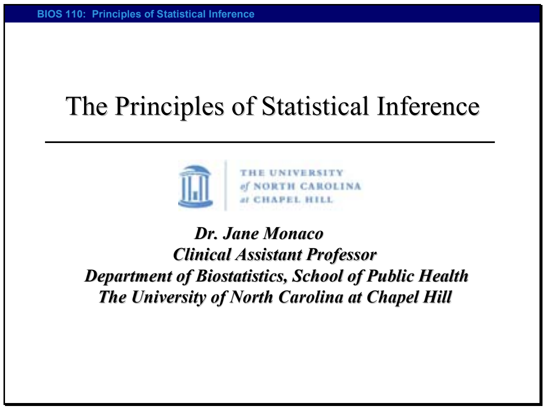
Slide 1



BIOS 110: Principles of Statistical Inference

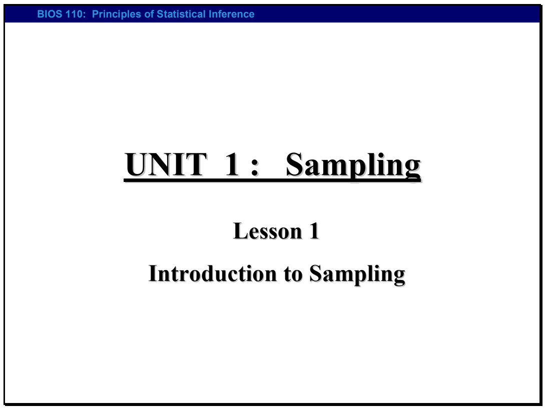
The Principles of Statistical Inference

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 to 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

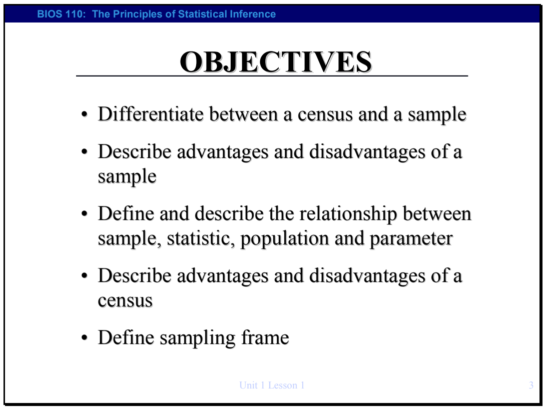
UNIT 1 : Sampling

Lesson 1

Introduction to Sampling

This unit addresses the topic of Sampling. Sampling is the way in which we select subjects from which to obtain data. The topic of Sampling is quite broad – this unit will contain only an overview of the topic of sampling, but hopefully this unit will inspire you to learn more about sampling! – Maybe even take a course in sampling.
We begin with Lesson 1 - an introduction to sampling.

Slide 3



BIOS 110: The Principles of Statistical Inference

OBJECTIVES

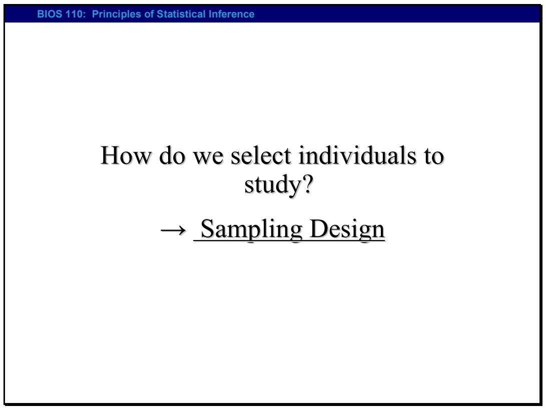
- Differentiate between a census and a sample
- Describe advantages and disadvantages of a sample
- Define and describe the relationship between sample, statistic, population and parameter
- Describe advantages and disadvantages of a census
- Define sampling frame

Unit 1 Lesson 1 3

When you complete Lesson 1, you should be able to:

*Differentiate between a census and a sample
*Describe advantages and disadvantages of a sample
*Define and describe the relationship between sample, statistic, population and parameter – you should be able to identify these in a study
*Describe the advantages and disadvantages of a census
*Define sampling frame

Slide 4



BIOS 110: Principles of Statistical Inference

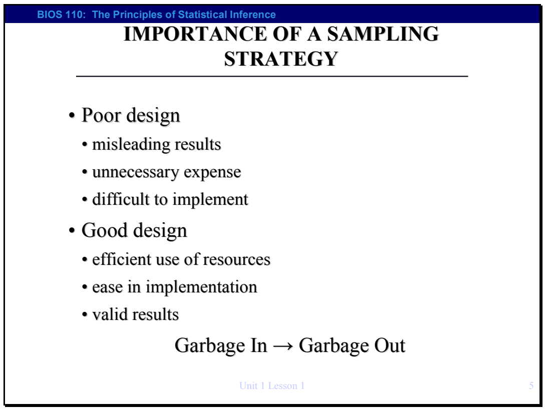
How do we select individuals to study?

