


1

UNIT 3B

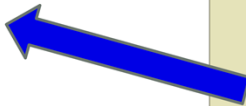
Sampling Distributions




Statistics we calculate from samples vary

**SO ...**

What does THEIR distribution look like?





We are about to embark on the most important theoretical topic so far. The concepts we will discuss here are abstract and many students find this material challenging to fully understand.

This topic will build the bridge from our foundation of producing data, exploratory data analysis, and probability distributions to our final goal of the course – statistical inference.

In this section, we will return our focus to situations involving ONE categorical or ONE quantitative variable. You might wish to quickly review material from Unit 1 covering these topics.

We will use these two simple cases to explore the theory behind sampling distributions and statistical inference.

Later in the semester, we will return to the two-variable cases and discuss inference in those scenarios. When we do, we will not delve too deeply into the theoretical derivation of the method but instead focus on the application of these methods to real data.

Before beginning this topic, let's quickly review some of the important ideas that we will be building on and ask some questions which we will address in this section.

Our motivating question for this section is:

Statistics we calculate from samples vary and so ... What does THEIR distribution look like?

## One Categorical Variable

| Diabetic Y/N |           |         |                      |                    |
|--------------|-----------|---------|----------------------|--------------------|
| DIABETES     | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| No           | 2142      | 92.89   | 2142                 | 92.89              |
| Yes          | 164       | 7.11    | 2306                 | 100.00             |

| BMI Category |           |         |                      |                    |
|--------------|-----------|---------|----------------------|--------------------|
| bmicat       | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Underweight  | 33        | 1.43    | 33                   | 1.43               |
| Normal       | 1013      | 43.93   | 1046                 | 45.36              |
| Overweight   | 962       | 41.72   | 2008                 | 87.08              |
| Obese        | 298       | 12.92   | 2306                 | 100.00             |

**Goal:** Use **Sample Proportion** (*known*) to estimate **Population proportion** (*unknown*)

In our discussion on one categorical variable, we presented these results from a subset of the Framingham data.

In the next unit on inference, we will be interested in estimating unknown population proportions using the data from our sample.

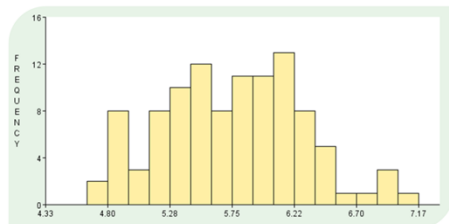
For example, here we may wish to estimate the true prevalence of diabetes or the true prevalence of obesity.

Although we can find the prevalence of diabetes in our sample (7.11%) and the prevalence of obesity in our sample (12.92%), we don't know how close these are likely to be to the true values.

In order to know that, we will need to learn more about how much this **sample proportion** might vary in repeated sampling. If we did this again how different might the result be?

We can't just "guess" at an answer!!

## One Quantitative Variable



N = 105  
 Mean = 5.4256  
 Standard Deviation = 0.5379  
 Min = 4.33  
 1st Quartile = 5.04  
 Median = 5.44  
 3rd Quartile = 5.8  
 Max = 6.81

**Goal: Use Sample **Statistic** (*known*) to estimate Population **Parameter** (*unknown*)**

For one quantitative variable, we learned to summarize our sample numerically and display the results of our sample visually including methods like those in the results presented here.

We have a histogram representing the acidity of rainwater in a sample of 105 measurements. We might be interested in using this sample to estimate or test hypotheses about aspects of the population from which this sample was taken.

Unlike the scenario with one categorical variable where we usually simply wish to estimate a proportion, or equivalently a percentage or probability, in the case of one quantitative variable, we have many aspects of the population distribution we might wish to investigate.

We will focus on the population mean as it is commonly of interest when comparing groups and investigating relationships in increasingly complex methods.

However, we could be interested in the population median, the population quartiles, the population standard deviation, etc.

Our goal will be to use the appropriate statistic from our sample to estimate the desired value in the population (called a **parameter**).

## Practice vs. Theory

### Inference – In Practice

Collect one sample and ask “what can we say about the population from which the sample was taken?”



### Now – Reverse – Theoretically

IF WE KNOW what the population looks like, “what can we expect the sample to look like?”

The previous two examples give you some idea of where we are headed. Specifically, in the next unit on statistical inference, we will want to estimate population parameters with confidence intervals or conduct a hypothesis tests about the parameter.

In inference, we will be using one sample to learn about our population.

In this section, we must approach it from the opposite direction! We label this as theoretical only because we will ASSUME we know the population for our discussions, which is unlikely in practice.

In fact, the ideas are relatively straight-forward once you realize we are studying the behavior of a particular statistic through contemplating (or simulating) the process of sampling from a theoretical – or “what if” – perspective.

Although statistical inference is more practical in that we can envision obtaining one sample from our population, it requires inductive reasoning which actually makes us take a bigger leap than we will need to make here in this study of sampling distributions! Maybe you will agree after you have looked carefully at both topics.

So, for now, we will assume a few things:

- 1) We will be taking simple random samples from the population (no dependent observations or complex sampling plans)

2) We will get to assume that we know the population under study

Then we will ask, if we repeat this process in exactly the same manner (with the same sample size for each sample)

1) What can we expect a sample to look like?

2) How much do samples vary?

3) How does a particular statistic calculated from these samples behave?

**4) What is its distribution?**

This last question is our main goal for this section.

## Parameters vs. Statistics

|                           | (Population) Parameter | (Sample) Statistic |
|---------------------------|------------------------|--------------------|
| <b>Proportion</b>         | $p$                    | $\hat{p}$          |
| <b>Mean</b>               | $\mu$                  | $\bar{x}$          |
| <b>Standard Deviation</b> | $\sigma$               | $s$                |

We have used the terminology “Parameter” and “Statistic” a few times already. Let’s define them more precisely.

A parameter is a number that describes some aspect of the population. These are fixed quantities which are usually unknown.

Examples would be the population mean, the population proportion, and the population standard deviation but there are many many more.

Here we see the notations we will use to represent these concepts in our course.

To represent a population proportion, we will use the lower case letter “p.”

To represent a population mean, we will use the lower case GREEK letter “mu.”

And to represent a population standard deviation, we will use the lower case GREEK letter “sigma.”

We used these notations when we discussed the theoretical concept of discrete and continuous probability distributions. In fact, we used them because – in those sections we were also assuming we KNEW these values for the entire **population** (or at a minimum had very good estimates of the population values).

A statistic is a number that describes some aspect of the sample.

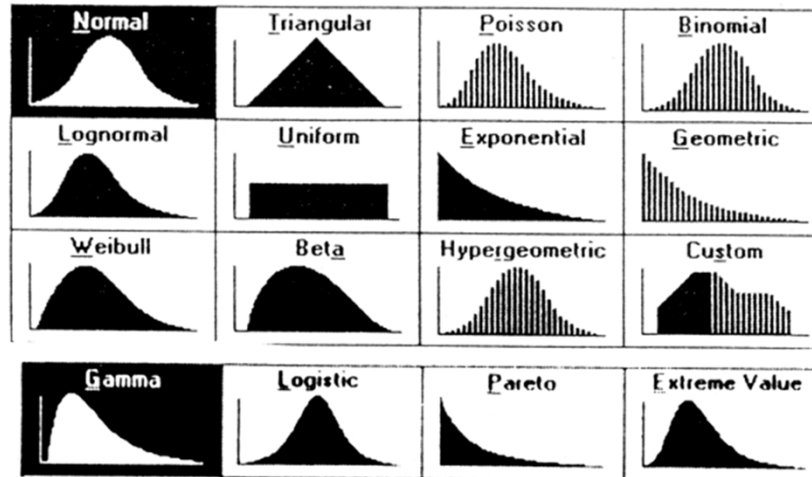
Examples would be the sample mean, the sample proportion, and the sample standard deviation but, again, there are many more.

For the sample mean and sample standard deviation, we use the same notation as in unit 1;  $\bar{x}$  represents the sample mean and a lower case “s” represents the sample standard deviation.

The only new notation is  $\hat{p}$ , which is used to represent the sample proportion which is the number of successes over the total number of trials in a binomial experiment.

So, parameters are numeric quantities that summarize some aspect of the entire population – clearly these are usually unknown to us in most research questions. And statistics are numeric quantities that summarize the sample.

## Distribution of Sample Statistics



We have seen that the results of quantitative variables in our samples can have a wide variety of shapes. This indicates that populations also are not always “normally” distributed.

Here are a few “named” statistical distributions. We studied the normal and binomial distributions in the previous section.

Our goal in this section is to study the behavior of sample statistics. Since they vary and are numeric quantities, they are themselves random variables and hence have a distribution – called the **sampling distribution**.

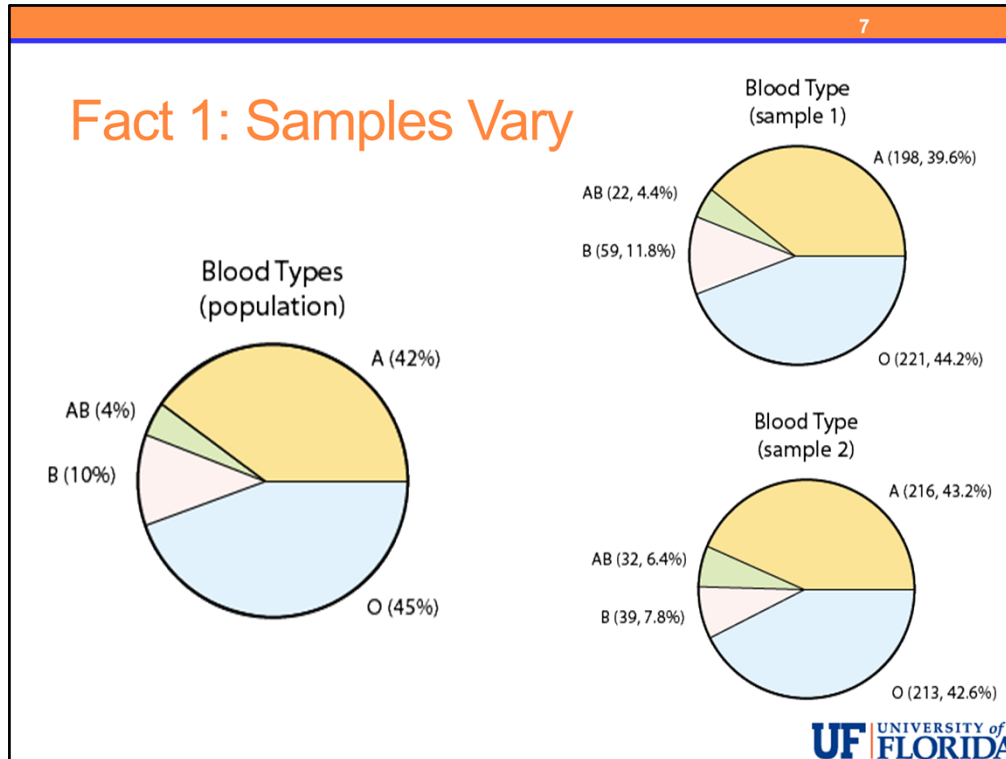
We want to study the same things as with any other distribution:

- 1) What shape does it have?
- 2) Where is it centered?
- 3) How spread out is it?
- 4) What values are common or rare?
- 5) How can we calculate probabilities or cutoff values more precisely?

