STAT 100 Notes March 23, 2010

The Day1 handout lists the following for week 10.

10. **Spatial distributions**. Clusters Real and Imagined. Travelling Salesman. Epidemics.

Spatial Distributions:

There were some spatial aspects to the distribution of tiger prey in the article by Gerow et al, and in the spread of the Africanized bee in Matis and Kiffe. And in the discussion of Zipf's Law, there was implicitly a comparison of urbanization patterns in North America and Oceania. Our further discussion of this topic is actually simpler: when we have a bunch of points scatter around an area, how can we measure or describe the patterns we see, and what difference does it make?

One common use of spatial distributions is to study the way plants are distributed in the wild. Consider a plant like the Ocotillo which is common in the Sonoran desert.

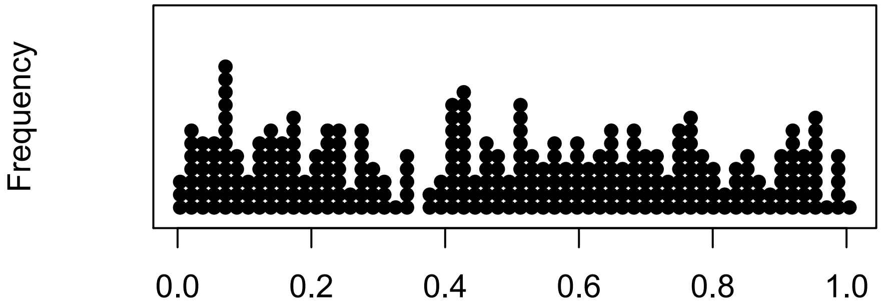
Photo: L. Weldon

It is sometimes quite common:

One question to ask is: Does it grow in clusters or randomly over a large area? What does *randomly* mean in this context?

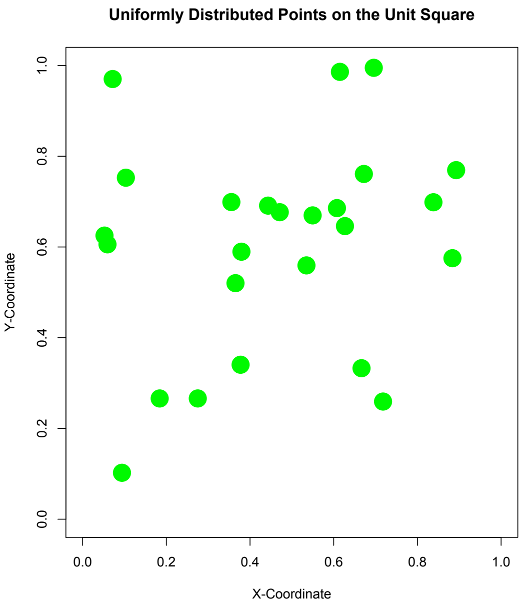
To approach this topic, lets look at an example of randomly distributed points:

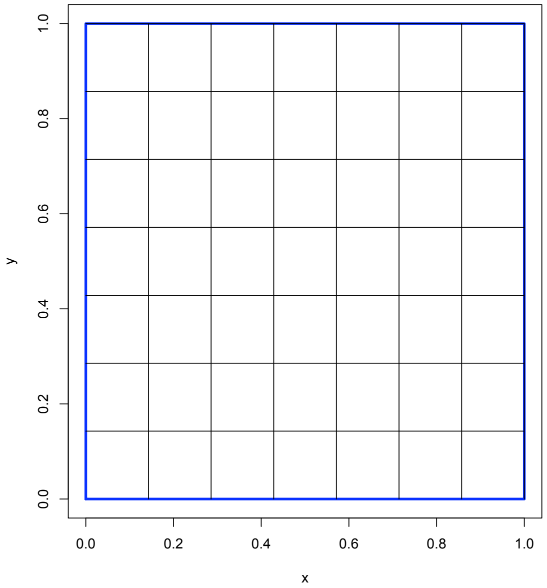
First, in one dimension, a sample from a uniform distribution will look like this:



In other words, equal frequencies across its range of values. We can use this kind of distribution for both the x and the y coordinates on the unit square.

