Ch 3 – Sections 3,4,5 - Sampling Theiory

Ch 6-9 are about "Inference" – how to generalize from a sample to a population. Ch 5, and especially the sections we cover today are about the sampling theory that underlies all these inference methods. The sampling theory required the machinery of probabilities (Ch 2-Ch 4) as well as the tools of descriptive statistics (Ch 1). Descriptive Statistics and Probability Theory have uses independent of inference, but they are also essential for inference per se.

Sample variation – the variety of values we can get in
i) sample selection, or in
ii) measurement.

Note the population in i) is real, tangible, concrete, ...whereas in ii) it is hypothetical. The distinction is important in practice, but not for most of our course.

In both cases, we describe our sample data as a sequence of IID random variables

$X_1, X_2, ..., X_n$

and the sample values are usually denoted, for theoretical discussion,

$x_1, x_2, ..., x_n$

You should think of the X_i as the unknown outcome of an experiment, and the x_i as a particular value that was obtained in one instance.

Typically, we use the data $\{x_i: i=1,2,...,n\}$ to try to infer the common distribution of the X_i, or some feature of the distribution such as the value of a parameter (like μ or σ).

When we are trying to estimate a parameter value, we concentrate the information in the sample data by calculating a function of the n data value – this function is called a statistic, and sometimes the value of the statistic is also called a statistic (a bit confusing, admittedly).

Some notation to reduce confusion:

μ and σ always represent the population values of the parameters mean and standard deviation.

$\overline{X}$ or $\bar{x}$, and S or s, always represent the sample values of the mean and standard deviation. $\overline{X}$ and S are "statistics" in the sense of functions of the sample space, and $\bar{x}$ and s are statistics in the sense of values that $\overline{X}$ and S might take.

Note the definition of a random sample p 228. If you think of drawing tickets from a hat, with replacement, you will have the idea. The independence condition is important.

Consider the following example: population of 100,000 men, 100,000 women, draw a random sample of 100 women and 100 men. Is the sample of size 200 a random sample from the population of 200,000? Answer is NO, even though the chance is 1/1000 for every one of the 200,000 in the population to be selected into the sample. Note that samples of 99 mean and 101 women are impossible in this sampling scheme, even though they would be possible in a direct random sample from the 200,000.

Sampling Distributions

When a sample of size n is selected from a population, and a statistic is computed from that one sample (to estimate a population parameter say), the whole process has produced one number, the value of the statistic. Now think of doing this whole process over and over again so it produces many numbers. This collection of numbers would have a distribution, and this distribution is called the **sampling distribution of the statistic**.

Note: A sampling distribution is different from a sample distribution. A sample distribution is just the distribution from a single sample, wheras the sampling distribution of a statistic requires many samples to produce. Of course, in practice we do not taqke many samples of size n - but we still conceive of it so we can make probability statements about how variable our statistic might be (based on a single sample of size n).

The idea of the sampling distribution of a statistic is difficult but very important. Spend some time sorting it out.

Using Simulation to determine the Sampling Distribution of the mean:
Lets use our risky company payback distribution as our "population".

Payback($)	Probability	Net Profit($)
0.00	0.25	-1.00
0.50	0.25	-0.50
1.00	0.25	0.00
4.00	0.25	3.00

(Usually we do not know the population when we are trying to learn about it, but in this case we are studying the method of recovering some aspect of the population so starting with a known population makes sense.)

Recall that it was the mean of the distribution that was important – the average return was computed to be $0.38 per dollar invested, and now lets see how close we might come to estimate this based on a sample of say size 10.

Here are a few random samples of size 10 from this population:

```
risky.sample()
 [1] -0.5  3.0  0.0 -1.0 -0.5  0.0 -1.0 -0.5 -0.5  0.0
 [1] -1.0 -0.5 -1.0 -1.0 -0.5 -0.5  3.0 -1.0 -1.0 -1.0
```

```
 [1] -0.5 -0.5 -0.5 -1.0  0.0  3.0 -1.0 -1.0  3.0 -0.5
 [1] -0.5  3.0 -0.5 -1.0 -1.0  0.0 -1.0 -0.5  0.0 -0.5
 [1]  3.0 -1.0 -0.5 -1.0  0.0  0.0  0.0 -0.5  3.0  3.0
[1] "and the sample means are"
[1] -0.10 -0.45  0.10 -0.20  0.60
>
```

