 International Journal of Epidemiology 2002;**31**:463–467

A nomogram for single-stage cluster-sample surveys in a community for estimation of a prevalence rate

Rajeev Kumar and Abhaya Indrayan

Background	Proper assessment of the magnitude of the problem is essential for devising adequate allocation of available resources and for developing future strategies to combat a disease. The cluster random sampling (CRS) technique is commonly used for rapid assessment of public health problems in developing countries. Our objective is to devise a nomogram that can instantly provide the number of clusters of specified size needed to estimate the prevalence rate of a disease in a community with given precision, ratio of design-effect to cluster size and confidence level. This would be applicable only to single-stage CRS.
Methods	We use a logarithmic transformation to linearize the relation between the number of clusters (C) on one side and design-effect (D), cluster size (B), precision (L), anticipated prevalence rate (P) and confidence level (α) on the other. By using this relation, we construct a nomogram using established methods.
Results	A nomogram is obtained that can be used to determine the number of clusters needed in a survey with the help of only a ruler when other parameters are known. This is a 6-in-1 figure as it gives the number of clusters C corresponding to any combination of α from among the popularly used 0.05, 0.10 and 0.20, and precision 10% of P or 20% of P. Using a very simple calculation, the number of clusters for the other values of α and L can also be obtained.
Conclusion	This nomogram can be a useful aid in instantly providing the number of clusters required to rapidly estimate the prevalence rate of a disease in a community when the ratio of design-effect to cluster size, confidence level, and precision are specified. However, it is not applicable to intervention studies where interest mainly focuses on testing a hypothesis rather than estimation.

Keywords	Cluster random sampling, nomogram, number of clusters, rate of homogeneity, precision
Accepted	6 November 2001

Proper assessment of magnitude is essential to monitor the status of a disease in an area. This assessment is also required for adequate allocation of available resources and for developing strategies to combat the problem.

Sampling technique plays a vital role in providing an adequate estimate of the magnitude of a disease. Besides reliability, cost and time factors, as well as administrative and available resource barriers should be taken into consideration when choosing a technique. Large numbers of sampling techniques are available in the literature,[1,2] such as simple random sampling (SRS), stratified random sampling, cluster random sampling (CRS) and multistage random sampling. Their suitability depends upon the situation to which they are applied. Cluster random sampling is commonly used in developing countries as a rapid assessment method. The advantage is that a relatively large group of subjects is surveyed at one place. This helps to increase the size of the sample without a corresponding increase in the cost. One of the most popular uses is for evaluating immunization coverage.[3] Cluster sampling has also been used for rapid assessment of prevalence of age-related cataract blindness.[4]

In CRS, the population is divided into a number of clusters, generally of nearly equal size, and the desired number of clusters is randomly sampled.[2] Equal size is preferred not only for operational convenience but also in order to minimize bias and

Division of Biostatistics and Medical Informatics, University College of Medical Sciences, Delhi, India.

Correspondence: Rajeev Kumar, Division of Biostatistics and Medical Informatics, University College of Medical Sciences, Dilshad Garden, Delhi-110 095, India. E-mail: dbmi@ucms.ernet.in

for efficiency considerations. Important questions in this context are the optimal cluster size (B) and the number of clusters (C) needed in the sample to provide an adequate estimate of the prevalence rate. These parameters depend upon design-effect (D), precision required (L) (generally stated in terms of percentage of prevalence rate), size of critical region (α) and the anticipated prevalence rate (P). Among these, design-effect perhaps needs some explanation. We briefly explain this for CRS and also later on explain that it varies from situation to situation.

The World Health Organization (WHO) recommended 30 clusters of 7 children each to estimate immunization coverage in the Expanded Programme on Immunization (EPI) among young children within ±10 percentage points of the true proportion with 95% confidence. This assumes 50% coverage. The design-effect was assumed as 2. For practical and logistic reasons, the number of clusters (30) was considered adequate.[5] Limburg *et al*.[4] suggested that 37 clusters of size 30 each or 28 clusters of size 40 each may be appropriate for estimation of prevalence of cataract blindness in old age. The size of the cluster is generally decided by non-statistical considerations such as convenience in completing a survey of one or two clusters in a day by one team. Different precision levels and different confidence levels provide different number of clusters for a fixed cluster size. This calculation is done by a formula incorporating parameters L, α, D, P and B. The calculation is not simple. Recomputation becomes necessary if any of them is changed. To facilitate this process, we have devised a nomogram. The primary objective of this nomogram is to be able to immediately read the number of clusters required to be surveyed for given values of L, α, P and ratio of D to B.

