STAT 100 **Notes** **Jan. 26, 2010**

Schedule for week 4

Assignment #4 – Sample Midterm
Recap of Sampling Theory for Means
Insurance – Spreading the Risk
Diversification of Investments
Data-Based Prediction: Vintage-Wine Prices

The variability of sample Means

The reason for introducing SDs and Normal distributions is to describe the variability of a sample mean (from one sample of size n to another sample of size n).

Here are five new samples of size 25 (from N(0,1)) summarized by mean & SD:

Sample	Mean	SD
1	0.19	0.92
2	0.18	1.1
3	0.04	1.0
4	0.08	0.8
5	-0.1	0.91

Suppose again that we do not know the mean =0 and SD=1 of the population that generated these samples. And, suppose after the first sample is observed, we want to guess the variability of the Mean in subsequent samples. Sounds impossible but it can be done pretty well.

From the formula we showed before

SD of sample mean = SD of population / $\sqrt{}$sample size

Now the "variability of the Mean in subsequent Samples" is measured by "SD of sample mean" so the formula should help. But the problem is we are supposing we do not know the SD of the population. However, the sample SD should be close to it so we use it instead. Then

SD of sample mean = (approx) SD of sample / $\sqrt{}$sample size
$0.92/\sqrt{25} = 0.92/5 = .184$

OK. How good an approx is it? Look at the other means {0.18,0.04,0.08,-0.1}. The SD of these 4 means is .116. My claim is that the .184 should be a reasonable estimate of the SD of new sample means, ands in our example here the target was

only .116. Seems like a bad estimate, until you realize that, if we round these to one decimal place, our estimate 0.2 is pretty close to the observed 0.1. So if we knew nothing about the population sampled in that first sample, we could still get a rough estimate of how much subsequent sample means would vary based on that first sample alone.

Do you see why there should be some surprise in this accuracy? I have used ONE sample to provide information about FOUR other samples. The information I extracted was the variability of the means of the FOUR samples, and I based it on a completely different sample. Of course, it was important that all five samples were random samples from the same population. That population happened to be normal but that was not really a requirement of the method.

Now why is the SD of the sample mean a useful thing to estimate? When a sample from a population is obtained, it is common to want to estimate the mean of the population. For example, if Statistics Canada wants to be able to provide information on household incomes by province, it might take a information from a random sample of households in each province, and provide a table of mean incomes by province. This would be easier to see the overall provincial differences this way than to look at 10 histograms. So means are useful for making group comparisons. Another typical example is to look at the average response to two drugs that are being used to treat an ailment. For example, to compare the pain relief from aspirin vs ibuprofin. The mean responses are one way to compare the overall effects. Here is another application in a bit more detail, to tie this means sampling theory with the idea of hypothesis testing.

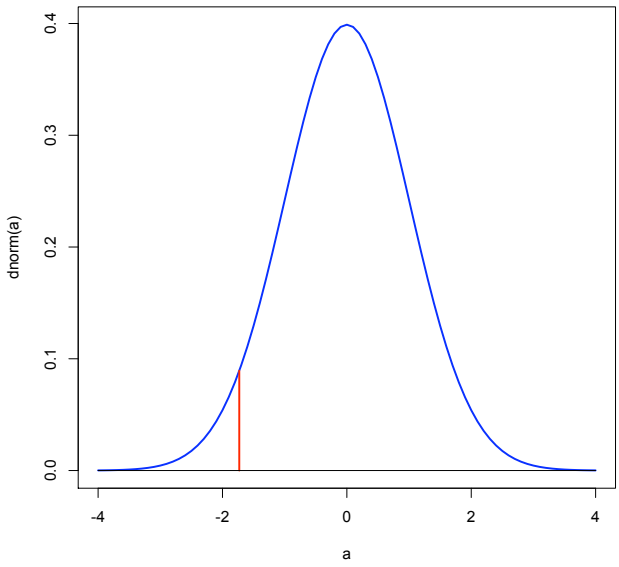
Applications of the Method

An example:

You own an auto-repair shop and you record your revenue each day. A large factor in your revenue is the hourly charge to customers for labour. You have been charging the same rate for several months but you would like to increase the rate so you can pay more to your deserving staff. But you are worried that a raise in the labour rate will cut into your revenues if your customers spread the word about your higher rate.
Prior to the rate rise your revenues averaged $2000 with a SD of $200 and was fairly stable over several months. After the hike your revenues averaged $1900 with a SD of $190 based on only twelve days of operation. Can you decide already that the rate hike is hurting revenues, or might the drop be simply random variation?

