The Central African Journal of Medicine

Supplementary Issue to 1992 Volume 38,
1991 University of Zimbabwe Annual Research Day

effects of cluster sampling in an African urban setting

P FERRINHO,[2*] A VALLI,[2] T GROENEVELD[4]
E BUCH,[3] D COETZEE[1,3]

SUMMARY

Cluster sampling was popularised by the sampling procedure promoted by the WHO/UNICEF for the evaluation of the expanded programme of immunisation (EPI). Without a clear understanding of the limitations of the sampling strategy used, this sampling strategy has been extended to other types of surveys.

This article shows how to approach the assessment of cluster sampling techniques scientifically by calculating design effects (DEFFs) and rates of homogeneity (roh) and illustrates this scientific assessment with three case studies from Alexandra in South Africa. We report on the DEFFs and rohs for variables studied in these surveys.

The DEFF for all the variables relating to housing tended to exceed two and was as high as 6,99 for the

[1]*Alexandra Health Centre and University Clinic and Institute for Urban Primary Health Care*
[2]*Centre for Epidemiological Research in Southern Africa of the South African Medical Research Council*
[3]*Department of Community Health, Faculty of Medicine, University of the University of the Witwatersrand*
[4]*Institute of Biostatictics of the Medical Research Council*
[*]*Current address:*
Department of Tropical Public Health
Institute of Hygiene and Tropical Medicine
Universidade Nova de Lisboa
Rua da Jungueira, 96
1300, Lisbon
Portugal

variable new development. The variables relating to health service utilisation and health practices, namely immunisation status, nutrition status, presence of Road to Health Cards (RTDCs), breast-feeding and knowledge of diarrhoea and oral rehydration all had a DEFF close to one. The variables relating to contraception use, literacy and schooling had DEFFs close to one and a half.

For a few variables the DEFFs were below one and the rates of homogeneity less than zero. The highest values of roh were for environment factors (all above 0,1433). Rohs for factors related to utilization of PHC services were mostly between 0,0200 and 0,0499. No single class of factors seemed to be related to very low values of roh. These results are then discussed.

The sampling strategy used for evaluation of the expanded programme of immunisation: Cluster sampling was popularised by the sampling procedure promoted by the WHO/UNICEF for the evaluation of the expanded programme of immunisation (EPI).[1,2] Without a clear understanding of the limitations of the sampling strategy used, this sampling strategy has been extended to other types of surveys.[3]

This article shows how to approach assessment of the cluster sampling technique scientifically by calculating design effects (DEFFs) and rates of homogeneity (roh) and illustrates this scientific assessment with three case studies from Alexandra in South Africa.

Design effects and rates of homogeneity: Standard statistical methods have been developed on the assumption of simple random sampling. Independent selection of elements (hence independence of observations) greatly facilitates statistical analysis. However, much research can only be accomplished with complex sample designs, including cluster sampling.

Cluster sampling denotes methods of selection in which the sampling unit contains more than one population element, with relative homogeneities within clusters that negate the independence of sample elements and introduce complexity into statistical analysis. There has been considerable attention to the problem of randomising clusters rather than individuals. The consequence of randomising clusters is a reduction in effective sample size, with the variance of any estimated outcome larger than for a simple random design. This variance depends on intra-cluster dependence of the variable, on the variability between clusters and on the number and size of clusters selected for the survey. Another factor in terms of repeated or longitudinal surveys is the impact of differential interventions on the variance of the variables being measured.

