ALLOCATION USING EMPIRICAL BAYES ESTIMATORS

Harry M. Marks, U.S. Department of Agriculture
Stephen M. Woodruff, Bureau of Labor Statistics

I. INTRODUCTION

The allocation problem in survey design as usually stated is the problem of allocating a sample over a set of strata, in order to minimize a variance expression, subject to a constraint on cost or other considerations. Many times the problem stated in this way is incomplete, because the variances used are estimated, and thus are subject to error. If the estimates are of poor reliability, then the knowledge or the amount of information known about the variances is not greatly increased over that of no information. Thus, the allocation of the sample over the strata should be "close" to that of the allocation over strata that would be if no information was available.

In many surveys, more than one characteristic is of interest. The allocation that is optimal for one of the characteristics may not be optimal for the other characteristics. Of importance are the probabilities that an allocation will yield an estimate with reliability better than a given reliability. The allocation problem can be viewed as operating on a function of these probabilities, a function which considers the desires of the sponsor of the survey.

In this paper we estimate the posterior variance, using empirical Bayesian estimation techniques, under an assumed underlying population distribution for variances. The estimated variance is a weighted sum of the usual unbiased estimate of variance, and a "prior" estimate of variance based on a prior assumption of the relationships between the variances over the strata. These "prior estimates are derived using the observed data. We then estimate the posterior probability distribution empirically.

Let j be the index for a stratum; j=1,...J, and let n_j be the sample size in stratum j. Let x_j be the estimator for the j^{th} stratum, with variance= σ_j^2/n_j. The allocation problem is to find $n_1,...n_J$ such that the variance of $x = \sum_{i=1}^{J} W_j x_j$, where the W_j are fixed, is minimized, subject to a contraint $n = \sum_j n_j$. Suppose a simple random sample without replacement in each stratum. The variance of x can then be written as:

$$(1) \qquad V_x = \sum_j W_j^2 (1-f_j) \sigma_j^2/n_j$$

where f_j is the finite population correction factor, and σ_j^2 is the "unit" variance for the j^{th} stratum, j=1,...,J.

Data, from a past survey is available, and an estimator of σ_j^2, Z_j is used. Suppose that $p(Z_j/\sigma_j^2)$ is the conditional probability function of Z_j given σ_j^2. Further suppose that $p(\sigma_j^2)$ is the proper or improper prior distribution function of σ_j^2. Then the posterior distribution of σ_j^2 given a realized value of Z_j is:

$$(2) \qquad P(\sigma_j^2/Z_j) \;\alpha\; P(Z_j/\sigma_j^2)\, P(\sigma_j^2)$$

The expected value of σ_j^2 given Z_j is:

$$(3) \qquad \sigma_{j,B}^2 = \frac{\int \sigma_j^2\, p(Z_j/\sigma_j^2)\, P(\sigma_j^2)\, d\sigma_j^2}{\int p(Z_j/\sigma_j^2)\, p(\sigma_j^2)\, d\sigma_j^2}$$

which we shall take to be the estimator of σ_j^2.

At this point, many times the complete determination of $\sigma_{j,B}^2$ is not possible, because the prior distribution $p(\sigma_j^2)$ depends upon unknown parameters, say θ. If θ is not dependent upon j, then by considering the variables, $Z_1,...,Z_J$, one can estimate θ using any of the sampling or bayesian estimation procedures. The likelihood of θ, given $Z_1,...,Z_J$ is:

$$(4) \qquad L(\theta/Z_1,...,Z_J) \;\alpha\; \prod_{j=1} p(Z_j/\sigma_j^2)\; p(\sigma_j^2)$$

Once an estimate of θ is determined, this can be inserted into equation (3) to derive the estimate of σ_j^2, and to determine the posterior distribution of σ_j^2 given $Z_1,...,Z_J$. The estimator σ_{jB}^2, determined in this fashion is termed the "empirical" Bayes estimator of σ_j^2 and the probability function of the posterior distribution $p(\sigma^2/Z_j)$ will be termed the "empirical" p.d.f.

We can now view (1) as a sum of random variables σ_j^2, with known constant $W_j^2 (1-f_j)/n_j$ as weights, where the designer has "control over the values $(1-f_j)/n_j$. Thus the p.d.f. of V_x can be estimated using the empirical p.d.f. of the σ_j^2, and the probabilities.

