BASICS OF SAMPLING

M. E. Swisher, 2018

Use this document when you try to complete the Flow Chart for Reports You Read or the Flow Chart for Your Research Designs. This document discusses key terms in the sampling literature. Some of the statements you will read in the literature are confusing, including some of the readings assigned in this class. Many people use term interchangeably and I would say that relatively few researchers are very careful about distinguishing between terms. None of the readings or videos for this class include all of the terms clearly defined in one document or video. This is my attempt to put all of the key ideas in one document and avoid ambiguity or confusion. **Use the terms I say to use as I define them in this document in this class.**

KEY TERMS

Theoretical population

The group of people (population) to which you want to be able to extend (generalize, apply) your conclusions, defined by shared characteristics of interest with regard to the specific research question and research objectives. The point of scientific research is to reach science-based ***conclusions.*** Whether the author explicitly uses theory or ever uses the term "theoretical population," a scientist almost always wants to reach conclusions that refer to some group of people that are defined by specific, theoretically-based traits. These traits are often general – they are based on the constructs in a theory, not variables. The theoretical population is the ***entire population everywhere*** that you can reach conclusions about "in theory," even though you cannot ever hope to take a sample from some pool of all of them. It's just not possible. The theoretical population could indeed be global – all youth everywhere in the world to whom we can apply the socio-ecological theory of youth development. That really is ALL youth in the world. The conclusions you want to generalize are ***theory-based conclusions*** – e.g., the effects of family interactions with the child on early emotional development. The author will probably not state "The theoretical population for this study is…" You have to figure this out for yourself most of the time. We never really can (at least in my experience) take a sample that we can reasonably assume is representative of the entire theoretical population ***with regard to traits and characteristics that could affect the results of the study***, but we can, over time with many studies, arrive at some confidence about the impact of various factors. For example, we can reach ***theoretical conclusions*** about the impacts of violence in the home on early childhood emotional development. The specific types of the impacts may differ by culture, place, intensity of the violence, etc. but we actually do have a "mountain of data" that supports the theoretical linkage between violence in the home and "not good" emotional development for children. Many people use the terms target population, population of interest, and theoretical population interchangeably. You will have to distinguish between these in your assigned readings. This is an unfortunate outcome of our lack of precision and agreement on terminology that you will see time after time in this class.

Accessible Population or Target Population

Both terms, and sometimes the term population of interest as well, are commonly used interchangeably. The accessible population is a group that is representative of (like) the theoretical population with ***regard to traits or characteristics that could affect the results of the study*** that you can actually reach***.*** Research takes a lot of time and money. Researchers

have to limit where they sample, except perhaps in the case of "internet based" sample (see below for a discussion of the growing concerns about internet sampling). Therefore, we conduct research with a smaller, local population that we can actually gain access to – that we have the money and time to reach. This accessible population should be "like" the theoretical population with regard to traits that are important from a theoretical perspective -- like children who experience trauma or violence in the home. It is also true that we usually define the accessible population in terms of variables, not broad constructs. E.g., we go from "everyone, everywhere in the theoretical population to people in X area who have the specific traits A, B, and C that represent the theoretical construct Z. The variables are supposed to represent the constructs of interest in a particular study context. For example, we might say that the population of interest for a study consists of youth between the ages of 14 and 18 with repeated episodes of behaviors leading to sentencing under the Juvenile Justice System in Florida. The theoretical population might be much broader and, if you could really talk to the researcher, s/he would say something like "I am trying to understand the behavioral outcomes of youth exposure as young children to antisocial behavioral patterns within the family setting." That's the theoretical population which in term of traits we use to identify an accessible population could translate to "youth in trouble with JJS more than once." Use the term **accessible population in this class when you refer to a specific group of people who are used to represent a theoretically defined population.**

Sampling Frame

A sampling frame is a list or a map or some other tool that literally lets you identify every possible member of the accessible population. Usually, however, researchers apply additional criteria to create the sampling frame. It very rarely includes the entire accessible population. Assume you are conducting a study of that is based on benefit-risk theory. For example, assume you want to study whether benefit-risk theory can be used to understand high-risk decision-making. You decide to focus on hurricane preparedness because deciding how much to prepare for a potential event of this magnitude and type is a good example of having to make a decision under conditions of high risk – but also high uncertainty. You decide to focus on people in the Southeast United States. So now you have an accessible population defined to some degree, but this is still a LOT of people. You decide to limit the study to Florida. Florida is frequently under risk of hurricane and you live here. Still a LOT of people. So now you decide to trim the accessible population to Florida property owners. They should be very concerned about property damage so this makes sense. Renters might make the decision to just leave more easily because they do not own the property that is at risk. This is still a LOT of people. You limit further to coastal counties only. Finally, you start to try to create the list and you realize that a lot of people own coastal property but do not live in Florida year-round. Finding out who really lives in Florida year-round just is not possible. You end up with a list of people who own property in coastal counties in Florida who *filed for the homestead exemption tax rate in the previous year.* This is a subset of the list of property owners in the coastal counties. But it is the list you can actually get. It will consist of people who signed a legal document (you can get in big trouble if you lie) that says that the property in Florida is the signee's primary residence. This gives you some confidence that the list will contain the year-round residents. However, look how far you have come from your original theoretical population or even the original accessible population. Finally, your sampling frame is a sort of "proxy" for property owners in counties at high risk of hurricanes who are year-round residents in Florida. It is as much "like" the accessible population as you can get. These kinds of multiple steps are the main reason why most sampling frames do not contain the entire accessible population.

