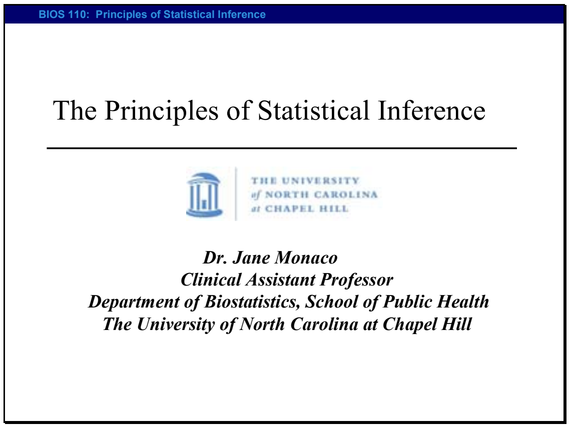
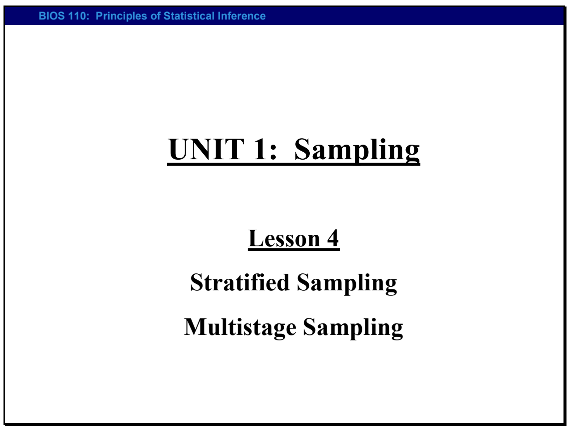
Slide 1



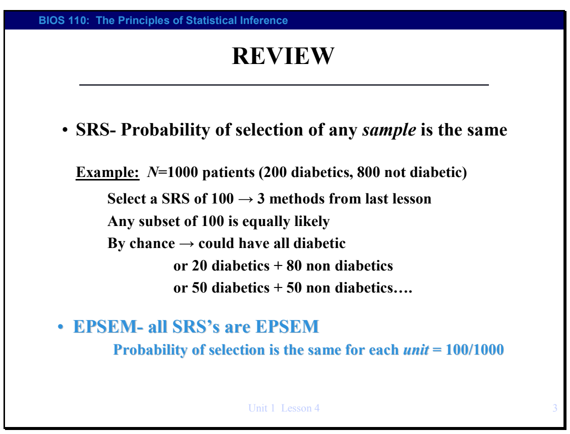
Welcome back to The Principles of Statistical Inference- the online version of the introductory course in biostatistics offered by the department of Biostatistics.

Slide 2



We continue in Unit 1 to discuss the topic of Sampling. How do you pick units to study? In Lesson 4, we will learn about two more sampling designs… stratified sampling and multistage sampling.

Slide 3



Let's review a bit.
Recall that a SRS is when the probability of selection of any sample is the same as the probability of selecting any other sample.
We discussed the example when we had 1000 patients in a clinic -200 were diabetic and 800 were not.
We could use any of the three methods discussed in the last lesson to select a SRS. Important to note that any sample, any subset, of 100 subjects is equally likely to be chosen as any other sample. By chance, we might get all diabetics, we might get 20 diabetics and 80 not diabetic, for example.
We also discussed the concept of EPSEM. Recall all SRS are EPSEM. EPSEM means the probability of selection is the same for each member.
So far so good….but….
What if we want to ensure at least some diabetics are included in sample?
What if we want to ensure equal numbers of diabetics and nondiabetics?

What if we want to select a disproportionate number of diabetics so that analyze them separately?
Solution is stratified random sampling.

Slide 4

BIOS 110: The Principles of Statistical Inference

OBJECTIVES

- **Identify and understand a stratified sample**
  - **Advantages and disadvantages**
  - **Produce a stratified sample**
  - **Calculate an unbiased estimate of a parameter of stratified sample**
- **Identify and understand a multistage sample**
  - **Advantages, disadvantages**
- **Differentiate between SRS, systematic, stratified and multistage sampling**
- **Combine different sampling types**

Unit 1 Lesson 4 4

The objectives of this lesson are to identify a stratified sample by providing advantages and disadvantages, producing a stratified sample and calculating an unbiased estimate of parameter of stratified sample.
Other objectives are to describe and understand a multistage sample,

to differentiate between different sampling types,
to combine different sampling types.