A nomogram is a chart consisting of three or more lines (sometimes curves) so arranged that the required reading can be made with the aid of a straightedge ruler without resorting to calculations. More accurate calculations can be done easily on a computer with the help of a spreadsheet such as MS-Excel but the availability of computers and statistical expertise is not universal and both are restricted in developing countries where cluster samples are generally used for rapid assessment. A nomogram obviates the need for a computer and the formula. Computer literacy, that might be at premium in some developing countries, is also not needed. The nomogram can be carried easily anywhere since it is just a piece of paper and can be used repeatedly with the help of a ruler only. These are still popular in developing countries due to their simplicity. Every time a specification changes, there is no need to recalculate but only to shift the ruler to the required position on the nomogram. It can also be used to assess the cost and precision implications of different numbers of clusters.

A nomogram is particularly useful for solving repetitive problems related to engineering, design and in the medical sciences. In the latter it has been used to calculate the number needed to treat in a therapeutic trial against values of absolute risk in the absence of treatment,[6] to calculate the predicted exercise capacity in METS (metabolic equivalent) for untrained normal adults against the age,[7] and to calculate body mass index from height and weight.[8] The last is an extremely simple calculation yet a nomogram has been devised and is included in a WHO Technical Report.[8] Thus the nomogram seems to be a useful tool even for simple calculations.

Design-effect

When sampling techniques other than SRS are used, the variance of the estimate generally increases because of the restrictions imposed on the sample. In case of CRS,

$$\text{Var(CRS)} = \text{Var(SRS)}[1 + (B - 1)\rho], \qquad (1)$$

where
B = cluster size,
and ρ = intra-cluster correlation.

Design-effect D is the ratio of the variance of an estimate in a particular sampling method relative to the variance in sample random sampling. In our context D = Var(CRS)/Var(SRS). Thus it measures the efficiency of CRS relative to the SRS. This design-effect is likely to be more than one because in cluster sampling the size of cluster is always more than one and intra-cluster correlation is very rarely negative or zero. The elements within a cluster generally are relatively homogeneous which leads to a positive intra-cluster correlation. For single-stage CRS as envisaged in this communication intra-cluster correlation is equivalent to the rate of homogeneity (roh). Thus equation (1) can be written as

$$D = 1 + (B - 1)(\text{roh}). \qquad (2)$$

The ROH is the variability across different segments of the population. The ROH, and consequently design-effect, for infectious or common diseases like measles and pertussis is much higher than for rarer[9] and non-infectious diseases. This is because infectious diseases generally have focal occurrence whereas non-infectious and rarer diseases are more evenly scattered. The ROH for morbidity from infectious diseases can go up to 0.3.[10] For health care services, such as immunization coverage roh varies from 0.1 to 0.3.[10] In case of cataract blindness, the roh was found to lie in a narrow range from 0.011 to 0.016.[11] The ROH can be reasonably estimated for many situations where cluster sampling might be used as a rapid assessment methodology. If roh is known, relation (2) can be used to find D corresponding to chosen B.

Method

In single-stage cluster sampling, the number of clusters C that will estimate the prevalence rate with a required precision can be computed by the formula[10]

$$C = \frac{P(1 - P)D}{B} \times \frac{Z^2_{1-\alpha}}{L^2} \qquad (3)$$

where
C = number of clusters,
P = anticipated prevalence rate,
D = design-effect (ratio of variance of P for cluster sampling design to variance of P for simple random sampling),
α = size of the critical region ($1 - \alpha$ is the confidence level),
$Z_{1-\alpha}$ = standard normal deviate corresponding to the specified α,
L = precision required,
and B = cluster size (number of subjects in a cluster).

Since the sample size in a rapid assessment survey would necessarily be large, Gaussian theory can be safely applied. This explains $Z_{1-\alpha}$ in equation (3). The nomogram we have developed would not be applicable in the case where the sample size is small. This would be rare in community surveys for estimation of prevalence rates. It is customary to state precision L of the estimate of a prevalence rate in terms of its percentage. That is, $L = kP/100$. If L is 10% of P then $k = 10$. With this, equation (3) becomes

$$C = \frac{10^4 \times (1/P - 1) \times Z^2{}_{1-\alpha}}{k^2} \times \frac{D}{B}. \quad (4)$$

A nomogram is easy to prepare when the relationship between various parameters is linear. In order to linearize equation (4), we take logarithm and get

$$\log C = \log U + \log T \quad (5)$$

where

$$U = \frac{100 \times (1/P - 1) \times Z^2{}_{1-\alpha}}{k^2},$$
$$T = \frac{100 \times D}{B}.$$

The procedure followed by us to construct the nomogram is the one described by Adams[12] and Molnar.[13] Briefly, this is as follows.

