


1

UNIT 4A


Statistical Inference
Part 1



Now we know more about
how **statistics** *estimated*
from **samples** *vary*

SO ...

How does this help us
learn about *unknown*
population parameters?



{Unit 4a - Statistical Inference Part 1}

Now we begin our final look into the underlying theory of statistical methods with a discussion of statistical inference in the simple cases of one population proportion or one population mean.

Our discussions will still be more theoretical in nature than practical in nature but now from a statistics perspective rather than a probability perspective. In the section on sampling distributions, the results relied on the concepts of random sampling, probability, and independence. In this section we will see that important statistical terms such as bias and variation rely on sampling distributions for their definitions.

The more you understand the concept of sampling distributions, the more the underlying concepts of statistical inference will make sense to you.

Our motivating question for this unit is:

Now that we know more about how **statistics** *estimated* from **samples** *vary*, how does this help us learn about *unknown* **population parameters?**

Statistical Inference

**DRAWING CONCLUSIONS
ABOUT A POPULATION
BASED (ONLY) ON THE DATA
OBTAINED FROM ONE
PROBABILITY SAMPLE
CHOSEN FROM
THAT POPULATION**

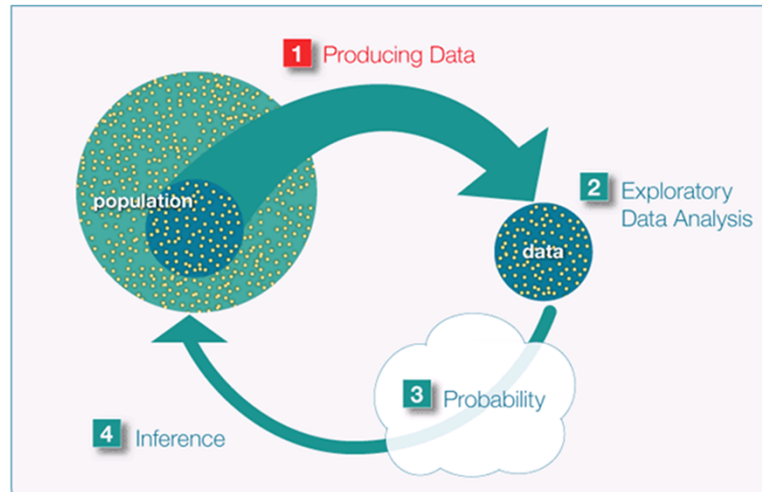


Statistical inference is: drawing conclusions about a population based (only) on the data obtained from one probability sample chosen from that population.

Notice this is the opposite of the approach taken for sampling distributions where we KNEW the truth in the population and we could then learn about the behavior of certain statistics calculated from simple random samples from that KNOWN population.

Here our goal is to do our best to learn about the population based only upon the information from our single sample.

Outline of the Process of Inference



At the beginning of the semester, we introduced the big picture of statistical inference. We have looked at it a few times during the semester.

In step 1, we produce our data.

In Unit 2, we introduced you to some of the principles of sampling and study design. For the inferential methods we will learn, we are assuming a simple random sample of independent observations from the population of interest. For complex sampling plans or dependent observations, some adjustments to methods may be necessary and may provide estimates with less bias or smaller variability. These principles are important to keep in mind as the results we obtain will not hold if there were flaws in our sampling plan or study design.

