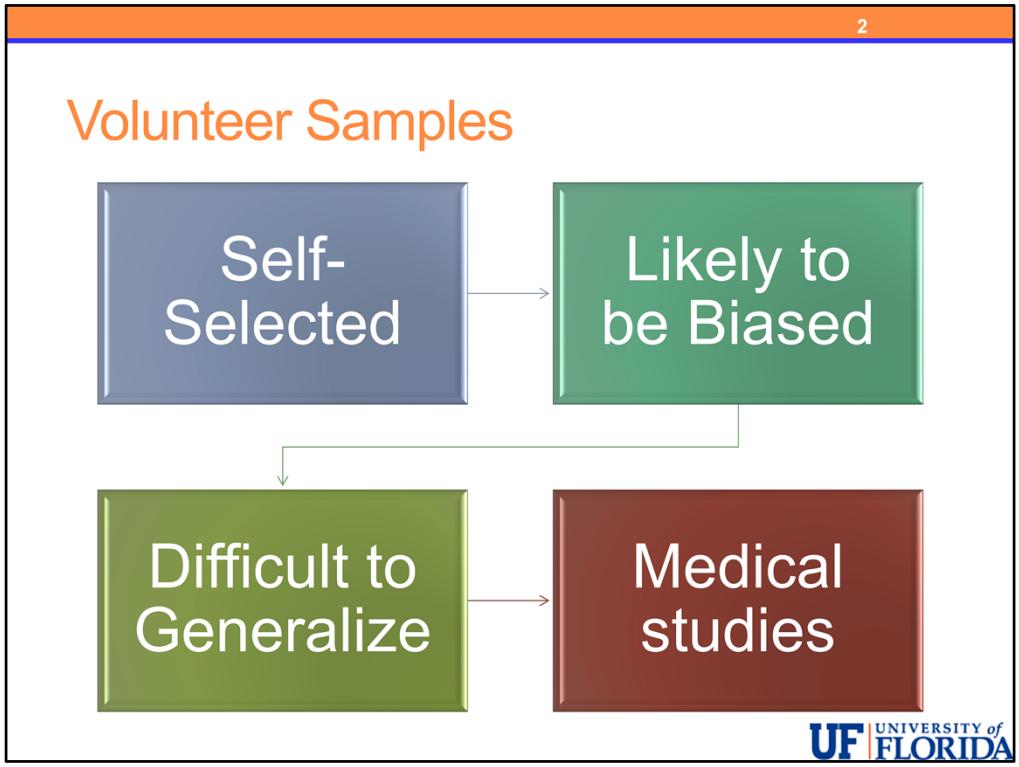
SAMPLING PLANS

Unit 2: Producing Data

UF UNIVERSITY of FLORIDA

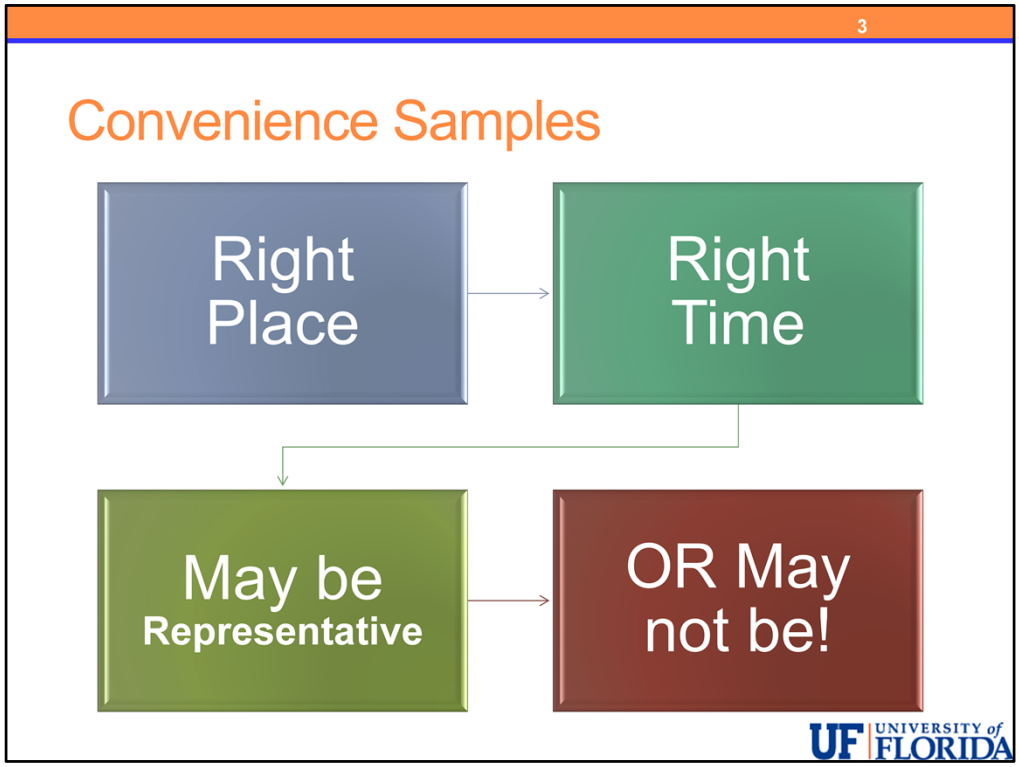
Now we will define some common sampling plans and discuss their strengths and limitations.