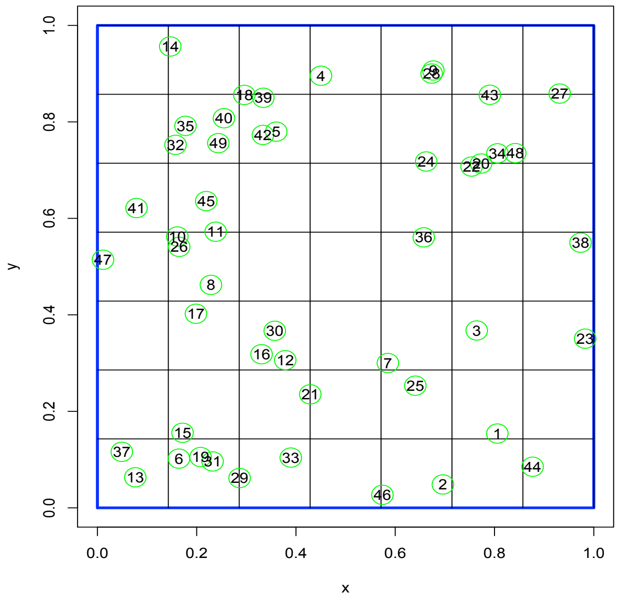
For a discussion of clustering over an area, we use the unit square (a square with sides of length 1) and simulate a sample of 25 points.





These were simulated using a simulator that treats every location in the unit square as equally likely, for each point. There is no dependence from one point to the next, and yet it looks like there might be some.

Next consider the 7 x 7 grid of the same square. If I generate 49 points, will there be one in each square? Not likely if they are truly random! Here is what really happens, typically:



There appear to be large gaps without points and clusters where there are several.

It looks like there is some proximity process clustering the points. But actually the process generating the points is independent and uniform over the whole square.

How can we summarize the degree of clustering in such a display? One method is to look at the distribution of the number of points in each small square. For the result in the simulation above:

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]   0    1    0    1    2    0    1
[2,]   0    4    4    0    1    3    0
[3,]   1    2    0    0    0    2    0
[4,]   1    3    0    0    1    0    1
[5,]   0    1    3    0    1    1    1
[6,]   0    1    0    1    1    1    0
[7,]   2    3    2    0    2    0    1
```

←--- This is the count in each square. Note the range of values: 0,1,2,3,4

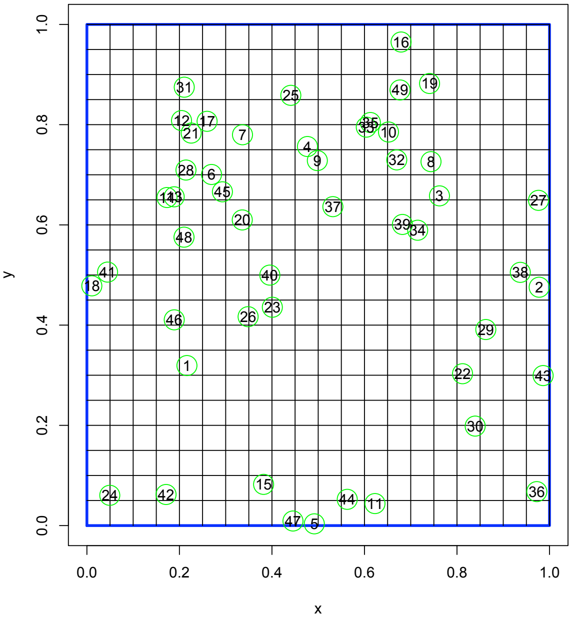
```
count
 0  1  2  3  4
20 17  6  4  2
```

This is the square-count distribution.

In a uniform distribution sample, there are quite a few "empty spaces".

Q: What would happen to the proportion of empty squares if we decrease the square size by increasing the number of grid cells?

Suppose we had the same sample size (n=49) but a 20 x 20 grid size ….



```
count
  0   1   2
352  47   1
```

So 352/400 cells are empty.

So even if every spot is equally likely to be selected in a uniform sample, the sample certainly does not look uniform in the sense of "regular".

There is a new distribution that mathematics can show will be an excellent model for the number of cells that have each of 0,1,2,3, … points. Of course it depends on the sample size and the number of cells. Actually, if there are N cells and n points then the average number of points per cell is n/N. Call this m (it is a mean number of points per cell). The Poisson distribution with mean m is the model that works in this situation. Here are a few examples of counts from a sample of 400 cells with a mean of 49/400 in each cell:

```
> table(rpois(400,49/400))

  0   1   2 
354  41   5 
> table(rpois(400,49/400))

  0   1   2 
355  42   3 
> table(rpois(400,49/400))

