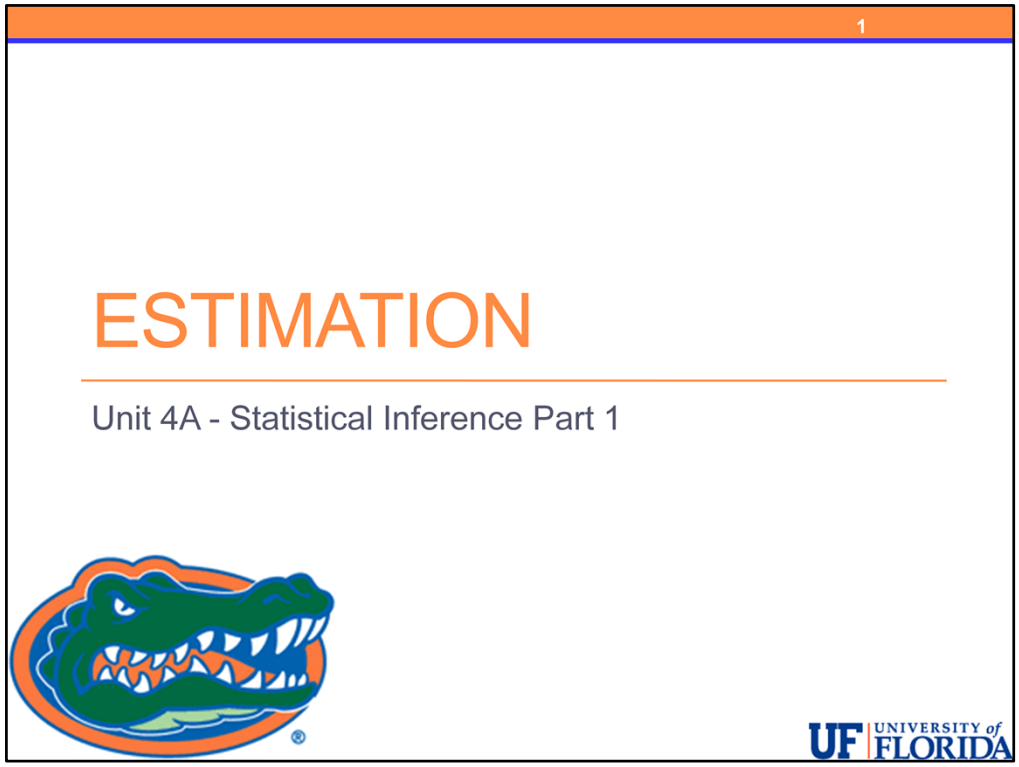
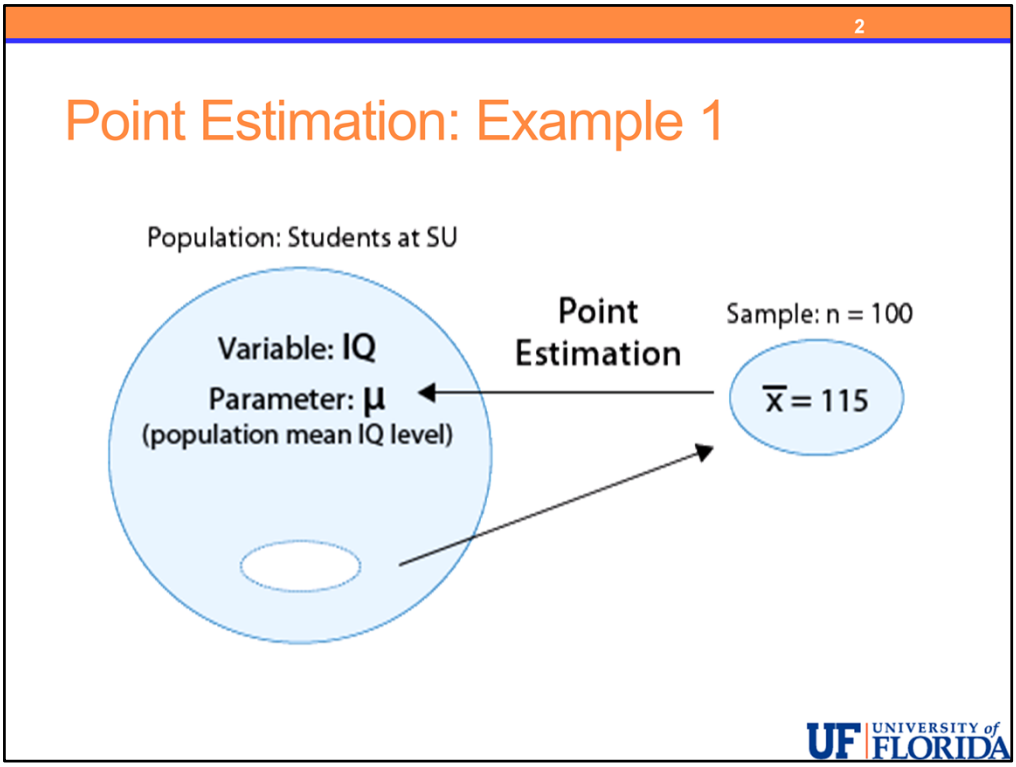


