INDEPENDENT VERSUS DEPENDENT QUALITY CONTROL OF CLERICAL CODING: ONE ENLIGHTENING EXPERIENCE

Robert T. O'Reagan, Maureen P. Lynch, Susan M. Odell *
Bureau of the Census

1. INTRODUCTION

Almost any organization that processes data has applications that require a coder or clerk to substitute numerical codes for free verbal responses to questions. Concise numerical codes are much more tractable for computer manipulation and aggregation.

More generally, assume that one has a set of N responses, designated $r_1, r_2, \ldots, r_N$. Each of these responses truly belongs in one and only one of K mutually disjoint categories called codes, $c_1, c_2, \ldots, c_K$. Coding is the act of associating response r_i with c_j; it is conceived of as a many-to-one mapping.

Experiences over several decades have shown that coding of any complexity is quite error prone. It has not been unusual for the coding error to reach ten or more times the magnitude of the sampling error; compare the error rates quoted in this document with the Current Population Survey sampling error. The coder performance has displayed considerable variance both between coders in a given time period and within the same coder's work across time periods. Insufficiencies in the coding structure, the reference manuals, the training, the responses, or the good intentions of some coders may influence these variations. Coding is therefore a popular candidate for the application of statistical quality control. Statistical samples are selected and the codes verified in order to estimate the error level present in each coder's work. When a coder's performance appears to decline below some pre-stipulated threshold, corrective action is taken. The corrective action could include rework of the product, retraining of the coder, or even removal of the coder.

For the purpose of this paper, it will be convenient to regard all quality control schemes as belonging in one of two categories based merely on the verification process used: dependent verification or independent verification.

Dependent verification was described by Dalenius (1968) in somewhat the following way:

> The production coder (the coder carrying out the ongoing coding to be verified) assigns a response into a category by placing one or more digits on a document carrying the information collected from this respondent. This document is then presented to the verifier, who examines the code classification and either approves or disapproves of it. A disapproved code is called an error.

This is the contrasting example of independent verification from the same source:

> The production coder and one or more additional coders classify the same response as to category, but each coder operates without knowledge of the category assigned by the other(s). This verification procedure takes advantage of a comparison of the outcomes of the various coders.

A more comprehensive treatment of the contrasts of the two verification schemes can be found in Hansen, et al (1962).

It is quite reasonable to assume that under dependent verification, the verifier is influenced to some degree by the decision of the original production coder, and that even when that decision is actually incorrect, the verifier may fail to identify it as such. This is borne out in Minton (1969) where some examples of this "miss rate" are as high as 90 percent; that is, the verifier recognizes only 10 percent of the errors to which he is exposed. Indeed, the miss rate has always been the major objection of the U. S. Bureau of the Census against dependent plans. It was primarily to overcome this deficiency that the Census Bureau invented independent verification, according to Lyberg (1983). Dependent verification yields a biased or understated estimate of the production error rate. Early textbooks on statistical quality control, such as Grant (1946), indicated no awareness of this phenomenon. Lavin (1946) recognized the problem but his message was overlooked.

It was not until Minton (1972) that serious attention was given to the fact that there are additional consequences to dependent verification. The Operating Characteristic Curve (OC) and the Average Outgoing Quality Curve (AOQ), two functions which are intensely studied by quality control professionals when selecting a control plan, are affected in ways which are significant and less than obvious. Even were it known that a dependent process understated the error rate by say one half, it would not be legitimate simply to transform the error rate by doubling it and replacing the OC and AOQ curves with those for the transformed error value. Tables such as those from Mersch and Dyke (1978) have long been available for assistance in the selection of quality control plans, but the tables never acknowledge a miss rate.

This paper deals with an unusual opportunity afforded us to compare the dependent and independent verification approaches side by side, when, over a ten month period, they were both applied to the same coding operation.

2. THE PARTICULAR EXPERIENCE

The Current Population Survey (CPS), has been conducted monthly since 1942 in response to a need that emerged in the late 1930s for reliable and up-to-date estimates of unemployment. Throughout the 46 years the survey has been in operation, revisions and additions have been made in the data collected. We are concerned with the 70,000 or more persons who are coded each month as to their industry and occupation categories.