(5) $\phi_x (V_{ox}) = P_r (V < V_{ox})$

where V_{ox} is a constant, as a function of $n_1 \ldots, n_J$. If $X_1, \ldots, X_k$ are K different characteristics each for which a sponsor wants all estimates to be equally reliable, i.e., V_{ox} are all equal, then the choice of $n_1, \ldots, n_J$ to be made might be that which minimizes a maximum of a function which measures the variability of the probabilities.

The Bayesian method of deriving optimal designs has been discussed by various authors. See article by Ericson and Rao. The method of estimating the parameters of the prior distribution will depend much on the prior belief of certain relationships. Without prior beliefs, then an empirical approach will depend upon data analysis on the raw observations. Many articles on "Empirical Bayes" methods of estimation have been written. See particularly an article by Enfron and Morris, 1975 in Journal of the American Statistical Association (JASA).[3]

II. ESTIMATION

Let s_j^2 be the sample sum of squares about the mean associated with the j^{th} stratum, and V_j the degrees of freedom of s_j^2. Let $Z_j = s_j^2/V_j$, and assume that

(6) $$p(Z_j/\sigma_j^2) = \frac{1}{\Gamma(b_j)} \left(\frac{b_j}{\sigma_j^2}\right)^{b_j} Z_j^{\;b_j - 1} \cdot \exp(-b_j Z_j/\sigma_j^2)$$

where $b_j > 0$, This is a gamma distribution and

(7) $$E(Z_j/\sigma_j^2) = \sigma_j^2$$
$$Var(Z_j/\sigma_j^2) = \sigma_j^4/b_j$$

Thus, b_j can be thought of as a generalization of the degrees of freedom i.e., if (s_j^2/σ_j^2) was a chi-square with v_j degrees of freedom, then $b_j = v_j/2$. We shall assume that b_j can be estimated, using the raw data, or using a replication technique.

For the computations done here we took $b_j = \frac{1}{2} V_j (1 + \lambda_4/2)$ where λ_4 is the standardized measure of the kurtosis, i.e., the fourth cumulant divided by the square of the variance. This is the first term of the Roy-Tika approximation of the distribution of Z_j using gamma functions and Laguerre polynomials.[4]

We shall take as a prior distribution of σ_j^2 the natural conjugate, the inverted gamma with parameters a_j and v, i.e.,

(8) $$p(\sigma_j^2/a_j, v) = \frac{a_j^{\,v-1} (\sigma_j^2)^{-v}}{\Gamma(v-1)} \exp\left(-\frac{a_j}{\sigma_j^2}\right)$$
$$v > 1.$$

Thus, the posterior distribution is:

(9) $$p(\sigma_j^2/Z_j, a_j, v) \,\alpha\, (\sigma_j^2)^{-(b_j + v)} Z_j^{(b_j - 1)} \cdot a_j^{\,v-1} \exp\left(-\frac{a_j + b_j Z_j}{\sigma_j^2}\right)$$

which is an inverted gamma distribution. Hence

(10) $$E(\;_j^2/Z_j, a_j, v) = (a_j + b_j Z_j)/(b_j + v-2)$$
$$Var(\sigma_j^2/Z_j, a_j, v) = \left(\frac{a_j + b_j z_j}{b_j + v-2}\right)^2 \cdot \frac{1}{b_j + v-3}$$

The expected value can be seen to be weighted average of the unbiased sample estimate, Z_j, and the "prior" expected value of $\sigma_j^2 = \frac{a_j}{v-2}$. Moreover, the rel-variance of the posterior distribution = $(b_j + v-3)$ for $v > 3$ is smaller than the rel-variance of (Z_j/σ_j^2).

