14_15_CLT.pdf	Michael Hallstone, Ph.D. hallston@hawaii.edu

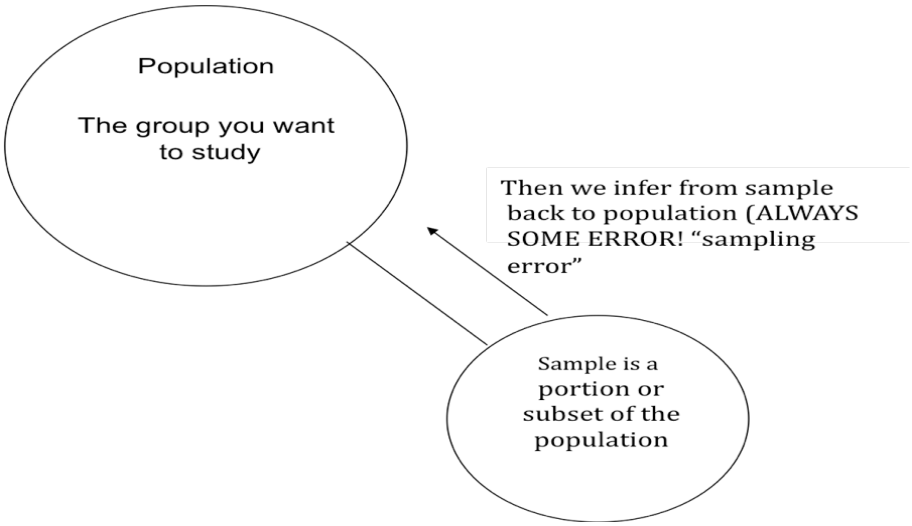
Lecture 14: Sampling Distribution of Means & Lecture 15 Central Limits Theorem

Introduction

This is probably the most important lecture of an introductory statistics course! We will learn the theory that provides the basis of much of inferential statistics.

Quick review of sampling (see diagram below)

Remember we want to study a group (typically made up of people) called a population, but due to issues of time, money, and feasibility we cannot possibly interview or speak to everyone in the population so we study a portion of the people in the population called a sample. We study the sample and hope that it is representative of the population. We infer from the sample to the population. This is called "inferential statistics." In this lecture we will learn the theoretical basis for inferential statistics. To put it crudely, the Central Limits Theorem tells us "why" or "how" we can use samples to infer to populations



Very Briefly: what are random and non-random samples

This is not a methods course so we will be brief…but be aware that sampling is a whole skill and "art" to itself. They have whole graduate level courses on the subject! When you taste a pot of sauce while you are cooking it, you are taking a sample! Hopefully that taste is representative of the way the rest of that pot is going to taste right?

Inferential statistics requires a REPRESENTATIVE random sample.

Although I will use the term "random sample" for the rest of this lecture for simplicity's sake, technically speaking inferential statistics requires a representative random sample. In order to "do" inferential statistics we need a sample that is not only random, but also representative of the population. It is possible to have a sample that is random, but not representative of the population. More on that in a second.

So when you read "random sample" for the rest of this lecture, what I really mean is "representative random sample."

simple random sample = the most basic kind of "random sample"

When we use inferential statistics we MUST have a random sample. The Central Limits Theorem (CLT) only "works" if we have a random sample. To visualize what a random sample is think of the most basic kind of random sample – called a "simple random sample." Bingo or keno use simple random samples. In bingo all the balls are placed in that rotating cage thing and they are drawn randomly. Thus all bingo balls have an equal chance of being chosen at the start of the bingo game. Well imagine bingo balls were people in a population. In a simple random sample all persons in population have equal chance of being included in sample. If you put everyone's name from the class and drew names from the hat, that could be considered a simple random sample.

Again, the central limits theorem, REQUIRES a random sample. The most basic assumption of inferential statistics is that the data comes from a random sample. If you do not have a random sample the probability rules of inferential statistics do not work. It is as simple as that. So, all *inferential statistics* are based upon the idea of a random sample.

Random but not representative sample?

Since you are a UHWO student, and smarter than the average college student, you are wondering how a sample could be randomly drawn, but not representative of the population.

