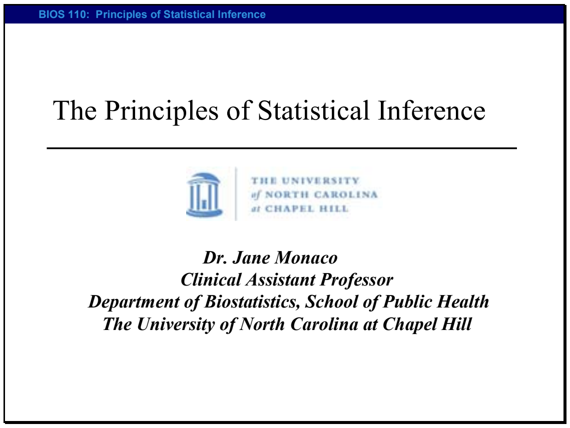
Slide 1



BIOS 110: Principles of Statistical Inference

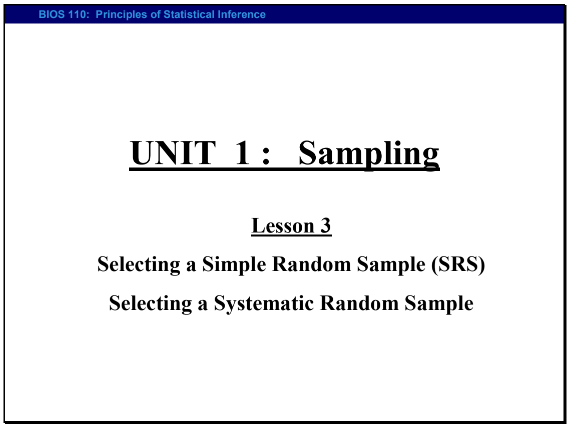
The Principles of Statistical Inference

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Dr. Jane Monaco
Clinical Assistant Professor
Department of Biostatistics, School of Public Health
The University of North Carolina at Chapel Hill

Welcome back to The Principles of Statistical Inference- the online version of the introductory course in biostatistics offered by the Department of Biostatistics at The University of North Carolina at Chapel Hill.

Slide 2



BIOS 110: Principles of Statistical Inference

UNIT 1 : Sampling

Lesson 3

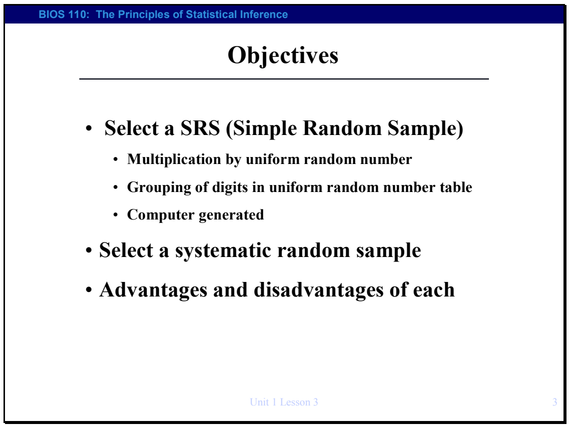
Selecting a Simple Random Sample (SRS)

Selecting a Systematic Random Sample

We continue the unit on Sampling, in other words, “How do we select a group of individuals to study?”

This lesson, Lesson 3, will address how to select a simple random sample (SRS) and how to select a systematic random sample.

Slide 3



BIOS 110: The Principles of Statistical Inference

Objectives

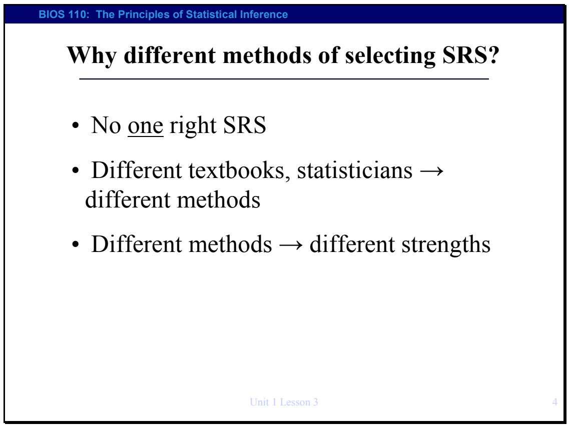
- **Select a SRS (Simple Random Sample)**
  - **Multiplication by uniform random number**
  - **Grouping of digits in uniform random number table**
  - **Computer generated**
- **Select a systematic random sample**
- **Advantages and disadvantages of each**

