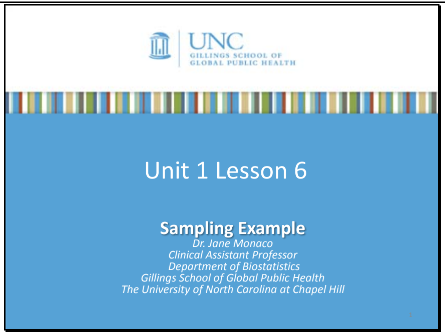
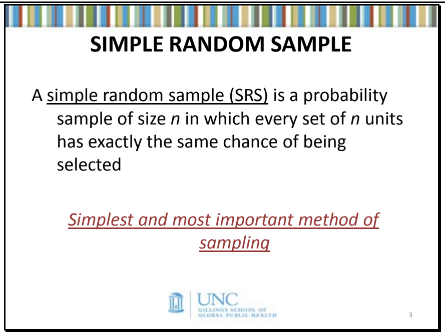
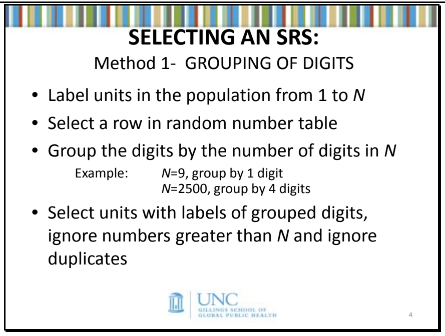
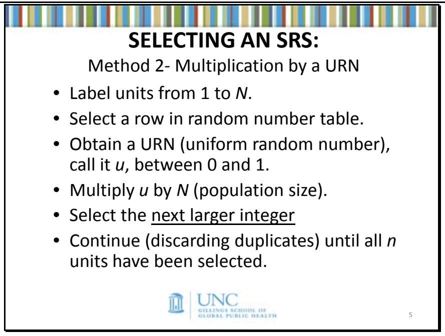
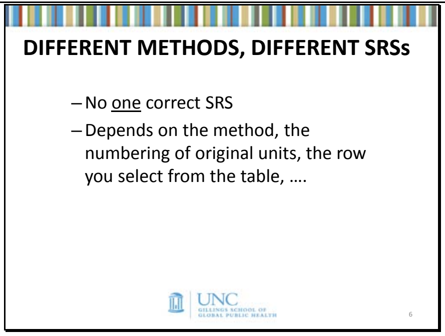
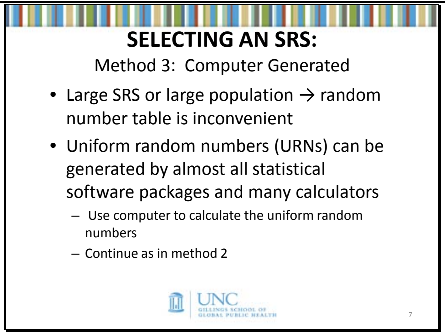
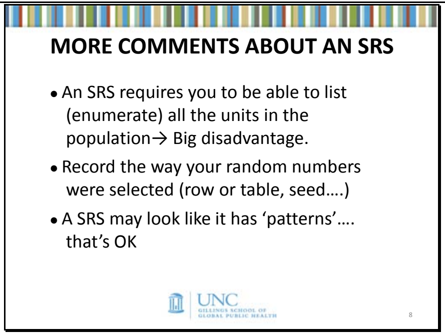
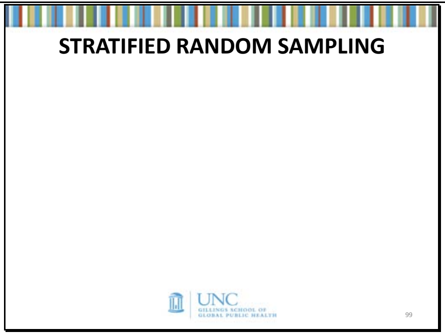
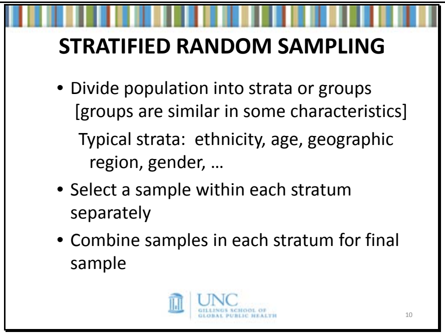
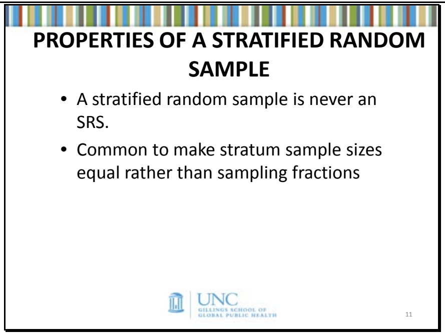
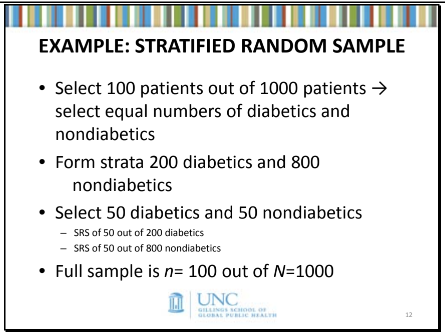
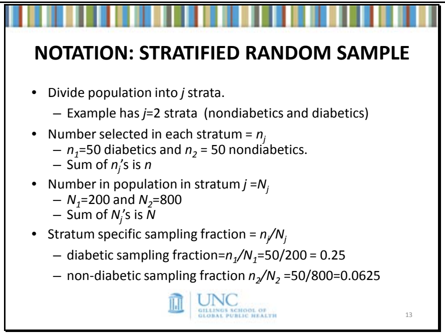
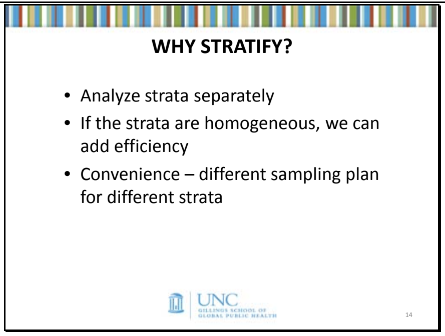
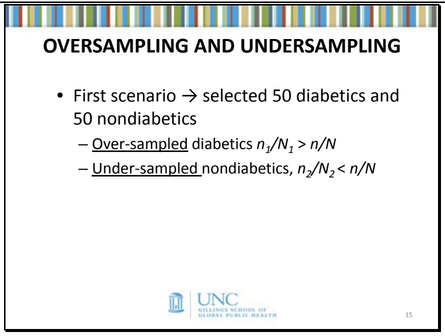
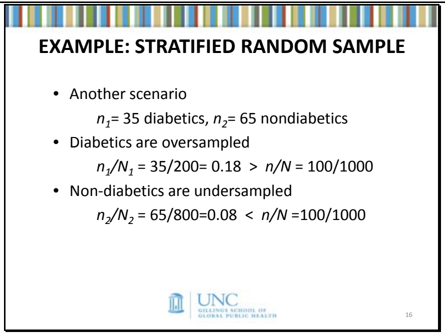
Slide 1	 UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH Unit 1 Lesson 6 **Sampling Example** *Dr. Jane Monaco* *Clinical Assistant Professor* *Department of Biostatistics* *Gillings School of Global Public Health* *The University of North Carolina at Chapel Hill*	
Slide 2	 **Objectives** ▪ Review Simple Random Sampling, Stratified Sampling and Multi-stage Sampling ▪ Understand the role of sampling in motivating example: Quality of Life and Obesity in School Children in Australia 2	▪Review Simple Random Sampling, Stratified Sampling and Multi-stage Sampling ▪Understand the role of sampling in motivating example: Quality of Life and Obesity in School Children in Australia – one of our global health examples.
Slide 3	 **SIMPLE RANDOM SAMPLE** A simple random sample (SRS) is a probability sample of size *n* in which every set of *n* units has exactly the same chance of being selected *Simplest and most important method of sampling* 3	Recall that a probability sample is one in which every element has a known non-zero probability of selection and that probability is determined by chance. Most important example of a probability sample is a simple random sample or SRS. Many other methods use a SRS as a starting point. An SRS is a probability sample of size *n* in which every set of *n* units has exactly the same chance of being selected. [This is an important definition to know and understand.] An SRS is the simplest and most important method of sampling, because it is a building block we need other methods and when we talk about randomization. An SRS is a probability sample, but not the only kind of probability sample.