So if we were to use the sample mean to guess the population mean (which we know is 0.38), we could be quite a way off – even in these five instances, we might have been off by as much as .83 (with -0.45). Maybe we need a bigger sample. Try n=25 and lets see how consistent they are:

```
risky.sample(n=25)
 [1]  0.0  0.0 -1.0  3.0  0.0 -0.5  3.0  3.0  0.0  3.0 -0.5  0.0  0.0 -0.5 -1.0
[16]  0.0  0.0  3.0  0.0  3.0 -1.0  3.0  3.0 -0.5 -0.5
 [1] -0.5 -1.0  0.0  3.0 -1.0  3.0  3.0  3.0  3.0 -1.0 -0.5  0.0  0.0 -0.5 -0.5
[16]  3.0  0.0  0.0 -0.5  3.0  3.0 -1.0  0.0 -1.0 -1.0
 [1]  0.0  3.0 -1.0 -0.5  3.0 -0.5 -1.0  0.0 -0.5 -1.0  3.0  3.0  0.0 -1.0  0.0
[16]  0.0  0.0 -1.0  0.0  3.0  3.0  3.0 -1.0  0.0  0.0
 [1]  0.0  3.0 -1.0 -0.5 -1.0  0.0 -1.0  3.0  0.0 -1.0 -0.5  3.0  3.0 -0.5  3.0
[16] -0.5  0.0  3.0 -0.5 -0.5 -0.5  3.0 -0.5 -1.0  3.0
 [1] -1.0 -1.0 -1.0  0.0 -1.0  0.0 -1.0  0.0  0.0  0.0  3.0 -0.5 -1.0  3.0 -0.5
[16]  0.0 -0.5 -1.0  3.0 -1.0 -0.5 -1.0 -1.0  3.0  3.0
[1] "and the sample means are"
[1] 0.74 0.62 0.54 0.60 0.12
```

This is better – the worst of the five estimates this time is 0.74 and it is only .36 away from the true value.

Lets try n=100.

In this case the 5 averages are

0.435 0.525 0.370 0.200 0.225

and we can see that the worst one this time is only .18 away from the true value.

So we have

n	worst error of estimate in 5 samples
10	.83
25	.36
100	.18

Of course, these are just simulations, and if we redo the experiment, we would get something different:

like

n	worst error of estimate in 5 samples
10	.113
25	.34
100	.145

Nevertheless, it is clear that the larger samples give better estimates of the mean, since the errors are smaller for larger samples. Now the question is, can we anticipate, without having to do this simulation, how bad the error might be, and in other words, how could the estimate of the mean would be, based on a **single** sample of size n. The answer is yes. The surprising thing is that our one sample will provide an estimate of the mean, **and,** it will provide an estimate of how good that estimate of the mean is!

The standard deviation of the sample mean is estimated by the standard deviation of the sample divided by the square root of n.

$sd(\overline{X}) = sd(X)/\sqrt{n} = sd(x)/\sqrt{n}$ approximately.

For example, if our one sample of n=25 values is

```
> risky.sample(n=25,m=1)
 [1]  3.0  3.0  3.0 -0.5  0.0  0.0  0.0  0.0 -0.5 -1.0 -0.5 -1.0  3.0  3.0 -1.0
[16] -0.5  0.0 -0.5 -1.0  0.0  0.0  0.0 -0.5  3.0  3.0
[1] "and the sample means are ="
[1] 0.56
[1] "and the sample sds are ="
[1] 1.589811
```

Then our estimate of the population mean is the sample mean 0.56 and our estimate of the sd of this sample mean is 1.59/5 = .32.

So for samples of only n=25, with this population, we will have to live with typical errors in our estimate of about .32.

Here is a redo of this with n=100.

```
> risky.sample(n=100,m=1)
  [1]  3.0 -0.5 -0.5  3.0 -1.0 -0.5  0.0 -1.0  0.0 -1.0 -1.0 -1.0 -1.0  0.0  3.0
 [16]  3.0  3.0 -0.5  0.0 -0.5 -1.0  0.0 -1.0 -1.0 -1.0 -0.5  3.0  0.0  0.0 -0.5
 [31]  3.0  0.0 -0.5 -1.0 -1.0 -1.0 -0.5  0.0  0.0  3.0 -0.5 -0.5 -0.5 -0.5 -1.0
 [46]  3.0  3.0  0.0 -0.5  3.0  3.0  3.0  0.0  0.0  3.0 -0.5 -1.0  3.0  3.0  3.0
 [61] -1.0  0.0  3.0 -0.5 -0.5  3.0 -1.0  3.0 -0.5 -0.5 -1.0 -1.0 -0.5 -1.0 -0.5
 [76]  3.0  3.0  3.0 -1.0 -1.0  0.0  3.0  3.0 -1.0 -1.0  3.0 -1.0  0.0 -1.0  0.0
 [91]  3.0  0.0  0.0  0.0  0.0 -0.5  0.0  0.0  3.0 -1.0
[1] "and the sample means are ="
```

[1] 0.46
[1] "and the sample sds are ="
[1] 1.630951

This time our estimate of the population mean is 0.46 (better, remember the real one is 0.38) and we estimate that a typical error now is about 1.63/10 = .16. Obviously the n=100 estimate is better than the n=25 estimate, but the useful thing is we can see how much better and decide if it is good enough for practical purposes.

