


Slide
1

BIOS 115: 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 115: Principles of Statistical Inference

UNIT 1: SAMPLING

Lesson 2
Probability samples
Special Case: Simple Random Sample

We continue with the unit on Sampling– “How do we select individuals to collect data from?” This lesson, Lesson 2, will address probability samples as compared to other types of samples. Then we’ll talk about the most important example of a probability sample, the Simple Random Sample.

Slide
3

BIOS 115: The Principles of Statistical Inference

OBJECTIVES:

- Define and give examples of anecdotal evidence, voluntary response samples, other samples which are not probability samples
- Define probability sample
- State importance of randomness
- Identify issues with sampling and data collection
- Define Simple Random Sample (SRS) and state advantages and disadvantages
- Define and understand “EPSEM”

Unit 1 Lesson 2

Define and give examples which are not probability samples, such as anecdotal evidence, voluntary response samples, other samples

Define probability sample

State importance of randomness

Identify issues with sampling and data collection

Define Simple Random Sample (SRS) and state advantages and disadvantages

Define and understand EPSEM

Slide
4

BIOS 118: The Principles of Statistical Inference

Probability Sample Definition

- **Probability sample:** every element in the population has a known, nonzero probability of selection
- Probability of selection is determined by chance (at random)

Randomness \neq haphazard

Unit 1 Lesson 2

In a probability sample, every element in the population has a known, nonzero probability of selection.

This probability is determined by chance (randomly)

There are some words which we use in every day language which have very specific definitions in statistics.

“At random” is one of these terms. In everyday language, we may use the term “randomly” to mean “haphazardly”. In statistics, “at random” does **not** mean “haphazardly”.

Words like “significant”, “correlated”, “independent” “at random” are all common words that have a very specific meaning in statistics and should be treated with care in this class.

We will come back to this definition with examples of probability samples, but sometimes the best way to show what something is, is to show what it is not. Let's begin with some examples which are **not** probability samples.

Slide
5

BIOS 118: The Principles of Statistical Inference

NOT PROBABILITY SAMPLES

- **Anecdotal evidence:** Case(s) selected because interesting or unusual
- **Selection by data collector:** Data collector selects units by some method other than random selection (convenience, haphazard, ...)
- **Selection by expert:** Knowledgeable investigator selects units deemed typical

Unit 1 Lesson 2

Before going in depth with “what is a probability sample?” Let's give some examples of some samples which are not probability samples, keeping in mind the definition that we just gave of a probability sample.

Anecdotal evidence is not a probability sample – it is case (or cases) chosen not an random but specifically because it is unusual or to illustrate a point. Strictly speaking, anecdotal evidence is not even really a sample.

Sometimes a data collector selects units. This sample may be selected because the subjects are convenient, or in a haphazard manner such as “xray films in the front of the cabinet”, or every couple of subjects who seem “agreeable” and therefore likely to participate or subjects who are easy to contact.

Even when an expert selects subjects that he or she think are representative of the whole population, this is not a probability sample. This method of selection likely will not give as

good an estimate of the population value compared to a probability sample.

Favoritism may occur in these last two schemes. All of these schemes may result in a systematic error we will call bias.

Slide
6

BIOS 110: The Principles of Statistical Inference

MORE TYPES OF SAMPLES

- **Voluntary Response Sample:** Subjects respond based on general appeal

Example:

Ann Landers poll:
‘If you had it to do over again, would you have children?’
30% of 10,000 respondents said ‘yes’

Newsday poll
91% of 1373 randomly selected individuals said ‘yes’
(Large sample size does not eliminate the need for a well-designed study)

Unit 1 Lesson 2

Another example of a sample which is not a probability sample is a voluntary response sample. Examples are common “call-in” polls. News organizations often will ask viewers or readers to call in or write in responses to a general question. This sampling method can be quite biased because (among other things) those with strong opinion are more likely to respond.

An often cited example of the problems with a voluntary response sample is that only those who feel strongly about the issue are likely to participate. Ann Landers, an advice columnist once asked in her advice column, “If you had it to do over again, would you have children?” Readers wrote in and only 30% said “yes”. Later, a much smaller Newsday poll, using random sampling was conducted, and 91% of participants said they would have children if they “had it to do over again.” The randomly selected poll more accurately represented that population. Likely only those with strong opinions would write in to a voluntary response poll. Clearly, a voluntary response poll may not be representative of the population we are interested in. Note that sample size is less important in this example than the method of data collection!