→ Sampling Design

An important question in any kind of scientific investigation is, “How do we select individuals to study?”
The answer will likely have a considerable influence on the results of the study!
[By ‘individuals’, we could mean whatever unit we are studying ---people, lab rats, objects.]

Sometimes there is no ‘selection’ of subjects to participate in a particular study. For example, in some investigations, you may study the first subjects who meet the criteria for the study and who agree to participate – this is not sampling.
Other times, investigators select a subgroup of subjects to represent a larger group in which we are interested – this is sampling.
The way in which we select individuals for study is the topic of Sampling Design or just Sampling.

Slide 5



BIOS 110: The Principles of Statistical Inference

IMPORTANCE OF A SAMPLING STRATEGY

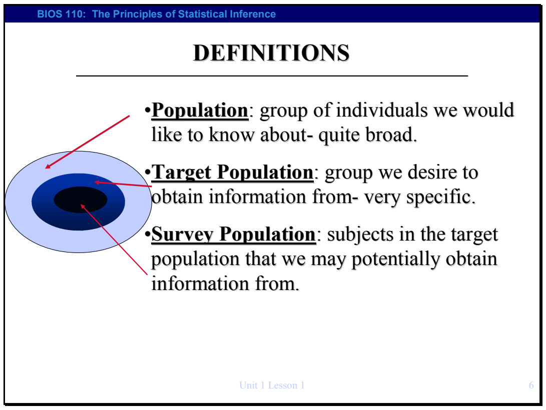
- Poor design
  - misleading results
  - unnecessary expense
  - difficult to implement
- Good design
  - efficient use of resources
  - ease in implementation
  - valid results

Garbage In → Garbage Out

Unit 1 Lesson 1 5

Why should we be concerned about how individuals are selected?
It’s because poor design can give us misleading results. A poor design can cause unnecessary expense. It can also be difficult to implement a poor design.
On the other hand, good design can give us valid results. It is an efficient use of resources and is easy (relatively easy?) to implement.
A common saying is “Garbage In -- Garbage Out”. A study in which the subjects are chosen with a poor design likely cannot overcome the problem. The results of a poorly designed study are suspect.
One consideration in conducting (or evaluating!) a study is to ask the questions: ‘how were the subjects chosen?’, and ‘are they representative?’

Slide 6



BIOS 110: The Principles of Statistical Inference

DEFINITIONS

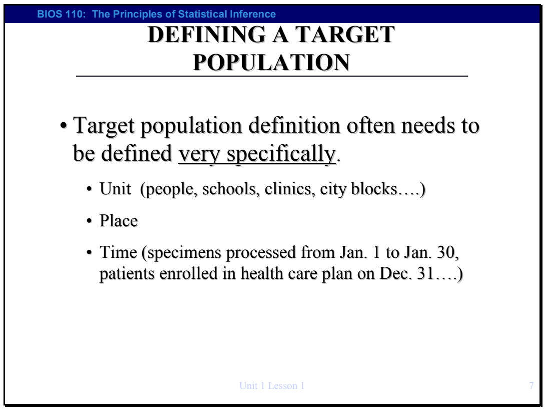
- **Population**: group of individuals we would like to know about- quite broad.
- **Target Population**: group we desire to obtain information from- very specific.
- **Survey Population**: subjects in the target population that we may potentially obtain information from.

Unit 1 Lesson 1 6

A few definitions to get started:
A population is a group of individuals we would like to know about. In an ideal world, we'd like to know about each member of the group.
The target population is the group we desire to obtain information from. We would like to have information about the entire population but usually this goal is quite unrealistic. So we limit the population to a target population- We restrict the population by, say geographic region or time or age…. I'll give an example in a minute.
The survey population contains the members of the target population we may potentially obtain information from. This is a further limitation (usually made for us!) of the target population.
I usually think of the target population as being a restriction of the population imposed by investigators, and the survey population as being further restricted by logistics.
(Again, here 'individuals' could mean humans, cities, vaccine vials, etc.)
The other common terms for individuals are subjects, units, elements, members of population.

Often, we'll just say "population" for simplicity, but keep in mind that investigators could mean any one of the above three when they say "population."