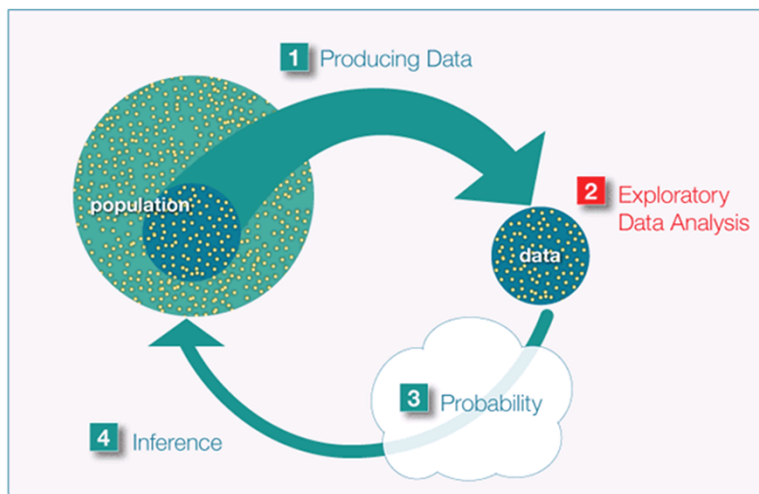
Once we have collected our data, there is little we can do at that point to correct any inherent flaws in the data!

We also learned that randomization is essential, not only because it prevents bias, but also because it permits us to rely on the laws of probability, the scientific study of random behavior.

We use randomization when we

- Take simple random samples (or use other probability sampling plans)
- Randomize subjects into treatment groups

Outline of the Process of Inference



In step 2, we conduct exploratory data analysis on our sample data.

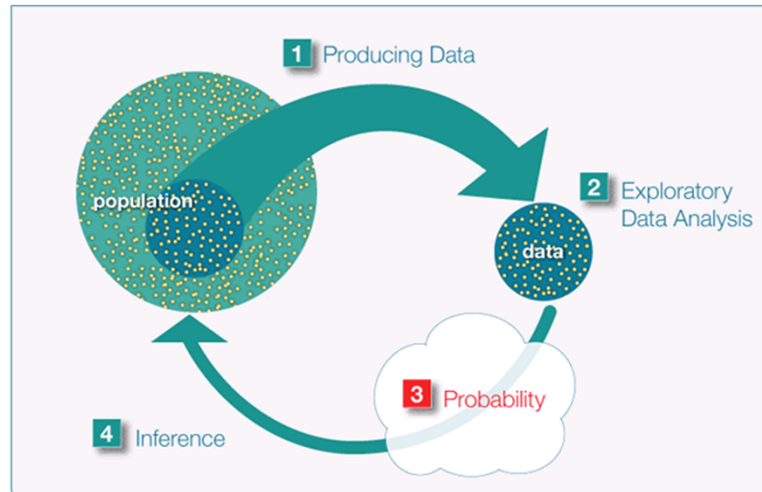
In Unit 1, we learned the exploratory data analysis methods for situations involving one categorical variable or one quantitative variable where we were interested in examining and describing the distribution of our single variable. Then we looked at exploratory data analysis in the three main two-variable cases CC, CQ, and QQ, where we were interested in examining and describing the relationship between two variables.

Once again, we will not formally address analyzing research questions involving case QC in this course, however, to assess the **association** between variables in case QC, we are still able to use methods developed for case CQ. We will review exploratory data analysis in each of these scenarios as we return to them to add statistical inference components.

Regardless of whether we had one or two variables, when we looked at exploratory data analysis, we did not attempt to make claims about the larger population from which the data were obtained even though such generalizations were a long-term goal of the course.

Hopefully you have a good understanding of exploratory data analysis as it is an important and useful skill for its own sake and it continues to be crucial when conducting more advanced statistical analysis.

Outline of the Process of Inference



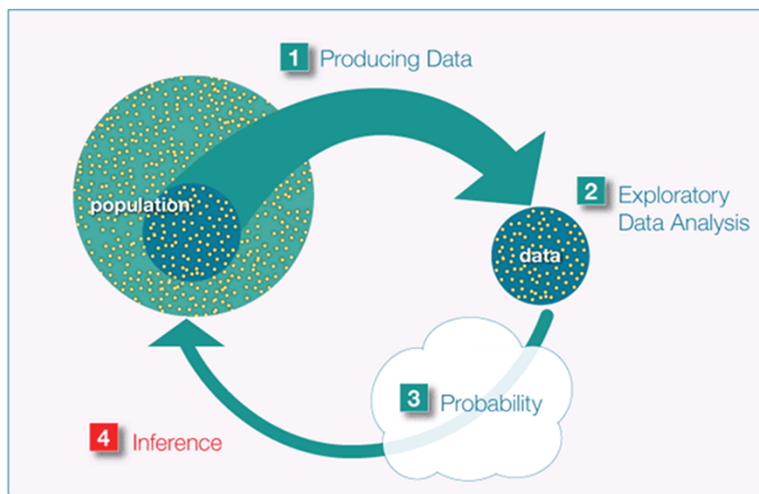
In step 3, we use probability – how?