For volunteer samples individuals are self-selected. Participants decide to include themselves in the study. Common examples would be internet surveys, restaurant rating cards, call-in surveys, etc.

These types of samples are almost guaranteed to be biased and hence we cannot generalize the results to a larger group.

One should realize that volunteer samples are used often in medical studies as individuals must agree to participate – voluntarily – and we will see that this is not necessarily problematic in an experiment designed to compare treatments.



Individuals in a convenience sample happen to be at the right place at the right time to suit the researcher's schedule.

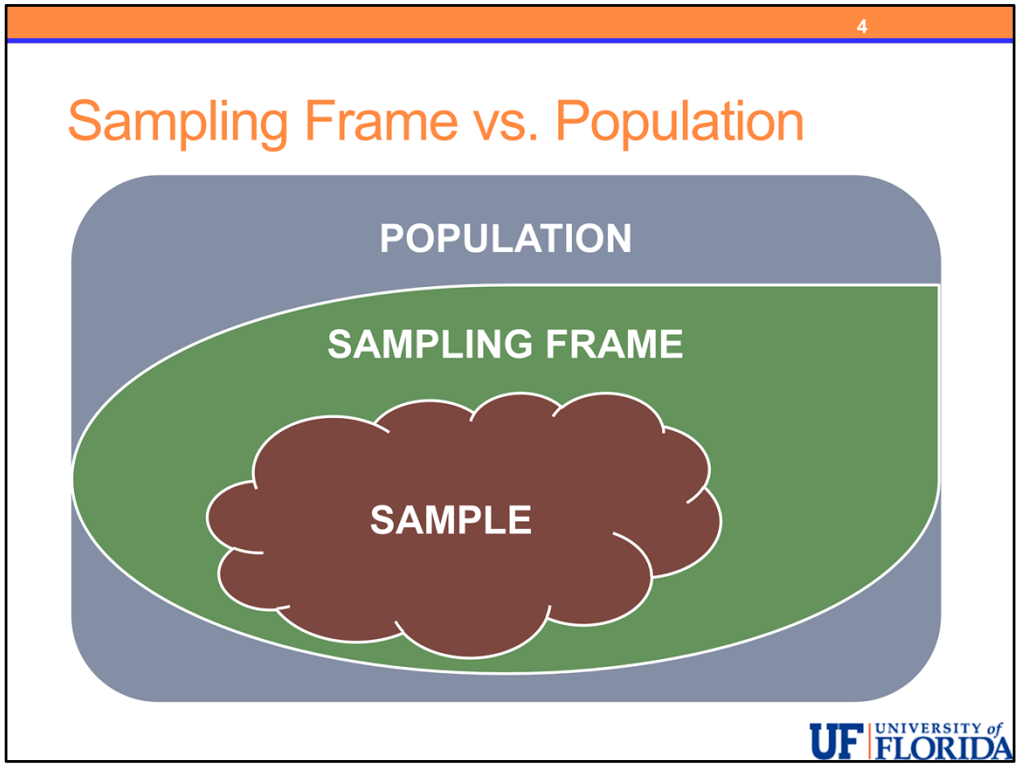
Depending on the questions under study, such a sample may or may not be representative of the larger population of interest.

A convenience sample may be susceptible to bias because certain types of individuals are more likely to be selected than others and in extreme cases some individuals may have no chance of being selected.

Suppose we use a student organization which is affiliated with a particular political party as our sample.

If we wish to study heights and weights of college students, this convenience sample may still provide a relatively representative sample, even though students who are not members of this organization have no chance to be selected.

However, if we wish to study opinions of college students on a social topic, you would likely not obtain a representative sample by selecting all students in a student organization that is affiliated with a particular political party. In this case, the fact that a particular party may have a particular stance on social topics could very well produce results which are not representative of the larger student population.



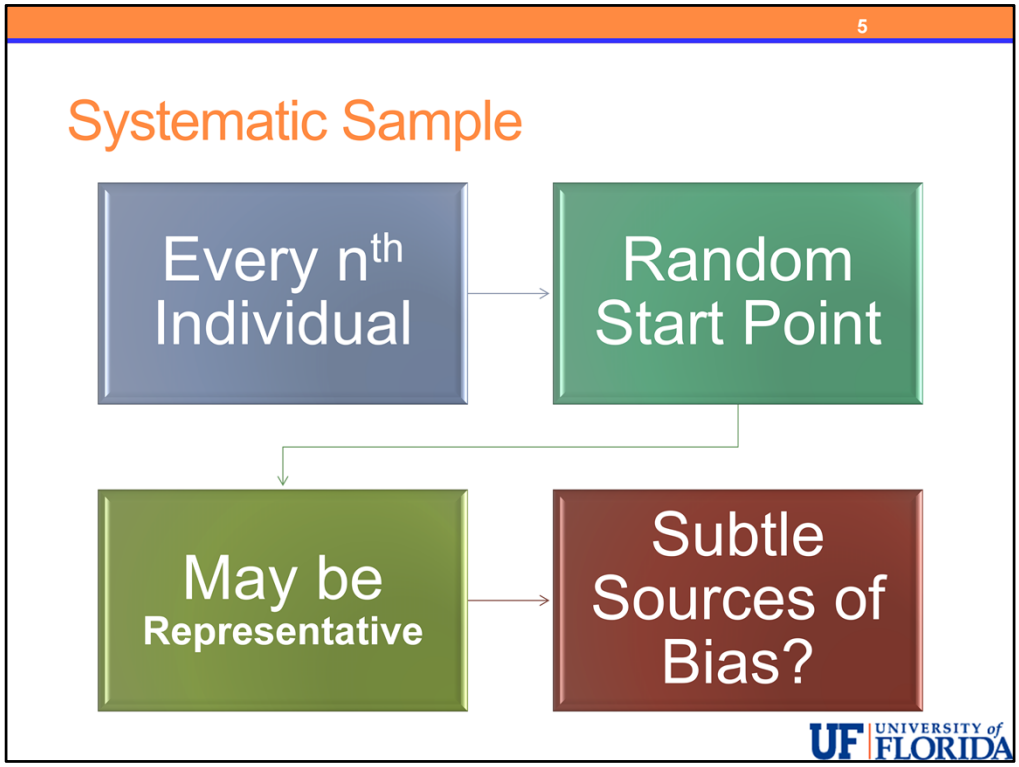
We define the **sampling frame** as the list of potential individuals to be sampled.

