1pops_samples.pdf	Michael Hallstone, Ph.D. hallston@hawaii.edu

Lecture 1: Why do we use statistics, populations, samples, variables,

why do we use statistics?

- interested in understanding the social world
- we want to study a portion of it and say something about it
    - ex: drug users, homeless, voters, UH students

Lies, Damn Lies, and Statistics

If you have not heard the Mark Twain joke, you've at least heard the trope "statistics can prove anything" and all the nonsense against science, experts, etc. There is this whole – very silly and stupid- notion that there is no objective reality and people should not trust science or scientists or main stream journalism. [Public Service Announcement: Do not get your news from cable TV and whatever you do NOT get your news from Facebook or other social media! You will develop opinions not consistent with known facts. Yes people facts still exist!]

Science and scientists are biased!

Bull excrement! Big fat stinking piles of bull excrement. A more accurate notion is that stupid scientists or politically motivated scientists [often, but not exclusively, employed by privately funded "think tanks"] can create bad or biased science.

Proviso: To be fair, especially in the mid 1900's, there was so called scientific "proof" in the social sciences for prejudices held by the dominant classes like girls are not as smart as

boys in math, descendants of plantation workers in Hawai'i are not as smart as descendants of Europeans [except the Portuguese], dark skinned people [especially who were descendants of Africa] are not as smart as Europeans, etc. But we've learned a lot about how the methods we use to collect data can create "biased science" in recent decades and these sorts of findings are less and less common. In fact the scientific peer review process is how we discover these errors. [Which speaks against the value of non-peer reviewed think tank papers!]

Statistics is useless in the "real world!"

Really? Hmmm. Let's see if we can think of a recent example that shows what an utterly vapid [think "airhead'], stupid, ignorant statement that is.

Even though this is a social science oriented statistics text, here is a medical science example you may relate to. Pretend we want to know something about a group of people called humans who could catch a respiratory illness that spreads through the air as people exhale and infects a lot of people. Pretend it is a virus so that means antibiotic drugs do not kill it.

Weirdly most people who get it won't know it but they will pass it on to others. That means it will spread far and wide because most people who are infected will not know it and thus unintentionally pass it on to other humans.

For those who get symptoms it's not a fun illness as it clogs up the lungs but it does not seem to kill most healthy young folks. The weird part is that it does tend to kill a pretty fair share of grandparents and gets way more lethal with age. It is especially dangerous for people of any age who are overweight, have heart disease, diabetes, cancer, and other diseases that are wide spread in many societies. When it kills it is a horrible death in that it's like drowning and it's so dangerous people die alone because their loved one's can't visit them in the hospital.

Let's review. There is a virus that we catch by breathing and it spreads like wildfire because most people who spread it never knew they were contagious and just went about their normal lives.

Now pretend there are no treatments or vaccines for this illness of the lungs. Such an illness would be very dangerous and could shut down whole cities, nations, and indeed the world economy.

Ring any bells?

We can't give an experimental drug to everyone with COVID people!

I had a unique viewpoint of the development of mRNA vaccines as one of my oldest and dearest friends happened to work in vaccine research for Pfizer. His job was to oversee the collection of data for experimental vaccines.

So as you have figured out you don't want to kill people on accident with an experimental vaccine. You start with a small group of people to make sure it is not dangerous and does not have any real nasty side effects. Once you figure out the vaccine is safe then you give it to a larger group to see if it "works." If the vaccine works for the larger group we infer that it will work for all humans. ***The way you figure this out uses inferential statistics.***

Read the paragraph above again. It describes a sample and a population although I did not use those words. You give the vaccine a group of people called a sample. If it works for the sample we infer it works for the population of all humans.

And this, my friends, is an example of why impulsive, shallow criticisms of science and statistics are stupid.

Populations and Samples

So once again, in the example above Pfizer and Modern gave their experimental vaccines to a sample or small group of humans. But they ultimately wanted to know if the vaccine would work on a much larger group of humans called the population. This is just a unique example where the population or the group the scientists were interested in studying was every single human on the planet. Most times populations are composed of much smaller groups of people. [In fact, as discussed below populations can be made up of things besides people.]

