Survey Methodology, December 1991
Vol. 17, No. 2, pp. 131–141
Statistics Canada

Combining Estimates from Surveys

PODURI S.R.S. RAO and I.M. SHIMIZU[1]

ABSTRACT

For estimating the proportion and total of an item for the present occasion, independent estimates at the current and previous occasions are combined through three different procedures. In the first one, trend over the occasions is utilized. For the second one, the One-Way Random Effects Model is employed. The third procedure uses the Empirical Bayes approach. All the three procedures are seen to perform better than the sample estimates obtained from the data of the current occasion alone. Advantages of these methods and their limitations are discussed. All the procedures are illustrated with the data from the National Health Discharge Survey.

KEY WORDS: Trend; Weighted least squares; Random effects; Improved estimation; Biases; Mean square errors.

1. INTRODUCTION

In several national surveys, independent samples are obtained at successive time periods. In this paper, information from the past surveys is utilized to improve the estimates for the current period. For the sake of illustration, we consider the National Health Discharge Survey (NHDS) in the U.S. In this survey, which has been recently redesigned, a three stage sampling design is used with geographical regions as the Primary Sampling Units (PSU's) at the first stage. Hospitals and discharges are selected at the second and third stages respectively. The survey collects information on various items of the patients like age, sex, racial characteristics, length of stay, diagnosis, and surgical and non-surgical procedures. The selected PSU's and hospitals remain in the study for a certain number of years. Independent samples of discharges are obtained every year from the selected hospitals. Shimizu (1987) presents further details of the redesign of the NHDS.

At present, for a given hospital, estimates of the proportions for the different items for the current year are obtained only from the data of this year. National estimates are obtained by suitably weighting these proportions with the reciprocals of the probabilities of selection of the hospitals and the PSU's. However, Bean (1987) found that for most of the items the estimates are somewhat correlated over the years. For the sake of illustration, sample proportions obtained from the NHDS for 1977-86 for Acute Myocardial Infraction (AMI) and Mental Disorders (MDS) are presented in Table 1 for three hospitals and they are exhibited in Figures 1 and 2. Examination of the proportions for these three and 17 more hospitals suggested that the inclusion of past information can increase the precision of the estimates for the current year.

It should be cautioned that the sample proportions in Table 1 or Figures 1 and 2 should not be used to make inferences regarding the increase or decrease of AMI or MDS in the entire population.

[1] Poduri S.R.S Rao, Department of Statistics, Hylan 703, University of Rochester, Rochester NY,14618 U.S.A., and I.M. Shimizu, National Center for Health Statistics, Office of Research and Methodology 1-68, 3700 East-West Highway, Hyattsville MD, 20782, U.S.A.

Table 1

Data from the National Health Discharge Survey for 1977-86
Sample totals and proportions for Acute Myocardial
Infraction (AMI) and Mental Disorders
(MDS) for three hospitals

Year	No. of discharges N	Sampled No. of discharges n	AMI		MDS	
			Total	Sample proportion	Total	Sample proportion
1977	9,416	276	5	.018	37	.13
1978	10,234	266	7	.026	24	.09
1979	9,354	294	9	.031	39	.13
1980	10,372	327	9	.028	41	.13
1981	10,712	342	8	.023	45	.13
1982	10,683	309	9	.029	43	.14
1983	10,935	360	7	.019	46	.15
1984	10,090	330	6	.018	50	.15
1985	10,431	297	8	.027	41	.14
1986	10,247	264	4	.015	35	.13
1977	6,720	474	9	.019	18	.04
1978	6,710	470	14	.030	25	.05
1979	6,970	495	8	.016	28	.06
1980	6,794	466	14	.030	29	.06
1981	7,055	486	9	.019	34	.07
1982	6,265	442	9	.020	24	.05
1983	6,234	442	10	.023	28	.06
1984	6,221	439	9	.021	15	.03
1985	6,063	375	8	.021	19	.05
1986	5,781	371	4	.011	12	.03
1977	6,400	606	21	.0347	41	.0677
1978	6,286	635	23	.0362	42	.0661
1979	6,494	554	12	.0217	27	.0487
1980	6,813	571	17	.0298	25	.0438
1981	7,430	729	14	.0192	32	.0439
1982	7,267	712	20	.0281	39	.0548
1983	7,110	694	23	.0331	43	.0620
1984	7,268	718	35	.0487	29	.0404
1985	6,716	657	19	.0289	45	.0685
1986	6,464	655	21	.0321	33	.0504

In this article, we examine three procedures for improving the estimates for a specified hospital by utilizing the information from the current and the previous years. In the first method, estimates of the proportions are obtained from the linear trend over the years and the Weighted Least Squares Method. If there is a significant positive or negative trend over the years, this method will have higher precision than the sample estimate of the current period. If the trend is not pronounced, the increase in precision will be negligible, as expected.