Inter-cluster variability is measured by the rate of homogeneity (roh). Roh is a measure of homogeneity that takes into account all the stages of the sample design. The sample designer can reduce roh by selecting larger clusters or by increasing the distance between the study units. Roh will be increased by poor standardisation of data collection. Roh is a summary statistic portable from one survey to another of similar or different design.[4]

A variable which is very homogeneously (randomly) distributed across the population will have roh close to zero, whereas one which is heterogeneous or naturally clustered, will have a higher value of roh up to one. Values of roh below zero can be considered as due to sampling error and treated as zero, although very occasionally (particularly in stratified samples) negative values of roh can occur when clustering produces more uniformity than would otherwise be produced by simple random sorting.[5]

However, the increase in the variance of an estimator due to clustering depends not only on the heterogeneity of the variable across the population, but also on the clustering imposed by the study design. The design effect (DEFF) incorporates both these sources of variation and as such can be used for other surveys of the same design, for the same variable or for variables of similar roh.[6] The DEFF was first described by Cornfield in 1951.[7] It is the variance of the estimated outcome under cluster sampling relative to the variance under simple random sampling.[5,8] Occasionally with an odd clustering effect DEFF is less than one, but generally is more than one. When roh is positive DEFF exceeds one. Even a relatively small positive roh can have a large effect on the variance if the sample cluster is large.[5]

Case studies from Alexandra: Alexandra, on the north border of Johannesburg, is a small urban area of approximately five square kilometres. It has a population of over 200 000 living in shacks, hostels and a formal housing sector.[10]

The Alexandra Health Centre and University Clinic (AHC) has been, since the 1920s, the main

provider of comprehensive preventive, promotive, curative and rehabilitative health care. A local authority government clinic providing a selective, mainly preventive service, started in 1986.[10] There are also 19 general practitioners in private practice who provide a predominantly curative service.[11] To guide planning and service development, the AHC has conducted a number of surveys to determine the health status of the local community, using methodologies similar to the one recommended for evaluation of the EPI.[12-16]

Three cluster sample surveys dealing with child health problems have been completed in Alexandra Township since 1988.[12-14] The accurate calculation of standard errors of the variables measured and hence their DEFFs and rohs would assist with a more scientific approach to planning sampling strategies in future surveys. Some of these results have already been reported.[17]

In this article we report on the DEFFs and rohs for variables studied in three of the surveys mentioned.

MATERIALS AND METHODS

The three studies are revised for their aims, inclusion criteria, sampling methods, samples sizes, variables measured and field work (Table I). The methods of calculation of standard errors, DEFFs and rohs are also described.

All the studies were cross -sectional, used a cluster sampling method and had a number of variables in common, including details and possession of a Road to Health Card (RTHC), which allowed for comparisons over time. All the surveys, besides the 1989 survey on oral rehydration therapy (ORT) and diarrhoeal diseases in children, measured immunisation coverage. The major differences related to sample size and to field procedures (Table I).

The 1988 and 1990 surveys collected immunisation data from Road to Health Cards (RTHC) and mothers were interviewed for PHC indicators (1988, 1990), including use of family planning and duration of breast feeding (1990). In 1990 heights and weights were also measured. The ORT study was a descriptive cross-sectional assessment of knowledge, attitude and skill regarding ORT.

Standard interview schedules were used for all the surveys and data captured into a computer for analysis.

Health workers who were also residents of Alexandra did all the interviewing. They received rigorous training at the AHC and under actual field conditions. The field workers worked in teams of three to four supervised by an experienced field worker.

The method of selecting the clusters was specific to areas with a predominance of either shacks, old brick houses, or new residential areas (new developments). Alexandra is laid out in a regular grid fashion with 91 blocks of similar population size. Most blocks have 40 plots, most of which have a central house, subdivided for multiple occupancy and a variable number of shacks. About 40 people are resident on each plot. Most of the squatting in Alexandra is integrated into the existing plots. However, discrete areas of informal shelters have started to spring up within the community.

The first stage of sampling was to define clusters, of about 500 households, from the bricks and informal settlements areas using a map, a recent aerial photograph and field inspection. Forty-five clusters were then randomly selected. A random plot number was selected on each block. At the time of the interview a random starting dwelling on the plot was selected from a bag of numbers carried by the field workers who then moved according to well defined rules until the specified number of children in the required age group were found.

