.*

   (a) *Derive the G function for a two-dimensional homogeneous Poisson process.*

   If the process has rate λ, then the G function is

$$\begin{aligned} G(h) &= P(H \leq h) \\ &= P(\text{at least 1 event in a circle with radius } h) \\ &= 1 - P(0 \text{ events in a circle with radius } h) \\ &= 1 - \exp\left(-\lambda\pi h^2\right) \frac{\left(\lambda\pi h^2\right)^0}{0!} \\ &= 1 - \exp\left(-\lambda\pi h^2\right). \end{aligned}$$

   (b) *Find the probability density function of H.*

   The probability density function of H is

$$g(h) = \frac{dG(h)}{dh} = 2\lambda\pi h \exp\left(-\lambda\pi h^2\right).$$

   (c) *Give the mean and variance of H. Hint: Before you start evaluating a gnarly integral or two take a close look at the pdf and see if you cannot identify the family of distributions. If you can do that you can use this knowledge to find the mean and variance.*

   H follows a Weibull $(2, \lambda\pi)$ distribution. According to the back of Casella & Berger (who use a different parameterization from both the notes and R), the mean and variance are

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

   and

$$Var(H) = \left(\frac{1}{\lambda\pi}\right)\left(\Gamma(2) - \Gamma^2\left(\frac{3}{2}\right)\right).$$

4. *The time in days to the development of a tumor for rats exposed to a carcinogen follows a Weibull distribution parameterized as in the notes with $\alpha = 2$ and $\lambda = 0.001$.*

   (a) *What is the probability a rat will be tumor free at 30 days? 60 days?*

   The notes give the survival function as $S(t) = \exp\left(-0.001t^2\right)$, so

   $$P(T > 30) = S(30) = \exp\left(-0.001 \times 30^2\right) = 0.4066$$

   and

   $$P(T > 60) = S(60) = \exp\left(-0.001 \times 60^2\right) = 0.0273.$$

   (b) *Find the hazard rate of the of the time to tumor appearance at 30 days. 60 days.*

   The notes give that hazard function as $h(t) = 0.002t$, so

   $$h(30) = 0.002 \times 30 = 0.06$$

   and

   $$h(60) = 0.002 \times 60 = 0.12.$$

   (c) *Find the median time to tumor development.*

   The median time is $t_{0.5}$ such that

   $$\begin{aligned} P(T > t_{0.5}) = S(t_{0.5}) &= 1 - 0.5 \\ \exp\left(-0.001t_{0.5}^2\right) &= 0.5 \end{aligned}$$

   so $t_{0.5} = \sqrt{-\dfrac{\log(0.5)}{0.001}} = 26.33$ days.