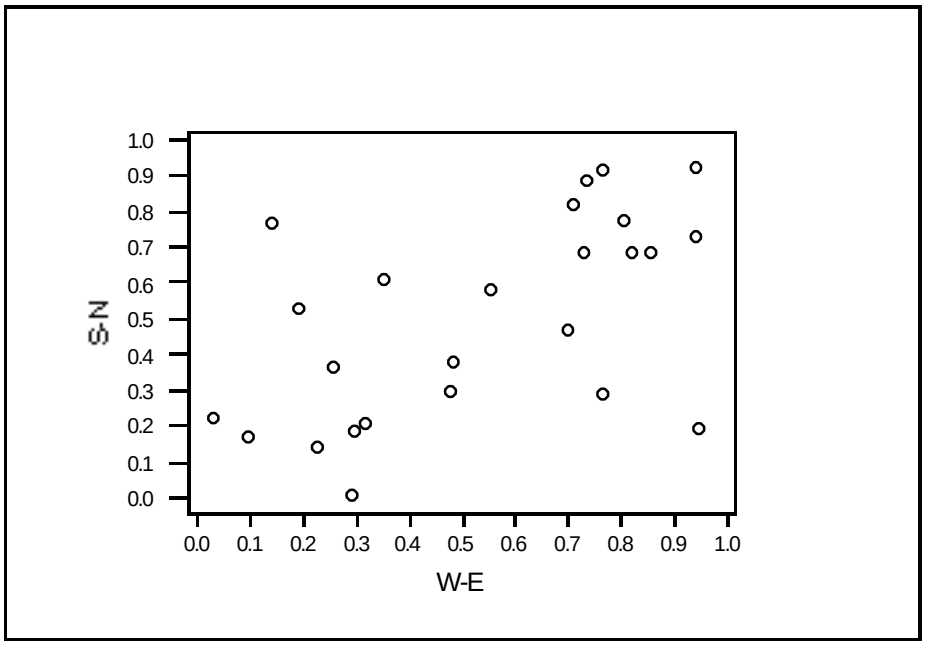
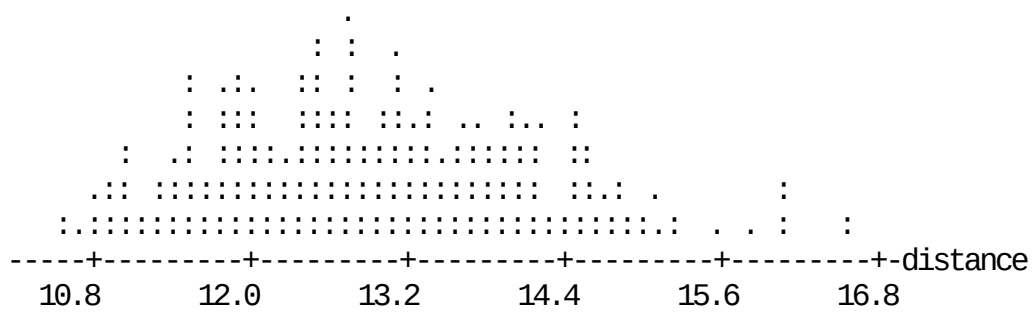
Today: More on Traveling Salesman Problem (and Optimization)
Extension to Spatial Point Distributions and applications
More on applications of Regression Prediction

Recall from last time:

