Statistics Canada
Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

WORKING PAPER NO. BSMD-88-027E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-88-027E

DIRECTION DE LA MÉTHODOLOGIE

A NON-PARAMETRIC EMPIRICAL BAYES APPROACH FOR ESTIMATING A PROCESS AVERAGE IN QUALITY CONTROL

by

J.H. MacMillan and W.V. Mudryk

August 1988

UNE APPROCHE EMPIRIQUE NON-PARAMÉTRIQUE "BAYES" POUR ESTIMER UNE MOYENNE DE TRAITEMENT DU CONTRÔLE DE LA QUALITÉ

RÉSUMÉ

À Statistique Canada, l'échantillonnage d'acceptation est utilisé à titre de méthode de contrôle de la qualité lors des opérations de traitements d'enquête. Les plans d'échantillonnage qui sont utilisés assurent un minimum d'inspection à un niveau spécifique d'erreur à l'entrée. Ce niveau d'erreur est estimé par une quantité connue sous le nom de moyenne de traitement. C'est un paramètre inconnu qui est habituellement estimé à partir de résultats d'inspection courantes, mais l'estimation est souvent difficile à réaliser à cause des petites tailles d'échantillons. Une plus grande précision de l'estimation peut être produite en utilisant plus de données des échantillons précédents afin d'améliorer le résultat de l'échantillon courant. Un estimateur empirique non-paramétrique "Bayes" de la moyenne de traitement est présenté. Un intervalle de confiance approximatif est aussi construit. Des exemples sont donnés.

Mots Clés: contraction, moyenne combinée, meilleur prédicteur linéaire sans biais.

A NON-PARAMETRIC EMPIRICAL BAYES APPROACH FOR ESTIMATING A PROCESS AVERAGE IN QUALITY CONTROL

J.H. MacMillan and W.V. Mudryk, Statistics Canada
J.H. MacMillan, 11D R.H. Coats Building, Tunney's Pasture
Ottawa, Ontario, K1A 0T6

ABSTRACT

At Statistics Canada, acceptance sampling is used as a method of quality control for survey processing operations. The sampling plans which are used will ensure minimum inspection at a specific incoming error level. This error level is estimated by a quantity known as the process average. It is an unknown parameter which is usually estimated from current inspection results, but frequently the estimation is difficult because of small sample sizes. Greater accuracy in the estimate may be produced by using more data from previous samples to improve upon the current sample result. A non-parametric empirical Bayes estimator of the process average is presented. An approximate confidence interval is also constructed. Examples are provided.

KEY WORDS: shrinkage, pooled average, best linear unbiased predictor

1. INTRODUCTION

Survey processing operations at Statistics Canada include processes such as data capture, coding, editing, transcription, assessment, corrections etc. Because this processing is largely manual and repetitive in nature, it is often subject to high levels of error. Quality control methods, applied at the individual producer (i.e. in this case, operator) level are used to control these errors to acceptable levels.

Acceptance Sampling has been found to be an effective method for controlling the quality of survey processing operations at Statistics Canada (Mudryk, 1988). Typically Dodge-Romig sampling plans are used, which provide average quality protection and minimum inspection for each operator at a specific incoming error level. This level is estimated by the process average which is defined as an individual's underlying error level at a specific point in time. For most applications it is expressed as a fraction or percent defective. The true process average is unknown and it is therefore desirable to estimate it as accurately as possible since inspection will only be minimized at the true value.

Estimation using only current data often yields inaccurate results because sample sizes are small. This is improved somewhat by including data from other recent processing periods using the Quality Control Processing System (QCPS), (Mudryk, 1988). The QCPS is quality control software developed at Statistics Canada which maintains quality control data for a four month period for each individual operator. The previous data on the system's historical files is combined with the current sample data using a non-parametric empirical Bayes approach in an attempt to produce a more accurate estimate of an operator's process average. This estimate is then used to select more efficient sampling plans for the operator.