For the second procedure, the One-Way Random Effects Model with unequal variances is used to combine the information. Yates and Cochran (1938) and Cochran (1954) suggested this type of procedure for combining information from experiments conducted at different time periods and locations. While the Analysis of Variance (ANOVA) method had been used

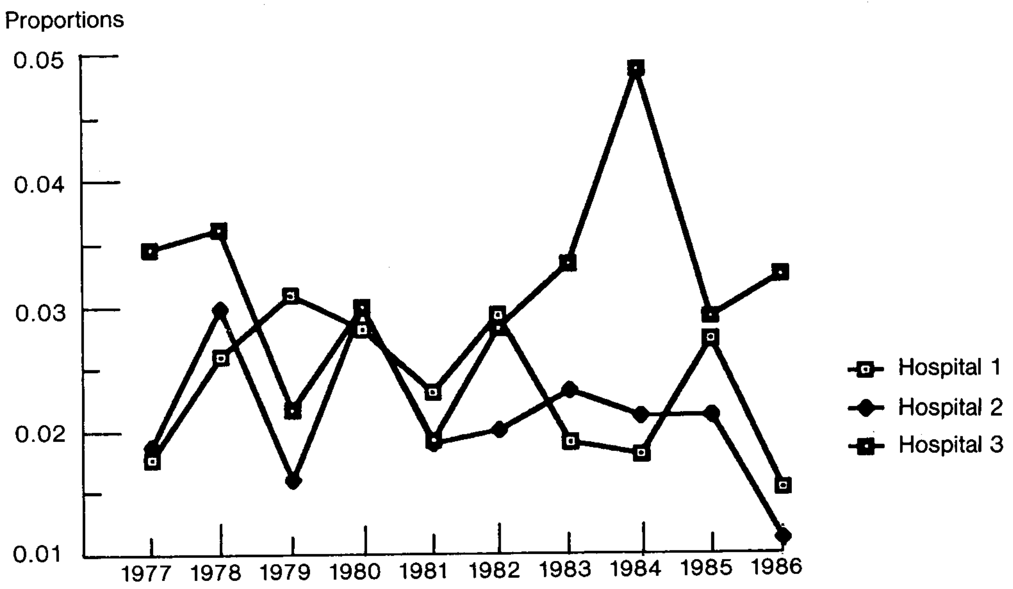

Figure 1. Proportions for AMI: 1977-86

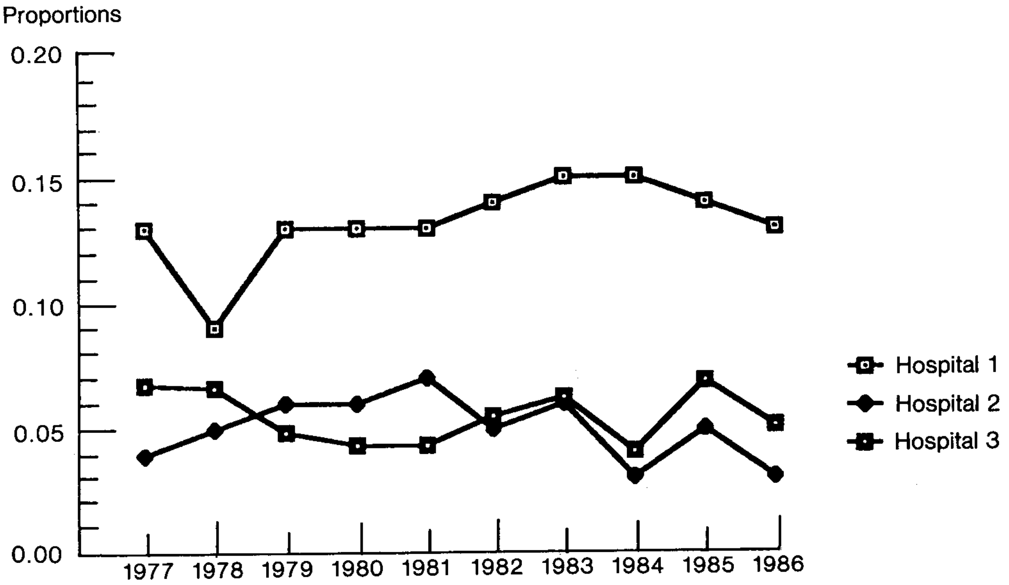

Figure 2. Proportions for MDS: 1977-86

for quite some time for this purpose, C.R. Rao (1970) suggested the Minimum Norm Quadratic Unbiased Estimation (MINQUE) and demonstrated its advantages. P.S.R.S. Rao, Kaplan and Cochran (1981) examined the relative merits of the ANOVA, MINQUE and several related procedures. We have employed the estimation procedures related to these methods. The estimate for the proportion obtained by any of these procedures is a weighted combination of the estimates of the different time periods. The weights depend on both the between and within variances of the time periods. In the third procedure, the Empirical Bayes approach is used to estimate the proportions for the current period.

We denote the above three procedures by TR, VC and EB respectively. The notation is presented in Section 2. The sample estimator for the proportion and its variance are given in Section 3. The above three estimation procedures along with the expressions for their Standard Errors (S.E.'s) are presented in Sections 4, 5 and 6. We have used these expressions to compute for 1986 the sample proportions, the above three types of estimates, and their S.E.'s for 20 hospitals in the NHDS. These estimates for the three hospitals mentioned earlier are presented in Table 2 for AMI and Table 3 for MDS. Results from the entire study are described in Section 7. The final section contains a discussion of the results and topics for further research.

Table 2

Estimates of the Proportions for 1986 and S.E.'s (bottom figures) for Acute Myocardial Infractions (AMI)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.0152 .0070	.0196 .0046	.0224 .0026	.0224 .0003
2	.0108 .0048	.0162 .0036	.0204 .0031	.0203 .0003
3	.0321 .0060	.0319 .0038	.0304 .0028	.0309 .0037

Table 3