Table I: Summary of methodology of surveys done in Alexandra

Survey	Sampling strategy	Number of clusters	Number of elements	Sample size	Selection of starting point	Age of children	Interview mothers	Objectives
1988	Cluster*	45	7	315	Random	12-23 Months	Yes	Vaccination/PHC
1989	Cluster*	45	10	450	Random	6-23 Months	Yes	ORT/Diarrhoea
1990	Cluster*	45	10	450	Random	12-23 Months	Yes	Vaccination/PHC
						12-23 Months	No	Vaccination

**One stage cluster sampling with probability to size*

Calculation of DEFF and roh: For the purposes of calculation of DEFF and rohs all the variables were dichotomised as defined in Tables II-IV. The SAS(185) system was used to calculate the standard errors of the variables using the formula specific for cluster sample surveys:

$$sec=[c/\Sigma xi]\sqrt{\{[\Sigma yi^2 - 2p\Sigma xiyi + p^2\Sigma xi]/[c(c-1)]\}}$$

sec=standard error for cluster sample.
p=proportion = $\Sigma yi/\Sigma xi$
c=No. of areas surveyed
yi=No. with observed variable in the i'th cluster
xi=No. of sample units in the i'th cluster

The DEFFS were then calculated according to the formula:[6]

$DEFF=sec^2/s^2$

s=standard error for simple random sample

$$= p\sqrt{(1-p)/N}$$

N=sample size

Roh was also calculated for each of the variables according to the formula:[6]

roh=(DEFF - 1)/(b - 1)

b=mean No. of final units sampled per cluster.

RESULTS

The results are shown in Tables II-IV.

The DEFF for all the variables relating to housing tended to exceed two and was as high as 6,99 for the variable new development. The variables relating to health service utilisation and health practices, namely immunisation status, nutritional status, presence of RTHCs, breast-feeding and knowledge of diarrhoea and oral rehydration all had a DEFF close to one. The variables relating to contraception use, literacy and schooling had DEFFs close to 1,5.

For a few variables the DEFFS were below one and the rates of homogeneity less than zero.

The highest values of roh were for environmental factors (all above 0,1433). Rohs for factors related to utilisation of PHC services were mostly between

Table II: Immunisation survey, 1988

Variable	dichotomisation (frequencies)	sec*	s**	DEFF	roh
No. on door of house	yes (182) vs no (139)	0,0414	0,0277	2,25	0,1746
type of housing	brick (215) vs other (114)	0,0444	0,0262	2,86	0,2543
shack or not	shack (155) vs (174)	0,0408	0,0275	2,20	0,1644
ANC attendance	none (18) vs some (312)	0,0140	0,0125	1,25	0,0336
	AHC (203) vs other (127)	0,0280	0,0268	1,09	0,0124
place of birth	home (20) vs supervised (310)	0,0148	0,0131	1,27	0,0362
	AHC (150) vs other (180)	0,0301	0,0274	1,21	0,0284
length of stay in Alexandra	<=5yrs (129) vs >5yrs (201)	0,0264	0,0269	0,97	-0,0045
	<=1yr (97) vs >1yr (233)	0,0248	0,0251	0,98	-0,0032
other children under five years	0(180) vs 1-4 (146)	0,0313	0,0275	1,29	0,0401
length of schooling	0-3yrs (100) vs 4-9yrs (230)	0,0316	0,0253	1,56	0,0758
literacy in vernacular	literate (294) vs not (36)	0,0183	0,0172	1,14	0,0186
	good literacy (262) vs not good or none (68)	0,0258	0,0223	1,34	0,0465
literacy in English	literate (266) vs not (64)	0,0259	0,0218	1,42	0,0573
	good literacy (213) vs not good or none (117)	0,0333	0,0263	1,60	0,0811
possession of road to health card	no (10) vs yes (320)	0,0096	0,0094	1,03	0,0034
	AHC (217) vsw other (107)	0,0298	0,0261	1,30	0,0414
BCG at three months	received (107) vs not (223)	0,0244	0,0258	0,89	-0,0146
Measles at one year	received (145) vs not (185)	0,0306	0,0270	1,28	0,0387
Fully immunised at one year	received (132) vs not (198)	0,0316	0,0273	1,34	0,0466

sec=standard error for the study sample taking into account the cluster design
****s=standard error for the study sample assuming random sampling***