Unit 1 Lesson 3 3

The objectives are to be able to select an SRS and a systematic random sample.
We will learn several ways to select a Simple Random Sample (SRS). We can multiply by a uniform random number, we can group the digits from uniform random numbers table, or we generate a random sample by computer.
Another method of sampling is called systematic random sampling, and we’ll learn how to select this type of sample as well.
Then you should know that the different types of sampling may have advantages and disadvantages depending on the sample and population of interest.

This is a good time to note that the abbreviation ‘SRS’ refers to a simple random sample. There are several other sampling types which start with “S”. So the abbreviation can be confusing. We’ll just need to make the best of it- nothing we can do about it. “Systematic Random Sample” doesn’t have an abbreviation (because “SRS” is already taken). We’ll need to always write out “Systematic Random Sample”.

Slide 4



BIOS 110: The Principles of Statistical Inference

Why different methods of selecting SRS?

- No one right SRS
- Different textbooks, statisticians → different methods
- Different methods → different strengths

Unit 1 Lesson 3 4

I think I just heard someone groan when I said that we’d learn a couple different methods for selecting an SRS….

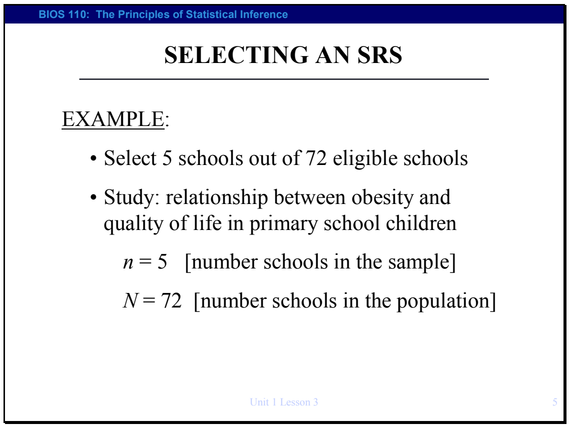
That person may have asked, “After all, this is a statistics / math-type class. Isn’t there a right or wrong answer?....

Why do we have different methods of selecting a SRS? Shouldn’t there just be one right method?”

There are a couple of reasons to learn different ways. We want to illustrate that there is no **one** right SRS. Also, different textbooks use different methods so depending on the textbook, different methods may be presented. In addition, there are often reasons to use one method rather than another because there are advantages and disadvantages to the different methods.

Usually we don’t refer “selecting **the** SRS”, rather we would say “selecting **a** SRS”.

Slide 5



BIOS 110: The Principles of Statistical Inference

SELECTING AN SRS

EXAMPLE:

- Select 5 schools out of 72 eligible schools
- Study: relationship between obesity and quality of life in primary school children

$n = 5$ [number schools in the sample]

$N = 72$ [number schools in the population]

Unit 1 Lesson 3 5

Let’s look at an example.

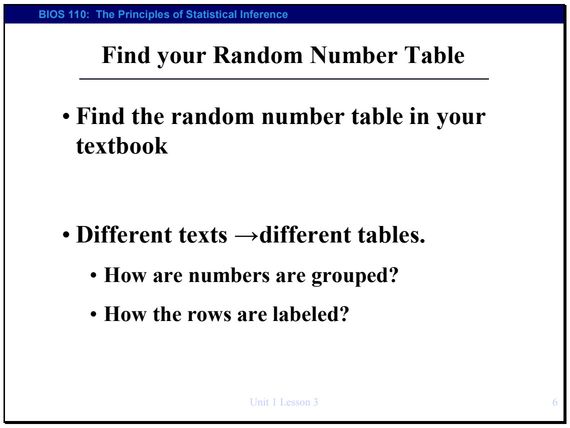
Suppose that you wish to investigate the relationship between childhood obesity and the quality of life in a school-age population. You wish to select 5 schools out of a total of 72 eligible schools in which to conduct your survey. In this example, the sample size is 5 and the population size is 72.

n=5 and *N*=72. What is the sampling fraction?

n/N=5/72.