The key question is how well do the twelve post-hike revenues estimate the post-hike average that your shop will experience if the hike persists for a longer time? In other words, is the $1900 average the result of sample variation from a probability distribution with mean $2000 and SD $200?

We can compute an estimate of the variability of the mean of twelve random values from the N(2000, 200) distribution using our formula. SD of the average = approx $200/\sqrt{12} = 57.7$. That means the observed sample mean of $1900 is just under 2 SDs below the mean of $2000. Actually (2000-1900)/57.7 = 1.73 - in other words, 1.73 SDs below the mean. Let's see where this is on our Normal distribution:



So, $1900 is a pretty low value if the true mean of the distribution underlying it had mean $2000. But is it low enough to provide strong evidence that our assumption (mean=$2000) is untenable? While this is a personal decision, statisticians usually require a value in the 5 % tail before rejecting the assumption of the calculation. By looking carefully at the graph, or looking up the value via calculator or computer, the tail area to the left of the red line above is 0.042 – i.e. 4.2% of the probability is to the extreme of the observed value. So if we use the traditional 5% critical value, we would take this data as evidence that the rate hike has actually reduced revenues. The observed $1900 is far enough below $2000 to "surprise" us with a rare event, but on reconsideration we think the result would be less surprising if in fact the underlying mean of the distribution that generated the $1900 were actually less than the originally assumed $2000. Got it?

A recap of theory

The above discussion is an example of hypothesis testing. Remember the logic "If something happens that is unusual under ordinary circumstances, then this is evidence that the circumstances are not ordinary." The additional complication involving means, the SD, and the Normal distribution is very often involved in this "hypothesis testing" process.

Why the Normal distribution? Isn't it just a model that could be very wrong? A piece of theory that we will visit frequently in future explains that Normal distributions pop up naturally when we are examining means, no matter what the sampled distribution is. More on that later under the banner of "The Central Limit Theorem".

A rough table of the Normal Distribution:

x=SDs above the mean	Probability less than x	For Remembering (approx)
-4	0.00003	0
-3	0.0013	0
-2	0.023	2.3%
-1	0.16	16%
0	0.50	50%
+1	0.84	84%
+2	0.977	97.7%
+3	0.9987	100%
+4	0.99997	100%

Or, even more simply and approximately:

Within 1 SD of the mean 68%
Within 2 SDs of the mean 95%
Within 3 SDs of the mean 100%

Insurance: Spreading the Risk

Everyone needs to understand insurance: home insurance, auto insurance, travel medical insurance, life insurance (i.e. death insurance). The key to understanding how it works is the theory we just outlined: the variability of a sampling mean.

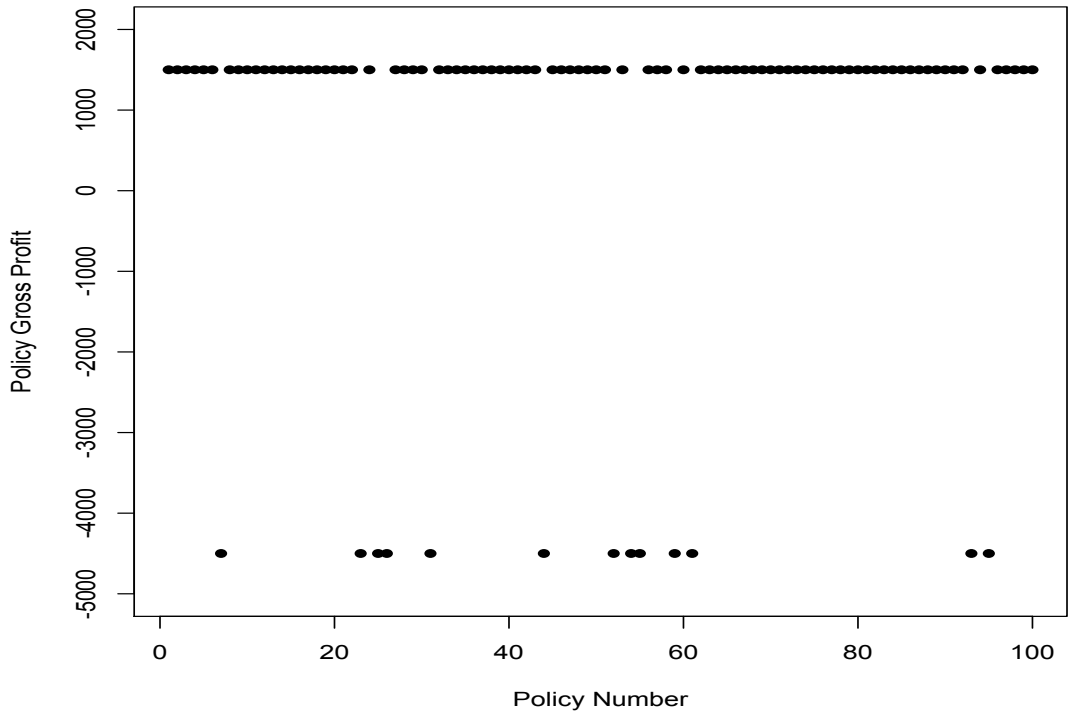
To simplify things, let's consider auto insurance, and suppose that we want a $0 deductible policy for one year. It might cost $1000 or more and if we don't have a claim in the year, that might seem to be $1000 down the drain! We are just enriching those big bad insurance companies. But lets look at it from the point of view of the insurance company, as an investor, say. Would we want to "underwrite" a policy for auto insurance for a 25 year old male for one year? It might cost us $1,000,000 if we are unlucky.