Table III: Oral Rehydration Survey, 1989

Variable	Dichotomisation (frequencies)	sec*	s**	DEFF	roh
no on door of house	yes (296) vs no (154)	0,0423	0,0223	3,19	0,2592
type of housing	brick (318) vs other (132)	0,0405	0,0214	3,57	0,2570
shack or not	shack (177) vs other (271)	0,0407	0,0230	3,12	0,2129
number of children	1 or 2 (269) vs 3+ (182)	0,0250	0,0231	1,17	0,0169
number of children dead	0 (366) vs 1+ (81)	0,0214	0,0181	1,39	0,0397
other children under 5 years	0 or 1 (445) vs 2+ (6)	0,0051	0,0054	0,91	-0,0910
	0-2 (414) vs 3+ (37)	0,0154	0,0129	1,41	0,0412
length of stay in Alexandra	0-4yrs (148) vs 5+yrs (303)	0,0289	0,0221	1,71	0,0708
length of schooling	<=6yrs (73) vs >6yrs (339)	0,0243	0,0180	1,48	0,0521
literacy in English	literate (363) vs not (88)	0,0224	0,0187	1,43	0,0434
	good literacy (296) vs not good or none (155)	0,0357	0,0224	2,54	0,1539
literacy in vernacular	literate (398) or not (53)	0,0153	0,0152	1,01	0,0011
	good literacy (353) vs not good or none (98)	0,0263	0,0194	1,83	0,0825
diarrhoea in last two weeks	yes (218) vs no (233)	0,0244	0,0235	1,08	0,0076
>3 loose stools daily in last two weeks	yes (220) vs no (209)	0,0255	0,0235	1,18	0,0183
possession of road to health card	yes (361) or no (90)	0,0196	0,0188	1,09	0,0088
awareness of ORS	yes (416) or no (35)	0,0130	0,0126	1,07	0,0070
knew best treatment for simple diarrhoea	ORT (273) vs rest (178)	0,0290	0,0230	1,59	0,0587

sec=sampling error for the study sample taking into account the cluster design
****s=sampling error for the study sample assuming random sample***

Table IV: Immunisation, nutrition and PHC indicator survey, 1990

Variable	Dichotomisation (frequencies)	sec*	s**	DEFF	roh
number of door of the house	yes (263) vs no (158)	0,0361	0,0236	2,34	0,1433
type of housing	brick (292) vs other (129)	0,0390	0,0225	3,01	0,2148
shack or not	shack (138) vs other (283)	0,0455	0,0229	3,96	0,3163
new development	yes (69) vs no (352)	0,0477	0,0180	7,00	0,6411
possession of road to health card	AHC (300) vs no or other (121)	0,0248	0,0221	1,27	0,0285
	yes (420) vs no (1)	0,0024	0,0024	1,00	0,0000
ANC attendance	AHC (256) vs no or other (139)	0,0283	0,0240	1,38	0,0436
	yes (381) or no (14)	0,0085	0,0093	0,83	-0,0194
place baby delivered	home (48) vs supervised (347)	0,0161	0,0164	0,95	-0,005
	AHC (142) vs other (253)	0,0196	0,0241	0,66	-0,0390
period breastfed	<6 mnts (88) vs >6 mnts (307)	0,0216	0,0210	1,00	0,000
	<1yr (115) vs >1yr (280)	0,0255	0,0229	1,24	0,027
PNC attendance	yes (348) vs no (47)	0,0173	0,0163	1,13	0,014
	AHC (249) vs other (146)	0,0263	0,0243	1,17	0,0197
use of family planning	yes (274) vs no (117)	0,0291	0,0232	1,58	0,066
BCG at three months	received (332) vs not (89)	0,0226	0,0199	1,29	0,0311
Measles at one year	received (288) vs not (133)	0,0256	0,0227	1,28	0,0300
Fully immunised at one year	received (246) vs not (175)	0,0235	0,0240	0,95	-0,0049
weight for age of child	<50th percentile (301) vs >=(120	0,0226	0,0220	1,06	0,0063
height for age of child	<3rd percentile (186) vs >=(235)	0,0252	0,242	1,08	0,0085

