Sample Design for a Survey of Immunization Levels of Children in Licensed Day-Care Facilities in North Carolina

ALTHOUGH IMMUNIZATION PROGRAMS have resulted in virtual elimination of some infectious diseases and drastically reduced the incidence of others, continued protection requires constant surveillance and appropriate reinforcement. Thus, awareness of the need for repetitive studies of the immunization levels of populations has become widespread. In North Carolina, such studies have been directly related to the State's compulsory immunization law (General Statute 130–87), which requires three DTP inoculations and three doses of oral poliomyelitis vaccine before a child's first birthday and immunization against red measles (rubeola) before the second birthday. Further, General Statutes 110–91 and 130–90 require that all children attending licensed day-care facilities or any public, private, or parochial school in the State be in compliance with 130–87. The only allowable exemptions are for documented medical reasons or religious beliefs.

Statewide sample surveys of the immunization levels of 2-year-olds were conducted in 1972 and 1974 (*1*,*2*). In 1973 an immunization census was taken of first graders in public schools (*3*), and in 1974 the census was expanded to include kindergarten children and private and parochial schools (*4*). It is expected that an immunization census of kindergarten and first-grade children will be taken annually in all schools.

In each of the surveys mentioned, information was elicited concerning compulsory immunization. In addition, in the surveys of 2-year-olds attempts were made to document levels of immunization against rubella. Although not required by law, both rubella and mumps vaccines are recommended in the immunization schedule for 1-year-old children (*5*).

☐ *Mrs. Surles and Mr. Johnson are statisticians in the Public Health Statistics Branch of the North Carolina Division of Health Services. Tearsheet requests to Kathryn B. Surles, State of North Carolina Department of Human Resources, P.O. Box 2091, Raleigh, N.C., 27602.*

In 1975, immunization levels were assessed further by means of a statewide survey of children aged 1–5 who were enrolled in licensed day-care facilities. The logistics of that effort are presented here in the hope that they may benefit others whose budgets do not allow for the services of sampling experts and who otherwise may find the application of sampling methodology a painful, if not impossible, process.

Choice of a Survey Plan

Licensure of a day-care facility in North Carolina requires the facility to maintain records of the medical examinations that are required of its enrollees. Because fear of license revocation or other reprisal might preclude an operator's honest examination of enrollees' records, it was considered unfeasible to ask facility operators to report immunization information. Thus, in lieu of a census such as that conducted in schools, a sampling scheme was devised whereby field staff of the immunization program would visit selected facilities and abstract the desired information from their records.

Age-specific counts of children enrolled in 1,808 facilities on each facility's most recent renewal date were available from the North Carolina Office of Child Day Care Licensing. Although this data set allowed various sampling plans, the following standard textbook methods were not acceptable: (*a*) simple random sampling of children because of extensive time and travel requirements; (*b*) sampling of facilities (with either equal probability or probability proportional to size) because of the time factor associated with complete measurement in large facilities; and (*c*) multistage sampling because of the level of mathematical difficulty required for estimation of variances. On the other hand, the plan devised, single-stage sampling of equal-sized clusters of children, was not associated with any of the preceding constraints. This sampling scheme offered the computational ease of a self-weighted sample and provided for coverage of a larger number of facilities than would either the preceding (*b*) or (*c*). Also, depending upon choice of cluster size, the plan allows the user some flexibility in customizing the sample to fit his particular situation with respect to survey resources, albeit the user must be aware that a large sampling unit (cluster) usually gives less accurate results than a small unit.

The Sample Design

Sampling frame. The sampling frame consisted of clusters of children aged 1–5 who were enrolled in licensed day-care facilities on April 23, 1975. With clusters of expected size 10 (a smaller cluster size would have necessitated more travel than available resources could bear) and the use of enrollment counts as of each facility's last licensure, the number of clusters in a facility was the total number of children aged 1–5 expressed as the nearest multiple of 10. For example, a facility with 54 children contained 5 clusters, whereas 1 with 55 children contained 6 clusters. The 1,808 facilities contained a total of 5,610 clusters of children. Thus, with a total population of 55,869 at the time of each facility's last license renewal, the final expected cluster size was $55{,}869 \div 5{,}610 = 9.9588 \cong 10$. Identification of the sample clusters is discussed later.

Sample size. Based on the last known age-specific counts of children in each facility and an assumption that immunization levels were a conservative 50 percent (because this yields maximum required sample size), a sample size of 186 clusters was needed to estimate with 95 percent confidence and within 10 percent of the true proportion a binomial proportion for each of several regions and single years of age. (At the stated confidence and precision levels, binomial theory required 100 children in each domain of interest. Representing 5.39 percent of the initial population, 1-year-olds were the smallest domain. With clusters of expected size 10, the required number was $186 = \frac{100}{(.0539)\ (10)}$.) However, the number was increased to an actual sample size of 250 because (*a*) the sample size required with cluster sampling was unknown, but it is probably greater than the estimate based on binomial theory and (*b*) it seemed likely that recent economic conditions may have resulted in the closure of some day-care facilities and reduced enrollment in others. It was hoped that the additional clusters would cover any sample deficiencies due to the preceding (*a*) or (*b*).

