STAT 100 Review April 6, 2010

Week 12 Review

1. (**Sample Means** - 115 choices) I will review the theory and application of the distribution of sample means, including the direct implications for the investment portfolio and insurance applications.
2. (**Spatial Distributions** - 85 choices) This recent material needs some repetition including the discussion of plant clusters, cell counts, the traveling salesman problem, and path length problems.
3. (**Time Series** - 67 choices) Several topics fit here: random walks, the stock market, smoothing, simulation, forecasting time series, and a few others.

Sample Means: Theory and Applications

Sampling: Selecting a portion of a group ("population") in order to learn something about the group.
Note: Usually we are interested in some feature of the population, and we take measurements on the sample, related to this feature, in order the learn about feature in the population.
e.g. We may be interested in the income distribution in a certain community. We take a sample of persons in the workforce of the community and determine their incomes. We summarize this income data (histogram, or perhaps mean and SD) in order to describe the income distribution in the population. Note that we did not sample the entire community, only a relatively small sample of persons.

Does this actually work? It depends on how we selected our sample of persons. The one method we have described in detail is called "random sampling".

Random Sampling is a method of selecting a sample from a population – the "random" part specifies that the method used must be such that all possible samples of a certain size must have the same chance of being selected from the sample as any other sample of the same size. With random sampling, it is possible to assess, from the sample itself, how accurate the sampling estimates (of histogram, or of mean and SD) are as estimates of the population features.

Population features that are numbers (such as mean, SD, for example) are usually called "parameters". So we use "sample estimates" to estimate "population parameters".

The most useful theory we cover in this course concerns "sample means". Sample means estimate population means: a sample mean is an estimate of a population parameter, the population mean.
In statistics, whenever we use a sample estimate to estimate a population parameter, we like to have some idea of the precision of the estimate. "Precision" here means the closeness to the population mean of the sample mean.

If we say that the average family income in a community is $41,234, based on a random sample of 100 families, and perhaps there are 10,000 families in the community, it is important to know how close $41,234 is to the actual average income of all 10,000 families. With the information I have recorded so far, there is no way to know this. But actually, the sample values that gave us the mean of $41,234 can also be used to compute the sample SD. This is very useful because

i) the sample SD estimates the population SD

ii) the population SD determines the variability of the sample mean

Before we move on, lets examine these points in more detail …

i) should be intuitively unsurprising – the fact that the sample SD estimates the population SD. If it is surprising to you, look at the formula for the sample SD – it is based on the typical deviation of the sample values from the sample mean, and of course the population SD is based on the typical deviation of population values from the population mean.

ii) is not quite so intuitively obvious. But it turns out that the variability of the sample mean(based on a random sample of size n), as measured by the SD of the sample mean, is the population SD divided by √n. So, the larger the population SD, the larger the sample we need to take to get a desired precision of estimate.

Now lets apply this result to the example of estimation of the average family income in our community of 10,000 families, based on our random sample of 100 families. The theory says that the SD of the sample mean (we have the one value $41,234 in our sample) is the population SD divided by √100.

Terrific! But, we don't know the population SD. However, we do have an estimate of it in the sample SD. I have not mentioned the value yet but it could be calculated from our sample of 100 family incomes: so lets say that the same sample that gave us the mean of $41,234 also gave us SD = $5,678. We could then guess (i.e. estimate) that the population SD is about $5,678, and therefore the SD of the sample mean from this sample of 100 family incomes is $5,678/√100 = $568. This tells us that the process that produced the mean $41,234, if repeated many times, would produce other sample means that, if we were able to calculate their SD, would have an SD of about $568.

OK, but how does the SD of the sample means relate to the precision of the sample mean as an estimate of the population mean? It is actually the same number!
To see why we need one more fact: The sample mean is an "unbiased" estimate of the population mean. That is, some sample means will be larger that the population mean, and some will be smaller than the population mean, but *on average* , there is no tendency one way or the other. Put another way, the average of the sample means, if we were to take a huge number of samples of size n, would be the population mean. So the SD of sample means (typical deviation of sample values around the sample mean) estimates the typical deviation of sample values around

the population mean. In other words, the precision of the sample mean as an estimate of the population mean is estimated by the SD of the sample divided by $\sqrt{n}$. In our example, we have that the mean family income in the community of 10,000 families is estimated as \$41,234 ± 568.

But if 568, the estimated SD of the sample mean, is only a typical deviation – a typical difference between the population mean and the sample mean – might the difference be quite a bit larger than 568? Yes, but rarely would it be larger than twice this number, and almost never thrice this number. (This is a Central Limit Theorem result, which we will cover soon.)