  0   1   2 
360  37   3 
> table(rpois(400,49/400))
```

```
  0   1   2
350  49   1
```

So you can see that it is pretty close to the one we got from the spatial distribution.

How do we use this model to answer a question about the actual spatial distribution of plants? We can summarize the spatial distribution of the plants in a convenient square grid, and then see how close it is to a Poisson distribution. If there are more empty cells than a Poisson would have, then that means the clusters a tighter than if they were generated by a uniform distribution. If there are fewer empty cells that would be suggested by a Poisson distribution, then we have evidence of segregation of the plants – the idea that a plant in a square tends to prevent another plant in the same square. Such findings can be useful for biologists.

Poisson Distribution: $P(X=x) = e^{-m}m^{x}/x!$ x=0,1,2,3,…. Where m= mean and e=2.718…

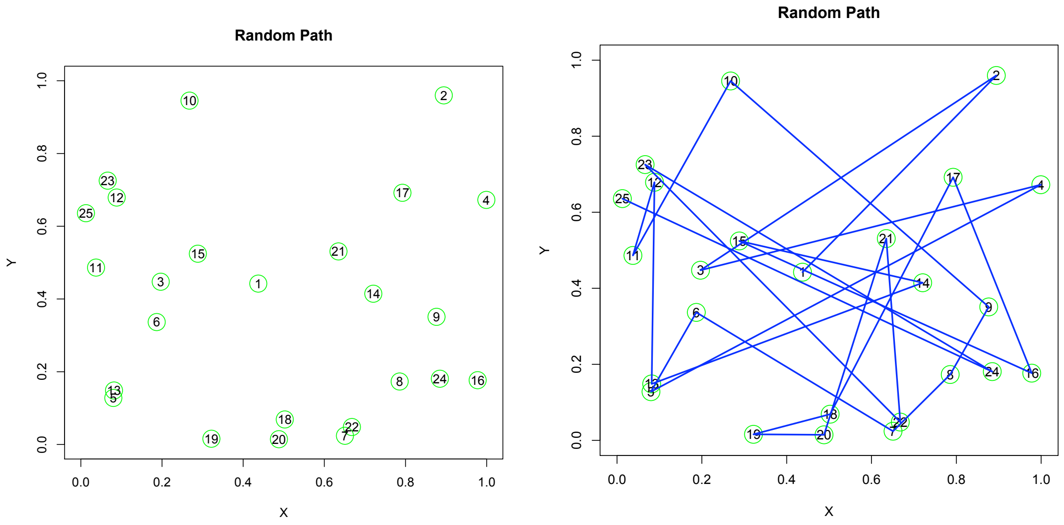
You don't have to know this formula but I present it here for your reference. There are many useful applications of the Poisson distribution.

Note the possible application of this model to the detection of infectiousness of an epidemic. Are adjacent cases there by chance, or by disease transmission?

Travelling Salesman Problem:

Another feature of spatial distributions is the path that joins them. For example, if the points are locations that a travelling salesman has to visit, what is the most efficient route (minimum distance) through all the locations?