First determine the minimum and maximum value of anticipated prevalence rate (P) and ratio of design-effect to cluster size (D/B). Then find the corresponding minimum and maximum value of U and T from equation (5). Length of anticipated prevalence rate line (P-line) and ratio of design-effect to cluster size line (D/B-line) can be selected arbitrarily. Distance between P-line and D/B-line can also be selected arbitrarily. But distance between P-line and number of clusters line (C-line) is determined by a scaling factor as described by Adams[12] and Molnar.[13] Calibrations also depend on this scaling factor.

The nomogram so obtained is shown in Figure 1. This in fact is a 6-in-1 figure because the lines corresponding to different α and L are depicted in the same figure. It consists of seven parallel lines. First on extreme left is P-line that depicts the anticipated prevalence rate of a disease. Second, third and fourth are C-lines for $\alpha = 0.05$, 0.10 and 0.20 respectively. The left side of each line is calibrated for $L = 10\%$ of P and the right side for 20% of P. These are the real operational lines that give the number of clusters needed to estimate P with a precision either 10% of P or 20% of P, and level of confidence 95%, 90% or 80%. The last three lines in Figure 1 correspond to the three different confidence levels and are calibrated for ratio of design-effect to cluster size. The scale of the last three lines is the same but their distance from the P-line varies.

Results

To find the number of clusters (C) needed to estimate the prevalence rate with specified precision, place a ruler joining the anticipated prevalence rate (P) with the ratio of design-effect to cluster size in Figure 1. Read the value of number of clusters where the ruler cuts the corresponding line of number clusters for your chosen α and L. For example, for $P = 0.04$, $L = 10\%$ of P, $D/B = 0.050$ and $\alpha = 0.05$, the number of clusters is nearly 460 as shown in Figure 1. Similarly, for $P = 0.04$, $\alpha = 0.10$, $L = 20\%$ of P (right side of $\alpha = 0.10$ among C-lines) $D/B = 0.10$ ($\alpha = 0.10$ among D/B-lines), the number of clusters is nearly 160. Also for $P = 0.06$, $\alpha = 0.20$, $L = 20\%$ of P (right side of $\alpha = 0.20$ among C-lines), $D/B = 0.030$ ($\alpha = 0.20$ among D/B-lines), the number of clusters is 18. We expect that in all cluster sample surveys, B would be known *a priori* since it is decided on the basis of survey convenience. If there is no information on design-effect D corresponding to a cluster size B, this can be obtained from rate of homogeneity (roh) as explained earlier.

It is easily seen from expression (4) that the number of clusters at precision 10% of P is exactly four times the number required for precision 20% of P for the same confidence level. That is, if the precision level is halved, the number of clusters becomes one-fourth. Also from equation (4),

$$C_1 = C_0 \times \left(\frac{k_0}{k_1}\right)^2. \quad (6)$$

where
C_0 = number of clusters when required precision is k_0% of anticipated prevalence rate,
and C_1 = number of clusters when required precision is k_1% of anticipated prevalence rate.

For example, if a surveyor estimates $C = 460$ from the nomogram for $L = 10\%$ of P, $D/B = 0.050$, $\alpha = 0.05$ and $P = 0.04$, he can estimate the number of clusters for $L = 25\%$ of P when B, α and P remain the same. This, from equation (6), is

$$C_1 = 460 \times \left(\frac{10}{25}\right)^2 = 74.$$

Thus nomogram generated values can be used to find the number of clusters for those values of L also that are not shown in the nomogram. The only calculation required is multiplication by the square of the ratio of the two ks.

Similarly a surveyor can also use this nomogram to estimate the number of clusters needed for a critical level other than $\alpha = 0.05$, 0.10 and 0.20 already given. For this, note from equation (3) that

$$C_1 = C_0 \times \left(\frac{Z_{1-\alpha_1}}{Z_{1-\alpha_0}}\right)^2, \quad (7)$$

where C_1 is the number of clusters for confidence level $(1 - \alpha_1)$ and C_0 for confidence level $(1 - \alpha_0)$. The only calculation now required is multiplication by the square of the Z values corresponding to the two αs. Thus, the nomogram can be used for any α and L—the additional calculation required is minor.

