ST433/533 Applied Spatial Statistics

Lab activity for 11/3/2020

A. CLARIFICATION QUESTIONS

(1) In the quadrat test, when calculating the test statistic, what did you mean by (m-1 dof) ?

DOF is the “degrees of freedom”. This is something like the number moving parts in the analysis and affects the sampling distribution.

(2) When looking for hotspots, what is the most conservative test that we can run?

You can be as conservative as you wish by setting the threshold for the p-value close to zero.

(3) Slide 8: In CRS test, the expected rate is proportional to area r(B). Can it be proportional to other variables, i.e. population of B?

Yes! See the discussion question (3) below.

(4) Scan statistic can test for whether there is a hot spot but the exact location of hot spot is unknown, so does 'hotspot' has any realistic meaning or we just use it as a tool to explore the dependence?

For the scan statistics the hot spot is formally defined as a region with higher sampling rate than the background. There is not one unique hot spot of course, but the concept is well defined.

(5) I'm not really understanding of the final test of CRS $T = \max_B t(B)$

Instead of computing the text statistic for one region, B, you compute for many possible regions and then take the largest value you observe.

(6) For the scan statistic, is there a parameter for how the window should be moved. If so, wouldn't this be the best statistic as you could decrease this parameter (small steps) for local effects and increase this parameter (big steps) for global effects? Am I misunderstanding something here?

Some decisions that have to be made are the shape (e.g., circle or square) and size (e.g., area) of the potential hot spot, and the number of potential B you will consider when computing the test statistic. The number of potential B is usually just taken to be as large as can be computed without taking forever. The shape and size are important tuning parameters, see student discussion question (7).

(7) We know that for scan statistics, under CRS, the rate is proportional to area. But this is also true for inhibition, right? So if we do not reject H0, it is possible the point pattern is inhibition, not CRS.

I think that’s right, the scan statistic has very little power to detect inhibition. You would use a different test for this.

(8) For spat stat, does that means the number of points in the window is fixed?

When you simulate datasets to approximate the p-value, each simulated dataset has the same number of locations.

(9) This week I have two questions about our final project: (1) Should we calculate the percentage as a total of all party votes or just the 2 parties? (2) Should the weights be a function of time or not? I.e. later polls, which is closer to the election day, have larger weights.

(1) Up to you, but be consistent in your definitions of X and Y. (2) Yes, this makes sense.

(10) Exam questions: (1) The response Bit, not Xit, correct? (2) What do we do with the 2020 data since Yit data is not released yet? It's said that the election results will not be final until Jan. 6 when Congress counts the Electoral Colleges votes (which take place Dec. 14).

(1) The response is Y-X, and you want to test whether its mean, B, is zero. (2) The near final percentages will be available this week, this is fine for our purposes.

(11) For the final exam, what should we do about unbalanced datasets between our training sets for the years 2012/2016. One is relatively smaller than the other.

Up to you!

(12) Bayesian statistics were mentioned in the lecture as playing a big role in these tests, but the tests themselves all seem relatively straightforward and don't need fancy stats or models. How do models/Bayesian stats come into play?

I don’t remember the context of this statement, but we will not do Bayesian stats for point patterns although these methods exist.

(13a) Was there a typo on slide 6? The chi stat decision rules had the same inequality twice.

(13b) In slide 6, should it be "Reject H0 if $X < X(m-1, 0.025)$ or $X > X(m-1, 0.975)$"?

Yes, my bad.

(14) Where can we find more details and examples about?

We will do some today (below). The textbook is also really thorough on this section of the material. The textbook authors created the spatstat package we’re using.

B. STUDENT DISCUSSION QUESTIONS

(1) What happens when there are clusters that are repulsed? For example, seats in a cafeteria are grouped around tables but tables are usually spaced to not be too close to each other. Would a K function show the seats as clustered or repulsed?

Yes, K should be high for small distances and low for longer distances.