Slide 7



BIOS 110: The Principles of Statistical Inference

DEFINING A TARGET POPULATION

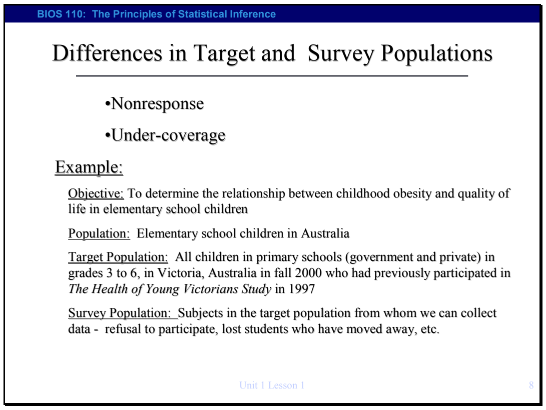
- Target population definition often needs to be defined very specifically.
  - Unit (people, schools, clinics, city blocks….)
  - Place
  - Time (specimens processed from Jan. 1 to Jan. 30, patients enrolled in health care plan on Dec. 31….)

Unit 1 Lesson 1 7

The target population is a subgroup of the population that we wish to study. In the beginning of a study or in evaluating a study in a journal article, we must be very clear about the target population. After this definition of the target population is established, we may be a little more lax in referring to the population, but a very specific definition of the target population in the beginning is needed .

For example, define the sampling unit (people, animals, days, petri dishes,….), the location of the population (geographic location –city, clinic, country,..) and an exact time or time frame.

Slide 8



BIOS 110: The Principles of Statistical Inference

Differences in Target and Survey Populations

- Nonresponse
- Under-coverage

Example:

Objective: To determine the relationship between childhood obesity and quality of life in elementary school children

Population: Elementary school children in Australia

Target Population: All children in primary schools (government and private) in grades 3 to 6, in Victoria, Australia in fall 2000 who had previously participated in *The Health of Young Victorians Study* in 1997

Survey Population: Subjects in the target population from whom we can collect data - refusal to participate, lost students who have moved away, etc.

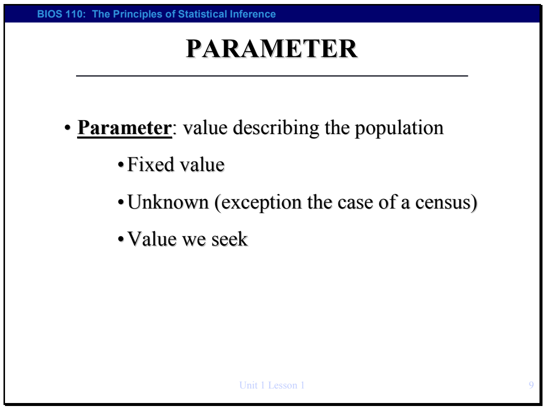
Unit 1 Lesson 1 8

The target population is a restriction of the population by time, place, age, … The target population and survey population differ due to factors such as nonresponse and under-coverage.

Nonresponse can be when the subject refuses, when the investigator has contact information but is unable to contact the individual (no answer on the phone), or when the subject is contacted but just doesn't reply (doesn't return the form or phone call),… Undercoverage is when the individual has no contact information (missing address or phone number) or is unable to be contacted (moved away) or omitted mistakenly by incomplete listing (listing of cases of ear infections in pediatrician's practice may omit cases seen by ER or urgent care doctors) ...

This example is based on an actual study, but simplified a bit. (We'll see the example again.) Consider a study in which the objective is to determine the relationship between childhood obesity and quality of life in elementary school children. Do children who are obese or overweight have a different quality of life compared to children of normal weight?
The population we would like to know about is elementary school children in Australia. This group is quite broad, so we narrow down the population of interest.
The target population may be all children in primary schools (government and private) in grades 3 to 6, in Victoria Australia in the fall of 2000 who had previously participated in The Health of Young Victorians Study in 1997. As specific as investigators can make the target population, the better.
The survey population is a further restriction. The survey population has the subjects that we can actually collect data on. The survey population will omit students that don't complete responses (refusal or non-response) and will omit students who have moved out of the area, etc.
Note: we'll talk later about under-sampling which is different from under coverage. So be aware now that these terms are different.