Note that the distribution of sample means has an SD estimated as \$568, and that this is much less than the SD of the incomes in the population (which is \$5678).

Theory of Sample Means Applications

1. **Investment Portfolio**

   We imagined we knew the probability distribution of returns to a \$1 investment, and we described this as:

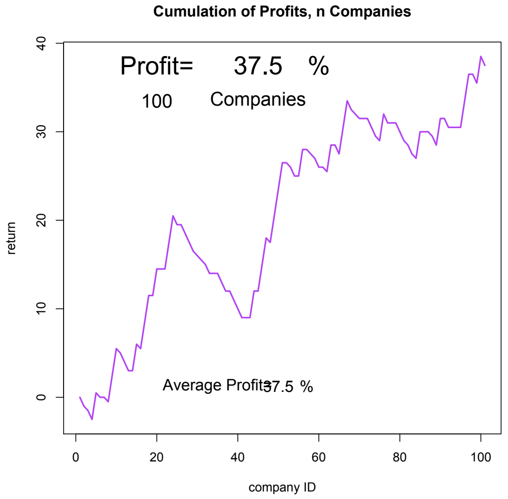
   -1 with probability .25
   -.5 with probability .25
   0 with probability .25
   3 with probability .25

   This probability distribution can also be thought of as a "population" as long as we sample "with replacement".

   We took a random sample from this population (think "random sample with replacement") of size 100.

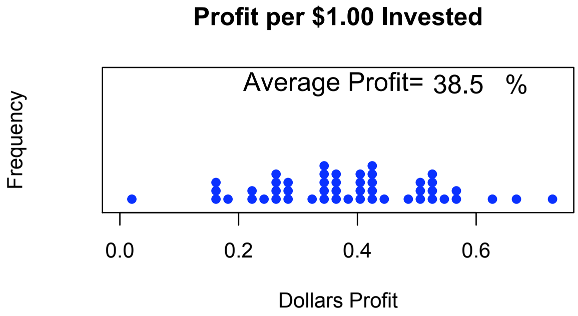
   The result was a sample mean of somewhere in the approximate range 0 – 0.7, so the return to a \$1 investment tended to be positive.

   Here is a record of what happened to the 100 similar companies – we scan through companies 1 to 100 and cumulate the profit as we go.



In this one simulation of 100 companies invested in, the average profit (that is per $1 invested) was $0.375. This might be a bit surprising since each individual company had a good chance of losing money.

Now lets see how typical this is – let's do the whole thing 50 times (that is , 50 simulations of 100 companies).



This dotplot shows that in every one of the 50 scenarios simulated, a good profit was made. So combining the 100 profits from the risky companies appears to have eliminated the risk. We will have more to say about this but first let's see how the theory of means could have predicted this result.

Each dot in the dotplot above is computed from the mean of the returns of 100 companies. The smallest dot is at about $0.02. The mean of the 100 company profits must have been $0.02. Similarly for the other dots. Each dot is a sample mean, and because we have insisted that the population underlying all the sample is the same (defined by 0.25 prob of -1, etc) the population underlying all these dots is also the same.

Note that the population values were {-1, -.5, 0, 3.0} . In other words they ranged from -1 to 3.0. But the average of 100 of these only ranged from 0.02 to 0.72 which is much less. (-$1 to +$3 is $4, but $0.02 to $0.72 is only $0.70). The average of a sample of 100 from this population was much less variable than the original population. In fact the range of sample mean values was only about 1/6 of the range of population values. *The narrowing of this distribution of sample means and the positivity of population average is what ensures the profitability of the portfolio.* We have demonstrated this with simulation. But what would our theory predict?

First of all it is better to use SD, rather than range, to measure variability. Now the SD of the population can be computed easily to be 1.70 (SD of {-1, -.5, 0, 3.0}). So the SD of the sample mean of 100 sample values is 1.70/√100 = 0.17. The population mean was 0.375 (mean of {-1, -.5, 0, 3.0}), so we would predict our sample means distribution would be 0.375 ± 0.17, or if we use 2 SDs, 0.375 ± 0.34, which is 0.035 to 0.725.
This is in fact what we observe!

Consider what we have just shown: the theory predicted that the sample mean of the profit (per company) from 100 companies would vary from one sample of 100 companies to another sample of 100 companies, but that for any one sample, the sample mean is likely to be in the range 0.035 to 0.725. We calculated this directly from our knowledge of the population (known in this demonstration), using the known distribution of sample means from this population. But to make sure we understand this theoretical result, we also found similar information by simulating the distribution of sample means – that is, simulating 50 samples of size 100, and recording the sample mean in each case. It is of course not surprising that we get similar results from the two approaches – the reason for doing it this way is to show that the result of random sampling (on the distribution of sample means) is really what the theory predicts.