Estimates of the Proportions for 1986 and S.E.'s (bottom figures) for Mental Disorders (MDS)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.1326 .0205	.1431 .0115	.1292 .0060	.1292 .0010
2	.0323 .0087	.0437 .0056	.0500 .0039	.0427 .0057
3	.0504 .0080	.0496 .0049	.0534 .0032	.0523 .0048

It should be mentioned that for the problem considered in this paper, the samples are drawn independently at the different time periods. Secondly, the population proportions for the previous periods are not known. Because of these reasons, the usual ratio and regression methods cannot be employed to improve the accuracy of the estimators for the current period. For the same reasons, the estimation procedures suggested in the literature for the rotation sampling schemes cannot be used in this situation. In spite of these difficulties, the three methods considered in this paper can be used to estimate the population quantities with a high accuracy. When summary figures at the different periods are available, public and private users can obtain these estimates and their standard errors without much difficulty. These procedures can also be used when there is nonresponse during some years - some of the hospitals do not provide information to the survey during some years.

2. NOTATION

We present in this section the notation for a selected PSU. Let y_{itj} denote the jth observation on the sampled discharge on an item like the number of surgical cases at time $t = (1, 2, \ldots, T)$, from the ith hospital, $i = (1, 2, \ldots, K)$, which has N_{it} discharges. Note that K may change over the years due to nonresponse or the addition of new hospitals.

The total and mean at time t are

$$Y_{it} = \sum_{1}^{N_{it}} y_{itj} \tag{1}$$

and

$$\bar{Y}_{it} = Y_{it}/N_{it}. \tag{2}$$

The total and mean of the sample of size n_{it} from the N_{it} discharges are

$$y_{it} = \sum_{1}^{n_{it}} y_{itj} \tag{3}$$

and

$$\bar{y}_{it} = y_{it}/n_{it}. \tag{4}$$

To estimate the total number and proportion for a specified item, let $y_{itj} = 1$ if the observation belongs to that item, and zero otherwise. With this notation, the total and proportion for an item at time t can be written as A_{it} and $P_{it} = A_{it}/N_{it}$. Note that P_{it} is the same as $\bar{Y}_{it}$.

In the following four sections, for the sake of convenience, we suppress the subscript i and describe the estimators for a given hospital.