Taking the product of (6) and (7), and intergrating with respect to σ_j^2, we have that the p.d.f. of Z_j given b_j, a_j and V is:

(11) $$p(Z_j/b_j, a_j, v) \,\alpha Z_j^{\,b_j - 1} a_j^{\,v-1} / (b_j Z_j + a_j)^{(b_j + v-1)}$$

This is of the form of an inverted beta distribution. The expected value and variance are:

(12) $$E(Z_j/b_j, a_j, v) = a_j/v-2$$
$$Var(Z_j/b_j, a_j, v) = (a_j/v-2)^2 \frac{(b_j + v-2)}{b_j\;(v-3)}$$

This implies that the rel-variance $(b_j + V-2)\; b_j\, (v-3)$ is independent of $E(Z_j)$. An estimate of this quantity is:

(13) $$R_j^2 = (Z_j - \hat{Z}_j)^2 / \hat{Z}_j^2$$

where $\hat{Z}_j$ is an estimate of the expected value of Z_j. Let R^2 be the average of R_j^2, so that we have

(14) $$R^2 = \frac{1}{J} \sum_{j=1}^{J} R_j^2$$
$$= \frac{1}{v-3} \left(1 + \frac{V-2}{J} \sum_{j=1}^{J} 1/b_j\right)$$

From this an estimate of V can be derived, i.e.

$$(15) \qquad V\text{-}2 = (R^2 + 1) / (R^2 - \Pi b).$$

$$\text{where } \Pi b = \frac{1}{J} \sum_{j=1}^{J} 1/b_j.$$

We are considering maximum likelihood estimates, which would involve iterative solutions, where the initial values are those derived from the regression and equation 15. If time permits, an effort will be put forth to derive a maximum likelihood solution.

Let $\hat{a}$ and $\hat{v}$ be the estimates of a and v. The empirical Bayes estimate of σ_j^2 can be written.

$$(16) \qquad \sigma_j^2\text{, Bayes} = \lambda_j Z_j + (1 - \lambda_j)\, \hat{a}/\hat{v}\text{-}2, \text{ where}$$

$$(17) \qquad \lambda_j = (b_j \hat{R}_j^2 - 1) / (b_j + 1)\, \hat{R}_j^2 \text{ and}$$

where

$$R_j^2 = \frac{1}{V-3} \left(1 + \frac{\hat{V}-2}{b_j}\right)$$

Note that

$\lim \lambda j = 1$

$\hat{v} \rightarrow 2^+$

$\lim \lambda_j = 1$

$b_j \rightarrow \infty$

Thus small values of $\hat{V}$ and large b_j give more weight to Z_j, the unbiased estimates of σ_j^2, as should be the case since $\hat{V}$ is a measure of the precision of the "prior" distribtuion and b_j is a measure of the relative precision of Z_j.

From equation (14), we get the empirical Bayes estimates of V_x, i.e.,

$$(18) \qquad V_x \text{ (Bayes)} = \Sigma w_j^2 (1-f_j)\, \sigma_j^2\text{, Bayes} / n_j$$

Minimizing this, subject to $\sum_j n_j = n$, we have to solve an equation using iteration, in order that $n_j < N_j$, the universe size for the j^{th} stratum. Assuming an infinite size universe, we have $n_j \,\alpha\, w_j \hat{\sigma}_j$.

This shows that the value of n_j is a weighted sum of the usual Neyman allocation, say n_j (Z) and the allocation using $\hat{a}$, i.e. proportional allocation say $n_j(p)$. That is, we can write:

$$n_j = F_j n_j \cdot (Z) + G_j n_j (p)$$

where F_j and G_j are numbers, which, however, do not add to 1 for a given j.

We can express V as a sum, $V_x = \Sigma C_j V_j'$ where C_j is a constant, a function of n_j, and $V_j' = \sigma_j^2$. We shall assume that V_j' are independent, and distributed with posterior distribution p (σ^2/z_j). In the case we are concerned with, V_j' is distributed as an inverted gamma, with parameters, $(a_j + b\, z_j) = d_j$, and $(b_j + v) = g_j$, i.e.

$$(19) \qquad p(V_j') \;\alpha\; (V_j')^{-g_j} \exp(-d_j/V_j')$$

Sufficiently high moments of V_j' do not exit i.e. the r^{th} moment about zero of V_j' is:

$$(20) \qquad \mu_r' = d_j^2/(g_j - 2)(g_j - 3) \ldots \ldots (g_j - (r + 1))$$

valid for $r < g_j - 1$. The variance of V_j' is:

$$(21) \qquad \mu_2 = d_j/(g_j - 2)^2 (g_j - 3).$$

Hence the ratio,

$$(22) \qquad \mu_r'/\mu_2^2 = 0(g_j^{1/2}), \quad r < g_j - 1$$

If g_j is large enough, so that fourth moments exist, and are measured with a certain degree of reliability, then by computing the first four moments, one can estimate the p.d.f., by determining which of the "Pearson" type of p.d.f. best fits the p.d.f. of V_x. We note that V_j' is a type V as classified by Kendall VOL. 1 "The Advanced Theory of Statistics".[5]

