Different types of sampling method education essay

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of individuals intended to yield some knowledge about the population of concern, especially for the purposes of making predictions based on statistical inference. Sampling is an important aspect of data collection.

Researchers rarely survey the entire population for two reasons (Adèr, Mellenbergh, & Hand, 2008): the cost is too high, and the population is dynamic in that the individuals making up the population may change over time. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller is possible to ensure homogeneity and to improve the accuracy and quality of the data.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, survey weights can be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice. In business and medical research, sampling is widely used for gathering information about a population

Process

The sampling process comprises several stages:

Defining the population of concern

Specifying a sampling frame, a set of items or events possible to measure

Specifying a sampling method for selecting items or events from the frame

Determining the sample size

Implementing the sampling plan

Sampling and data collecting

Reviewing the sampling process

population definition

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes that which defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality to be released to the customer, or should be sentenced for scrap or rework due to poor quality. In this case, the batch is the population.

Although the population of interest often consists of physical objects, sometimes we need to sample over time, space, or some combination of these dimensions. For instance, an investigation of supermarket staffing could examine checkout line length at various times, or a study on endangered penguins might aim to understand their usage of various

hunting grounds over time. For the time dimension, the focus may be on periods or discrete occasions.

In other cases, our ‘ population’ may be even less tangible. For example, Joseph Jagger studied the behaviour of roulette wheels at a casino in Monte Carlo, and used this to identify a biased wheel. In this case, the ‘ population’ Jagger wanted to investigate was the overall behaviour of the wheel (i. e. the probability distribution of its results over infinitely many trials), while his ‘ sample’ was formed from observed results from that wheel. Similar considerations arise when taking repeated measurements of some physical characteristic such as the electrical conductivity of copper.

This situation often arises when we seek knowledge about the cause system of which the observed population is an outcome. In such cases, sampling theory may treat the observed population as a sample from a larger ‘ superpopulation’. For example, a researcher might study the success rate of a new ‘ quit smoking’ program on a test group of 100 patients, in order to predict the effects of the program if it were made available nationwide. Here the superpopulation is “ everybody in the country, given access to this treatment” – a group which does not yet exist, since the program isn’t yet available to all.

Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc (see below). Sometimes they may be entirely separate – for instance, we might study rats in order to get a better understanding of

human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Time spent in making the sampled population and population of concern precise is often well spent, because it raises many issues, ambiguities and questions that would otherwise have been overlooked at this stage.

Sampling frame

In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats. Where voting is not compulsory, there is no way to identify which people will actually vote at a forthcoming election (in advance of the election).

These imprecise populations are not amenable to sampling in any of the ways below and to which we could apply statistical theory.

As a remedy, we seek a sampling frame which has the property that we can identify every single element and include any in our sample. The most straightforward type of frame is a list of elements of the population (preferably the entire population) with appropriate contact information. For example, in an opinion poll, possible sampling frames include:

Electoral register

Telephone directory

Not all frames explicitly list population elements. For example, a street map can be used as a frame for a door-to-door survey; although it doesn’t show individual houses, we can select streets from the map and then visit all houses on those streets. (One advantage of such a frame is that it would include people who have recently moved and are not yet on the list frames discussed above.)

The sampling frame must be representative of the population and this is a question outside the scope of statistical theory demanding the judgment of experts in the particular subject matter being studied. All the above frames omit some people who will vote at the next election and contain some people who will not; some frames will contain multiple records for the same person. People not in the frame have no prospect of being sampled. Statistical theory tells us about the uncertainties in extrapolating from a sample to the frame. In extrapolating from frame to population, its role is motivational and suggestive.

To the scientist, however, representative sampling is the only justified procedure for choosing individual objects for use as the basis of generalization, and is therefore usually the only acceptable basis for ascertaining truth.

-Andrew A. Marino

It is important to understand this difference to steer clear of confusing prescriptions found in many web pages.

In defining the frame, practical, economic, ethical, and technical issues need to be addressed. The need to obtain timely results may prevent extending the frame far into the future.

The difficulties can be extreme when the population and frame are disjoint. This is a particular problem in forecasting where inferences about the future are made from historical data. In fact, in 1703, when Jacob Bernoulli proposed to Gottfried Leibniz the possibility of using historical mortality data to predict the probability of early death of a living man, Gottfried Leibniz recognized the problem in replying:

Nature has established patterns originating in the return of events but only for the most part. New illnesses flood the human race, so that no matter how many experiments you have done on corpses, you have not thereby imposed a limit on the nature of events so that in the future they could not vary.