Additional Screening Criteria

These are traits or characteristics that the individual unit of study (house, cat, person, community, school) must have or in some cases must not have to be included in the study. You define some of these early in the process of identifying the theoretical population, accessible population and sampling frame. However, they are often additional criteria that are used to reduce "noise" in the sample. I have never conducted a study that did not employ screening criteria. For example, one of my students conducted a study of the environmentally responsible behaviors (ERBs) of college students. Her two comparison groups were incoming freshmen and seniors. She wanted to know whether the exposure to UF's "pro-environmental" atmosphere was associated with more ERBs. Part of the question had to do with the age of this group – college is a time when young people "discover who they are, start developing their own ideas about how to act independent of their families, etc." E.g., they "grow up." She was specific about this – adoption of ERBs by emerging adults in four-year institutions of higher education in the United States. So one screening criteria was that the freshmen had to be between the ages of 18 and 21 and the seniors between the ages of 21 and 24. We have freshmen at UF that are much older. By using this criteria, the student eliminated students who are not "emerging adults." She also screened for nationality. Only students born in the U.S. were included. Why? Because the process of "emerging adulthood" may be of limited value in many places. Some researchers argue that such an extended process of achieving adulthood and particularly the role of leaving home to go to college are limited to post-industrial places like the US. Therefore, the whole idea might not apply to many international students. Finally, she was looking at the role of the University environment (our sustainability programs and such) versus what one heard or learned from their family in determining whether students practice ERBs or not. ERBs of the sort we were concerned with are also emphasized in "post-industrial, wealthy places of high consumption," another reason for limiting the students to those born in the US. These were all criteria used to make sure the students included in the sample were "appropriate for the domain of the theory and the research question." We did not want a sample of "all students at UF." Quite the contrary, that kind of general sample would have made it impossible to reach any theory-based conclusions.

Sample

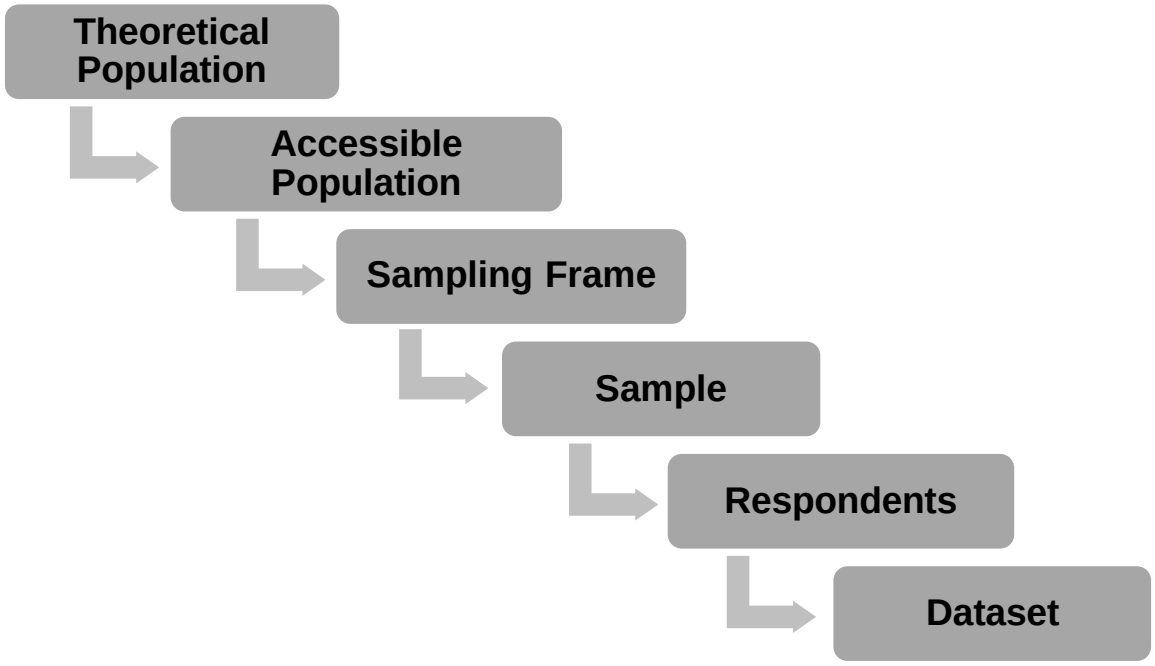
The sample is the final list of people, cats or bears that you select, preferably randomly, from the sampling frame. The sample consists of the individuals that you actually try to get to participate in your study. You contact them and they say "yes" or "no."

