

Sampling and Experiments

8.1 Sampling: In statistics, quality assurance, & survey methodology, **sampling** is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, weights can be applied to the data to adjust for the sample design, particularly stratified sampling. 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.

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
- Data which can be selected

Population definition[edit]

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 what 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[edit]

Main article: 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^[1]^[2]^[3]^[4]. 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 an electoral register and a telephone directory.

Probability and nonprobability sampling[edit]

A **probability sampling** 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.

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:

1. Every element has a known nonzero probability of being sampled and
2. 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 give rise to exclusion bias, placing 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.

Non-probability sampling methods include accidental sampling, quota sampling and purposive sampling. In addition, non-response effects may

turn *any* probability design into a non-probability design if the characteristics of non-response are not well understood, since non-response 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 [edit]

In a simple random sample (SRS) of a given size, all such subsets of the frame are given an equal probability. 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 over represent one sex and under represent the other. (Systematic and stratified techniques), 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" addresses this weakness of SRS.

Systematic sampling

Systematic sampling relies on arranging the study 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 k th element from then onwards. In this case, $k = (\text{population size}/\text{sample size})$. 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 k th 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.

For example, suppose we wish to sample people from a long street that starts in a poor area (house No. 1) and ends in an expensive district (house No. 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.

For 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 neighboring houses - but because this method never selects two neighboring 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. 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). 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

1. Variability within strata are minimized
2. Variability between strata are maximized
3. The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods

1. Focuses on important subpopulations and ignores irrelevant ones.
2. Allows use of different sampling techniques for different subpopulations.
3. Improves the accuracy/efficiency of estimation.
4. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages

1. Requires selection of relevant stratification variables which can be difficult.
2. Is not useful when there are no homogeneous subgroups.
3. 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, the data are stratified on the target and a sample is taken from each stratum 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. These 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 drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections.

Systematic sampling theory can be used to create a probability proportionate to size sample. This is done by treating each count within the size variable as a single sampling unit. Samples are then identified by selecting at even intervals among these counts within the size variable. This method is sometimes called PPS-sequential or monetary unit sampling in the case of audits or forensic sampling.

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 more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.) For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

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. In the example above, the sample only requires a block-level city map for initial selections, and then a household-level map of the 100 selected blocks, rather than a household-level map of the whole city.

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. For this reason, cluster sampling requires a larger sample than SRS to achieve the same level of accuracy - but cost savings from clustering might still make this a cheaper option.

Cluster sampling is commonly implemented as multistage sampling. This 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 subsamples of preceding random samples.

Multistage sampling can substantially reduce sampling costs, where the complete population list would need to be constructed (before other sampling methods could be applied). By eliminating the work involved in describing clusters that are not selected, multistage sampling can reduce the large costs associated with traditional cluster sampling. However, each sample may not be a full representative of the whole population.

Quota sampling

In **quota sampling**, the population is first segmented into mutually exclusive subgroups, 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 several years.

Accidental sampling

Accidental sampling (sometimes known as **grab, convenience or opportunity sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. 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 were 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 were 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:

1. Are there controls within the research design or experiment which can serve to lessen the impact of a non-random convenience sample, thereby ensuring the results will be more representative of the population?
2. 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?
3. 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. Some variants of snowball sampling, such as respondent driven sampling, allow calculation of selection probabilities and are probability sampling methods under certain conditions.

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 (potentially the same) information several times over a period of time. Therefore, each participant is interviewed at two or more time points; each period of data collection is called a "wave". The method was developed by sociologist Paul Lazarsfeld in 1938 as a means of studying political campaigns. This longitudinal sampling-method allows estimates of changes in the population, for example with regard to 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 to help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel data, including MANOVA, growth curves, and structural equation modeling with lagged effects.

Replacement of selected units

Sampling schemes may be *without replacement* ('WOR'—no element can be selected more than once in the same sample) or *with replacement* ('WR'—an element may appear multiple times in the one sample). For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g., if we eat the fish), this becomes a WOR design.

Sample size

Formulas, tables, and power function charts are well known approaches to determine sample size.