In Unit 3, we learned about basic probability, probability distributions, and sampling distributions. Of these topics, the last is what we will need for statistical inference. Although not related to statistical inference directly, we do need to understand basic probability.

We learned to use data to provide empirical estimates of unknown probabilities which can be useful in many practical problems. And clearly, in order to understand sampling distributions, we needed to understand probability distributions, which in turn required our understanding of basic probability calculations. We will elaborate more on how sampling distributions are used in the next two topics of estimation and hypothesis testing.

In our discussion of sampling distributions, we focused on two statistics, the sample mean (\bar{x}) and the sample proportion (\hat{p}). We studied the behavior of the center and spread of these distributions and the conditions under which the sampling distribution of these two statistics will be normally distributed.

Outline of the Process of Inference



In step 4, where we finally are now, we will learn the theoretical mechanisms used in statistical inference and then we will end the semester by discussing inferential methods for relationships between two variables.

Let's go through the outline of the process of statistical inference as we will be presenting it in our two simple examples regarding the population mean and population proportion.

- A simple random sample is taken from a population of interest. (This is step 1 in the big picture)
- (Step 2 in the big picture) In order to estimate a population parameter, a statistic is calculated from the sample. For example we might calculate:
 - The Sample Mean or
 - The Sample Proportion
- (Step 3 – the cloud of probability!) We then learn about the DISTRIBUTION of this statistic in repeated sampling (theoretically). We defined the distribution of a sample statistic as the sampling distribution of the statistic.
- (Step 4) Using THIS sampling distribution we can make inferences about our population parameter based upon our sample statistic.

Our next goal is to discuss how this last step works in the two simple cases of one

population proportion and one population mean.

In real-life situations all we know is what we see in the sample, and we want to use this information to say something concrete about the larger population

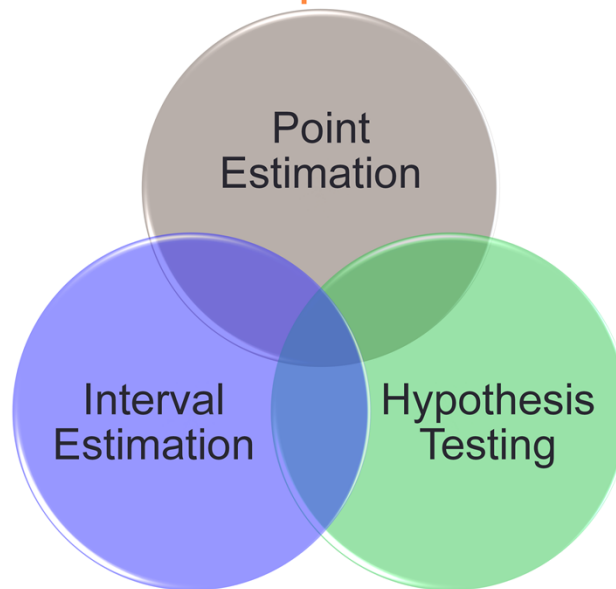
Probability theory has set the stage to accomplish this: learning what to expect from the value of sample mean, given that population mean takes a certain value, teaches us (as we'll soon learn) what to expect from the value of the unknown population mean, given that a particular value of sample mean has been observed.

If I know the sample mean and sample size, we will know something about how far away from the sample mean the fixed value of the true population mean might be.

This process of working backwards, inferring something about the population based on what is measured in the sample — is **statistical inference**, and it is very backwards! Always keep this in mind as you go through the rest of the course.