Respondents

Respondents are people who (1) agree to participate in your study *and* (2) actually complete the process. The ***response rate*** is the percentage of people you contacted in the sample who become respondents. Assume you have a sample of 500 people. Of those, 450 agree to participate and 50 refuse. The 50 who refused to participate are ***non-respondents***. The ***response rate*** is 450/500 or 90%. Of the 450 who agree to participate, 425 actually complete your study, but 25 drop out at some point. For example, in an on-line instrument, they might just quit answering the questions at some point. You can use various techniques to replace an occasional missing response for a respondent, but not when this phenomenon of "drop out" or ***mortality*** occurs. (Morality means "they quit answering or left the study.") You will have to throw out their responses because they dropped out. Now you have a final total database of 425 sets of responses. It is important to distinguish between non-response and mortality. A low response

rate creates concerns about the quality of the sample. One way to address this problem is to select a sample of the non-respondents, re-contact them, and ask why they did not want to participate. I use standard categories like “not enough time, do not like to provide information about myself to others, never respond to surveys, and did not like the topic of the study.” The first three are no cause for concern. Many people just “don’t want to do it.” The latter issue of the topic is a concern if many people select this option because it implies that the non-respondents could well have answered your questions differently than the respondents. Mortality is always a concern if more than a few people drop out. Once people commit to answering your questions or participating in your study, they tend to remain committed. They drop out, typically, because something “irritates” them about working with you or answering your questions. It may be too onerous. They may find the topic disturbing. They may find your data collection procedures difficult to manage. These red flags indicate that you may need to make adjustments to study procedures.

Figure 1: The Trail from Theoretical Population to Dataset



As you can see, the trail from theoretical population that you want to generalize about to the actual dataset you will analyze has many steps. There are potential problems at every step and there are procedures we can use to reduce or eliminate these problems. The following pages discuss common misconceptions about sampling, why probability samples are so important, how sampling affects your ability to generalize your results to people or things that you did not actually study (the accessible and theoretical populations), and what you can do to make both probability and non-probability samples “better.”

Comparison Groups and Sampling

Experiments create comparison groups after the sample is selected by applying a treatment to some members and not to others. In other designs, comparison groups are based on pre-existing differences among people. Comparison groups are usually based on one or more

constructs in a theory in theory-based research. In purely descriptive research, they may be based on a wide variety of variables, including demographic factors. Theory-based treatments or interventions are often the basis for comparison groups in true and quasi-experiments. Comparison groups are often based on age or other cohorts in longitudinal designs. Different age groups or people who started college at different ages are examples. Comparison groups can be based on an individual or a combination of constructs in cross-sectional and case study designs. However, unfortunately, you will find that many cross-sectional and case study designs fail to employ comparison groups. Whenever comparison groups are used, the sampling approach should explain the procedures used to sample for each theoretical population. If possible, the same or similar procedures are employed, but this is not always possible. Researchers also sometimes divide one sample of one theoretical population or population of interest into groups *after they have collected the data, even though they originally had no plan to compare two or more groups.* This is called *post hoc* identification of comparison group. In this case, there is one theoretical population and only one sampling procedure. It is important in these cases to understand how the groups were ultimately defined since this was not a planned comparison.