Population = all individuals of interest.

If the sampling frame is not the same as the population, there may be **bias** arising because of this discrepancy.

It is always best to have the sampling frame match the population as closely as possible.

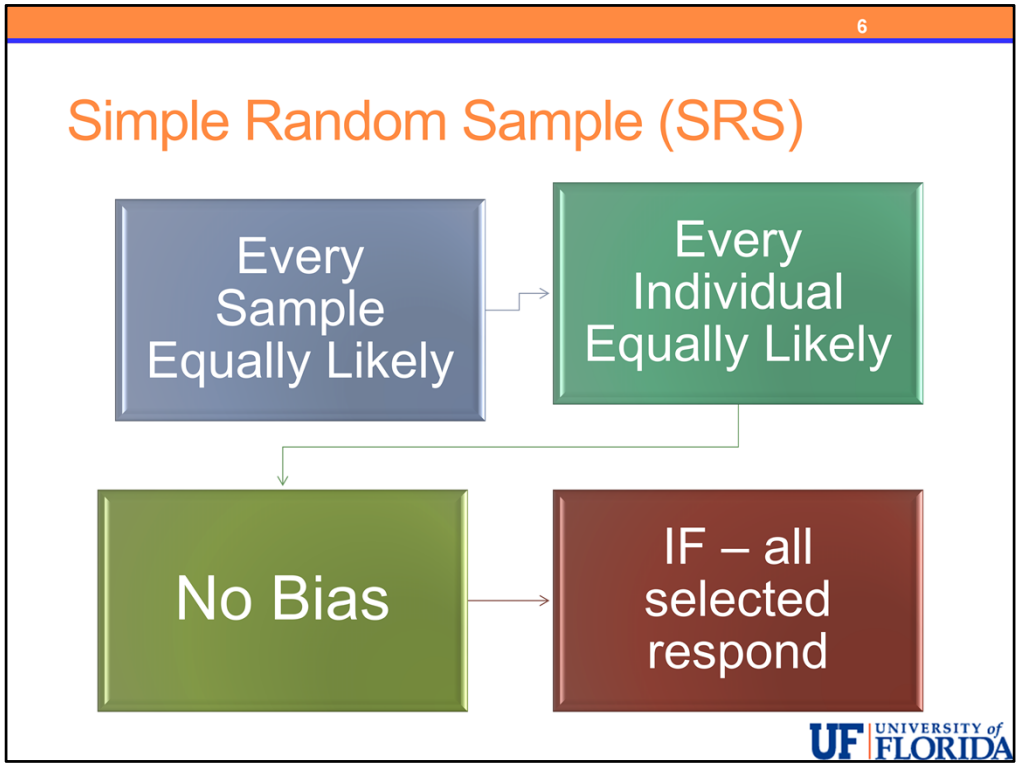
However, it is often impossible to achieve a sampling frame which matches the population exactly.



In a systematic sample, every n-th individual is selected from the sampling frame. In this method, we need to pick a random start point before starting the selection process (otherwise – it isn't a random sample at all!!).

For example, if we have the entire student roster in alphabetical order for a particular college, and we select every 50th student from the list after randomizing a starting point, we have a systematic sample.

Although these types of samples may not be subject to any clear bias, the fact that individuals with the same last name are not likely to be chosen together in a systematic sample – whereas they would be equally like in a completely random sample – may cause subtle sources of bias.