The answers to these questions will be EXTREMELY important to us for statistical inference!

Let’s begin with a few examples of “how statistics from samples vary”





On the left, we have a representation of the blood types of people in the entire U.S. population. In this population 42% have blood type A.

Our population parameter of interest is the population proportion of individuals with blood type A which is:

$$p = 0.42$$

Then on the right, we have two samples.

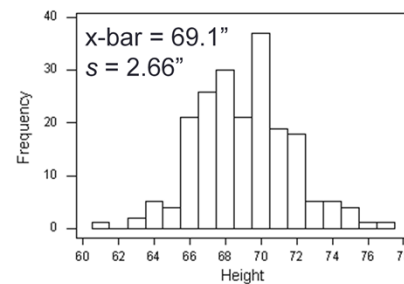
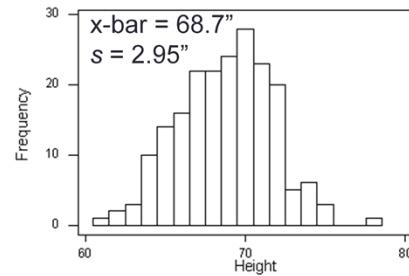
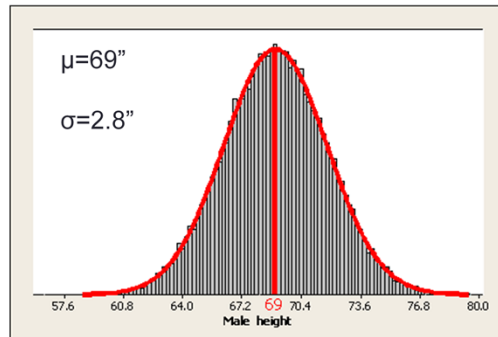
In the first sample, 39.6% have blood type A. Our sample statistic is  $\hat{p} = 0.396$ , the sample proportion of individuals from this sample with blood type A.

In the second sample, 43.2% have blood type A. Our sample statistic is  $\hat{p} = 0.432$ , the sample proportion of individuals from this sample with blood type A.

If we repeat this over and over again, we would start to gain a picture of how this statistic,  $\hat{p}$ , behaves. Where is it centered? How spread out is it? How can we calculate probabilities about it, etc.

This is the idea of sampling variability.

## Fact 1: Samples Vary



Let's look at sampling variability for a quantitative variable.

Now looking at the heights of males in the entire U.S. adult population.

On the left we have an extremely large sample and the theoretical distribution drawn. It is reasonable that we have good approximations for the mean and standard deviation of heights in this population. And heights are also well approximated by a normal distribution. In this population (U.S. Adult Males) we have the population parameters:

- 1) The population mean: " $\mu$ " = 69 inches
- 2) The population standard deviation: " $\sigma$ " = 2.8 inches

Then on the right, we have two samples.

In the first sample, the sample mean,  $\bar{x}$ , is 68.7 and the sample standard deviation is 2.95. These sample statistics are different from the exact values in the population, which is not surprising.

In the second sample, the sample mean,  $\bar{x}$ , is 69.1 and the sample standard deviation is 2.66. These sample statistics differ from both the population parameters and the statistics in the first sample.

If we repeat this over and over again, we would start to gain a picture of how these

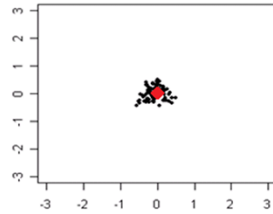
statistics behave.

We will be most interested in the behavior of the sample mean.

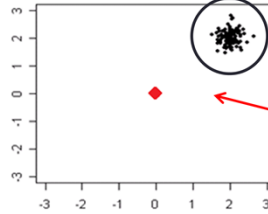
Before we move on, let me point out that these histograms give a good demonstration of how we might expect samples from a normal distribution to look. Notice they are not perfectly normal or even perfectly symmetric. This is why we must be cautious about making strict judgments about the shape of our population based upon our sample – without careful study. Our samples will vary and although, on average, the process will represent the population well, we can observe rare instances where it does not.

## Accuracy/Bias vs. Precision/Variability

Small Bias, Small Variability



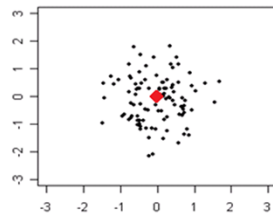
Large Bias, Small Variability



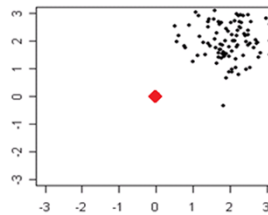
Result of  
Statistic for  
Different  
Samples

Parameter

Small Bias, Large Variability



Large Bias, Large Variability



UF UNIVERSITY of FLORIDA

We used this slide during our introduction to Unit 2 where we elaborated a little further on the concepts of **bias** and **variability**.

Both of these concepts rely first on this idea of repeating the process – repeated sampling.

Consider the red diamond in the center to be the true – fixed – population parameter and each black dot to represent the result of the statistic for different samples.

In the top-left plot, we have estimates which, on average, hit the target – and so are **unbiased**. And the points are very close to the target and thus the process is producing estimates with relatively small variability.

In the top-right plot, we have estimates which, on average, DO NOT hit the target – and so are **biased**. The points are still very close together though and thus the process still produces estimates with relatively small variability.

In the bottom-left plot, we have estimates which, on average, hit the target – and so are **unbiased**. Here, however, the points TEND TO VARY further from the target and thus the process produces estimates with relatively large variability.

In the bottom-right plot, we have estimates which, on average, DO NOT hit the target – and so are **biased**. The points TEND TO VARY further from the target and thus the process also produces estimates with relatively large variability.

It is clear that we would like an estimator that is

- 1) Unbiased – meaning – on average it will not overestimate or underestimate the target.  
Another statistical way of saying this is that the EXPECTED VALUE or mean of the statistic is EQUAL TO THE PARAMETER.

And we want our estimator to

- 2) Vary as little as possible – have the minimum possible standard deviation

If we know something about how variable our statistic will be, it will help us make informed decisions about the location of the parameter.

For example, if we know we are in the top-left case, we will be more certain that our estimate is closer to the true value than in the bottom-left case where the variation is much greater. They are both unbiased but we would clearly prefer less variation!

In practice – as we will see shortly – decreasing this variation usually requires increasing our sample size and this will result in increased costs to conduct the study – so a balance is often found that does the best possible with the resources available.

# SIMPLE EXAMPLES

---

We will use a few simple examples to give an overview of the idea of sampling distributions as well as a quick introduction to how they will be used to conduct hypothesis tests.

## Five-Question MC Test – Guessing?

|       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|
| NNNNN | NNNNY | NNNYN | NNYNN | NYNNN | YNNNN | NNNYY | NNYNY |
| NYNNY | YNNNY | NNYYN | NYNYN | YNNYN | NYYNN | YNYNN | YNNNN |
| NNYYY | NYNYY | YNNYY | NYYNY | YNYNY | YYNNY | NYYYN | YNYYN |
| YYNYN | YYNYN | YYYYN | YYYNY | YYNYY | YNYYY | NYYYY | YYYYY |

| Value of X | p-hat | Probability |
|------------|-------|-------------|
| 0          | 0.00  | 0.3277      |
| 1          | 0.20  | 0.4096      |
| 2          | 0.40  | 0.2048      |
| 3          | 0.60  | 0.0512      |
| 4          | 0.80  | 0.0064      |
| 5          | 1.00  | 0.0003      |

Suppose a student takes a 5-question multiple choice test where each question has 5 possible answers.

If we assume the student is guessing, there is a  $1/5 = 0.2$  probability of successfully answering each question.

There are 5 possible questions and if the student is guessing for each, the answers on each question are independent of each other.

So ... we can use the binomial distribution to obtain the distribution of the random variable:

$X$  = the number of questions correct out of the 5 answered.

In our current terminology, notice that here our “sample” consists of 5 questions and our “statistic” is  $X$  = the number of questions correct.

We could convert  $X$  to  $p$ -hat by dividing by 5 which in this case would represent the proportion of questions answered correctly.

In the last set of material, we would have called this the probability distribution of  $X$ . And we discussed the mean and standard deviation of the random variable  $X$ .

Since  $X$  is a statistic, we can also call this distribution **the sampling distribution of  $X$** .

Notice that it doesn't represent what happens to one particular student but the entire set of possibilities if we were to have many students take the quiz and completely guess on each question.

Around 33% of the time we expect students to get zero questions correct, about 41% of the time we expect students to get 1 question correct, and so on.

So that is the idea, we will be determining formulas for the mean and standard deviation of the sampling distribution of certain statistics (namely  $\hat{p}$  and  $\bar{x}$ ).

We will use these sampling distributions to make inferential decisions. For example, suppose you are an instructor and you have given a 5 question MC quiz like this. If you have a student who scores a 4 or a 5 on this quiz, do you think the student is guessing on every question?

This seems unreasonable, given that – if the student were guessing – the chance he or she could have scored so high is only  $0.0064 + 0.0003 = 0.0067$ . Thus it is reasonable to INFER that the student wasn't guessing.