(2) Can you combine two CRS tests to test for both global homogeneity and local features? Back to the hyenas example in the last lab: a group of territorial animals hunt together and form a cluster within their territory but mains a distance from other predators in the area. Can you use the quadrat test to pick up global repulsion dependence and use Clark/Evans to pick up local clustering dependence? If so, any problems for doing this?

(3) Would it be worth employing multiple types of CRS tests (K-function, Quadrat, Clark/Evans, Spat stat) in a single study? Would characterizing the difference in results between these tests be meaningful in any way?

(2) and (3): In theory, sure. It might be tricky to account for multiple testing and maintain a test with Type I error equal 0.05. But likely this is a potentially good idea. You could try rejecting if any of the p-values are less than 0.05/4, although because the tests are likely correlated this would be conservative.

(4) Given the advantages and disadvantages, how to choose the CRS test?

It's probably best to try them all to get a full picture. Spatstat is a bit of an outlier because it tests for a very specific alternative. K or C/E are the best for testing clustering/inhibition, and…

(5) How do you test for inhomogeneity?

The quadrat test is probably the best for testing inhomogeneity because is specifically compares the rates in subregions.

(6) All of the four tests don't account for the margin effect (points at the margin of the window). When the window is small, this can be a problem. How can we deal with it?

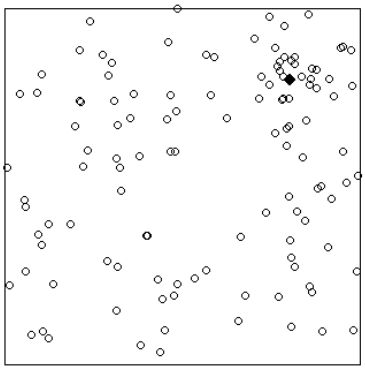
While some test statistics don't explicitly account for edge effects, the p-values based on samples from the null distribution should be OK because the same algorithm is applied to the real and simulated data.

(7) For the scan statistics, it seems like the size of the window could have a large influence on whether a cluster is found. Obviously you could try the test for many different sized windows, but then you'd potentially lose power due to a multiple comparison correction. Is there any other solution for this?

One option is to include regions of different sizes in the test statistics. For example, you first scan with radius one, then two, then three etc and take the test stat to be the maximum over all circle centers and radii.

C. BRIAN'S DISCUSSION QUESTIONS

(1) Archeologists scanned a region and marked the locations where they found ancient artifacts (open dots below). Based on the plot of these locations, they hypothesized that a region near a fresh water source a (solid diamond below) is a spatial cluster.



(a) In the context of this problem, what would a test for a completely random sample tell us?

BR: If you found a cluster then you might conclude there was a settlement near the water source.

(b) Interpret the results of the quadrat test below. Is m=16 a good number of quadrats?

BR: We reject the null hypothesis of a CRS. The expected count in each quadrat is 7.5, which is greater than 5 so this m seems OK. The largest observed value is 23 which is the potential hot spot near the water source (top right).

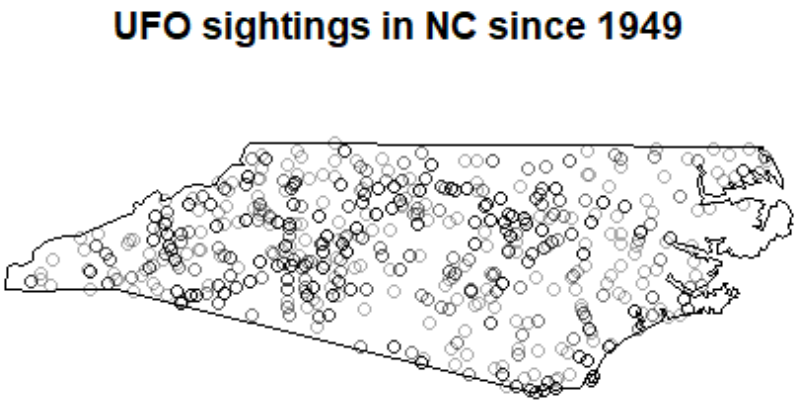
```
        Chi-squared test of CSR using quadrat counts

X2 = 50.4, df = 15, p-value = 2.071e-05

Quadrats: 4 by 4 grid of tiles
```