3. SAMPLE PROPORTION

An unbiased estimator of the proportion P_t for an item like AMI or MDS is

$$\hat{P}_t = a_t/n_t, \tag{5}$$

where a_t is the number of cases of that item observed in the n_t sample discharges. The variance of $\hat{P}_t$ and its unbiased estimator are

$$V(\hat{P}_t) = \frac{N_t - n_t}{N_t - 1} \frac{P_t(1 - P_t)}{n_t} \tag{6}$$

and

$$v(\hat{P}_t) = (1 - f_t)\frac{\hat{P}_t(1 - \hat{P}_t)}{n_t - 1}, \tag{7}$$

where $f_t = n_t/N_t$. Note that $\hat{P}_t$ is the same as $\bar{y}_t = \sum_1^{n_t} y_{tj}/n_t$.

4. LINEAR TREND

The sample observations y_{tj}, $j = (1, 2, \ldots, n_{tj})$ can be written as

$$y_{tj} = \mu_t + \epsilon_{tj}, \tag{8}$$

where μ_t is the mean for the ith hospital at the tth period, and ϵ_{tj} is the random error with expectation zero and variance $\sigma_t^2 = P_t(1 - P_t)$. Since the samples are drawn independently during each year, the errors ϵ_{tj} are uncorrelated from one year to another.

With the assumption of a linear trend, the sample mean can be expressed as

$$\bar{y}_t = \alpha + \beta x_t + \bar{\epsilon}_t, \tag{9}$$

where $x_t = t$ and $\bar{\epsilon}_t = \sum_1^{n_t} \epsilon_{tj}/n_t$. Further, $V(\bar{\epsilon}_t) = (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = 1/W_t$. Note that with the zero-one notation, $\bar{y}_t$ is the same as $\hat{P}_t$. The WLS estimators of β and α are

$$\hat{\beta} = \frac{\sum W_t(x_t - \bar{x})\bar{y}_t}{\sum W_t(x_t - \bar{x})^2} \tag{10}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{11}$$

where $\bar{x} = \sum W_t x_t / \sum W_t$ and $\bar{y} = \sum W_t \bar{y}_t / \sum W_t$.

Estimator of μ_t is

$$\begin{aligned}\hat{\mu}_t &= \hat{\alpha} + \hat{\beta} x_t \\ &= \bar{y} + \hat{\beta}(x_t - \bar{x}).\end{aligned} \tag{12}$$

This is the Trend Estimator (TR). Estimators of this type for infinite populations have been examined in the literature; see Carroll and Rupert (1988), for instance.

We have obtained the estimate of μ_t from this expression by replacing W_t with $w_t = (1 - f_t)\,\hat{\sigma}_t^2/n_t$, where $\hat{\sigma}_t^2 = \hat{P}_t(1 - \hat{P}_t)$. If it can be assumed that for large N_t the distribution of y_{tj} is normal, $\bar{y}_t$ will be independent of w_t. In this case, the expression in (12) remains unbiased for μ_t. Even if the assumption of normality is not valid, it can be seen that w_t approaches W_t for large n_t and hence the expression in (12) with the estimated weights approaches μ_t.

The variance of the above estimator is

$$V(\hat{\mu}_t) = \frac{1}{\sum W_t} + \frac{(x_t - \bar{x})^2}{\sum W_t(x_t - \bar{x})^2}. \tag{13}$$

We have estimated this variance by replacing W_t by w_t. The bias in the resulting estimator will be small for large n_t.

For the illustration in this article, $t = (1, 2, \ldots, 10)$, that is, $T = 10$. For 1986, we have found the estimate for the proportion of an item and its S.E. from (12) and (13) with $x_t = 10$.

5. VARIANCE COMPONENTS MODEL

Examination of the proportions for the AMI and MDS of the 20 hospitals for the ten years showed no specific linear or nonlinear trend. For all of them the patterns somewhat resembled those of the three hospitals, presented in Figures 1 and 2. These observations indicated that the proportion for AMI or MDS for the current year can be obtained by combining the information from all the ten years. The One-Way Random Effects Model can be used for this purpose.