Slide 9



BIOS 110: The Principles of Statistical Inference

PARAMETER

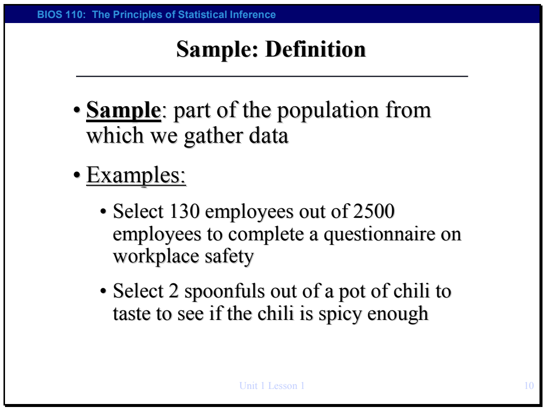
- **<u>Parameter</u>**: value describing the population
  - Fixed value
  - Unknown (exception the case of a census)
  - Value we seek

Unit 1 Lesson 1 9

A few more definitions, which are important….
A <u>parameter is the value which describes the population</u>. When I say “value”, I am talking about a number like an average, a proportion, a count, etc.
Some examples: the average number of colds per child per season…. The proportion of children using car seats …. Or the number of vaccinations given in the clinic during a month ….
A parameter is a value which is <u>fixed</u>. It doesn’t change. It is one value which describes the population (at a particular time). Usually a parameter is <u>unknown</u>. Unless you ask everyone in the population, which is census, you can’t know this value exactly. Since taking a census is likely impractical, you must estimate the parameter by taking a sample, or subgroup from which to obtain data. (we haven’t defined census yet, but you probably know – a census is collecting data from each member of the population.)

A parameter is usually “<u>what we want to know</u>.” Since it is difficult if not impossible to collect data from each member of the population, we must settle for a sample from which we can calculate an estimate for the parameter.

Slide 10



BIOS 110: The Principles of Statistical Inference

Sample: Definition

- **<u>Sample</u>**: part of the population from which we gather data
- <u>Examples:</u>
  - Select 130 employees out of 2500 employees to complete a questionnaire on workplace safety
  - Select 2 spoonfuls out of a pot of chili to taste to see if the chili is spicy enough

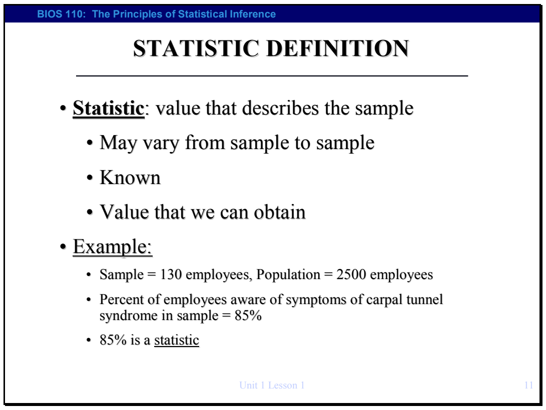
Unit 1 Lesson 1 10

This whole unit is about sampling, so we better get a definition.

A <u>sample</u> is the part of the population from which we gather data. It is a subgroup of the population.
Of course “sample” can be a noun or verb - we can talk about “the sample that we collect from the population” or could say “we now sample the population to determine…..”

You can sample employees in the workplace from all employees in order for the sample (the subgroup) to complete a detailed questionnaire.
You can sample chili out of the pot!

Slide 11



BIOS 110: The Principles of Statistical Inference

STATISTIC DEFINITION

- **Statistic**: value that describes the sample
  - May vary from sample to sample
  - Known
  - Value that we can obtain
- Example:
  - Sample = 130 employees, Population = 2500 employees
  - Percent of employees aware of symptoms of carpal tunnel syndrome in sample = 85%
  - 85% is a statistic

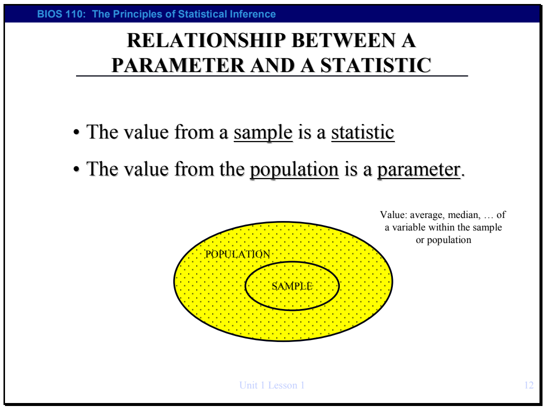
Unit 1 Lesson 1 11

We'd love to know that parameter… but we usually settle for a statistic. A statistic is a value that describes the sample.
[Again, by value we mean an average ..a proportion… for example]

A statistic usually varies from sample to sample, depending on the sample one happens to select (Compare this to a parameter which is fixed.) We would like for the statistic to be a good estimate of the parameter– otherwise it is not of much use. Much of whether the statistic is a good approximation of the parameter depends on the sampling scheme- how we select the sample.
A statistic is known. You calculate it from the sample. (Compare this to a parameter which is usually unknown)
A statistic is a value that you can obtain. After you've taken your sample, you simply calculate the value from the data from the sample (compared to a parameter which is difficult to obtain).