Notice that we could be incorrect. The result could be a rare event. The student COULD be guessing, but if so, the chance is extremely small that we could have seen a score of at least 4 on the quiz.

This process of inductive reasoning is the basis of inference and relies on sampling distributions as we introduced in this example – which is our first attempt at a hypothesis test!



## Five-Question T/F Test – Guessing?

|       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|
| NNNNN | NNNNY | NNNYN | NNYNN | NYNNN | YNNNN | NNNYY | NNYNY |
| NYNNY | YNNNY | NNYYN | NYNYN | YNNYN | NYYNN | YNYNN | YYNNN |
| NNYYY | NYNYY | YNNYY | NYYNY | YNYNY | YYNNY | NYYYN | YNYYN |
| YYNYN | YYYNN | YYYYN | YYYNY | YYNYY | YNYYY | NYYYY | YYYYY |

| Value of X | p-hat | Probability |
|------------|-------|-------------|
| 0          | 0.00  | 0.03125     |
| 1          | 0.20  | 0.15625     |
| 2          | 0.40  | 0.3125      |
| 3          | 0.60  | 0.3125      |
| 4          | 0.80  | 0.15625     |
| 5          | 1.00  | 0.03125     |

What if the quiz was a 5-question true/false quiz?

Here we have the sampling distribution of  $X$  = the number of correct questions.

Now, if a student scores a 5/5 on the quiz, there is a 3% chance this could happen if the student were guessing and although this isn't a large probability – it is not extremely rare. We would be less comfortable concluding that the student is not guessing based upon such a quiz and result.

These examples are simply to help you to see the connection between where we just were (probability distributions), where we are now (sampling distributions), and where we are headed (inference).

The probabilities we have mentioned (3% for this example and 0.0067 for the previous example) will be what we will call a p-value for a hypothesis test in the next unit on inference.

# IMPORTANT RESULTS

## Sampling distributions



Before presenting the two sets of equations, let's formally define the sampling distribution and the standard error.

**The sampling distribution of a particular sample statistic is the distribution of that statistic when the sampling process is repeated (using the same sample size).**

We will look at the two statistics  $\hat{p}$  (the sample proportion) and  $\bar{x}$  (the sample mean) and we will often use the phrases:

The sampling distribution of  $\hat{p}$  and the sampling distribution of  $\bar{x}$  – to be specific as to which sampling distribution we mean. Instead of  $\hat{p}$  or  $\bar{x}$ , we can also use their descriptions (the sample proportion or the sample mean respectively).

**The standard error of a statistic is the standard deviation of the sampling distribution of that statistic.**

We will use the phrases: standard error of  $\hat{p}$  and standard error of  $\bar{x}$  but realize these are in fact the standard deviation of the distribution of the statistic in repeated sampling!!

Although the terminology causes students some difficulty, you can see why such shortened terminology is helpful. The full phrases to define the sampling distribution and standard error are quite long.

It is extremely important to develop an understanding of both of these ideas in order to be able to understand the framework of statistical inference.

We will leave the formal definition of the standard error for the beginning of the next unit but wish to mention it here so that you will begin to understand what about the current topic of sampling distributions will be most important as we proceed – it is this idea of the variation of our statistic – which is commonly measured in practice by the standard error of that statistic.

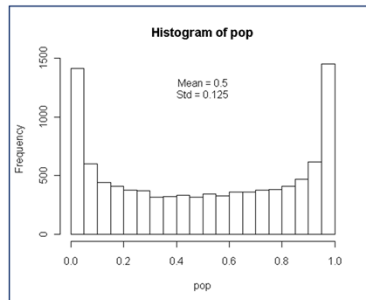
## Central Limit Theorem

- For large samples:
- Sampling distribution of sample means of the same sample size will be approximately normally distributed
- If the population from which the sample was drawn is itself normal, the above is true for all  $n$  (not just large  $n$ )

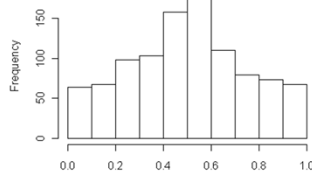
Both of the results we will discuss fall under the broader result of the central limit theorem which states that if we take repeated simple random samples of the same size from a particular population, and calculate the sample mean for each sample, the sampling distribution of these sample means will be approximately normally distributed.

Although it is also true that if the population itself is normally distributed then the sample mean will be normally distributed for any  $n$ , this is not as useful as very few populations are in fact normally distributed in practice.

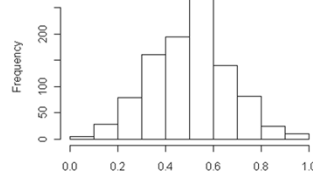
## Example - CLT



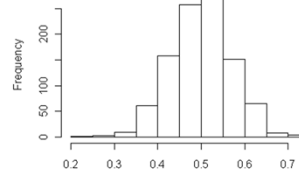
N = 2



N = 5



N = 25



Here we see a population (outlined in black) and the approximate sampling distributions of  $\bar{x}$  for samples of size  $n = 2, 5$ , and  $25$ .

As the sample size increases, we can see that the samples are getting increasingly normally distributed (despite the crazy lack of normality in the population) and are become less variable.

Let's present the two sets of results. Then we will convince you of these results and look at a few examples.

## Sampling Distributions

| Variable                                           | Parameter                                                                  | Statistic                        | Sampling Distribution |                           |                                                                     |
|----------------------------------------------------|----------------------------------------------------------------------------|----------------------------------|-----------------------|---------------------------|---------------------------------------------------------------------|
|                                                    |                                                                            |                                  | Center                | Spread                    | Shape                                                               |
| Categorical<br>(example:<br>left-handed or<br>not) | $p$ = population<br>proportion                                             | $\hat{p}$ = sample<br>proportion | $p$                   | $\sqrt{\frac{p(1-p)}{n}}$ | Normal if $np \geq 10$<br>and $n(1-p) \geq 10$                      |
| Quantitative<br>(example: age)                     | $\mu$ = population<br>mean, $\sigma$ =<br>population<br>standard deviation | $\bar{x}$ = sample<br>mean       | $\mu$                 | $\frac{\sigma}{\sqrt{n}}$ | Normal if $n > 30$<br>(always normal if<br>population is<br>normal) |

<http://www.stat.tamu.edu/~west/ph/sampledistrib.html>



This table summarizes all of the important information in this section. From this and what we know about normal distributions from the previous section, we can answer the kinds of questions which will be of interest to us.

For sample proportions ( $\hat{p}$ ), we have a formula for the mean (labeled as the center) and standard deviation (spread) of the sampling distribution.

This states that the sampling distribution will be centered at the true value of  $p$  (the population proportion) and will have a standard deviation of the square root [ $p$  times  $(1-p)$  divided by  $n$ ]. It is not a coincidence that the formula for the standard deviation is similar to what we learned for binomial random variables but in this course we will not delve into the mathematics of this connection.

We sometimes use notations for these values:  $\mu_{\hat{p}}$  is designed to denote “the mean of the sample proportions in repeated sampling” or “the expected value of the sample proportions in repeated sampling” and similarly the notation:  $\sigma_{\hat{p}}$  denotes “the standard deviation of the distribution of sample proportions in repeated sampling” which is also known as “the **standard error**” of  $\hat{p}$ .

The formulas for the mean and standard deviation of the sample proportion will always apply but the normal approximation will only be reasonable for large enough samples.

In the case of sample proportions, we must have  $np$  and  $n(1-p)$  be at least 10. In order to

apply the normal approximation for smaller probabilities, we need correspondingly larger samples.

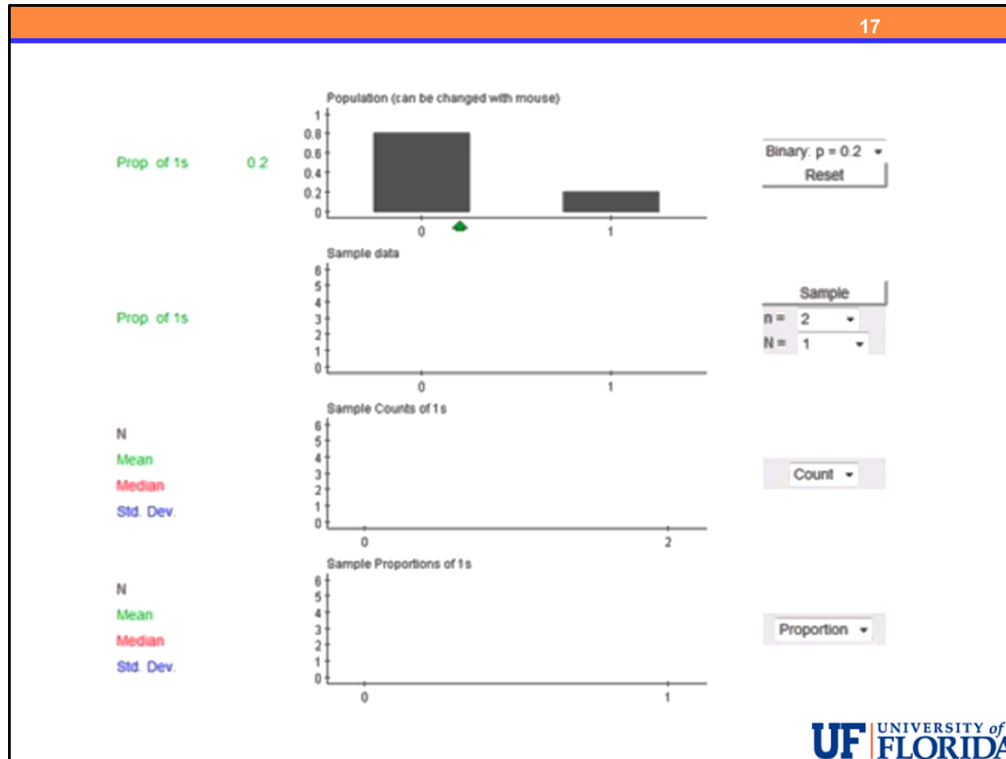
Notice that these results also show that the sample proportion,  $\hat{p}$ , is an unbiased estimator since, on average,  $\hat{p}$  will be centered at the true population proportion,  $p$ .

It may also be clear that if the sample size,  $n$ , increases, the standard deviation of  $\hat{p}$  will decrease, resulting in a distribution which is less variable and a statistic which will, on average, be closer to the true value of  $p$  than that for a smaller sample.