{Estimation}

We begin with a discussion of point estimation where we use a single number to estimate an unknown quantity.

We already know how to find point estimates but in this section we will formalize a few properties of good point estimates and discuss their limitations.



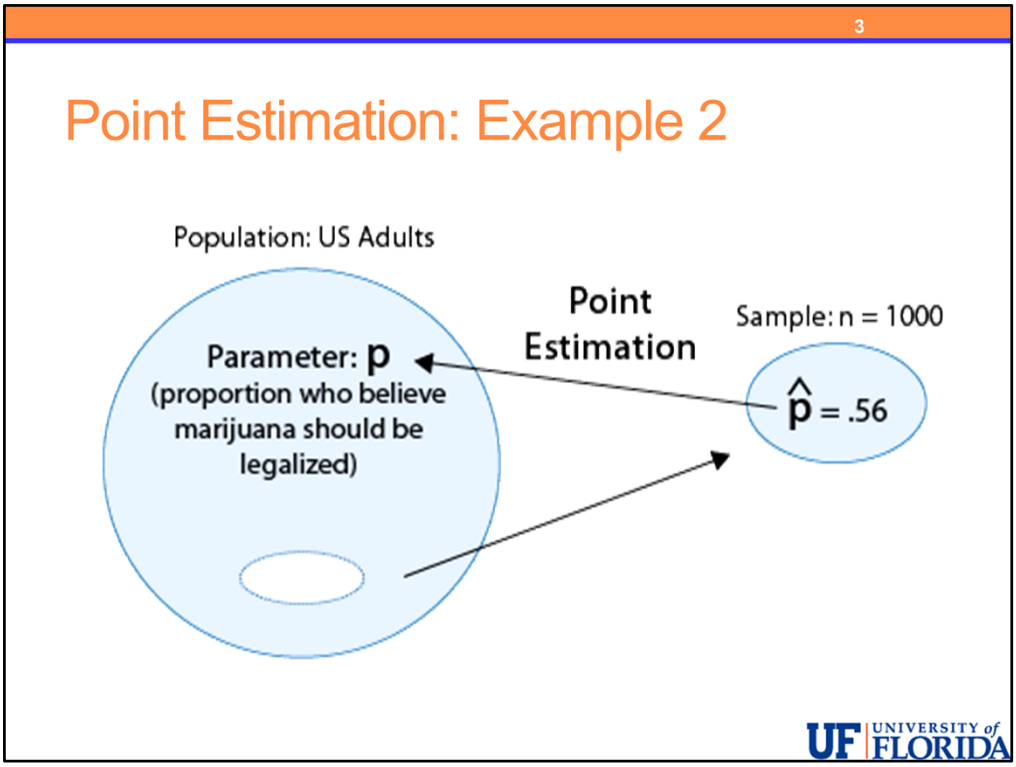
Suppose we are interested in studying the IQ levels of students at Smart University (SU). In particular (since IQ level is a quantitative variable), we are interested in estimating μ, the mean IQ level of all the students at SU. A random sample of 100 SU students is taken and the (sample) mean IQ level (x-bar) was found to be 115.

If we wanted to estimate μ, the population mean IQ level, by a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample mean which is 115

We say that 115 is the **point estimate** for μ, and in general, we'll always use the sample mean (x-bar) as the **point estimator** for μ

Note: when we talk about the **specific** value (115), we use the term **estimate**, and when we talk in general about the **statistic**, the **random variable** x-bar, we use the term **estimator.**

Also, here we are still in a less realistic scenario as we are assuming we know sigma, the population standard deviation. Although for SAT scores this may be realistic, in general we will not have this luxury. We are approaching these problems this way initially so that we can use the normal distribution in our demonstration of these concepts. Later, we will look at how to handle the more realistic situation when sigma is not known and we must approximate it from our data.