In practice, we have unknown population parameters which are of interest to us, however, the statistics that are computed from the sample vary from sample to sample due to sampling variability. Therefore we need know how much they vary before we can have any idea what the result from our sample can tell us.

Main Inference Topics



We will define and discuss three main forms of inference.

Point estimation, interval estimation, and hypothesis testing.

In Point Estimation we estimate an unknown parameter using a **single number** that is calculated from the sample data. This, we already know how to do!!

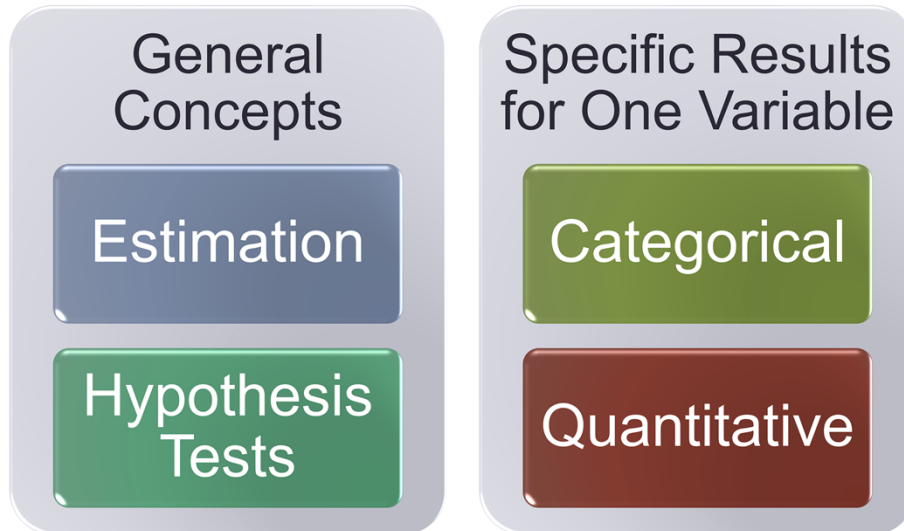
In Interval Estimation we estimate an unknown parameter using an **interval of values** that is likely to contain the true value of that parameter (and state how confident we are that this interval indeed captures the true value of the parameter)

In Hypothesis Testing: we begin with a claim about the population, and we check **whether or not the data** obtained from the sample **provide evidence AGAINST this claim**.

Interval estimation and hypothesis testing methods often require assumptions about the population distribution.

Methods that require no distributional assumptions are called non-parametric or distribution free. We will discuss some of these alternative methods later in the semester.

Current Focus



In Unit 4A we want you to:

Learn the general concepts of estimation and hypothesis testing which are applicable to all future inferential methods in this course and many others.

We will present the complete theory of the big picture in the simple scenario of one variable problems then we will focus on application of methods and minimize discussion of the underlying theory.

Just as in Unit 1 when we were examining distributions, the first thing we must determine is whether the variable is categorical or quantitative as this will determine which methods we apply.

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

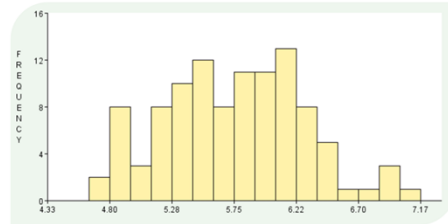
Goal: Use **Sample Proportion** (*known*) to estimate **Population proportion** (*unknown*)

When the variable of interest is **categorical**, the population parameter that we will infer about is the **population proportion (p)** associated with that variable

Our Goal: to use the **Sample Proportion** (which is *known*) to estimate the **Population proportion** (which is *unknown*)

Here again we have examples from the Framingham data related to diabetes and BMI categories where we might be interested in estimating or testing hypotheses about the true proportion with diabetes or the true proportion within each BMI category in the population based upon the results of this sample.