For sample means ( $\bar{x}$ ), the sampling distribution will be centered at the population mean,  $\mu$ , and will have a standard deviation of  $\sigma$  divided by the square root of  $n$ . Although 30 is not always large enough, for this course, we will consider the normal approximation possible for samples sizes larger than 30.

Before moving on, now is a good time to mention that the sample proportion is in fact nothing more than a special case of the sample mean when the data come from a sample containing nothing but 0's and 1's.

Now we will use an applet to illustrate these concepts in each case.



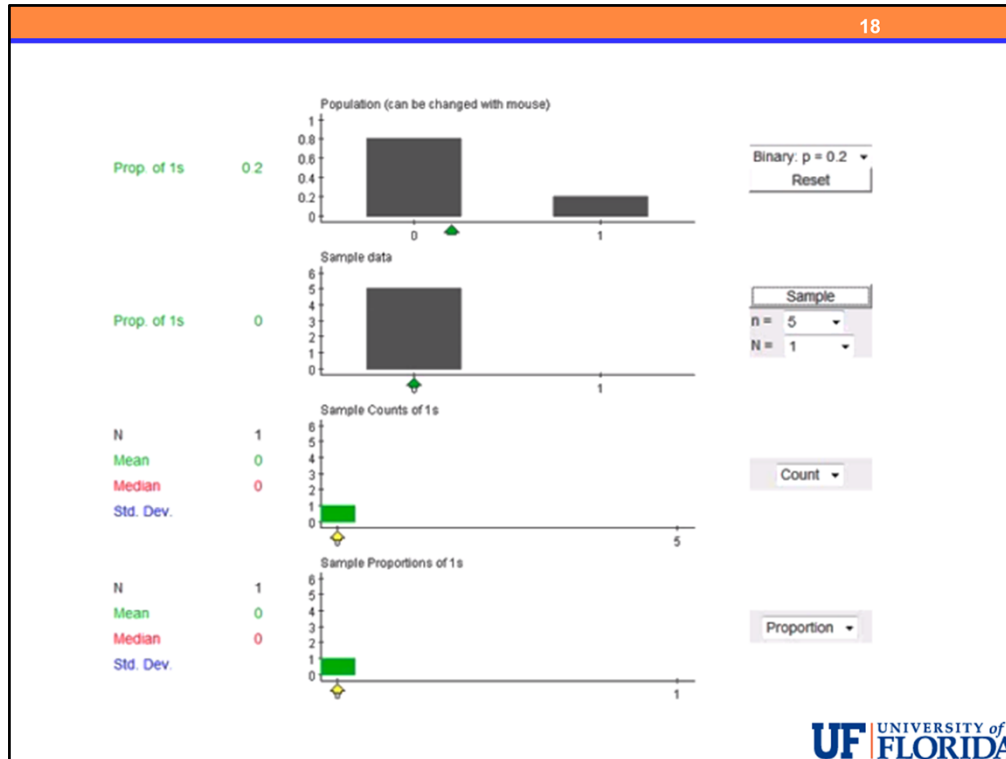
Let's start with the first of our earlier examples with the binomial which was a 5 question multiple choice test ( $n = 5$ ) where each question has 5 choices. If students guess on each question, the probability of a correct guess for a particular question is  $1/5$  or  $0.2$  ( $p = 0.2$ ).

This applet will let us sample from this scenario by choosing the population to be binomial  $p = 0.2$  and then setting  $n = 5$ .

Capital N represents, in this case, the number of students going through this process - taking the test and guessing on each choice.

We will start with one student. Notice the two distributions are for the count (we called this X - the number of successes - when we discussed binomial random variables) and proportion - this is  $\hat{p}$  - which we are studying in this section.



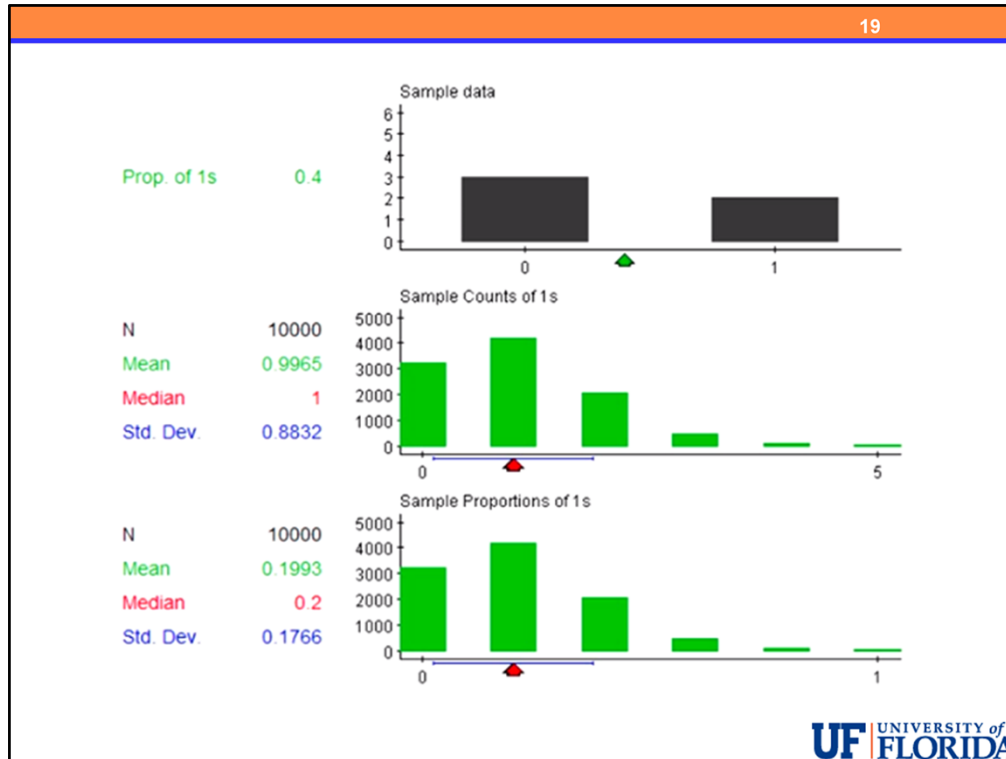


When we hit sample once, we see the data displayed for this student in the second plot - how many 0's (wrong answers) and how many 1's (correct answers).

Then it drops the value of  $X$  and  $\hat{p}$  into the two lower graphs. In this case the student scored 0 on the quiz. Notice that  $\hat{p}$  truly is simply the average of the 0's and 1's!

Let's repeat that a few more times to see the process. (See video)

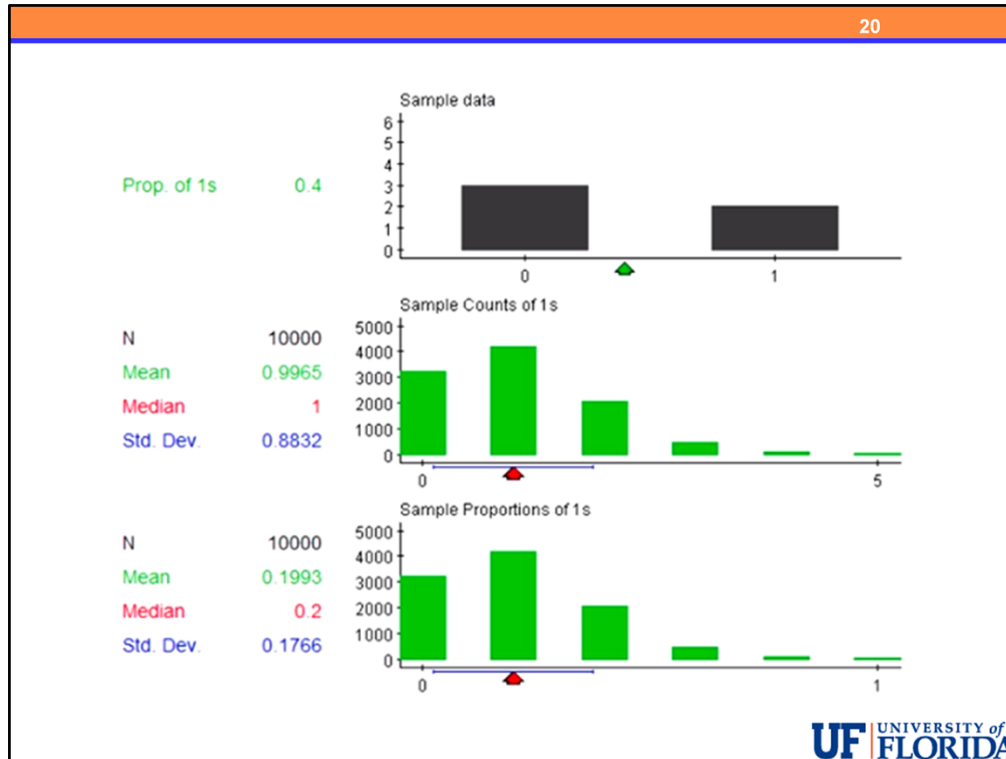
To speed up the process, we can change capital  $N$  to a much larger number and simulate what would happen if we had 10, 100, 1000, or 10000 students take the quiz at once.



Let's reset and change capital N to 10000 which will give us a good approximation of the true distributions of  $X$  and  $\hat{p}$ . We already know the distribution of  $X$  should follow a binomial distribution and that we could calculate these probabilities exactly using the binomial formula.

If you compare these probabilities to the binomial calculator - they are very similar.

In this case, the sample size for each sample is only 5 and thus the central limit theorem is not applicable. It is clear that a normal distribution would not approximate the bottom picture well.



Before we move on, notice that the distribution for  $X$  or  $\hat{p}$  looks identical except for how the x-axis is labeled.

We are interested in learning about the center and spread of the bottom distribution. We have already learned the formulas for center and spread for the distribution for the counts -  $X$ .

In the previous section we learned the mean is  $n$  times  $p$  which is  $5 \cdot 1/5 = 1$ . In our simulation we have 0.9965 - very close! If we continued to sample we could get even closer. See if you can calculate the standard deviation and compare it to the value in our simulation of 0.8832.