This example is based on an article in reference at the end of the slides. We’ll use this example to illustrate the different methods.

Slide 6



BIOS 110: The Principles of Statistical Inference

Find your Random Number Table

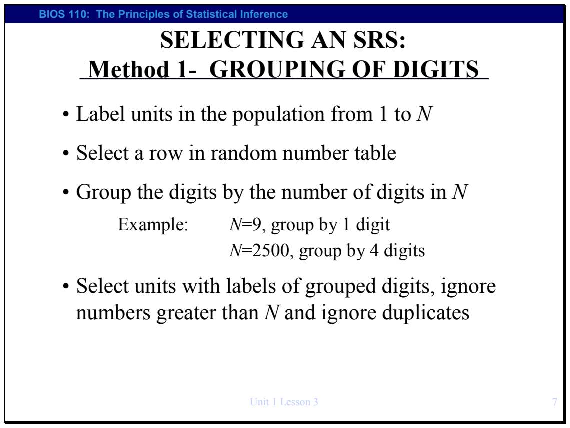
- **Find the random number table in your textbook**
- **Different texts →different tables.**
  - **How are numbers are grouped?**
  - **How the rows are labeled?**

Unit 1 Lesson 3 6

You'll need to find the random number table in your textbook. Almost all introductory statistics textbooks will have a random number table. Before proceeding, please pause the presentation and find the random number table in your textbook.
Different texts have different tables usually in the back or an appendix, but take a look at your book … in the random number table see how the numbers are grouped and how the rows are labeled.

Did you pause the presentation and find it? Good! If not… Go ahead, look for the table… I'll wait.

Slide 7



BIOS 110: The Principles of Statistical Inference

SELECTING AN SRS:
Method 1- GROUPING OF DIGITS

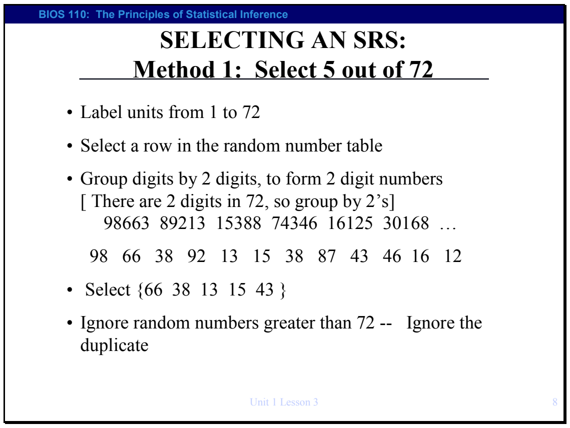
- Label units in the population from 1 to *N*
- Select a row in random number table
- Group the digits by the number of digits in *N*

  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

Unit 1 Lesson 3 7

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 the example will show that it is pretty straightforward.

Slide 8



BIOS 110: The Principles of Statistical Inference

SELECTING AN SRS:
Method 1: Select 5 out of 72

- Label units from 1 to 72
- Select a row in the random number table
- Group digits by 2 digits, to form 2 digit numbers
  [There are 2 digits in 72, so group by 2's]
  98663 89213 15388 74346 16125 30168 …

  98 66 38 92 13 15 38 87 43 46 16 12
- Select {66 38 13 15 43 }
- Ignore random numbers greater than 72 -- Ignore the duplicate

Unit 1 Lesson 3 8

For example,
Select 5 schools out of 72.
Label schools from 1 to 72 (the schools can be ordered or not ordered –alphabetical, size, etc.).
Group digits in random number table by 2 digits. [There are 2 digits in 72, so group by 2's] . The default is to group by 5 digits in my table (Other tables may be different). Suppose that the row in the random number table that we happen to select starts out:

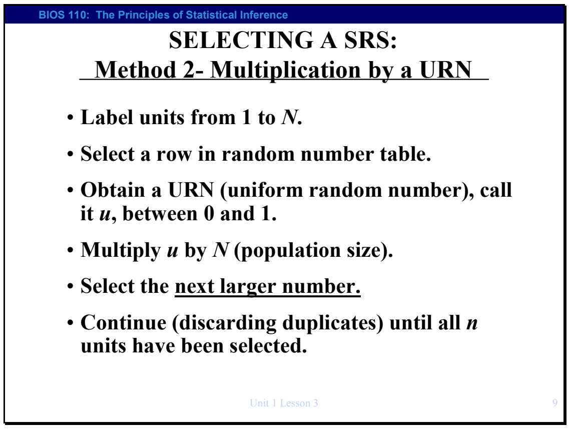
98663 89213 15388 74346 16125 ….