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

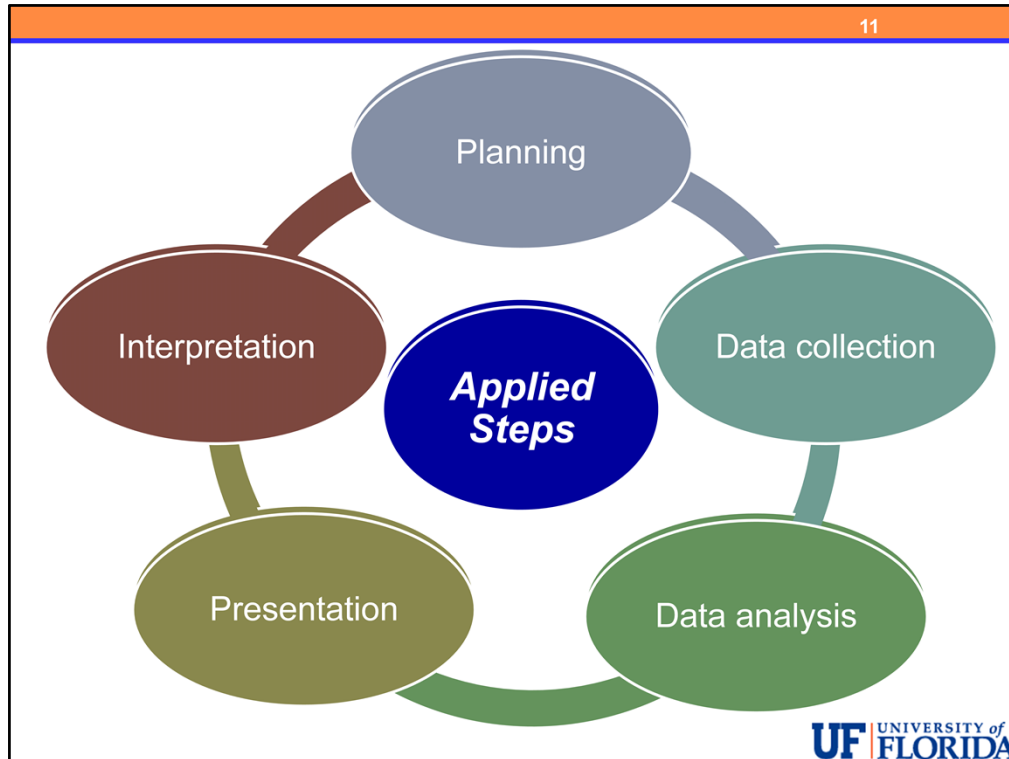
Goal: Use Sample **Statistic** (*known*) to estimate Population **Parameter** (*unknown*)

When the variable of interest is **quantitative**, the population parameter that we infer about is the **population mean (μ)** associated with that variable

Although there are many other aspects we could consider, we will focus on the center of the distribution as measured by the population mean.

Our Goal: to use the **Sample Mean** (which is *known*) to estimate the **Population Mean** (which is *unknown*)