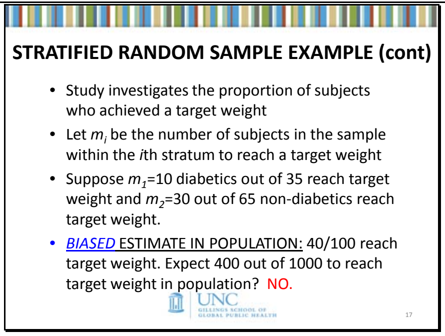
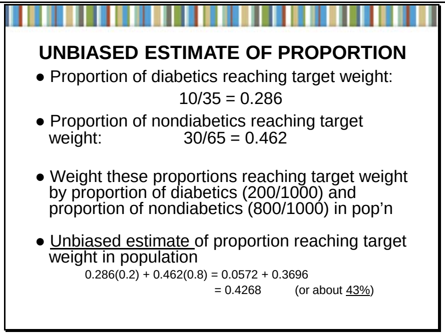
Slide 4	 **SELECTING AN SRS:** Method 1- GROUPING OF DIGITS • Label units in the population from 1 to *N* • Select a row in random number table • Group the digits by the number of digits in *N* Example: *N*=9, group by 1 digit *N*=2500, group by 4 digits • Select units with labels of grouped digits, ignore numbers greater than *N* and ignore duplicates	The first method for selecting an SRS that we'll discuss is to group digits in the random number table. Here's the method in general, and then we'll look at in an example…. First, label units from 1 to *N* where "*N*" is the population size. Then, select a row in random number table available in most introductory statistics text books (or random number book). (Can you imagine a more boring book than a random number book? Yet …there is such a thing…) You can select any row… we'll talk more about that in a minute. Then, group the digits in the table by the number of digits in *N* [*N*=9, group by 1 digit, *N*=2500 group by 4 digits, *N*= 72 group by 2 digits….] Finally, select units with labels of grouped digits, ignoring numbers greater than *N* and ignoring duplicates. It sounds complicated when you write down the steps, but we've seen it is pretty straightforward.
Slide 5	 **SELECTING AN SRS:** Method 2- Multiplication by a URN • Label units from 1 to *N*. • Select a row in random number table. • Obtain a URN (uniform random number), call it *u*, between 0 and 1. • Multiply *u* by *N* (population size). • Select the next larger integer • Continue (discarding duplicates) until all *n* units have been selected.	The next method uses multiplication by the uniform random numbers. Like the previous method, you start by numbering the members of the population from 1 to *N*. Next you select a row in the random number table. You then insert a decimal before the first digit, to view the random number as a number between 0 and 1. This value is called a uniform random number (URN). You will multiply the random number (between 0 and 1) by *N*, the population size. Take the next largest integer (in other words, round up). Select this unit of the population. Repeat these steps, ignoring duplicates until you have selected *n* units for the sample. Again, when you just read the steps, it looks confusing and hard, but it is not hard in practice.
Slide 6	 **DIFFERENT METHODS, DIFFERENT SRSs** – No one correct SRS – Depends on the method, the numbering of original units, the row you select from the table, ….	Do we get same simple random samples using these methods? In general, no - There is no one correct SRS. The SRS will depend on, among other things, the way you originally ordered/labeled the population, the method you use, the row you happen to select in random table, …. So there are lots of possible valid SRSs from a given population.