Clearly we need to underwrite many policies so that the revenue from the premiums will more than cover the claims that occur in the year. But how many claims do we need and how much should the premium be? This is a sophisticated business and the gathering of claims history and client characteristics is very complicated. But we can get the basic idea by considering a simplified scenario.

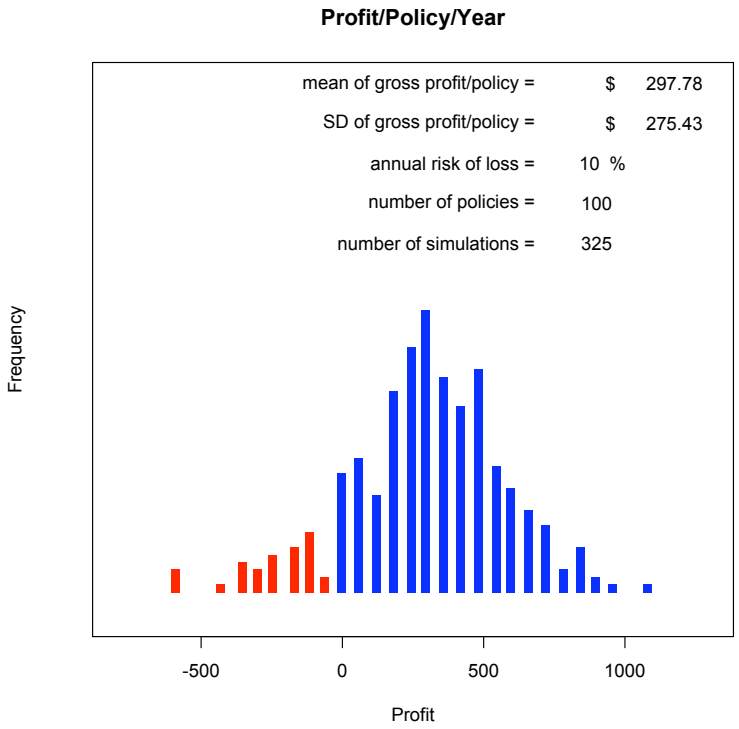
Auto Insurance Scenario:

Suppose that the insurance company experiences costs of $6000 on average for a claim, and that, in the class of clients considered (say, 25 year-old males in Vancouver), the chance of a claim during the year is 20%. Obviously, the premium needs to be at least $1200 if the insurance company will make a profit. Also, considering the costs of maintaining head office, billing, advertising, paying salespersons, legal costs, paying for investment advice, the premium would have to be much higher than $1200. Let's say the premium is set at $1500. One other crucial aspect of this scenario is the number of such policies we underwrite: for now let us suppose that in this category of client we only have 100 policies. What might the outcome be for the year's experience of these 100 policies?

This is the kind of question that a simulation in R can answer easily. Here is a sample output for one year's experience, showing gross profit for each policy.



But what we would really like to know, as an investor in this insurance company, what is the total gross profit likely to be, and how often will the company actually lose money? Here is a summary of 300 simulations of the 1-year experience:



Note that we expect the "gross" profit to be about $300 per policy, and in this simulation it is $297.78. Also note that, even with an average gross profit on each policy of $300, the probability of losing money on the year's experience is estimated to be about 10%. Of course, the situation is worse even than this implies since we have not allowed for business expenses yet (gross profit has to pay for all those other expenses mentioned above, like paying salespersons salaries).