If we wish to study the opinions of U.S. adults regarding legalizing the use of marijuana then, in particular, we may be interested in the parameter *p*, the proportion of U.S. adults who believe marijuana should be legalized.

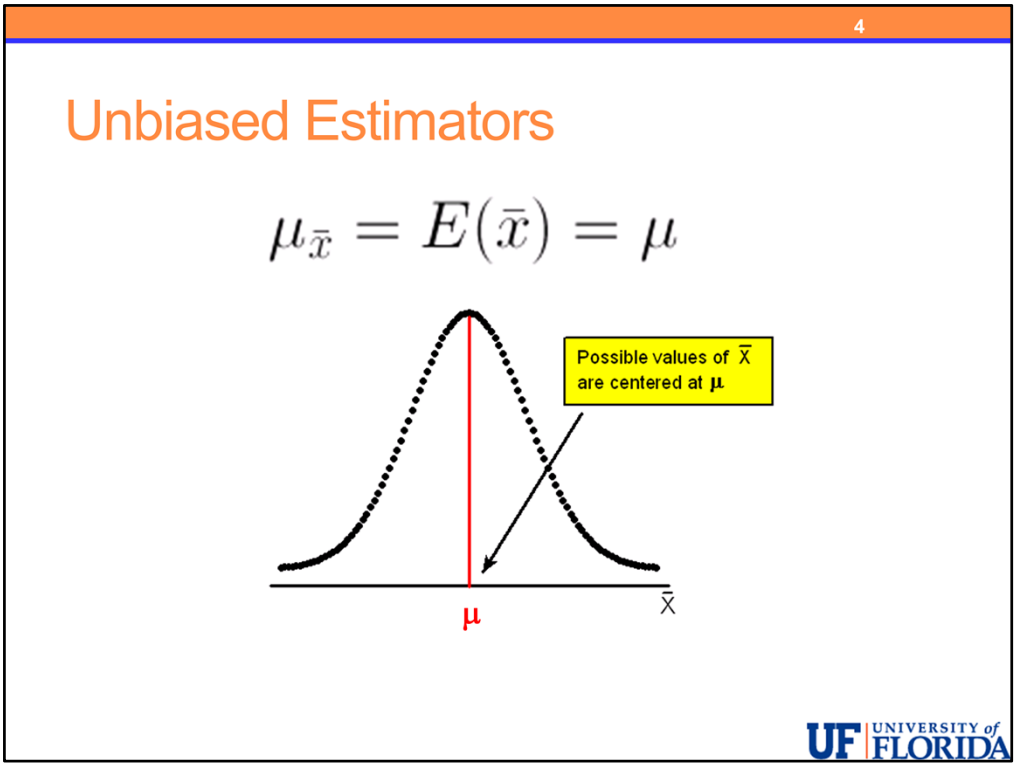
Suppose a poll of 1,000 U.S. adults finds that 560 of them believe marijuana should be legalized.

If we wanted to estimate *p*, the population proportion, using a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample proportion

p-hat = 560/1000 = 0.56

We say in this case that 0.56 is the **point estimate** for *p*, and in general, we'll always use *p*-hat as the **point estimator** for *p*

Note again: when we talk about the **specific value** (0.56), we use the term **estimate**, and when we talk in general about the **statistic** *p*-hat, we use the term **estimator.**



Point estimation is very intuitive, certainly, our intuition tells us that the best estimator for μ should be x-bar, and the best estimator for p should be p-hat