5 7.5 -0.91	7 7.5 -0.18	7 7.5 -0.18	23 7.5 5.7
6 7.5 -0.55	10 7.5 0.91	2 7.5 -2	13 7.5 2
7 7.5 -0.18	5 7.5 -0.91	3 7.5 -1.6	10 7.5 0.91
5 7.5 -0.91	8 7.5 0.18	3 7.5 -1.6	6 7.5 -0.55

(c) How would you test that the intensity of artifacts is higher near the fresh water source?

BR: To test this specifically, you could set B to be a circle centered on the source and let Y_1 and N_1 be the number of observations in the circle and the area of the circle. Similarly, let Y_2 and N_2 be the number of observations outside the circle and the area outside the circle. The model could be $Y_j \sim \text{Poisson}(N_j\lambda_j)$ for $j=1,2$ and we could test whether H_o: $\lambda_1=\lambda_2$ versus H_a: $\lambda_1>\lambda_2$.

(2) Below is a plot of UFO sightings in NC (and part of SC because spatstat is weird) taken from Kaggle.com.



(a) Visually, does there appear to be clustering, inhibition or inhomogeneity?

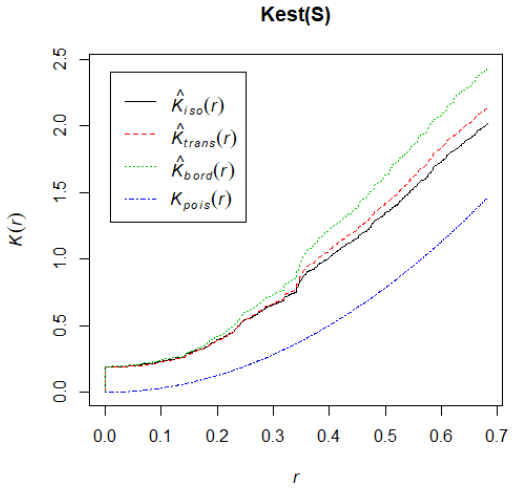
BR: It looks fairly random to me, but maybe there are more sightings in the Triangle and Charlotte?

(b) Why might a UFO researcher want to test if this is a completely random sample?

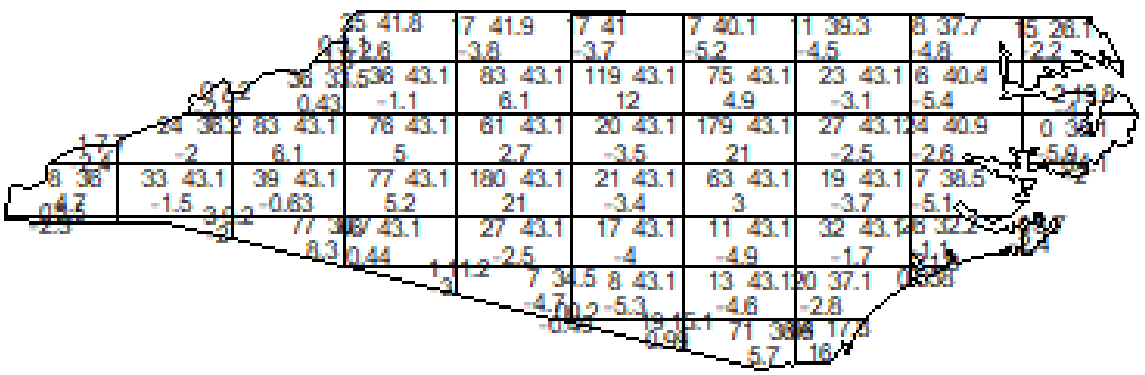
BR: A hot spot might indicate a significant alien landing base or something.

(c) What does the output below tell us about the distribution of these sightings?