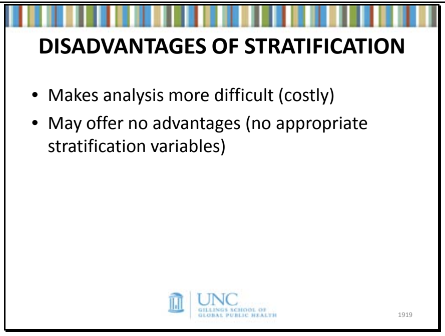
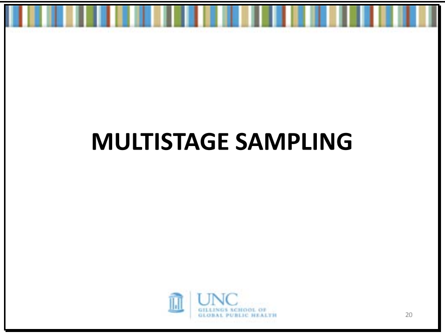
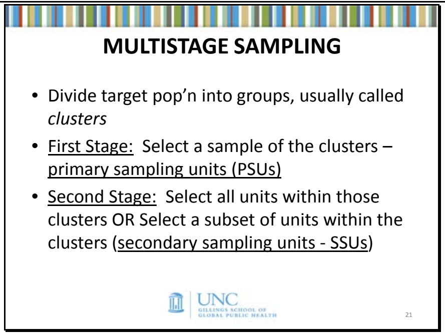
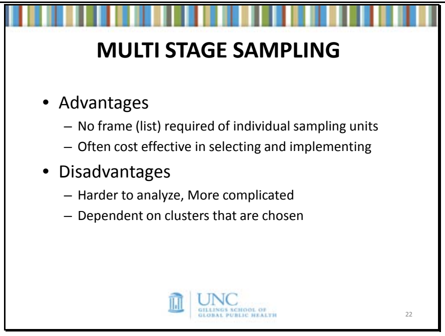
Slide 7	 **SELECTING AN SRS:** Method 3: Computer Generated • Large SRS or large population → random number table is inconvenient • Uniform random numbers (URNs) can be generated by almost all statistical software packages and many calculators – Use computer to calculate the uniform random numbers – Continue as in method 2	The most common method of selecting a SRS in practice is computer-generated. So instead of using a random number table, you use a computer to generate the random numbers. This computer method is almost the same as Method 2… In this method, you let the computer or calculator generate the uniform random numbers between 0 and 1. Any statistical software and lots of calculators have this feature. Then, you multiply the uniform random number by *N* (population size) and take next larger integer, as in method 2. With a little programming, you can set up a loop to select even big sample very quickly.
Slide 8	 **MORE COMMENTS ABOUT AN SRS** • An SRS requires you to be able to list (enumerate) all the units in the population→ Big disadvantage. • Record the way your random numbers were selected (row or table, seed….) • A SRS may look like it has 'patterns'…. that's OK	A couple of general notes about SRSs: 1: One big disadvantage of the selecting a SRS in general is that you must be able to list all the units in the population. This is a pretty big disadvantage which will be addressed in future lessons with multi-stage sampling. 2: Remember to make sure that you record how you selected your sample- in other words how you found the random number(s). For example, if you are using the tables, record the row number. If using the computer, record the "Seed". - So you can reproduce the sample, if necessary. 3: I noted previously that you can't tell by looking at a sample whether it is selected at random. This is true. However, I sometimes do an exercise in a residential class that shows that selection at random may seem like it has a pattern. I have half the class get together and use one of the methods with a random number table to select a sample. Then, I ask the other half of the class to gather and select a sample by selecting sample which "looks" random to them. They write down IDs so that the IDs look random. I leave the room while they complete the task. When I return, I try to guess which sample was chosen truly at random (using a table) and which was created to look random. Usually I am able to do it – the truly random sample may look like it has anomalies… IDs selected in a row (like 5,6,7), or many more even numbers in a row, or have big gaps, for example. I recently read a Newsweek article that lots of people think their iPods have favorites – when the iPod is supposed to shuffle the tunes at random, it may seem repeat one artist or music type while ignoring a particular tune for what seems like forever. I digress… but the point is… use truly random numbers from a table or computer. Even if your sample seems to have a pattern, that is OK. True random sampling, like random shuffling on an iPod, may 'feel' like it plays favorites.

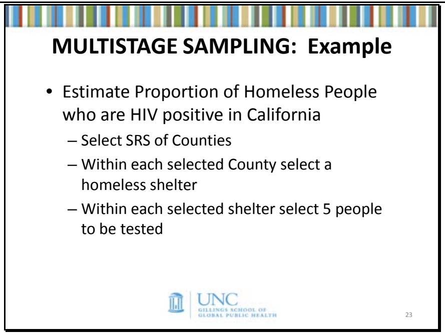
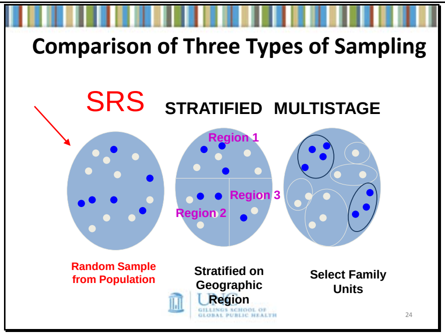
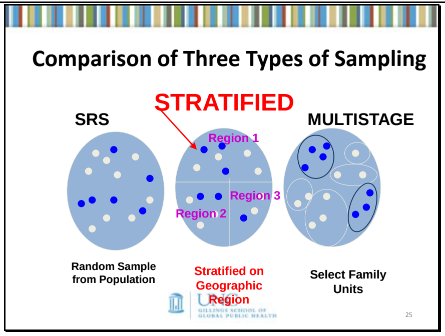
Slide 9	 STRATIFIED RANDOM SAMPLING	Sometimes, our study calls for a more complicated sampling scheme than SRS. One such scheme is Stratified Random Sampling.
Slide 10	 STRATIFIED RANDOM SAMPLING • Divide population into strata or groups [groups are similar in some characteristics] Typical strata: ethnicity, age, geographic region, gender, … • Select a sample within each stratum separately • Combine samples in each stratum for final sample	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. Recall that with a stratified sample, we are not selecting a subset of the strata – we are selecting observations WITHIN each the strata. In other words, if we stratify on ethnicity (5 groups) we don't, say, select 2 out 5 ethnicity, rather we select observations within each of the 5 ethnicities. But we can have some flexibility about how we select within each different ethnicity/strata.
Slide 11	 PROPERTIES OF A STRATIFIED RANDOM SAMPLE • A stratified random sample is never an SRS. • Common to make stratum sample sizes equal rather than sampling fractions	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. With a Stratified Random Sample, there are some samples that are not possible (probability of selection is zero). It's common to make stratum sample sizes equal rather than sampling fractions equal.