For example:
Select a sample 130 employees out of 2500 in a workplace. We are interested in the percent of employees aware of the symptoms of carpal tunnel syndrome. We calculate the percent of the sample (=130) aware of the symptoms. Let's say that 85% of the sample know the symptoms. Equivalently we could say that 110 out of 130 employees. 85% is a statistic.
(We would like to know the percent of the population (=2500) educated about the symptoms of carpal tunnel syndrome. If we could calculate this percent in the population, that value would be a parameter.)
Convince yourself that this statistic, 85%, can 1) vary from sample to sample 2) is known 3) is a value we can obtain.

Slide 12



BIOS 110: The Principles of Statistical Inference

RELATIONSHIP BETWEEN A PARAMETER AND A STATISTIC

- The value from a sample is a statistic
- The value from the population is a parameter.

Value: average, median, … of a variable within the sample or population

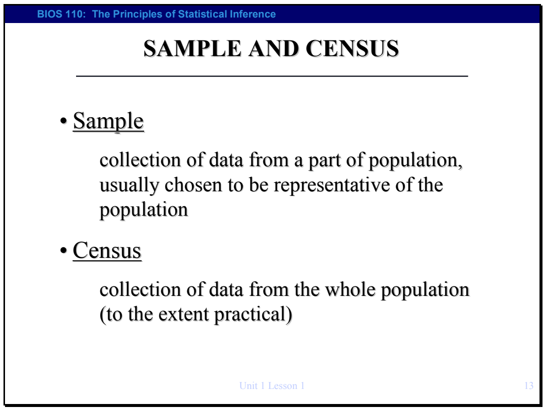
POPULATION

SAMPLE

Unit 1 Lesson 1 12

An easy way to remember the relationship is that the value from a sample is a statistic, the value from the population is a parameter.
If you calculate a value from a subset of the data (a sample) , it is a statistic. The value that could be calculated from the population is a parameter. (This is often impossible or impractical).
This is pretty simple…… it will come back, though, (come back to haunt you?) later in the course when we talk about confidence intervals and p-values. We will make hypotheses about a parameter and then use a statistic to draw conclusions about those hypotheses.

Slide 13



BIOS 110: The Principles of Statistical Inference

SAMPLE AND CENSUS

- Sample

  collection of data from a part of population, usually chosen to be representative of the population

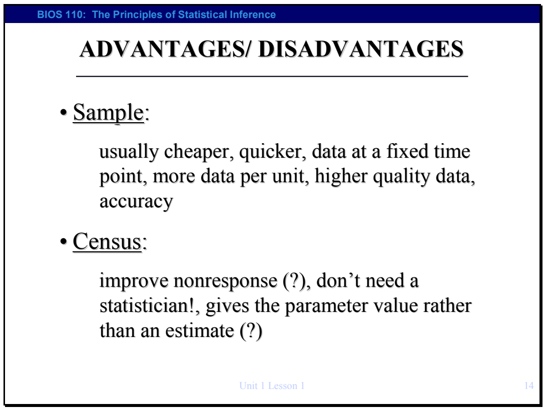
- Census

  collection of data from the whole population (to the extent practical)

Unit 1 Lesson 1 13

We've said it a couple of times, but let's give a definition… a sample is a collection of data from part of the population, usually chosen to be representative of the population. A sample is a subset of the population.
A census is a collection of data from the whole population (to the extent practical).
We are usually familiar with a census as being the US Census taken every 10 years, but the term census can apply to any population such as the population of children who are not immunized within a particular region in a developing country, or the 'population' of charges which are not reimbursed by Medicare in a particular health center.

Slide 14



BIOS 110: The Principles of Statistical Inference

ADVANTAGES/ DISADVANTAGES

- Sample:

  usually cheaper, quicker, data at a fixed time point, more data per unit, higher quality data, accuracy

- Census:

  improve nonresponse (?), don't need a statistician!, gives the parameter value rather than an estimate (?)

Unit 1 Lesson 1 14

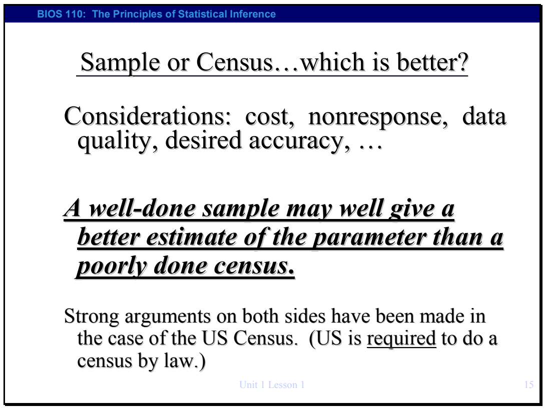
What are advantages and disadvantages of taking a sample compared to taking a census?