Imagine you had a simple random sample of all the approximately 1 million people who live on Oahu. That means every person who lives on Oahu would have an equal chance of being selected into your sample. But since the vast majority of people live in the urban core of Honolulu (see map below) then those folks would be over represented in a simple random sample of Oahu. Those folks who live in the less populated areas would be under represented in the sample. So the sample would not represent all of Oahu accurately.

non-random samples

There are many kinds of non-random samples. Most student projects use some sort of a non-random sample. One kind of non-random sample is when the researcher purposively selects people to be in their study. The people are not selected randomly. If you have ever seen a student collecting data on campus at UHWO their sample is NOT random.

We are doing inferential statistics in this course!!! So, what does this say about the data for any project that was collected using a non-random sample? It says that technically speaking the rules of statistics do not apply or are not valid. But it is okay, because as students we are just learning how to "do" statistics.

***On the take home exams I ask "*Mention the assumptions you have violated (if any)" and all of you should write something like, "I violated the assumption of a random sample."**

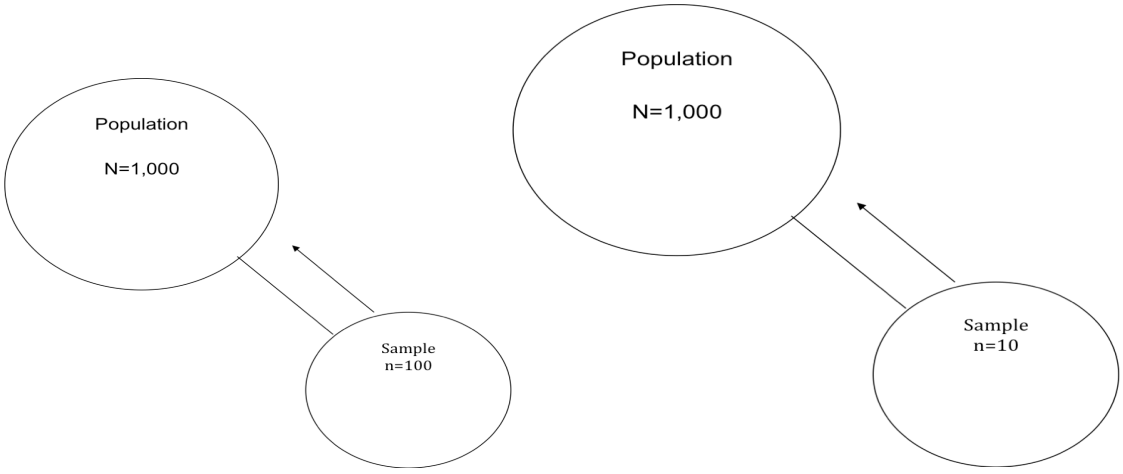
Our "purpose in life" in statistics

We want to be able to "say something" about the population that we are studying right? We want to infer from the sample to the population right? Well how can we collect some data that will be representative of that population?

Bigger is better with samples – and sample means tend towards population means

My first job in higher education was being the TA for a wonderful professor from Maui Community College by the name of Dr. Lynn Yankowski. He was on a sabbatical and teaching statistics at UH Manoa where I was a graduate student. He uses a wonderful example where all students can see that it intuitively makes sense that it would be better to take large samples. You don't have to be a statistician or mathematician or a rocket scientist to "get it."

If you had a population of 1000 people (such as UHWO students), would you rather have a sample of 10 people or a sample of 100 people upon which to draw your conclusions? Think about it. If you had to say something about 1,000 people would you rather base that on interviews with 10 people from that population or 100 people from that population? Choose one.



If you chose 100, you understand the basic concept of sampling error.

The following example is from another professor at UH Manoa for whom I was also a TA. Le Lei's example shows how that it might be even better to take many samples. Best yet, take many large samples and then take the mean off all of those! Does this not make intuitive sense that if we did this last thing the closer we might be able to estimate the real population mean? Remember – that is "our purpose in life" in statistics!!!! See the data on the next page and then the explanation that follows.