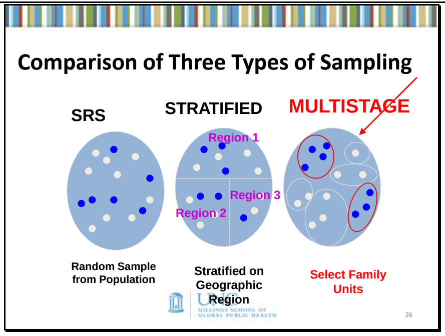
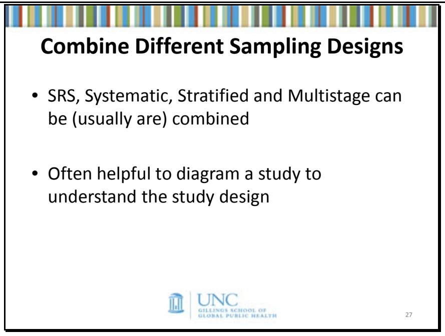
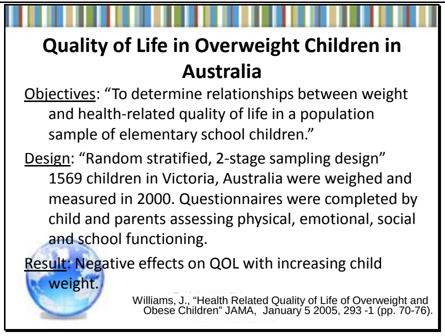
Slide 12	 **EXAMPLE: STRATIFIED RANDOM SAMPLE** • Select 100 patients out of 1000 patients → select equal numbers of diabetics and nondiabetics • Form strata 200 diabetics and 800 nondiabetics • Select 50 diabetics and 50 nondiabetics – SRS of 50 out of 200 diabetics – SRS of 50 out of 800 nondiabetics • Full sample is *n*= 100 out of *N*=1000 12	We saw in an example that we may want to stratify on diabetic status (Diabetic vs. Non-Diabetic) Suppose there are 1000 patients and we wish to select 100 patients for our sample. Suppose the population has 200 diabetics and 800 nondiabetics – these are our two strata. Suppose we need to select equal numbers of diabetics and nondiabetics. So select 50 diabetics and 50 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 13	 **NOTATION: 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 13	Recall our notation for stratified samples. 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 14	 **WHY STRATIFY?** • Analyze strata separately • If the strata are homogeneous, we can add efficiency • Convenience – different sampling plan for different strata 14	Why do we stratify? One reason, is so that we can analyze strata separately- we can assure that we have sufficient sample size within the strata. We can force by design to have enough observations in a certain strata (diabetic status, ethnicity, etc). (In the previous example we assure that we have enough diabetics to analyze separately.) If strata are homogeneous, we can add efficiency – this means that we can get smaller estimates of variance with the stratified sample as compared to using a SRS (under some assumptions). The details go beyond what we cover in Bios 600, but one thing to remember is that stratification can potentially be beneficial not just logistically but can be beneficial by decreasing the necessary sample size or decreasing variability of estimates if the strata are “homogeneous” (meaning similar with respect the outcome of interest). We also know that with stratified sampling, we could use not

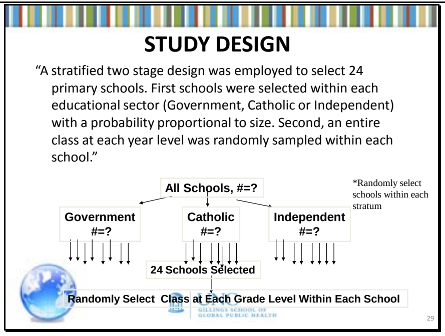
		only different sampling fractions but also different sampling methods or logistical details. Perhaps we stratify on age group. Then not only can we have a different sampling fraction for, say, the older age group but we could have a totally different sampling method….we could use systematic sampling for the older age group and a SRS for the other age group… In general we want a small number of strata that are homogeneous.
Slide 15	 **OVERSAMPLING AND UNDERSAMPLING** • First scenario → selected 50 diabetics and 50 nondiabetics – Over-sampled diabetics $n_1/N_1 > n/N$ – Under-sampled nondiabetics, $n_2/N_2 < n/N$ UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 15	We have seen it is possible to oversample or undersample the different strata. 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. 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 n/N = 50/800= 0.0625. OK… So the diabetics are over-sampled and the non-diabetics are under-sampled. NOTE: Under-coverage (in a previous lesson) and under-sampling are different concepts.
Slide 16	 **EXAMPLE: STRATIFIED RANDOM SAMPLE** • Another scenario n_1= 35 diabetics, n_2= 65 nondiabetics • Diabetics are oversampled n_1/N_1 = 35/200= 0.18 > n/N = 100/1000 • Non-diabetics are undersampled n_2/N_2 = 65/800=0.08 < n/N =100/1000 UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 16	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.

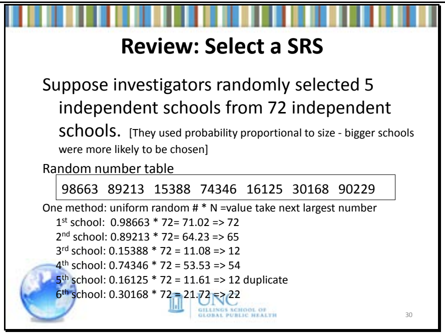
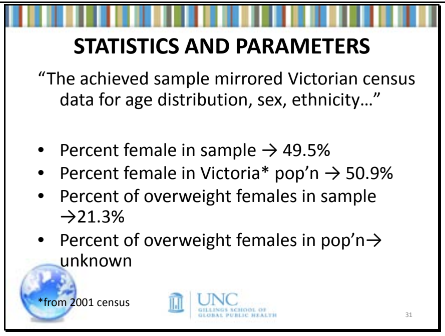
Slide 17	 **STRATIFIED RANDOM SAMPLE EXAMPLE (cont)** • Study investigates the proportion of subjects who achieved a target weight • Let m_i be the number of subjects in the sample within the *i*th stratum to reach a target weight • Suppose m_1=10 diabetics out of 35 reach target weight and m_2=30 out of 65 non-diabetics reach target weight. • *BIASED* ESTIMATE IN POPULATION: 40/100 reach target weight. Expect 400 out of 1000 to reach target weight in population? NO. UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 17	Suppose that a study investigates the proportion of subjects who achieved a target weight. When we talk about sampling a good bit of time is spent in how to select the sample, but remember eventually we have some dependent variable/ some outcome variable that we want to measure…. And then we want to know not just about the outcome in the sample, but the outcome in the population. 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. Reviewing the notation…. let m_i be the number of subjects in the sample within the *i*th 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 (incorrectly) 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. So you need to take into consideration the stratified sampling design.
Slide 18	 **UNBIASED ESTIMATE OF PROPORTION** ● Proportion of diabetics reaching target weight: 10/35 = 0.286 ● Proportion of nondiabetics reaching target weight: 30/65 = 0.462 ● Weight these proportions reaching target weight by proportion of diabetics (200/1000) and proportion of nondiabetics (800/1000) in pop'n ● Unbiased estimate of proportion reaching target weight in population 0.286(0.2) + 0.462(0.8) = 0.0572 + 0.3696 = 0.4268 (or about 43%)	We can compute the unbiased estimate of proportion in population reaching target weight. 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 in the sample. 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 + 0.3696 = 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.