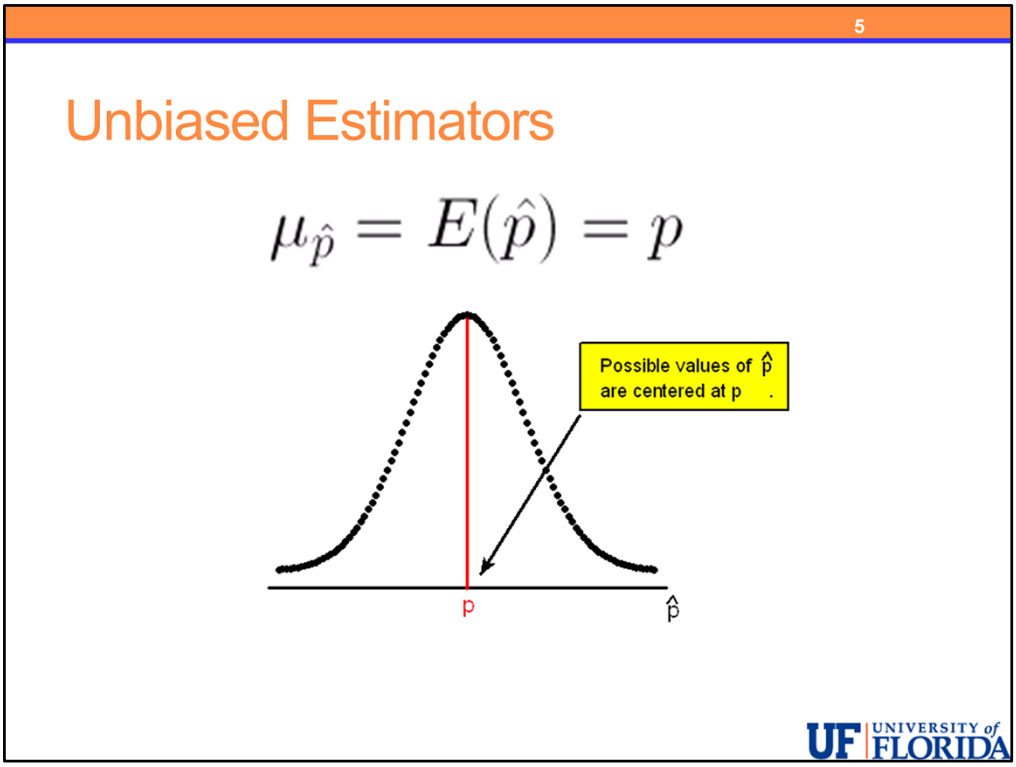
Probability theory does more than this; it actually gives an explanation (beyond intuition) **why** x-bar and p-hat are the good choices as point estimators for μ and p, respectively

From our study of sampling distributions we found that **as long as a sample is taken at random**, the distribution of sample means is exactly centered at the value of population mean

X-bar is therefore said to be an **unbiased** estimator for μ

Any particular sample mean might turn out to be less than the actual population mean, or it might turn out to be more, but in the long run, such sample means are "on target" in that they will not underestimate any more or less often than they overestimate

We have discussed bias a few times from a logical perspective but the true definition of an unbiased estimator is that the **mean**, also called the **expected value**, of the statistic (the mean of the sampling distribution) is equal to the target population parameter.



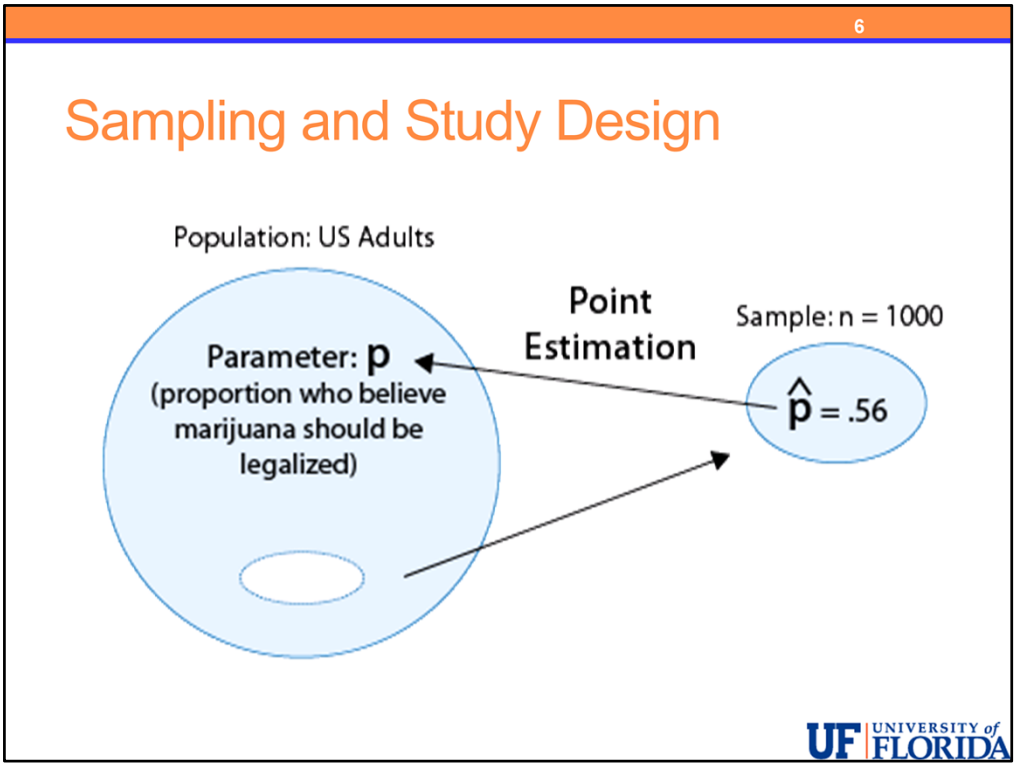
Likewise, we learned that the sampling distribution of the sample proportion, *p*-hat, is centered at the population proportion *p* (as long as the sample is taken at random), thus making *p*-hat an unbiased estimator for *p*.