TYPES OF SAMPLES

Random Samples

Unless there are good reasons for not doing so, take a RANDOM sample. All else equal, random samples are going to give you the greatest ability to draw sound conclusions and generalize them. Many statistical tests require a random sample. See my statistics cheat sheet. I list the assumptions you must meet to use each test. You will see that very commonly used parametric procedures like analysis of variance and t-tests require random sampling. Strictly speaking, failure to take a random sample makes it impossible to generalize your conclusions to other people, places, cats, communities that you did not actually study. Note that I say conclusions, not results. Results are specific to a study. Usually we really do not want to generalize the specific results – the actual numbers like a difference of 10 points on a pre- and post-test of knowledge before and after training. Conclusions grow out of the results in a study, but they are the broad theoretical and explanatory contributions that we make to the body of knowledge. For example, you conduct a study about differences between males and females in the U.S. with regard to life goals at the age of 16-20. Your theoretical hypothesis is that females will demonstrate have more limited life goals than males due to patriarchal norms in the society. The statistical hypothesis states that females' scores on an index than measures scope of life goals will be lower than that of males. You run a statistical test (t-test) with your data and confirm the statistical hypothesis – women score 15 points lower than males on the 100-point index. ***If you took a random (probability) sample,*** you can conclude that your data support the theoretical hypothesis and you can generalize the conclusion to 16-20 year old males and females in the US as whole. Researchers usually want to generalize the theoretical conclusion – not the specific numbers that they got – like the 15-point difference in this example. There are many kinds of random samples. A completely random sample may not be the best choice and often requires much more effort to achieve than some other random sampling procedures (see the video by Ortlieb and my statistics cheat sheet, Guide to Statistics). With regard to your ability to draw conclusions and generalize them, it does not matter what specific type of random sample you take. There are several types of random samples, including completely random, stratified random, systematic random, clustered random, and multi-stage random. See Types of Samples for a description of each.

Random-like Samples

If you did not have a sampling frame, you do not have a probability sample, although you may have a random sample. However, the difference between random and non-random sampling is not as clear as it might seem. You will see differences in definitions of non-random samples in the videos, the readings, and my cheat sheet. It just is not easy. Intercept sampling is a good example of a "random-like, non-probability sample." My students and I use intercept sampling quite often when we need a sample that we just cannot locate through any other way of sampling. We stop people at various venues and ask them to answer some questions. We use bus stops, waiting for the parade at homecoming, actually riding around on buses, the places where students come and go like Plaza of the Americas or the recreation center, and parking lots. I usually want to perform statistical tests on the data from these samples. (You cannot conduct a long personal interview at a bus stop. Therefore, the instruments are usually "check the box" orally administered by the researcher. You get perhaps 10 minutes or less of the respondent's time.) Therefore, I ask myself "Is this sample "random enough" to meet the requirements for statistical tests? Are these samples random? Honestly, I do not know. I think of them as "quasi-random" because they can have some traits of random samples ***if they are well done***. They can only be "well done" if the researcher takes every possible step to reduce systematic bias. If you have a set of "rules to the game," you can have more confidence that you will be able to analyze data statistically. You can also make a much better presentation of your sampling procedures in your dissertation or publications. Here are examples of "rules of the game." Stop every 5th person. Go to different venues to sample. Go at different times of the day, different days of the week. Get others to help in some venues. One student was sampling people who go to local sporting tournaments (civic, not university). He got three or four colleagues to help him cover all of the entrances to the venues instead of just him at one entrance because he could get more responses and because there could be differences in the traits of people who enter Door 1 versus Door 4 just due to the cost of seating or something like that. These procedures help ensure that you avoid introducing systematic bias – for example, asking people who made eye contact or seemed "approachable." I am usually willing to say that these samples have some weaknesses, are unlikely to produce generalizable conclusions, and are not truly random ***but that they are not*** systematically biased. I am cautious about the conclusions we draw and we are careful to say that the conclusions cannot be generalized to the theoretical (or even the accessible) population.

Non-Random Samples

Finally, there are a variety of samples that are not random or "random-like," the true = non-random samples. These include quota samples, snowball or referral samples, volunteer samples, and judgmental or purposeful samples. ***A convenience or haphazard sample is not acceptable by scientific standards of evidence and is not acceptable in the designs you create in this class.*** People use the term convenience in ways that I would not. I do not think that everyone who says they used a convenience sample wants to imply that they made no effort to reduce bias. I have seen some very well done intercept samples, for example, called "convenience samples." The other non-random samples are described in the document Types of Samples (Swisher documents list). Two videos discuss them, and several readings for the second module on sampling provide good material about non-random sampling.

Note that ***only a subset of the non-random samples are judgmental or purposive samples.*** People use these terms sloppily. See the write-up about what constitutes a judgmental sample on the Swisher document Types of Samples. A true judgmental sample requires that you know

something about the specific individuals that you select – something that is more than just screening criteria or generally available information. I think of this as "deep and personal information." The most common way to get this information is from previous research with the individuals or because you or other researchers can identify them as "experts" in the topic of interest. You select the people "on purpose" as individuals. I often use judgmental samples in multi-stage sampling approaches. Typically, I use these samples under two sets of conditions. (1) I contact individuals that ***I know have specific knowledge, experience or insights*** about the topic of the research to help me figure out what factors to include in research instruments before using them in the study itself. This is part of instrument development and testing. For example, I would contact people with a great deal of experience in strawberry production to decide what to ask about in interviews with strawberry producers. (2) The second use is in multi-stage sampling. I select a judgmental sample from a random or non-random sample of respondents based on prior knowledge from the original study or part of my study that lets me identify individuals who will be able to provide distinctive information, different perspectives, or exceptional experiences (see the Malterud reading for the concept of information power). I use the information from the judgmental sample ***to better interpret the results and improve the conclusions I can reach from the larger first-stage sample.*** E.g., I do not treat the judgmental or purposive sample as "stand alone" evidence.