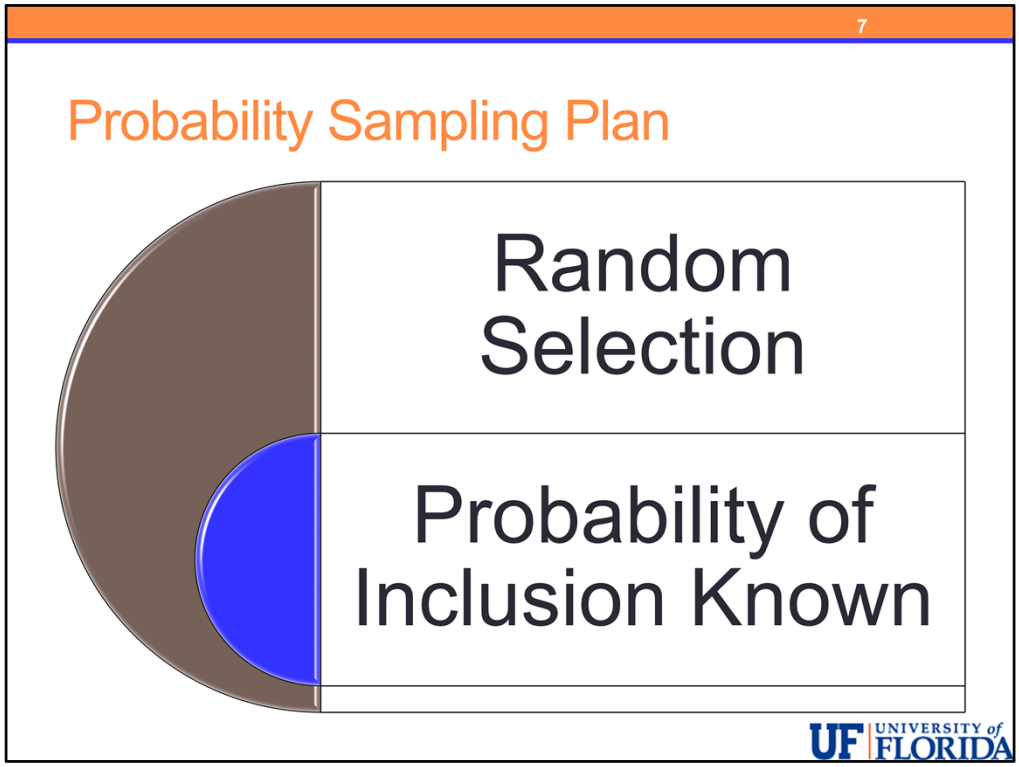
In a simple random sample, the process is conducted in such a way that every possible sample is equally likely to be chosen. In this case, it is also true that each individual is equally likely to be chosen. Like choosing names out of a hat.

For example, if we have the entire student roster in alphabetical order for a particular college, and we select 500 students completely at random.

If every student selected participates then we have a sample which is not subject to any bias and should succeed in being representative of the population of interest.

On the other hand, if we send the selected students a survey which they then choose whether or not to complete and return. Then we have a simple random sample which is also subject to volunteer response which has some of the same concerns as a volunteer sample. Non-response is a large concern with these types of surveys.

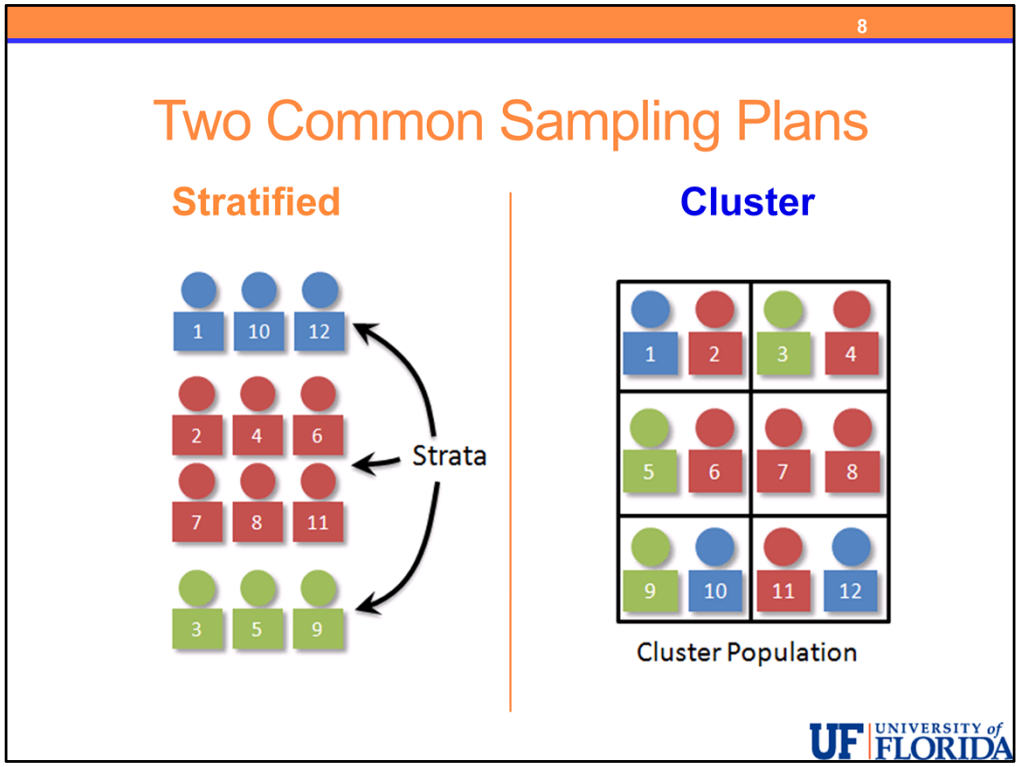
Sometimes we might follow-up with reminders to try to increase the response rate.