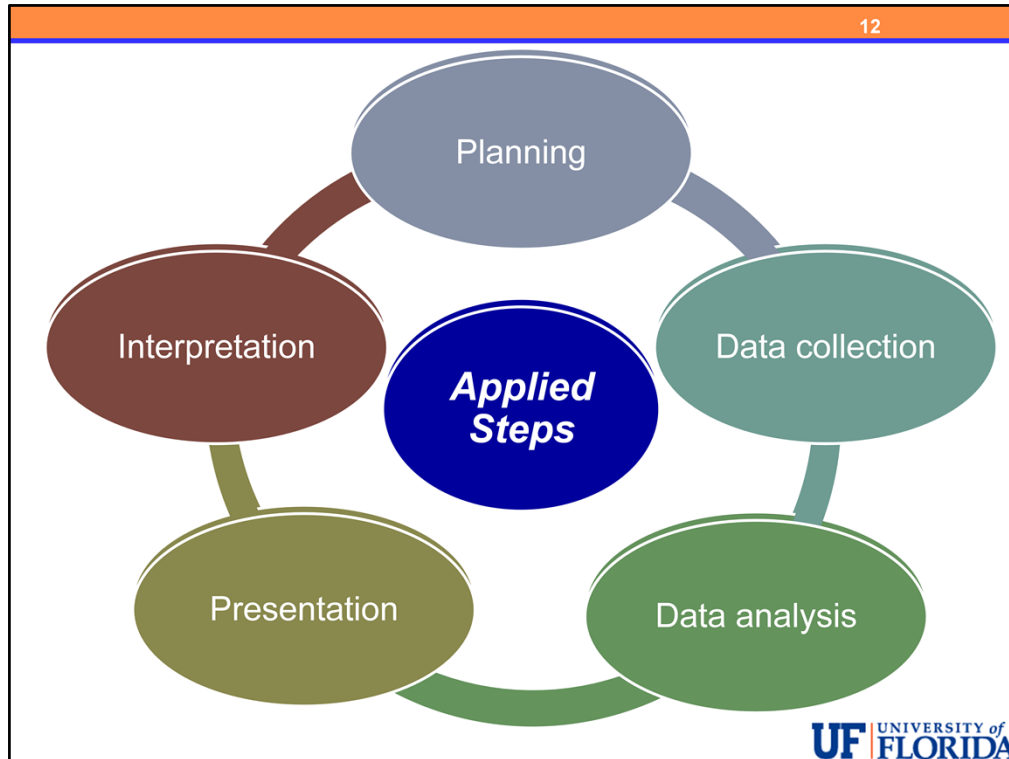
For example, here we have the acidity of rainwater in a sample of 105 measurements. We might be interested in using this sample to estimate or test hypotheses about the population mean acidity of rainwater.



Note that the theoretical process of statistical inference is only a small part of the applied steps in a research project. Previously, in our discussion of the role of biostatistics, we defined these steps to be:

- Planning
- Data collection
- Data analysis
- Presentation
- Interpretation

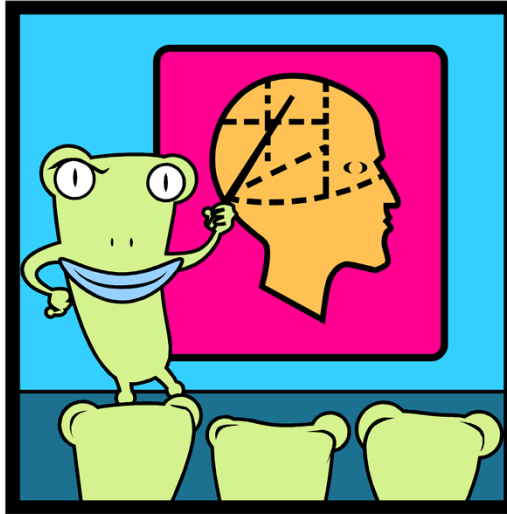
You can see that, in effect, both exploratory data analysis and inferential methods will fall into the category of "Data Analysis" in this list.



Probability isn't evident in our steps in a research process. Where is it hiding?

- In “planning” through the determination of significance levels, confidence levels, and desired power for important tests – topics we will soon discuss.
- In “data collection” in the form of probability samples (such as simple random sampling, cluster sampling, stratified sampling).
- In “data analysis” in the form of sampling distributions and possibly the estimation of probabilities of interest.
- In “presentation” and “interpretation” in the form of p-values and confidence intervals (which we will also discuss shortly) and in certain situations, we will provide and interpret estimated probabilities of interest.

How do Researchers Think?



- Population?
- Question/Problem?
- How to sample?
- How to analyze data?
- How to report results?

AFTER know what to do

THEN collect data

For researchers, the following represent some of the important questions to address when planning a study.

What is the population of interest?

What is the question or statistical problem?