If we group these numbers by 2's, we will get

98 66 38 92 13 15 …

Starting from the left, skip 98… select 66… select 38… skip92… etc.
select {66 38 13 15 43 }
Ignore random numbers greater than 72 and ignore the duplicates.

Slide 9



BIOS 110: The Principles of Statistical Inference

SELECTING A 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 <u>next larger number.</u>**
- **Continue (discarding duplicates) until all *n* units have been selected.**

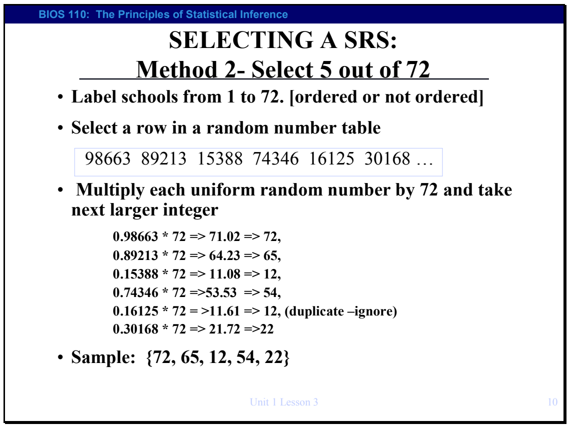
Unit 1 Lesson 3 9

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 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. The example, I hope, will clarify the instructions.

Slide 10



BIOS 110: The Principles of Statistical Inference

SELECTING A SRS:

Method 2- Select 5 out of 72

- **Label schools from 1 to 72. [ordered or not ordered]**
- **Select a row in a random number table**

98663 89213 15388 74346 16125 30168 …

- **Multiply each uniform random number by 72 and take next larger integer**

0.98663 * 72 => 71.02 => 72,
0.89213 * 72 => 64.23 => 65,
0.15388 * 72 => 11.08 => 12,
0.74346 * 72 =>53.53 => 54,
0.16125 * 72 = >11.61 => 12, (duplicate –ignore)
0.30168 * 72 => 21.72 =>22

- **Sample: {72, 65, 12, 54, 22}**

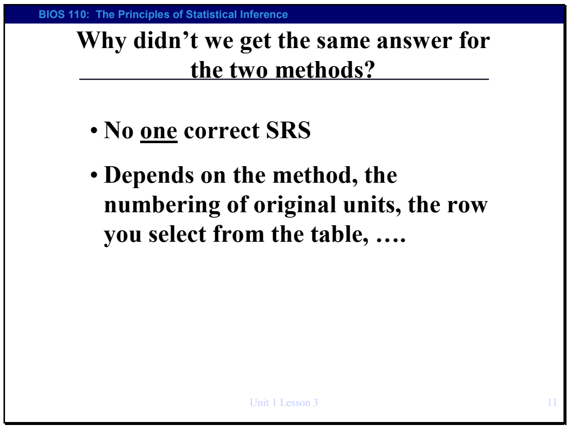
Unit 1 Lesson 3 10

These methods are much easier to explain using an example. So, here’s an example of method 2….
Return to the scenario in which we are interested in selecting 5 out of 72 schools. Label the schools from 1 to 72, and select a row in the random number table.
I am going to view the first numbers in this row as a number between 0 and 1 by inserting a decimal (0.98663). [Depending on the several things, you may wish to use a different number of digits, but we’ll not get into some of these details unless you are just curious…] Let’s group this row by the default of 5 digits.

Now multiply 0.98663 * 72 = 71.02 . Then you round up (take next largest integer) to 72. Select unit 72 to be in the sample.

Continue…. Noting that we are rounding up and <u>if we get a duplicate, just ignore it and move on</u>. Continue until we have selected 5 schools out of 72.
Our sample is {72,65,12,54,22}.