Populations, Sampling Elements, Frames, and Units

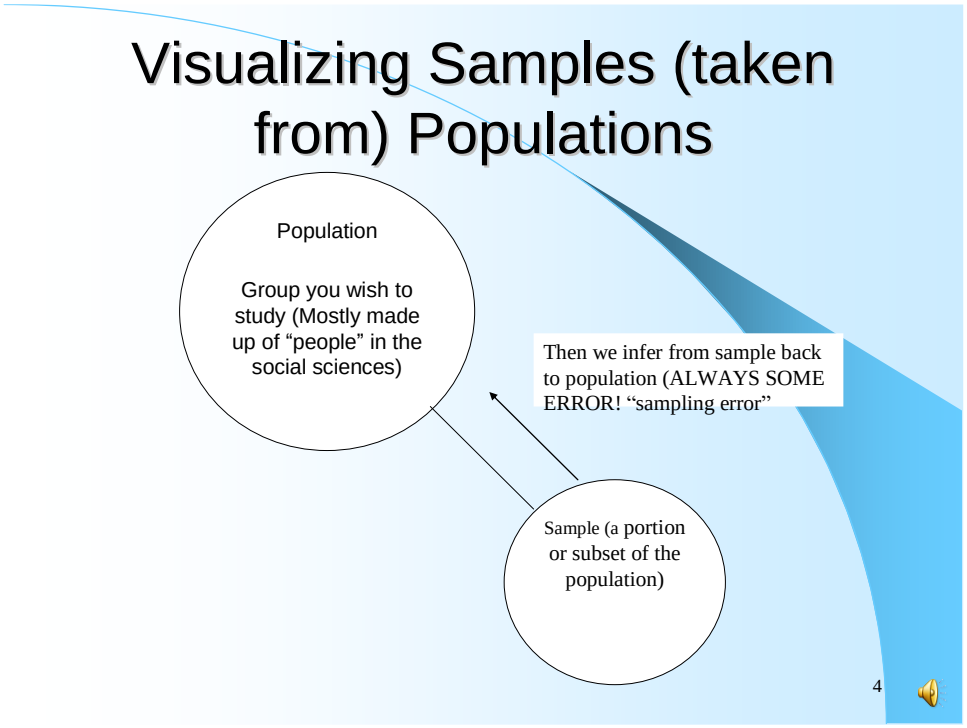
A researcher defines a group, "list," or pool of cases that she wishes to study. This is a **population.** Another definition: population = complete collection of measurements, objects or individuals under study.

sample = a portion or subset taken from population

funny circle diagram

so we take a sample and infer to population

Why? feasibility – all MD's in world , cost, time, and stay tuned for the central limits theorem...the most important lecture of this course.



This population is made up of the things she wishes to actually study called **sampling elements. Sampling elements** can be people, organizations, schools, whales, molecules, and articles in the popular press, etc. **The sampling element is your exact unit of analysis.** For crime researchers studying car thieves, the sampling element would probably be individual car thieves – or theft incidents reported to the police. For drug researchers the sampling elements would be most likely be individual drug users.

Inferential statistics is truly the basis of much of our scientific evidence. We hope that what we study (a sample) is representative of the population from which the sample was drawn. You are in a statistics class to learn about the process of inferential statistics, to learn what are the "mathematical rules" that allow us to infer from a sample to a population

There are two main types of sampling methods that you should be aware of: probability vs. non-probability samples. If we use a method that ensures a probability sample we know in advance how likely (what is the probability?) it is that a sampling element will be selected from its population. If our method does not allow us to know this likelihood (or probability) it is non-probability sample.

Inferential statistics depends upon a random sample or a probability sample. The best scientific data "out there" is based upon probability sampling. However, for some subjects probability sampling is very difficult or even unethical. For example, most of the drug research (prescription and illegal) is based upon non-random samples. Ethnography has been used by researchers to study crime.

Simple Random Sample= all of the people (or sampling elements) in the population have an equal opportunity of being selected into the sample. Idea of drawing names out of a hat, or balls from lottery game. And actually there is a difference between a random sample and a **random representative sample**. For this class we will say that one of our assumptions for inferential statistics is a "random sample," but technically speaking what we really mean is a "random representative sample."

How a simple random sample of Oahu would not be "representative" of Oahu.

Imagine you wanted a sample to represent all people living on Oahu and you put everyone's name on a ping pong ball (like keno or bingo) and had a machine that would spit out 2,000 balls randomly (again like keno or bingo but with way more balls in the cage or hopper). So sample would be random, but not representative of all of Oahu. Why? Because most of the people on Oahu live in the "urban core" say between Diamond Head and Pearl City. So your sample would tend to get a whole lot of those people and not enough people from the less populated areas such as Kahuku or North Shore, etc. You could apply the same logic to the population of voter in the US. If you did a simple random sample, your sample would be made up mostly of people who live in the big cities on the

east and west coasts and not enough people from the sparsely populated states such as North Dakota, Wyoming, Idaho, etc.