Various types of non-random samples, including quota, volunteer and "snowball" (referral is a more technical term) samples. They are used more than you might think based on what you have seen in the readings for this module. For example, medical research must often rely on what some would call "quasi-volunteer" samples. Assume we have a new treatment for attention deficit syndrome (ADS). We want to compare this treatment to the existing "best practice" for youth between the ages of 13 and 17 who experience ADS. We ask doctors who treat ADS patients to place flyers about our study in their waiting rooms and to give information about our study to patients and/or the parents of patients who currently take the "existing treatment" and meet the age (screening) criteria for the study. This is a volunteer sample because we wait for the patient/parent to contact us. This approach is very common. You have probably seen such flyers in doctor's offices, seen or heard announcements on the internet, television or radio about "participating in a study. We do generalize based on the results in clinical trials using volunteer samples. When we can generalize is not easy to determine, even though everything I say and the readings say may make it seem so.

Design becomes very critical in determining whether you can generalize. We can generalize from medical trials with small sample sizes and a volunteer sample because the design was a true (or sometimes quasi) experiment. Experiments provide protections against reaching invalid conclusions (internal validity) that other designs do not offer, and this does affect our ability to generalize the conclusions we reach (external validity). Design choices, sampling procedures and decisions, and how you analyze the data are highly interrelated and produce combined effects on internal and external validity, and on explanatory power. Look at the Swisher document on "Comparative Characteristics of Design Groups." Note that there are two very different kinds of sampling "logic" – statistical sampling logic (what we have been talking about here) and replication sampling logic. I am not saying that random sampling is irrelevant to replication sampling logic. However, the way we use screening is very different in the statistical and replication sampling logics and the very extensive use of screening criteria in replication sampling logic ***reduces the variance in the sample*** on purpose. I realize this is rather contrary to what I wrote probability sampling and capturing "the full range of variance in the population." We will discuss this in detail when we discuss designs with controls and interventions and how

they differ fundamentally from the cross-sectional, longitudinal and case study designs. The compound effects of sampling and design are critical.

Common Misperceptions about Sampling

Wrong: The theoretical population typically means "everybody" in some geographic location (state, nation, or city). This is rarely true. Researchers are typically interested in some group besides "everybody in X place." Further, while the accessible population is often limited to one place, the theoretical population is rarely limited to one place.

Wrong: The theoretical population is defined by demographic traits (race, gender, ethnicity, etc.). In fact, very often these traits are irrelevant to the theoretical basis for the study and to the research question. Researchers typically compare the demographic traits of the people who agreed to participate in the study to the demographic traits of the accessible or study population (not the theoretical population unless race was a specific component in the definition of the theoretical population). They provide a kind of "check" that can help the researcher identify potential bias in the sample if there are major differences between the traits in the accessible population as a whole and in the sample. For example, assume we have a research question relating to the experiences of incarcerated youth ages 18-21 (theoretical population). We conduct our study in Jacksonville, FL (accessible or study population). We know from court records that 80% of youth ages 18-21 incarcerated in Jacksonville are African American males, but only 50% of the youth who participate in our study are African American males. This should cause us to examine our sample closely to make sure that the difference between the study population and the sample who participated in our study is nothing more than a statistically unlikely sample – one that happens to exhibit a lot of variance with regard to the race of participants purely by chance. We cannot automatically assume that the sample is biased – even though it is not representative of the study population with regard to race. However, we examine our records and learn that a disproportionately large percentage of the young African Americans whom we asked to participate in the study declined. Our study has nothing to do with race as the topic of research. Race is not a component in the theoretical framework for the study. However, it **is** true that African American youth are more likely to be incarcerated, even for minor crimes, than youth of other races. Therefore, in this case, we would conclude that there may well be a difference between the study population and the youth who agreed to work with us ***that could affect the outcomes of the study, a bias in the sample***. It is not that race per se is a predictor of the outcomes, but the difference in response rate for African American youth and other groups could easily be an indirect effect of young African Americans not wanting to respond to questions about incarceration. Since the difference would be likely to affect the outcomes of the study, there is bias in the sample.

Wrong: A convenience sample can represent a theoretical population. Haphazard or convenience samples incorporate none of the required elements of a representative sample. By definition, they cannot be treated as representative of the theoretical population. I think if more people would use the term "haphazard" instead of the nicer sounding term "convenience," we would see the term less in publications. Imagine saying "I sampled haphazardly, putting the minimum possible effort into getting the sample, paying no attention to the requirements of a representative sample at all. However, I want you to accept my conclusions." Who would really write **that** in a publication? However, you will see many articles that use the term convenience sample. Unfortunately, this is more common in studies that use qualitative data analyses than those using quantitative data analyses. Some of these studies also exhibit little concern about other aspects of research design.