It is not suggested the estimates of variances be based on Bayesian methods, but that only these methods be used for the purposes of allocating a sample and for predicting the unknown variances, (the result of the next period survey). However, when smoothing, or generating variances, the empirical Bayes estimation should be considered, because of the properties associated with them (i.e. the admissibility or the decrease in the MSE, under certain conditions.) The form of the estimator need not be exactly like (14) but in many cases will end up to be a weighted average between the prior and "unbiased" estimates, where the weights are inversely proportional to the variance of the estimates.

If one approaches the problem from the sampling theory point of view then by minimizing the MSE, one derives a similar kind of result, with a slight shrinkage toward the unbiased estimator, due to the correlation that exists between the estimates of the prior and unbiased estimates.[6]

III. An Empirical Study

Data for a one month period (March, 1978) from the Labor Turnover Survey conducted by the Bureau of Labor Statistics was available. This survey measures the rate of new hires and separations, on a monthly basis, for establishments classified in the manufacturing standard industrial classification (SIC) scheme. Approximately 36,000 establishments are sampled each month, at the national level. The statistic of interest is the number of turnovers.

For our study, we considered the allocation of the sample for each of seven two digit SIC classes, for which a sufficient sample size for our purpose was available. The strata that are commonly used by the Bureau for both internal analysis, and publication are "size classes", where size is the number of employees in the establishment; and ranges from 1-3; 4-9; 10-19; 20-49; 50-99; 100-249; 250-499; 500-999; 1000 +; 9 strata in all. The estimate under consideration can be written:

$$(1) \qquad r = \sum_j E_j (X_j/y_j)$$

where E_j is the "benchmark" employment for the j^{th} size class, X_j is the estimated number of new hires, and y_j is the estimated number of employees in the j^{th} size class. By linearizing the estimate, we can express the variance by

$$(2) \quad \text{var}(r) = \sum_j E_j^2 \,\text{Var}(X_j/y_j)$$
$$= \Sigma E_j^2 (1-f_j)\, \sigma_j^2/n_j$$

where n_j is the number of establishments in the sample in the j^{th} size class, σ_j^2 is the "unit" variance of the "linearized value" for the establishments in the j^{th} size class, and f_j is the finite correction factor (we are assuming a simple random sample without replacement). An estimate of σ_j^2 was made by computing

$$(3) \quad \hat{\sigma}_j^2 = \frac{1}{\bar{y}_j^{\,2}} \sum_{i=1}^{n_j} (X_{ij} - r_j y_{ij})^2/(n_j - 1)$$

where x_{ij}, y_{ij} are the new hires and number of employees for the i^{th} establishment, $r_j = x_j/y_j$ and $\bar{y}_j$ the average number of employees per establishment.

On examining the data, we noticed that $\hat{\sigma}_j^2$ was correlated with j. In fact, when averaging $\hat{\sigma}_j^2$ over the SIC's, for a given j, and then taking this average and regressing against j, a simple linear regression resulted, with a correlation of -0.95. In view of this, we assumed apriori that "a_j" of equation (8) of section II, was linearly related with j, i.e.

$$(4) \qquad \left(\frac{a_j}{v-2}\right) = a + b\, j$$

The sample was randomly divided into five subsamples, by SIC and size class. Each subsample was used to derive an allocation by the "empirical Bayes" method of allocation, outlined in Section II, for the Neyman allocation, and for an allocation using the regressed value as the estimate of the variance described below. Thus, for each SIC we have five comparisons; and with seven SICs, this gave us 35 comparisons.

The values of α, β and V (equation 8 of Section I) were computed by regressing log (Z_j), on j, using a non-parametric method.(7) We determined all possible slopes between all the pairs of points, and used the median of these to be the slope and we had the regression line pass through the medians. Then an estimate of V was computed from equation (15) of Section I.