Slide 11



BIOS 110: The Principles of Statistical Inference

Why didn’t we get the same answer for the two methods?

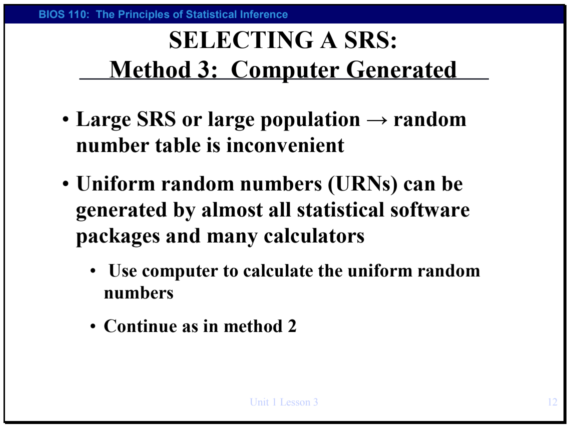
- **No <u>one</u> correct SRS**
- **Depends on the method, the numbering of original units, the row you select from the table, ….**

Unit 1 Lesson 3 11

Did we get same simple random samples with the two methods? No.

There is no one correct SRS. The sample will depend on, among other things, the way you originally ordered/labeled the population, the method, the row you select in random table,

Slide 12

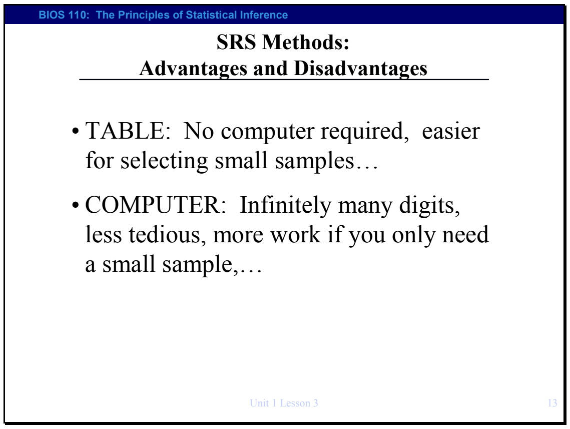


SELECTING A SRS:
Method 3: Computer Generated

- **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.
Using the previous methods are fine, and quite common, if you need just a small sample, but those methods are tedious otherwise. The computer generated way is easier to reproduce and perhaps less likely to produce errors.
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.
Are you going to have to do this on a test? No… I mention it because it is important to realize what is used in practice.

Slide 13



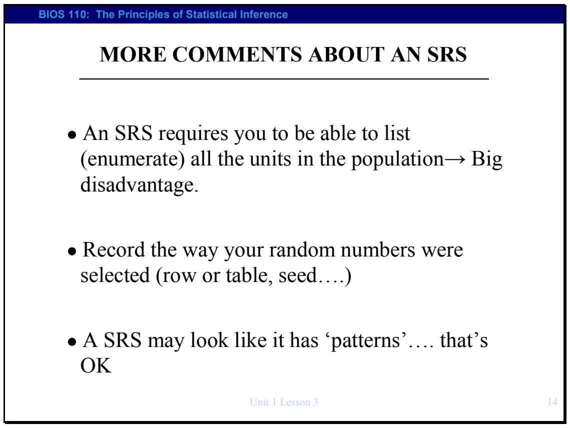
SRS Methods:
Advantages and Disadvantages

- TABLE: No computer required, easier for selecting small samples…
- COMPUTER: Infinitely many digits, less tedious, more work if you only need a small sample,…

The advantages of using the tables are that no computer is required, and it is easier if you just need a small sample. A data collector ‘in the field’ can do it ‘on the fly’. Also it is transparent what is happening.