However there is still a step in the theory that we have not demonstrated. Recall that our original theory of means arrived at a statement about the population mean, not only about the distribution of sample means. Again, we use the distribution of sample means to imply the location of the sample mean. Since the sample mean is unbiased, we say that the population mean will lie in the interval

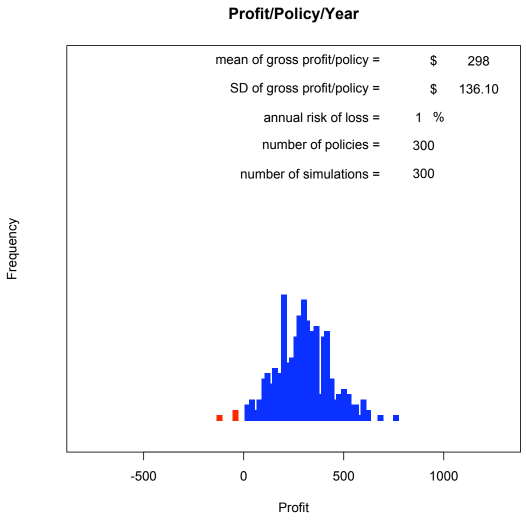
Sample mean ± 2*SD/√n

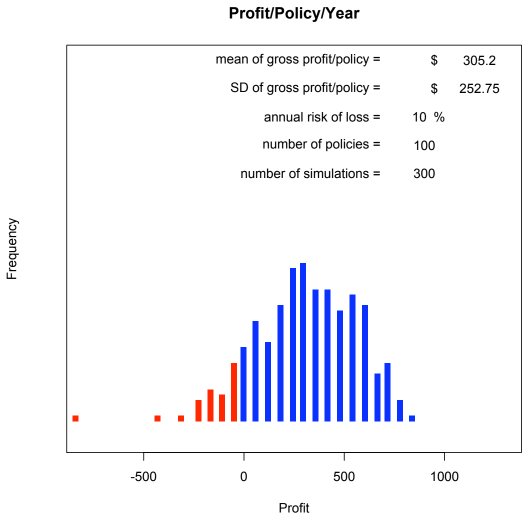
95% of the time. The 2*SD/√n can be thought of as the distance of the sample mean from the population mean (i.e. from the mean of sample means), or, equivalently, it can be thought of as the distance from the sample mean from the population mean. This latter interpretation leads to the idea of a 95% Confidence Interval for the population mean.

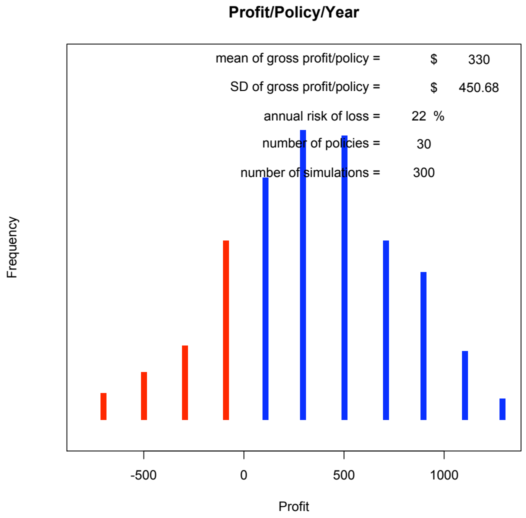
But we still need to justify the 95%. But that is a result of the approximate Normality of the distribution of sample means. All Normal distributions have the property that 95% of the distribution lies within 2 SDs of the mean of the distribution. And the fact that the distribution of means is approximately Normal is just the "Central Limit Theorem".

2. Insurance

The same theory is what produces the big-company advantage in the insurance business. Recall the difference in outcomes from our simulations for a small insurance company and a large insurance company.







While the overall mean does not change much (from simulating the annual experience 300 times in each case, using the same population for each sample), the distribution of sample means from each year narrows as the number of policies (sample size) gets larger, and since that narrowing is closing in on a positive value, the larger company has much less red – that is, chance of losing money in one year. This affect can be accentuated if the population mean is even closer to 0, but still positive, and reducing the premium a little will have this affect. So if the larger company wants to drive out the smaller company, it would squeeze the premium down closer to the average cost of a policy, and the smaller companies would have a big chance of losing money while the big company has a small chance of losing money.