A simple random sample is the easiest way to base a selection on randomness. However, there are other, more sophisticated, sampling techniques that utilize randomness that are often preferable in real-life circumstances due to the complexities of the problems under study and the logistics of carrying out the sampling process.

Any plan that relies on random selection is called a **probability sampling plan (or technique)**.

Inferential methods require full understanding of the underlying probability of inclusion for each individual in the population (or more likely the sampling frame).



Weblink: {http://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html}

Now let's talk about two other common sampling plans.

The question of which to use usually depends either on the goals of the study or the feasibility of the sampling procedure or both.

For both of these methods, we utilize some naturally occurring grouping of the individuals in our population. The difference between these groupings is whether the individuals inside each group are naturally similar to each other or if the groups or more diverse.

For stratified sampling – it is either easier or is of interest to divide the population into groups called strata – inside each strata, individuals are similar. For example, in a class of students, I could stratify by gender and take a simple random sample of 20 students from each gender.

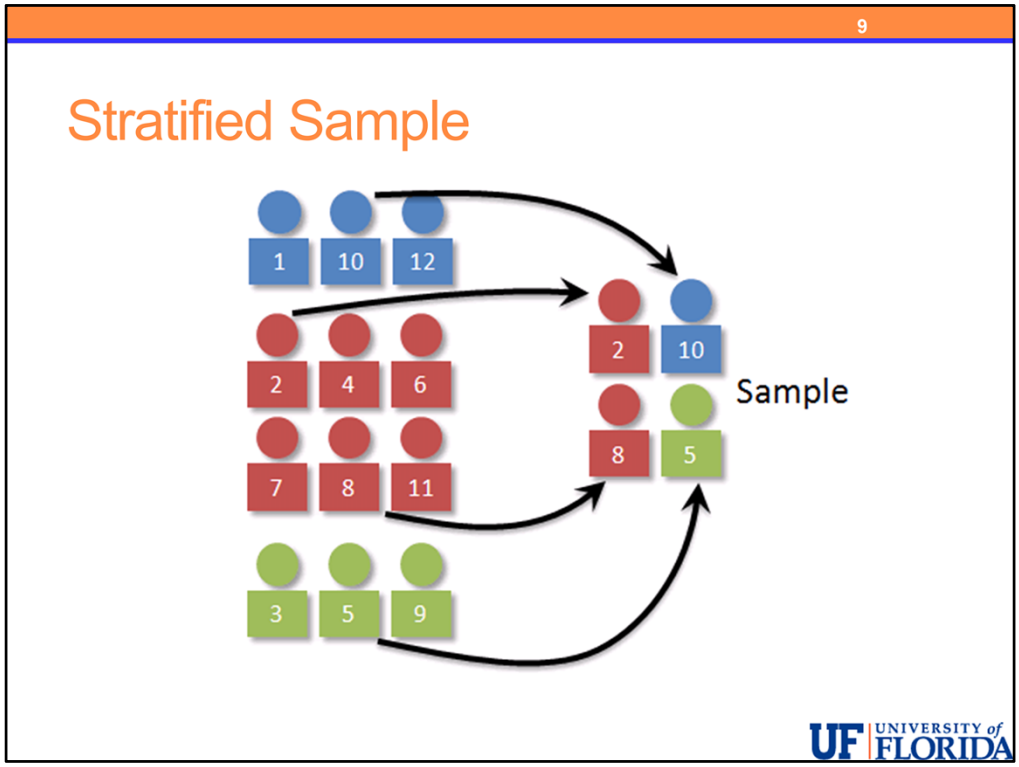
In that case, the population is the whole class, the strata are the two groups of students formed by female students and males students respectively. Individuals in the strata are similar in gender.

In a case-control study – there is stratification of the population into cases and controls – subjects are then sampled from these two strata.

Alternatively, for cluster sampling, the population is again divided up into groups – now called clusters – however individuals in clusters are diverse and more representative of the population as a whole. Here we randomly select which clusters to sample. In a simple cluster sample, every member of the cluster is chosen.

For example, in a survey of college students, we might consider each course a cluster and randomly sample which courses to use in our study. This would be considerably easier than randomly selecting students and then needing to contact and survey each selected student individually.

In a survey of people across the U.S., we might cluster people by county, zip code, or even a section of a street – although as we will mention shortly, in reality, if our cluster is county or zip code, we still will need to select only a subset of the people in a cluster in most cases.



In this illustration, the population is divided naturally by color which represent our strata.