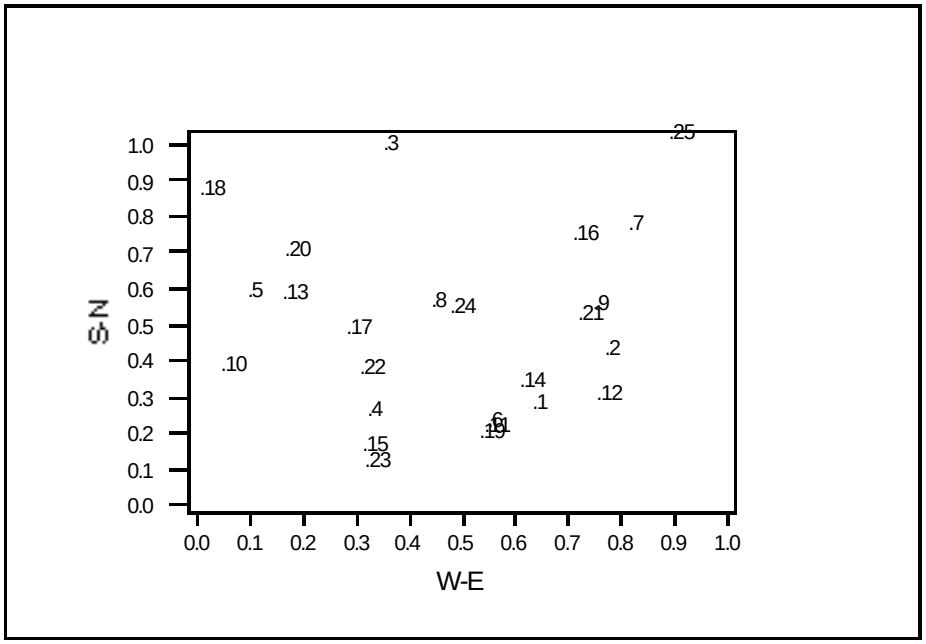




With 25 points selected at random from the unit square, we could not do much better than a distance of 10 units for a circuit through the 25 points. This "best route" is not very good – in fact it can easily be improved manually. But consider the population sampled here – all possible distances (a population of size 25! = 25 x 24 x 23 x …x 2 x 1). Our sample is random but only a time proportion of the population. This is OK for estimating a mean, but not for estimating a minimum.

Our next strategy for the traveling salesman is to use common sense! Move to the nearest neighbour.

Here is another example:



and the random paths for this graph have total path distances in the range 9-14 usually :

```
    .  .   . . :..   : .:  :..... : ..: ..   .      .
-------+---------+---------+---------+---------+--------distance
    9.60      10.40     11.20     12.00     12.80     13.60
```

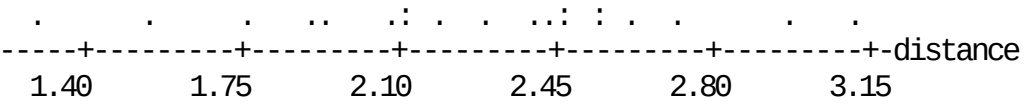
These distances are for random orderings of the points in a single sample of 25 points. But this range is also usual for random point distributions of this size, n=25, on the unit square. However, if we use the nearest neighbour approach, on a variety of samples of size 25, we get the total path distances to be much less:

```
                    .
    ..   .     . :   :  :  .  .  .  ..    .   ..   .
-------+---------+---------+---------+---------+--------distance
    4.00      4.50      5.00      5.50      6.00      6.50
```

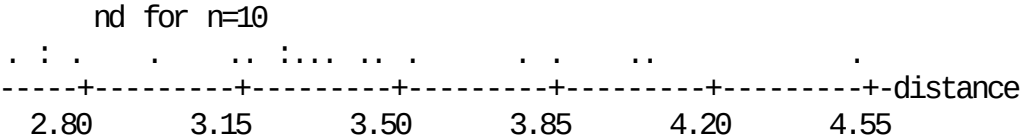
So this simple rule gives a big improvement over random paths.

Do we need to worry that the particular sample of points we are working with might be atypical? No, because we have shown the results for many different selections of n=25 points.
Let us now explore the relationship between the number of sites to visit, and the distance required (using the nearest neighbour approach to determine a path). For n=25, it looks like a distance of 4.0 units can be easily achieved. What about n=5, 10 or 50?