2. EMPIRICAL BAYES METHODS

The notion of using a Bayesian approach to estimate the process average was prompted by the work of Hoadley (1979 and 1981) at Bell Labs. He was able to show an increase in the accuracy of estimates required for quality assurance, by using an empirical Bayes technique.

In empirical Bayes inference the values of the hyper-parameters of the prior distribution are unknown and are estimated from the marginal distributions of the observed data. Hoadley's model specifies a Poisson likelihood for the observed data, and then takes the conjugate prior of the Poisson, the Gamma, as the prior distribution. Some fairly complex numerical integration is then required. It is possible to avoid this by using a non-parametric approach. This means essentially, that an unrestricted prior is being used. Apart from being simpler computationally, it has the advantage that it will often provide a more intuitive form for an estimator. This was the approach taken in developing an estimator for the process average.

3. ESTIMATOR OF THE PROCESS AVERAGE

Assume that the estimate of an operator's error rate may be expressed as $\hat{P}_i = P_i + e_i$ where e_i is the sampling error and P_i is the operator error rate for the ith period. It is assumed that the operator's error rates are random and homogeneous, analogous to a one way layout with random effects, i.e., $P_i = P + a_i$, $i = 1, \ldots, t$, where P is the average error rate and a_i is the ith period effect with $E(a_i) = 0$. The process variance, denoted by A, is defined as the variance of the period effect, $A = V(a_i)$. The process average estimate is required for period t since this is the time at which an operator's sampling plan will be revised. The weight (shrinkage factor) for period t is constructed as

$$W_t = D_t/(D_t + A), \tag{3.1}$$

where D_t is the sampling variance for period t. The non-parametric empirical Bayes (NPEB) estimate of the process average for period t is then given by

$$\hat{P}_t^* = (1 - \hat{W}_t)\hat{P}_t + \hat{W}_t\hat{P}, \tag{3.2}$$

where $\hat{W}_t$ is the estimate of the weight for the period t, $\hat{P}_t$ is the estimate of the operator's error rate for period t and $\hat{P}$ is the estimate of the average error rate over t periods.

Heuristically, if the sampling variance is small relative to the process variance, the weight on the average error rate will be small because the current sample result is accurate. If the converse is true which means that the sampling variance is large relative to the process variance, the weight on the average error rate will be large because the current sample result is inaccurate in comparison.

Morris (1983) notes that a choice between an estimate computed from the prior data and an estimate computed from the current data may be interpreted as a choice between $A = 0$ and $A = \infty$. The shrinkage estimator of the process average that is proposed is richer because it allows values of A between zero and infinity. This is the true benefit of the model and how the weight W_t is estimated is far less critical.

If the average error rate P is estimated through weighted least squares:

$$\hat{P} = \frac{\Sigma_i \hat{P}_i/(\hat{A} + \hat{D}_i)}{\Sigma_i 1/(\hat{A} + \hat{D}_i)}, \qquad (3.3)$$

where $\hat{D}_i$ is the estimated sampling variance, the estimate of the process average in (3.2) is approximately a Best Linear Unbiased Predictor (BLUP). That is, in the class of all unbiased predictors of the process average, it has minimum variance (Rao, 1986).

In practice, there are situations where the weighted least squares estimate is impossible to obtain because estimates of the process variance and sampling variance for one of the periods i are both zero. Zero estimates of the process variance are a consequence of high sampling variances and/or too few periods of data while a zero estimate of the sampling variance is a degenerate case caused by the absence of errors in the sample.

However, a pooled average calculated through ordinary least squares, is another unbiased estimate which is always available:

$$\hat{\hat{P}} = \Sigma_i \ell_i \hat{P}_i, \qquad (3.4)$$

where $\ell_i = N_i/\Sigma_i N_i$ and N_i is the number of units an operator has processed for the ith period.

Note also that when the estimate of the process variance is unavailable, the weight (3.1) cannot be calculated either. This problem is discussed under implementation in section 5.