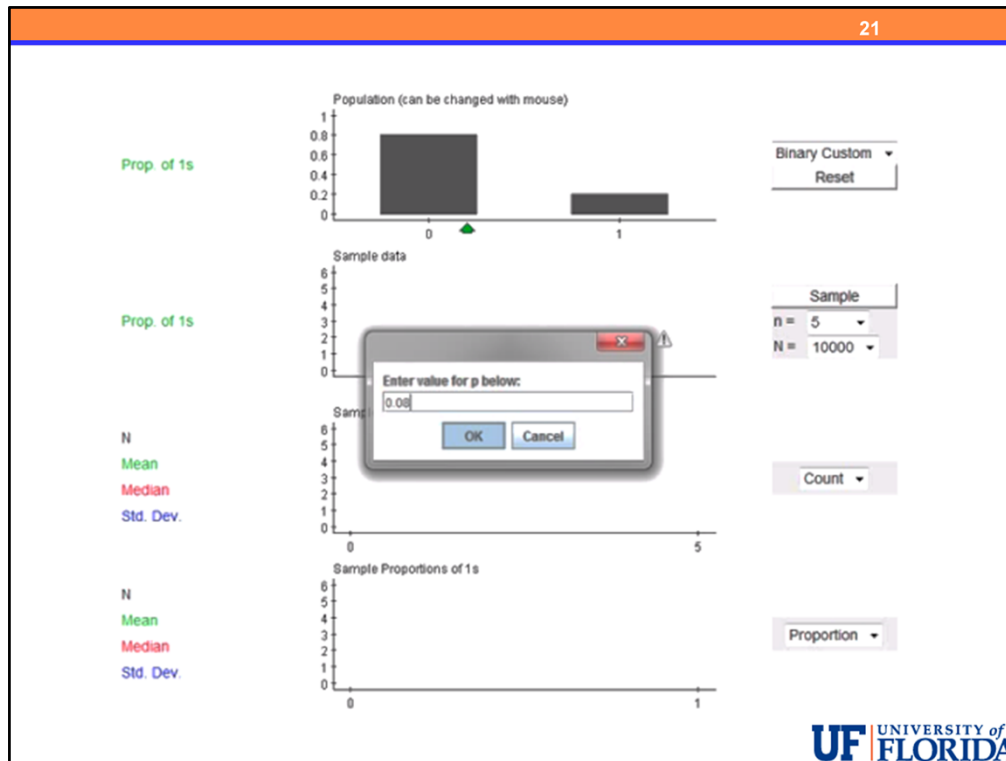
What we care about now is the center and spread of the distribution of  $\hat{p}$  - the bottom distribution in the applet.

For the mean - it should be centered at the true population value which is 0.2. In the simulation we get 0.1993.

The standard deviation is  $\sqrt{p(1-p)/n} = \sqrt{0.2 \cdot 0.8/5} = \sqrt{0.032} = 0.1789$  and in our simulation we get 0.1766.

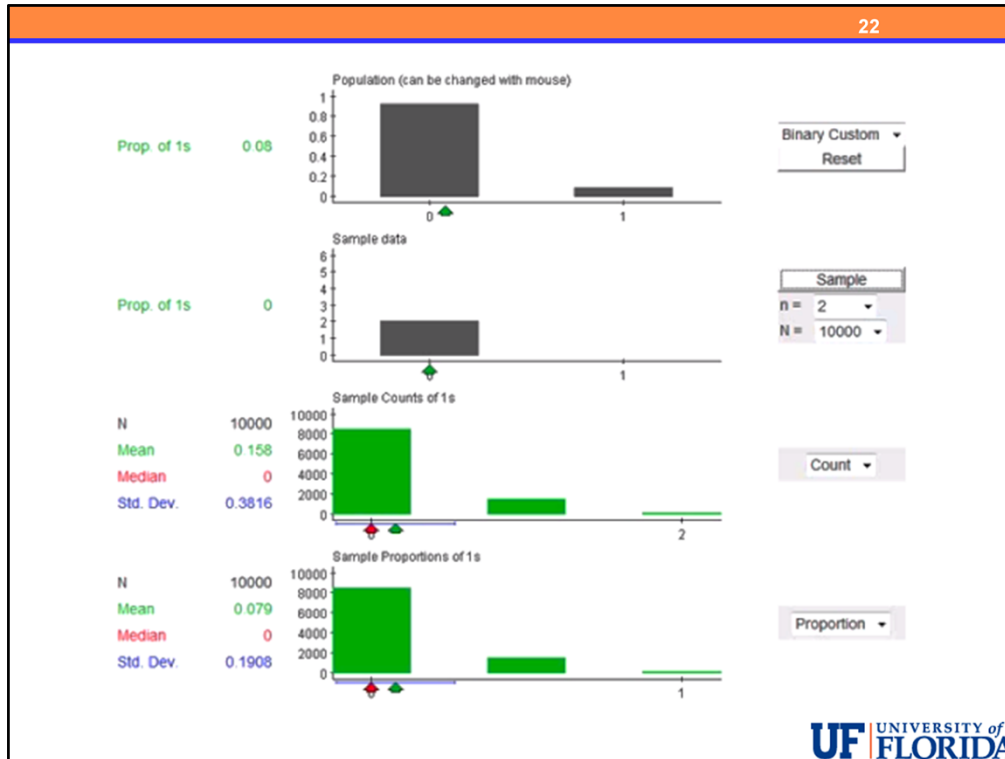
Notice these values continue to tell us the same things as always.

Where is the "center" of the data and how far, on average, do we expect random observations - in this case values of  $\hat{p}$  - to fall from the center.

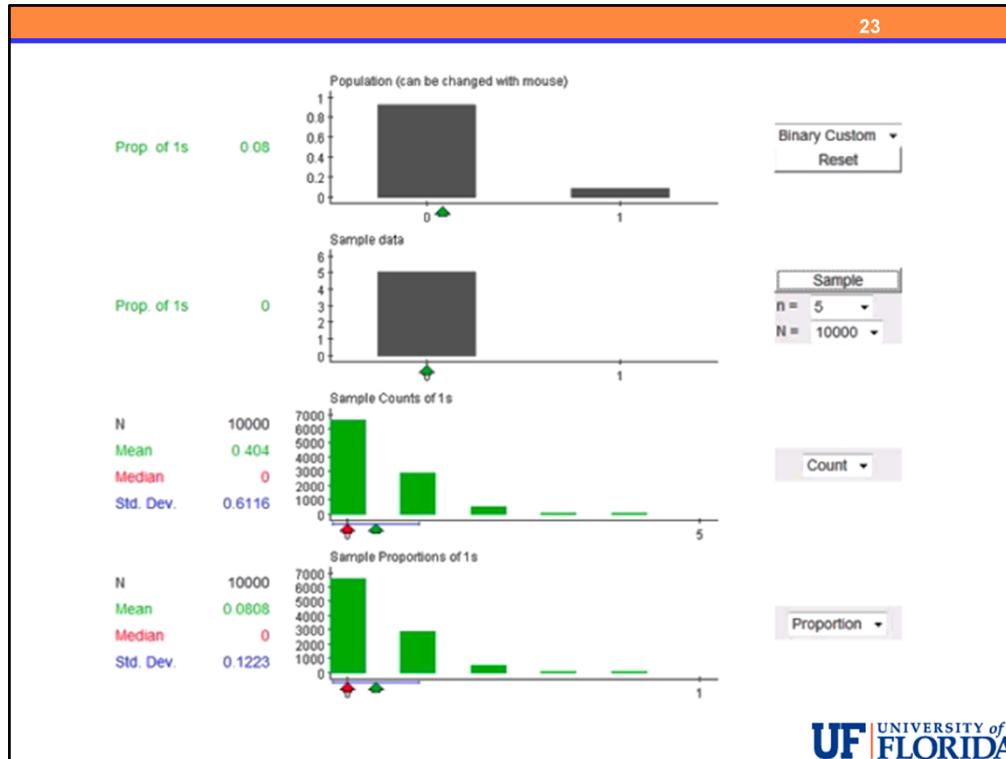


Now let's look at a more realistic example - suppose the prevalence of diabetes in a certain population is 8%.

We are going to illustrate the sampling distribution for increasing sample sizes using this applet. We need to enter a custom probability in the applet.

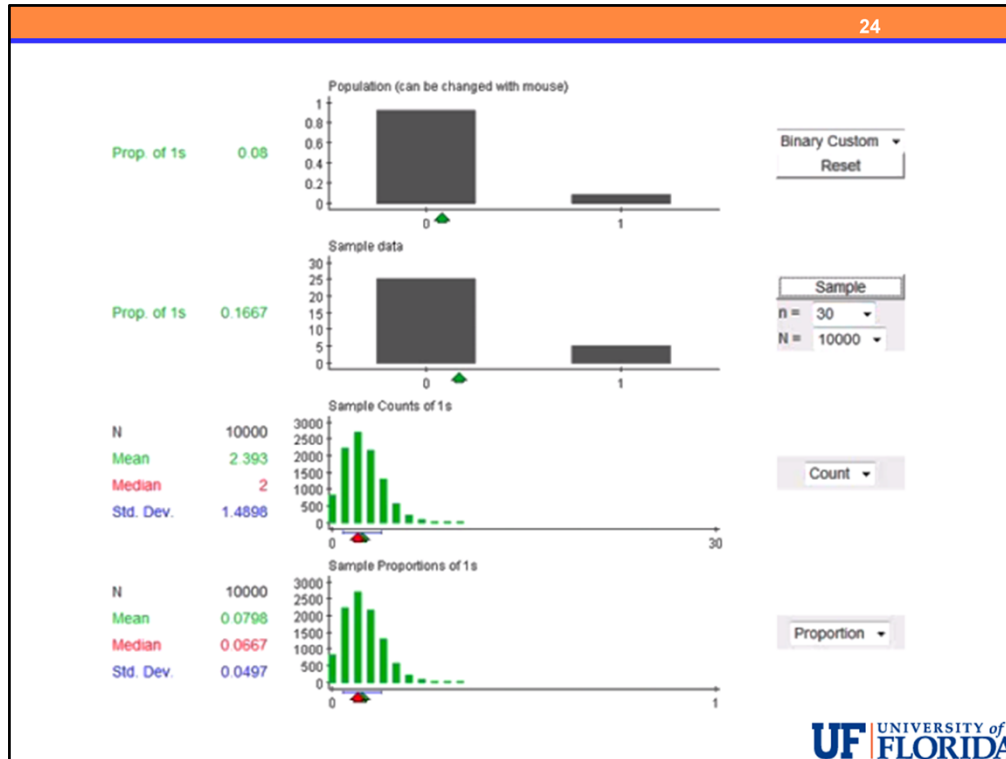


We will start with a sample of size  $n = 2$ , we will skip straight to repeating this process 10000 times. The sampling distribution of  $\hat{p}$  (at the bottom) is clearly heavily skewed right - this is reasonable since for a random sample of 2 people we would most often expect neither of them to be diabetic, rarely 1, and even more unlikely 2.