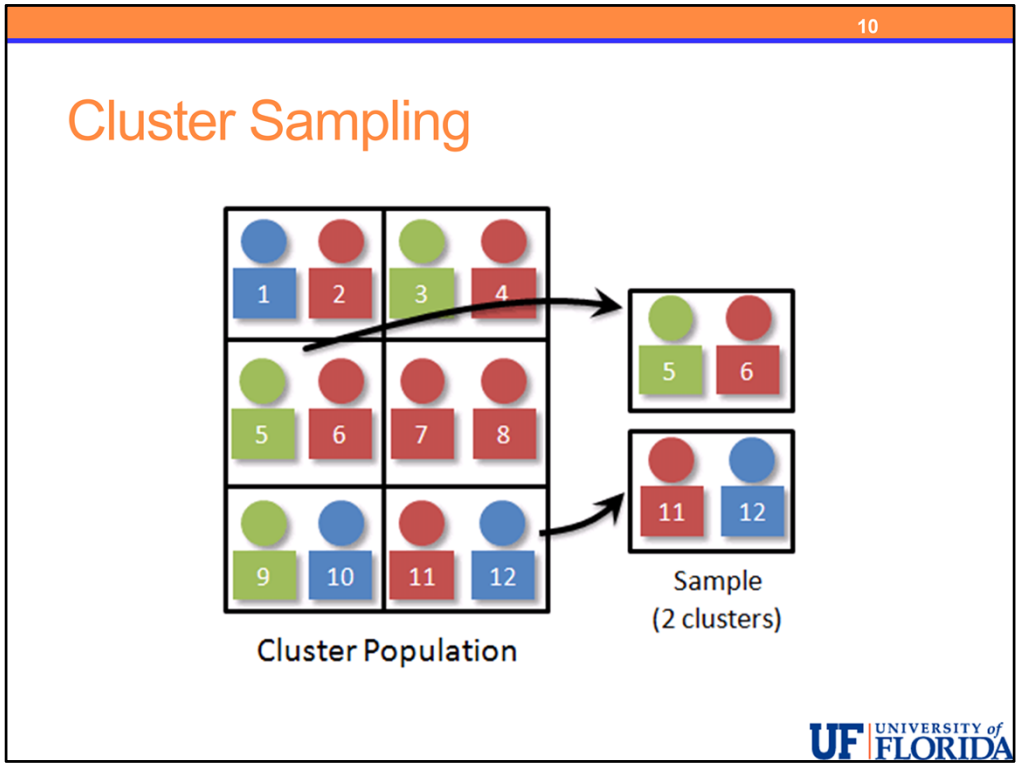
Since there are twice as many red (in the middle) as blue or green (on top and bottom), in order to represent the population, we should sample twice as many red as blue or green. Which has been done in the illustrated sample.

Even if the sampling is not proportional – as long as the proportion is known – the proper adjustments can be made if you know how (or know someone who knows how).

Complex sampling plans are common in health surveys!

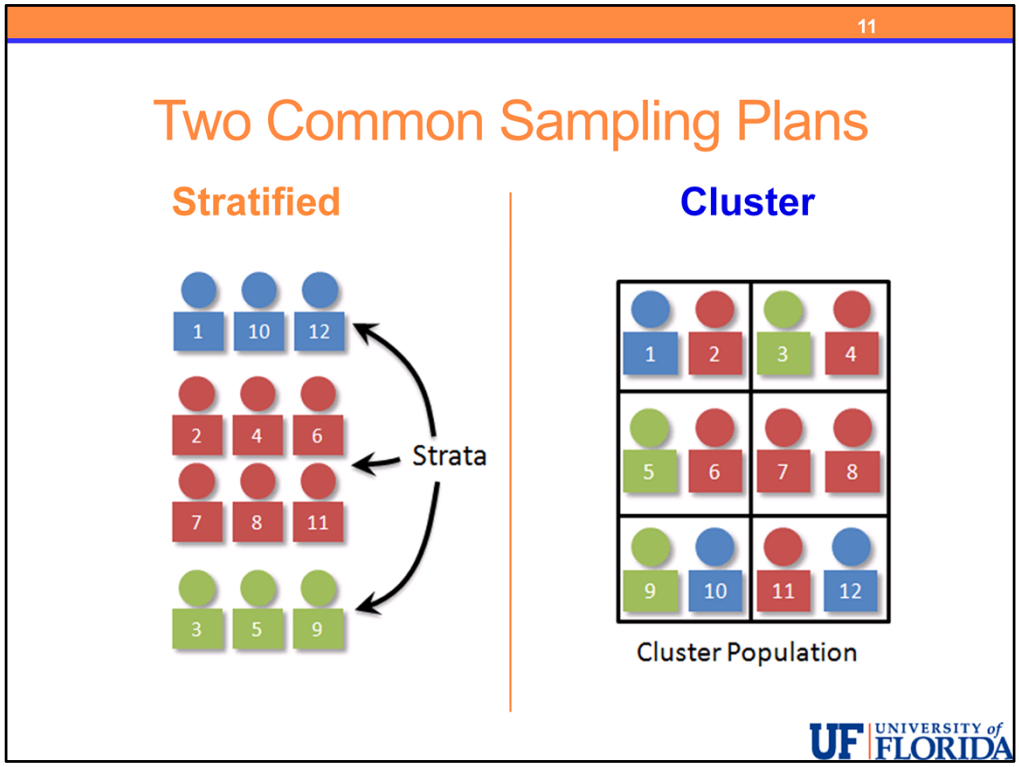
One reason for stratification is the possibility that you wish to have a particular number from each strata (equal males and females or twice as many cases as controls).

For example, if we take a simple random sample or cluster sample, we would not be guaranteed to get a particular number of individuals in each gender.



Here, we have individuals grouped into clusters where the members within each cluster are not necessarily similar – although they may be either randomly or due to some particular effect of that cluster.