Steps for using sample size tables

1. Postulate the effect size of interest, α , and β .
2. Check sample size table^[9]
 1. Select the table corresponding to the selected α
 2. Locate the row corresponding to the desired power
 3. Locate the column corresponding to the estimated effect size.

4. The intersection of the column and row is the minimum sample size required.

Sampling and data collection

Good data collection involves:

- Following the defined sampling process
- Keeping the data in time order
- Noting comments and other contextual events
- Recording non-responses

Errors in sample surveys

Survey results are typically subject to some error. Total errors can be classified into sampling errors and non-sampling errors. The term "error" here includes systematic biases as well as random errors.

Sampling errors and biases

Sampling errors and biases are induced by the sample design. They include:

1. **Selection bias:** When the true selection probabilities differ from those assumed in calculating the results.
2. **Random sampling error:** Random variation in the results due to the elements in the sample being selected at random.

Non-sampling error

Non-sampling errors are other errors which can impact the final survey estimates, caused by problems in data collection, processing, or sample design. They include:

1. **Overcoverage:** Inclusion of data from outside of the population.
2. **Undercoverage:** Sampling frame does not include elements in the population.
3. **Measurement error:** e.g. when respondents misunderstand a question, or find it difficult to answer.
4. **Processing error:** Mistakes in data coding.
5. **Non-response:** Failure to obtain complete data from all selected individuals.

After sampling, a review should be held of the exact process followed in sampling, rather than that intended, in order to study any effects that any divergences might have on subsequent analysis. A particular problem is that of *non-response*.

Two major types of non-response exist: unit non-response (referring to lack of completion of any part of the survey) and item non-response (submission or

participation in survey but failing to complete one or more components/questions of the survey). In survey sampling, many of the individuals identified as part of the sample may be unwilling to participate, not have the time to participate (opportunity cost), or survey administrators may not have been able to contact them. In this case, there is a risk of differences, between respondents and nonrespondents, leading to biased estimates of population parameters. This is often addressed by improving survey design, offering incentives, and conducting follow-up studies which make a repeated attempt to contact the unresponsive and to characterize their similarities and differences with the rest of the frame. The effects can also be mitigated by weighting the data when population benchmarks are available or by imputing data based on answers to other questions.

Non-response is particularly a problem in internet sampling. Reasons for this problem include improperly designed surveys, over-surveying (or survey fatigue), and the fact that potential participants hold multiple e-mail addresses, which they don't use anymore or don't check regularly.

Survey weights

In many situations the sample fraction may be varied by stratum and data will have to be weighted to correctly represent the population. Thus for example, a simple random sample of individuals in the United Kingdom might include some in remote Scottish islands who would be inordinately expensive to sample. A cheaper method would be to use a stratified sample with urban and rural strata. The rural sample could be under-represented in the sample, but weighted up appropriately in the analysis to compensate.

More generally, data should usually be weighted if the sample design does not give each individual an equal chance of being selected. For instance, when households have equal selection probabilities but one person is interviewed from within each household, this gives people from large households a smaller chance of being interviewed. This can be accounted for using survey weights. Similarly, households with more than one telephone line have a greater chance of being selected in a random digit dialing sample, and weights can adjust for this.

Weights can also serve other purposes, such as helping to correct for non-response.

Methods of producing random samples

- Random number table
- Mathematical algorithms for pseudo-random number generators
- Physical randomization devices such as coins, playing cards or sophisticated devices such as ERNIE

History

Random sampling by using lots is an old idea, mentioned several times in the Bible. In 1786 Pierre Simon Laplace estimated the population of France by using a sample, along with ratio estimator. He also computed probabilistic estimates of the error. These were not expressed as modern confidence intervals but as the sample size that would be needed to achieve a particular upper bound on the sampling error with probability 1000/1001. His estimates used Bayes' theorem with a uniform prior probability and assumed that his sample was random. Alexander Ivanovich Chuprov introduced sample surveys to Imperial Russia in the 1870s.

In the USA the 1936 *Literary Digest* prediction of a Republican win in the presidential election went badly awry, due to severe bias. More than two million people responded to the study with their names obtained through magazine subscription lists and telephone directories. It was not appreciated that these lists were heavily biased towards Republicans and the resulting sample, though very large, was deeply flawed.