Bigger samples are better and sample means "tend" towards population means

Pretend this is a population of 30 students with a population mean age of 22.01 years. We have 3 random samples of size 10 (n=10) and three random samples of size 15 (n=15).

I.D. No.	Age										
1	21.53			Random	Sample	Of 10	Students				
2	21.23			Sample 1		Sample 2		Sample 3		Sample 4	
3	23.66		Size	I.D. No.	Age	I.D. No.	Age	I.D. No.	Age	I.D. No.	Age
4	19.38		1	2	21.23	3	23.88	6	23.24	6	23.24
5	23.56		2	4	19.38	4	19.38	7	21.87	10	18.57
6	23.24		3	6	23.24	6	23.24	8	19.57	11	23.81
7	21.87		4	10	18.57	7	21.87	13	21.83	12	18.82
8	19.57		5	16	23.13	8	19.57	14	18.45	13	21.83
9	24.89		6	18	20.75	12	18.82	15	21.03	17	25.4
10	18.57		7	19	21.63	13	21.83	16	23.13	20	21.91
11	23.81		8	20	21.91	17	25.4	19	21.63	21	23.42
12	18.82		9	24	20.5	19	21.63	24	20.5	24	20.5
13	21.83		10	25	23.79	21	23.42	27	24.48	28	23.16
14	18.45		Mean		**21.41**		**21.9**		**21.57**		**22.07**
15	21.03										
16	23.13			Mean 1-4	**21.74**						
17	25.4										
18	20.75			Random	Sample	of 15	Students				
19	21.63			Sample 5		Sample 6		Sample 7		Sample 8	
20	21.91		Number	I.D. No.	Age	I.D. No.	Age	I.D. No.	Age	I.D. No.	Age
21	23.42		1	3	23.66	2	21.23	3	23.66	5	23.56
22	20.46		2	6	23.24	4	19.38	5	23.56	6	23.24
23	21.39		3	8	19.57	5	23.56	6	23.24	8	19.57
24	20.5		4	10	18.57	6	23.24	7	21.87	9	24.89
25	23.79		5	11	23.81	7	21.87	13	21.83	11	23.81
26	25.52		6	13	21.83	8	19.57	14	18.45	12	18.82
27	24.48		7	15	21.03	10	18.57	15	21.03	13	21.83
28	23.16		8	17	25.4	15	21.03	16	23.13	16	23.13
29	19.87		9	18	20.75	16	23.13	17	25.4	17	25.4
30	23.45		10	20	21.91	20	21.91	21	23.42	18	20.75
Mean	**22.01**		11	23	21.39	22	20.46	22	20.46	19	21.63
			12	24	20.5	24	20.5	23	21.39	21	23.42
			13	25	23.79	25	23.79	24	20.5	22	20.46
			14	26	25.52	26	25.52	25	23.79	23	21.39
			15	28	23.16	27	24.48	28	23.16	24	20.5
			Mean Age		**22.28**		**21.88**		**22.33**		**22.16**
				Mean 5-8	**22.16**						

Above Example Shows Bigger is Better and Sample Means "Tend Towards" Population Means.

Here we can figure out the population mean, because it is a small population. But taking samples of 10 and 15, it is unlikely that any of your samples will be EXACTLY equal to the population mean. Sample means should tend to be close to the population mean though…..we see that in the above example. Sample means tend toward the population mean and larger samples tend to be even closer to the population mean!

However we need to make an estimate of the population mean based upon a single sample! Our sample mean is just one of literally thousands of possible sample means out there! Will it be close to the population mean?

For a population size of 15 and a sample size of 6 there are 5,005 possible different samples that could be taken. Thus 5,005 sample means. STOP AND MAKE SURE THIS CONCEPT IS UNDERSTOOD.

total # of possible combinations= $= \frac{N!}{n!(N-n!)} = \frac{15!}{(6!)(9!)} = 5{,}005$

So there are 5,005 possible sample means out there and we don't know which one we are going to get!

Above is an example with N=30 and we took samples of n=10. Can you imagine how many combinations of possible sample means there are in that situation? It is way more than 5,000! It is mind-boggling. It is more than 30 million!