Slide
7

BIOS 110: The Principles of Statistical Inference

Examples : Not probability samples

- Advantages: convenient, cheap
- Disadvantages: may be biased, may not represent the population

Previous examples (not probability samples) do not contain an element of randomness

Unit 1 Lesson 2

Advantages of a these non-probability sample are that the sampling may be convenient and/or cheap, but disadvantages include that the results are likely biased (or least serve a different purpose).

Rather than just discounting these non-probability samples as useless, I prefer to just be aware of what they represent... ..anecdotal evidence is a special case and should not be viewed as a typical case, call-in polls represent those with strong opinions not the population in general, etc.

Slide
8

BIOS 119: The Principles of Statistical Inference

PROBABILITY SAMPLE

Probability Sample: A sample in which every element in the population has a known, nonzero probability of selection. That probability is determined by chance.

(Sometimes called 'scientific' or 'statistical' sampling)

Note: You cannot tell by looking at the sample data whether the sample is a probability sample. You must know how sample was selected.

Unit 1 Lesson 2 8

Let's go back – we've seen some examples which are NOT probability samples... what is a probability sample? I'll repeat the definition.... it is a sample that is selected in such a way that every element in the population has a known, nonzero probability of selection and that probability is determined by chance. Sometimes it is called 'scientific' or 'statistical' sampling. You cannot tell by looking at the sample data itself whether a sample is a probability sample. It can look random...but you must know how sample was selected to determine if it is a probability sample.

Slide
9

BIOS 119: The Principles of Statistical Inference

PROBABILITY SAMPLING

- **Advantages**
 - Eliminate sampling bias
 - Control sampling variability
 - Can never be eliminated (Unless you take a census)
- **Disadvantages**
 - Can be more expensive than, say, anecdotal evidence
 - Not a census

Unit 1 Lesson 2 9

The advantages of probability sampling are that it can help eliminate sampling bias. It can control sampling variability although this variability can never be eliminated unless you take a census.

Disadvantages of a probability sample are that it can be more expensive to conduct. Also as with any sample, it is not a census, so you get an estimate of the parameter which, by chance, may or may not be close the parameter.

Slide
10

BIOS 119: The Principles of Statistical Inference

VARIABILITY

- Can be reduced by taking a larger sample
- With a probability sample, we can measure the variability and calculate sufficient sample size to produce acceptable level of variability

Unit 1 Lesson 2 10

Sampling variability, which is how far away a statistic is away from the parameter, can be reduced by taking a larger sample.

With a probability sample, we can measure the variability and calculate sufficient sample size to produce acceptable level of variability. Compare this to a data collector picking data – there is no way to find out how many units are needed to get a close estimate of the parameter.

Slide
11

BIOS 118: The Principles of Statistical Inference

Types of Error

- **Sampling error** – error due to taking a sample instead of census
- **Nonsampling error**– other types of error (logistical, poor planning, imperfect nature of sampling)

Unit 1 Lesson 2 11

Sampling error is an error due to taking a sample. After all, depending on the sample you take, you will likely not calculate the parameter exactly... (you are likely to get a value which may be higher or lower) just due to the fact that you selected only part of the population.

Nonsampling errors are other types of errors which can be caused by logistical issues, poor planning or imperfect nature of sampling.

Slide
12

BIOS 118: The Principles of Statistical Inference

Nonsampling Errors

Probability sampling cannot overcome some problems such as nonsampling error

- Errors in measurement
- Processing errors
- Wording of question
- Nonresponse
- Refusal
- Undercoverage

Unit 1 Lesson 2 12

Probability sampling cannot overcome some problems such as nonsampling error... for example errors in measurement, processing errors, wording of question, nonresponse, refusal and undercoverage...

Let's investigate each one of these issues in turn.

Slide
13

BIOS 118: The Principles of Statistical Inference

TYPES OF ERROR

- **ERRORS IN MEASUREMENT:**
 - Insufficient training of interviewers or staff
 - Mistakes in measurements
- **PROCESSING ERROR:**
 - Physical problem with data – “hanging chad”
 - Computer problems such as incorrect programming

Unit 1 Lesson 2 13

Suppose that interviewers were trained insufficiently in obtaining results such as not probing patients' responses appropriately or consistently. If techniques of interviewers vary between interviewers, bias may result. Similarly, if medical personnel are trained differently or incorrectly in taking, say, skinfold measurements or blood pressure measurements, bias may result. Different labs may give slightly different values for cholesterol values or white blood cell count – some more accurate than others.