Wrong: If I place any limitations on the traits of people included in the theoretical population, I have a biased sample. On the contrary, defining the theoretical population is critical for research – unless your intention is to generalize what you conclude to the entire global population. In fact, you will often find that it is better to define the theoretical population narrowly. This is true in experiments, for example. The purpose of an experiment is to test whether some intervention has any effect – does it work or not? It is much easier to determine if the intervention works at all, has any effect on the outcome, if you reduce the effects of all sorts of other traits of people. If you cannot detect an effect with a very well defined population where the treatment "really should work," it is very unlikely to have an effect when the population exhibits more variance ***with regard to traits or characteristics that could affect the outcome of the study.*** One of the references for this week makes a statement that could lead you to believe that this constitutes creating a biased sample. This is actually **not** what the author means and he clarifies later, but the statement itself is not easy to understand. Defining the theoretical population is fine – in fact, desirable. However, you cannot place additional restrictions on the sample other than those that you used to define the theoretical population. You cannot "add" traits of interest or traits to eliminate (screen out) people. The sample must represent the theoretical population ***as you defined it.*** The statement also implies that any screening criteria are "bad." Again, this is not true. You must screen potential participants to make sure they do fit the criteria you used in defining the theoretical population.

Incorrect ideas about the accessible population

Wrong: The fact that you do the study somewhere, with some specific people means that it is a case study. If that were true, ***all scientific research designs*** would be case studies because all scientific research occurs ***somewhere***. Even physics research occurs somewhere.

Wrong: Any place is fine. This is not true either. You have to make sure that the site for your research does not have traits that would tend to produce a biased sample. An example would be conducting a study about the value of incarceration in Starke, Florida. There is a large state prison in Starke. Many people in the town work at the prison. The economy of the town is tied to the prison. This is not a suitable place to conduct the study about the value of incarceration. Most research can be conducted in many places without danger of creating a biased sample just because of the location. However, you must examine your research question and determine whether there are logical reasons to believe that people in Place Z would differ from the theoretical population with regard to traits or characteristics that could affect the results of the study. Only you can make that decision and it is largely based on seeking out the relevant information about the locale in which you plan to conduct a study.

Wrong: Researchers do not have to be concerned about identifying an accessible or study population for internet-based studies (like "on-line surveys"). Even internet-based sampling poses the need to define a study population. For example, many researchers use Qualtrics or some other company to secure a list of e-mails. This list is the sampling frame and it was extracted from a study population. Usually, you specify the traits you want in the people who participate in your study. You want to compare attitudes toward corporal punishment among rural and urban residents who are "between the ages of 35 and 60, own their own home, and have been residents in the same location for at least five years." You would define "rural and urban" to the company. They give you two lists – but how did they find all those e-mail addresses and know all this information about the people? They often use people who register with them as "willing to respond to surveys" or participate in studies for a fee (they are paid to do so). This is a volunteer sample, not a probability sample – by definition open to bias. In this

case, the study or accessible population consists of people who have computers, somehow found out that they can earn money for completing “surveys” on line, and have time to answer surveys regularly. Those traits would generate bias in many studies. On the other hand, there are existing databases that consist of people who have responded to previous research. They already provided data – often as part of some nationwide survey of by a government entity or because they used some service like a hospital for childbirth. Those are fine study populations for many projects. You simply select the individuals who have the traits you require in your research. Payment can also be an incentive to answer questions, particularly when you have no way of knowing anything about the study population because companies do not provide that. They select the list of e-mails to use. Even if a respondent honestly is not able to answer well because s/he does not have the right experience or knowledge, there is no way to control who is on the list and who responds. I think this is often true of internet-based studies that use payments to attract participants, especially when the study population consists of a pool of people who registered as potential participants in on-line studies.

Judging the Quality of the Sample

Ask Yourself…

Q1: Was the sample good enough to permit generalization of* any *sort? A sample of six nonprofit organizations that are all located in Gainesville, Florida probably does not provide ***any basis for generalization*** about the goals and objectives of non-profit organizational management. Unless there is something really fascinating about the study for some other reason, it probably belongs in the “do not read” list.

Q2: Was the author (including you) forthcoming and clearly describe the sampling procedures used? If the author fails to make accurate, clear, detailed statements about the sampling procedures, you should be very cautious about accepting either the results or the conclusions of the study. It is especially important that the author explain how and why sampling decisions were made, especially if the sample is not a probability sample. You have to decide whether the arguments are persuasive or not. Try not to be overcritical, but do not accept “any old haphazard sampling approach” either.