Probability theory plays an essential role as we establish results for statistical inference. We stated that probability was the foundation and sampling distributions, the bridge, to statistical inference.

Our assertion above that sample mean and sample proportion are unbiased estimators is our first step on that bridge.

The definition of an unbiased estimator is a statistics definition that relies on probability theory.

There are many other examples of this idea. Any parameter I may want to estimate from the population has a sample counterpart that we can study in a similar way.

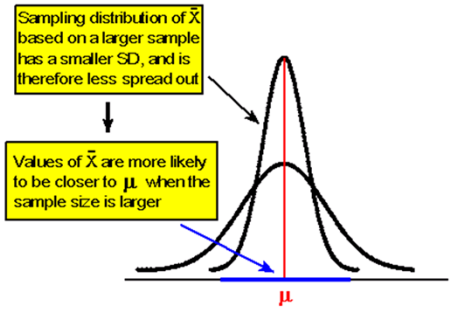


Even though what we assert is true, when the samples are chosen randomly, if, in our example about opinions on legalization of marijuana, the sample of U.S. adults was not random, but instead included predominantly college students, then 0.56 would be a biased estimate for *p*, the proportion of all U.S. adults who believe marijuana should be legalized.

If the survey design were flawed, such as loading the question with a reminder about the dangers of marijuana, or a reminder about the benefits of marijuana for cancer patients, then 0.56 would be biased on the low or high side, respectively.

Our point estimates are truly unbiased estimates for the population parameter **only if the sample is random and the study design is not flawed.**

Sample Size and Standard Error



$$\frac{\sigma}{\sqrt{n}}$$

$$\sqrt{\frac{p(1-p)}{n}}$$

UF UNIVERSITY of FLORIDA

Not only are sample mean and sample proportion on target as long as the samples are random, but they become less variable as the sample size increases, in other words, their precision improves as sample size increases.

We have equations for the standard error in the two cases we are currently considering.

Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

	(Population) Parameter	(Sample) Statistic
Proportion	p	$\hat{p}$
Mean	μ	$\bar{x}$
Standard Deviation	σ	s

UF UNIVERSITY of FLORIDA

Another example of a point estimate is using sample standard deviation (s) to estimate population standard deviation, σ

We will not be concerned with estimating the population standard deviation for its own sake, but since we will often substitute the sample standard deviation (*s*) for σ when standardizing the sample mean, it is worth pointing out that *s* is an unbiased estimator for σ, in fact the reason that we divide by n-1 instead of n is because dividing by n-1 results in an unbiased estimator!