Errors in measurement can be just mistakes in collecting the data... using pounds instead of kilograms...equipment which is not calibrated correctly....

Processing error describes error which occurs after the data are collected. A classic example is the “hanging chad” issue in the 2000 presidential election in Florida. Computer glitches, faulty programming, or even physically losing or inadvertently changing the data are all possible problems.

Slide
14

BIOS 110: The Principles of Statistical Inference

WORDING OF QUESTION

- Biased results from wording of a question
- Some terms elicit strong responses
 - 'Abortion' vs. 'termination of pregnancy'
 - 'Welfare' vs. 'aid to the poor'
- Poor wording can lead to confusion

Unit 1, Lesson 2 14

Biased results may occur when sampling questions contain wording which is prejudicial. For example, wording such as "abortion" vs. "termination of pregnancy" or "welfare" vs. "aid to the poor" may evoke strong reactions. Most folks are willing to spend resources on aid to the poor, but welfare has a negative connotation. In evaluating the results of a poll, readers would benefit from knowing the exact wording of a question. Often slight wording differences may produce very different results. Didn't Bill Clinton say something like "it depends on the meaning of the word 'is' is"?. The wording could be prejudicial intentionally or unintentionally.

So seemingly clear statement may have different meanings or connotations

Slide
15

BIOS 110: The Principles of Statistical Inference

NONRESPONSE, REFUSAL, UNDERCOVERAGE

- **NONRESPONSE**- individuals can be contacted but do not respond to pleas for participation
 - **REFUSAL**- individuals are contacted but refuse to participate
- **UNDERCOVERAGE**- individuals are not on sampling frame, thus can not be contacted... bias can occur *when individuals not on the sampling frame differ from those on the sampling frame*

Unit 1, Lesson 2 15

More types of error can include...

NONRESPONSE occurs when individuals can be contacted but do not respond to pleas for participation. Surveys which do the best job, often track down the nonresponders aggressively, attempting contact them multiple times, almost to the point of "badgering" the potential subject. It is very important, if possible to find out how nonresponders are different from responders.

REFUSAL is when individuals are contacted but refuse to participate. Refusal can be thought of as nonresponse... but you actually do have some information about these subjects.... You know they don't want to participate! For people who refuse to participate, it is often helpful to collect **any** sort of information, either from them or from records, such as demographic information to assess whether refusers are different from participants.

UNDERCOVERAGE usually occurs when individuals are not on sampling frame, thus can not be contacted. Bias occurs when individuals not on the sampling frame differ from those on the sampling frame in a systematic way (for example geographically different, age or ethnicity, etc.).

We have talked about a lot of different types of

error. Always it's easier to note weaknesses in other investigators' studies.

While it's important to investigate possible implication of short comings, it is important to realize that time and cost are involved in eliminating errors. In an ideal world, we would love to have perfect frame and track down all nonresponders for example. Realistically this may be impractical. I prefer to consider not only the shortcomings of samples but perhaps more importantly the implications of the problems. In other words, if bias is likely in the study, which direction is the study likely biased?

Slide
16

BIOS 110: The Principles of Statistical Inference

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

An SRS is a probability sample, but not the only kind of probability sample

Unit 1 Lesson 2 16

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.]

It is a good idea to pause the lecture and look at that one more time... I'll wait.

An SRS is the simplest and most important method of sampling. Many other methods rely on SRS at some point.

Slide
17

BIOS 110: The Principles of Statistical Inference

SIMPLE RANDOM SAMPLE (SRS)

EXAMPLE:

- Population: 1000 patients ($=N$) in the clinic
- Sample : 100 patients ($=n$)
- 20% of pop'n diabetic, 80% not diabetic.

(1) Select at random 100 patients. Probability of selection of any one subject is $100/1000 = 0.1$.

(2) Probability of selection of any one set of 100 patients is the same as the probability of selection for any other set of 100.

Unit 1 Lesson 2 17

Let's look at an example.

Suppose you wish to conduct a study concerning patients with a clinic. There are 1000 patients in the clinic (this is the population of interest), and you have determined a sample size of 100 patients is sufficient to obtain the needed level of precision (more about sample size later in the course). $N=1000$, $n=100$. About 20% of patients are diabetic, 80% are not diabetic. So about 200 patients are diabetic and 800 are not diabetic.