The above simulation tells us that the gross profit for a group of 100 policies can be quite variable – in fact its SD is estimated to be $275.43 in the above simulation. But note that the gross profit in one year is actually an average of the result of 100 policy outcomes, and so our theory about the variability of averages should apply here. Even though the policy outcomes have just the two values -$4500 and +$1500, the annual gross profit only varies from about -$500 to + $1000.

Using the Theory of Sample Averages:

Was this simulation outcome predictable from our theory? We know the average gross profit is +$300 ($1500 premium less a cost of $6000 20% of the time). Also, if we had 20% at -$4500 and 80% at $1500, it can be computed that the SD of the policy outcomes over one year would be 2412.1. So that means the SD of the annual gross profit itself (an average) is $2412.1/\sqrt{100} = 241.21$. So, even before looking at the simulation, we would have guessed the mean gross profit was $300 and its SD was about $241.

Q: What happens if we were studying the same insurance context except that the number of policies was 500 instead of 100?

Q: Can a larger insurance company drive a smaller company out of business by reducing the premium?

The simulation of 300 years of this group of policies estimated that the mean of the annual gross income was $297.98 and the SD of the annual gross income was $275.43. So the calculation based on a guessed outcome (20% -4500 and 80% 1500), along with our theory of the variability of averages, gave a pretty good indication of the likely outcomes even before the simulation was done.

A preview of the Central Limit Theorem

Did you notice that the simulated distribution looked roughly bell-shaped, like the Normal distribution? This was no accident. The population we were sampling here was a population with 80% $1500 and 20% -$4500 – certainly not normal. And yet, the average over one hundred randomly selected values from this population has the bell shaped distribution. This is the Central Limit Theorem at work: it says, roughly, that **averages tend to have a Normal distribution, and this tendency increases as the sample size increases.** (Our sample size here was 100).

Diversification of Investments

"Don't put all your eggs in one basket". It does not take a statistician to point out the merit of spreading your investments around different investment types, different companies, different industries, even different countries. But quantifying the advantage to a diversified portfolio can reveal some surprising consequences of this widely known advice.

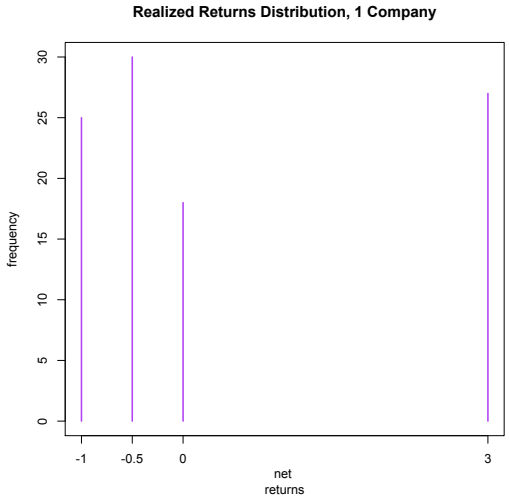
Why do some people invest in high-risk "penny" stocks that frequently become worthless? The answer is that in the unusual circumstance that the company with the penny stock survives its early years, its stock may appreciate to many times its early cost, and result in huge profits for the shareholder. So such stocks are a big gamble. Conservative investors stay away from these kinds of investments. But we will investigate a possibly profitable strategy that could be used with a portfolio of high-risk companies.

Suppose you are offered shares in a company that has the following 1-year prospects:

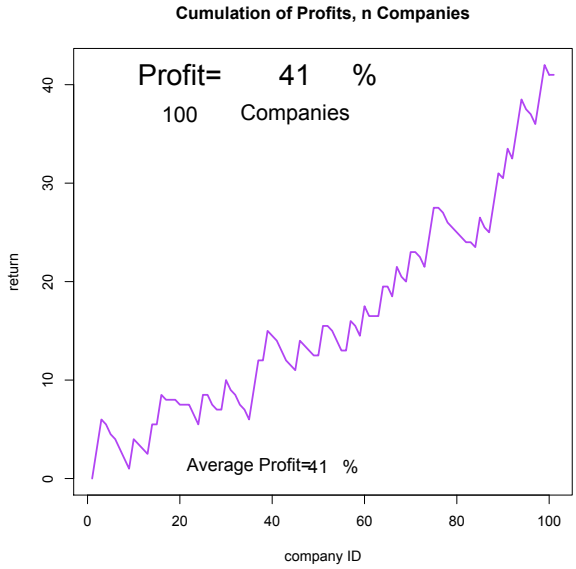
The shares are for sale for $1.00

At the end of one year, the stock will be equally likely to sell for $0, $0.50, $1.00, or $4.00. In other words, there is a 25% chance for your net income per share to be: -$1.00, -$0.50, $0.00, or $3.00. So there is a good chance (50%) that you will lose money, and a very good chance that you will not make any money (75%). There is small chance for a fairly good payoff of $3.00 (25%). So this would normally be thought of as a risky investment.