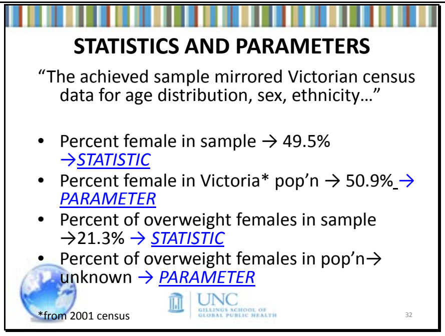
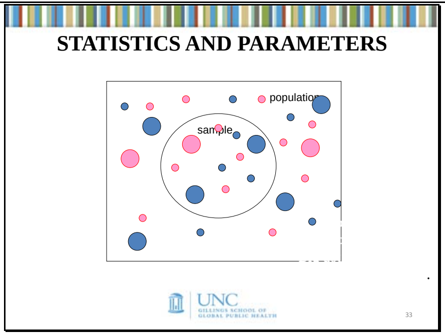
Slide 19	 DISADVANTAGES OF STRATIFICATION - Makes analysis more difficult (costly) - May offer no advantages (no appropriate stratification variables)	Makes analysis more difficult (costly) Stratification may offer no advantages – there may be no variable on which is makes sense to stratify – no variable where the observations are “homogeneous”– or may be expensive measure the stratification variable.
Slide 20	 MULTISTAGE SAMPLING	Another method of sampling that is often used is multistage sampling. Another term for this sampling design is **cluster sampling.**
Slide 21	 MULTISTAGE SAMPLING - Divide target pop’n into groups, usually called *clusters* - First Stage: Select a sample of the clusters – primary sampling units (PSUs) - Second Stage: Select all units within those clusters OR Select a subset of units within the clusters (secondary sampling units - SSUs)	Moving to a different sampling strategy, we will investigate multistage sampling. In multistage sampling, the target population is separated into groups, often called clusters. In the first stage, we randomly select a sample of the clusters (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 a subset of units within the clusters. These selections are called secondary sampling units or SSUs.
Slide 22	 MULTI STAGE SAMPLING - Advantages – No frame (list) required of individual sampling units – Often cost effective in selecting and implementing - Disadvantages – Harder to analyze, More complicated – Dependent on clusters that are chosen	On the positive side, with multistage sampling, no frame (list) required of individual sampling units Also, multistage sampling is often cost effective in selecting and implementing. However, multistage design are harder to analyze and just more complicated (expensive). Of course, this method is dependent on which clusters are chosen – so you choose your clustered carefully and in an unbiased way (you do this always, anyway!), but special care needs to be taken in cluster sampling.

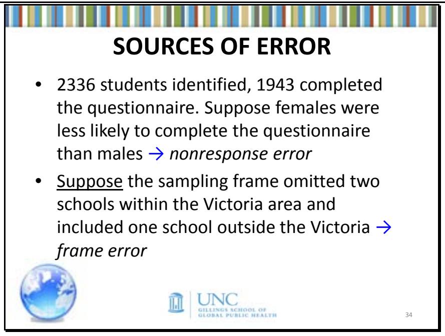
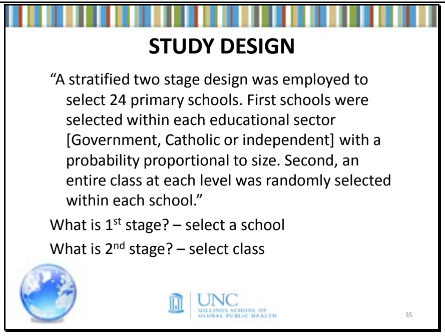
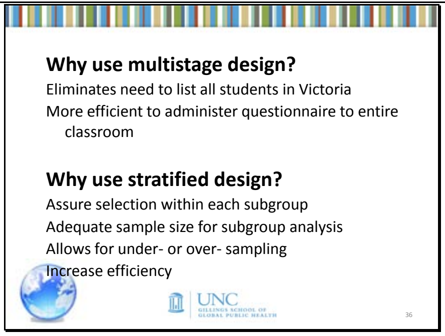
Slide 23	 **MULTISTAGE SAMPLING: Example** • Estimate Proportion of Homeless People who are HIV positive in California – Select SRS of Counties – Within each selected County select a homeless shelter – Within each selected shelter select 5 people to be tested UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 23	Estimate Proportion of Homeless People who are HIV positive in California Select SRS of Counties Within each selected County select a homeless shelter Within each selected shelter select 5 people to be tested
Slide 24	 **Comparison of Three Types of Sampling** SRS STRATIFIED MULTISTAGE Region 1 Region 3 Region 2 Random Sample from Population Stratified on Geographic Region Select Family Units UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 24	We've reviewed several types of sampling… let's compare three of the types that we've discussed. Suppose each of these ovals represents the population– copies of the SAME population…the dark blue dots are units within the population that are sampled and the white 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 can think of each dot as a person… and each person is part of a geographic region and a family which may play a part in the sampling scheme. 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 dark blue dots.
Slide 25	 **Comparison of Three Types of Sampling** SRS STRATIFIED MULTISTAGE Region 1 Region 3 Region 2 Random Sample from Population Stratified on Geographic Region Select Family Units UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 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.