For the majority of cases, where an estimate of the process variance is available, an estimate can be obtained by noting that the expected value of the weighted residual sum of squares of the error rates about their average value will be equal to the degrees of freedom; Carter and Rolph (1974). Therefore, we solve the following non-linear equation

$$\sum_i \frac{(\hat{P}_i - \hat{P})^2}{(\hat{A} + \hat{D}_i)} = t - 1, \qquad (3.5)$$

for $\hat{A}$ to get an estimate of A.

Morris (1983) suggests solving this equation iteratively by

$$\hat{A}_{n+1} = \frac{\Sigma_i (t/(t - 1) (\hat{P}_i - \hat{P})^2 - \hat{D}_i) / (\hat{A}_n + \hat{D}_i)}{\Sigma_i 1/(\hat{A}_n + \hat{D}_i)}, \qquad (3.6)$$

for $t \geq 4$. If the solution $\hat{A} < 0$, set $\hat{A} = 0$. Convergence to a level of accuracy of 1% of the difference between successive iterates can be be expected in 4 to 8 iterations (Morris, 1983). A starting value for the algorithm may be determined by calculating a simple quadratic unbiased estimate of A as

$$\hat{A}_1 = \frac{\Sigma_i (\hat{P}_i - \hat{\hat{P}})^2}{t - 1} - \hat{\hat{D}}. \qquad (3.7)$$

where $\hat{\hat{P}} = \Sigma_i \hat{P}_i/t$ and $\hat{\hat{D}} = \Sigma_i \hat{D}_i/t$. Since the iterative algorithm requires a positive starting value, if the above method for obtaining a starting value produces a negative result, a small positive initial value may be used.

4. APPROXIMATE CONFIDENCE INTERVAL

In the case of equal variances, ($D_i = D$, $i = 1, \ldots, t$), when the number of periods is large, the process average estimator becomes a Bayes estimator and then the Bayes posterior variance of each process average is nearly equal to D * (1 - W). Although most of the development for empirical Bayes confidence intervals has been done for the case of equal variances, experience for many QC applications has shown that an operator's sampling variance fluctuates considerably from one period to the next. It is therefore necessary to consider an approximate confidence interval which does not assume equal sampling variances.

Another consideration is that the number of periods of data to be used in the calculation of the process average will be at most 4. In fact an operator will typically process for more than 4 periods, but only 4 periods are used because this is the maximum number of periods of data which the Quality Control Processing System maintains. This small value of t will cause higher variances in the process average because of estimation of the average error rate and the uncertainty in estimating the process variance A. The expression for the process average variance must account for both sources.

A result due to Morris (1983) is applied to give the estimate of the variance of the process average for the current period t as:

$$s^2(P_t^*) = \hat{D}_t [1 - (t - \hat{r}_t)/t \hat{B}_t], \qquad (4.1)$$
$$+ \hat{v}_t (\hat{P}_t - \hat{P})^2.$$

where

$$\hat{r}_t = \frac{t}{(\hat{A} + \hat{D}_t)} (\sum_i \frac{1}{(\hat{A} + \hat{D}_i)})^{-1}, \qquad (4.2)$$

$$\hat{B}_t = \frac{(t - 3)}{(t - 1)} \hat{W}_t, \qquad (4.3)$$

$$\hat{v}_t = \frac{2}{(t - 3)} \hat{B}_t{}^2 \left(\frac{\hat{D}^w + \hat{A}}{\hat{D}_t + \hat{A}} \right), \qquad (4.4)$$

and $\hat{D}^w = \dfrac{\Sigma_i \hat{D}_i/(\hat{A} + \hat{D}_i)}{\Sigma_i \; 1/(\hat{A} + \hat{D}_i)}$.