How did I figure out it was 30 million? (feel free to skip this section!)

Actually for you inquiring minds:

total # of possible combinations= $= \frac{N!}{n!(N-n!)} = \frac{30!}{(10!)(20!)} = 30{,}045{,}015$

So for a N=30 and n=10 there are 30 million possible sample means!

If N=100 and n=50? 10,890,000,000,000,000,000,000,000,000 possible sample means (1.0089×10^{29}) That is 10,890 septillions or quadrillions! A quadrillion is 1,000 trillion dollar bills! So if that were dollars and there was such thing a bill worth 1,000 trillion dollars you would have 10,890 of those bills!

A note from Mike to Mike for in person class (Online students can ignore this):

You can tell the story of how in the real world they can estimate the population of US Voters with a n of about 2,500 or so. You can also tell the story of the Presidential election in the 1930's or so when they had a representative sample of telephones and a poll that predicted the Republican candidate would win. Problem of "rich" people owing phones.

Back to the plot….(resume reading here!)

Suffice it to say that in most cases in the real world there are "bazillions" of possible sample means out there and we only get to take one sample! Yikes! There are literally trillions upon trillions of possible sample means and if you do a study you will get only one sample mean.

Out of all the trillions of possible sample means you will have to estimate the population mean based upon one single sample mean. Will your sample mean be a good estimate of the population mean or will it be a poor estimate?

A good workable definition of the sampling distribution of means

To butcher the book's definition, all of the possible sample means from a population is called the "Sampling Distribution of Means." Think of it as a distribution of all the possible sample means that are possible give your population size and your sample size. [So recall if N=15 and n=6 there would be 5,005 possible sample means – that list of 5,005 sample means would be the "sampling distribution of means."]

Let's go back to the problem of research. We will only take one sample and use its mean. We will get to use one mean out of billions, trillions, (or even more)! Will we get one close to the population mean or will we get one "far away" from the population mean? Wouldn't it be great if we had some way to find that out? Well if the sampling distribution of means follows a "normal distribution" then we could apply some of those probability concepts we just learned about normal distributions. (The z-score exercises were all about making probability statements using a normal distribution.)

Sampling Distribution of Means

Pretend you were in a classroom and you could arrange the chairs in the room to be in the form of a normal distribution. Pretend each seat represents an individual sample mean.

Now pretend there were 5,005 seats in that classroom. This would be the number of possible random samples possible if the population N=15 and the sample size n=6. So if you had a population of 15 people and took random samples from it with 6 people in them (and took the mean of each sample), there would be 5,005 possible combinations of sample means. In a sampling situation of that size there are 5,005 possible sample means you could get. So each of the 5,005 seats in the classroom represent a possible sample mean you would get if you had a population of N=15 and took a random sample n=6.

What if we took all the possible samples of a certain size (n) of a known population got means from each of those? (For a finite population of size N there is a finite number of combinations of samples of a certain size n.)

This nauseatingly long process is called the **sampling distribution of means:** "is the distribution of arithmetic means of all the possible random samples of size n that could be selected from a given population" (from text book). So in the N=15 and n=6 example above, there are 5,005 sample means in the sampling distribution of means.

Exercise for in person class: Select a population of 15 students then take a few random samples of n=6. Write down age and sample means. To exhaust all possible combinations of a sample size of n=6 from a N=15, there would be 5,005 sample means! So if there were 5,005 seats in my classroom then we could pretend each of the seat represented an individual sample mean in the sampling distribution of means.

Grand mean

Using that 5,005 example, pretend you could add up all of the 5,005 means in the sampling distribution of means and then divided that number by 5,005. You would have the "mean of the sampling distribution of means." The mean of all of the means in the sampling distribution of means = "grand mean."

It just so happens that the mean of the sampling distribution of means (grand mean) is equal to the population mean!!!! The book proofs this for you and they have done simulated trials with computers to prove this too.

"Grand mean is = to population mean" (This is a leap of faith unless you want to do a whole bunch of proofs like in book.)

STOP. Class Exercise:

Write down and define: 1) sampling distribution of means 2) grand mean 3) grand mean = to what?