The advantages for the computer-generated method are that there are infinitely many digits in the URN and that this method is less tedious if you need to select a bigger sample. The computer-generated method is not worth the trouble though if you just need a small sample.
SRS is like picking numbers for bingo --- label the pingpong balls in a bin and pull out “n” balls

Slide 14



BIOS 110: The Principles of Statistical Inference

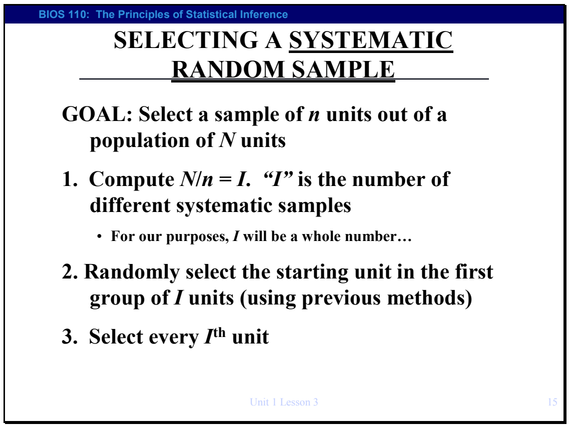
MORE COMMENTS ABOUT AN SRS

- 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

Unit 1 Lesson 3 14

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 calculated 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”.
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 15



BIOS 110: The Principles of Statistical Inference

SELECTING A <u>SYSTEMATIC RANDOM SAMPLE</u>

GOAL: Select a sample of *n* units out of a population of *N* units

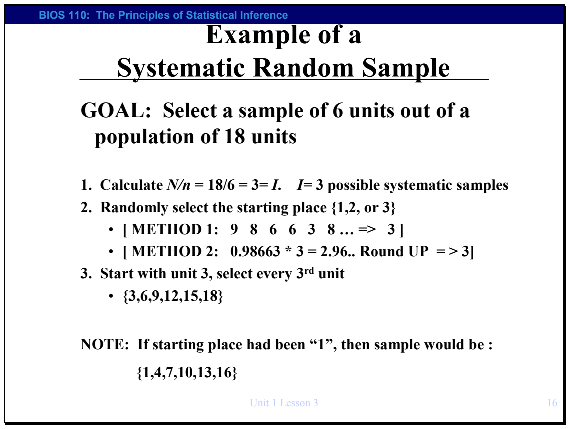
1. **Compute *N*/*n* = *I*. *"I"* is the number of different systematic samples**
   - **For our purposes, *I* will be a whole number…**
2. **Randomly select the starting unit in the first group of *I* units (using previous methods)**
3. **Select every I^{th} unit**

Unit 1 Lesson 3 15

Now we turn our attention to another type of sampling… a systematic random sample (<u>not</u> abbreviated SRS which we reserve for Simple Random Sample).

First the method, and then an example….
Suppose you wish to select a sample of *n* units out of a population of *N* units.
Start by computing *N/n=I*. *I* is the number of possible systematic samples. "*I*" will be an integer in our case. (There are other rules for when *I* is not an integer that we won't discuss here… again we'll just stick to a straightforward case and save more details for a course devoted completely to sampling…)
Select the starting unit by one of the previous methods.
Select every *I*th unit.

Slide 16



BIOS 110: The Principles of Statistical Inference

Example of a Systematic Random Sample

GOAL: Select a sample of 6 units out of a population of 18 units

1. **Calculate *N/n* = 18/6 = 3= *I*. *I*= 3 possible systematic samples**
2. **Randomly select the starting place {1,2, or 3}**
   - **[METHOD 1: 9 8 6 6 3 8 … => 3]**
   - **[METHOD 2: 0.98663 * 3 = 2.96.. Round UP = > 3]**
3. **Start with unit 3, select every 3rd unit**
   - **{3,6,9,12,15,18}**

NOTE: If starting place had been "1", then sample would be :

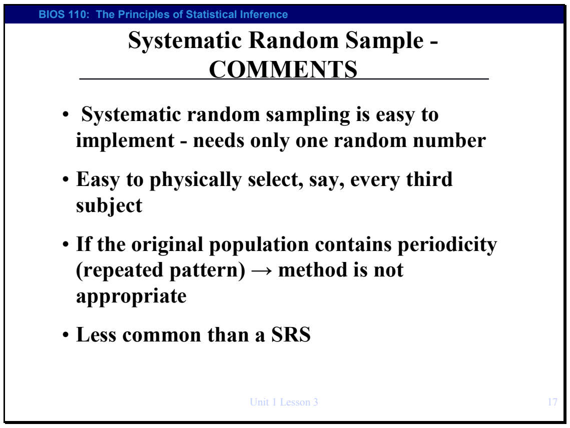
{1,4,7,10,13,16}

Unit 1 Lesson 3 16

Let's take a new example… suppose we wish to select 6 out of 18 units.
We calculate 18/6 = 3. *I*=3 is a whole number. (We'll only look at cases when *I* is a whole number. There are special rules when I is not an integer.) Out of the first (*I*=)3 units, I need to select one unit. You can use any of the previous methods that makes you happy to select the starting unit.
You can group the random numbers in this example by one digit… (you are selecting 1 out 3). Using the same row in the random number table in previous slides…. You get 9 8 etc. The first number between 1 and 3 is "3". So using this method 1, the starting place is 3.

