In the language of sampling:

-a **population** is the entire collection of people or things you are interested in;

-a **census** is a measurement of all the units in the population;

-a **population parameter** is a number that results from measuring all the units in the population;

-a **sampling frame** is the specific data from which the sample is drawn, e.g., a telephone book;

-a **unit of analysis** is the type of object of interest, e.g., arsons, fire departments, firefighters;

-a **sample** is a subset of some of the units in the population;

-a **statistic** is a number that results from measuring all the units in the sample;

-statistics derived from samples are used to estimate population parameters.

  

For example, to find out the average age of all motor vehicles in the state in 1997:

Population=all motor vehicles in the state in 1997

Sampling frame=all motor vehicles registered with the DMV on July 1, 1997

Design=probability sampling

Unit of analysis=motor vehicle

Sample=300 motor vehicles

Data gathered=the age of each of the 300 motor vehicles selected in the sample

Statistic=the average age of the 300 motor vehicles in the sample

Parameter=the estimate of the average age of all motor vehicles in the state-1997

**Why Sample?**

    Sometimes "measuring" or "testing" something destroys it. The government requires automakers who want to sell cars in the U.S. to demonstrate that their cars can survive certain crash tests. Obviously, the company can't be expected to crash every car, to see if it survives! So the company crashes only a sample of cars.

    Another reason for sampling is that not all units in the population can be identified, such as all the air molecules in the LA basin. So to measure air pollution, you take a sample of air molecules. Also, even if all those air molecules could be identified, it would be too expensive and too time consuming to measure them all. 

**Types of Samples:**

**Non-probability (non-random) samples:**

    These samples focus on volunteers, easily available units, or those that just happen to be present when the research is done. Non-probability samples are useful for quick and cheap studies, for case studies, for qualitative research, for pilot studies, and for developing hypotheses for future research.

**Convenience sample**: also called an "accidental" sample or "man-in-the-street" samples. The researcher selects units that are convenient, close at hand, easy to reach, etc.

**Purposive sample**: the researcher selects the units with some purpose in mind, for example, students who live in dorms on campus, or experts on urban development.

**Quota sample**: the researcher constructs quotas for different types of units. For example, to interview a fixed number of shoppers at a mall, half of whom are male and half of whom are female.

Other samples that are usually constructed with non-probability methods include library research, participant observation, marketing research, consulting with experts, and comparing organizations, nations, or governments.

**Probability-based (random) samples:**

    These samples are based on probability theory. Every unit of the population of interest must be identified, and all units must have a known, non-zero chance of being selected into the sample.

**Simple random sample**: Each unit in the population is identified, and each unit has an equal chance of being in the sample. The selection of each unit is independent of the selection of every other unit. Selection of one unit does not affect the chances of any other unit.

    For example, to select a sample of 25 people who live in your college dorm, make a list of all the 250 people who live in the dorm. Assign each person a unique number, between 1 and 250. Then refer to a table of random numbers. Starting at any point in the table, read across or down and note every number that falls between 1 and 250. Use the numbers you have found to pull the names from the list that correspond to the 25 numbers you found. These 25 people are your sample. This is called the table of random numbers method.

    Another way to select this simple random sample is to take 250 ping-pong balls and number then from 1 to 250. Put them into a large barrel and mix them up, and then grab 25 balls. Read off the numbers. Those are the 25 people in your sample. This is called the lottery method.

**Systematic random sampling**: Each unit in the population is identified, and each unit has an equal chance of being in the sample.

    For example, to select a sample of 25 dorm rooms in your college dorm, make a list of all the room numbers in the dorm. Say there are 100 rooms. Divide the total number of rooms (100) by the number of rooms you want in the sample (25). The answer is 4. This means that you are going to select every fourth dorm room from the list. But you must first consult a table of random numbers. Pick any point on the table, and read across or down until you come to a number between 1 and 4. This is your random starting point. Say your random starting point is "3". This means you select dorm room 3 as your first room, and then every fourth room down the list (3, 7, 11, 15, 19, etc.) until you have 25 rooms selected.