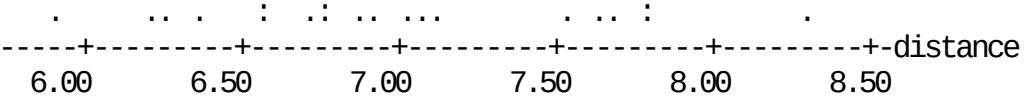
Here are the results for n=5:



nd for n=10



And for n=50:



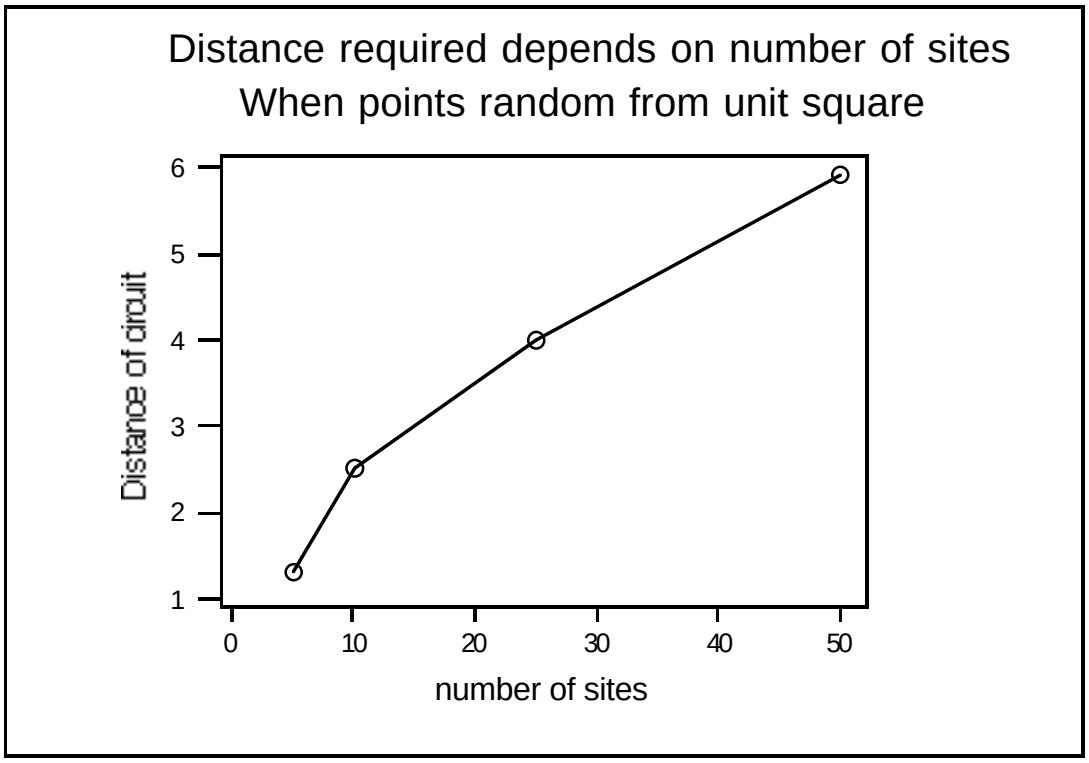
A rough summary would be:

n	distance (that nearest neighbour finds)
5	1.3
10	2.5
25	4.0
50	5.9

This suggests a relationship between the number of sites and the distance required for a circuit. Is this useful? It might be used to estimate workload for service personnel (or salesmen) with different numbers of accounts (assuming a random spatial distribution). The size of the area can be taken into account by multiplying by the above distance by the actual length of the square's edge.

What about rectangular regions? Is a 1 x 4 square more or less work to cover than a 2 x 2 square?

Where are we? We first looked at random paths through a fixed set of sites – we could see a wide range of path distances but not much help in finding a short circuit path. Then we looked at nearest neighbour paths for several different simulated site distributions. This allowed us to predict the length of a short path without having to know the exact sites.