Selection method. Clusters in the 1,808 facilities were sequentially numbered from 1 to 5,610, and random numbers from 1 to 5,610 were generated. The first 250 random numbers identified the clusters comprising the sample.

Further preliminary work was undertaken to insure that the 250 sample clusters contained a sufficient number of 1-year-olds (the smallest group for which estimates were desired). This aspect of the fieldwork involved contacting the 233 facilities containing 1 or more of the 250 sample clusters to ascertain each facility's current enrollment by single years of age. Application of the 1-year-old counts to each facility's sampling rate indicated that the expected number of 1-year-olds well exceeded 100, the minimum number judged to be adequate. Otherwise, plans were to successively increase the sample size by use of previously generated random numbers to identify additional clusters. (The actual sample yielded 145 one-year-olds.)

Identification of clusters. When the final sample was determined, each facility's sampling rate—the number of sample clusters in the facility divided by the total number of clusters—and a random start number then determined the selection of children from the facility's file of children aged 1–5. To illustrate, if a facility contained 20 clusters and 1 of these was a sample cluster, the sampling rate was 1/20. If the randomly selected start number (between 1 and 20) were 4, then the children numbered 4, 24, 44, 64, and so on would comprise the sample cluster from that facility. This procedure prevailed regardless of the facilities' file arrangements, which were either in alphabetical or enrollment-date order. In either instance, it is clear that a sample cluster may not include siblings, although there may be many siblings in the universe. We resolved to accept this amount of bias because we considered it minimal.

Post-stratification. Single-stage sampling of equal-sized clusters results in a sample that is self-weighted so that survey estimates can be obtained easily by post-stratification for various regional structures. The resulting population and sample sizes for four regions of the State were as follows:

Region	*Population size*	*Sample size*
Western	2,088	93
North central	1,319	54
South central	1,446	65
Eastern	757	38
Total	5,610	250

Nonresponse. Data were not obtained for all clusters. As shown in the following table, facilities containing 18 clusters of children had closed by April 23, the effective date of the survey, and 3 clusters were void. A void cluster may occur when a facility's sampling rate and start number (based on enrollment at last licensure) are inconsistent with present enrollment. For example, suppose that the randomly selected start number (between 1 and 20) is 16; if fewer than 16 children are presently enrolled, the sample cluster is void.

	Clusters of children			
Region	*Facility Interviewed*	*Facility closed*	*Cluster void*	*Data not obtained*
Western	82	7	2	2
North central	45	5	1	3
South central	56	4	0	5
Eastern	36	2	0	0
Total	219	18	3	10

Thus, a total of 21 clusters contained no sample children. These clusters do not constitute nonresponse; rather, they contribute zero children to the sample estimates. The zero clusters reflect, as anticipated, facility closures and reduced enrollments probably resulting from depressed economic conditions.

Although a concerted effort was made to obtain data for all viable clusters, delays in the survey resulted in the loss of six clusters because of summer closures before June. Data for four other clusters were not obtained owing to lack of cooperation by the facilities' operators. These 10 clusters constituted nonresponse.

Based on enrollment counts obtained in preliminary fieldwork, 94 children comprised the 10 nonresponse clusters (10 clusters in 10 facilities). The 240 response clusters (in 223 facilities) yielded a total of 2,351 children. Nonresponse for the survey was therefore 3.8 percent (94÷2,445). Response and nonresponse by region and age were as follows:

Region and age	*Response*	*Nonresponse*	
		Number	*Percent*
Region:			
Western	855	4	0.5
North central	457	41	8.2
South central	656	49	7.0
Eastern	383	0	0.0
Age (years):			
1	144	1	0.7
3	284	10	3.4
2	512	21	3.9
4	717	29	3.9
5	694	33	4.5

Population and sample comparisons. The age and regional distributions of children from the sample are compared with the distributions of children comprising the initial population in the following table. The close agreement of these percentage distributions indicates that a representative sample was obtained.

Region and ages	*Population*	
	Initial	*Sample*
Region:	*(55,869)*	*(2,445)*
Western	37.5	35.1
North central	23.7	20.4
South central	25.6	28.8
Eastern	13.3	15.7
Age (years):		
1	5.4	5.9
2	12.5	12.0
3	22.7	21.8
4	30.4	30.5
5	29.1	29.7

Estimation of population totals. The equations presented for population estimates in this and the following section are given by Cochran (*6*).

An estimate of the total number of children aged 1–5 enrolled in licensed day care facilities on April 23, 1975, is

$$\hat{Y} = \left[\sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i \right] X = \frac{y}{x} X$$

where n = number of clusters in the sample,
y_i = number of childern observed in the i^{th} cluster,
X = number of children in the initial population, and
x_i = number of children expected in the i^{th} cluster $= \frac{X}{N}$ where N is the number of clusters in the universe.

