Wiley StatsRef:
Statistics Reference Online

Systematic Sampling

By *Graham Kalton*

Keywords: *simple random sampling, stratification, probability proportional to size sampling, collapsed stratum technique, successive differences*

Abstract: Systematic sampling is a simple and flexible way of selecting a probability sample from a finite population. It was introduced in the early days of probability sampling in survey research and it remains in widespread use today. A systematic sample is obtained by selecting a random start between 1 and *k* from a list of the population and then taking every *k*th element thereafter. The ordering of the list determines the precision of the survey estimates; a random order produces a sample that has the same precision as a simple random sample, whereas an order that is monotonically associated with the variables under study gives greater precision than a simple random sample. A single-stage systematic sample is equivalent to the selection of a single cluster and, hence, variance estimation is not possible without a model assumption. Systematic sampling is widely used to select units with unequal probabilities without replacement, as, for instance, with probability proportional to size sampling in multistage surveys. In such cases, the collapsed stratum technique is frequently used for variance estimation. Systematic sampling is often applied in two dimensions in spatial surveys of geographical areas.

Systematic sampling has been widely used in many diverse applications since the early days of the use of probability sampling in survey research. When the elements of the survey population are listed in some order, the method essentially consists of selecting every *k*th element from the list, starting with a random start. An early example was the pioneering survey of living conditions of working class families in the town of Reading in England in 1912. For that survey, a sample of 1 in 10 buildings (later reduced to 1 in 20 buildings) was selected systematically from a local directory of buildings listed in alphabetical order of the streets[1,2]. Among other examples, Stephan[3] reports an even earlier application of systematic sampling for a survey of Norwegian workers conducted by A.N. Kaier in 1895.

In recognition of the practical importance of systematic sampling, a considerable amount of research was carried out around the middle of the past century on the theoretical properties of the technique and its properties relative to simple random sampling and stratified sampling. Some key papers include those by Cochran[4], Madow and Madow[5], Madow[6], Madow[7,8], and Yates[9]. Buckland[10] provides a review of this early research.

Westat, Rockville, MD, USA

Update based on original article by Graham Kalton, Wiley StatsRef: Statistics Reference Online © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat03380.pub2

Simple random sampling and stratified sampling require a pre-existing list of known population size from which to draw the sample. This requirement does not apply with systematic sampling. Provided that a reasonable estimate of the population size can be made, the sampling interval k can be determined, a random start can be drawn, and the sampling interval can be applied successively until all the population is covered. This feature is important when the population is identified on a flow basis as, for instance, when sampling visitors entering or leaving a museum or for exit polls, sampling electors as they leave the polling booths[11]. With such applications, the sample size achieved is not fixed but depends on how close the estimate of the population size is to the actual size.

To introduce some notation, suppose that a sample of n elements is required from a population list of N elements and assume for now that $k = N/n$ is an integer. A systematic sample is obtained by taking every kth element on the list; k is the *sampling interval*. As a rule, the first sampled element is determined by the selection of a random number between 1 and k, say r. The selected sample then comprises the rth, $[r+k]$th, $[r+2k]$th, … , and $[r+(n-1)k]$th elements on the list. The use of a random start gives every population element the same selection probability $1/k$. The joint selection probability for the ith and jth population elements is $1/k$ if $i = j + mk$, where m is an integer, and 0 otherwise.

In practice, the sampling interval k is seldom an integer. Noninteger values of k may be handled in several ways. One is to round k to the nearest integer and apply the preceding procedure. The resulting sample size will differ from the initial choice, but in many cases, this will be acceptable. A second way is to round k down to an integer, to select a random start throughout the whole population, and then select $(n-1)$ additional elements by successively adding the rounded-down k to the random start; the list is treated as circular with the last element being followed by the first. A third way is to randomly remove a number (t) of elements from the population so that $(N-t)/n$ is an integer, after which the procedure previously described can be applied. A fourth way is to employ a fractional sampling interval k, choosing a fractional random number, and successively adding k to it; the resulting numbers are then rounded down to identify the selected elements. See, for example, Kish[12].

Systematic sampling is equivalent to the selection of a single cluster in **cluster sampling**. Each of the k possible random starts defines a population cluster, with the jth cluster comprising the population elements j, $[j+k]$, $[j+2k]$, … , $[j+(n-1)k]$. The random start chooses one of the population clusters to be the sample. Let Y_{ij} denote the value of the variable Y for element i in cluster j, let $\overline{Y}_j$ denote the mean for the elements in cluster j, let $\overline{y} = \sum y_i/n$ be the sample mean (i.e., the mean of the selected cluster), and assume that $k = N/n$ is an integer. Then $\overline{y}$ is an unbiased estimator of the population mean $\overline{Y} = \sum\sum Y_{ij}/N = \sum \overline{Y}_j/k$, and its variance is

$$V(\overline{y}) = \sum (\overline{Y}_j - \overline{Y})^2/k$$

As systematic sampling selects only one cluster and replication of a sampling process is needed for unbiased variance estimation, $V(\overline{y})$ cannot be estimated from the sample without invoking some assumption about the formation of the clusters or equivalently about the order of the list. A variety of alternative variance estimators has been proposed based on different assumptions about the list order. Wolter[13,14] reports on a theoretical and empirical comparison of eight of these variance estimators. See also Valliant[15].