We select at random 100 patients. Probability of selection of any one individual is $100/1000 = 0.1$. We referred to this in the last lesson as the sampling fraction.

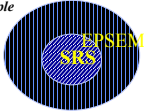
Also probability of any one set of 100 patient is the same as the probability of selection for any other set of 100. This particular statement is exactly what it means to be a SRS.

Slide
18

BIOS 110: The Principles of Statistical Inference

SRS Compared to EPSEM Sample

- **EPSEM** (Equal Probability of Selection of Each Member)
 - Probability sample in which the probability of selection is the same for each *individual*
- **SRS**-
 - Probability sample in which the probability of selection is the same for each *sample*



Unit 1 Lesson 2

For example, the probability of selection of any one set of 50 diabetics and 50 non-diabetics is the same as the probability of selecting any one set of 100 diabetics which is the same as selecting any particular group of 20 diabetics and 80 nondiabetics.

We will talk more about how to **select** a SRS in another lesson.

Let's look a little more closely at this example. A characteristic of the previous example is referred to as EPSEM.

EPSEM means "Equal Probability of Selection of Each Member" A probability sample is EPSEM if the probability of selection is the same for each individual. In the previous example, the probability of selection for any one individual is 0.1 – the same for each individual. It is characteristic (1) on the previous slide. It may be a good time to pause the lecture and look back at the last slide.

SRS- recall the definition of SRS. The probability of selection of any particular **SAMPLE** is the same. This is characteristic (2) on the previous slide. Go ahead and look back... I'll wait.

So, the previous example was an SRS and EPSEM.

All SRS samples are also EPSEM. But, you can have an EPSEM sample which is not an SRS.

If this is not crystal clear at this point... this concept is not easy... I'll try to clear it up with more explanation and examples...

Slide
19

BIOS 110: The Principles of Statistical Inference

EPSEM COMPARED TO SRS (Our Example)

- **EPSEM**- (Equal Probability of Selection of Each Member)
 - Example: the probability of selection of any unit is the same as the sampling fraction: $n/N=100/1000=0.1$
- **SRS**- Probability of selection of any sample of 100 is the same as the probability of selection of any other sample of 100

Our example is EPSEM and an SRS

Unit 1 Lesson 2

Returning to the example...

In an EPSEM sample, the probability of selection is the same for any individual. "Equal Probability of Selection of Each Member" – addresses the probability of selection for '**members**'. In the diabetic/nondiabetic example, the probability of selection of any unit is sampling fraction: $n/N= 100/1000=0.1$.

In SRS sample, the probability of selection of any **sample** of 100 is the same as the probability of selection of any other sample of 100. There are lots of different groups of 100 out of 1000. Selection of each of these groups is equally likely.

BIOS 119: The Principles of Statistical Inference

Why is this example an SRS?

- **1000 subjects: IDs 1 to 1000**
 - IDs 1 to 200 are diabetic • IDs 201 to 1000 are not diabetic
- **Select 100 IDs from 1000**
 - (1) select ID values { 1,2,...,100}
 - (2) select ID values {2,3,...,101}
 - (3) select ID values {901,902,...,1000}
 - (4) select ID values {1,2,...,20, 201 ...280}
 - (5) select ID values {51,...,100, 251,..., 300}
 - EACH SET IS EQUALLY LIKELY- THIS EXAMPLE IS A SIMPLE RANDOM SAMPLE -- SRS.
- **Sample could contain all diabetics (1) and (2), could contain all non diabetics (3), 20% diabetic and 80% nondiabetic (4), or ½ (5) anywhere in between...**

Unit 1 Lesson 2

To drive the point home about why this example is a SRS, let's look a little closer.

Suppose that the subjects have IDs 1 to 1000. We know that subjects with IDs 1-200 are diabetic. Subjects with IDs 201 to 1000 are not diabetic. (20% diabetic, 80% not diabetic). -- we've just ordered them that way.

We select a sample of 100 subjects out of 1000. How likely is it that you would select IDs {1,2,3,...,99,100}? (it is not very likely... I could calculate the probability -- you would not like the calculations! - the value of this probability is not important.) Well, selecting {1,...,100} is equally likely as selecting IDs {2,3,...,99,100, 101}. Here, order doesn't matter... they are all in your study... it doesn't matter the order that they enter the study.... How likely is it that you would select for your sample of 100 the set {901,902,...,999,1000}? How likely is it that you would select IDs {1,...,20,200,..., 280}? Each of these sets is equally likely to be selected. The probability of selecting any set of 100 IDs is the same as the probability of selecting any other set of 100. This is exactly the definition of SRS!!