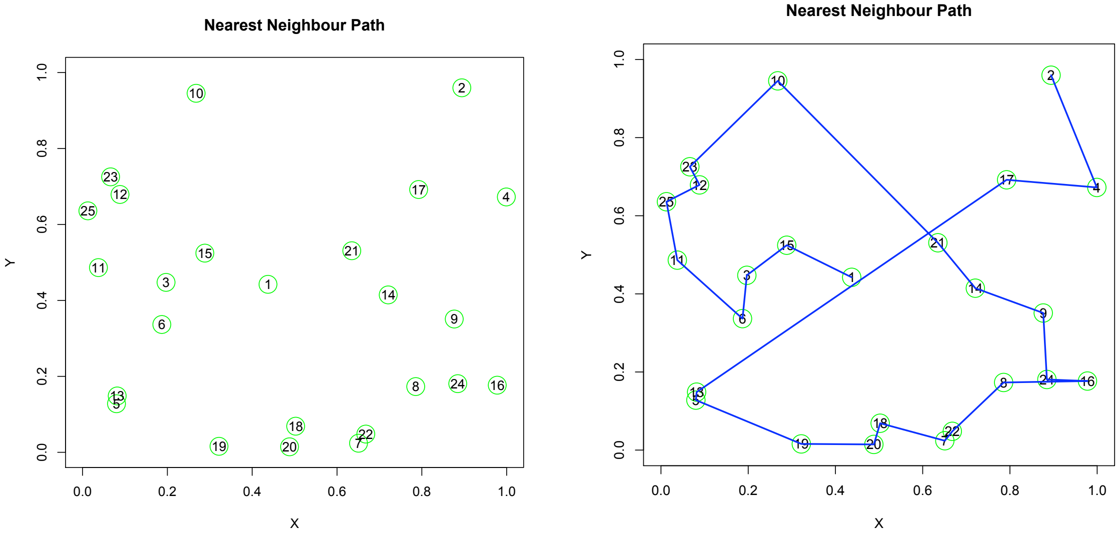
We again turn to a simulation to see what might happen in this situation. Suppose there are 25 locations spread uniformly (in the sense we have discussed) over the sales area. The salesman would not likely follow the route defined by his initial list of locations. If he did the length of his path would be about 14 units (1 unit is defined by the side of the square being 1.)



Q: What strategy would be a better one to minimize distance travelled?

One idea is to start at a random point (say point number 1) and move to nearest neighbours until all points have been visited. The length of the line in this case is 4.8, much smaller than the 14 we had with the unsorted path. (See graph next page).

However, the nearest neighbour criterion does not produce the optimal path. Can you improve on it?



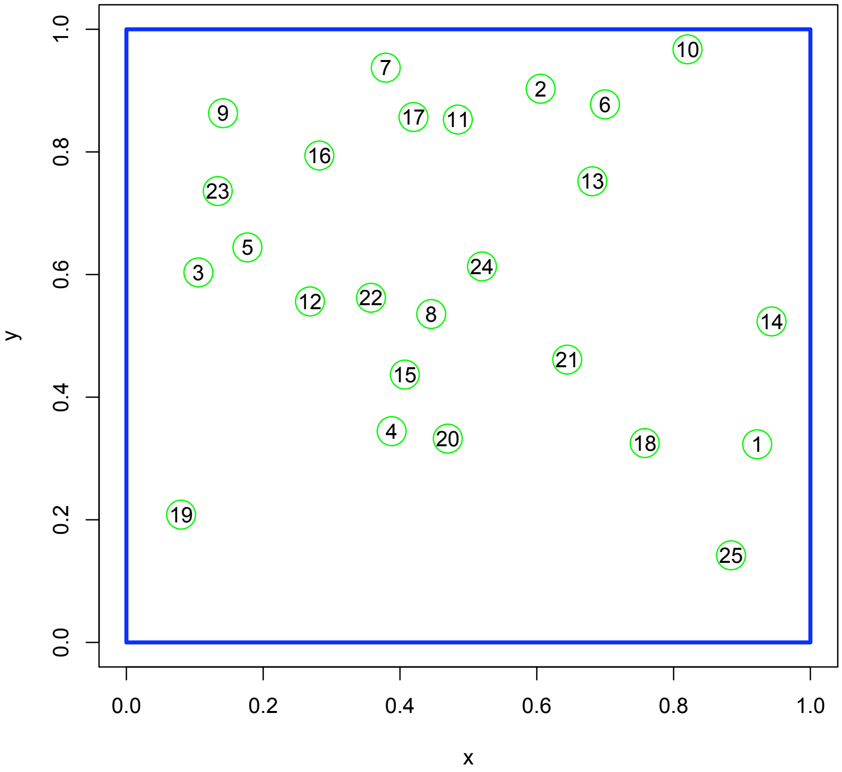
Hint: How about a link from 10 to 2 and 17 to 21 to avoid a long link from 13 to 17?

The lesson here that the travelling salesman problem is a difficult problem. However, it is a problem for which solutions are available. All we can do at this stage is to understand what the problem is. Of course, it is also useful to think of other applications of the technology such as security-monitoring routes, wholesale distribution routes (like to food stores or gas stations), road building among communities, postal delivery routes, and many more.