So if you do a simple random sample your sample will end up being disproportionately made up of people from large urban areas, right?

Real representative surveys use "disproportionate stratified random samples"

To oversimplify it, researchers will essentially use known characteristics of the population to purposefully "oversample" these smaller groups that are less likely to be selected in a sampling frame. They then use mathematical formulas to weight the individuals so they achieve a truly representative random sample. Take the example of politics. We know rural areas tend to be largely Republican, but pretend we know there is a small but significant subset of "non-white Democrats" living in rural areas of the Midwest. The researcher makes sure to over-sample them and use math to "weight" them so the final sample is representative of all voters in the rural Midwest.

Good random [representative] samples require a good sampling frame

Good random samples depend upon having a **sampling frame** that is representative of the population.

When we make a "list" that operationalizes our population and closely approximates all of the elements in our population we have created a **sampling frame.** A sampling frame could be telephone numbers [be they landline, mobile, or both], DMV records, voter registration lists, all of the people who frequented the school common area at the time you collected your data, etc. Many times it is **quite difficult** to find a **sampling frame that closely approximates all of the elements in your "targeted" population**. This is **especially true** with **"deviant' or "hidden" populations** such as drug users, criminals, homeless, etc.

A good sample is usually dependent upon a very **good match between the sampling frame and the elements**. As you might imagine, it is difficult to come up with a good

sampling frame for of criminals, or drug users. This is true for any “hidden” or “deviant” population.

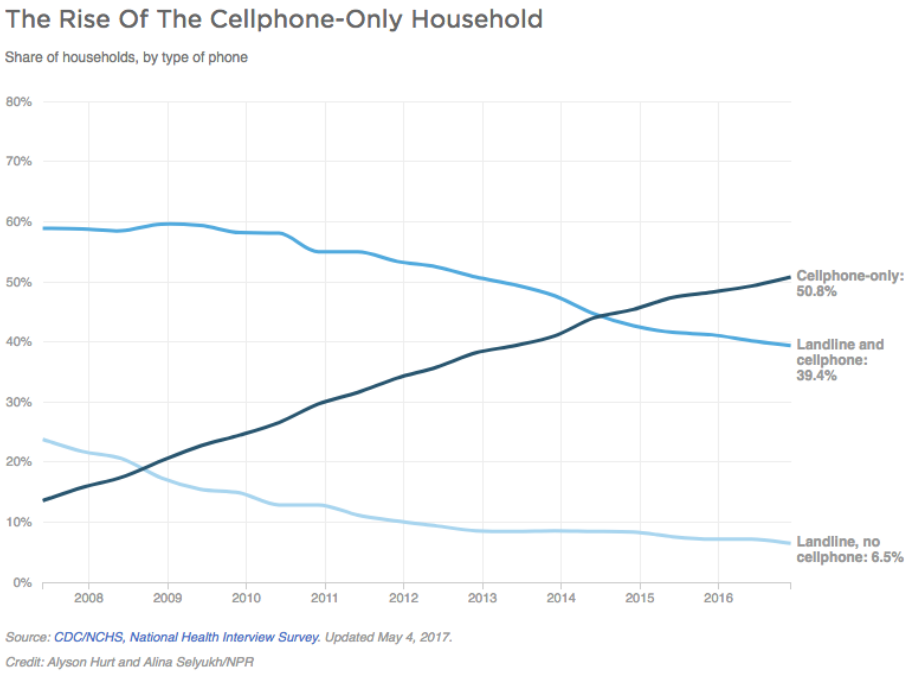
Good and bad matches between sampling frames and the elements in the population

Imagine voting registration records to make a sampling frame of voters, or listed phone numbers in a phone book to approximate homeowners. These are pretty good matches.

Bad Matches

Imagine using DMV records (car ownership) to reach jurors (as in CA), or voting registration lists to reach the homeless. These are extreme examples of poor mismatches but hopefully they illustrate the potential problems.

Back in the day, about 95% of the household population had landline telephones. Therefore the books used to teach that landlines are a pretty good sampling frame. That has changed.



So imagine using landlines to reach the population of registered voters? What would be wrong with that? Well in 2008 when President Obama was elected some polls did that. As late as 2017 it was illegal for pollsters to use computerized random digit dialing to reach cell phones. You could use it to reach landlines, but to reach a cell phone you had to have a real person dial the number. [Hint: it is more expensive to pay a person than a computer to dial phone numbers.] So if you had a sample of registered voters that came from landlines in 2008, what type of voters would you systematically miss? The young. Why? Because at that time the young were more likely to have completely dropped landline service and use a cell phone as their only phone.