Slide 5

BIOS 110: Principles of Statistical Inference

STRATIFIED SAMPLING

Let's begin with Stratified Sampling

Slide 6

BIOS 110: The Principles of Statistical Inference

STRATIFIED RANDOM SAMPLE

- **Divide population into strata or groups [groups are similar in some characteristics]**
  - **Typical strata are ethnicity, age, geographic region, gender, etc.**
- **Select a sample within each stratum separately**
- **Combine SRSs in each stratum for final sample**

Stratified Random Sample may or may not be EPSEM

Unit 1 Lesson 4 6

In a stratified random sample, we divide the population into strata (or groups). These groups are similar in some characteristic… some common examples are to stratify by ethnicity, age, geographic region, etc.
Within each stratum, we select a sample (usually a SRS).
Then, we combine these samples from each stratum to form our final sample.
NOTE: A stratified random sample, may or may not be EPSEM.
I'll explain all this with an example….

Slide 7

BIOS 110: The Principles of Statistical Inference

Stratified Random Sample: Example

- Select 100 patients out of 1000 patients → select equal numbers of diabetics and nondiabetics
- Form strata 200 diabetics and 800 nondiabetics
- Select 50 diabetics and 50 nondiabetics
  - SRS of 50 out of 200 diabetics
  - SRS of 50 out of 800 nondiabetics
- Full sample is n= 100 out of N=1000

Unit 1 Lesson 4 7

Let's return to our example.
We've seen this example before in a previous lesson– now let's look at it in the context of stratified sampling.
Suppose we need to select equal numbers of diabetics and nondiabetics. So select 50 diabetics and 50 nondiabetics
We first forming strata 200 diabetics and 800 nondiabetics.
Then, select a SRS of 50 out of 200 diabetics and select a SRS of 50 out of 800 nondiabetics. In other words we force that "50/50" split. You can use any method for selecting an SRS.
Thus the full sample is n= 100.

Slide 8

BIOS 110: The Principles of Statistical Inference

NOTATION FOR STRATIFIED RANDOM SAMPLE

- Divide population into j strata.
  - Example has j=2 strata (nondiabetics and diabetics)
- Number selected in each stratum = n_j
  - n_1=50 diabetics and n_2 = 50 nondiabetics.
  - Sum of n_j's is n
- Number in population in stratum j =N_j
  - N_1=200 and N_2=800
  - Sum of N_j's is N
- Stratum specific sampling fraction = n_j/N_j
  - diabetic sampling fraction=n_1/N_1=50/200 = 0.25
  - non-diabetic sampling fraction n_2/N_2 =50/800=0.0625

Unit 1 Lesson 4 8

These are some notations we will need:
Divide population into j strata. (We had j=2 strata. Diabetic and non-diabetic)
Number selected in each strata is $n_{j.}$ (We selected n_1=50 diabetics and n_2= 50 nondiabetics. Convince yourself that the sum of the little nj's are little n)
Number in population in stratum j is $N_{j.}$ (N_1= 200, N_2= 800)
The stratum specific sampling fraction is $n_j/N_{j.}$
We just consider the sampling fraction for each group, each stratum.. Our example has diabetic sampling fraction=50/200 and nondiabetic sampling fraction 50/800.

Slide 9

BIOS 110: The Principles of Statistical Inference

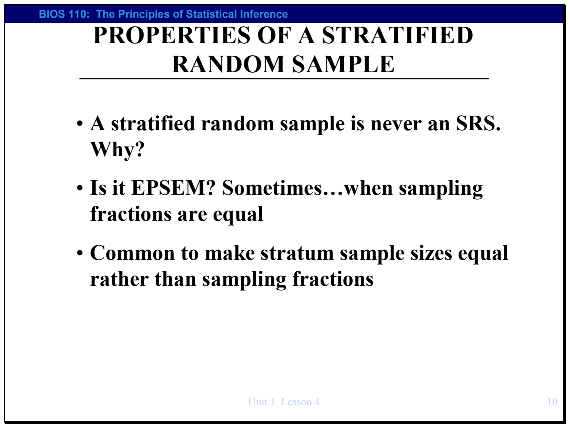
Why stratify?