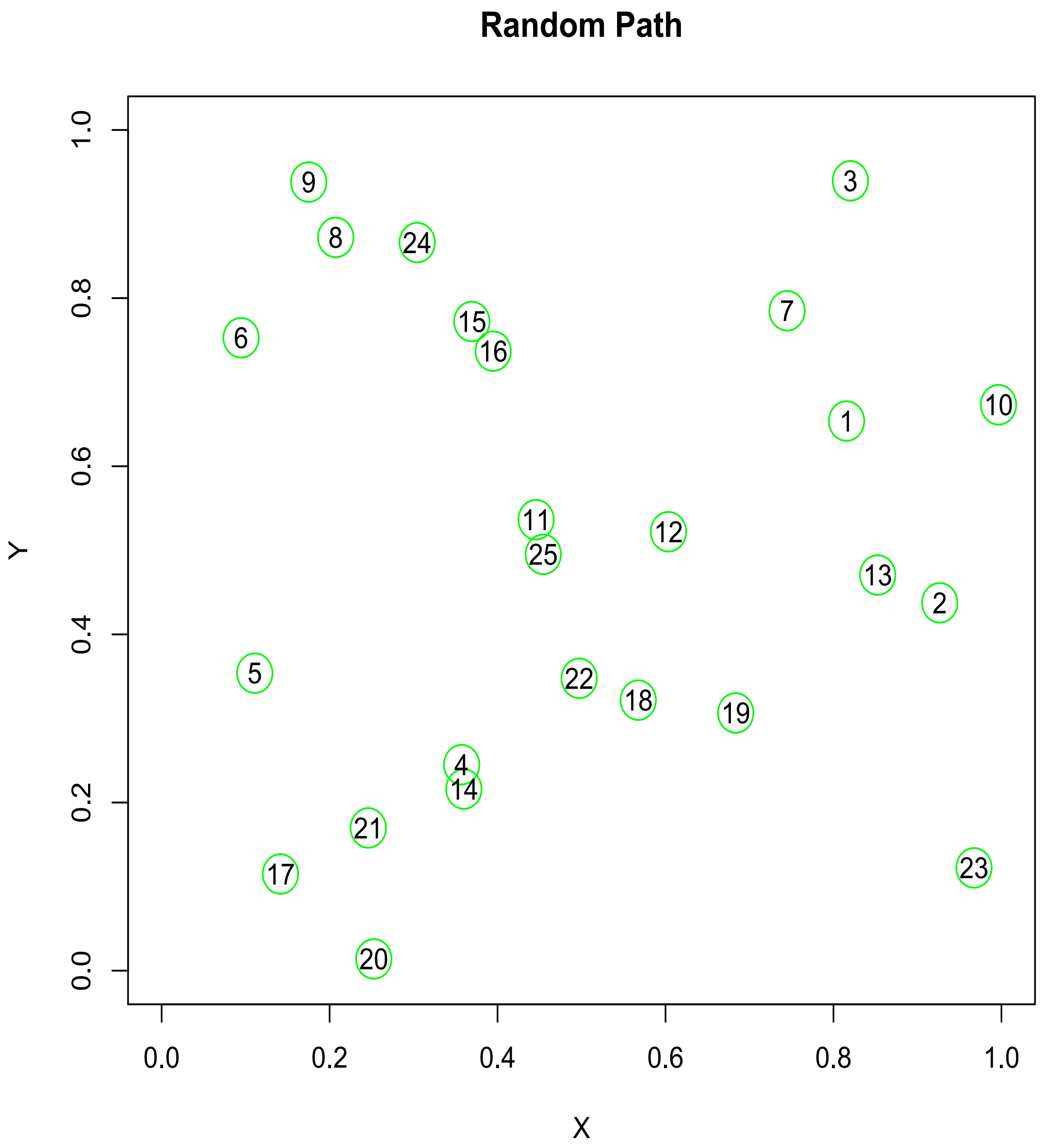
Assignment #8 (Due March 30, 2010, a 4 pm)

1. The survival time data we summarized from the students attending class on Sept. 18 turned out to be approximately a straight line through the origin and with slope 1/180.
a) What does this imply about the survival time distribution itself? Explain.
b) Would the straight line summary work well if the class had a wider range of ages (and exposure durations)? Explain.

2. For a Poisson distribution with mean 0.25, the percentage of zeros in a sample of size 100 is about 78%. ($e^{-.25}$=.78). Use the spatial sample below to tally the actual distribution of points per cell, and comment on whether the Poisson model is an adequate predictor of the 78%.



3. For the point spread below, the ordered path length is 10.7. By choosing a different order, make the path, starting at point “1” and touching all the other points, that has a much smaller total distance. Aim for at least 4.4 or less. The units of measurement are given by the side of the square. (if your graph is not quite square, you need to make it square, but the size of the square is not important.) Your answer should consist of a sequence of integers starting with 1, and including the other integers in the range 2 to 25, and tabulate the rough measurement of distance for each link in the chain, as well as the total length.


Random Path
Y
X