Again, *The narrowing of this distribution of sample means and the positivity of population average is what ensures the profitability of the larger company.*

There is one assumption in everything said so far today that needs to be admitted at this point and explained: *independence.*

When we assume a random sample is selected from a population, we imply that the selection of one value in the sample does not affect the selection of the next selection in the sample – that the values selected are "independent" of each other.

When we apply random sampling theory to insurance, we have to ask if the outcome of one policy affects the outcome of another policy. It does seem reasonable to assume this is true, although it is certainly not exactly true. A big disaster would be a counter example.

However, in the stock portfolio application, the independence assumption is more suspect. It is likely that many companies shares will rise and fall together – not independently. Nevertheless, the phenomenon shown survives the violation of this assumption – but to get the portfolio result, we need to try to find companies in different markets so there is some degree of independence in the outcomes.

This completes the review of the distribution of sample means.

Spatial Distributions

There are two features of spatial distributions that we will review:

1. The amount of clustering in a scatter of points
2. The length of a path through a scatter of points

In both cases we will use a uniform spatial distribution as our benchmark.

1. Clusters

Compare the following summary tables of three spatial distributions:

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

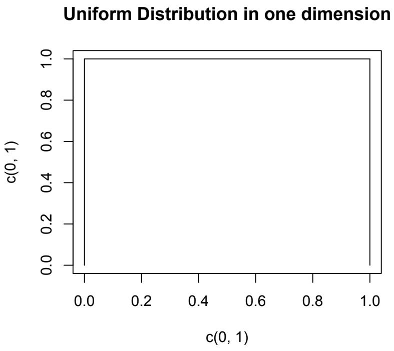
1	0	0	4	0
0	1	0	1	2
0	1	3	0	0
1	2	2	1	0
1	2	1	0	2

0	0	5	0	0
0	0	4	0	0
0	0	0	0	0
0	4	0	0	0
5	7	0	0	0

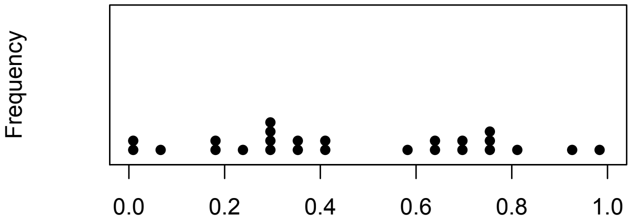
Each “field” has 25 “plants” in it. The first one has a very regular distribution, the third one seems very clustered and the middle one is in between. What we showed in the earlier lecture was that the second pattern could occur even if the underlying distribution (i.e. the population distribution) was “uniform”. The apparent clustering in the middle one could result (in a sample) even when there is no tendency to cluster. We need some way to tell if the apparent clustering is from a tendency to cluster, or just a result of chance selection from a uniform distribution.

To understand this last paragraph, it is important to understand the idea of “uniform distribution” .

The histogram, or density curve, of a uniform population looks like this:

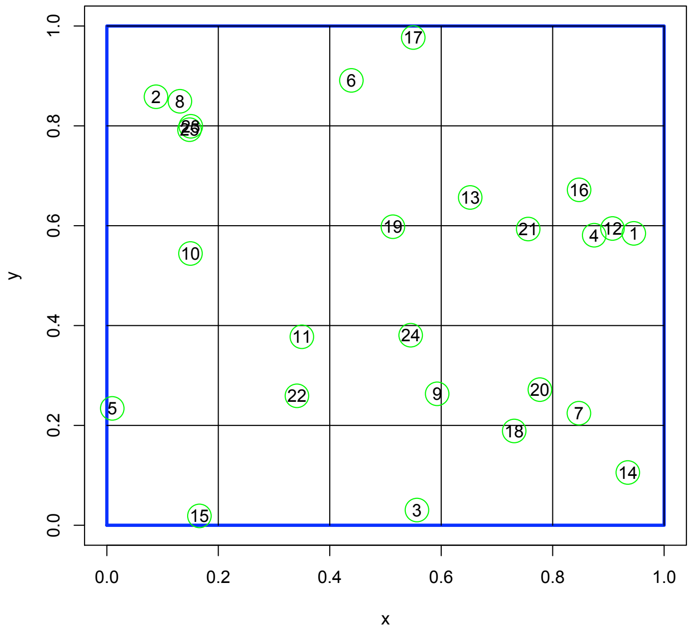


What this means is that every value in the population, between 0 and1, occurs equally often. As sample from this distribution looks like this (as a dotplot):