In our statewide survey, $n = 250$, $x \cong 2,500$ ($x_i = 9.9588 \cong 10$ for all clusters), $y = 2,445$ y and $X = 55,869$. Thus, the estimated population is $\hat{Y} = 54,640$ children aged 1–5.

Since all x_is are equal, the formula for the estimated variance of this estimate, with the finite population correction (fpc) ignored, reduces to

$$v(\hat{Y}) = N^2 s^2 / n$$

where $N =$ number of clusters in the universe

$$\text{and } s^2 = \sum_{i=1}^{n} (y_i - y)^2 / n - 1.$$

With $N = 5{,}610$, the variance of the population total is 3,420,170 which yields a 95 percent confidence interval of 54,640 ± 3,625.

Under whatever assumption one makes concerning nonresponse and missing values, as discussed in the following section, any percentage figure from the sample may be interpreted as an estimate for the entire population. Therefore, statewide categorical totals may be obtained by multiplying the statewide percentage estimates by 54,640. Similarly, regional percentage estimates may be applied to regional population totals.

Estimation of population proportions. In the estimation of immunization levels, one may deal with nonresponse and missing data (child's immunization record not on file) in alternative ways, that is, one can assume that all the children were immunized, that none were immunized, or that they had some in-between level of immunization. In the computation of variances, the assumption of "in-between level" requires assignment of immunization values to each child for whom data were not obtained. To avoid such assignment, and in the interest of being conservative in our estimates of the proportion of children immunized, we assumed nonimmunization for all nonresponse and missing records in the estimates that follow.

Structurally, the sample estimate of a proportion is a typical ratio estimate computed as

$$p = \sum_{i=1}^{n} w_i \bigg/ \sum_{i=1}^{n} y_i = \frac{w}{y}$$

where $n =$ number of clusters in the universe

$y_i =$ number of children observed in the i^{th} cluster,

and $w_i =$ number of children in the i^{th} cluster who have the attribute of interest.

For the estimated variance (approximate), we have

$$v(p) = \frac{\Sigma w_i^2 + p^2 \Sigma y_i^2 - 2p \Sigma w_i y_i}{N n y^2}$$

where $y = \dfrac{y}{n}$ is the average number of children in a cluster and the fpc is ignored.

An example of the ratio esitmate of a proportion is the proportion of the 1-year-old population with three DTP inoculations and three oral poliomyelitis vaccine doses. This ratio is $p = 114/145 = 0.786$, where $n = 250$. With $N = 5{,}610$ and $y = 0.58$, the estimated variance of the proportion is 0.000056325 with a 95 percent confidence interval of 0.786 ± 0.015. In the estimation of proportions and their variances, all statewide estimates will have $N = 5{,}610$ and $n = 250$.

Comments

Our search of some 3,000 medical and health-related publications (*7*) revealed a dearth of information concerning applied survey methodology. With respect to immunization surveys per se, the only methodology recently reported in the United States was that developed by the Tennessee Department of Public Health for sampling 2-year-old populations by use of birth certificate information (*8*). Although that methodology has proved satisfactory in North Carolina, it is less applicable to older populations and, in any event, could not be applied to an institutional population.

Although there were some procedural problems in our survey of the day-care population, particularly with respect to the recordkeeping instruments and practices encountered in the facilities, the sample design was considered highly efficient in terms of the reliability of survey estimates (sampling error). The design was simple and economical to execute—the survey was conducted by six fieldworkers in approximately 6 weeks. The design also offered the computational ease of a self-weighted sample. The only problem encountered was the classic one of nonresponse, which might have been eliminated entirely had there been an earlier completion date for the survey.

References

1. North Carolina State Board of Health: North Carolina immunization level survey of two-year-old children. Division of Epidemiology, Immunization Program, Raleigh, December 1972.
2. North Carolina Department of Human Resources: North Carolina immunization level survey of two-year-old children. Division of Health Services, Epidemiology Section, Communicable Disease Control Branch, Raleigh, September 1974.
3. North Carolina Department of Human Resources: North Carolina immunization level survey of first grade children (public schools). Division of Health Services, Epidemiology Section, Immunization Program, Raleigh, Oct. 29, 1973.
4. North Carolina Department of Human Resources: North Carolina kindergarten and first grade immunization census report (public and private schools). Division of Health Services, Epidemiology Section, Immunization Program, Raleigh, Jan. 31, 1975.
5. North Carolina State Board of Health: Immunization schedule for normal infants and children. Raleigh, undated.
6. Cochran, W. G.: Sampling techniques. John Wiley and Sons, Inc., New York, 1953, pp. 112, 115, 124–125.
7. National Library of Medicine: Medline, a computerized retrieval system for bibliographies indexed since January 1, 1974. Bethesda, Md., August 1975.
8. Hutcheson, R. H., Jr., and Barid, S. J.: An inexpensive methodology for immunization surveys. Public Health Rep 89: 315–319, July–August 1974.