The theoretical result just used is given in the box on p 237 and proved on p 244. But it is very simple to verify the square root law directly. Assume X_1 and X_2 are IID with $V(X_i) = \sigma^2$.

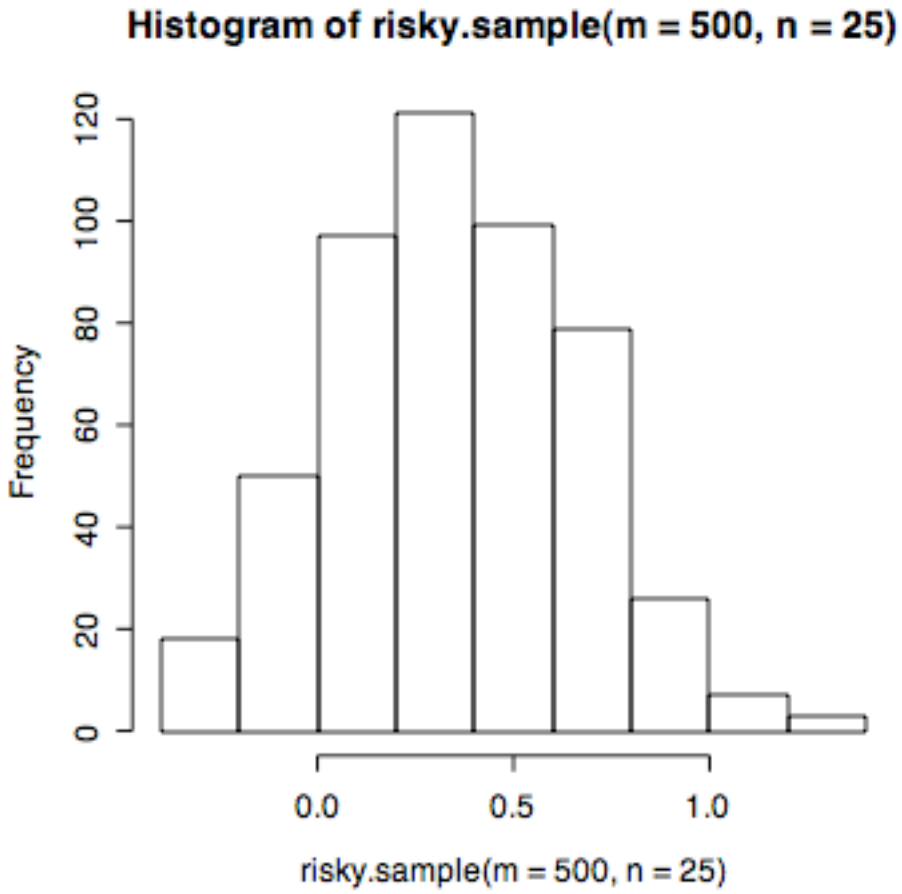
$V(X_1+X_2) = E(X_1+X_2)^2 - E^2(X_1+X_2) = E(X_1^2)-E^2(X_1) + E(X_2^2)-E^2(X_2)$
since independence -> 0 covariance. (p 223)
$= V(X_1) + V(X_2) = 2\sigma$

