Stratified sampling

From Wikipedia, the free encyclopedia

In statistics, **stratified sampling** is a method of sampling from a population.

In statistical surveys, when subpopulations within an overall population vary, it is advantageous to sample each subpopulation (stratum) independently. **Stratification** is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Then simple random sampling or systematic sampling is applied within each stratum. This often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population.

In computational statistics, stratified sampling is a method of variance reduction when Monte Carlo methods are used to estimate population statistics from a known population.

Contents

Example

Assume that we need to estimate average number of votes for each candidate in an election. Assume that country has 3 towns: Town A has 1 million factory workers, Town B has 2 million office workers and Town C has 3 million retirees. We can choose to get a random sample of size 60 over entire population but there is some chance that the random sample turns out to be not well balanced across these towns and hence is biased causing a significant error in estimation. Instead if we choose to take a random sample of 10, 20 and 30 from Town A, B and C respectively then we can produce a smaller error in estimation for the same total size of sample.

Stratified sampling strategies

1. Proportionate allocation uses a sampling fraction in each of the strata that is proportional to that of the total population. For instance, if the population consists of *X* total individuals, *m* of which are male and *f* female (and where *m* + *f* = *X*), then the relative size of the two samples (*x1 = m/X* males, *x2 = f/X* females) should reflect this proportion.
2. Optimum allocation (or Disproportionate allocation) - Each stratum is proportionate to the standard deviation of the distribution of the variable. Larger samples are taken in the strata with the greatest variability to generate the least possible sampling variance.

Stratified sampling ensures that at least one observation is picked from each of the strata, even if probability of it being selected is close to 0. Hence the statistical properties of the population may not be preserved if there are thin strata. A rule of thumb that is used to ensure this is that the population should consist of no more than six strata, but depending on special cases the rule can change - for example if there are 100 strata each with 1 million observations, it is perfectly fine to do a 10% stratified sampling on them.

A real-world example of using stratified sampling would be for a political survey. If the respondents needed to reflect the diversity of the population, the researcher would specifically seek to include participants of various minority groups such as race or religion, based on their proportionality to the total population as mentioned above. A stratified survey could thus claim to be more representative of the population than a survey of simple random sampling or systematic sampling.

Advantages

The reasons to use stratified sampling rather than simple random sampling include[1]

1. If measurements within strata have lower standard deviation, stratification gives smaller error in estimation.
2. For many applications, measurements become more manageable and/or cheaper when the population is grouped into strata.
3. It is often desirable to have estimates of population parameters for groups within the population.

If the population density varies greatly within a region, stratified sampling will ensure that estimates can be made with equal accuracy in different parts of the region, and that comparisons of sub-regions can be made with equal statistical power. For example, in Ontario a survey taken throughout the province might use a larger sampling fraction in the less populated north, since the disparity in population between north and south is so great that a sampling fraction based on the provincial sample as a whole might result in the collection of only a handful of data from the north.

Randomized stratification can also be used to improve population representativeness in a study.

Disadvantages

Stratified sampling is not useful when the population cannot be exhaustively partitioned into disjoint subgroups. It would be a misapplication of the technique to make subgroups' sample sizes proportional to the amount of data available from the subgroups, rather than scaling sample sizes to subgroup sizes (or to their variances, if known to vary significantly e.g. by means of an F Test). Data representing each subgroup are taken to be of equal importance if suspected variation among them warrants stratified sampling. If subgroups' variances differ significantly and the data need to be stratified by variance, then there is no way to make the subgroup sample sizes proportional (at the same time) to the subgroups' sizes within the total population. For an efficient way to partition sampling resources among groups that vary in their means, their variances, and their costs, see "optimum allocation". The problem of stratified sampling in the case of unknown class priors (ratio of subpopulations in the entire population) can have deleterious effect on the performance of any analysis on the dataset, e.g. classification.[2] In that regard, minimax sampling ratio can be used to make the dataset robust with respect to uncertainty in the underlying data generating process.[2]

Mean and Variance

The mean and variance of stratified random sampling is given by,[1]

$$\mu_s = \frac{1}{N} \sum_{h=1}^{L} N_h \mu_h$$

$$\sigma_s^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$$

where,

$N =$ Size of entire population, should equal to sum of all stratum sizes

$N_h =$ Size of each stratum

$n_h =$ Number of observations in each stratum

$L =$ Count of strata

$\sigma_h =$ sample standard deviation of stratum h

$\mu_h =$ sample mean of stratum h

Strata Size Calculation

In general the size of the sample in each stratum is taken in proportion to the size of the stratum. This is called proportional allocation. Suppose that in a company there are the following staff:[3]

- male, full-time: 90
- male, part-time: 18
- female, full-time: 9
- female, part-time: 63
- Total: 180

and we are asked to take a sample of 40 staff, stratified according to the above categories.

The first step is to find the total number of staff (180) and calculate the percentage in each group.

- % male, full-time = 90 ÷ 180 = 50%
- % male, part-time = 18 ÷ 180 = 10%
- % female, full-time = 9 ÷ 180 = 5%
- % female, part-time = 63 ÷ 180 = 35%

This tells us that of our sample of 40,

- 50% should be male, full-time.
- 10% should be male, part-time.
- 5% should be female, full-time.
- 35% should be female, part-time.

- 50% of 40 is 20.
- 10% of 40 is 4.
- 5% of 40 is 2.
- 35% of 40 is 14.

Another easy way without having to calculate the percentage is to multiply each group size by the sample size and divide by the total population size (size of entire staff):

- male, full-time = 90 × (40 ÷ 180) = 20
- male, part-time = 18 × (40 ÷ 180) = 4
- female, full-time = 9 × (40 ÷ 180) = 2
- female, part-time = 63 × (40 ÷ 180) = 14

See also

- Opinion Poll
- Statistical benchmarking
- Stratified sample size
- Stratification (clinical trials)

References

1. "6.1 How to Use Stratified Sampling | STAT 506". *onlinecourses.science.psu.edu*. Retrieved 2015-07-23.
2. Shahrokh Esfahani, Mohammad; Dougherty, Edward R. (2014). "Effect of separate sampling on classification accuracy". *Bioinformatics* **30** (2): 242–250. doi:10.1093/bioinformatics/btt662.
3. Hunt, Neville; Tyrrell, Sidney (2001). "Stratified Sampling". *Webpage at Coventry University*. Retrieved 12 July 2012.

Further reading

- Särndal, Carl-Erik; et al. (2003). "Stratified Sampling". *Model Assisted Survey Sampling*. New York: Springer. pp. 100–109. ISBN 0-387-40620-4.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Stratified_sampling&oldid=700608199"

Categories: Sampling (statistics) | Sampling techniques | Statistical terminology | Variance reduction

- This page was last modified on 19 January 2016, at 15:18.

- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.