One frequently used assumption for variance estimation is that the list is randomly ordered with respect to the survey variables. Then systematic sampling is equivalent to **simple random sampling**, and $V(\overline{y})$ may be estimated by the simple random sampling formula $(N-n)s^2/(Nn)$, where $s^2 = \sum (y_i - \overline{y})^2/(n-1)$.

Often, the list is ordered in groups, as, for instance, when a firm's employees are listed in the departments (groups) in which they work. Systematic sampling from such a list ensures that each group is represented in the sample in approximately the same proportion as in the total population. Assuming a random ordering of elements within groups, the sample design closely resembles proportionate stratified sampling (*see* **Proportional Allocation**), the groups being the strata. $V(\overline{y})$ may then be estimated using the formula for proportionate stratified sampling. With proportionate stratification, the desired stratum sample sizes are

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat03380.pub2

generally fractional and, hence, have to be rounded to the nearest integer; when the sample sizes are very small, rounding may cause distortions. As systematic sampling avoids the need for this rounding, it is often used when a detailed stratification is required: The elements are listed by strata, with careful attention to the ordering of the strata, and then a systematic sample is taken throughout the list, yielding an "implicit stratification."

In many cases, the list is deliberately ordered by some factor that is continuous in nature, such as by geography or by time. In multistage samples (discussed later), the primary sampling units (PSUs) could be ordered, for example, by the percentage of poor households or the proportion of minority households (perhaps within explicit regional strata). In such cases, the design can be viewed as a one-per-stratum design with the set of elements in each sampling interval defining an implicit stratum. However, design-based variance estimation is then not possible[16].

A common solution for variance estimation with the one-per-stratum design is to use the collapsed stratum approach by treating the systematic sample as implicitly stratified into strata with two selections per stratum. The usual way to effect this outcome is to pair the first and second selections into the first implicit stratum, the third and fourth selections into the second implicit stratum, and so on, throughout the sample. Variances are then computed as for a proportionate design. This approach fails to capture the effect of the finer stratification within the implicit strata and, hence, overestimates the variances of the survey estimates. In fact, the expected value of an estimated variance from this collapsed stratum approach is larger than the expected value of the unbiased variance estimate that would have been obtained had the sample been explicitly stratified into strata with two selections made at random within those strata (see, e.g., Kish[12], Section 8.6B). This finding has led to some debate about the use of systematic sampling rather than a design that permits unbiased variance estimation, particularly for sampling the PSUs in a multistage design. It has also led to recent research on nonparametric ways of estimating variances with one-per-stratum designs[17].

When the basic collapsed stratum technique is applied to a systematic sample, the degrees of freedom for the variance estimator are $n/2$. When n is small, it may be beneficial to obtain more degrees of freedom using a successive difference method, pairing sampled units 1 and 2, 2 and 3, 3 and 4, and so on (Kish[12], p. 119; see also **Successive Differences**). Fay and Train[18] describe the application of a successive difference replication approach with the U.S. Current Population Survey, and Opsomer *et al*.[19] extend the approach to a two-phase design.

The methods of variance estimation already described require assumptions about the order of the population list because a systematic sample selects only a single cluster. The need for such assumptions can be avoided by selecting several clusters, that is, by taking several random starts. Thus, for instance, instead of a single systematic sample with an interval of k, c systematic samples could be selected with intervals of kc, starting with c different random starts from 1 to kc. Assuming that the population size is a multiple of kc, the variance of the overall sample mean from the c samples ($\bar{y}$) may then be estimated by

$$[1 - (1/k)]\sum(\bar{y}_\gamma - \bar{y})^2/\{c(c-1)\}$$

where $\bar{y}_\gamma$ is the mean of the γth subsample.

Another approach is to randomize the order of the list before sampling. In this case, the resultant systematic sample can be treated as a simple random sample for variance estimation. Yet another approach is to employ a partially systematic sample design in which one part of the sample is a systematic sample and the other part is a simple random sample[20].

A number of theoretical and empirical studies have been conducted to examine the efficiency of systematic sampling in specific situations. A simple theoretical model for the order of the list has the Y values following a linear trend. Under this model, the sample mean from a systematic sample is more precise than the mean from a simple random sample of the same size but less precise than the mean from a

DOI: 10.1002/9781118445112.stat03380.pub2

proportionate stratified sample in which one element is selected from each of n equal-sized strata (the first k population elements comprising the first stratum, the next k elements the next stratum, etc.) (Cochran[21], Chapter 8).