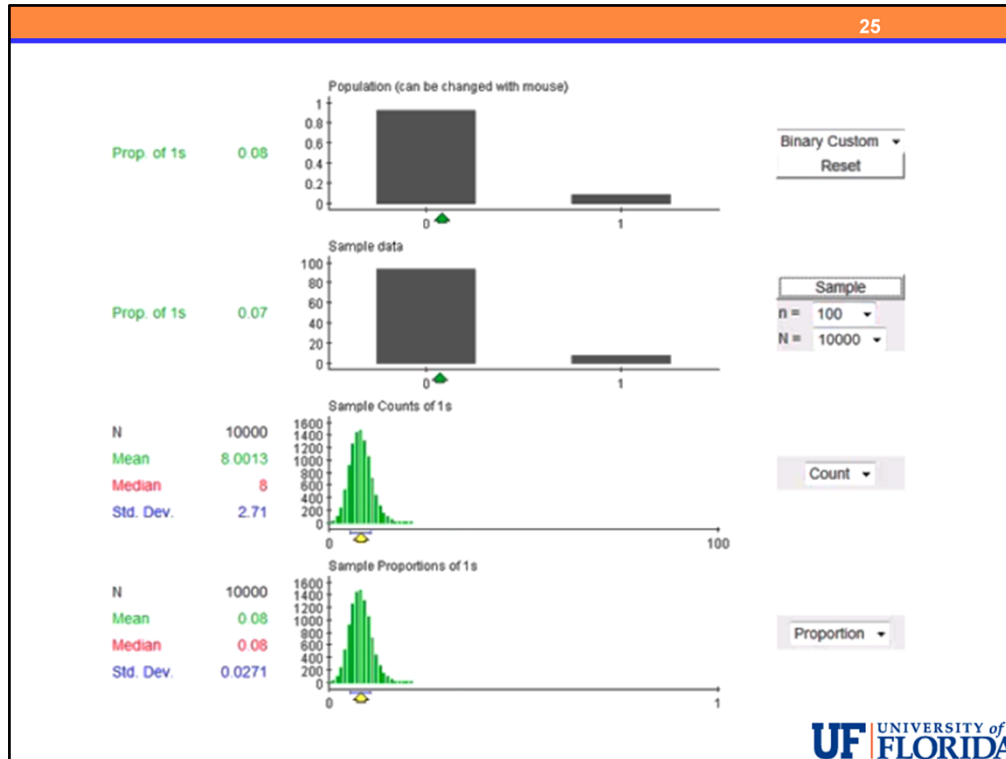
Now let  $n = 5$ . Still skewed right.

By the rules we discussed - in order for the normal approximation to apply, we need  $n$  times  $p$  and  $n$  times  $(1-p)$  to both be at least 10. Since  $p = 0.08$ , we would need  $n$  to be at least 125 to satisfy this requirement as  $125(0.08) = 10$ .



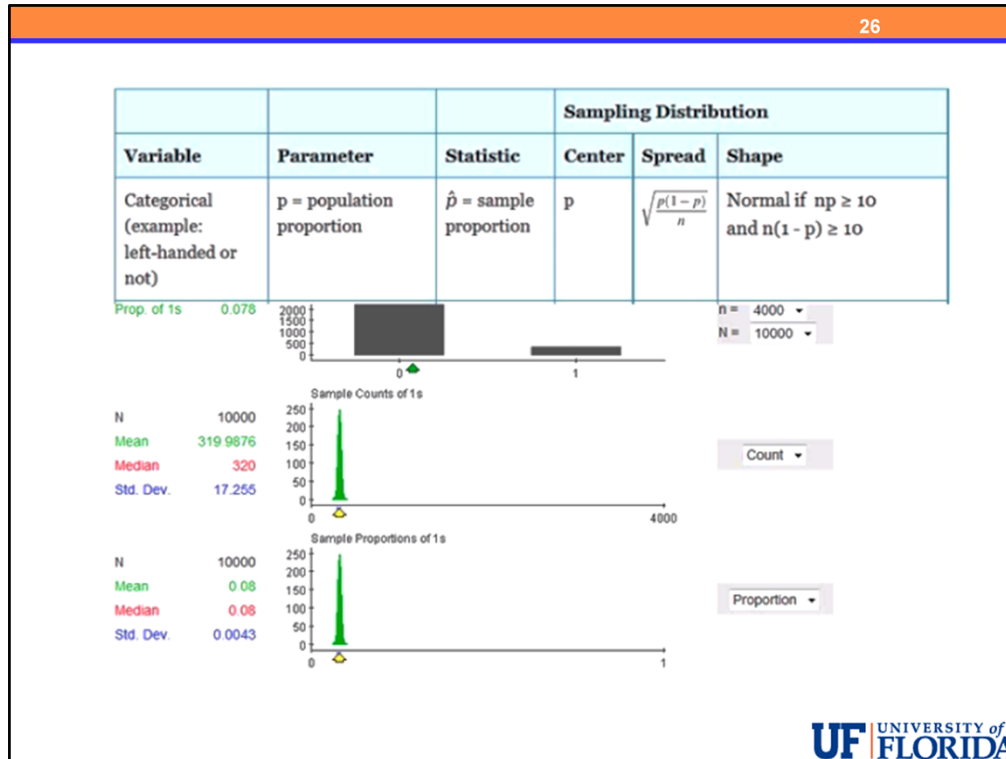
So let's skip to 30 - we can see it is getting better - much more symmetric but still not well approximated by a normal distribution.





If we go to 100, we can really begin to see how the central limit theorem works.

If you review the graphs they are all centered at around 0.08 but as our sample size increases (little n). The variability is decreasing AND the distribution is becoming increasingly normally distributed!



Let's take a sample of size  $n = 4000$  - the largest possible. This takes a little longer and now ... notice how the distribution is a tiny sliver around the true value of 0.08!!

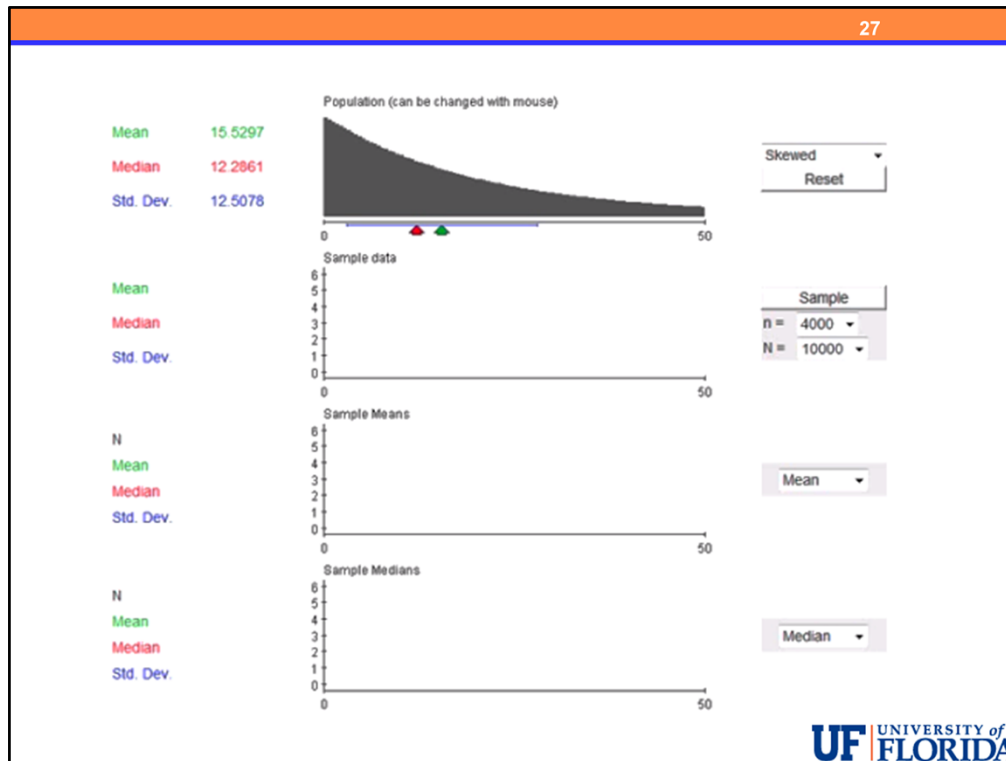
If you can take good random samples - increasing the sample size will provide estimates with less and less variation - they will be closer to the truth.

In every one of these results, we can see how close the mean and standard deviation of the simulated results comes to the theoretical values.