Q3: Did the author explicitly discuss the implications of the limitations in sampling on generalization? If not, perhaps the author does not understand the limitations or chooses to ignore them and generalize as though they did not exist. Many authors specifically indicate the limitations in a separate section in the article. They often make clear samples warning the reader about the degree to which the conclusions may be restricted to the sample.

Q4: Did the author take steps to make the sample “as good as possible under the conditions,” or was s/he seemingly willing to use haphazard or convenience sampling? We all face challenges in sampling. Even when a probability sample is not possible, there are many ways to improve non-probability samples, some discussed below. An author who shows no effort to get the “best possible” sample should be very, very cautious about trying to generalize in any way, especially recommendations regarding practice.

Q5: Was the author (including you) circumspect or conservative about the nature of the conclusions that s/he generalized? There is a relationship between the breadth or scope of the conclusions an author wants to make and the sampling process. In general, authors should limit the conclusions they try to reach to be realistic given sampling limitations. I am much more

willing to accept limited conclusions than broad, sweeping ones when the sample is non-probable, especially if the authors acknowledge the flaws and purposefully limit the scope of their conclusions.

SYSTEMATIC BIAS

Think critically about sampling procedures that will introduce *systematic bias* into the sample. Generally, most readers regard a sample as "good enough" to reach sound conclusions and perhaps even to generalize them when there is little danger that the sample is systematically biased. Any sample, even a true probability sample, can be unrepresentative of the population. This occurs because any one sample could include many people (cats, communities) that are atypical of the population purely by chance. This is why we use the term "probability" sample – you try to reduce the probability that your sample consists of atypical cases. However, you never eliminate that possibility. One feature of probability samples is that repeated samples will capture the full range of variance in the traits of interest in the population as a whole. This is the key to their power. Of course, you only get to take one sample in most cases, and you can always get that atypical (extreme) sample. However, you have a "good probability" of getting a sample that is ***"unbiased" and "captures the full range of variance in the population." A systematically biased sample, on the other hand, is one that has a consistent bias in it that is a result of sampling procedures.*** It is a "bad" sample because it violates the key purposes for probability sampling. Here is an example.

> Assume I want to study alcohol consumption by students at large public universities in the southeastern U.S. UF students are the study (accessible) population. There is no reason that I can find in the literature to think that UF students have traits or characteristics that would make them poor representatives of the theoretical population. Now I need a sample that is "good enough" to reach conclusions about the weekend alcohol consumption of UF students as a study population. I could go to bars and ask "people who look like students" whether they are enrolled at UF and how many drinks they have had as they leave the bar. This is a terrible procedure. I am making a judgment about "who looks like a student" and I am obviously capturing students who go to bars. If I do this at 10:00 PM on Friday night, I am probably getting people who have already been drinking for a while and may not be in the best mental condition to answer my questions. These are all sources of systematic bias. Other factors can make it even worse. For example, if I did this at the end of the semester (finals time), I would introduce a new source of systematic bias because people could be studying a lot and hanging out at bars less. This bias would presumably lower the estimations I would make regarding alcohol consumption. I would get yet another source of systematic bias if I did this in the hours prior to a football game.

In this case, I would have to choose between alternatives, all of which have some disadvantages.

(4) One is to take a probability sample from the registrar's database. This is an excellent *sampling* procedure. However, now I will have to ask students to estimate how many drinks they consume these kinds of estimations are fraught with error. People like to think they drink less than they do, for example. My questions will force them to try to generalize across all the different kinds of occasions on which they drink – tailgating before a Gator football game, during finals week, when they are out with friends on Friday night, during weeknights. If I get more specific about all the possible occasions, the mortality (drop out) rate will go up because people just cannot sort all of this out

easily – too many questions that are too specific. Ultimately, most will "generalize" even if they do answer all the specific questions. All of these are sources of ***sampling error.*** The probability sample is the best, but the data generated may not be the best.

(2) I could try to go to all of the venues where students consume alcohol at all the different kinds of occasions when they are at these venues – Friday night at 10:00 pm, football game tailgating, exam week, etc. This is probably impossible and I will still get a sample that contains systematic bias that overestimates alcohol consumption.

(3) I could take repeated samples from the registrar's database on occasions that coincide with presumably different patterns of alcohol consumption and ask specifically "How many alcoholic beverages did you consume on…?" Take sample 1 during a "normal" week and ask about alcohol consumption on Tuesday night. Take sample 2 right after a big home Gator football game and ask about consumption in the 8 hours prior to the game. Take sample 3 during finals week – and so forth. More samples would be better. This will require more effort, but the procedure will produce more accurate data than a single sample, thereby reducing sampling error. Multiple probability samples are very unlikely to produce random effects due to atypical respondents.