The dotplot has limited resolution so it lumps together values that are close. But note the values are spread fairly evenly across the interval (0,1). But not perfectly evenly … that is the nature of random samples from a uniform distribution. Now we want to move into a spatial scatter of points that are obtained as a sample from a spatial uniform distribution. We get a uniform spatial population by making the two coordinates uniform in one dimension. And, we get the same apparent "clustering" in two dimensions as we see in one dimension, *even when there is no clustering at all in the population.*

For example, if we use a unit square as our "field" then the distribution of plants from a spatial uniform distribution (that is, uniform on (0,1) for both the x-coordinate and the y-coordinate), we might get:



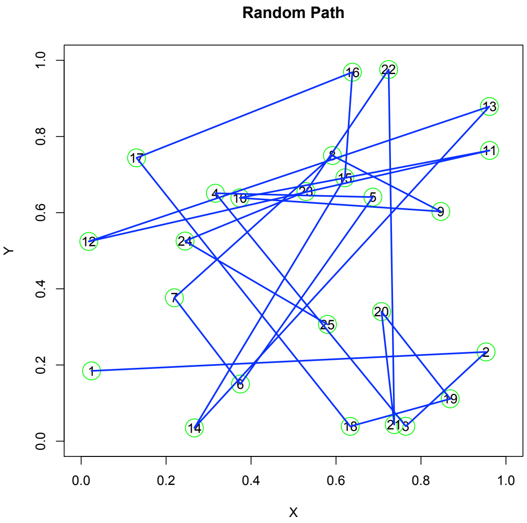
Note that there are several empty cells and there does seem to be some clustering – but this clustering is not from the existence of clusters in the population, since that population is a spatial uniform distribution. To conclude that the population has clusters in it, we would need more tightly formed clusters in the sample than we see here.

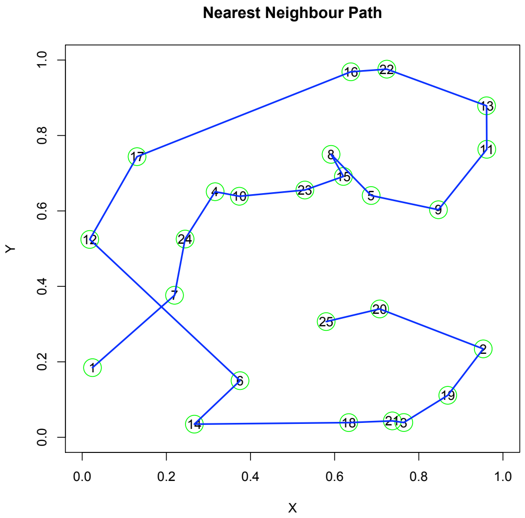
But how can we characterize the amount of clustering that we get when the population has no clusters? That is where the cell count distribution is useful. In the above we have 0,1,2,3 points per cell with frequencies 7,12,5,1, respectively. How does this compare with what the spatial uniform usually produces? We use the theory that says that the distribution 7,12,5,1 for 0,1,2,3, should be well approximated by the Poisson with mean 1 and sample size 25. The expected frequencies for this Poisson can be compared with what we got to judge the fit. This step is beyond this course, but we had a simple version of it in the assignment.

The important thing is to realize that the simulation of the spatial uniform gives us a benchmark with which to compare an actual spatial scatter, and from this comparison we can in principle decide whether the population is uniform, regular, or clustered. In the case of distribution of a certain plant like the Ocotillo, the nature of the clustering could be important for biological study of the plant.

2. Path Length

Another feature of the uniform spatial distribution is the length of the path needed to touch each point. The problem of how to minimize this is sometimes called the travelling salesman problem. It is a difficult problem that we will not solve. However, we can make some progress with the problem by using the nearest neighbour criterion – that does certainly do better than using the natural sequence of the points (the order in which they were generated). And further, we can improve on the nearest neighbour path by doing an ad hoc fix as was asked of you in your assignment.





The random path length is 13.5, the nearest neighbour one is 4.6, and we can imrove on the nearest-neighbour path by modifying the nearest path by 1-12-17-…4-24-7-6-14-….25. We substitute 1->12 for 1->7 which is a tiny bit longer, but then use 7->6 which is much shorter than 12->6. The distance would then be less than 4.5.

The point of the two spatial exercises is simply to give you an awareness of the nature of randomness in this context. As with other contexts we have studied, there are some surprising features caused by randomness.

The next topics I will cover are Time Series (including random walks, the stock market, smoothing, forecasting, and population growth) and Regression (including weather forecasting, data mining, covariates, residual plots, and correlation).

KLW 2010/04/06