Well, taking a sample is usually cheaper and quicker than a census. That makes sense – you are collecting data on fewer "individuals" or units compared to a census.
With a sample, it is easier to define a time point which the data represent. Since a census takes longer to conduct, the population may be a 'moving target' during the data collection phase. The data values may be a moving target. With a sample, investigators can more easily capture data at a fixed time period rather than spreading data collection over a long period when the population or the values are changing.

With a sample, it is usually possible to collect more data per person and more accurate data from the members of the sample compared to similar funding for a census. In other words, there is a tradeoff between quality of data (more data per person and more in depth data) and the number of subjects in an investigation. For example, with a fixed budget, investigators may be able to collect more demographic or lab values and higher quality data (use one –on- one interviewers vs. questionnaires) compared to a census (more subjects, but less in-depth, quality data).
A data from sample may be more accurate than the data from a census. By accurate, I mean that compared to a census, the quality control is likely better in a sample, just due to cost. You could more easily track down, say, unusual values in a sample rather than in the whole population.

A census, on the other hand has advantages. Some may argue that the census will improve nonresponse (by definition there is no nonresponse in a census). We'll talk more about this. Certainly a census is easy to interpret - you don't need a statistician to calculate or interpret the results. "It is what it is". [I tend to think of this as a disadvantage since it puts my job at risk □.] Finally a census gives the true parameter value. We'll also talk about whether this is always true.

Slide 15



BIOS 110: The Principles of Statistical Inference

Sample or Census…which is better?

Considerations: cost, nonresponse, data quality, desired accuracy, …

A well-done sample may well give a better estimate of the parameter than a poorly done census.

Strong arguments on both sides have been made in the case of the US Census. (US is required to do a census by law.)

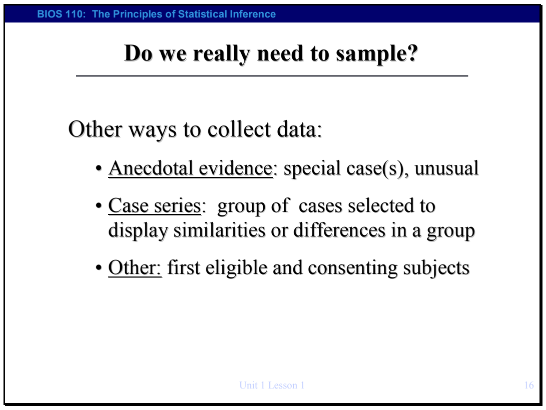
Unit 1 Lesson 1 15

Of course, an important question is "which is better?".
With a small population, a census can be practical, efficient and undoubtedly a good choice.
For larger populations, which is 'better' depends on the circumstances. Given a choice, there may be reasons to use a sample rather than a census!
Well, certainly a sample is likely to be less expensive.
Let's go back to "Census has less nonresponse". In a perfect census, the census by definition would have NO nonresponse. In the real world, when the subjects are people, a census will have

nonresponse. There may be a tendency to have more participation in a census… psychologically subjects may think "Everyone is doing it, I better do it too! " "I better not refuse… it will make me stand out in the crowd." OR NOT there may be more participation in a sample "Hmm. I was selected for this sample, I want to participate because my response is important." Arguments can be made both ways. Much depends on the information collected by a particular study. Also, let's return to the statement that an advantage of a census is that it gives the true parameter rather than a statistic which is just an estimate. This is certainly true in a perfect census. Unfortunately a census rarely is perfect. *A well-done sample may well give a better estimate of a parameter than a poorly done census.* Even a well conducted census just may not have the resources to give accurate values. This is in part due better ability to address data quality issues and nonresponse and assuring a representative group by using a smaller sample.

Much has been written about using a census compared with a sample, particularly in the case of the US Census. I won't go into all the arguments. Those same arguments can be made in general, not just for the US Census. The US is REQUIRED to do a census every ten years… but many argue that a sample would not only be cheaper but also give more accurate results.   Just be aware that a census is not always better.

Slide 16



BIOS 110: The Principles of Statistical Inference

Do we really need to sample?