Summary – Point Estimation

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

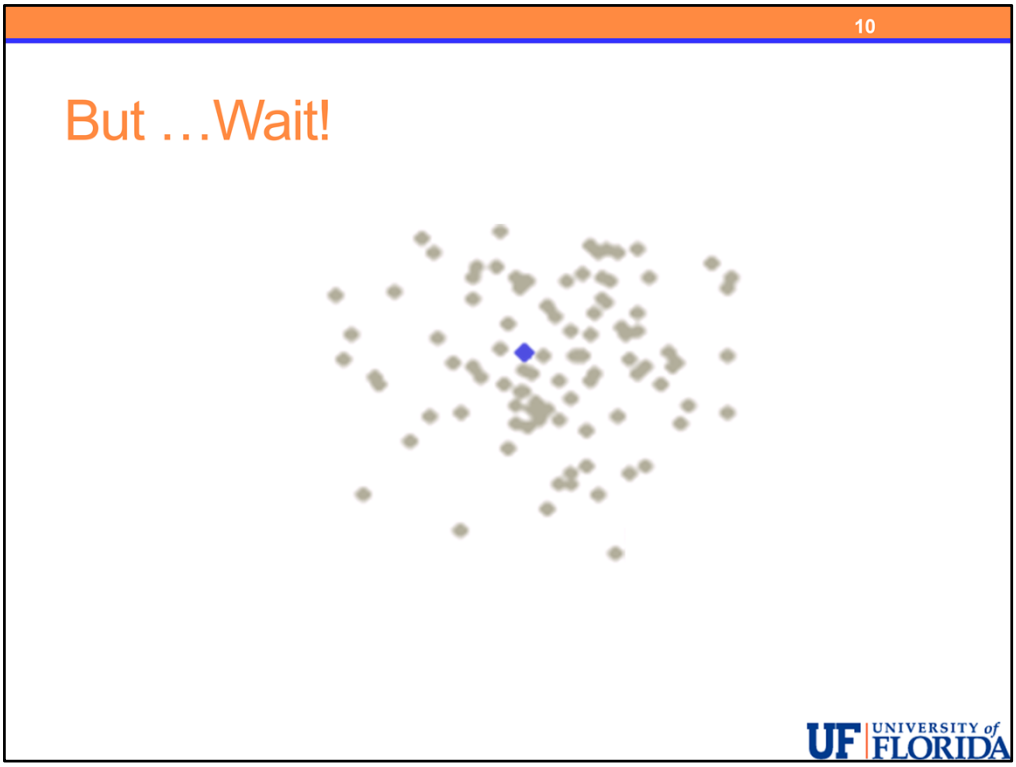
UF UNIVERSITY of FLORIDA

To summarize,

For categorical variables, we use *p*-hat (sample proportion) as a point estimator for *p* (population proportion). It is an unbiased estimator: its long-run distribution is centered at *p* for simple random samples.

For quantitative variables, we use x-bar (sample mean) as a point estimator for μ (population mean). It is an unbiased estimator: its long-run distribution is centered at μ for simple random samples.

In both cases, the larger the sample size, the more precise the point estimator is. In other words, the larger the sample size, the more likely it is that the sample mean (proportion) is close to the unknown population mean (proportion)



When we estimate μ by the sample mean x-bar we are almost guaranteed to make some kind of error!

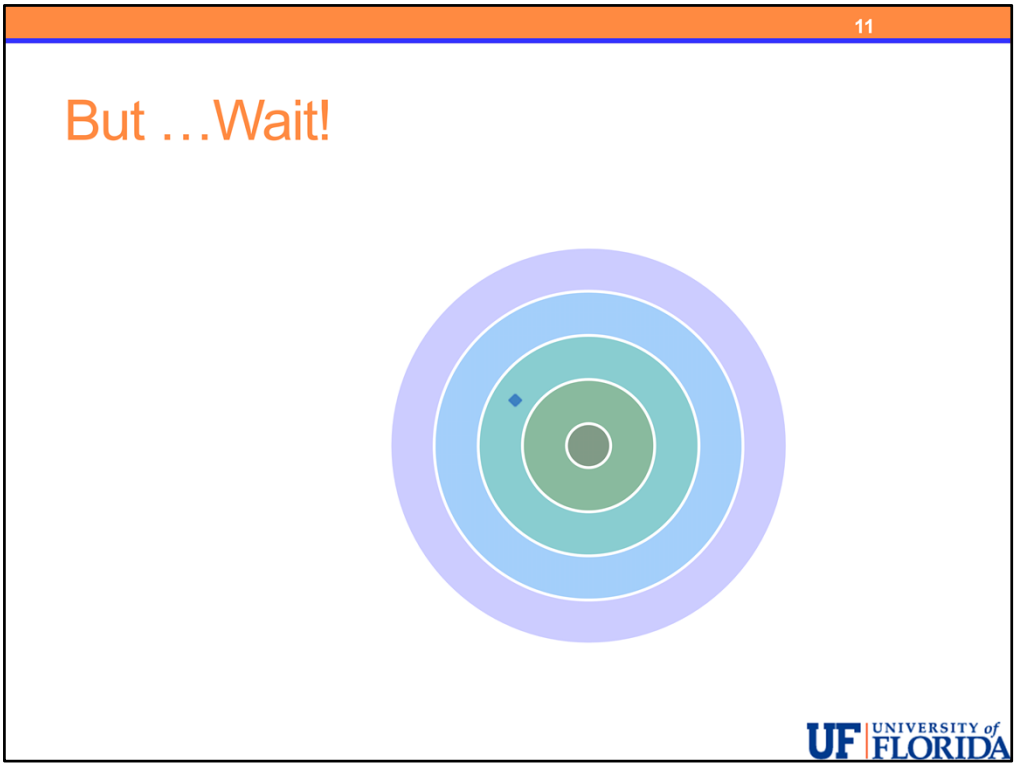
Even though we know that the values of x-bar fall around μ, it is very unlikely that the value of x-bar will fall exactly at μ.