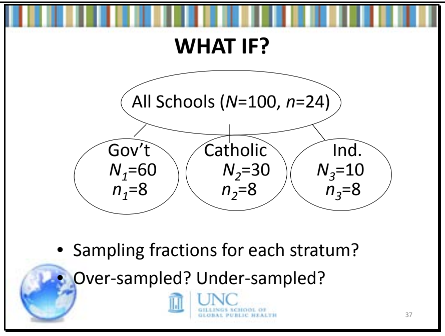
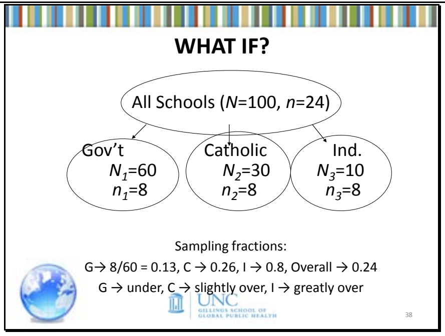
Slide 26	 Comparison of Three Types of Sampling SRS STRATIFIED MULTISTAGE Region 1 Region 2 Region 3 Random Sample from Population Stratified on Geographic Region Select Family Units 26	We could also use multistage sampling. Here we view the population as having clusters… say family units are clusters 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	 Combine Different Sampling Designs • SRS, Systematic, Stratified and Multistage can be (usually are) combined • Often helpful to diagram a study to understand the study design 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.
Slide 28	 Quality of Life in Overweight Children in Australia Objectives: "To determine relationships between weight and health-related quality of life in a population sample of elementary school children." Design: "Random stratified, 2-stage sampling design" 1569 children in Victoria, Australia were weighed and measured in 2000. Questionnaires were completed by child and parents assessing physical, emotional, social and school functioning. Result: Negative effects on QOL with increasing child weight. Williams, J., "Health Related Quality of Life of Overweight and Obese Children" JAMA, January 5 2005, 293 -1 (pp. 70-76).	Let's turn to a global health example regarding the Quality of Life of Overweight Children in Australia which used many of the sampling topics we've discussed. For the best understanding, please read articles referenced at the end of the slide presentation if possible. This is the same example that we've referred to in several of the lessons. The relevant portions are summarized for you, but you can read the articles listed in the reference for more information. Very briefly, let's start with the objectives of the study. The primary objective was to determine relationship between weight and health related quality of life in a population sample of elementary school children in Australia. The design was a random two-stage sampling design. 1569 children in Victoria Australia were weighed and measured in the year 2000. The children and the parents completed questionnaires about physical, emotional, social and school functioning. The questionnaire results showed that increasing child weight was associated with decreased quality of life in this population. More results will be given near the end of this presentation. Let's look at the sampling design of this study and how it illustrates much of what we've learned (although I may make/change some assumptions about the original study or simplify the design to make a point about what we've talked about in sampling.)

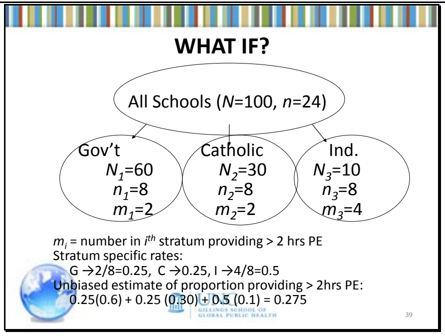
Slide 29	 STUDY DESIGN “A stratified two stage design was employed to select 24 primary schools. First schools were selected within each educational sector (Government, Catholic or Independent) with a probability proportional to size. Second, an entire class at each year level was randomly sampled within each school.” All Schools, #=? *Randomly select schools within each stratum Government #=? Catholic #=? Independent #=? 24 Schools Selected Randomly Select Class at Each Grade Level Within Each School 29	A quote from the article….. “A stratified two-stage design was employed to select 24 primary schools. First schools were selected within each educational sector (Government, Catholic or Independent) with a probability proportional to size. Second, an entire class at each year level was randomly sampled within each school.” Let’s see if we can get a mental picture of this description. Investigators obtained a sampling frame of all primary schools in Victoria. The schools were categorized as either Government, Catholic or Independent. These three categories were the strata. Within each stratum, schools were selected for a total of 24 schools. We cannot know from the information that I’ve given so far how many schools were selected within each stratum – just that the total number of schools is 24. It could be that eight schools were selected in each stratum, or not. In the diagram I show 8 school from each stratum, but we really don’t know this from the article….. We also don’t know how many schools were on the sampling frame. This is often the case in the real world – investigators are limited in the amount of information they can include in an article so the reader is left to wonder a bit. So, we can see, so far, this is a stratified design. Also, we can see that the study is a multistage (2 stage) design. First schools are selected (they are the PSU’s the primary sampling units). The second stage is to select an entire class at each year level with in the school. So, classrooms are SSU (secondary sampling units). The description indicates schools were selected with a “probability proportional to size”. This means that clusters (schools) which are bigger have a greater probability of being selected and that probability is proportional to the size of the cluster (school). Some questions to think about so far….what are the sampling frames in this example? Why did they use a multistage design? Why did they use a stratified design? What are the implications if a school did not participate? Was that school replaced by another school in that stratum? Did they use SRS to select classrooms within the schools? How many primary schools were in Victoria and is 24 a pretty good representation? (We will of course talk about sample size more formally, but I also think it is helpful to just “ball park” it - 24 school were selected out of how many?) How did they define ‘primary school’? In reading an article, or designing a study yourself, careful consideration of questions like this can be helpful (instructive?).

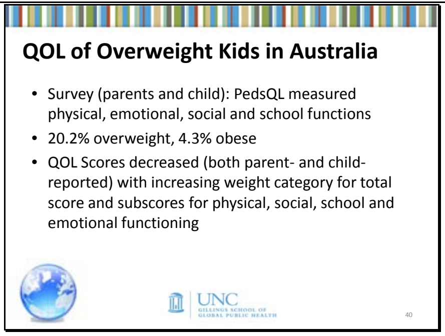
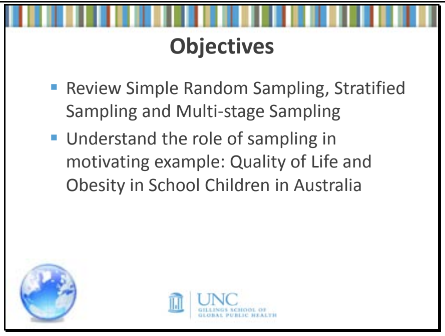
Slide 30	 **Review: Select a SRS** Suppose investigators randomly selected 5 independent schools from 72 independent schools. [They used probability proportional to size - bigger schools were more likely to be chosen] Random number table 98663 89213 15388 74346 16125 30168 90229 One method: uniform random # * N =value take next largest number 1st school: 0.98663 * 72= 71.02 => 72 2nd school: 0.89213 * 72= 64.23 => 65 3rd school: 0.15388 * 72 = 11.08 => 12 4th school: 0.74346 * 72 = 53.53 => 54 5th school: 0.16125 * 72 = 11.61 => 12 duplicate 6th school: 0.30168 * 72 = 21.72 => 22 30	This is a review of how to select a SRS. Let's suppose investigators randomly selected a SRS of independent schools from 72 independent schools. [The actual study used probability proportional to size where bigger schools were more likely to be chosen- but we'll simplify that and use simple random sampling] Let's list the first 5 schools that would be selected, out of the total sample size (like 8) Once we got uniform random numbers from a random number table, we multiply the number by N and then take the next larger integer. The first 5 schools to be selected are {72,65,12,54,22,…}
Slide 31	 **STATISTICS AND PARAMETERS** "The achieved sample mirrored Victorian census data for age distribution, sex, ethnicity…" • Percent female in sample → 49.5% • Percent female in Victoria* pop'n → 50.9% • Percent of overweight females in sample →21.3% • Percent of overweight females in pop'n→ unknown *from 2001 census 31	Let's talk a bit about statistics and parameters. The article states: "The achieved sample mirrored Victorian census data for age distribution, sex, ethnicity…" Why is this important? Why did the authors take the opportunity to tell the reader this? Well, the authors want to show that their sample is a good representation of the population of interest. In other words, from the census we know the age distribution, sex and ethnicity distribution etc." If our sample is representative, we would expect the sample to have similar ages, genders and ethnic background as the population. Suppose that: The percent female in from the final sample was 49.5%. And suppose that the percent female in Victoria was 50.9% (from 2001 census). Here we are not using a statistic to estimate a parameter – the parameter is known. We are using the statistic(s) to show that the sample is pretty representative. The percent of overweight females in sample is 21.3% according to the article. The percent of overweight females in population is unknown. [We can talk about whether 50.9% and 49.5% are statistically significant, but for now let's just say that they appear to be pretty close.]