- Analyze strata separately
- If the strata are homogeneous, we can add efficiency

Unit 1 Lesson 4 9

Why do we stratify?
It's because we can analyze strata separately- we can assure that we have sufficient sample size within the strata.(In the previous example we assure that we have enough diabetics to analyze separately.)
If strata are homogeneous we can add efficiency (we won't get into why… we'll save this for another course).

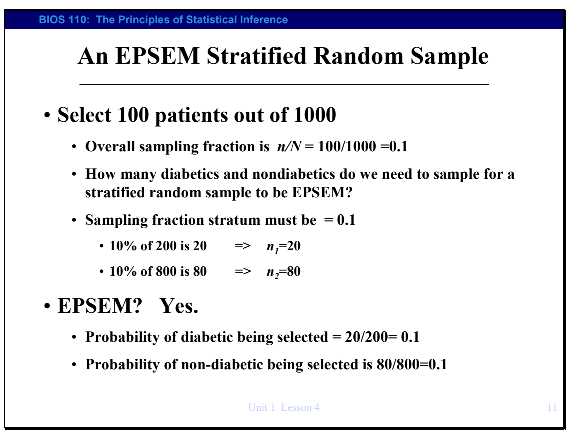
Slide 10



The properties of a stratified random sample are :
A stratified random sample is never SRS. Why?
Recall the definition of an SRS… the probability of selection for any sample is the same. We've mentioned this example before… the probability of selecting 100 diabetics in our example is, well… 0. We are stratifying and are assured of getting nondiabetics.
Is it EPSEM? Sometimes it is when sampling fractions are equal. On the next slide we'll discuss this.

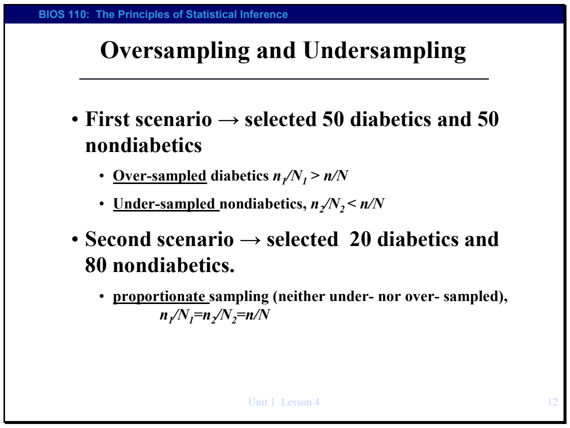
It's common to make stratum sample sizes equal rather than sampling fractions equal.
In the above example, we made the stratum sample size equal (n_1=n_2=50) rather than sampling fractions.
The sampling fractions are NOT equal in the previous example n_1/N_1=50/200, is not equal n_2/N_2=50/800

Slide 11



How can we select a stratified random sample which is EPSEM?
Recall we are selecting 100 patients out of 1000. Now (instead of 50 diabetics and 50 nondiabetics) let's investigate how we can make the sampling fractions the same.
Overall sampling fraction is n/N = 100/1000 =0.1.
How many diabetics and nondiabetics do we need to sample for a stratified random sample to be EPSEM? The stratum specific sampling fractions must be the same as each other and the same as the overall sampling fraction, 0.1.
Sampling fraction diabetics = 0.1, n_1=20.
Sampling fraction nondiabetics=0.1, n_2=80.
Check that this is EPSEM. The probability of selection for each member is the same.
Probability of selection for a diabetic is 0.1.
Probability of selection for a nondiabetic is 0.1.