To summarize - to find the sampling distribution of  $\hat{p}$  we need to

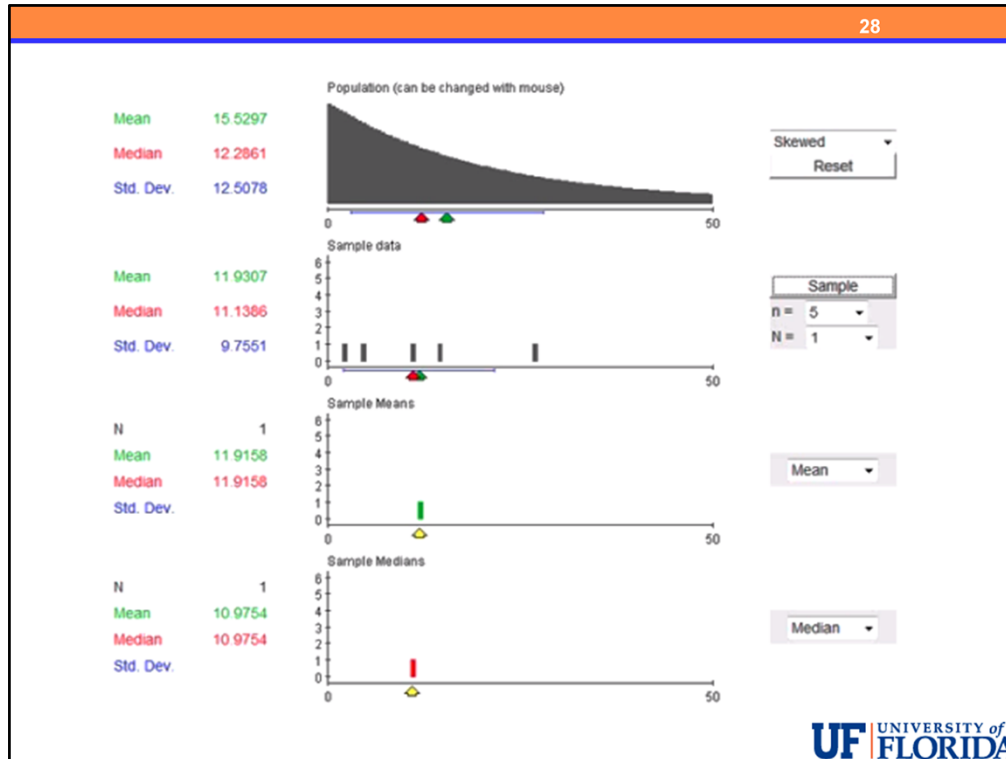
- 1) find the mean of the sampling distribution - this is simply the true population proportion - no work to do!
- 2) find the standard deviation of the sampling distribution - which we will soon call the standard error of  $\hat{p}$  - this is the square root of the quantity  $p$  times  $(1-p)$  divided by  $n$ .

Remember in these simulations capital  $N$  represents the repeated samples - our simulation size - and little  $n$  represents the sample size of each individual sample.



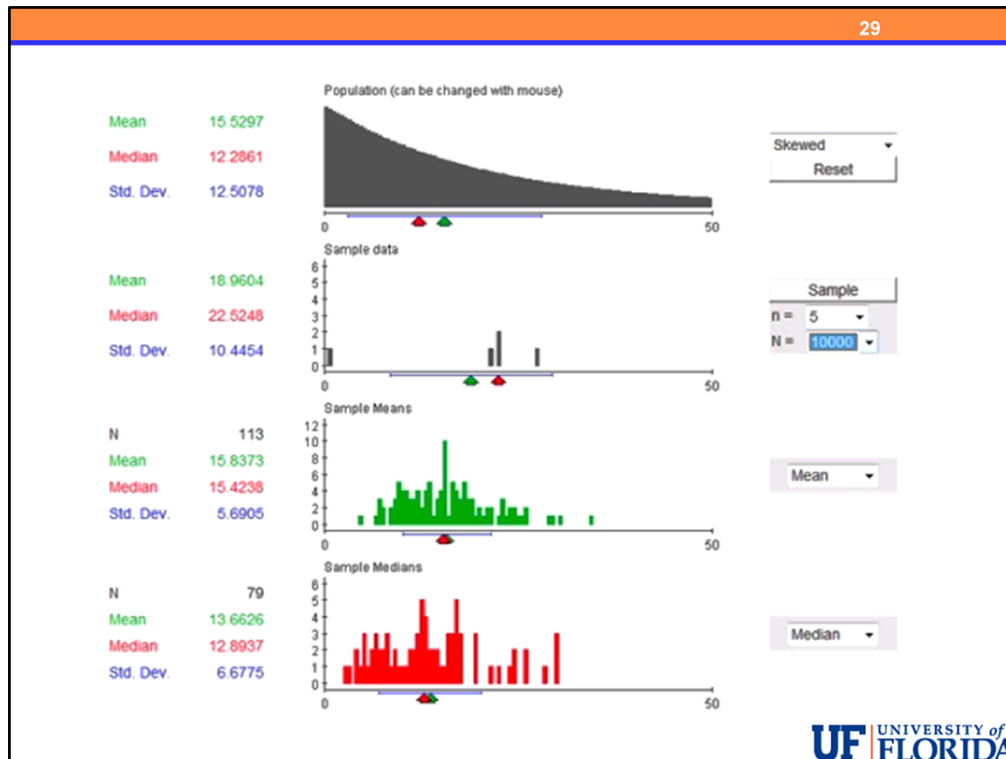
Now let's look at a situation for the sample mean,  $\bar{x}$ . For this we will use the skewed population which is available in the applet. This is also a skewed right distribution and we will see similar results except that the nature of the original population is continuous rather than discrete.

We will go through the same progression.



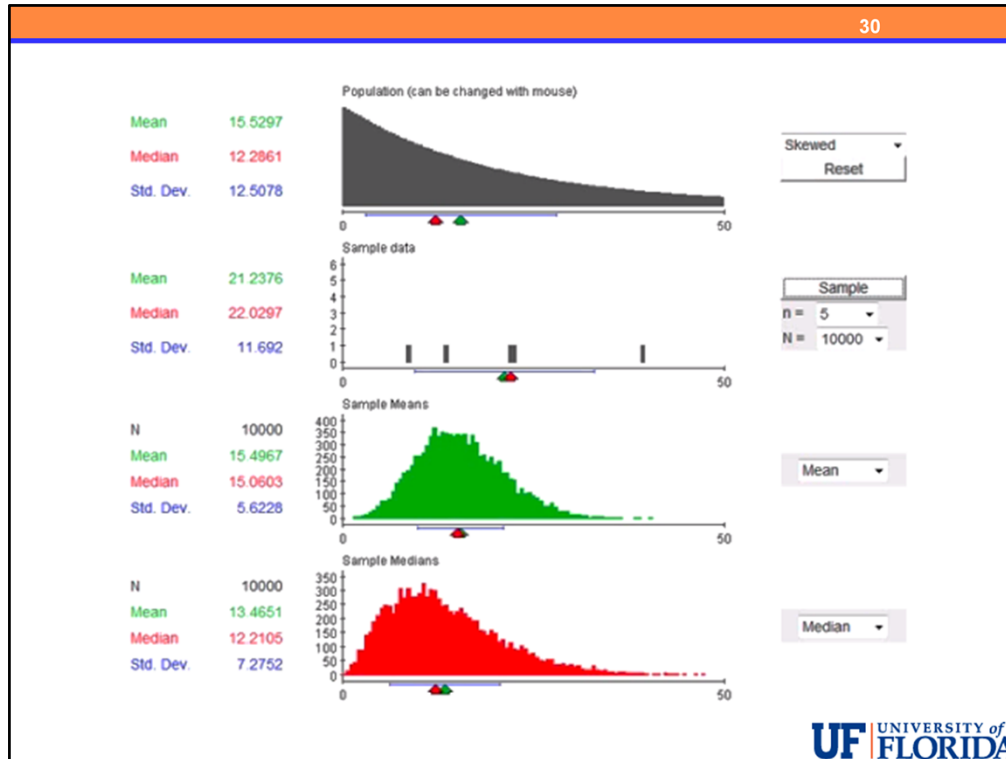
First we set  $n$  to 5 and illustrate the sampling process again for capital  $N = 1$ .

When we take one sample of size 5, we see the values in the second plot and then the MEAN drops into the third plot. A comparison to the median is also provided so the median drops into the final plot.

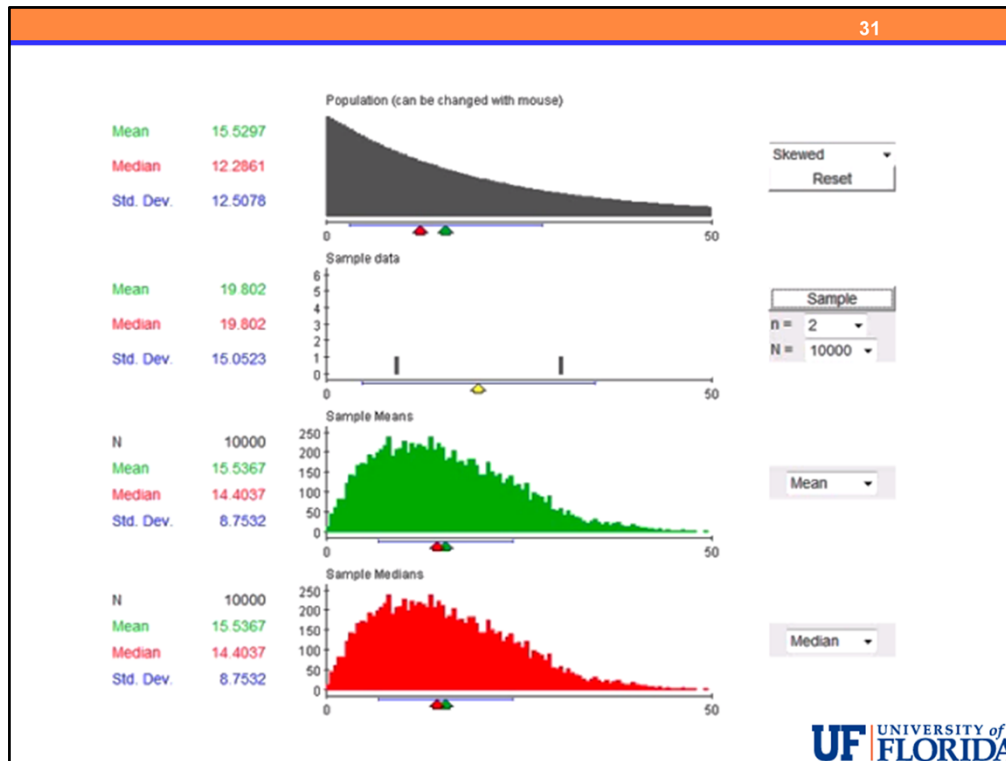


Each time we sample, we see the process repeat. If we continue to do this, the picture will eventually take shape! (repeat – See video)