So $V(\bar{X}) = \sigma^2/2$ and $SD(\bar{X}) = \sigma/\sqrt{2}$

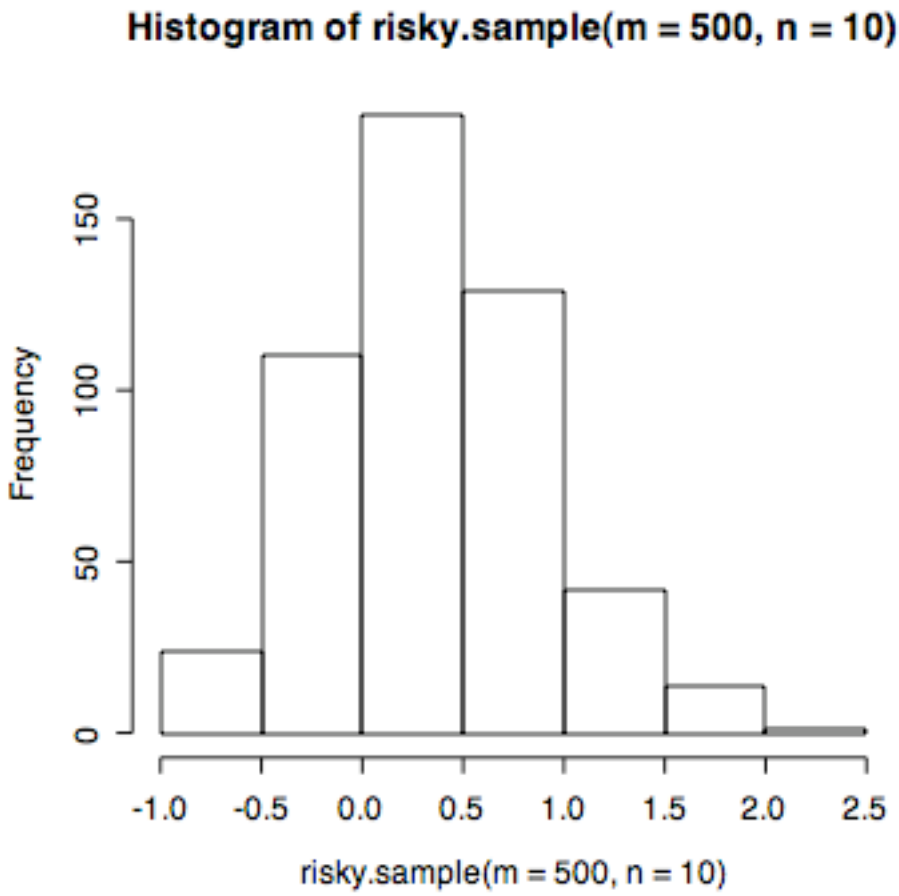
(qed for n=2, and this argument extends easily to n>2.)

We can see that the sample mean $\bar{X}$ has a distribution whose mean is the population mean and whose sd is the (popualtion SD)/ $\sqrt{n} = \sigma/\sqrt{n}$. But what can we say about the shape of the sampling distribution of $\bar{X}$? A very important theorem called the Central Limit Theorem (CLT) says it is approximately normal, with the approximation becoming increasingly precise as n gets larger. See precise statement p 239. Note that we do not need the CLT for the results about mean and SD, even though they are included in the theorem on p 239. It is the normality that is the main result.

The population we were working with above – the risky company – had a distribution of returns that was very non-normal: if X=Profit, P(X=-1)=.25 = P(X=-0.5)=P(X=0)=P(X=3). But lets look again at the mean of a sample of size 25 – we will simulate doing this 500 times and take a look at the distribution of the sample mean.