    This method is useful for selecting large samples, say 100 or more. It is less cumbersome than a simple random sample using either a table of random numbers or a lottery method. For example, you might have to sample files in a large filing cabinet. It is easier to select every 17th file than to pull out all the files and number them, etc.

    However, you must be aware of problems that can arise in systematic random sampling. If the selection interval matches some pattern in the list (e.g., each 4th dorm room is a single unit, where all the others are doubles) you will introduce systematic bias into your sample.

**Stratified random sampling**: Each unit in the population is identified, and each unit has a known, non-zero chance of being in the sample. This is used when the researcher knows that the population has sub-groups (strata) that are of interest.

    For example, if you wanted to find out the attitudes of students on your campus about immigration, you may want to be sure to sample students who are from every region of the country as well as foreign students. Say your student body of 10,000 students is made up of 8,000 - West; 1,000 - East; 500 - Midwest; 300 - South; 200 - Foreign.

    If you select a simple random sample of 500 students, you might not get any from the Midwest, South, or Foreign. To make sure that you get some students from each group, you can divide the students into these five groups, and then select the same percentage of students from each group using a simple random sampling method. This is proportional stratified random sampling.

    However, you may still have too few of some types of students. Instead, you may divide students into the five groups and then select the same number of students from each group using a simple random sampling method. This is disproportionate stratified random sampling. This allows you to have enough students in each sub-group so that you can perform some meaningful statistical analyses of the attitudes of students in each sub-group. In order to say something about the attitudes of the total student population of the university, however, you will have to apply weights to the findings for each sub-group, proportional to its presence in the total student body.

**Cluster sampling**: cluster sampling views the units in a population as not only being members of the total population but as members also of naturally-occurring in clusters within the population. For example, city residents are also residents of neighborhoods, blocks, and housing structures.

    Cluster sampling is used in large geographic samples where no list is available of all the units in the population but the population boundaries can be well-defined. For example, to obtain information about the drug habits of all high school students in a state, you could obtain a list of all the school districts in the state and select a simple random sample of school districts. Then, within in each selected school district, list all the high schools and select a simple random sample of high schools. Within each selected high school, list all high school classes, and select a simple random sample of classes. Then use the high school students in those classes as your sample.

    Cluster sampling must use a random sampling method at each stage. This may result in a somewhat larger sample than using a simple random sampling method, but it saves time and money. It is also cheaper to administer than a statewide sample of high school seniors, because there are many fewer sites to obtain information from.

The differences between Probability (Random) Sampling and Non-Probability (Non-Random) Sampling are summarized below.

|  |  |
| --- | --- |
| **Probability (Random) Sampling** | **Non-Probability (Non-Random) Sampling** |
| Allows use of statistics, tests hypotheses | Exploratory research, generates hypotheses |
| Can estimate population parameters | Population parameters are not of interest |
| Eliminates bias | Adequacy of the sample can't be known |
| Must have random selection of units | Cheaper, easier, quicker to carry out |

**How Big a Sample Do I Need?**

    The size of the sample depends on the type of research design being used; the desired level of confidence in the results; the amount of accuracy wanted; and the characteristics of the population of interest. Sample size has little to do with the size of the population, however.

    Random sampling procedures are based on probability theory; this is why they are also called probability sampling methods. Say we are interested in knowing what is the average monthly income of all the full-time students at our university. There are 5 full-time students each with a different monthly income as follows: $500; $650; $400; $700; $600. This is our population of students. Say we take a simple random sample of 2 students and figure the average for the sample.