Slide 12

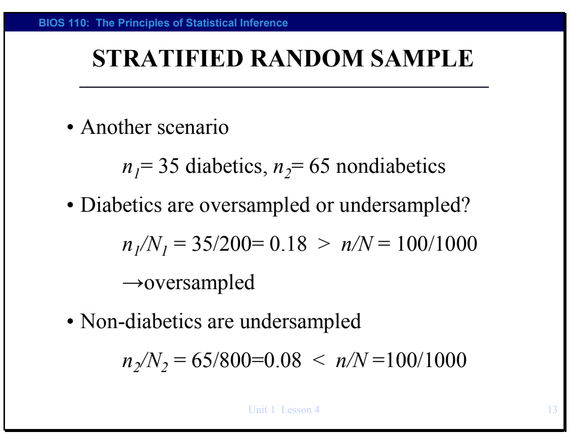


Sometimes we may want to over-sample a stratum or under-sample a stratum.
Over-sampling is when the sampling fraction for the stratum is bigger than the overall sampling fraction.
Under-sampling is when the sampling fraction for the stratum is smaller than the overall sampling fraction.
Why would you want to do such a thing? Because we may wish to analyze this stratum separately or a stratum may have more variability than another stratum. To reduce variability of overall estimate, we over-sample in certain subgroups.

Consider the first scenario…. We selected 50 diabetics and 50 non-diabetics… the overall sampling fraction was 0.1. The sampling fraction for diabetics was 50/200 = 0.25. The sampling fraction for the nondiabetics was 50/800= 0.0625. Check me! Don't ever trust my math!
OK… So the diabetics are over-sampled and the non-diabetics are under-sampled.
Consider the second scenario…When the sampling fractions are the same for all the strata (and equal to the overall sampling fraction), we call this <u>proportionate</u> sampling. There is neither over- nor under- sampling. An example is when we selected 20 diabetics and 80 non-diabetics.

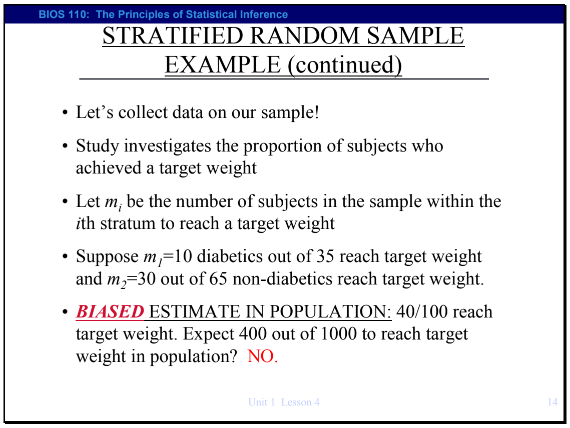
NOTE: Under-coverage (in a previous lesson) and under-sampling are different concepts.

Slide 13



Here is another example, suppose that n_1= 35, n_2= 65.
This may be because you want to analyze the diabetics separately– the necessary sample size within each stratum would be determined by a statistician to ensure sufficient sample size to analyze diabetics separately.
Check that the sampling fraction for the diabetics is bigger than the sampling fraction for the non-diabetics. So the diabetics are oversampled. .

Slide 14



Continuing with this particular example, suppose we want to collect data on this sample! So far we've really just talked about selecting the sample, but eventually you'll want to collect data about them. □

Suppose that a study investigates the proportion of subjects who achieved a target weight.

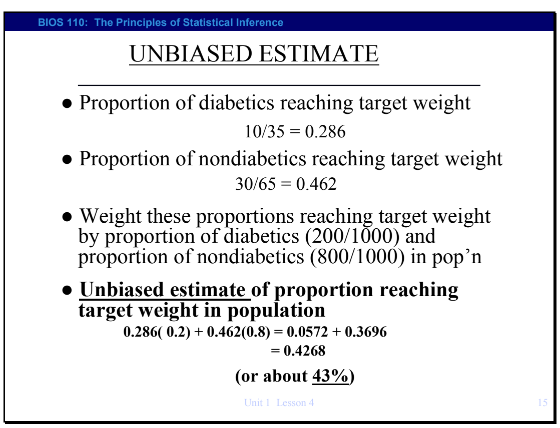
You'll have some diabetics who reach their target weight, some non-diabetics who reach their target weight and the proportions within the subgroups may be different. We are still considering the sample to be 35 diabetics and 65 non diabetics.

Some MORE notation…. let m_i be the number of subjects in the sample within the ith stratum to reach a target weight.

Suppose m_1=10 diabetics out of 35 reach target weight and m_2=30 out of 65 non-diabetics reach target weight.

