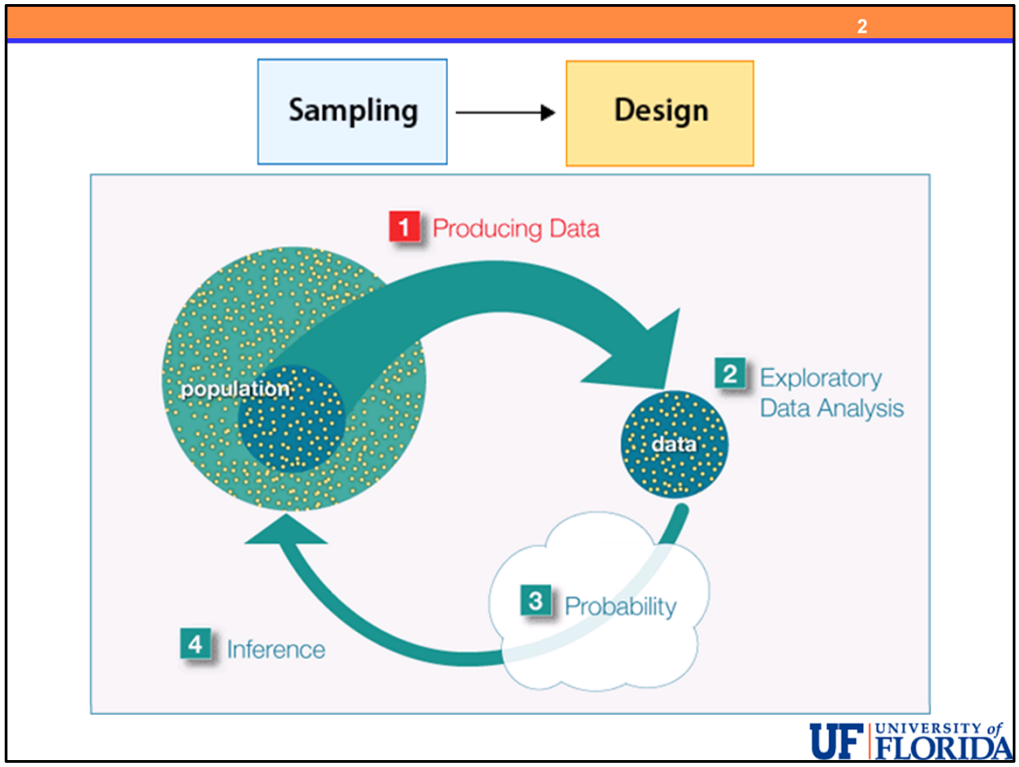
INTRODUCTION

Unit 2: Producing Data

UF UNIVERSITY of FLORIDA

Now that we have discussed exploratory data analysis, we take a step backwards and discuss some concepts and issues related to producing data.



We go back to Step 1 in the Big Picture of Statistical Inference.

There are two components in producing data.

The first stage in this process is **sampling** and the second stage is **study design**.

For the final part of the big picture, inference, to work reliably, it is important to have a representative sample from the population of interest.

Therefore, it is important to ensure that the sampling procedure will produce a sample that represents the population.

Having a representative **sample**, is only the first step. In order to answer our research questions effectively, we must also consider the **design** for producing data from our sample.

In particular, if you wish to show a causal relationship between variables, the design of the study should be carefully planned.

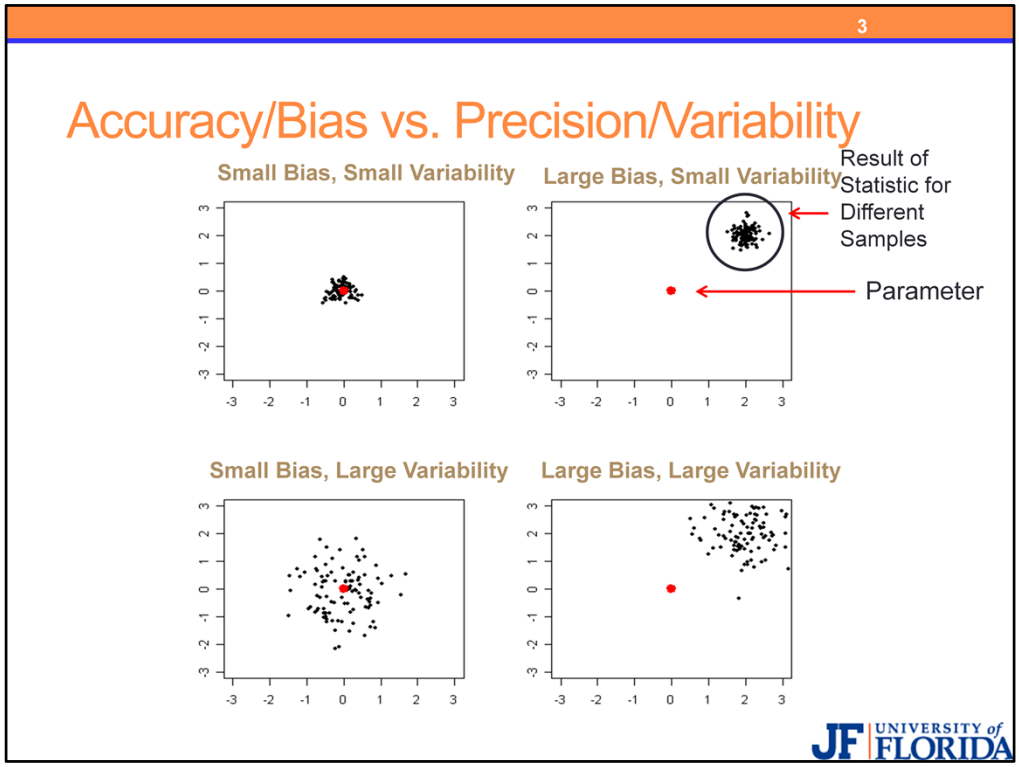
In this unit we will establish guidelines for the IDEAL production of data.