    It is entirely possible that we could take a simple random sample 2 students from the 5 students above and get an average as low as $450 per month. It is equally possible that we could take a different simple random sample of 2 students and get an average as high as $675 per month. Try it with the following figures. There are 10 possible samples of two students:

$500 + $650 = $575   
$500 + $400 = $450   
$500 + $700 = $600   
$500 + $600 = $550   
6500 + $400 = $525   
$650 + $700 = $675   
$650 + $600 = $625   
$400 + $700 = $550   
$400 + $600 = $500   
$700 + $600 = $650

    We know from probability theory that if we took all possible combinations of samples of 2 full-time students from our population of 5, found the average monthly wage for all possible samples, and took the average of all those averages, we would find the exact typical monthly income of all 5 students.

    The average monthly wage of the 5 students in the population = $570.   
    The average of the 10 samples of 2 students each = $570.

    Now in this example, of course it would be easier to just find the average monthly wage for all five students in the population. However, we can apply this same principle to much larger populations, where it would be nearly impossible to measure every unit in the population.

    Say we wanted to find the average monthly wage of all 10,000 full-time students at our university. We can take a simple random sample of 150 students, find the average monthly wage for the 150 students in the sample, and then use that number (a sample statistic) to estimate the average monthly wage for the entire population of students (a population parameter).

    We know from probability theory that if we took a very large number of simple random samples of 150 students from our student population, and found the average monthly wage for each sample, that those averages would tend to distribute themselves in the pattern of a "bell-shaped" curve, also called "the normal curve." That curve has well-established properties.

    For example, approximately 68% of the sample averages would fall within plus or minus one standard deviation of the true population average. We also know that approximately 95% of the sample averages would fall within plus or minus two standard deviations of the true population average. And finally, we know that approximately 99% of the sample averages would fall within plus or minus three standard deviations of the true population average.

    Using these established principles, we do not have to take repeated simple random samples (fortunately!). Instead, we can use these principles to estimate how well our sample statistic estimates the population parameter. We can also use these principles to select an adequate sample size for our research.

    Say we want to know what proportion of the support of students at our university support the death penalty. To calculate sample size, we must make four decisions:

    First, are we doing a true experimental design (e.g., control-group, pretest-posttest design) or a non-experimental design (e.g., a cross-sectional survey)? The former can use smaller sample sizes, while the latter require larger sample sizes. In this case we are doing a survey.

    Second, how sure do we want to be that we could get the same results if we did the study multiple times? Do we want to be 50% sure, 90% sure, 95% sure, or 99% sure? This is called the confidence level. The more sure we want to be, the larger the sample size needs to be. In this case, we want a confidence level of 95%.

    Third, how accurate do we want to be at estimating the population parameter? Will a margin of error of (plus or minus) 5% be acceptable, or 4%, 3%, 2%, or 1%? This is also called the confidence interval. In this case, we want an accuracy of plus or minus 4%.   This means that if we find that 66% of the students oppose the death penalty, we really mean that we have found that 66% plus or minus 4% oppose the death penalty.

    Fourth, how is the population distributed on the variable of interest? That is, in a yes/no situation, how many do we think will say yes? How many will say no? The most conservative way to approach this is to guess that the population is split 50/50 on the question. In this case we guess that 50% of the students will support the death penalty, and 50% will oppose it.

If we are doing a survey of a population, and are not interested in sub-samples within the population, and will accept a 95% confidence level, and a 4% margin of error, and assume a probability of .5 on the variable (.5 will say yes), then the formula for sample size is as follows:

the square root of     =     square root of     x     confidence level divided   
the sample size                 (p) x (1-p)                 by the margin of error

Solving for the sample size, we have

the square root of sample size = [the square root of (.5) x (1-.5)] x 1.96/.05 =   
the square root of sample size = the square root of .25 x 1.96/.05 =   
the square root of sample size = the square root of (.5) x 49 =   
the square root of sample size = the square root of 24.5   
Squaring both sides, we have   
the sample size = 24.5 squared =   
the sample size = 600.25 (round off to 600)

    As the margin of error decreases, the sample size will need to increase (and vice versa). If we wanted to change the margin of error to plus or minus 3%, (keeping the confidence level at 95%), the required sample size increases to 1,067. If we could afford to use a margin of error of plus or minus 5%, the sample size would decrease to 384.

