Sampling

CO-2: Differentiate among different sampling methods and discuss their strengths and limitations.

Video

Video: Sampling (12:38)

Sampling Plans

As mentioned in the introduction to this unit, we will begin with the first stage of data production — sampling. Our discussion will be framed around the following examples:

Suppose you want to determine the musical preferences of all students at your university, based on a sample of students. Here are some examples of the many possible ways to pursue this problem.

EXAMPLES: Sampling

Example 1: Post a music-lovers' survey on a university Internet bulletin board, asking students to vote for their favorite type of music.

This is an example of a **volunteer sample**, where individuals have selected themselves to be included. Such a sample is almost guaranteed to be **biased**. In general, volunteer samples tend to be comprised of individuals who have a particularly strong opinion about an issue, and are looking for an opportunity to voice it. Whether the variable's values obtained from such a sample are over- or under-stated, and to what extent, cannot be determined. As a result, data obtained from a voluntary response sample is quite useless when you think about the "Big Picture," since the sampled individuals only provide information about themselves, and we **cannot generalize to any larger group at all.**

Comment:

- It should be mentioned that in some cases volunteer samples are the only ethical way to obtain a sample. In medical studies, for example, in which new treatments are tested, subjects must choose to participate by signing a consent form that highlights the potential risks and benefits. As we will discuss in the next topic on study design, a volunteer sample is not so problematic in a study conducted for the purpose of comparing several treatments.

Example 2: Stand outside the Student Union, across from the Fine Arts Building, and ask the students passing by to respond to your question about musical preference.

This is an example of a **convenience sample**, where individuals happen to be at the right time and place to suit the schedule of the researcher. Depending on what variable is being studied, it may be that a convenience sample provides a fairly representative group. However, there are often subtle reasons why the sample's results are **biased**. In this case, the proximity to the Fine Arts Building might result in a disproportionate number of students favoring classical music. A convenience sample may also be susceptible to bias because certain types of individuals are more likely to be selected than others. In the extreme, some convenience samples are designed in such a way that certain individuals have no chance at all of being selected, as in the next example.

Example 3: Ask your professors for email rosters of all the students in your classes. Randomly sample some addresses, and email those students with your question about musical preference.

Here is a case where the **sampling frame** — list of potential individuals to be sampled — does not match the population of interest. The population of interest consists of all students at the university, whereas the sampling frame consists of only your classmates. There may be **bias** arising because of this discrepancy. For example, students with similar majors will tend to take the same classes as you, and their musical preferences may also be somewhat different from those of the general population of students. It is always best to have the sampling frame match the population as closely as possible.

Example 4: Obtain a student directory with email addresses of all the university's students, and send the music poll to every 50th name on the list.

This is called **systematic sampling**. It may not be subject to any clear bias, but it would not be as safe as taking a random sample.

If individuals are sampled completely at random, and without replacement, then each group of a given size is just as likely to be selected as all the other groups of that size. This is called a **simple random sample (SRS)**. In contrast, a systematic sample would not allow for sibling students to be selected, because of having the same last name. In a simple random sample, sibling students would have just as much of a chance of both being selected as any other pair of students. Therefore, there may be subtle sources of bias in using a systematic sampling plan.

Example 5: Obtain a student directory with email addresses of **all** the university's students, and send your music poll to a **simple random sample** of students.

As long as all of the students respond, then the sample is **not subject to any bias**, and should succeed in being representative of the population of interest.

But what if only 40% of those selected email you back with their vote?

The results of this poll would not necessarily be representative of the population, because of the potential problems associated with **volunteer response**. Since individuals are not compelled to respond, often a relatively small subset take the trouble to participate. Volunteer response is not as problematic as a volunteer sample (presented in example 1 above), but there is still a danger that those who do respond are different from those who don't, with respect to the variable of interest. An improvement would be to follow up with a second email, asking politely for the students' cooperation. This may boost the response rate, resulting in a sample that is fairly representative of the entire population of interest, and it may be the best that you can do, under the circumstances. **Nonresponse** is still an issue, but at least you have managed to reduce its impact on your results.

So far we've discussed several sampling plans, and determined that a simple random sample is the only one we discussed that is not subject to any bias.