IV. RESULTS

We computed variances for the optimal allocations for a 20 percent, 10 percent and 5 percent sample using the three different methods of estimating variances. We found that the Bayes method of estimating variances was the best of the three methods. Though, it did not give the allocation with the minimum variance in 60 percent of the cases, it was at least the second best of the three methods in all cases. For the Neyman optimal allocation, and the optimal allocation using the regressed values of variances, the regression method was slightly better, but both exhibited a tendency for wide variation, even within the same SIC.

For each sample size, and for each case (35 in all), we ranked the variances the lowest getting a rank of one, the highest a rank of three. The results can be summarized in the chart below:

20% Sample	Neyman	Regression	Bayes
Number of 3's	24	11	0
Number of 2's	6	8	21
Number of 1's	5	16	14
Sum of Ranks	89	65	56

10% Sample	Neyman	Regression	Bayes
Number of 3's	20	15	0
Number of 2's	8	6	21
Number of 1's	5	16	14
Sum of Ranks	81	73	56

5% Sample	Neyman	Regression	Bayes
Number of 3's	19	16	0
Number of 2's	8	5	22
Number of 1's	8	14	13
Sum of Ranks	81	72	57

With regards to the probability distribtuion based on the empirically Bayes estimation, we computed the first four moments of V_x, using equation (20), and using the fact that the summands V_i are independent. From this we classified the curve according to one of the Pearson types. The Pearson type that seemed to described the distribtuion was the Pearson type IV as classified by Kendall and Stuart.(5) More research is needed in this area.

As suggested during the meetings we are providing the following table. The entries are the average differences between the variance computed for a given type of allocation and the optimal variance divided by the optimal variance for the different sample sizes. That is, we computed the average over the 35 cases of the quantity.

(Variance(type)—Optimal)/Optimal

where type = Bayes, Regression or Neyman.

	Bayes	Regression	Neyman
20%	.05341	.04804	.11887
10%	.0499	.05487	.09955
5%	.04376	.04752	.08636

It is apparent that Bayes and Regression are clearly superior to Neyman and that Bayes is slightly superior to Regression except in the 20% case.

V. CONCLUSION

Other methods and distributional assumptions can be considered, and we hope that more research is done concerning the questions that we have addressed. The important thing in the allocation problem is really the ratio of variances, so an approach estimating these ratios, with some prior estimates might be fruitful. Further, the above concepts can be extended to more complex designs.

It seems to us that the empirical Bayes method of estimation has built into it certain "safety" features, which allows one to calculate the degree in which an estimate yields information. Thus, we believe such methods, even if they do not yield the optimal, are more sturdy and consistent on repeated uses. This may be, since we are using more information from the available data.

We would like to express our thanks to the staff of the Bureau of Labor statistics, for all their help in the typing of the paper, with particular appreciation to Anna Taylor and Alfreda Reeves who typed the final paper.

REFERENCES

1. Rao, J.N.K. and G.Hangurde P.D., "Bayesian Optimization in Sampling Finite Populations", Journal of the American Statistical Association, 67, June 1972, pg. 439-443.

2. Ericson, W.A., "Optimum Stratified Sampling Using Prior Information", Journal of the American Statistical Association, 60, September 1965, pg. 750-771.

   "Optimum Allocation in Stratified and Multi-Stage Samples Using Prior Information", Journal of the American Statistical Association, Vol. 63, September 1968, pg. 964-982.

3. Efron, B. and Morris C., "Data Analysis Using Stein's Estimator", Journal of the American Statistical Association, Vol. 70, June 1975, pg. 311-319.

4. Tan W.Y. and Wong S.P., "On the Roy-Tika Approximation to the Distribution of the Sample Variances from Non-Normal Universes" Journal of the American Statistical Association, 72, December 1977, pg. 875-880.

5. Kendall, Maurice G., "The Advanced Theory of Statistics", Vol. I Charles Griffin & Co., Ltd., 1969, Chapter 6.

6. Marks, Harry, "Composite Estimation Techniques, used for the Consumer Price Index Revision (CPIR) Weights", American Statistical Association 1978 Proceedings of Section on Survey Research Methods.

7. Hollander, Myles and Wolfe, Douglas A., "Nonparametric Statistical Methods", John Wiley and Sons, 1973.