BIASED ESTIMATE OF TARGET WEIGHT FOR POPULATION: Some might say that therefore we'd expect in general that 40% of people in the population to reach their target weight. This is because 40 out of 100 reached target weight in the study- in the sample. But the diabetics made up a larger proportion in the sample than in the population. So in this biased estimate the diabetics contribute too much to the estimate. (Not helpful…)

Slide 15



How can we get an unbiased estimate?

First compute proportion of diabetics reaching target weight in the sample, 10/35 = 0.286 or we could say that 29% of diabetics achieved the target weight

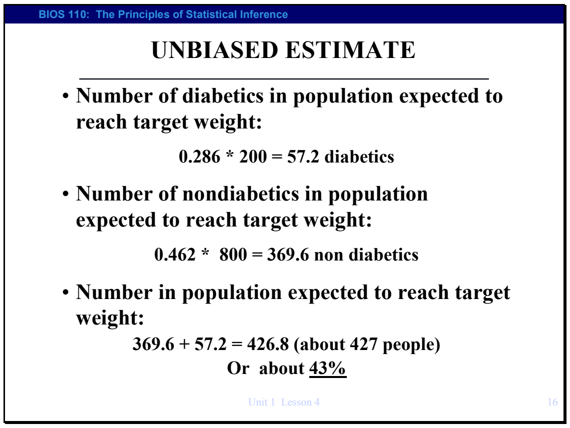
Proportion of nondiabetics reaching target weight in the sample, 30/65 = 0.462 (or 46%).

Now we want to weight these numbers above by proportion of diabetics (200/1000) and proportion of nondiabetics (800/1000) in the population.

Unbiased estimate = 0.286(0.2) + 0.462(0.8) = 0.0572 + 0.3695 = 0.4268

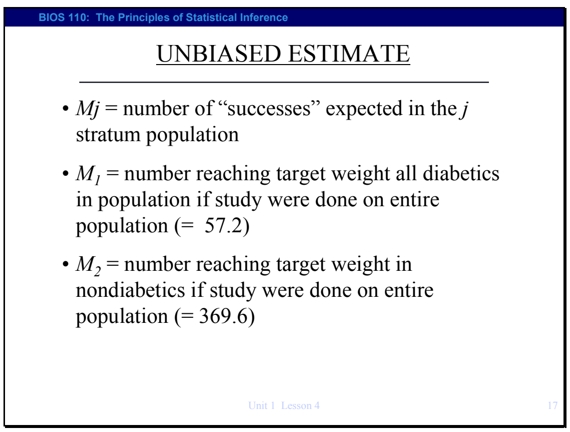
In other words, in the population of 1000 we would expect that about 43% would achieve their target weight if placed in the same program. If we conducted the study on all 1000 people, we'd expect about 427 of them to achieve their target weight.

Slide 16



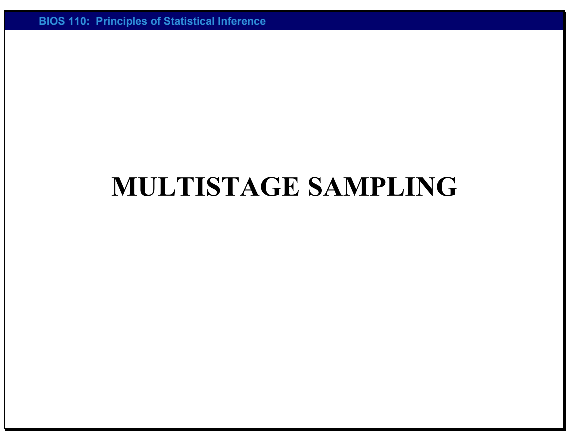
Another way to think about the unbiased estimate in the population is the following: First calculate the number of diabetics, out of 200, we'd expect to reach their target weight. This would be 28% of 200 which is equal to 57 people
How many nondiabetics would we expect to reach their target weight? Calculate 46% of 800 nondiabetics. We expect about 370 people to reach their target weight in this group.
MISTAKE IN THE AUDIO: I said 396 instead of 369
Finally how many total in the population might we expect to meet their target weight… 427 people out of 1000 which is 43%.