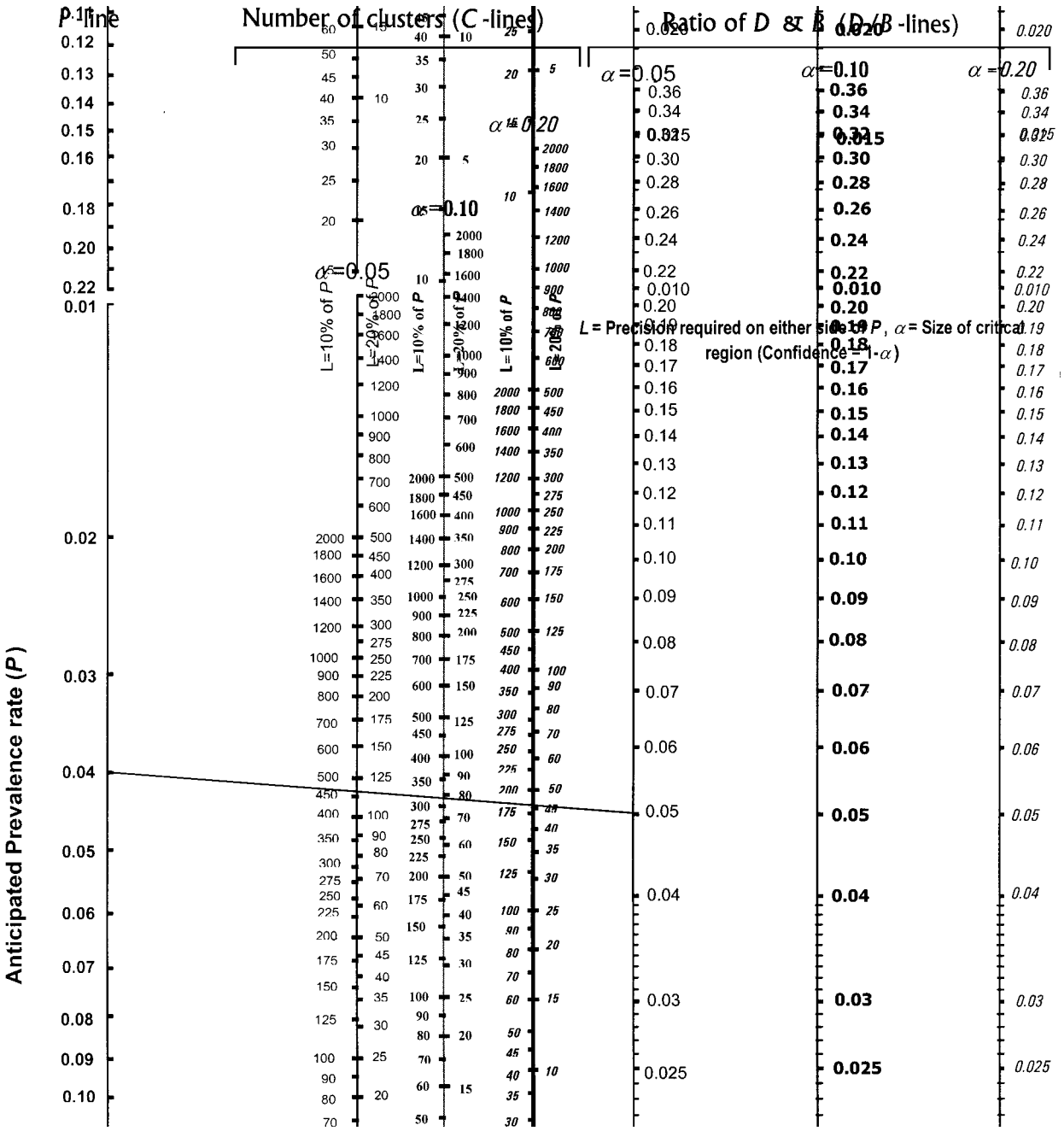

Figure 1 Nomogram for the number of clusters for a given value of *D/B* (see text)

Discussion

The aims and objectives of the investigation should be clearly specified before conducting any survey, and should be known to the investigator along with the relative cost, the time needed and the other resources available. These factors help in deciding and planning the sampling strategy including the size of the cluster in the case of CRS, the precision required and the level of confidence needed. Precision of a CRS estimate depends, among other things, on the size of the clusters. Other things being equal, a large number of clusters of small size gives a more precise estimate than a small number of clusters of large size. It also depends on how the characteristic is distributed in the population. An even distribution across various segments of the population would require small sample size whereas uneven concentration within the population would need a large sample. The design-effect is less for evenly distributed populations but greater for unevenly scattered populations. In a xerophthalmia prevalence survey, Katz *et al.*[14] calculated the design-effect by taking the ratio of variance under the EPI cluster sampling design to that of simple random sampling and found it to be 1.4, 1.5 and 5.5 for clusters of size 10, 15 and 100 units respectively. In the case of the cataract blindness survey design-effect varied from 1.26 to 2.13 for cluster of sizes 20 to 90.[11] The design-effect was found to be very close to 1 for marriage status and exposure variables and 1.4 to 1.7 for other demographic variables such as fertility behaviour, contraceptive knowledge, ever use of contraceptive and current use of contraceptives.[15] All these show that design-effect is not fixed but varies according to the population distribution pattern.