sec=sampling error for the study sample taking into account the cluster sampling
****s=sampling error for the study sample assuming random sampling***

0,0200 and 0,0499. No single class factors seemed to be related to very low values of roh.

The values of roh presented some trends for variables measured during more than one of the surveys.

The value of roh for residence in a shack increased and for residence in a brick house decreased. The value of roh for coverage by BCG increased. Roh for attendance for ante-natal care (ANC) at the AHC also increased, but decreased for non-attendance of ANC, site of delivery, possession of a RTHC, coverage with measles vaccine and for full immunisation by one year of age.

DISCUSSION

The values for DEFF and roh are lower than those usually reported in the literature for similar variables, probably reflecting a community where, although clustering of environmental variables is important, the other variables studied are reasonably randomly distributed. The values DEFF and roh for the variables reflecting utilisation of PHC services suggest that the AHC is reaching the community fairly homogeneously, with preventive and promotive services and the clustering one might expect with a health service that only meets the needs of a section of the community not encountered. To what extent these observations would apply to other similar communities is not known as not much data of this nature is available.

We found changes in roh over time, for variables measured in more than one survey, useful to reflect on the impact of changes in the community and on the impact of our health interventions (even taking into account that roh also decreases with an increase in the cluster size and that we cannot comment on the proportion of changes observed for roh that are attributable to the differences in cluster size between the surveys).

Rohs for attendance of PHC services have decreased (measles and full immunisation, site of delivery, possession of an RTHC) indicating a homogeneous penetration of the community by the PHC services in Alexandra. The decrease in roh for non-attendance of ANC is probably explained similarly. The exceptions on BCG coverage and ANC attendance at the AHC are probably easy to explain.

The increase in roh for BCG coverage probably reflects the many children not born in Alexandra, not receiving BCG at the of site of delivery, migrating to informal shelters in Alexandria where vaccination has until recently been provided by a mobile clinic where BCG was not available. Roh for attendance for ANC at the AHC has also increased, probably for reasons very similar to the ones for BCG coverage, but probably also reflecting a growing middle class cluster in the newly upgraded residential area and more use being made of private practitioners.

For the environmental variables, roh decreased for residence in brick houses and increased for residence in informal shelters. The first probably reflects the extensive building and upgrading programme in the township. while the second reflects the tendency for more clustering of the informal shelters, as the upgrading programme progressed.

Conclusions: The determination of DEFFs and rohs is a simple statistical procedure that should be carried out routinely. The wealth of empirical data so gathered would be useful not only to plan local or regional health surveys, but could even provide useful data to monitor the impact of community health interventions.

The finding that immunisation coverage variables had DEFFS close to one is important as it implies that analysis of cluster sample surveys of immunisation coverage in settings such as Alexandra may give reasonably valid results even if the data are analysed as if they were from a simple random sample.

The high DEFFs for housing variables suggest that the sample size may need to be relatively larger for cluster sample surveys in which the accurate measurement of housing variables are central to the objectives of the study.