Since sample means tend towards pop. means, and since sampling distribution of means has frequencies of occurrence it is essentially a probability distribution. We could place these means on a frequency distribution. Draw it out. Example of 5,005 from above. Place all 5,005 sample means on a frequency distribution! Make that frequency distribution in the shape of a normal curve.

If the sample size is sufficiently large (n>30), ***then the sampling distribution of means resembles a normal distribution regardless of whether the population itself is normal.*** *What? Yup.*

Say that again? ***Well, if we take a sufficiently large sample (n>30), then we can say the sampling distribution of means is normal.*** What? Yup.

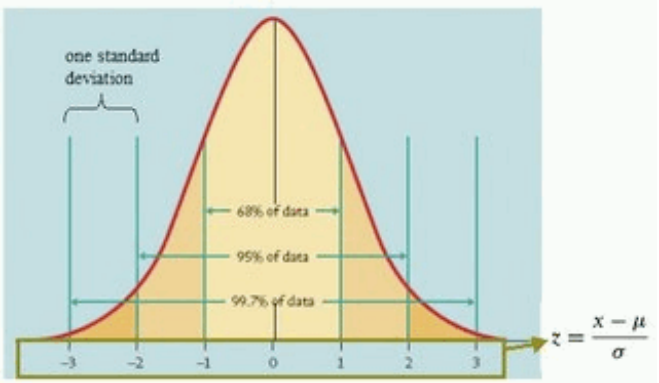
This is great! If we satisfy one of these conditions then we know that the sampling distribution of means resembles a normal distribution and we can apply all that probability stuff.

First if the distribution is normal and the grand mean = population mean then the population mean goes in the middle! Draw out a "bell shaped curve" and put a mean right in the middle. The grand mean and the population mean go right in the middle.

That means we have NO idea which sample mean we got from the sampling distribution of means but we can use the properties of the normal curve to make some probability statements about how "close" our sample mean is to the population mean! Remember that we used a standard normal distribution to determine probabilities. We used standard deviations to tell us probabilities didn't we?

Then we can use the properties of the standard normal curve to tell us things.

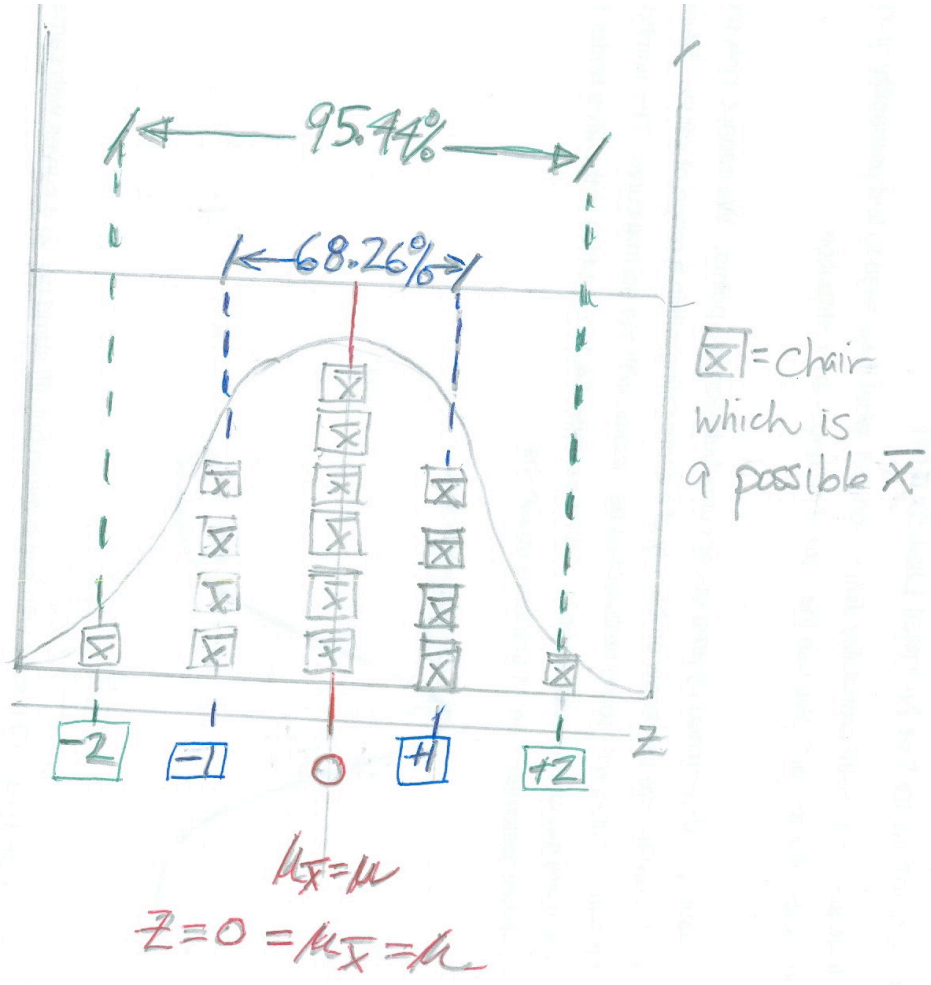
For example, we could say the probability of a single sample mean falling ±1 SD from the "true" or "real" population mean =68.26%. Or looking at our z table we could say that the probability of a sample mean falling ± 2 SD from the mean = .4772 X2 = 95.44%. Remember the pictures from z-scores?