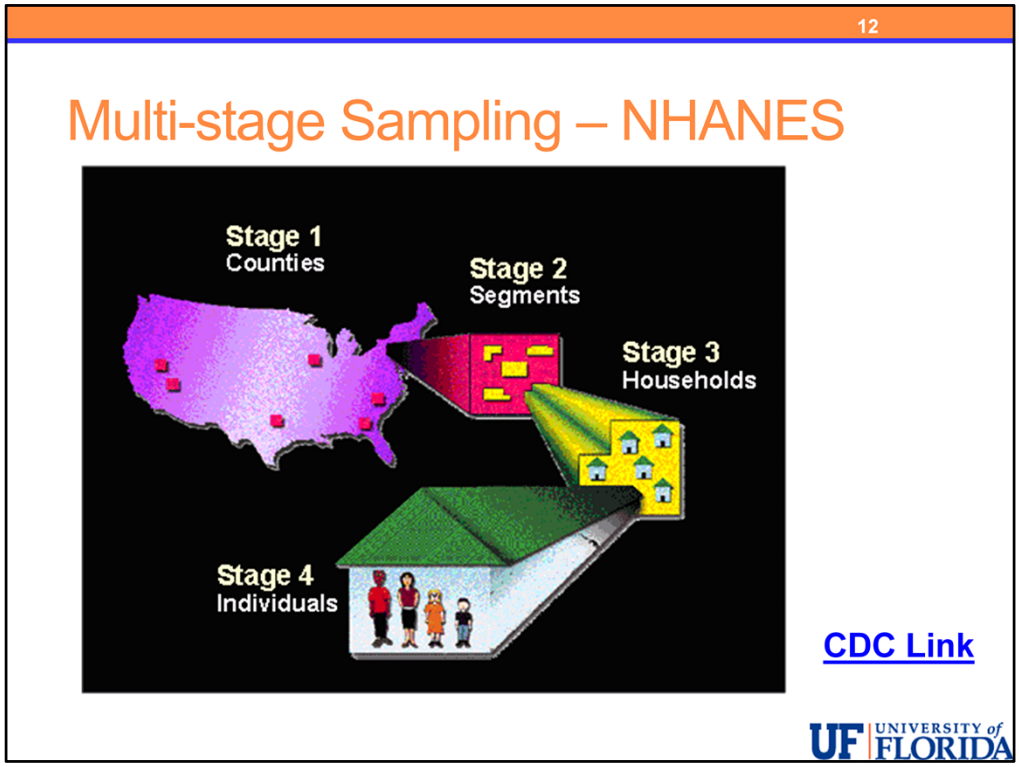
In this case we randomly sample two clusters – it is coincidental that in this particular case we still ended up with two reds and one each of blue and green. It could have ended up quite differently.



Often students find it difficult to distinguish between these two methods, since they both involve a natural grouping but if you think about

- Whether the individuals in the groups are similar on some variable of interest or if they are more diverse. Here,
    - Strata contain similar individuals
    - Clusters are more diverse – but may be similar in location or some other variable that allows for clustering to occur

- And whether we randomly sample which groups to select.
    - For strata we almost always sample from every strata
    - For clusters we almost always randomly select only a few clusters to sample



Weblink
{http://www.cdc.gov/nchs/tutorials/dietary/SurveyOrientation/SurveyDesign/Info1.htm}

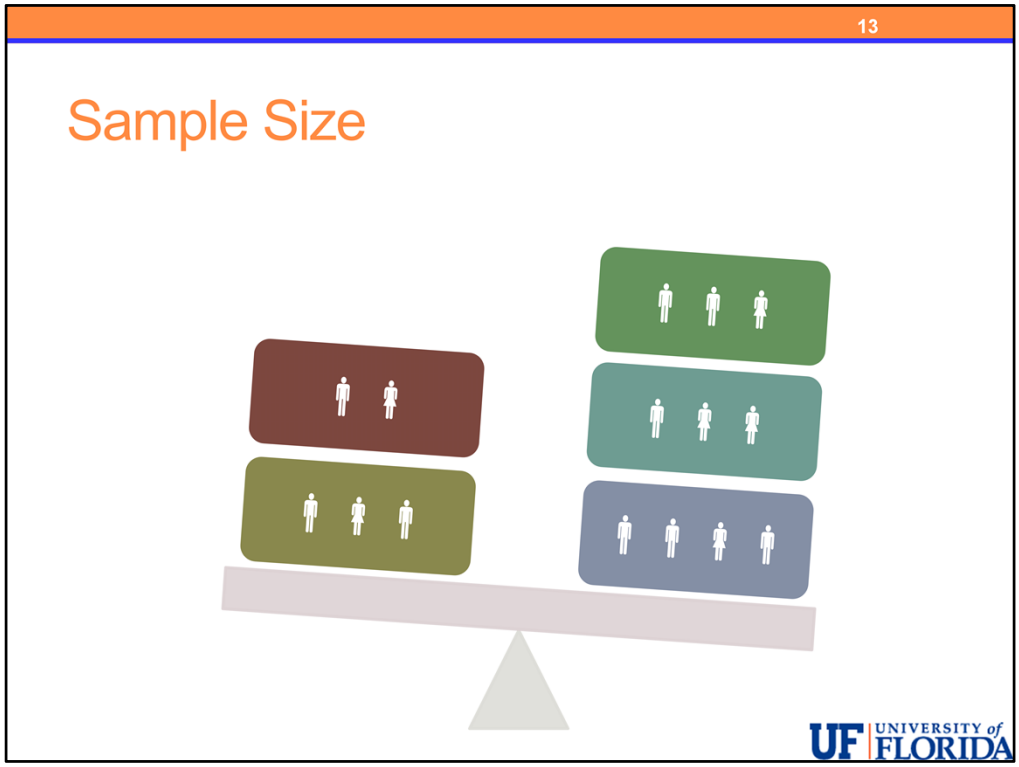
Multi-stage sampling plans are also common. In multi-stage sampling, various sampling methods may be used at any stage resulting in complex probabilities of selecting certain individuals into the sample. Usually simple random sampling is employed repeatedly in choosing clusters to sample from at a particular stage or in choosing which individuals to sample from a particular cluster or strata.