However, in reality, it is rarely possible to reach this ideal.

As researchers and reviewers of the research of others, you must also use common sense to evaluate which imperfections are acceptable and which may completely undermine a studies results.

You should consider carefully the admitted limitations in research studies. Many of these will be related to the sampling method or the study design.

When we produce data that is not representative due to systematic over-estimation or under-estimation of the values of interest we have **biased** results. Bias can result from a poor sampling plan or a poor study design. Thus we may sometimes hear the terms “biased sample,” “biased design,” “biased results,” and later we will talk about biased estimators of certain quantities of interest.



Let's elaborate a little further on the concept of **bias** (which is our **accuracy** at estimating a quantity of interest) and **variability** which is how **close** our estimates will be to **each other** if we **repeat** the **process**.

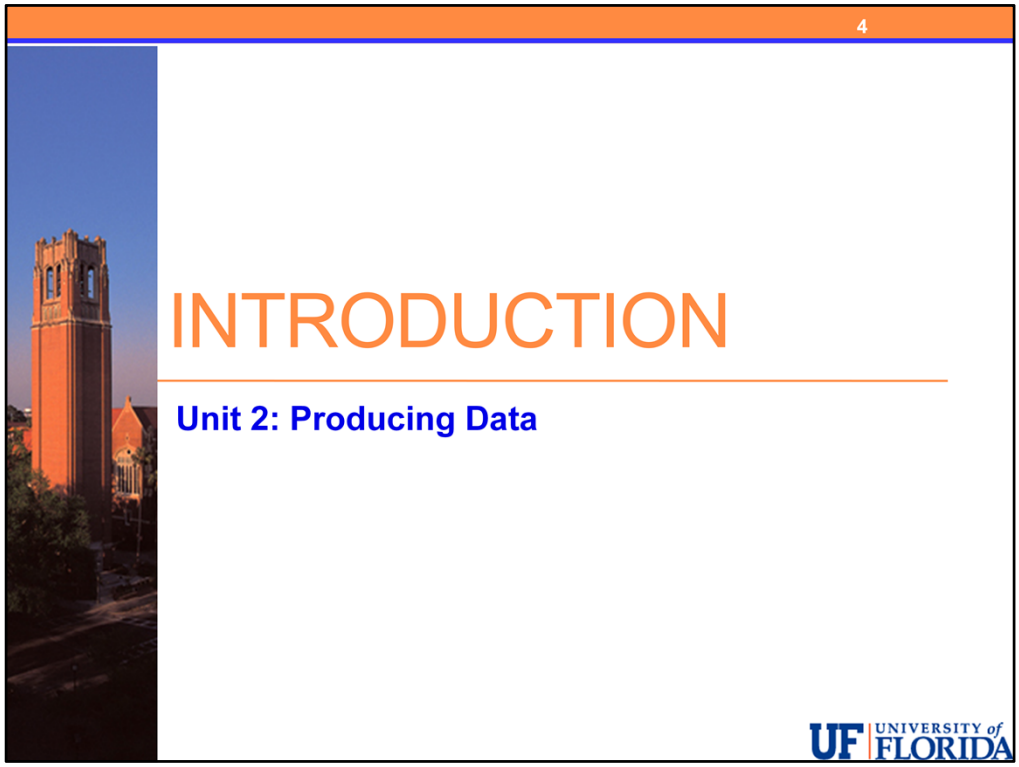
We will talk more about the concept of variability in Unit 3 but at this point we wish to point out that bias results in an estimator - produced by our data or process - which is inherently flawed in that it will miss the target in a systematic way.

In these displays, the idea is that the graphs on the left have no clear bias. The cloud of points representing what could happen if we were to repeat the process, is centered around the target (red dot).

The graph on the bottom shows an estimator which has more variation, but it is still unbiased – in that it is no more likely to over-estimate than under-estimate the quantity of interest.

The graphs on the right, however, do not hit the target at all and thus represent biased results.

If we know exactly how the bias will occur, it may be possible to fix it, like readjusting the sights on a rifle but if we cannot say EXACTLY how the results will be biased then there is little hope of recovering useful information from the data.



This unit contains many definitions and concepts. Our goal is not to teach you everything about these topics but to introduce you to these ideas.

If you conduct your own research in the future, you will need to spend time reviewing in more detail those topics which most apply to your situation.

For example, if you plan to conduct clinical trials, you will need to learn much more about sampling and design concerns specifically for clinical trials, yet survey design may never be extremely important to you. However, if you conduct research in which surveys are the primary method of acquiring data then you will care a great deal indeed about survey design and methodology.

Each of these topics could entail one or even more additional courses, just on that specific topic – clinical trials or survey methodology and there are many other possibilities.

The stage of producing data is perhaps the most critical stage of the research process since, after the data are collected, there is often little hope of repairing any problems that exist in our plan or process. Sometimes, the problems are unforeseen – in any case – the resulting data from a flawed process may be completely useless once it is collected – or worse – incorrect conclusion may be drawn which have a negative impact on patients or policy.