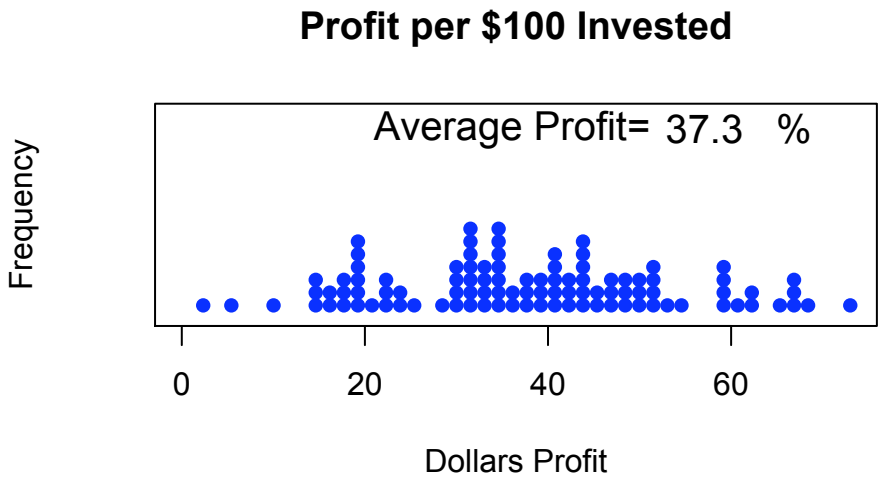
Let's look at a typical simulation of this stock: since the outcome is only a single number (-1,-1/2,0,3) it is more informative to look at 100 experiences to get an idea of what is likely.



The bars are fairly equal as one would expect. But lets look at what would happen to our net profit during a sequence of 100 such investments:



As we scan through the 100 investment outcomes, accumulating the results, we see a fairly systematic accumulation of profit: 41% more than our initial investment. But was this typical. The simulation is set up so we can do the same thing 100 times.



One does not always make 40% profit, but returns of 15%-50% are fairly typical. And all this based on an investment prospect that loses money half the time.

Is there a catch? Are all risky investments potential winners? Certainly not. We need two conditions:
1. The 100 companies (that we have each modeled in the same way) need to have *independent* outcomes. That is, whatever happens in one company is unrelated to what happens in the other companies.
2. The average return from the model company must be positive.

Re 1. The independence is hard to achieve in practice, but in forming a portfolio the best stability of returns is achieved by making the component investments as independent as possible.

Re 2. The average return in our example was 37.5 cents per dollar invested. To see this, note that (0 + .50 + 1.00 + 4.00)/4 = 5.50/4 = 1.375 so a $1 investment yields $1.375 on average for a positive return of 37.5%.

We can also compute the SD of company outcomes:
First compute the squared deviations $(0-1.375)^2+(.50-1.375)^2+(1.00-1.375)^2+(4.00-1.375)^2 = 9.7$. Then divide by n-1 = 3 and take the square root of the result.
$\sqrt{(9.7/3)} = 1.80$

The investment lesson survives these requirements: find companies with prospects that, on average, are positive, and choose companies in the portfolio for which the success is as independent as possible.

How does out theory of variability of averages predict these findings? Note that the portfolio is like a random sample from 100 companies each having prospects with mean $1.375 and SD $1.80 for each $1.00 investment. If we have 100 of these, the average return would be $1.375 and the SD of this average would be $1.80/√100 = $0.18. Now look back at the most recent graph. Did the simulation show that a typical deviation of a 100 company investment was about $0.18?

Here is another theory to check. The Central Limit theorem says that the distribution of the averages should have a Normal distribution. But we know that Normal distributions have 95% of the values within 2 SDs of the mean. Is it true that the distribution shown in the graph has 95% within $0.36 of $1.375? Count the dots between $1.015 and $1.735. Well in this particular simulation it looks like about 99% is in that range. The theory suggested 95%. Still a useful approximation since we could have computed it without doing the simulation at all.

How do we simulate the outcome of one of these companies? We need a mechanism that produces the outcomes 0.00, 0.50, 1.00, and 4.00 each with probability 0.25. But the tossing of two fair coins would do that:

HH -> 0.00
HT -> 0.50
TH -> 1.00
TT -> 4.00

It is much easier to have the computer do the simulation. But it is easier still to do the calculation using our "square root law" for the variability of means.