Other ways to collect data:

- Anecdotal evidence: special case(s), unusual
- Case series: group of cases selected to display similarities or differences in a group
- Other: first eligible and consenting subjects

Unit 1 Lesson 1 16

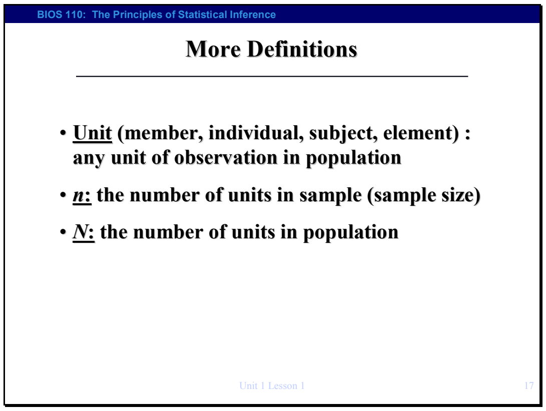
Selecting a sample is not the only way to collect data.
Anecdotal evidence is often reported… this is a special case or special cases. These cases often are noted because they are unusual. In sampling, we want to select subjects that are representative of the entire population. I don't want to minimize the importance of anecdotal evidence - it is just very different way (an important way!) of collecting or reporting results. Often, anecdotal evidence is mentioned in a derogatory manner. "That is just anecdotal evidence." Well, it has merit as long as we recognize what it is meant for. For example, the AIDs epidemic was first identified through anecdotal evidence when physicians noted unusual cases of an autoimmune disease primarily in homosexual men.

A case series is a group of cases in which similarities or differences are noted. Again, a very different way to collect or report data.

There are also other ways of selecting subjects on whom to collect data. You could take all eligible subjects such as in a clinical trial. This case is when the first eligible and willing subjects enter a study.

In the next lesson, we will say that these examples are not "probability samples". For now, just note that these are not what we usually mean by "collecting a sample".

Slide 17



BIOS 110: The Principles of Statistical Inference

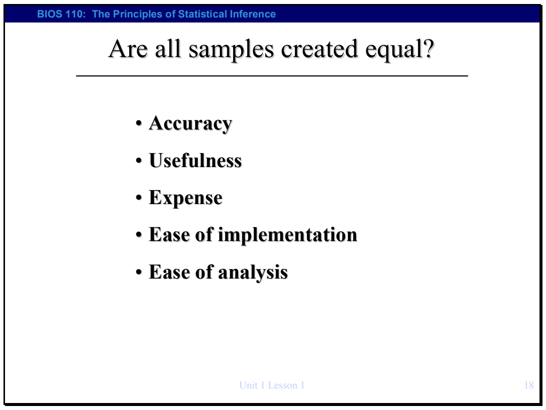
More Definitions

- **Unit (member, individual, subject, element) : any unit of observation in population**
- ***n*: the number of units in sample (sample size)**
- ***N*: the number of units in population**

Unit 1 Lesson 1 17

Some more definitions….
We've said that a unit can be a member, individual, subject element or any unit of observation in population.
'*n*' is the number of units in sample, that is, sample size.
'*N*' is the number of units in population.

Slide 18



BIOS 110: The Principles of Statistical Inference

Are all samples created equal?

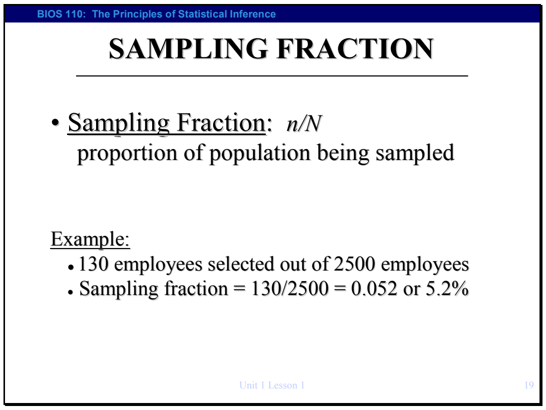
- **Accuracy**
- **Usefulness**
- **Expense**
- **Ease of implementation**
- **Ease of analysis**

Unit 1 Lesson 1 18

Are all samples created equal?
Of course not. Some are more accurate than others…. some are more useful than others…. More or less expensive to conduct… easier or harder to conduct…..easier or harder to analyze.

There are trade-off as with anything. Investigators must balance all of these factors to conduct the best study within any given constraints.