Big picture:

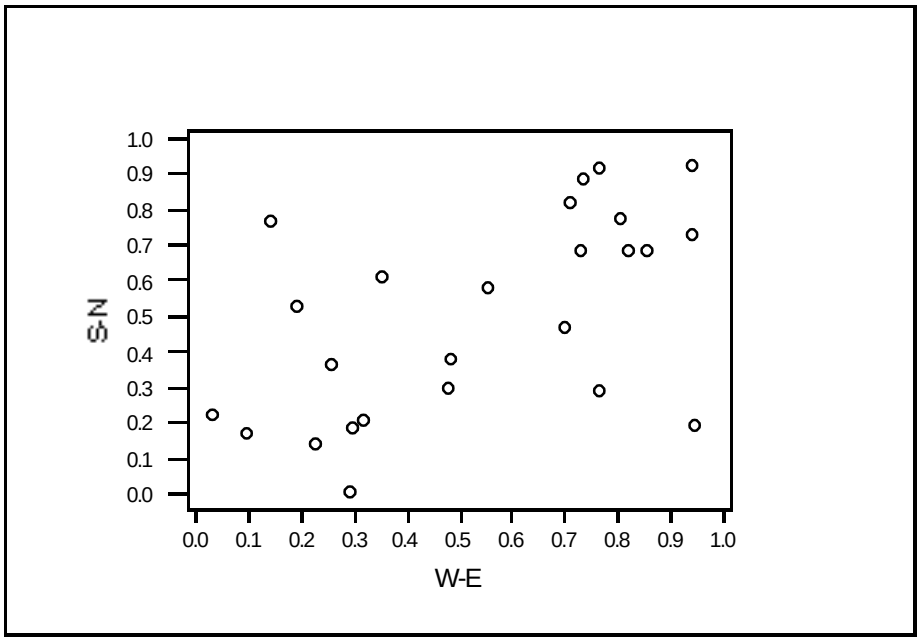
1. Random selection can be used in various ways to approach certain deterministic problems. It is important to use contextual knowledge and common sense when deciding what is a good approach.

2. Optimization is the usually domain of the mathematician, but simulation can give sub-optimal but useful approximations.

Reminder: See Tanur article which discusses some additional optimization applications, as well as the general problem of the use of chance in optimization.

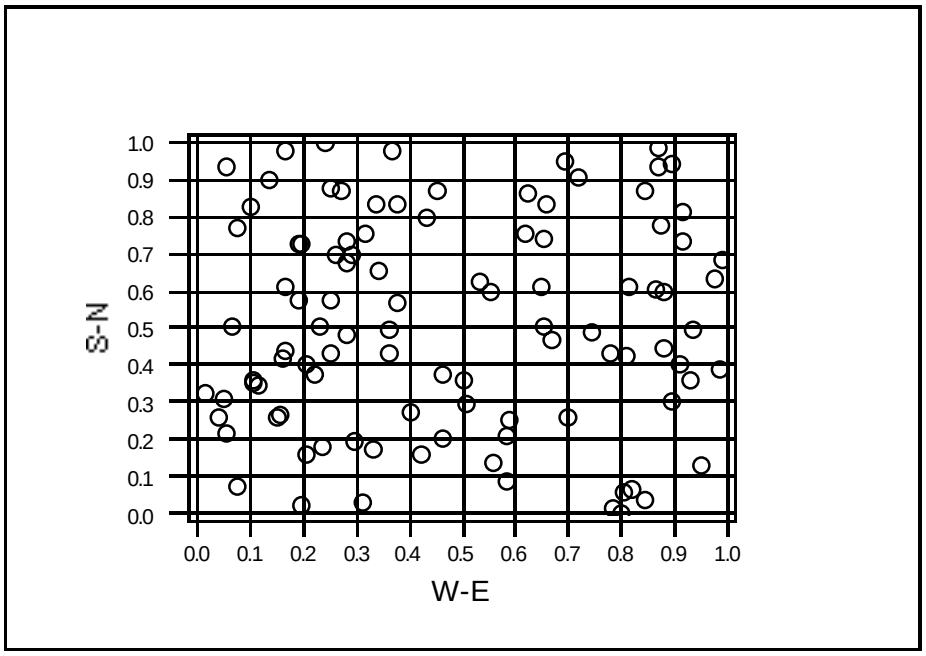
The prediction of distance from number of sites is a “regression” problem. Although we have only discussed regression prediction using a straight line fit to points, a similar approach is possible for curved functions.