Given that such errors are a fact of life for point estimates, these estimates are in themselves of limited usefulness, unless we are able to quantify the extent of the estimation error.

My favorite analogy for estimation is the reverse game of darts!

The parameter we are trying to "hit" with our estimate is a single number, the tip of a dart on the wall.

Now consider trying to hit the tip of a dart with another dart! That is point estimation! It is very unlikely, no matter how good we are at throwing darts, that we can hit it exactly.



Interval estimation addresses this issue.

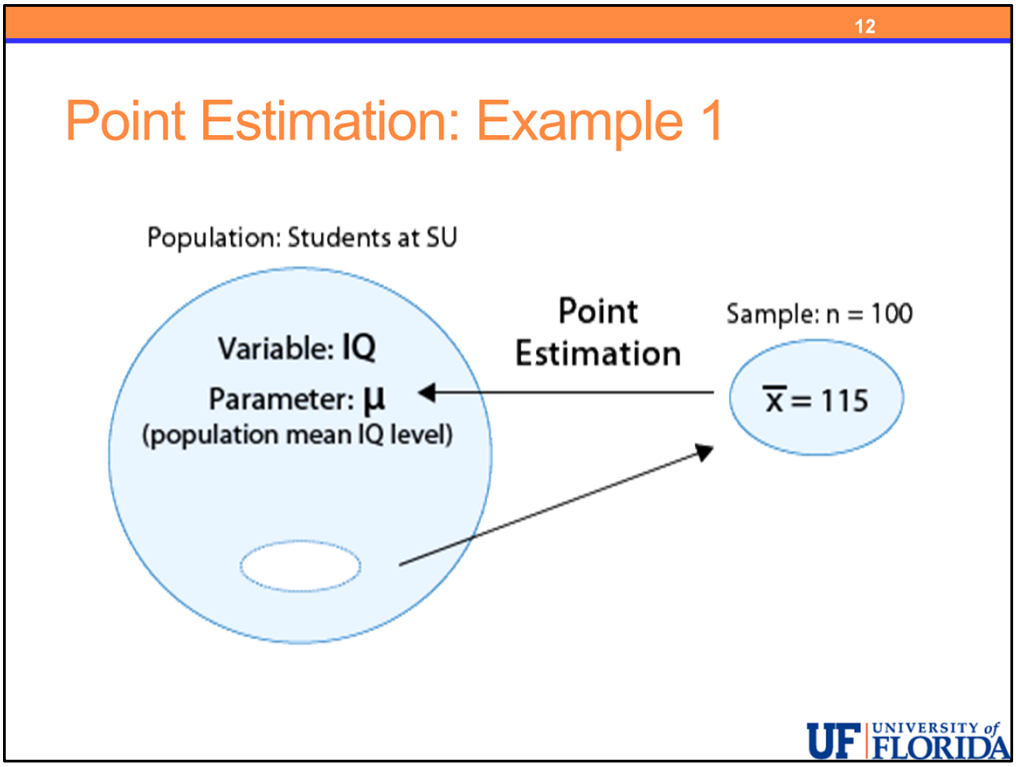
The idea behind **interval estimation** is, therefore, to enhance the simple point estimates by supplying information about the **size of the error attached.**

We want to quantify the potential error in using our estimate from our sample to represent the population value.