In the NHANES example – simplistically – first the counties are randomly selected – then each selected county is broken into segments and a certain number of segments are chosen randomly – then each selected segment is divided into households – and a random sample of households are selected. Finally within each sampled household an individual is randomly chosen.

In our course, the methods we use are traditionally applied for simple random samples. In the case of other sampling methods, adjustments to the standard methods are likely required.

The example here is from the NHANES survey – on the data information page linked here – you can see the warnings about properly accounting for the sampling design. These concepts are definitely beyond the scope of this course but some of you may be working with data like this in your future and will need to learn more about complex sampling and

how to adjust for it.



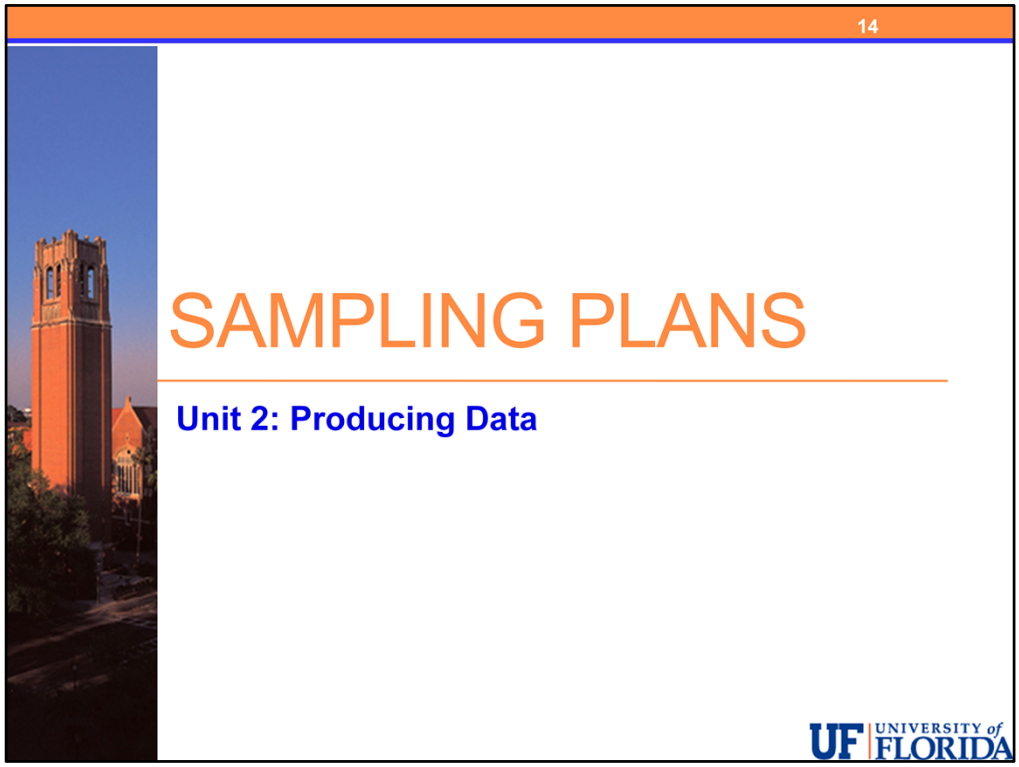
We will discuss sample size in more detail in the unit on statistical inference. Here we simply mention a few points.

First, our priority is to make sure the sample is representative of the population, by using some form of probability sampling plan

Second, it should be reasonable that a larger representative sample will do a better job at estimating the quantities of interest than a smaller one.

Since increasing the sample size will inevitably require greater resources, the question will be what is the smallest sample size we could use which will allow us to effectively answer our research question.

In practice, we must determine the appropriate sample size very early, if not entirely before, our data collection process. However, the methods which will be used to analyze the data must be known before these calculations can be carried out.



In this section, we defined a few common sampling methods. Among these were the probability sampling plans which include:

- Simple random samples (SRS)
- Cluster samples
- Stratified samples
- Multi-stage sampling
- And to a lesser extent, systematic samples

Simple random samples – if feasible to conduct – will result in a sampling plan that is free from any bias

We also discussed non-probability sampling plans – for which bias is likely to result - including:

- Volunteer samples
- Convenience samples

We defined the sampling frame – which is all possible subjects with the possibility to be included in a particular study – and discussed how this should match the population as closely as possible.

And finally we briefly mentioned the issue of sample size.