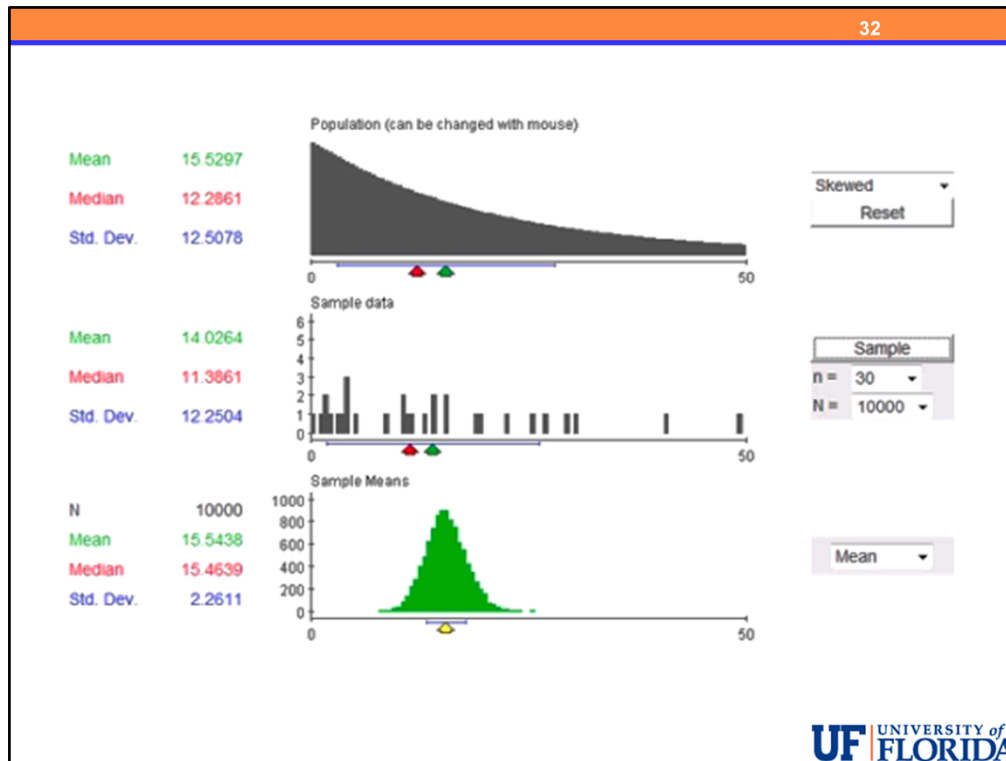
Now let's reset and let capital N be 10000 again.



Notice that already for a sample of size 5, the distribution of the sample mean is remarkably normally distributed even for such a skewed population. Notice that the distribution of the median on the bottom is still clearly skewed right.



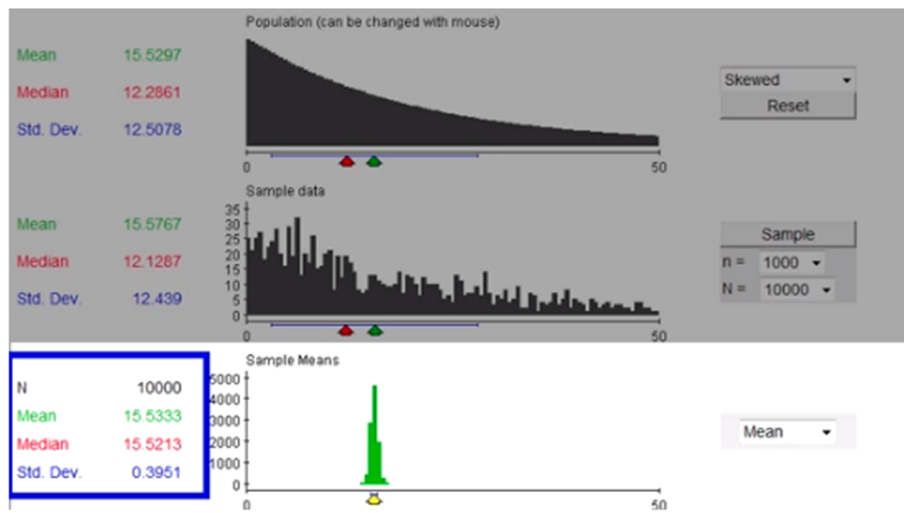
We could back up to  $n = 2$  or  $3$  to see a more skewed result for the sampling distribution of  $\bar{x}$ . And, each time we do the simulation, we will get slightly different results.



We stated that a sample size over 30 should be large enough to apply the normal approximation, regardless of the original population. When we set  $n = 30$  here we see that the sampling distribution is well approximated by a normal distribution.

We can also see that the variation is decreasing as the sample size, little  $n$ , increases.

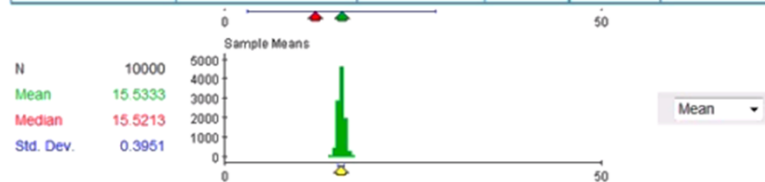




If we increase the sample size further, notice that the distribution gets extremely tight around the target value.

Again, we can compare the theoretical values of center and spread to those in each simulation for the mean (remember these are the third plot by the default settings).

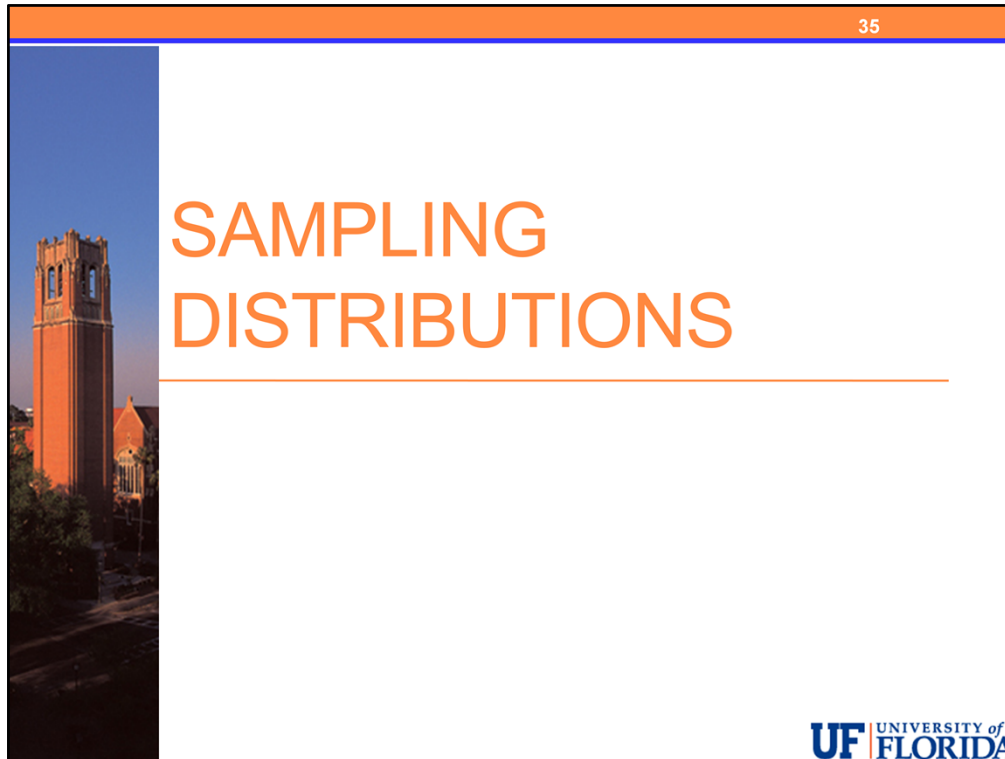
| Variable                       | Parameter                                                         | Statistic               | Sampling Distribution |                           |                                                               |
|--------------------------------|-------------------------------------------------------------------|-------------------------|-----------------------|---------------------------|---------------------------------------------------------------|
|                                |                                                                   |                         | Center                | Spread                    | Shape                                                         |
| Quantitative<br>(example: age) | $\mu$ = population mean, $\sigma$ = population standard deviation | $\bar{x}$ = sample mean | $\mu$                 | $\frac{\sigma}{\sqrt{n}}$ | Normal if $n > 30$<br>(always normal if population is normal) |



To find the sampling distribution of  $\bar{x}$ -bar we need to

- 1) find the mean of the sampling distribution - this is simply the true population mean - no work to do!
- 2) find the standard deviation of the sampling distribution - which we will soon call the standard error of  $\bar{x}$ -bar - this is the population standard deviation divided by the square root of  $n$ .

The applet can also be used to look at the standard deviation and variance.



Remember that in the simulations we used and the results we presented, we are assuming a simple random sample. Sampling distributions can be derived in increasingly complex situations but our goal is to introduce you to this theoretical concept using these two simple cases.

These will give you some understanding the KIND of process statisticians go through in developing new methods.

In this class, we want to help you to understand what is hiding in that cloud of probability every time you interpret a p-value or a confidence interval in the future.

In the next unit, we will learn about the theory behind confidence intervals and hypothesis tests in these simple cases and then, for future methods, we will rely on software to obtain the results for us and focus on interpreting those results.

Maybe now you can see how we have been leading to these ideas throughout the semester.

We need to

- understand samples, exploratory data analysis, and basic probability so we can
- understand probability distributions so we can
- understand the concept of the distribution of a sample statistic in repeated sampling – which we call the sampling distribution of the statistic

And, sampling distributions are the basis of the process of statistical inference that we are finally about to discuss!

In this section we have presented conditions under which the sampling distributions of the sample mean ( $\bar{x}$ ) and the sample proportion ( $\hat{p}$ ) will be approximately normally distributed and we have specified the mean and standard deviation of the sampling distribution in each case. It is very important to correctly identify each scenario.

Remember that in a proportion problem – there will be no mean or standard deviation provided for the population. The question will be about a proportion, percent, prevalence, or probability of some kind. The original data collected on a single individual will be categorical (diabetic or not diabetic).

For mean problems, you should be given the mean and standard deviation of the population and be working with a population that represents a quantitative measurement or count.

Our initial question: Statistics vary so ... what does THEIR distribution look like?

Since we will be using these statistics to make decisions, we must learn about how these statistics behave, what their distribution looks like, how much they vary, before we can have any idea whether the values we see in our sample have any interpretable meaning or not!

The concept of sampling distributions is among the most difficult to fully understand. You may need to review the materials a few times and try it yourself with the interactive applets to get the idea to sink in entirely!