Spatial Point Distributions: Look again at our traveling salesman sites ....



Does this look "uniformly" distributed on the unit square?

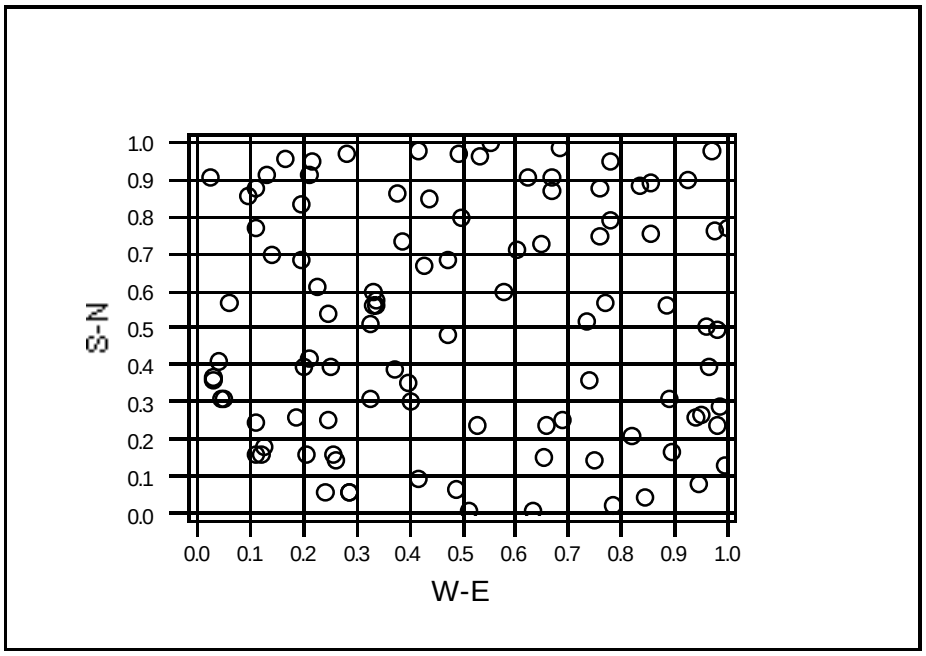
How much "clumping" does a uniform distribution have?



From which we have

Row	noincell	nocells
1	0	34
2	1	39
3	2	20
4	3	7
5	4	0
6	5	0
7	6	0
8	7	0

And here is another one:



and its frequency summary:

Row	noincell	nocells
1	0	39
2	1	35
3	2	17
4	3	6
5	4	2
6	5	1
7	6	0
8	7	0

The lesson here:

Lack of a reason for clumping does not prevent clumping. But through simuilation, one can judge when there is enough clumping to think that there is a reason for it.

Applications: Distribution of animals or plants or disease cases
Crime spatial analysis (e.g. credit card fraud, violent crimes, ..)
Astronomy
Traffic accidents

Further comment on Regression Analysis:

simplest form: fitting a line to points to predict "Y" from "X".

For example, we were predicting distance travelled from number of points (in units of the size of the square area.), and although we did not use a straight line, the idea is the same for curve-regression.

Other applications of regression-prediction:

Credit Card Fraud predicted by usage characteristics
Security Risk predicted by traveler characteristics
Marketing: Predicting buyer behavior based on information about potential buyers
Remote Sensing: Predicting illegal activites from spectral mix (vegetation, transprtation)
(Marine piracy?)

When predicting a variable "Y" using several characteristics, the technique is called multiple regression analysis, but the concept is a simple extension of the straight line prediction idea discussed last time.

Next topic: Quality Control – Read Tanur p 170-177 but more in class.

Next assignment (#7, Due Wed Nov 27) will be defined Monday& Wed Nov 18&20th