source: http://www.sci.sdsu.edu/class/psychology/psy271/Weeks/psy271week06.htm

So pretend there were 5,005 little chairs in a classroom arranged in the bell curve shape of the picture above. Each of the chairs represent a possible sample mean you could "draw" if you took a sample (n=6) from a population (N=15). Which chair will you get? Will it be sitting close to the middle (population mean) or will it be far away from the middle (far from the population mean)? Well you are going to infer from your sample mean to your population mean – you are going to say that your sample mean is "representative" of your population mean.

Below is a hand drawn picture where each of the sample means is a chair in the classroom where the seats are arranged in the shape of a normal distribution. We could say the probability of a single sample mean falling ±1 SD from the "true" or "real" population mean =68.26%. Or looking at our z table we could say that the probability of a sample mean falling ± 2 SD from the mean = .4772 X2 = 95.44%.



Lecture 15: Central Limits Theorem

The Central Limits Theorem (CLT) summarizes formally what we have been talking about regarding sampling distribution of means, standard error, etc.

The CLT gives us a way out of the problem all researchers face when they want to estimate the population mean. Remember our "gig in life" is to be able to say something about the average of a population. But since we can't study the whole population, we have to study a sample of the population, take the mean of that sample ("sample mean") and hope that it is "representative" of the population mean. Well, the problem is there is a whole big long list of possible sample means that we could get when we take a random sample from the population. That big long list of possible sample means is called the "sampling distribution of means." We only get to take one sample mean at random from that long list of possible sample means! That is the problem faced by all researchers! So now what? What do we do? Well, the CLT gives us a way to deal with that problem using probability statements.

So if we could say that list ("the sampling distribution of means") was normally distributed, then at least we could make some probability statements about how different our sample mean is likely to be from the population mean. For example, if the sampling distribution of means is normally distributed, then we could at least say, “I don't know which sample mean I'm going to get from that long list of possible sample means, but at least I can say that there is a 68.26 % chance that that sample mean I'm going to get will be no different from the true population mean than plus or minus 1 standard deviations." Or "I don't know which sample mean I'm going to get from that long list of possible sample means, but at least I can say that there is a 95.44 % chance that that sample mean I'm going to get will be no different from the true population mean than plus or minus 2 standard deviations." The CLT allows us to make these kinds of statements.

C.L.T. DEFINITION

- The mean of the sampling distribution of means (the "grand mean") is equal to the population mean.
- We have a way of figuring out a measure of the standard deviation of the sampling distribution of means (the "standard error of the mean"). In fact, there are four formulas for the standard deviation of the sampling distribution of means -- two for finite populations and two for infinite populations. They are in the book.