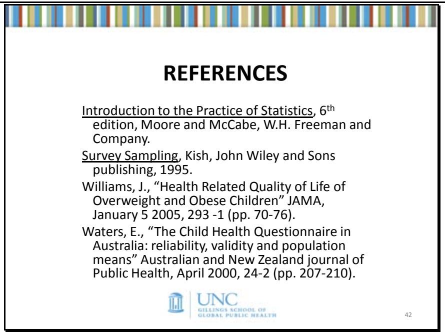
Slide 32	 STATISTICS AND PARAMETERS "The achieved sample mirrored Victorian census data for age distribution, sex, ethnicity…" • Percent female in sample → 49.5% →*STATISTIC* • Percent female in Victoria* pop'n → 50.9%_→ *PARAMETER* • Percent of overweight females in sample →21.3% → *STATISTIC* • Percent of overweight females in pop'n→ unknown → *PARAMETER* *from 2001 census UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 32	So to drive this point home, let's label them… 49.5 % female is a statistic (from the sample) 50.9 % female is a parameter (from the population) 21.3% overweight (within the females in the sample) is a statistic – it is measurable and describes the sample.. It varies from sample to sample… And the percentage overweight in the population of females is unknown - this value is one of the values we wish to estimate. Since the percent female in the sample is pretty close to the percent female in the population, we would like to think that the percent overweight in the sample is pretty close to the percent overweight in the population. It is a good sign that our sample is representative of the population.
Slide 33	 STATISTICS AND PARAMETERS UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 33	In conducting a sample, we want our sample to be representative of the population. Suppose the population only had 30 students and we are take a SRS of 10 students out of 30 students. In this picture, consider the pink dots to be female subjects, the blue dots to be males subjects, the big dots are overweight children and the small dots to be healthy weight children. There are 15 boys and 15 girls in the population. Go ahead and count the dots. Make sure I'm right! So in this (small) population there are 50% girls and 50% boys. If we've taken a representative sample then we'd expect our sample to have about the same proportions. Let's check.. In the sample (inside the ring), what is the percent of girls? (5 out of 10 or 50%) and the percentage of boys is 5 out of 10 or 50%. So the sample has the same gender distribution as the population. Our sample appears to be representative. Good! Convince yourself that the proportion of overweight females in the sample is 1 out of 5 (20%). You'll probably need to pause the slide show at this point. In the sample, there are 5 females and one is overweight…. The statistic is 20% of females in the sample are overweight. In the population we can see that 3 out of 15 females are overweight (=20%). Turning our attention to the fellas… 2 out of 5 in the sample are overweight (40%)… 6 out of 15 in the population are overweight (again 40%). You can also convince yourself that the proportion overweight in the sample (3 out of 10) is the same as the proportion overweight in the population (9 out of 30). This diagram is to show visually how a SRS should "look like" the population in the variable of interest. The sample should be representative of the population.

Slide 34	 **SOURCES OF ERROR** • 2336 students identified, 1943 completed the questionnaire. Suppose females were less likely to complete the questionnaire than males → *nonresponse error* • Suppose the sampling frame omitted two schools within the Victoria area and included one school outside the Victoria → *frame error* 34	Returning to the original study…. There are sources of error, let's look at a couple: Out of 2336 students identified, 1943 completed the questionnaire. Suppose females were less likely to complete the questionnaire than males. What type of error may occur? – nonresponse error Suppose the list of schools from which the sample was taken omitted two schools within the Victoria area and included one school outside the Victoria. This type of error is best described as what type of error?- frame error
Slide 35	 **STUDY DESIGN** "A stratified two stage design was employed to select 24 primary schools. First schools were selected within each educational sector [Government, Catholic or independent] with a probability proportional to size. Second, an entire class at each level was randomly selected within each school." What is 1st stage? – select a school What is 2nd stage? – select class 35	Let's go back to the study design "A stratified two stage design was employed to select 24 primary schools. First schools were selected within each educational sector [Government, Catholic or independent] with a probability proportional to size. Second, an entire class at each level was randomly selected within each school." Recall the 1st stage -- select a school The second stage is selecting a classroom within each grade level from the schools which we selected.
Slide 36	 **Why use multistage design?** Eliminates need to list all students in Victoria More efficient to administer questionnaire to entire classroom **Why use stratified design?** Assure selection within each subgroup Adequate sample size for subgroup analysis Allows for under- or over- sampling Increase efficiency 36	I asked these questions earlier, now let's answer them. Why do the investigators use multistage design? It eliminates need to list all students in Victoria. It is more efficient to administer questionnaire to entire classroom. Why do the investigators use a stratified design? In order to assure selection within each subgroup. It can provide a adequate sample size for subgroup (school type) analysis. It allows for under or over sampling within the strata. We can increase efficiency if the strata are homogeneous. You'll just have to trust me on that. There are downsides, of course. The analysis is more difficult. Also with this clustered (multistage) design a bigger sample size is usually needed. Again, we haven't gotten into the reasoning… just be aware there is "no free lunch"…. In statistics, as in most things, there are tradeoffs.