For example, Figure 1 shows both the theoretical and actual OC curves for a sample of size 172 and an acceptance number of 6, given an 80 percent miss rate. While the sanguine practitioner might suppose that a lot with a six percent incoming error has only a ten percent chance of lot acceptance, the actual probability of lot acceptance is nearly one. The quality control sample has practically no discriminatory power.

The AOQ curve which corresponds to the OC example above is shown as Figure 2. This assumes that the verifier misses an average of 80 percent of the errors in the quality control sample and that the rectification process overlooks a similar proportion. Instead of the residual error being less than one percent, as the theoretical curve would have it, virtually all errors pass through the system undetected. For practical purposes, there is no effective Average Outgoing Quality Limit (AOQL).

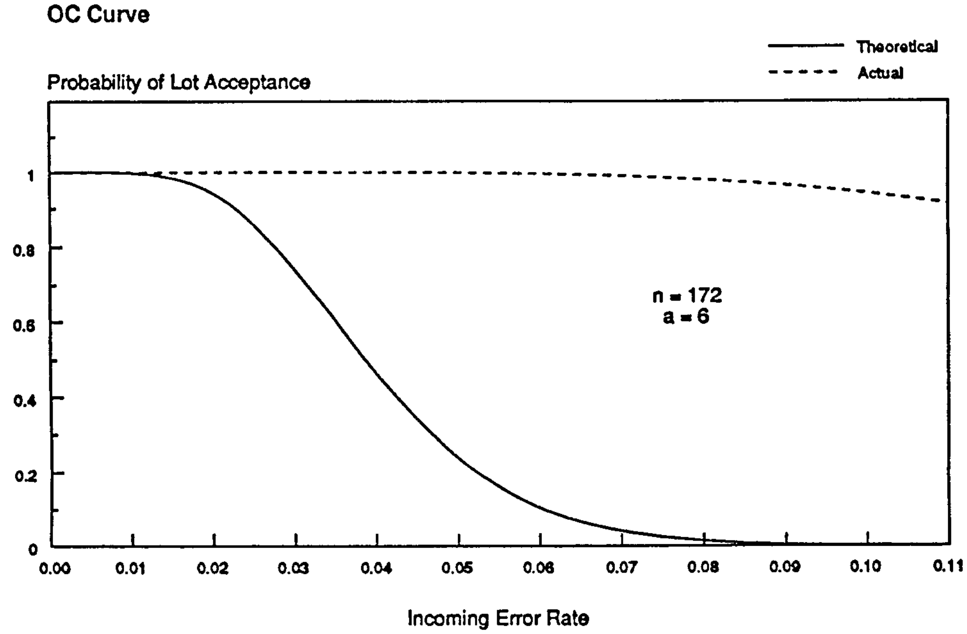

Figure 1: Operating, Characteristic Curve for lot size 2000, Sample Size 172, and 80 percent miss rate.

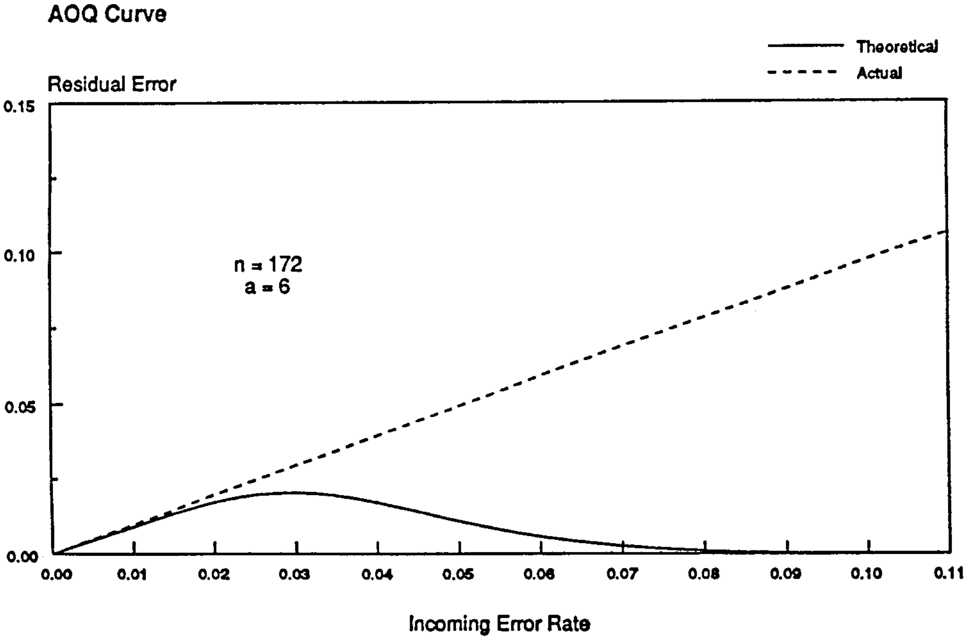

Figure 2: Average Outgoing Quality Curve, corresponding to OC curve shown.