A simple random sample is the easiest way to base a selection on randomness. There are other, more sophisticated, sampling techniques that utilize randomness that are often preferable in real-life circumstances. Any plan that relies on random selection is called a **probability sampling plan (or technique)**. The following three probability sampling plans are among the most commonly used:

- **Simple Random Sampling** is, as the name suggests, the simplest probability sampling plan. It is equivalent to "selecting names out of a hat." Each individual has the same chance of being selected.
- **Cluster Sampling** — This sampling technique is used when our population is naturally divided into groups (which we call clusters). For example, all the students in a university are divided into majors; all the nurses in a certain city are divided into hospitals; all registered voters are divided into precincts (election districts). In cluster sampling, we take a random sample of clusters, and use all the individuals within the selected clusters as our sample. For example, in order to get a sample of high-school seniors from a certain city, you choose 3 high schools at random from among all the high schools in that city, and use all the high school seniors in the three selected high schools as your sample.
- **Stratified Sampling** — Stratified sampling is used when our population is naturally divided into sub-populations, which we call stratum (plural: strata). For example, all the students in a certain college are divided by gender or by year in college; all the registered voters in a certain city are divided by race. In stratified sampling, we choose a simple random sample from each stratum, and our sample consists of all these simple random samples put together. For example, in order to get a random sample of high-school seniors from a certain city, we choose a random sample of 25 seniors from each of the high schools in that city. Our sample consists of all these samples put together.

Each of those probability sampling plans, if applied correctly, are not subject to any bias, and thus produce samples that represent well the population from which they were drawn.

Comment: Cluster vs. Stratified

- Students sometimes get confused about the difference between cluster sampling and stratified sampling. Even though both methods start out with the population somehow divided into groups, the two methods are very different.

- In cluster sampling, we take a random sample of whole groups of individuals taking everyone in that group but not all groups are taken), while in stratified sampling we take a simple random sample from each group (and all groups are represented).
- For example, say we want to conduct a study on the sleeping habits of undergraduate students at a certain university, and need to obtain a sample. The students are naturally divided by majors, and let's say that in this university there are 40 different majors.
  - In cluster sampling, we would randomly choose, say, 5 majors (groups) out of the 40, and use all the students in these five majors as our sample.
  - In stratified sampling, we would obtain a random sample of, say, 10 students from each of the 40 majors (groups), and use the 400 chosen students as the sample.
  - Clearly in this example, stratified sampling is much better, since the major of the student might have an effect on the student's sleeping habits, and so we would like to make sure that we have representatives from all the different majors. We'll stress this point again following the example and activity.

✔ EXAMPLE:

Suppose you would like to study the job satisfaction of hospital nurses in a certain city based on a sample. Besides taking a simple random sample, here are two additional ways to obtain such a sample.

1. Suppose that the city has 10 hospitals. Choose one of the 10 hospitals at random and interview all the nurses in that hospital regarding their job satisfaction. This is an example of cluster sampling, in which the hospitals are the clusters.

2. Choose a random sample of 50 nurses from each of the 10 hospitals and interview these 50 * 10 = 500 regarding their job satisfaction. This is an example of stratified sampling, in which each hospital is a stratum.

Cluster or Stratified — which one is better?

Let's go back and revisit the job satisfaction of hospital nurses example and discuss the pros and cons of the two sampling plans that are presented. Certainly, it will be much easier to conduct the study using the cluster sample, since all interviews are conducted in one hospital as opposed to the stratified sample, in which the interviews need to be conducted in 10 different hospitals. However, the hospital that a nurse works in probably has a direct impact on his/her job satisfaction, and in that sense, getting data from just one hospital might provide biased results. In this case, it will be very important to have representation from all the city hospitals, and therefore the stratified sample is definitely preferable. On the other hand, say that instead of job satisfaction, our study focuses on the age or weight of hospital nurses.

In this case, it is probably not as crucial to get representation from the different hospitals, and therefore the more easily obtained cluster sample might be preferable.

Comment:

- Another commonly used sampling technique is **multistage sampling**, which is essentially a "complex form" of cluster sampling. When conducting cluster sampling, it might be unrealistic, or too expensive to sample **all** the individuals in the chosen clusters. In cases like this, it would make sense to have another stage of sampling, in which you choose a sample from each of the randomly selected clusters, hence the term multistage sampling.