- If the sample size is sufficiently large (n>30) the sampling distribution of means approximates the normal probability distribution – this means we can use the probability properties of the standard normal curve! **Remember:** *When the sample size is over 30 then the sampling distribution of means will approximate a normal distribution (see below).*

- There is another way the sampling distribution of means will be a normal distribution. If the population is normally distributed, the sampling distribution of means is normal regardless of sample size.

The book provides proofs and example of this and we won't go into them here.

The relationship between sample size and standard error

How do we figure out the SD of the sampling distribution of means?

The standard deviation of sampling distribution of means is called **standard error of the mean** there are few formulas for it (see below).

Standard error is a measure of dispersion of sample means about the population mean. As dispersion decreases values become clustered around the mean. That increases the possibility that sample mean will be closer to population mean. Look at formula.

$$\sigma\bar{x} = \frac{\sigma}{\sqrt{n}}$$

$\sigma = pop.sd$ n= sample size

(In practice we do not know the population SD, because then we would know the population mean. Thus we typically substitute sample standard deviation for the population standard deviation.)

As n increases standard error decreases!

Let's plug some numbers into the formula and see what happens to the number *when the sample size (n) increases but the standard deviation remains constant.*

If σ=1 n= 4 then: $\sigma\bar{x} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{4}} = \frac{1}{2} = .5$

If σ=1 n= 16 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{16}} = \frac{1}{4} = .25$

If σ=1 n= 25 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{25}} = \frac{1}{5} = .2$

If σ=1 n= 64 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{64}} = \frac{1}{8} = .125$

If σ=1 n= 100 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{100}} = \frac{1}{10} = .1$

See how the $\sigma_{\bar{x}}$ gets smaller and n increases?

What happens to any fraction as the bottom number gets bigger? 1/1 ¼ ½ 1/5 1/8 1/10 1/100 *The number gets smaller!*

> If you had a population of 1000 people (such as UHWO students), would you rather have a sample of 10 people or a sample of 100 people upon which to draw your conclusions? Think about it. If you had to say something about 1,000 people would you rather base that on interviews with 10 people from that population or 100 people from that population? Choose one.

This is the mathematical basis for your gut feeling from the example above:

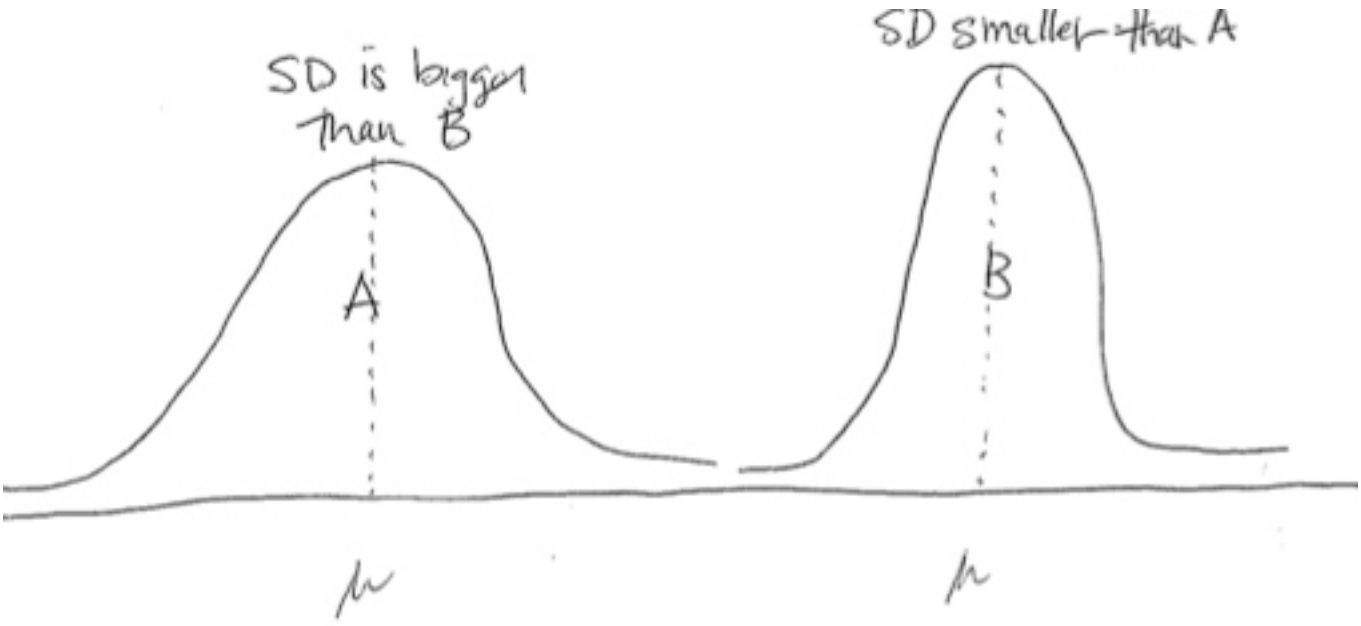
All of you intuitively knew that the sample of 100 would have less "error." Well you were right:

If σ=1 n= 10 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{10}} = \frac{1}{3.16} = .316$

If σ=1 n= 100 then: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = = \frac{1}{\sqrt{100}} = \frac{1}{10} = .1$

See page 218 in the sixth edition for examples if this does not make sense to you.

Therefore as sample size gets bigger, the greater the chances that a sample mean is going to accurately approximate the "true population mean. This is because as n grows larger the "standard deviation of the sampling distribution of means" gets smaller. As the standard deviation gets smaller all of the "chairs in the classroom get "squished closer together." Or as the sample size increases, the SD of the sampling distribution of means gets smaller and all the individual sample means get "squished closer together." As the SD gets smaller, the normal curve gets more "squished together" -- see my hand drawn picture:



Formulas for the standard error of the mean

These are 'real world' formulas and all assume that we do not know the population standard deviation!

Infinite population

$$\hat{\sigma}\bar{x} = \frac{s}{\sqrt{n}}$$

Finite population

$$\hat{\sigma}\bar{x} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

N = population size n= sample size.

As the size of the population increases to very large population, this finite factor becomes closer and closer to one and thus less and less important.

N=1000 and n=100

$$\sqrt{\frac{1000-100}{1000-1}} = \sqrt{\frac{900}{999}} = .949$$

N=200,000 (approximate size of US) n=2,000 (this example taken from sixth edition of *Statistics: A First Course* page 217)

$$\sqrt{\frac{200{,}000{,}000-2{,}000}{200{,}000{,}000-1}} = .9999999999$$

correction factor = .99999

As one textbook states, "…if the size of a finite population involved is large compared to the sample size, using the finite population correction factor has a minimal effect" (Sanders and Smidt: 217). You see the math that proves this quote above.

You can see how .99999999 is essentially one. It is not going to "correct" by much. The 0.94 above would correct a bit more. So multiply any number by .9999999 and then that same number by 0.94 and you will see the number changes less in the .999999 instance.

1 x 0.94= 0.94 or 1 x 0.999999 = 0.999999

So the .94 correction factor has a larger effect.

Practice:

There are not really any problems to compute for this lecture. However, you need to be familiar with the Central Limits Theorem. You need to know enough to understand the following:

- What is the sampling distribution of means?
- The mean of the sampling distribution of means (the "grand mean") is equal to the population mean.

- We have a way of figuring out a measure of the standard deviation of the sampling distribution of means (the "standard error of the mean"). In fact, there are four formulas for the standard deviation of the sampling distribution of means -- two for finite populations and two for infinite populations.
    - Know that as n increases the standard error of the mean decreases or gets smaller. Also know the implications of this dynamic in a sampling situation (what happens to the “standard deviation” of the sampling distribution of means (the standard error) as the sample size (n) increases? What does that imply? (See the picture of the two curves with different size SD’s above).
- If the sample size is sufficiently large (n>30) the sampling distribution of means approximates the normal probability distribution – this means we can use the probability properties of the standard normal curve! **Remember:** *When the sample size is over 30 then the sampling distribution of means will approximate a normal distribution (see below).*
- There is another way the sampling distribution of means will be a normal distribution. If the population is normally distributed, the sampling distribution of means is normal regardless of sample size.