There are lots of different groups of 100 out of 1000. Selection of each of these groups is equally likely.

Notice that your sample could contain

*all diabetics (1) or (2)

*all not diabetic (3)

*20% diabetic and 80% not diabetic (4)

½ diabetic and ½ not diabetic (5)

Or anywhere in between

I suggest that you stop the presentation and reread the last few slides. Assure yourself that you know the definition of EPSEM, SRS and the difference.

Slide
21

BIOS 118: The Principles of Statistical Inference

EPSEM Sample Which Is Not an SRS

- Within 200 diabetics, select 20
- Within 800 non diabetics, select 80

- **EPSEM? Yes. Probability of a diabetic being selected is $20/200 = 0.1$ = Probability of a nondiabetic being selected is $80/800 = 0.1$**
- **SRS? →?**

Unit 1, Lesson 2 21

- You may be asking yourself at this point..."OK, an SRS is always EPSEM, I understand that. Can you show me an example of an EPSEM sample which is not an SRS?"

You may be also asking yourself the question, "how long is this lecture going to last? do I have time to go get a drink?" or "is this going to be on the test?"

I think I'll address the first question!

Let's suppose we want to be assured that we get enough diabetics and enough non diabetics to draw some conclusions. We decide to take sample of 20 diabetics from all 200 diabetics, and we select 80 nondiabetics from all 800 nondiabetics . We force there to be 20 diabetics and 80 nondiabetics by sampling in that way – just list the 200 diabetics and select 20... We list the 800 non diabetics and take a sample of 80.

Is it EPSEM? Let's check the definition...Yes, this sample is EPSEM, because the probability of diabetic being selected is $20/200 = 0.1$. The probability of a nondiabetic being selected is $80/800 = .1$. The probability selection is same for all members- the sample is EPSEM.

Slide
22

BIOS 118: The Principles of Statistical Inference

EPSEM Sample Which Is Not an SRS (cont.)

- **SRS?**
 - Probability of selecting all diabetics in sample = 0
 - Probability of selecting all nondiabetics = 0
 - Probability of selecting a particular set of 20 diabetics and a set of 80 nondiabetics not equal zero
- **=> NO. The example is not a Simple Random Sample**

Unit 1, Lesson 2 22

Is this example an SRS? No, because in this scenario, the probability of selecting all diabetics in sample (of 100) = 0 (it can't happen), and the probability of selecting all nondiabetics = 0. We have force there to be 20 diabetics and 80 non diabetics. But the probability of selecting a particular set of 20 diabetics and a set of 80 nondiabetics is not zero. So, this example is not SRS.

We could calculate this probability, but for us it is enough to know the probability is not zero. This is an example of an EPSEM sample which is not an SRS.

Slide
23

BIOS 119: The Principles of Statistical Inference

Advantages and Disadvantages of SRS

- Advantages:
 - Conceptually simple
 - Straightforward to analyze
- Disadvantages:
 - Need a sampling frame
 - Insufficient sample size in subgroups for subgroup analysis by chance
 - Need a random number generator (or table)

Unit 1 Lesson 2 23

The advantages of SRS include that it is conceptually simple and straightforward to analyze. SRS is often a building block for more complicated sampling designs. The disadvantages include that the sample may, by chance, not contain certain subgroups or sufficient number in the subgroup for separate analysis, and that we need a list (sampling frame) of population of interest. Other sampling schemes we explore in future lessons may be preferable to SRS.

Slide
24

BIOS 119: The Principles of Statistical Inference

OBJECTIVES:

- Define and give examples of anecdotal evidence, voluntary response samples, other samples which are not probability samples
- Define probability sample- state importance of randomness
- Issues with sampling and data collection
- Define SRS and state advantages and disadvantages
- Define and understand EPSEM

Unit 1 Lesson 2 24

The objectives of SAMPLING - Lesson 2 are:
Define and give examples of anecdotal evidence, voluntary response samples, other samples which are not probability samples.
Define probability sample. State importance of randomness
Issues with sampling and data collection
Define and understand SRS and state advantages and disadvantages
Define and understand property of EPSEM

Slide
25

BIOS 119: 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.

Slide
26