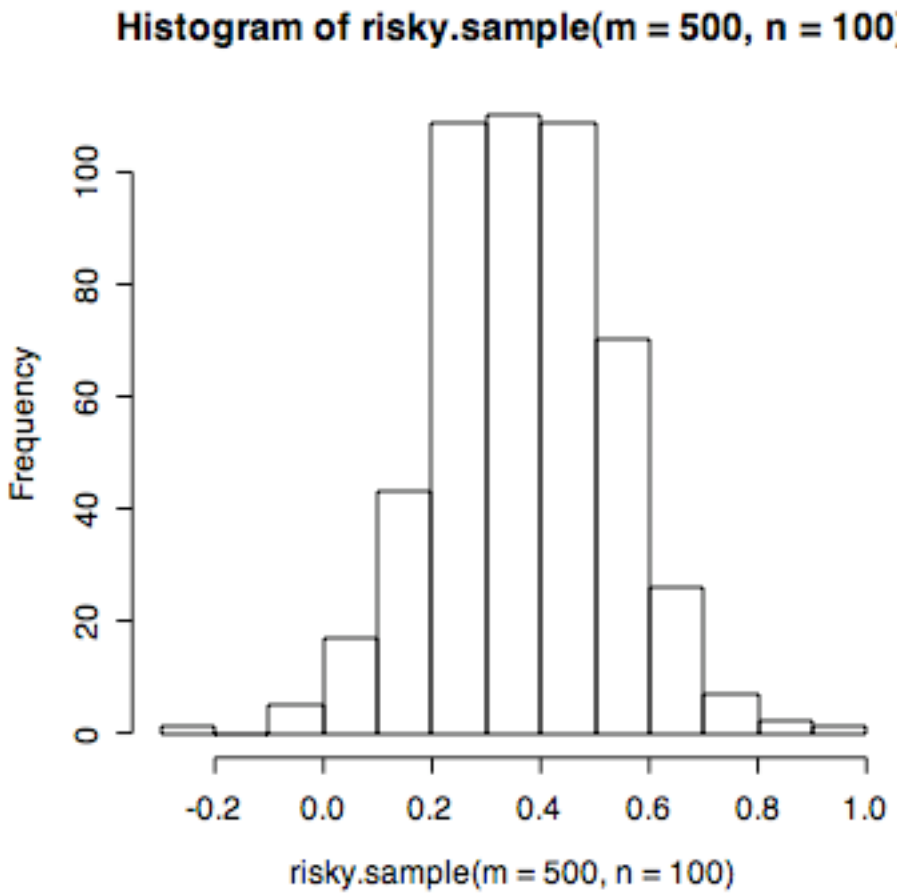


This certainly is much more symmetrical than the original profit distribution and it is reasonable to suppose it is approximately normal. There is mild evidence of a bit of right skew. Lets look at the n=10 version.



This one (n=10) is quite skewed and the normal approximation is bound to be pretty bad.

Now lets look at n=100.



The symmetry is much better now than for n=10 and even better than n=25.

The theorem says that as n grows the normal approximation gets better and better.

Note is n that must grow, not m. m is just the number of times we compute a mean and sample and it is large just to show the shape of the resulting distribution. m was 500 when n=10 gave that skewed result. A million sample means based on samples of size 3 will have a very skewed frequency distribution.

The sample mean is a particular linear combination of IID RVs. There are similar results for any linear combination of IID RVs. See p 244. In fact the theorem on p 244 does not require the variables to be identically distributed, although the most useful cases will assume this.

Now lets use the CLT to find the chance of losing money with 25 risky companies like the ones we simulated – we are still assuming these company outcomes are independent of each other (an ideal that is hard to achieve in practice).

Recall our "population" distribution

Probability(P(X=x)	Net Profit($) x
0.25	-1.00
0.25	-0.50
0.25	0.00
0.25	3.00

Now if we sample 25 outcomes (25 companies), our average gain in the sample should be about $0.38 and the sample SD should be about $1.56. These numbers 0.38 and 1.56 were not simulated, they were calculated from the above table. We can also calculate that the SD of the sample mean should be $1.56/\sqrt{25}$ = $0.31. In other words, the sample mean should have a mean of $0.38, an SD of $0.31 and the shape of the distribution should be approximately normal as a result of the CLT. That is useful info – we can compute, for example, P(profit >0) = P($\overline{X}$>0) = P(($\overline{X}$-.38)/.31>-.38/.31) = P(Z>-1.23) =1-,1093 = .89 from Table 3 in the text.

In other words, we can compute the probability that we make money if we have 25 of these risky companies and they operate independently. We have an 89% chance to profit from the 25 companies even though there is only a 25% chance that any single company will profit. If you refer back to the simulation recorded in the notes of Feb 14, where we simulated 100 instances of this 25 company portfolio, there were 87 that made money. So the theory worked pretty well. (We experienced 87% when the chance calculated was 89%).

So we can do something useful when we know the population – we can anticipate what the simulation would produce in this example. But note that ALL we used in the calculation was the mean and SD of the population. The actual formula for the population probabilities did not enter into the calculation except to give us the mean and SD. **This may suggest another use for this theory**: use the sample to estimate the mean and SD, and proceed as above assuming these mean and SD estimates are equal to the population ones. Our answer will be approximate, but still useful. If we had data on the profitability of 25 companies, but no knowledge of the probability distribution producing those profits, we could still do the above calculation, at least approximately.

It is this strategy of using data to test hypotheses about an unknown population from the the data is a random sample, that Ch 6-9 is about – this in "inference" of sample to population.

PS: One technical detail that I did not cover in the above is that if the population distribution is normal, then so are all possible linear combinations, and in particular the sample mean has a normal distribution. So a normal population gives a normal sampling distribution of the sample mean, and this is exact, not approximate. Any other distribution the normality of the sample mean is approximate, with the approximation improving as the sample size increases.

A bit of preview of Ch 6:

When we know (or assume) the class of models for the population we are sampling, like assume normal, or assume gamma, or assume Poisson, we need to use the functional form of the model to relate the data to the parameters. Often this reduces to using the sample mean and variance to estimate the population mean and variance, while the latter are known functions of the parameters. So this gives a way of estimating the parameters from the data. But this way does not always work. Ch 6 talks about estimation of parameters and in particular how to make use of the class of parametric models applicable to the situation (if known or assumed to be known).