In common with all nomograms, our nomogram also depicts the mathematical relationship between various parameters. Also, this can be used inversely. Where required, the ratio of design-effect to cluster size can be obtained corresponding to a specific number of clusters. Thus roh and design-effect can be estimated for a specific cluster size. The size of the critical region can be crudely estimated as between 5% and 10% or between 10% and 20% if other values are known. Similarly, precision level L can also be crudely estimated when α, the number of clusters and the ratio of design-effect to cluster size is known. When used in this manner, the nomogram is helpful in assessing the cost and precision implications of different sizes and number of clusters on the estimation of the prevalence rate.

Note that our nomogram does not incorporate Type II error. Thus this cannot be used for intervention studies where the main purpose is testing of a hypothesis rather than estimation. This nomogram is meant only for surveys where the objective is estimation of the magnitude of a problem in terms of proportion affected. As already stated, the nomogram is also not applicable if the sample size is small, however, we do not anticipate community surveys with small sizes for finding prevalence rates. Another limitation is that nomogram readings by their very nature are not fully accurate. In place of 462, one might read 460 from the line. But this minor deviation would rarely compromise the utility of the nomogram since the number of clusters in any case is based on an anticipated prevalence rate which might be in error.

KEY MESSAGES

- A nomogram was prepared that instantly provides the number of clusters needed for single-stage cluster sampling estimation of a prevalence rate in a community.
- Inputs required are anticipated prevalence rate, ratio of design-effect to cluster size, precision desired and confidence level.
- The nomogram is not applicable to intervention studies or to studies based on small sample sizes.

References

[1] Sukhatme PV. *Sampling Theory of Surveys with Applications*. Iowa: Iowa State University Press, 1954.

[2] Cochran WG. *Sampling Techniques, 3rd Edn*. Chichester: John Wiley & Sons, 1977.

[3] Henderson RH, Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bull World Health Organ* 1982;**60:**253–60.

[4] Limburg L, Kumar R, Indrayan A, Sundaram KR. Rapid assessment of prevalence of cataract blindness at district level. *Int J Epidemiol* 1997;**26:**1049–54.

[5] Lemeshow S, Robinson D. Surveys to measure programme coverage and impact: a review of the methodology used by the expanded progeamme on immunization. *World Health Stat Q* 1985;**38:**65–75.

[6] Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. *BMJ* 1996;**312:**426–29.

[7] Krishnaswami S, Chandy ST, George P. Exercise capacity in untrained normal adults: a nomogram for Indians. *Ind Heart J* 1996;**48:**357–60.

[8] World Health Organization. Physical status: the use and interpretation of anthropometry. *World Health Organ Tech Rep Ser 854,* 1995, p.434.

[9] Rothenberg RB, Labanov A, Singh KB, Stroh Jr G. Observations on the application of EPI cluster survey methods for estimating disease incidence. *Bull World Health Organ* 1985;**63:**93–99.

[10] Bennett S, Woods T, Livanage WM, Smith DL. A simplified general method for cluster-sample surveys of health in developing countries. *World Health Stat Q* 1991;**44:**98–106.

[11] DANPCB. *Rapid Assessment of Cataract Blindness in Persons of 50 Years and Older: Report of Survey Design Methodology and Results*. New Delhi: Danish Assistance for National Programme for Control of Blindness, 1997.

[12] Adams DP**.** *Nomography: Theory and Application*. Hamden: Connecticut Archon Books, 1964, pp.1–17.

[13] Molnar J. *Nomographs: What They Are And How To Use Them*. Ann Arbor: Ann Arbor Science, 1981.

[14] Katz J, Yoon SS, Brendal K, Wesr Jr KP. Sampling designs for xerophthalmia prevalence surveys. *Int J Epidemiol* 1997;**26:**1041–48.

[15] Ulusoy M. Sampling errors for selected variables from the 1988 Turkish population and health survey. *Turk J Popul Stud* 1991;**13:**33–55.