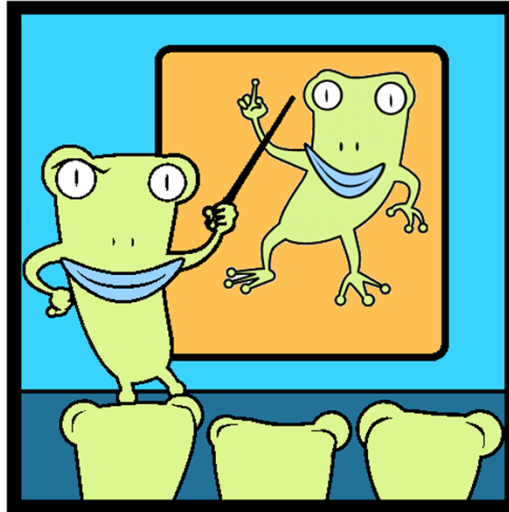
How should we sample to best address the question (given the available resources)?

How will we analyze the data?

How will we report the results?

AFTER we know what we are going to do, then we can begin collecting data!

How do Statisticians Think?



- Assumptions?
- Parameter?
- Statistic?
- Behavior?
 - Unbiased?
 - Variability?
 - Sampling Distribution?

Statisticians, on the other hand, need to ask questions like these:

What assumptions can be reasonably made about the population?

What parameter(s) in the population do we need to estimate in order to address the research question?

What statistic(s) from our sample data can be used to estimate the unknown parameter(s)?

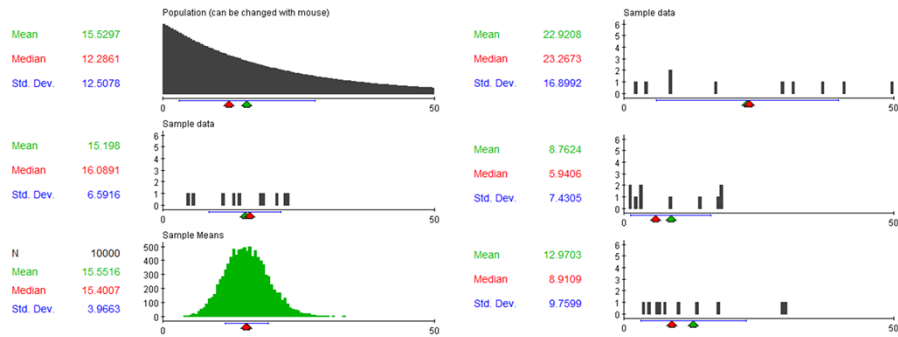
How does each statistic we will use behave?

Is it unbiased?

How variable will it be for the planned sample size?

What is the distribution of this statistic? (This is it's Sampling Distribution)

Standard Error of a Statistic



We have seen that:

Each time we take a sample, we obtain a different value for our statistic and thus statistics are themselves random variables and also have a distribution.

The distribution of a statistic (from a sample of a given sample size) is called the sampling distribution of the statistic.

The standard error of a particular statistic is defined as the standard deviation of the sampling distribution of that statistic. In the pictures shown here we have the sampling distribution of the sample mean from samples of $n = 10$ from this skewed right population. The simulation estimates the standard error to be around 3.97.

The standard error is the most common measure of the variability of a sample statistic.

Standard Error of a Statistic

Standard Error of p-hat

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

Standard Error of x-bar

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

We verified the equations for the standard error of x-bar and the standard error of p-hat in our section on sampling distributions.

Be sure to understand when to apply each of these results.

We use the standard error of p-hat when we have situations involving a sample proportion and the standard error of x-bar when we have situations involving a sample mean.

Notice both of these equations will require us to know something about some of the population or to estimate certain population parameters.

In the case of the standard error of p-hat, we need to estimate the population proportion, p . For the standard error of x-bar, we need to know or estimate the population standard deviation, σ .



STATISTICAL INFERENCE PART 1

Although we will only go through the theoretical derivation of confidence intervals and hypothesis tests for the two simple cases of one proportion or one mean, these ideas are underlying most inferential statistical methods used in practice.

We will see that confidence intervals and hypothesis tests are applicable in more advanced analyses but in such cases we will rely on software to perform the calculations.

Next we will formalize the concept of point estimates and confidence intervals.