Note that $\hat{P}_t$ is the operator error rate for the current period and $\hat{P}$ is the average error rate as previously defined. The terms of this expression may be interpreted by comparison with the Bayes posterior variance, D (1 - W). In that expression, replace the sampling variance D, common to all periods, with $\hat{D}_t$ for the current period. The weight W is replaced by (4.3) which is the weight for the current period with a correction factor applied.

Since the number of periods used to estimate the variance will be 4, the coefficient of $\hat{W}_t$ in (4.3) becomes a constant which is equal to 1/3. This is the correction for the curvature dependence of $\hat{W}_t$ on $\hat{A}$. Since $\hat{W}_t$ is a convex, non-linear function of $\hat{A}$, substitution of an unbiased $\hat{A}$ will still produce an estimate of W_t which has too large a bias. This factor will correct for the curvature.

The coefficient of $\hat{B}_t$ in (4.1), $(t - \hat{r}_t)/t$, accounts for the increase in the variance of the process average due to the small number of periods t. Expression (4.2), $\hat{r}_t$, represents the proportion of r for the current period. In Morris' more general framework, r is the rank of the X matrix of prior data. For the process average, since the prior data is summarized by the average error rate, r = 1.

The second term, $\hat{v}_t$, in the variance of the process average (4.1) (which is not present in the Bayes posterior variance) is the variance of the weight. This term accounts for the uncertainty of estimating the process variance A for a small number of periods. The expression $(\hat{D}^w + \hat{A})/(\hat{D}_t + \hat{A})$ reflects the increase in the variance of the process average resulting from unequal sampling variances.

An approximate empirical Bayes confidence interval for $\hat{P}_t^*$ is given by $\hat{P}_t^* \pm zS(\hat{P}_t^*)$, where z is the 100(1 -α/2)% point of a standard normal distribution, chosen based on the desired level of confidence.

The evidence for this empirical Bayes confidence interval is incomplete but there are three observations which support its use (Morris, 1983):

1. When the variances are equal, the above interval reduces to the one for equal variances, which is known to have the correct coverage probability.

2. This interval was derived using Bayes' theory. As the number of observations (periods) increases, $\hat{A}$ converges in probability to A, and consequently the correct coverage probability is guaranteed by the theory.

3. Computer simulations have shown the coverage probabilities to be valid.

5. IMPLEMENTATION

In most cases implementation of the proposed estimator of the process average may be achieved by simply evaluating (3.2). However, exceptions arise when an estimate of the process variance, A, cannot be calculated. This means that the average error rate cannot be estimated through weighted least squares, nor can the weight be estimated and hence the process average estimate is unavailable. The inability to estimate A is a result of high sampling variances leading to negative estimates of A.

For the average error rate, estimation is still possible by using the pooled average specified in (3.4). It is an unbiased estimate and the efficiency loss in using it is expected to be relatively small. Because of this, the decision was taken to simplify implementation even further by using only the pooled average; that is, even when an estimate of A is available.

On the other hand, for estimation of the weight, there is no alternative estimator for W_t which is independent of A. Consequently, the process average cannot be estimated using (3.2). To compensate for this the process variance is assumed to be equal to the sampling variance forcing the weight to take a value of 1/2. This means that the estimated error rates will contribute equally to the process average estimate. This is felt to be a reasonable decision in the sense that when the variabilities of the estimated error rates cannot be compared, neither estimate should dominate in the estimate of the process average.

There are also occasions when the weight cannot be estimated because the <u>sampling</u> variance is unavailable, as was noted in section 3. This situation is a consequence of the process average being measured as a proportion - when no errors are observed, the estimate of the sampling variance degenerates to zero which implies no sampling variability. Again it is assumed that the sampling variance is equal to the process variance which results in a weight of 1/2. The rationale is as stated above.

Finally, when there are fewer than 25 units for the current period, it is felt that the sampling variability will be too large to shrink between the current sample and pooled average error rates. Instead, the average error rate is taken as the estimate of the process average because it is more stable. Indirectly then, the value of the weight is assumed to be one. Note that the current sample result does figure in the estimate of the process average because the average error rate includes the current period data.