There are in total roughly 230 distinct industry codes and 500 occupation codes. For example, the industry code of 838 indicates a hospital, while the occupation code of 074 signifies dietitian. In the case of CPS, the error rate is defined as the number of errors in industry plus the number of errors in occupation, divided by the number of industry codes assigned plus the number of occupation codes assigned. Because of the larger number of code possibilities in occupation, it has historically had a higher error rate than industry, but we will ignore that for this investigation. We will also pass over the particular question phrasing which elicits the responses that the coders view.

For many years the CPS quality control verification scheme was dependent, and displayed an error rate (as defined above) just in excess of one percent. Assignment of a verifier to a coder was random and changed from month to month.

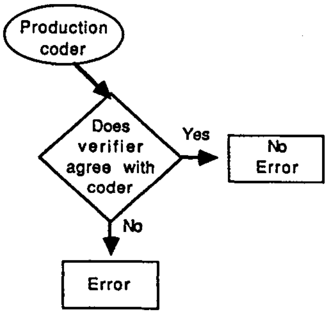

Figure 3: Dependent Verification Plan

Yielding to continued pressure from the quality control staff, the production managers finally agreed to the introduction of the more expensive independent verification plan. In this instance, that verification was a so-called "three way" independent plan; the production coder was charged with an error if two other coders assigned the same code which did not coincide with that of the production coder. If no "majority code" occurred, no error was charged.

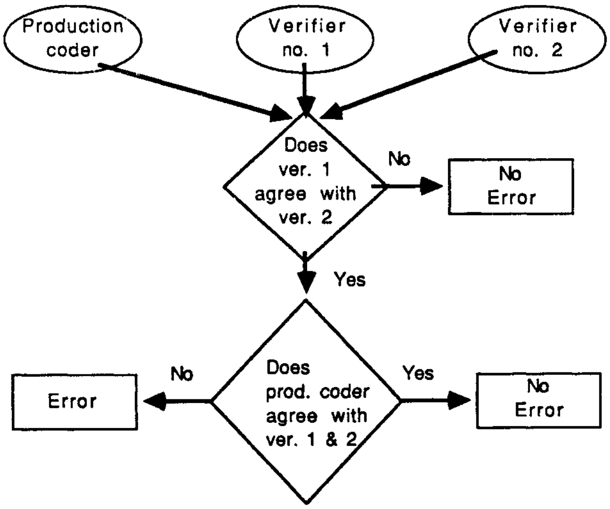

Figure 4: Independent Verification Plan

For an interim test of the procedure (ultimately a test of ten months duration), both quality control schemes were run in parallel. The quality control samples were selected independently, and their results were separately tracked.

It should be noted that the 37 production coders involved were trained and experienced; learning had already taken place and their performance was deemed to be stable. The results reported below would not in all aspects be those expected from newly employed coders. We should also observe that industry and occupation coding is widely regarded as one of the most difficult classification tasks that our employees are assigned; the learning curves for it rise for as much as 36 months.

Over the ten month period of overlap of the plans, the dependent verification continued to display a 1.2 percent error rate, while the estimate based on the independent scheme was 6.3 percent (it has more recently levelled at about 5.3 percent). The net miss rate of the dependent scheme was inferred to be 80 percent. This was a shock to some production managers, but more or less expected by the quality control personnel.

3. THE SURPRISES

Various correlations, ANOVAs, and means and variances were computed. For each kind of quality control verification there were 10 monthly averages, 37 coder averages, and 320 individual monthly/coder error estimates (only 13 coders worked every one of the 10 months). On an average, in a given month the quality control sample under the dependent scheme yielded 172 codes per coder. The sample size for the independent system over that time period averaged 350 observations per coder per month.