For example, say you would like to study the exercise habits of college students in the state of California. You might choose 8 colleges (clusters) at random, but you are certainly not going to use all the students in these 8 colleges as your sample. It is simply not realistic to conduct your study that way. Instead you move on to stage 2 of your sampling plan, in which you choose a random sample of 100 males and a random sample of 100 females from each of the 8 colleges you selected in stage 1.

So in total you have 8 * (100+100) = 1,600 college students in your sample.

In this case, stage 1 was a cluster sample of 8 colleges and stage 2 was a stratified sample within each college where the stratum was gender.

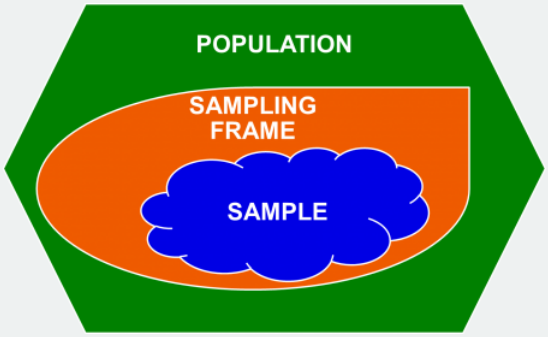
Multistage sampling can have more than 2 stages. For example, to obtain a random sample of physicians in the United States, you choose 10 states at random (stage 1, cluster). From each state you choose at random 8 hospitals (stage 2, cluster). Finally, from each hospital, you choose 5 physicians from each sub-specialty (stage 3, stratified).

Did I Get This?: Sampling

Overview So Far

We have defined the following:

Sampling Frame: List of potential individuals to be sampled. We want the sampling frame to match the population as closely as possible. The sampling frame is embedded within the population and the sample is embedded inside the sampling frame.



Biased Sample: A sample that produces data that is not representative because of the systematic under- or over-estimation of the values of the variable of interest.

Volunteer Sample: Individuals have selected themselves to be included.

Convenience Sample: Individuals happen to be at the right time and place to suit the schedule of the researcher

Systematic Sample: Starting from a randomly chosen individual in the ordered sampling frame, select every i-th individual to be included in the sample.

Simple Random Sample (SRS): Individuals are sampled completely at random, and without replacement. The result is that EVERY group of a given size is **just as likely to be selected** as all the other groups of that size. Each individual is also equally likely to be chosen.

Cluster Sampling: Used when "natural" groupings are evident in a statistical population and each group is generally representative of the population. In this technique, the total population is divided into these groups (or **clusters**) and a **sample of these groups** is selected. For example randomly selecting courses from all courses and surveying ALL students in selected courses.

Stratified Sampling: When subpopulations within an overall population vary, it can be advantageous to **take samples from each subpopulation (stratum) independently.** For example, take a random sample of males and a separate random sample of females.

Nonresponse: Individuals selected to participate do not respond or refuse to participate.

Sample Size

So far, we have made no mention of sample size. Our first priority is to make sure the sample is representative of the population, by using some form of probability sampling plan. Next, we must keep in mind that in order to get a more precise idea of what values are taken by the variable of interest for the entire population, a larger sample does a better job than a smaller one. We will discuss the issue of sample size in more detail in the Inference unit, and we will actually see how changes in the sample size affect the conclusions we can draw about the population.

EXAMPLE:

Suppose hospital administrators would like to find out how the staff would rate the quality of food in the hospital cafeteria. Which of the four sampling plans below would be best?

1. The person responsible for polling stands outside the cafeteria door and asks the next 5 staff members who come out to give the food a rating on a scale of 1 to 10.

2. The person responsible for polling stands outside the cafeteria door and asks the next 50 staff members who come out to give the food a rating on a scale of 1 to 10.

3. The person responsible for polling takes a random sample of 5 staff members from the list of all those employed at the hospital and asks them to rate the cafeteria food on a scale of 1 to 10.

4. The person responsible for polling takes a random sample of 50 staff members from the list of all those employed at the hospital and asks them to rate the cafeteria food on a scale of 1 to 10.

Plans 1 and 2 would be biased in favor of higher ratings, since staff members with unfavorable opinions about cafeteria food would be likely to eat elsewhere. Plan 3, since it is random, would be unbiased. However, with such a small sample, you run the risk of including people who provide unusually low or unusually high ratings. In other words, the average rating could vary quite a bit depending on who happens to be included in that small sample. Plan 4 would be best, as the participants have been chosen at random to avoid bias and the larger sample size provides more information about the opinions of all hospital staff members.