-Gottfried Leibniz

Kish posited four basic problems of sampling frames:

Missing elements: Some members of the population are not included in the frame.

Foreign elements: The non-members of the population are included in the frame.

Duplicate entries: A member of the population is surveyed more than once.

Groups or clusters: The frame lists clusters instead of individuals.

•

Probability and nonprobability sampling

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (In effect, the person who is selected from that household is taken as representing the person who isn't selected.)

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an ' equal probability of selection'

(EPS) design. Such designs are also referred to as ‘ self-weighting’ because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

Every element has a known nonzero probability of being sampled and

involves random selection at some point.

Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as ‘ out of coverage’/’undercovered’), or where the probability of selection can’t be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions place limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e. g. an unemployed person who spends most of their time

at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it’s not practical to calculate these probabilities.

Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling. In addition, nonresponse effects may turn any probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element’s probability of being sampled.

Sampling methods

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

Nature and quality of the frame

Availability of auxiliary information about units on the frame

Accuracy requirements, and the need to measure accuracy

Whether detailed analysis of the sample is expected

Cost/operational concerns

Simple random sampling

In a simple random sample (‘ SRS’) of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned.

Furthermore, any given pair of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimises bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn’t reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to overrepresent one sex and underrepresent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.

SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide subsamples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards. In this case, k=(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every 10th name from the telephone directory (an ‘ every 10th’ sample, also referred to as ‘ sampling with a skip of 10’).

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can make it efficient, if the variable by which the list is ordered is correlated with the variable of interest. ‘ Every 10th’ sampling is especially useful for efficient sampling from databases.

Example: Suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e. g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end; by randomly selecting the start between #1 and #10, this bias is eliminated.)

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate than simple random sampling.

Example: Consider a street where the odd-numbered houses are all on the north (expensive) side of the road, and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible’ to get a representative sample; either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side.

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to quantify that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses – but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is not ‘ simple random sampling’ because different subsets of the same size have different selection probabilities – e. g. the set {4, 14, 24,..., 994} has a one-in-ten probability of selection, but the set {4, 13, 24, 34,...} has zero probability of selection.

Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate “ strata.” Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected[3]. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). It is important to note that even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group’s size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

A stratified sampling approach is most effective when three conditions are met

Variability within strata are minimized

Variability between strata are maximized

The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods

Focuses on important subpopulations and ignores irrelevant ones.

Allows use of different sampling techniques for different subpopulations.

Improves the accuracy/efficiency of estimation.

Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

Requires selection of relevant stratification variables which can be difficult.

Is not useful when there are no homogeneous subgroups.

Can be expensive to implement.

Poststratification

Stratification is sometimes introduced after the sampling phase in a process called “ poststratification”. This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample’s estimates.

Oversampling

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling[4], the data are stratified on the target and a sample is taken from each strata so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

Probability proportional to size sampling

In some cases the sample designer has access to an “ auxiliary variable” or “ size measure”, believed to be correlated to the variable of interest, for each element in the population. This data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size (‘ PPS’) sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawbacks of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i. e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample size by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available – for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel’s number of rooms as an auxiliary variable. In some cases, an older measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates.

Cluster sampling

Sometimes it is cheaper to ‘ cluster’ the sample in some way e. g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense ‘ clustered’ in time – although this is rarely taken into account in the analysis.)

Cluster sampling is an example of ‘ two-stage sampling’ or ‘ multistage sampling’: in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.

This can reduce travel and other administrative costs. It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate.

Multistage sampling Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, It is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

•

Matched random sampling

A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups.

The procedure for matched random sampling can be briefed with the following contexts,

Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.

Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures. Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

Quota sampling

In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select

the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years

Convenience sampling

Convenience sampling (sometimes known as grab or opportunity sampling) is a type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a sample population selected because it is readily available and convenient. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer was to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing. Several important considerations for researchers using convenience samples include:

Are there controls within the research design or experiment which can serve to lessen the impact of a non-random, convenience sample whereby ensuring the results will be more representative of the population?

Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population?

Is the question being asked by the research one that can adequately be answered using a convenience sample?

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

Line-intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a “ transect”, intersects the element.

Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a “ wave”. This sampling methodology is often chosen for large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables

from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel sample data, including MANOVA, growth curves, and structural equation modeling with lagged effects. For a more thorough look at analytical techniques for panel data, see Johnson (1995).