Using the other method, multiply 0.98663 * 3 = 2.96, take the next largest integer =3. So using this method 2, the starting place is 3. (<u>It is just a coincidence that the two methods gave the same starting place.)</u>
A SYSTEMATIC RANDOM SAMPLE ONLY REQUIRES ONE RANDOM NUMBER.
Now start with unit 3 and select every 3^{rd} (*I*th)unit…. {3,6,9,12,15,18} is the sample of 6 out of 18.
[If our starting place were "1" then the sample would be {1,4,7,10,13,16}]

Slide 17



BIOS 110: The Principles of Statistical Inference

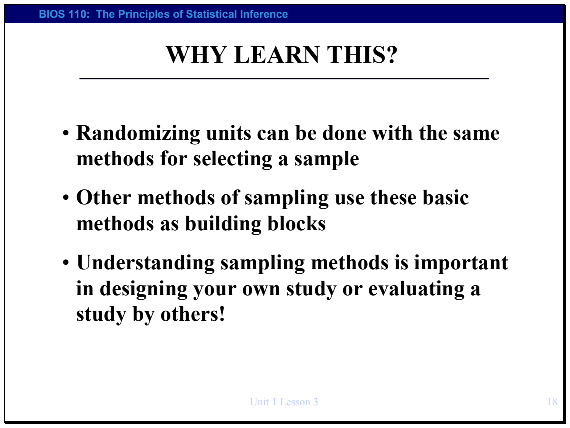
Systematic Random Sample - COMMENTS

- **Systematic random sampling is easy to implement - needs only one random number**
- **Easy to physically select, say, every third subject**
- **If the original population contains periodicity (repeated pattern) → method is not appropriate**
- **Less common than a SRS**

Unit 1 Lesson 3 17

Systematic random sample only requires one random number, easy to implement.
Easy to physically select, say, every third subject.
You may be thinking “Does it really matter? Is this going to be on the test? Does anyone really use these methods because it seems pretty silly to me? What am I going to have for dinner?”
I’ll address a couple of those questions.
Investigators very definitely use these methods! Sometimes you need a random sample of records, and it is much easier to pick every 5th record from a large filing cabinet than to use a random number generator and then comb through the filing cabinet for those IDs spread through out the file cabinet.
If the original list contains periodicity, this method is not (usually) appropriate – suppose you are selecting every 50th subject seen in a practice starting with patient number 8. Your systematic sample would be {8, 58, 108, …}. Suppose the practice sees about 50 patients a day. The patients then may have a periodicity (some repeated pattern) in the order they come in to the clinic …In that patients seen in the early morning may differ in ways of interest from patients seen in the afternoon. (More urgent?, Less likely to be full time employed?, New consults? …) Systematic random sampling may give you only patients seen in the morning for example which would not be representative of the population.
Systematic sampling is much less common than SRS.

Slide 18



BIOS 110: The Principles of Statistical Inference

WHY LEARN THIS?

- **Randomizing units can be done with the same methods for selecting a sample**
- **Other methods of sampling use these basic methods as building blocks**
- **Understanding sampling methods is important in designing your own study or evaluating a study by others!**