How about this crime example: I want to study drug dealers and I use as my sampling frame all of those convicted for that crime in Phoenix last year. Is this a good match? Why or why not? (It is not a good match because it might just include the really "bad" or "unlucky" or perhaps even "stupid" drug dealers.)

known vs. unknown sampling frames for populations

The sampling frames for some populations are intrinsically knowable while some are "unknowable." This is important, because when we seek to take a truly random sample from a population, we need to have a sampling frame that matches the population as closely as possible. The closer the match between the population and the sampling frame, the better the sample will be.

Some populations are "unknowable" and there is literally no way to come up with a truly matching sampling frame. If there is NO way to accurately count the members of the population it is unknowable. If there is "no list" of members of the population then it is best thought of as "unknowable." For example, almost all deviant or "hidden" populations out there are "unknowable" as there is no list of marijuana users, cocaine users, street prostitutes, homeless people, surfers, etc. So for populations that are "hidden" (some would be considered "deviant" and some could be quite "conventional"), there are no perfectly matching sampling frames that exist and a researcher has to "do the best they can." Try to imagine sampling frames for the hidden populations of drug addicts, drug

dealers, car thieves....these would all be considered “deviant” populations in some way and you cannot come up with a perfect match. (Politically correct note: deviant is NOT a moral statement or judgment: it is a statement about how conventional the group is considered to so called “main-stream” society). There are also many very conventional populations that are still “hidden.” There is nothing deviant about surfers, divers, fishermen, and mountain bike riders, but each of these populations are “hidden” if you think about it. There is no list of ALL surfers, divers, fishermen, or mountain bike riders that could be used as a sampling frame.

Some populations are “knowable” and there are perfectly matching sampling frames (e.g. using the sampling frame of “a voter registration list” to reach the population of “voters”) or very good matching sampling frames (e.g. using the sampling frame of phones to reach the population of “home owners.”) So, knowable populations are those where there is probably a list out there (even if you do not have access to it). Registered voters are registered on a list. So are homeowners, car owners, etc.

Now to complicate matters. Some populations are knowable and a perfect sampling frame exist for the, but they might as well be unknowable, because it is very unlikely that a researcher would ever gain access to it. So, just because some government agency has a list (a perfectly matching sampling frame), that does not mean that a researcher will have access to that list to use as a sampling frame. The classic example is children in public schools K-12. This is a knowable population and there is obviously a perfectly matching sampling frame for it: the list of registered students. But, since they are “minors,” it is exceedingly difficult for researchers to gain access to such lists. The same can be said for “juvenile delinquents,” as court records of juveniles are generally considered “private.” And even though there exists as perfectly matching sampling frame for the population of “registered UH students,” a generic researcher would be unlikely to gain access to that list due to privacy concerns. A researcher would need major political connections to be granted access to lists of school children or juvenile delinquents, even though these populations “have a list” and they are “knowable” in the theoretical sense.

Exercise for in person class: decide whether or not the following populations are knowable or unknowable/ whether or not there exists a "list" that could be used as a sampling frame for the following populations. Remember that some answers will be "it depends upon how the population is defined."

- all professional baseball players in the world (MLB in USA)

This one depends. All professional baseball players in the world is probably best described a "unknowable," as there are many countries with small "semi-pro" leagues that do not keep centralized lists.

- all registered students at UH

This is a knowable population and there exists a perfectly matching sampling frame for it, but the list for this population would most likely be "off limits" to most researchers.

- illicit drug users

If this population is defined as "all illicit drug users" then the population is unknowable and there is no list that could be used as a sampling frame. Now if you defined the population as "illicit drug users who were convicted of possession of illicit drugs" then there is such a list that is kept by state/federal court systems (but the researcher would need political connections to gain access to these sampling frames).

2)which are samples and pops

- people who drive cars
- registered voters who responded to CNN survey
- pot smokers in HI

The answer to all of these above is "it depends on how the population is originally defined by the researcher." Each could be either a sample or a population.

Variables - we use variables in statistics to study social world

When we study the social world using statistics we need to define characteristics that we wish to study so that they can be expressed using numbers.

A book definition of a variable: (p 42) "a characteristic of interest that can be observed." Examples include gender, , age, # of touchdowns,# of children born, t-shirt size.