(4) Another option would be to resample the same students on all of the occasions mentioned in option 3. This has the advantage of giving me the ability to understand how drinking patterns change in different settings (Gator football game versus finals week). However, the people in the sample will be apt to drop out (mortality) over the course of the semester or year. One might suspect that those who drink the most will be most apt to drop out because the memory task would be more difficult for them. A person who never drinks would mark "0" every time and be done with the questions. For people who do drink, there would be several questions – what kind of drink (beer, wine, whiskey, etc.), in what period of time (1 hour, less than 2 hours, 2-4 hours, etc.), and other details. This is onerous and at some point could become psychologically distressing to people who see that they are consuming alcohol more than they should. In either case, the mortality rate goes up. In the case of psychological distress, the data would show that this was systematic bias that reduces the estimate of alcohol consumption for those who drink the most.

As you can see, there are several options. They vary with regard to (1) the effort and expense required to get accumulate the data, (2) the kind of statistical analyses the researcher can perform, (3) the researcher's ability to draw conclusions, and (4) the generalizability of conclusions one can draw. Never make quick decisions about sampling. In my experience, it is **never simple.** Write down every decision you made and why you made it. Otherwise, you forget. This record allows you to provide a complete description in your publication, which forestalls criticism. Consider every aspect of your procedures carefully. The sampling decisions the researcher makes are crucial to all three of the critical aspects of scientific research that we have discussed in this class – internal validity (the quality of the conclusions), external validity (whether you can generalize), and explanatory power (how much your work contributes to the overall body of knowledge).

Other Potential Issues – A Tale of Too Many Responses

Inducements can be problematic, but are often used. I think payment was a problem in an internet-based study I did recently. There were indications that some respondents did not match the required traits – people who have experience with a specific kind of technology used in

farming. I have no way to be sure that this is true, but I think they were responding to the payment for participating. We were looking for farmers in a five-state area with experience using a specific agricultural technology that is not widely in use. This is the theoretical population. There is no list of such people. I wanted to get a list of potential participants from people who work regularly with farmers. My approach was to ask these agricultural advisors to send me a list of e-mail addresses only (no names, no addresses, nothing else) of farmers they know who currently use this technology or at least that they "are pretty sure" they use the technology or have used it in the past. The study population would have been all farmers that these technical advisors know (work with). The sampling frame would have been the lists of e-mails they sent me of "probably known users of the technology." There are problems with the study population, but those problems pale compared to the risks in other approaches. However, there was a lot of pressure to generate data quickly because this study was an added piece at the end of the project. My colleagues asked me to send the e-mail with information about the study to all the technical advisors in the region and ask them to forward it to their entire mailing lists of farmers. The first problem is that I now have a volunteer sample. I did not select participants. They saw the e-mail and decided on their own to participate in the study. Further, as you can imagine, we lost control over who received the e-mail. Who knows how many people sent the e-mail to their lists of farmers? Who knows how many times the e-mail was forward to more and more e-mail lists? I cannot possibly calculate the response rate because I have no way of knowing how many farmers received the e-mail flyer. In short, the decision to go from a controlled sampling procedure where we sent e-mails to people to an uncontrolled procedure was a very poor one. My colleagues thought that the potential for a larger database offset the problems.

Then it got worse. After two weeks during which a few responses per day appeared, we got 300 responses in three or four days. My program assistant was alert and suspected a problem. We closed the study. Fortunately, people had to provide us with some personal information, including a mailing address, to receive the payment, which we sent electronically. We started getting many requests for payment. My program assistant then identified specific suspicious cases. For example, some mailing addresses were P.O. Boxes. Others do not exist in data banks of actual physical mailing address. Some supposedly individual respondents had the same mailing address. We cannot prove that this was an attempt at fraud. We shut down the study, reported the incident to UF, and they told us what to do. However, we **can reduce the threat of systematic bias in the data**. We will have to compare their answers from "suspect respondents" to each other and to "non-suspects," particularly the early respondents that we have good reason to believe are legitimate because we sent out those e-mails directly. I suspect that the e-mail invitation passed among a group of people who know each other or share list-serves. As far as that goes, it could have been one person. At any rate, I hypothesize that the responses among the "suspect" respondents are very similar, e.g., demonstrate little variance. I also hypothesize that they will vary significantly from the responses of "non-suspect" respondents because you really do need some specific experience and knowledge to answer our questions reasonably. We will discard all "suspect data" – and, of course, have to explain all this in our reports and publications. We will probably end up with the same sample size we would have had with the originally proposed procedure of direct contact with potential participants (sigh…). Post-script. We were able to identify the responses with little variance and we removed them from the database.