From what we know about the sampling distributions of x-bar and p-hat combined with our ability to work with normal distributions, we can construct confidence intervals for the population mean (μ) and the population proportion (p)

Returning to our analogy of the reverse game of darts, now instead of throwing a single dart (my point estimate) at the fixed tip of a dart on the wall (my parameter), I get to throw the whole dart board (my interval estimate). Now we will increase our chances of being able to hit the target.

In many ways, we say statistics is backwards. This analogy illustrates this well as the idea of throwing a dart board at a dart is indeed very backwards!



Our understanding of sampling distributions will let us say things like:

"I am 95% confident that by using the point estimate x-bar = 115 to estimate μ (mu), I am off by no more than 3 IQ points"

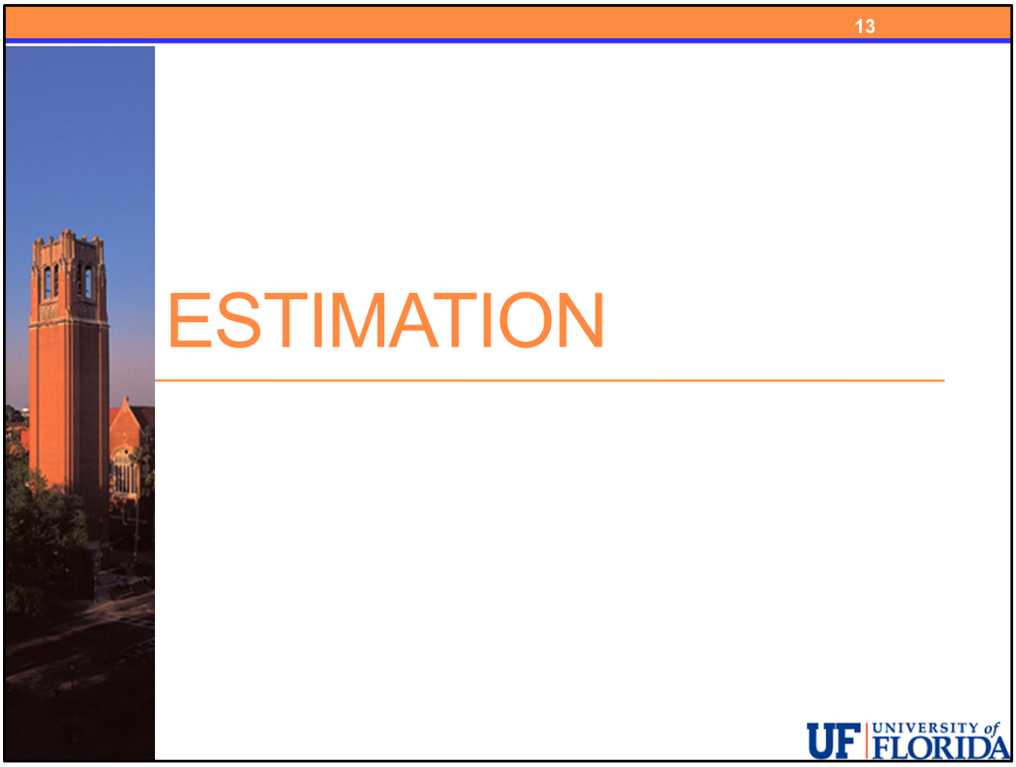
Which could be rephrased as:

"I am 95% confident that μ (mu) is within 3 IQ points of 115 (i.e., between 112 and 118)."

And again as:

"I am 95% confident that μ (mu) is somewhere in (or covered by) the interval (112, 118).

Soon, we will discuss how these intervals are created.



Estimation is an important aspect of statistical inference. Even when we are conducting hypothesis tests, there will still be the need for estimation.

We have discussed point estimates and the desired properties of being unbiased and less variable.

We have discussed the drawbacks of point estimates and introduced the idea of interval estimation.

In the next section we will outline the process of creating and interpreting confidence intervals.

We will see in the later modules that confidence intervals are useful whenever we wish to use data to estimate an unknown population parameter, even when this parameter is estimated using multiple variables (such as our cases: CC, CQ, QQ)

For example, we can construct confidence intervals for the slope of a regression equation or the correlation coefficient.

In doing so we are always using our data to provide an interval estimate for an unknown population parameter (the TRUE slope, or the TRUE correlation coefficient).