Slide 17



The notation here is ***Mj*** the number of 'successes' expected in the jth stratum for the population. Success is an event of interest… perhaps reaching target weight. "success" is not necessarily a positive event.
M_1 is, as we just said, the number of diabetics reaching their target weight the population
M_2 is the number of nondiabetics reaching their target weight in the population
As before our unbiased estimate for the population is 0.43 or about 43%.

Slide 18



Next, we are going to talk about multistage sampling. Another term for this sampling design is **cluster sampling.**

Slide 19

BIOS 110: The Principles of Statistical Inference

MULTISTAGE SAMPLING

- **Divide target pop'n into groups, usually called *clusters***
- **<u>First Stage:</u> Select a sample of the clusters –<u>primary sampling units (PSUs)</u>**
- **<u>Second Stage:</u> Select all units within those clusters OR Select units within the clusters (<u>secondary sampling units - SSUs</u>)**

Unit 1 Lesson 4 19

Moving to a different sampling strategy, we will investigate multistage sampling.
In multistage sampling, the target population is separated into groups, often called <u>clusters</u>.
In the first stage, we randomly select a sample of the clusters. These clusters are called primary sampling units, or PSUs
In the second stage, either we select ALL the units within the clusters which have already been selected OR you may select units within the clusters. These selections are called secondary sampling units or SSUs.

Slide 20

BIOS 110: The Principles of Statistical Inference

MULTI STAGE- Advantages and Disadvantages

- **Advantages**
  - **No frame (list) required of individual sampling units**
  - **Often cost effective in selecting and implementing**
- **Disadvantages**
  - **Harder to analyze**
  - **More complicated**

Unit 1 Lesson 4 20

With multistage sampling, no frame (list) required of <u>individual</u> sampling units
Also, multistage sampling is often cost effective in selecting and implementing
However, multistage design are harder to analyze and just more complicated

Slide 21

BIOS 110: The Principles of Statistical Inference

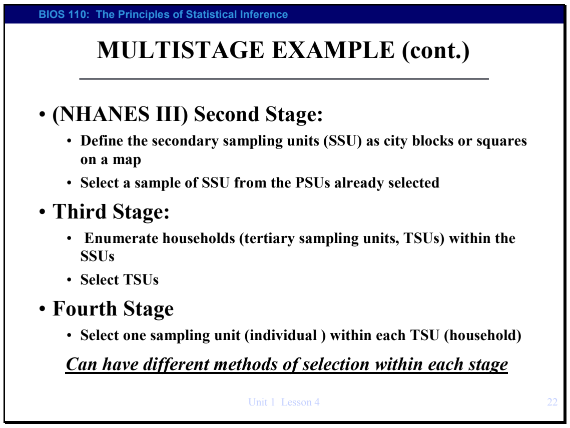
MULTISTAGE SAMPLING: Example

- **NHANES III: Third National Health and Nutrition Examination Survey**
  - **Target Population: Civilians age 2 months and older in 50 states and Washington DC**
- **First Stage: Select counties (SRS) within the US**
  - **PSUs are counties (81 counties out of all counties in US)**
  - **There is an accurate list of counties in the US**
  - **[In actuality, investigators selected a stratified sample of the PSUs!]**

Unit 1 Lesson 4 21

An example of a multistage sample is the NHANES.
The target population was civilians age 2 months and older in the 50 states and Washington DC.
Imagine the logistical problems of getting a list of all these individuals and taking a simple random sample- all civilians in the US! …. Not very practical.
The study design was to first select 81 counties within the US. So the primary sampling units were counties. There was an accurate list of counties in the US. In actuality, the investigators selected a stratified sample of the counties – they combined the sampling strategies which is often the case. As an illustration of multistage sampling, we can just think of selecting counties out of all counties in the US.

Slide 22



Continuing with this example, in the NHANES study, after the counties were selected it was not practical to interview all residents with these counties.

So, investigator selected <u>secondary sampling units</u> (SSUs). The secondary sampling units were city blocks or grid squares on a map.

Within each PSU (county), investigators created groups (SSUs) within the counties and selected then one or more of the SSUs.

The third stage is to select households within the SSUs. These households can be referred to as tertiary sampling units.