Several approaches have been proposed to improve the precision of systematic sampling in the case of a linear trend. One is to change the weights of the sampled elements that are lowest and highest in the order of the population list[9,22]. A second is to take a centrally located sample and instead of starting with a random number between 1 and k, select the middle element of the sampling interval. The sample mean of a centrally located systematic sample has a lower mean square error than that of a random start systematic sample when the population follows a monotonic trend[8]. However, a centrally located systematic sample is not a probability sample. A third approach, termed *balanced systematic sampling*[23], takes two balanced starts in the sampling interval $2k$ (with n even), the first being a random number (r) from 1 to k and the second being $(2k - r + 1)$. Another variant is to take one-half of the systematic sample working forward through the list and the other half working backward through the list from the end, using the same random start for both halves[24].

Systematic sampling fares badly if the survey variable followed a periodic variation in the population list and the sampling interval coincides with a multiple of the periodicity. When a periodicity is present, a sampling interval that is a multiple of the periodicity should be avoided. For example, in systematically sampling days for air pollution monitoring, Akland[25] points out the need to avoid sampling intervals that are multiples of seven because such sampling intervals would end up always sampling the same day of the week. In practice, regular periodic cycles are rarely encountered, and when they exist, they are generally readily identified and avoided.

Systematic sampling has been extended into two (and more) dimensions. This extension may be applied, for instance, in sampling geographical areas as in agricultural surveys. Consider a square field of n^2k^2 square unit areas of which n^2 are required for the sample. The field could be divided into n^2 subsquares of dimensions $(k \times k)$, with one unit area to be selected from each. The choice of two random numbers r and r' between 1 and k would fix the coordinates of the selected unit square in the top left-hand subsquare of the field. The remaining selections could then be determined by successively adding k to r and r'. This procedure, which results in the sampled unit squares having the same location in each subsquare, produces an *aligned sample*. A modification is to randomly choose different horizontal coordinates for the first row and different vertical coordinates for the first column, leading to an unaligned sample. Further extensions of this approach lead to lattice sampling. Gilbert[26] discusses the use of systematic sampling in one and two dimensions for environmental pollution monitoring. Flores *et al*.[27] discuss systematic sample designs for spatial surveys and Fewster[28] examines variance estimation for such designs.

The preceding discussion relates to applications of systematic sampling for sampling units with equal probability. It is also widely used for sampling units with unequal probability, such as sampling with **probability proportional to size (PPS)**. The procedure is best described by means of an example. Suppose that three units are to be selected from the following six units with probabilities proportional to their measures of size M_i:

Unit (i)	1	2	3	4	5	6	Total
M_i	10	3	14	12	9	18	66
Cumulative M_i	10	13	27	39	48	66	–

The cumulative totals of the size measures are calculated as in the last row; using these totals, unit 1 is associated with the numbers 1–10, unit 2 with the numbers 11–13, and so on. Dividing the overall

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat03380.pub2

total (66) by the number of units to be selected (3) gives the sampling interval of 22. A random start between 1 and 22, say 12, is chosen; adding 22 to the random start gives 34 and adding 22 again gives 56. The three selections are thus units 2, 4, and 6. Provided that the sizes of all the units are smaller than the sampling interval, no unit can be selected more than once. This systematic procedure provides a simple means of selecting units with unequal probabilities without replacement[29].

Systematic sampling features as a component of many sample designs. Chapters on the method are to be found in **survey sampling** texts such as Cochran[21], Hansen *et al*.[30], Kish[12], Konijn[31], Murthy[23], Sukhatme *et al*.[32], and Yates[33]. Other useful references are Bellhouse[34], Iachan[35], and Murthy and Rao[36].

Related Articles

Area Sampling; **Probability Proportional to Size (PPS) Sampling**; **Proportional Allocation**; **Survey Sampling**; **Systematic Sampling Methods**; **Partially Systematic Sampling**.