Slide 19



BIOS 110: The Principles of Statistical Inference

SAMPLING FRACTION

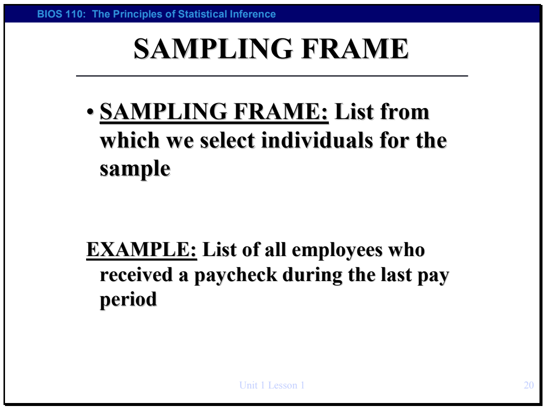
- Sampling Fraction: *n/N*
  proportion of population being sampled

Example:
- 130 employees selected out of 2500 employees
- Sampling fraction = 130/2500 = 0.052 or 5.2%

Unit 1 Lesson 1 19

Sampling fraction is n divided by N, and is the proportion of population being sampled.
In the example, when we select a sample of 130 workers out of 2500, the sampling fraction is 130/2500 = 0.052 or 5.2%

Slide 20



BIOS 110: The Principles of Statistical Inference

SAMPLING FRAME

- **SAMPLING FRAME: List from which we select individuals for the sample**

EXAMPLE: List of all employees who received a paycheck during the last pay period

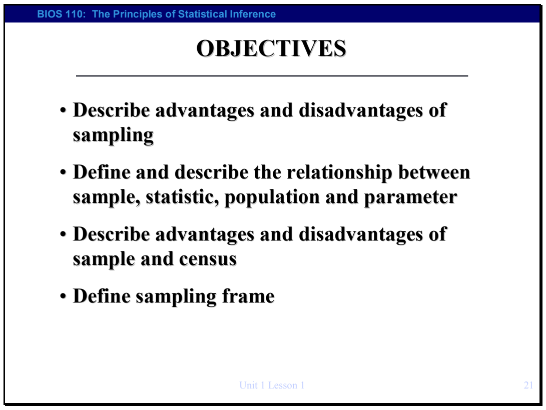
Unit 1 Lesson 1 20

Sampling frame is a list from which we select individuals for the sample.

Ideally, we would like a complete list of all the members of the population. From this list, we would select a random sample. Sometimes, a sampling frame (or list) is not available. We will discuss this issue in other lessons.

For example, in the example when we select a sample of 130 workers from 2500, we may get a sampling frame by contacting human resources and get a list of all employees who received a paycheck in the last pay period. This list would be our sampling frame.

Slide 21



BIOS 110: The Principles of Statistical Inference

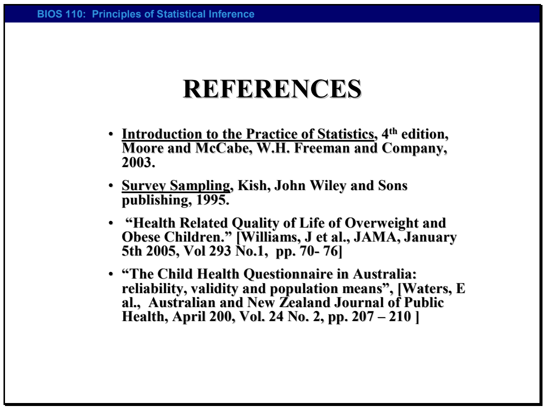
OBJECTIVES

- Describe advantages and disadvantages of sampling
- Define and describe the relationship between sample, statistic, population and parameter
- Describe advantages and disadvantages of sample and census
- Define sampling frame

Unit 1 Lesson 1 21

The topic of sampling is extensive.
In Unit 1 SAMPLING - Lesson 1 our objectives were:
Describe advantages and disadvantages of sampling
Define and describe the relationship between sample, statistic, population and parameter
Describe advantages and disadvantages of sample and census
Define sampling frame

Slide 22

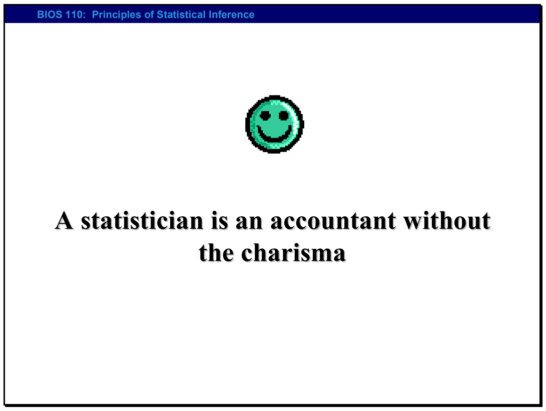


BIOS 110: Principles of Statistical Inference

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.
- “Health Related Quality of Life of Overweight and Obese Children.” [Williams, J et al., JAMA, January 5th 2005, Vol 293 No.1, pp. 70- 76]
- “The Child Health Questionnaire in Australia: reliability, validity and population means”, [Waters, E al., Australian and New Zealand Journal of Public Health, April 200, Vol. 24 No. 2, pp. 207 – 210]

Slide 23



BIOS 110: Principles of Statistical Inference

A statistician is an accountant without the charisma