Gender: 1= female 2=male

Age (in whole years): _____ [insert number]

# of touchdowns scored: _____ [insert number]

# of children born to a woman: _____ [insert number]

T-shirt size: 1= small 2= medium 3=large.

But note that a variable must vary. If you are doing a study of the population of teen mothers, then the characteristic of gender would NOT be a variable. It would not vary in the population as all the people in the population are, by definition, women. So a variable must vary!

"coding" variables

When we assign a number to a category of a variable we have "coded" it. Pretend we have a variable that measures "gender" 1= female 2=male. We assign the number 1 to female and the number 2 to males. Thus, we have "coded" men as 2 and coded females as 1.

Pretend we have a variable that measures age in whole years: Age (in whole years): _____ [insert number]. The number is the "coding" and is quite common sense. If you are 41 years old, you insert the number 41 and that is the coding. Many "natural number" variables code in a very common sense manner like this. Imagine *height in whole inches, weight in whole pounds, number of cars you own, number of dollars you paid for your house.* Each of these would "code" naturally with the number.

So in the lecture notes for this course you will often see variable expressed with their coding (as above):

- Age (in whole years): _____ [insert number]
- # of touchdowns scored: _____ [insert number]
- # of children born to a woman: _____ [insert number]
- T-shirt size: 1= small 2= medium 3=large.

Stay tuned for how we classify variables as “continuous vs. discrete” and “nominal, ordinal, interval, or ratio.”

parameter (pop) vs. statistic (sample)

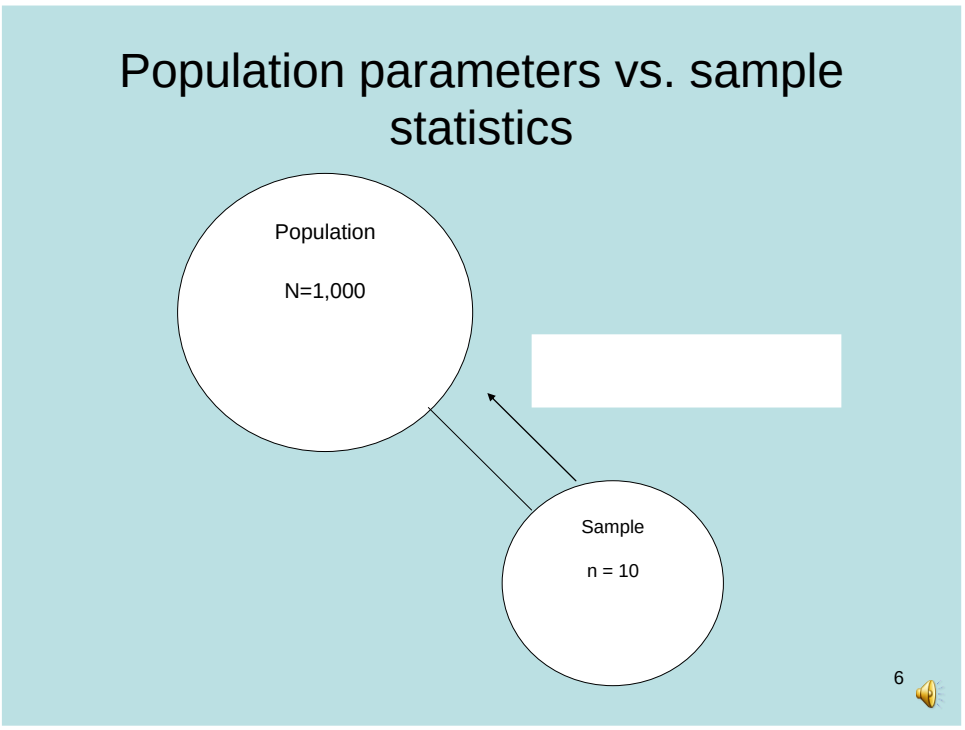
Whenever we refer to a number that describes a population we call it a “population parameter” or more commonly simply a “parameter.” For example when we describe the size of a population we refer to it by the CAPITAL letter N. [“N” = number] So, using statistical shorthand, a population with 1,000 members would be look like this:

N=1,000

Whenever we refer to a number that describes a sample we call it a “sample statisitc” or more commonly simply a “statistic.” For example when we describe the size of a sample we refer to it by the SMALL CASE letter n. [“n” = number] So, using statistical shorthand, a sample with 10 members would be look like this:

n=10

Below is our “funny circle diagram” to help you visualize what I am talking about:



You will see many many symbols that are statistical short hand for “parameters” and “statistics” in a statistics course.