Slide 37	 **WHAT IF?** All Schools (*N*=100, *n*=24) Gov't N_1=60 n_1=8 Catholic N_2=30 n_2=8 Ind. N_3=10 n_3=8 • Sampling fractions for each stratum? • Over-sampled? Under-sampled? 37	Let's have some more fun with this example by making some assumptions… Let's simplify this example in order to practice some of the material we've learned. Let's just look at the first part of the design the stratification. We'll just consider this as a stratified design– with the strata being school type. The investigators pick schools from each stratum (Government, Catholic, Independent). Further, for simplicity, let's suppose there are 100 schools in the ' population' of schools – it is nice round number. 100 primary schools in Victoria to choose from. We are going to select 24 schools out of 100 possible schools. Also, assume that there are 60 Government, 30 Catholic, 10 Independent (we don't actually know this number, but suppose this is the breakdown for conversation sake) (You can think of the as proportions (0.6 Government, 0.3 catholic and 0.1 independent) or percentages (60% Gov't, 30% Catholic and 10% Independent). Suppose we select 8 schools in each stratum. We've talked about previously how often investigators will make the sample size in each stratum be equal rather than the sampling fraction. That brings up an interesting question… what are the sampling fractions? Are there strata which are under or over sampled? Are we all comfortable with the notation so far? Try these questions on your own first and then I'll address them on the next slide.
Slide 38	 **WHAT IF?** All Schools (*N*=100, *n*=24) Gov't N_1=60 n_1=8 Catholic N_2=30 n_2=8 Ind. N_3=10 n_3=8 Sampling fractions: G→ 8/60 = 0.13, C → 0.26, I → 0.8, Overall → 0.24 G → under, C → slightly over, I → greatly over 38	Let's compute the stratum specific sampling fractions. We are going to calculate the sampling fraction within each group. ….“stratum specific” means for each stratum (each group) sampling fraction (recall the sampling fraction is sample size/ population size or n/N). For the gov't schools the sampling fraction is 8/60 = 0.13 (13%). For the Catholic schools the sampling fraction is 8/30 or 26% and the independent schools have a sampling fraction of 0.8. The overall sampling fraction is 24/100 or 0.24. The Government schools are undersampled (because their sampling fraction is less than 0.24. The independent schools are greatly oversampled -their sampling fraction is 0.8 much greater than 0.24.

Slide 39	 **WHAT IF?** All Schools (N=100, n=24) Gov't N_1=60 n_1=8 m_1=2 Catholic N_2=30 n_2=8 m_2=2 Ind. N_3=10 n_3=8 m_3=4 m_i = number in i^{th} stratum providing > 2 hrs PE Stratum specific rates: G →2/8=0.25, C →0.25, I →4/8=0.5 Unbiased estimate of proportion providing > 2hrs PE: 0.25(0.6) + 0.25 (0.30) + 0.5 (0.1) = 0.275 GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 39	Further, suppose may we collect data at the school level. Let's review calculating an unbiased estimate for a proportion in a stratified sample. We may ask them "Does your school provide greater than 2 hours of Physical Education for students per week?" Let *mi* be the number of schools in the *ith* stratum in the sample that provide 2 hours of PE for student per week. So in the government sample of 8 schools we might find that 2 schools provide this level PE (0.25) In the Catholic schools in the sample, also 2 schools out of 8 provide > 2 hours of PE (0.25) And among the 8 independent schools, 4 provide that level of PE (0.5) Next we want to calculate an unbiased estimate of the proportion of schools in the population providing that level of PE. We know the proportion in the sample 8 out of 24. We want to know about the population of 100 schools… Recall we weight stratum specific use of the program by the stratum specific sampling fractions. In English, we want to weight those values, 0.25 0.25 and 0.5 by the proportion in each stratum in the population. 0.25(0.6) + 0.25 (0.30) + 0.5 (0.1) = 0.15 +0.075 + 0.05 = 0.275 What does this mean… in the population of 100 schools, we'd expect 0.275 of the schools to be providing > 2 hrs of PE… or approximately 28 schools out of 100.

Slide 40	 **QOL of Overweight Kids in Australia** - Survey (parents and child): PedsQL measured physical, emotional, social and school functions - 20.2% overweight, 4.3% obese - QOL Scores decreased (both parent- and child-reported) with increasing weight category for total score and subscores for physical, social, school and emotional functioning 40	So what did this actual stratified, 2 stage sampling design tell us about the Quality of Life and Obesity among Children in Australia? Both the children and parents were given a survey called the PedsQL which measured physical, emotional, social and school functioning. The 1,456 children who participated in the survey were primarily 9- to 12- years old. Approximately 20.2% of children in the sample were classified as overweight, 4.3% were classified as obese. Obesity is not just a US problem but is an issue in other countries, particularly developed countries, as well. The paper reports that "Parent and child perceptions were strikingly similar with obese children having the lowest summary and subscale scores." Quality of Life Scores decreased (both parent- and child- reported) with increasing weight category for total score, physical, social, school and emotional functioning (although not all decreases were statistically significant.) Health related quality of life (as measured by the PedsQL instrument) was lower for children in the heavier weight categories with the obese children in general having the lowest scores in each of the subcategories. So childhood obesity (outside the US) is an important public health issue not only for the physical impact on the child's health but also for the emotional and psychological toll that associated with being overweight. The sampling methods we've learned are used in many studies – such as this one with global impact – to select subjects in sample to tell us about associations in the population we are interested in. [Williams, J., "Health Related Quality of Life of Overweight and Obese Children" JAMA, January 5 2005, 293 -1 (pp. 70-76).]
Slide 41	 **Objectives** - Review Simple Random Sampling, Stratified Sampling and Multi-stage Sampling - Understand the role of sampling in motivating example: Quality of Life and Obesity in School Children in Australia	

Slide 42	 **REFERENCES** Introduction to the Practice of Statistics, 6th edition, Moore and McCabe, W.H. Freeman and Company. Survey Sampling, Kish, John Wiley and Sons publishing, 1995. Williams, J., "Health Related Quality of Life of Overweight and Obese Children" JAMA, January 5 2005, 293 -1 (pp. 70-76). Waters, E., "The Child Health Questionnaire in Australia: reliability, validity and population means" Australian and New Zealand journal of Public Health, April 2000, 24-2 (pp. 207-210). UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH 42	