EXAMPLE:

Suppose a student enrolled in a statistics course is required to complete and turn in several hundred homework problems throughout the semester. The teaching assistant responsible for grading suggests the following plan to the course professor: instead of grading all of the problems for each student, he will grade a random sample of problems.

His first offer, to grade a random sample of just 3 problems for each student, is not well-received by the professor, who fears that such a small sample may not provide a very precise estimate of a student's overall homework performance.

Students are particularly concerned that the random selection may happen to include one or two problems on which they performed poorly, thereby lowering their grade.

The next offer, to grade a random sample of 25 problems for each student, is deemed acceptable by both the professor and the students.

Comment:

- In practice, we are confronted with many trade-offs in statistics. A larger sample is more informative about the population, but it is also more costly in terms of time and money. Researchers must make an effort to keep their costs down, but still obtain a sample that is large enough to allow them to report fairly precise results.

Learn By Doing: Sampling (Software)

Let's Summarize

Our goal, in statistics, is to use information from a sample to draw conclusions about the larger group, called the population. The **first step** in this process is to **obtain a sample** of individuals that are truly representative of the population. If this step is not

carried out properly, then the sample is subject to bias, a systematic tendency to misrepresent the variables of interest in the population.

Bias is almost guaranteed if a **volunteer sample** is used. If the individuals select themselves for the study, they are often different in an important way from the individuals who did not volunteer.

A **convenience sample**, chosen because individuals were in the right place at the right time to suit the researcher, may be different from the general population in a subtle but important way. However, for certain variables of interest, a convenience sample may still be fairly representative.

The **sampling frame** of individuals from whom the sample is actually selected should match the population of interest; bias may result if parts of the population are systematically excluded.

Systematic sampling takes an organized (but not random) approach to the selection process, as in picking every 50th name on a list, or the first product to come off the production line each hour. Just as with convenience sampling, there may be subtle sources of bias in such a plan, or it may be adequate for the purpose at hand.

Most studies are subject to some degree of **nonresponse**, referring to individuals who do not go along with the researchers' intention to include them in a study. If there are too many non-respondents, and they are different from respondents in an important way, then the sample turns out to be biased.

In general, bias may be eliminated (in theory), or at least reduced (in practice), if researchers do their best to implement a **probability sampling plan** that utilizes **randomness**.

The most basic probability sampling plan is a **simple random sample**, where every group of individuals has the same chance of being selected as every other group of the same size. This is achieved by sampling at random and without replacement.

In a **cluster sample**, groups of individuals are randomly selected, such as all people in the same household. In a cluster sample, all members of each selected group participate in the study.

A **stratified sample** divides the population into groups called strata before selecting study participants at random from within those groups.

Multistage sampling makes the sampling process more manageable by working down from a large population to successively smaller groups within the population, taking advantage of stratifying along the way, and sometimes finishing up with a cluster sample or a simple random sample.

Assuming the various sources of bias have been avoided, researchers can learn more about the variables of interest for the population by taking **larger samples**. The "extreme" (meaning, the largest possible sample) would be to study every single individual in the population (the goal of a census), but in practice, such a design is rarely feasible. Instead, researchers must try to obtain the largest sample that fits in their budget (in terms of both time and money), and must take great care that the sample is truly representative of the population of interest.

We will further discuss the topic of sample size when we cover sampling distributions and inferential statistics.

In this short section on sampling, we learned various techniques by which one can choose a sample of individuals from an entire population to collect data from. This is seemingly a simple step in the big picture of statistics, but it turns out that it has a crucial effect on the conclusions we can draw from the sample about the entire population (i.e., inference).

Caution

Generally speaking, a probability sampling plan (such as a simple random sample, cluster, or stratified sampling) will result in a nonbiased sample, which can be safely used to make inferences. Moreover, the inferential procedures that we will learn later in this course assume that the sample was chosen at random.

That being said, other (nonrandom) sampling techniques are available, and sometimes using them is the best we can do. It is important, though, when these techniques are used, to be aware of the types of bias that they introduce, and thus the limitations of the conclusions that can be drawn from the resulting samples.