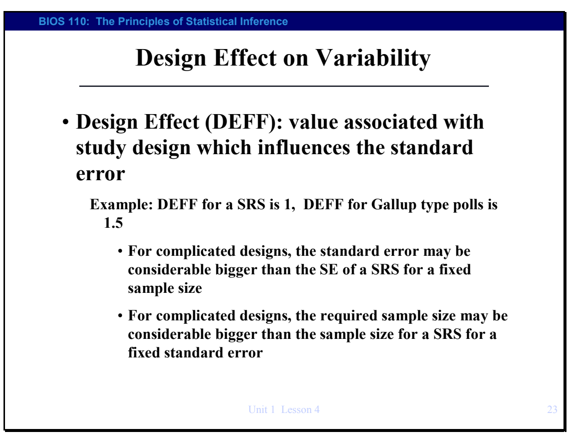
Finally within each household (TSUs), investigator selected one individual.

Whew… a complicated study design, but you can see why this was an efficient use of time for the investigators selecting the sample and the data collectors. Data collectors only had to visit 81 counties and collect data from a number of individuals within those counties and within those city blocks.

Note also that at any one of these stages, we could have a different method of selection… you could have had a stratified, SRS or systematic sample taken for example.

This multistage design did not require a sampling frame of individuals (all citizens in the US).

Slide 23



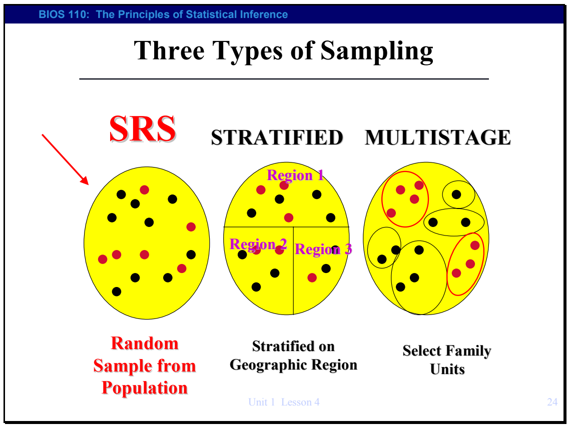
When we use these different kinds of sampling schemes, we may gain a ‘logistical advantage’ . .. It is often easier to implement a more complicated design or otherwise advantageous to use some sampling plan other than SRS.

However, these more complicated designs can have the effect of increasing the standard error of estimates. So for a fixed standard error, we may be required to have a bigger sample size for these complicated designs compared to other designs.

The DEFF (design effect) is a value associated with the study design which influences the standard error.

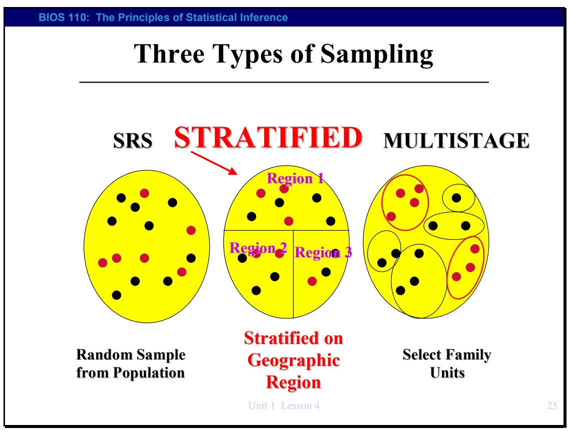
The DEFF of a SRS is 1. So, the standard error for these other designs are compared to the SRS. It is used as a yardstick. The design effect for a gallup poll is 1.5.

Slide 24



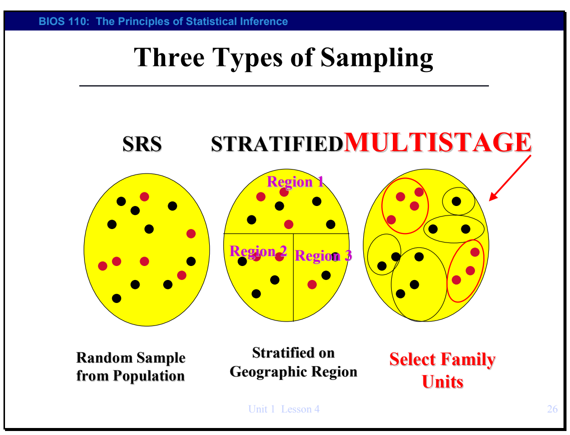
We've discussed several types of sampling… let's compare three of the types that we've discussed.
Suppose each of these yellow ovals represents the population– copies of the SAME population…the red dots are units within the population that are sampled and the black dots are the units in the population which are not sampled.
Each oval represents the **same** population – under the three different sampling strategies, different units are selected.
We'll start with SRS. In a SRS, over to left, we can see our dots which are sampling units. We just randomly select several sampling units to be our sample- these are the red dots.