The model in (8) can be written as

$$\begin{aligned} y_{tj} &= \mu + (\mu_t - \mu) + \epsilon_{tj} \\ &= \mu + \alpha_t + \epsilon_{tj}. \end{aligned} \tag{14}$$

If μ_t is considered to be randomly drawn from a population with mean μ, the random effect α_t will have mean zero and variance σ_α^2. It is assumed to be independent of ϵ_{tj}. The sample mean (proportion) can now be written as

$$\bar{y}_t = \mu + \alpha_t + \bar{\epsilon}_t, \tag{15}$$

where $\bar{\epsilon}_t$ has mean zero and variance $(1 - f_t)\ \sigma_\alpha^2/n_t$. Thus, from (15),

$$V(\bar{y}_t) = \sigma_\alpha^2 + (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = \frac{1}{U_t}. \tag{16}$$

The WLS estimator of μ is

$$\hat{\mu} = \frac{\sum U_t \bar{y}_t}{\sum U_t}. \tag{17}$$

This is the Variance Components Estimator (VC) and its variance is

$$V(\hat{\mu}) = 1/\sum U_t. \tag{18}$$

For obtaining the mean in (17) and its variance in (18), we have replaced σ_t^2 by its estimate $\hat{P}_t(1 - \hat{P}_t)$. Procedures like the ANOVA and MINQUE are available for estimating σ_α^2. The MINQUE depends on the *a priori* values r_t of $(\sigma_t^2/\sigma_\alpha^2)$. A related procedure called the Unweighted Sums of Squares (USS) method does not depend on r_t and it is described below. P.S.R.S. Rao, Kaplan and Cochran (1981) found that this method provides estimates for σ_α^2 comparable to the ANOVA and MINQUE, unless n_t or r_t is very small. The USS is computationally less cumbersome than the MINQUE. With $\bar{y}^* = (\sum \bar{y}_t)/T$, from (15),

$$E[\sum (\bar{y}_t - \bar{y}^*)^2] = (T - 1)\sigma_\alpha^2 + (T - 1)(\sum v_t)/T, \tag{19}$$

where $v_t = (N_t - n_t)P_t(1 - P_t)/(N_t - 1)n_t$. The USS estimator for σ_α^2 is

$$\hat{\sigma}_\alpha^2 = \sum (\bar{y}_t - \bar{y}^*)^2/(T - 1) - (\sum \hat{v}_t)/T, \tag{20}$$

where $\hat{v}_t = (1 - f_t)\hat{P}_t(1 - \hat{P}_t)/(n_t - 1)$. If N_t is large relative to n_t, the sampling fraction f_t can be set to zero. We have estimated U_t from (16) by estimating σ_α^2 from (20) and the second term by $\hat{v}_t$. Utilizing this estimate of U_t, we have estimated μ from (17) and its variance from (18). If σ_α^2 is much larger than v_t, the estimator $\hat{\mu}$ in (17) will be close to $\bar{y}^*$. In this case, estimation of U_t as described above can be expected to have almost no effect on $\hat{\mu}$. Since $\hat{\mu}$ depends only on the relative values of U_t, this conclusion can be expected to be valid even when σ_α^2 is not considerably larger than v_t. Thus, estimation of U_t can be expected to result in only a negligible bias for $\hat{\mu}$.

As is well-known, all the procedures for estimating σ_α^2 unbiasedly can result in negative estimates. In such a case, we have employed the usual practice of substituting a small positive quantity for the negative estimate. In Rao *et al.* (1981) it was found that unless σ_α^2 is very small, this adjustment results in only a negligible bias for $\hat{\sigma}_\alpha^2$ and an insignificant increase in its standard error. Further, unless σ_α^2 is small, the difference in the MSE of $\hat{\mu}$ for the USS and other methods of estimating U_t was found to be negligible.

6. BAYES' ESTIMATOR

The discussion in the beginning of Section (5) suggests that μ_t can be assumed to have a prior distribution with mean μ and variance σ_α^2. With the assumptions that for large N_t the distribution of y_{tj} is normal with mean μ_t and variance σ_t^2, and that the prior distribution of μ_t is also normal, the Bayes' Estimator for μ_t is

$$B_t = E(m_t \mid \bar{y}_t) = (1 - a_t)\bar{y}_t + a_t\mu, \tag{21}$$

where $a_t = v_t/(\sigma_\alpha^2 + v_t)$. The expression for v_t is the same as given in the previous section.

For given $\bar{y}_t$, the variance of the above estimator is

$$V(B_t) = \frac{1}{(1/\sigma_\alpha^2) + (1/v_t)}. \qquad (22)$$

With estimates $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_t^2$ and $\hat{\mu}$, the expression in (21) can be written as

$$\hat{B}_t = (1 - \hat{a}_t)\,\bar{y}_t + \hat{a}_t\hat{\mu}, \qquad (23)$$

where $\hat{a}_t = \hat{v}_t/(\hat{\sigma}_\alpha^2 + \hat{v}_t)$. This estimator may be called the Empirical Bayes' estimator (EB). Note that $\hat{\mu}$ is obtained from (17) with $\hat{\sigma}_\alpha^2$ and $\hat{v}_t$. The variance of this estimator may be obtained from (22) by replacing σ_α^2 and v_t with their estimates. For obtaining the EB and its variance, we have estimated σ_α^2 and v_t from the USS procedure described in the previous section.