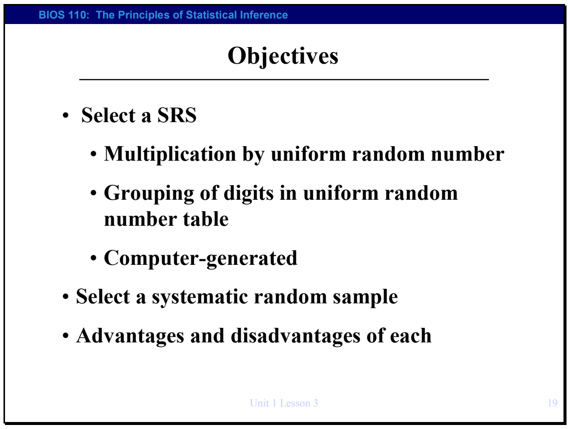
Unit 1 Lesson 3 18

We've looked at a couple of ways to select samples. Why do you need to learn this?
A few reasons, not necessarily in the order of importance..
We'll see these concepts again when we are talking about randomizing members to treatment groups. So please don't forget these methods even after we are done with the sampling unit. The methods for randomizing units is practically the same. Randomizing subjects to treatment or control can be thought of as selecting a sample out of the population to receive the treatment.

Also another reason for learning these methods is that other sampling methods we will study in coming lessons also will use SRS and systematic random sampling as a building blocks.

Finally, understanding sampling methods is important in designing your own study or evaluating a study by others! You will be surprised how often you will read in journal articles about sampling methods that were used – you should now be able to read these studies and identify the methods we've covered so far. You will also understand if the method that was chosen is appropriate and have insight into why it may have been chosen. Newspapers also are full of examples of sampling or polls – the sampling methods are often not reported but are often key to interpreting the results.
Good Job!

Slide 19



BIOS 110: The Principles of Statistical Inference

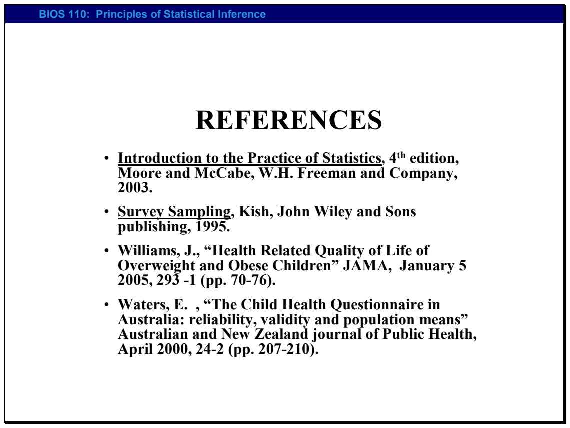
Objectives

- **Select a SRS**
  - **Multiplication by uniform random number**
  - **Grouping of digits in uniform random number table**
  - **Computer-generated**
- **Select a systematic random sample**
- **Advantages and disadvantages of each**

Unit 1 Lesson 3 19

Select a SRS
Multiplication by a uniform random number
Grouping of digits in URN table
Computer generated
Select a systematic random sample
Be able to identify the advantages and disadvantages of each as they apply to a public health study

Slide 20

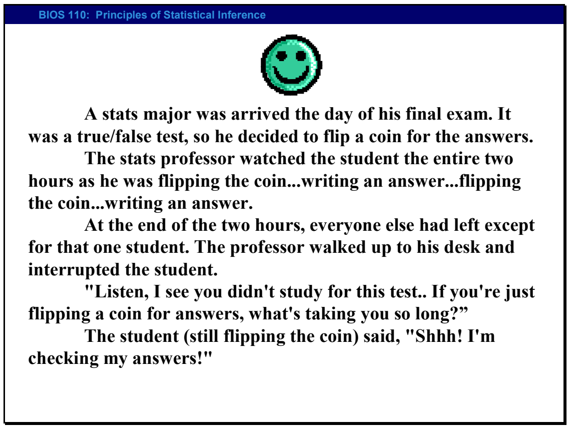


BIOS 110: Principles of Statistical Inference

REFERENCES

- Introduction to the Practice of Statistics, 4th edition, Moore and McCabe, W.H. Freeman and Company, 2003.
- 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).

Slide 21



BIOS 110: Principles of Statistical Inference

A stats major was arrived the day of his final exam. It was a true/false test, so he decided to flip a coin for the answers.

The stats professor watched the student the entire two hours as he was flipping the coin...writing an answer...flipping the coin...writing an answer.

At the end of the two hours, everyone else had left except for that one student. The professor walked up to his desk and interrupted the student.

"Listen, I see you didn't study for this test.. If you're just flipping a coin for answers, what's taking you so long?"

The student (still flipping the coin) said, "Shhh! I'm checking my answers!"