In conclusion, the EPI cluster sampling strategy is commonly used. It is not easy to analyse data derived from these surveys, because of the clustering effect implicit in the strategy. Calculation of DEFF is a useful factor to consider when analysing such data. DEFF and roh are also useful as baselines to plan sample sizes (total size, number of clusters and number of sampling units per clusters) in future surveys and possibly as useful measures to monitor the impact of community health interventions.

ACKNOWLEDGEMENTS

This article benefited from incisive comments by the Editorial Committee of the Centre for Epidemiological Research in Southern Africa.

REFERENCES

1. Henderson RH, Sundaresan T. Cluster sampling to assess immunisation coverage: a review of experience with a simplified sampling method. *Bull WHO* 1982; 60:253–260.
2. Lemeshow S, Robinson D. Surveys to measure programme coverage and impact: a review of the methodology used by the expanded programme of immunisation. *Wld Hlth Statist Quart* 1985;38:65–75.
3. Maru M, Getahun A, Hosana S. A house-to-house survey of neonatal tetanus in urban and rural areas in the Gondar region, Ethiopia. *Trop and Geographical Medicine* 1988; 40:233–236
4. Frerichs RR. Simple analytic procedures for rapid microcomputer-assisted cluster surveys in developing countries. *Public Health Reports* 1989; 104:24–35.
5. Kish L. Survey Sampling. New York, London, Sidney: John Wiley & Sons, 1965.
6. Bennet S, Woods AJ, Liyanage, Smith DL. A simplified general method for cluster sample surveys of health in developing countries. *World Health Statistics Quarterly* 1991; 44: 98–106. *Medicine*, 1989.
7. Cornfield J. Modern methods in the sampling of human populations. *Am J Public Health* 1951; 41:654–661.
8. Mickey RM, Goodwin GD, Constanza MC. Estimation of the design effect in community intervention studies. *Statistics In Medicine* 1991; 10:53–64.
9. Butcher B, Tarling R. Programs for analysing complex sample surveys. *The Professional Statistician* 1986; 5:3–6.
10. Ferrinho P, Robb D, Mhlongo A *et al*. A Profile of Alexandra. *S Afr Med J* 1991; 80:374–378.
11. Frame G, Ferrinho P, Wilson TD. The care of STDs in Alexandra: review of previous research and a survey of General Practitioners. *S Afr Fam Practice* 1991; 12:87–92.
12. Rees H, Buch E, Ferrinho P, Groeneveld H, Neethling A. Immunisation coverage and reasons associated with non-immunisation in Alexandra Township in September 1988. *S Afr Med J* 1991; 80:378–381.
13. Ratsaka M, Buch E. Oral rehydration knowledge and practice in Alexandra Township. Presented at the 8th Conference of the Epidemiology Society of Southern Africa, Durban, South Africa, 1989.
14. Coetzee D, Ferrinho P, Reinach SG. Vaccination and nutritional status of children 12–23 months of age in Alexandra Township. Ninth Conference of the Epidemiological Society of Southern Africa, East London, South Africa, 11–13 September 1990.
15. De Swardt R. Valli A. Immunisation coverage of 12 to 23 month-olds in the Southern Transvaal region. Unpublished report submitted to the Department of National Health and Population Development, September 1990. Department of Community Health, Faculty of Medicine, University of the Witwatersrand, Johannesburg, South Africa, 1990.
16. Cornielje H, Ferrinho P, Kemp S, Wilson TD, Coetzee D, Reinach SG. The development of a community-based rehabilitation programme for the community of Alexandra. Tenth Conference of the Epidemiological Society of Southern Africa, Cape Town, South Africa, July 1990.
17. Ferrinho P, Valli A, De Swardt R, Groeneveld HT, Coetzee D. A comparison of two vaccination coverage surveys in Alexandra using different cluster sampling methodologies. *Southern African Journal of Epidemiology and Infection* 1992; in press.
18. SAS Manuals, Version 6 Edition. SAS Institute Inc, Cary NC, 1985.