

 Lagunita is retiring and will shut down at 12 noon Pacific Time on March 31, 2020. A few courses may be open for self-enrollment for a limited time. We will continue to offer courses on other online learning platforms; visit <http://online.stanford.edu>.

Course > Producing Data: Sampling > Sampling > Sampling: Probability Sampling Plans

 Bookmark this page

Sampling: Probability Sampling Plans

Learning Objective: Identify the sampling method used in a study and discuss its implications and potential limitations.

Learning Objective: Critically evaluate the reliability and validity of results published in mainstream media.

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, while in stratified sampling we take a simple random sample from each group. 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.

Did I Get This

1/1 point (graded)

What sampling technique is being used in this scenario?

Voters are selected at random from an alphabetical list of all registered voters.

☐ cluster sampling☒ simple random sampling ✓☐ stratified sampling☐ systematic sampling**Answer**

Correct: Voters were selected directly from a list of the entire population (all registered voters).

Submit

Did I Get This

1/1 point (graded)

What sampling technique is being used in this scenario?

Voters are selected by choosing at random several of the city's zip codes and selecting all the voters from those selected zip codes.

☒ cluster sampling ✓☐ simple random sampling☐ stratified sampling☐ systematic sampling**Answer**

Correct:

The population of registered voters was divided into groups (zip codes). A number of those groups were chosen, and then all members of each chosen group were selected to participate in the study.

Submit

Did I Get This

1/1 point (graded)

What sampling technique is being used in this scenario?

Several pieces of fruit from each tree in an orchard are selected.

☐ cluster sampling

☐ simple random sampling

☒ stratified sampling ✓

☐ systematic sampling

Answer

Correct:

The population was divided into groups (trees), then some fruit from each group was selected. Since we do not know if all of the trees contain the same kind of fruit, one way to ensure that we will have a representative sample of fruit is to select some from each tree. Note: Suppose the trees are lemon, lime, orange, and tangerine. One technique would be first to stratify the orchard by type of trees (according to the kind of fruit it has), then select some of each type of tree. This would be a multistage sample, using strata first, and then clusters second.

Submit

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).

Scenario: Risk of Heart Disease

An insurance industry research foundation wants to study the quality of care given to all patients at risk for heart disease in the United States. Since not all those at risk seek treatment, the foundation randomly selects 3,500 claims only from among those at-risk patients who were actually treated for chest pain. The foundation obtains this sample in several stages. First, the foundation identifies 5 large companies that represent a broad cross-section of patients, chooses 2 of the 5 at random, and gains access to the claims of all the companies' patients. The two companies' claims are classified (depending on their origin) to 7 geographical regions (California, Florida, Great Lakes, Midwest, Northeast, Southern, and Southwest), and within each region, 5 counties are selected that represent a continuum spanning rural, suburban, and urban populations. (In total, then, patients from 35 counties are included).

Within each county (and for each company), claims of 25 male and 25 female patients treated for chest pain are randomly selected for the study, for a total of 3,500 patients.

Did I Get This

1/1 point (graded)

In this study, what is the population?

- ☐ A diverse group of U.S. patients treated for chest pain.
- ☒ All U.S. patients at risk for heart disease. ✓
- ☐ A subset of 3,500 from a diverse group of U.S. patients treated for chest pain.

Answer

Correct:

The insurance industry research foundation wants to study the quality of care given to all patients at risk for heart disease in the United States so this is the population of interest.

Submit

Did I Get This

1/1 point (graded)

In this study, what is the sampling frame?

- ☒ A diverse group of U.S. patients treated for chest pain. ✓
- ☐ All U.S. patients at risk for heart disease.
- ☐ A subset of 3,500 from a diverse group of U.S. patients treated for chest pain.

Answer

Correct:

The sampling frame is a list of potential individuals to be sampled. In this study, a diverse group of U.S. patients treated for chest pain represents the sampling frame.

Submit

Did I Get This

1/1 point (graded)

In this study, what is the sample?

- ☐ A diverse group of U.S. patients treated for chest pain.

☐ All U.S. patients at risk for heart disease.

☒ A subset of 3,500 from a diverse group of U.S. patients treated for chest pain. ✓

Answer

Correct: The sample is a subset of 3,500 from a diverse group of U.S. patients treated for chest pain.

Submit

Did I Get This

1/1 point (graded)

For the following stage in the multistage sampling plan of this study, what was the sampling technique that was used?

1) The research foundation identifies 5 large companies that represent a broad cross-section of patients, chooses 2 of the 5 at random, and gains access to the claims of all the companies' patients.

☒ cluster sampling ✓

☐ simple random sampling

☐ stratified sampling

Answer

Correct:

Indeed two whole groups of patients (all the patients from two companies) were randomly selected.

Submit

Did I Get This

1/1 point (graded)

For the following stage in the multistage sampling plan of this study, what was the sampling technique that was used?

2) The 2 companies' claims are classified (depending on their origin) according to 7 geographical regions, and within each region, the sampling continues.

☐ cluster sampling☐ simple random sampling☒ stratified sampling ✓**Answer**

Correct:

Indeed, the claims are divided into groups/strata (regions), and then the sampling continues from within each such group.

Submit

Did I Get This

1/1 point (graded)

For the following stage in the multistage sampling plan of this study, what was the sampling technique that was used?

3) From each region, 5 representative counties are selected. (In total, all the claims originating from 35 counties are examined.)

☒ cluster sampling ✓☐ simple random sampling☐ stratified sampling**Answer**

Correct:

Indeed, in this stage, whole groups of claims were selected. In particular, from each region, 5 groups of claims were selected (each group being all the claims originating from a certain county).

Submit

Did I Get This

1/1 point (graded)

For the following stage in the multistage sampling plan of this study, what was the sampling technique that was used?

4) Within each county (and for each company), claims of 25 male and 25 female patients are randomly selected.

☐ cluster sampling

☐ simple random sampling

☒ stratified sampling ✓

Answer

Correct:

Indeed, the claims are divided into groups/strata (by gender). Note that after the claims are divided into strata, 25 claims are chosen from each gender group using simple random sampling.

Submit

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.

Here is another example:

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.

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 nonrespondents, 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.

Open Learning Initiative 



 Unless otherwise noted this work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License .

© All Rights Reserved