References

[1] Bowley, A.L. (1913) Working-class households in Reading. *J. R. Stat. Soc.*, **76**, 672–701.
[2] Bowley, A.L. and Burnett-Hurst, A.R. (1915) *Livelihood and Poverty*, G. Bell and Sons, London.
[3] Stephan, F.F. (1948) History of the uses of modern sampling procedures. *J. Am. Stat. Assoc.*, **43**, 12–39.
[4] Cochran, W.G. (1946) Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Stat.*, **17**, 164–177.
[5] Madow, W.G. and Madow, L.H. (1944) On the theory of systematic sampling, I. *Ann. Math. Stat.*, **15**, 1–24.
[6] Madow, L.H. (1946) Systematic sampling and its relation to other sampling designs. *J. Am. Stat. Assoc.*, **41**, 204–217.
[7] Madow, W.G. (1949) On the theory of systematic sampling, II. *Ann. Math. Stat.*, **20**, 333–354.
[8] Madow, W.G. (1953) On the theory of systematic sampling, III. Comparison of centered and random start systematic sampling. *Ann. Math. Stat.*, **24**, 101–106.
[9] Yates, F. (1948) Systematic sampling. *Philos. Trans. R. Soc. London Ser. A*, **241**, 345–377.
[10] Buckland, W.R. (1951) A review of the literature of systematic sampling. *J. R. Stat. Soc. B*, **13**, 208–215.
[11] Kalton, G. (1991) Sampling flows of mobile human populations. *Surv. Methodol.*, **17**, 183–194.
[12] Kish, L. (1965) *Survey Sampling*, John Wiley & Sons, Inc., New York.
[13] Wolter, K.M. (1984) An investigation of some estimators of variance for systematic sampling. *J. Am. Stat. Assoc.*, **79**, 781–790.
[14] Wolter, K.M. (1985) *Introduction to Variance Estimation*, Springer-Verlag, New York.
[15] Valliant, R. (1990) Comparisons of variance estimators in stratified random and systematic sampling. *J. Official Stat.*, **6**, 115–131.
[16] Fuller, W.A. (2009) *Sampling Statistics*, John Wiley & Sons, Inc., New York.
[17] Breidt, F.J., Opsomer, J.D., and Sachez-Borrego, I. (2016) Nonparametric variance estimation under fine stratification: an alternative to collapsed strata. *J. Am. Stat. Assoc.*, **111**, 822–833.
[18] Fay, R.E. and Train, G.F. (1995) *Aspects of Survey and Model-based Postcensal Estimation of Income and Poverty Characteristics for States and Counties*. Proceedings of the Section on Government Statistics, American Statistical Association, pp. 154–159.
[19] Opsomer, J.D., Breidt, F.J., White, M., and Li, Y. (2016) Successive difference replication variance estimation in two-phase sampling. *J. Surv. Stat. Methodol.*, **4**, 43–70.
[20] Zinger, A. (1964) Systematic sampling in forestry. *Biometrics*, **20**, 553–565.
[21] Cochran, W.G. (1977) *Sampling Techniques*, 3rd edn, John Wiley & Sons, Inc., New York.
[22] Bellhouse, D.R. and Rao, J.N.K. (1975) Systematic sampling in the presence of a trend. *Biometrika*, **62**, 694–697.
[23] Murthy, M.N. (1967) *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.
[24] Singh, D., Jindal, K.K., and Garg, J.N. (1968) On modified systematic sampling. *Biometrika*, **55**, 541–546.
[25] Akland, G.G. (1972) Design of sampling schedules. *J. Air Pollut. Control Assoc.*, **22**, 264–266.
[26] Gilbert, R.O. (1987) *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York.

DOI: 10.1002/9781118445112.stat03380.pub2

[27] Flores, L.A., Martinez, L.I., and Ferrer, C.M. (2003) Systematic sample design for the estimation of spatial means. *Environmetrics*, **14**, 45–61.
[28] Fewster, R.M. (2011) Variance estimation for systematic designs in spatial surveys. *Biometrics*, **67**, 1518–1531.
[29] Hartley, H.O. and Rao, J.N.K. (1962) Sampling with unequal probabilities and without replacement. *Ann. Math. Stat.*, **33**, 350–374.
[30] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953) *Sample Survey Methods and Theory*, John Wiley & Sons, Inc., New York, vols. 1 and 2.
[31] Konijn, H.S. (1973) *Statistical Theory of Sample Survey Design and Analysis*, North-Holland, Amsterdam.
[32] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984) *Sampling Theory of Surveys with Applications*, 3rd edn, Iowa State University Press, Ames.
[33] Yates, F. (1981) *Sampling Methods for Censuses and Surveys*, 4th edn, Griffin, London.
[34] Bellhouse, D.R. (1988) Systematic sampling, in *Handbook of Statistics*, vol. **6** (eds P.R. Krishnaiah and C.R. Rao), Elsevier, Amsterdam, pp. 125–145.
[35] Iachan, R. (1982) Systematic sampling: a critical review. *Int. Stat. Rev.*, **50**, 293–303.
[36] Murthy, M.N. and Rao, T.J. (1988) Systematic sampling with illustrative examples, in *Handbook of Statistics*, vol. **6** (eds P.R. Krishnaiah and C.R. Rao), North-Holland, New York, pp. 147–185.

Wiley StatsRef: Statistics Reference Online, © 2014–2017 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat03380.pub2