6. EXAMPLES

This section presents examples from 5 typical quality control operations at Statistics Canada. For each operation, the process average was estimated over 10 periods for a single operator. The estimates include the sample error rate ,($\hat{P}_t$), pooled average error rate, ($\hat{P}$), and NPEB estimate, ($\hat{P}_t^*$).

Table 1 displays the average for each estimate as well as the average mean square error (MSE) over 10 periods.

The mean square error of the pooled average is derived under the model as:

$$MSE(\hat{\bar{P}}) = \Sigma_i \ell_i (A + D_i) + A - 2\ell_t A. \qquad (6.1)$$

The derivation is straight-forward and will not be given here. An estimate of the mean square error is obtained by substituting the estimates of A and the D_i's in (6.1). Denote an estimate of mean square error by mse.

Table 1: Average Estimates and Average Mean Square Error for 5 Individual Operators from 5 QC Operations.

Operation av. est. av. mse.	Estimator		
	$\hat{P}_t$	$\hat{\bar{P}}$	$\hat{P}^*_t$
LFS	1.123	1.108	.946
	.624	.286	.490
SEPH	1.824	2.543	1.945
	1.062	1.322	1.008
CODE3	3.156	3.345	3.040
	7.453	4.432	5.844
CODE4	8.734	8.034	7.489
	17.260	14.812	14.409
E62	2.726	2.212	2.469
	2.670	2.442	1.595

A comparison of the mses within each operation indicate that the sample error rate does poorly, and that there is not a large difference in the accuracy of the NPEB and pooled average estimates. Although it may be surprising that the pooled average estimate is so close to the NPEB estimate it must be noted that only the minimum number of observations ($t = 4$) is available to calculate the NPEB estimate. If a larger number of periods was used for the purpose of improving estimation of the process variance, the NPEB estimate would be expected to be considerably more accurate than the pooled average estimate. In fact for the small value of t, there is a good indication that the NPEB estimate performs quite well.

7. SUMMARY

The non-parametric empirical Bayes approach outlined in this paper is an attempt to improve estimation of the process average by borrowing strength from an additional source of information. It is convenient because it avoids the complex numerical integration which arises when a prior distribution is specified using a parametric empirical Bayes approach.

A heuristic adjustment procedure has been proposed to resolve cases where the shrinkage factor cannot be calculated. In these situations there is insufficient data to use a more rigorous approach. The assumption of the sampling variance being equal to the process variance is felt to be reasonable and conservative, in the sense that without an estimate of the weight, there is no evidence to support that one estimate of the error rate should dominate over the other in the estimate of the process average.

For smaller values of t, estimates of the variance for the NPEB are inflated, because of the uncertainty in estimating the process variance. Future work should therefore be directed at improving the estimates of the process variance. This could be achieved by using more data for each operator or perhaps by pooling data for a homogeneous group of operators, for the purpose of estimating the process variance.

ACKNOWLEDGEMENTS

The authors wish to thank Professor J.N.K. Rao for specifying the model described in this paper and for his guidance concerning the modifications which were incorporated. We are also grateful for many helpful comments received from J. Smith and J.G. Kovar.

REFERENCES

Carter, G.M. and Rolph, J.E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. Journal of the American Statistical Association 69, 880-885.

Hoadley, B. (1979). An empirical Bayes approach to quality assurance. Proceedings of the 33rd Annual Technical Conference of the American Society for Quality Control, 257.

Hoadley, B. (1981). The quality measurement plan (QMP). The Bell System Technical Journal 60, 215-273.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. Journal of the American Statistical Association 78, 47-54.

Mudryk, W.V. (1988). Quality Control Processing System at Statistics Canada. Presented at the Fourth Annual Research Conference, Arlington, Virginia, March 21-23.

Rao, J.N.K. (1986). Sampling theory and methods. Unpublished course notes. Carleton University, course no. 70:552.