BR: The K function shows strong clustering (I think there are multiple observations at some locations and this doesn't show up on the plot above) and the quadrat test rejects the hypothesis of a CRS.




```
> QT <- quadrat.test(S,nx=10,ny=7)
> plot(QT)
> QT
```



```
        Chi-squared test of CSR using quadrat counts

X2 = 1993.3, df = 56, p-value < 2.2e-16

Quadrats: 57 tiles (irregular windows)
```

(3) In the hotspot/scanstat lecture, we used the example of testing for a region with high cancer rates and assumed the population was constant across space. Let's explore how this might go without this assumption. First some notation:

- Let the sampling window D be all of NC.

- For region B, denote n(B) as the population of B and Y(B) be the number of cancer cases in B.

- Let B^c be B's complement, so that n(B^c) and Y(B^c) are the number of people and cases in NC but not in B

(a) If we performed this test in NC without considering population, what might be the result?

BR: We would almost certainly find hotspots in the large cities just because they have more people and thus a higher expected number of cases.

(b) Define a test statistic t(B) for the test that the cancer rate is the same inside and outside B.

A reasonable choice is the ratio of the rate inside (Y(B)/n(B)) and outside (Y(B^c)/n(B^c)) of B,

$$t(B) = [Y(B)/n(B)] \,/\, [Y(B^c)/n(B^c)].$$

(c) The overall test statistic is $t^* = \max_B t(B)$. How would you compute a p-value for this test statistic?

The assumption of the null hypothesis is that the expected rate in region B is proportional to n(B), so you would sample datasets from an inhomogeneous Poisson process with intensity n(B). We'll cover this in the final lecture.

(4) Read the article “Can we finally agree to ignore election forecasts” by Zeynep Tufekci at

https://www.nytimes.com/2020/11/01/opinion/election-forecasts-modeling-flaws.html

(a) Identify from this article challenges in forecasting the election results.

1) Behavioral correlation between people reading polls and making actual votes.

2) Polls not reaching full population.

3) Interpopulation variance (opinion varies widely even within demo groups) is high

4) Differences in turnout between poll responders and voters

5) Social media effect is hard to account for (previous models are not valid anymore)

6) Low replication (only 10 or so elections)

7) Weather doesn’t change (at least in the short term), the electorate does

8) Fundamentals might be useful, but voters are unpredictable. Local effects are hard to understand.

(2) Discuss measures that could be taken to address these challenges in a future election.

1) Quantifying how divisive the election will be might help. Candidates use polls to target populations too. Could change the way you report the results to minimize bias. How to prove it’s a real effect? Maybe study a population that hasn’t seen the polls?

2) Study response bias. Try to use cell phones (assuming they’re not). Spend more time “on the ground” rather than online.

3) With a large enough sample, maybe it balances out. Need a statistical analysis to find the right (homogeneous) groups.

4) Similar to (1) and (2), could study divisiveness in poll responders versus voters. Boots on the ground would be helpful. Make it easier for people to vote!

5) Weight groups by their social media use (quite users, active users, non-users). Study differences in social media use by group. (could be tricky due to self-reporting of media use) (issue in question 1 complicates this too) (also, social use changes way faster than a four cycle)

6) Spatiotemporal interpolation could help. Consider this in uncertainty quantification. Other counties elections. Test procedure by predicting polls.

8) First, figure out why voters are unpredictable (maybe effect of media, and level of trust in media and social media…see 5). From a stats perspective, it seems hard to untangle.

Focus less on polls, more on candidates (13 keys, https://hdsr.mitpress.mit.edu/pub/xhgpcyoa/release/2)

Study which party is more likely to change vote. (13 keys)

Unpredictability might increase over time. More issues, hard to figure out what people focus on.

(3) Why is modeling spatial correlation in the forecast errors important? How to estimate say the correlation in errors between NC and SC?

Spatial correlation is important because nearby states/counties have similar errors because people are similar across states/counties. This helps with prediction, but also in quantifying the probability of extreme scenarios like all states have positive errors. I guess you could look at errors from the previous year?