7. PERFORMANCE OF THE ESTIMATORS

We have computed the estimates of P_t for 1986 for the 20 hospitals through the different procedures described in the previous sections. Since the population values of P_t are not known, as described earlier, we have found the S.E.'s for the different procedures by substituting the sample proportion $\hat{P}_t$ in the place of P_t. Since the sample sizes n_t are not small, the resulting biases in estimating the variances or S.E.'s of the estimators can be expected to be small.

For the three hospitals, the estimates of P_t and the S.E.'s of the different procedures are presented in Tables 2 and 3 for AMI and MDS respectively.

As can be seen from these tables, S.E.'s of TR, VC and EB are smaller than the S.E. of the sample proportion. As expected, utilizing the data from the previous periods has helped reduce the S.E. of the estimate for the current period.

Both VC and EB have smaller S.E.'s than TR. However, TR does not require the estimation of σ_α^2. We have found the S.E. of TR to be usually less than 50 percent of the sample proportion.

The EB has smaller S.E. than VC, as expected. Note that VC estimates the overall proportion, whereas EB estimates the proportion of the conditional distribution. The S.E. of the EB becomes close to that of the sample proportion if the sample size is large.

It is interesting to observe from Tables 2 and 3 that for both AMI and MDS the difference between the VC and EB estimates is negligible. The reason for this result is that $\hat{a}_t$ is close to unity, which indicates that σ_α^2 is small relative to v_t.

The estimates for the total number of cases for 1986 and their S.E.'s can be obtained by multiplying the estimates of the proportions in Tables 2 and 3 by the corresponding number of discharges N_t given in Table 1.

8. DISCUSSION

As described in the above section, the results of this investigation recommend the TR, VC or EB methods for estimating the proportions and totals for the current period.

For estimating the S.E.'s of the different procedures, we have utilized the sample proportions. Further investigation is needed to examine the biases and MSE's of these S.E's.

For estimating σ_α^2 and v_t, we have employed the USS. The effects of the ANOVA and the MINQUE procedures for this purpose can also be examined. However, the investigation in Rao *et al.* (1981) showed that different procedures of estimating σ_α^2 may not have a significant effect on the estimation of μ or its S.E.

Further investigation is needed to determine the effect of the different procedures of estimating the variances on the EB for μ_t.

We have substituted a small positive quantity for a negative estimate of σ_α^2. As can be seen, this adjustment may result in a small S.E. for both the VC and EB, and may present too optimistic a view about the estimates of μ and μ_t. Further examination of this problem is needed.

We have assumed a linear model for the proportion. The logit or probit transformation can be used before using this model. However, large population and sample sizes are needed to justify the estimates that can be obtained through these transformations. The estimates proposed in this article can be obtained by the public and private users by using any simple computer program.

Improved estimates for each hospital are considered in this paper. The national estimates for a given item like AMI or MDS can be obtained by suitably weighting the above estimates by the reciprocals of the probabilities with which the hospitals were selected. Such a procedure is expected to improve the precision of the national estimates.

Time series methods like the ARIMA can be used as suggested for instance by Blight and Scott (1973) and Scott and Smith (1977) for estimating the proportions and total numbers. These methods will result in different models for different items. Secondly, the available package programs for these approaches assume large population sizes and equal error variances, and the same sample sizes for all the time periods. Such assumptions are not satisfied for the problem we have considered in this article. As mentioned in Section 1, the TR,VC and EB methods can also be used when there is nonresponse during some years.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and a referee for their helpful comments.

REFERENCES

BEAN, J.A. (1987). NHDS variance and covariance estimation of year to year differences. National Center for Health Statistics, research report.

BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

CARROLL, R.J., and RUPERT, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

RAO, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.

RAO, P.S.R.S., KAPLAN, J., and COCHRAN, W.G. (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76, 89-96.

SCOTT, A.J., and SMITH, T.M.F. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

SHIMIZU, I.M. (1987). Specifications for the redesigned NHDS sample. National Center for Health Statistics, research report.

YATES, F., and COCHRAN, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Sciences*, 28, 556-580.