Slide 25



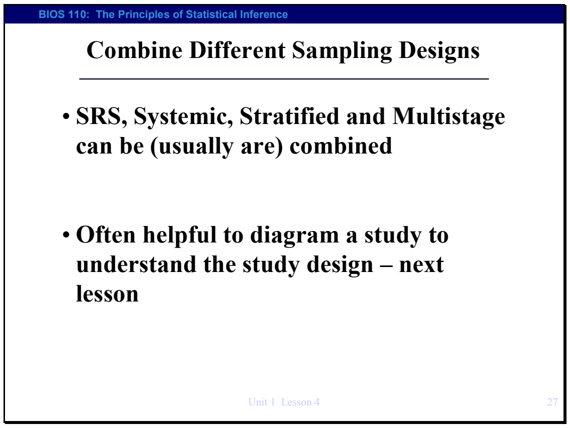
In a stratified sample, we think of the population as divided into groups called strata. Here let's say that the data are stratified by geographical location. There are three strata (say, geographical region 1, 2, and 3). Then within each strata we select using some method (SRS? Systematic?) individuals within each region. How many we select within each region will be determined by the goals of the study. The strata may be over sampled, under sampled or proportionately sampled.

Slide 26



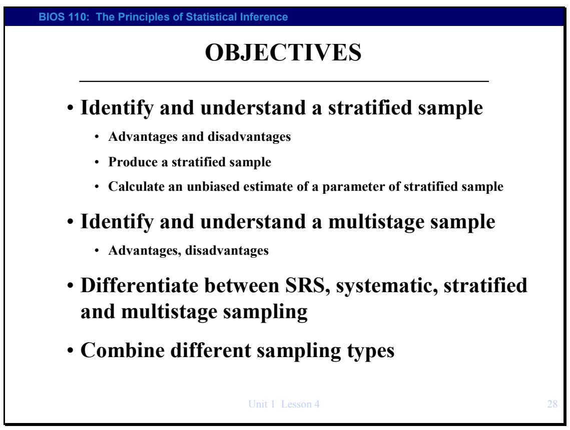
We could also use multistage sampling. Here we view the population as having clusters… say family units are clusters in this example. Each family is represented here by dots which are circled.
We select two family units from all family units. The family units we select are marked in red. Then within these family units (PSUs or clusters) we select all family members. [You could, alternatively, add another stage and select individuals within each family.]

Slide 27



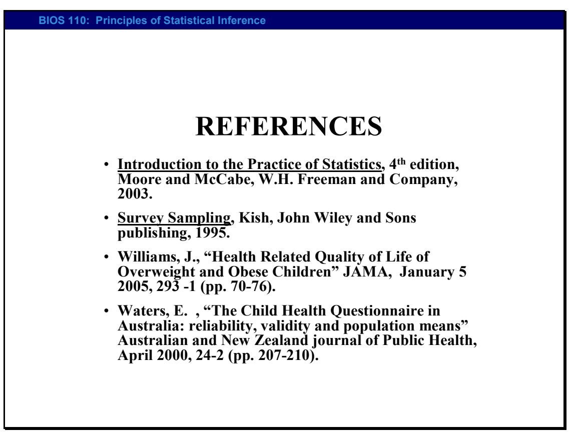
As we have said a few times, different sampling designs are often combined. Sampling designs can be quite complicated, so often it is helpful to diagram a study design. We'll look at this more in the next lesson.

Slide 28



The objectives of this lesson are to identify a stratified sample by providing advantages and disadvantages, producing a stratified sample and calculating an unbiased estimate of parameter of stratified sample.
Other objectives are to describe and understand a multistage sample, to differentiate between different sampling types, to combine different sampling type.

Slide 29



Slide 30