    Similarly, if the confidence level increases, the sample size will need to increase. If we increase the confidence level to 99%, the sample size increases to 1,036 (with the margin of error remaining at 4%). If the confidence level decreases to 90%, the sample size decreases to 413.

    If you have a fixed sample size, you can increase the confidence level and decrease the accuracy, or you can increase the accuracy and decrease the confidence level, but you cannot do both.

    As the variability in the population on the variable of interest increases, the sample size increases. A probability of 50/50 demonstrates the greatest variability in the population. If the variability decreases to 60/40, or 70/30, then a smaller sample size will result.

The following table summarizes the calculations for sample sizes for survey research, assuming a probability of 50/50 on a dichotomous question, and no sub-populations. 

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy (+/-)**  **(Margin of error)** | **Confidence Level** | | |
| **90%** | **95%** | **99%** |
| **1** | 6,765 | 9,604 | 16,576 |
| 2 | 1,691 | 2,401 | 4,144 |
| 3 | 752 | 1,067 | 1,848 |
| 4 | 413 | 600 | 1,036 |
| 5 | 271 | 384 | 663 |
| 10 | 68 | 96 | 166 |
| 20 | 17 | 24 | 41 |

If the researcher wants to study sub-populations as well as the whole population, then larger sample sizes will be needed. In addition, if more than one variable is being studied at the same time, then the rule of thumb is to have a total of at least 10 cases per variable.

    If the research is to be a controlled experiment, then smaller sample sizes can be used. However, it is recommended to use samples of no smaller than 30 for each group in the experiment (e.g., experimental and control groups). Many common statistics are based on sample sizes of a minimum of 30; for sample sizes of less than 30, other special statistics must be used. 

**Sample Quality**

    Sampling error arises from two principal sources: random error, and non-random error. Random error results from taking a sample from a population, instead of measuring the entire population. It is predictable, using probability theory. It is the reason that sample statistics only provide estimates of population parameters, but the amount of random error is known.

    Non-random error results from bias being introduced into the sample from some flaw in the design or implementation of the sample. For example, using a telephone book as the sampling frame for all the residents of a city will result in some bias, because some people are not listed in the directory or do not have telephones. People who refuse to take part in a study (which is their right) also may introduce bias into the sample. Some people may provide erroneous information, which also biases the results. Finally, mistakes in computing the required sample size, in identifying the actual units to be included in the sample, or other errors can introduce bias into the sample.

    To assess whether an adequate sample was used in a piece of research, ask the following questions:

**Size**--was the size adequate for the purpose of the study, especially if there were many sub-groups included in the analysis, or many variables used simultaneously?

**Representativeness**--was the sample selected randomly from the population, using probability theory?  Was the sampling frame adequate?

**Implementation**--was the sampling plan carried out carefully, was it adequately supervised, was there some quality control plan, did it result in a good response rate? 

**Exercises**

    For each of the following, suggest an appropriate sampling design and sample size.

1. A public health official wants to estimate the number of babies who are being born infected with HIV.   
    
2. The City Manager wants to know, by next Monday, the extent of pothole damage in the city caused by the latest El Nino storms.   
    
3. The Chief County Librarian wants to know what the patrons think of the county's branch libraries, and whether they would be willing to support a tax increase for the libraries.   
    
4. The economic development analyst wants to know what reason people give for not attending the city's job readiness training, even though they are eligible to do so.   
    
5. Why do some residents in nursing homes get broken hips when they fall, while other residents fall but do not break their hips, and other residents do not fall at all?   
    
6. Do older Hispanic women who live in neighborhoods with higher proportions of Hispanic residents get fewer preventive health care checkups than older Hispanic women who live in neighborhoods with lower proportions of Hispanic residents?