Computing the linear correlation between the two error rate estimates across coders and months, we found it to be just under +.24, not significant at the α=.05 level. It appeared that the dependent plan was not merely downward biased compared to the independent plan, but that the two estimates were not correlated. Further, a Spearman rank correlation was computed utilizing the combined 10 months of sample observations for each coder. Coders were ranked according to their overall error rate by the dependent estimate and ranked again by their overall error rate estimate under the independent procedure. The correlation between the two ranks was barely significant. Under either system, the same coder ranked best and the same coder ranked worst. Actually, if that single best coder (about 3 sigma from the means) and that single worst coder (over 3 sigma from the means) are omitted, the correlation between the two rankings is .25, not significant at the α=.05 level.

Month to month correlations were computed under each plan; i.e. January error rates were correlated with February error rates, February with March, and so on. The average correlation within the independent system was .33, and with the dependent system was .20. These are not significant correlations at the α=.05 level for the degrees of freedom involved. In other words, this month's error estimates are not good

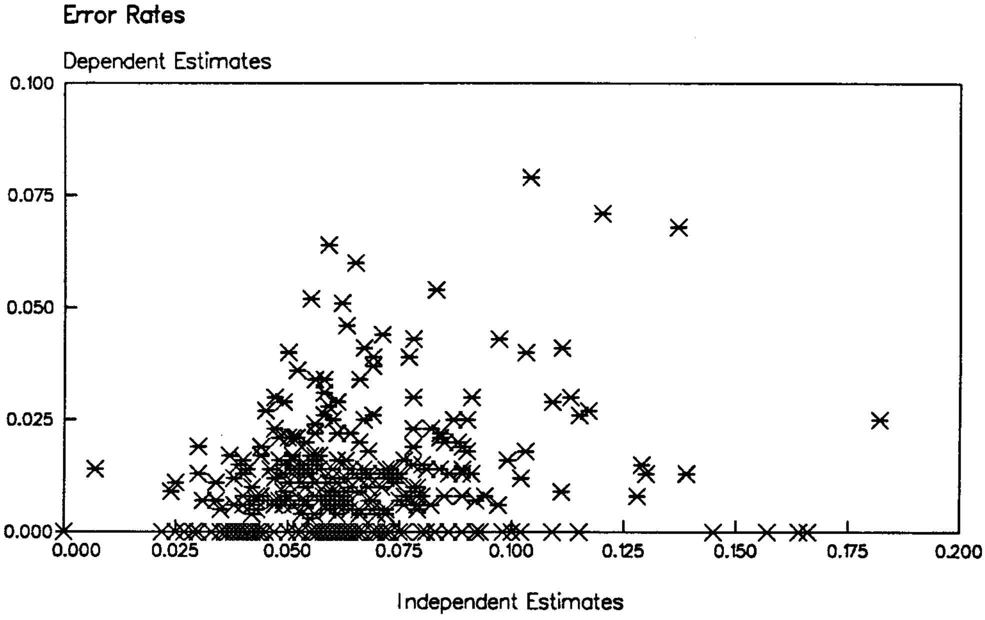

Figure 5: Scatterplot of Dependent Error Rate Estimates versus Independent Error Rate Estimates

predictors of next month's performance.

Various ANOVA runs were made, some using 3 factors and all coders, others restricted to fewer factors and just the 13 coders who worked every one of the ten months. By far the most significant factor was which QC plan (dependent, independent) produced the error estimate. That came as no surprise. The coder is likewise significant, but much more notably (smaller α) for the independent plan. Significant variability is attributable to the month factor only under the dependent plan, where across-month variance within coder almost matches within-month variance across coders. Remember that verifier assignments rotated.

Other examples of uncorrelated estimates exist. In the processing operations for the 1980 Censuses of population and Housing there were three major coding operations, of which industry/occupation coding was one. The decision strategy was to control the overall quality by quickly identifying and removing bad coders. A safety net provided that if the quality for a lot was estimated to be extraordinarily bad, that lot would be rectified. In other words, the possible decisions that could be made on a work unit based on the QC sample were A, for accept; R, for reject; and E for extremely bad/rectify. E decisions, a subset of R decisions, were given only to work units with error rates estimated to exceed 30 percent; E decision lots were 100 percent rectified. For all three coding operations, including industry/occupation, the error rates estimated from 100 percent rectification failed to correlate with estimates of the lot error rate from the QC sample at the α=.05 level. For example, in industry/occupation coding the quality control sample estimate of the error rate for the work lots which did undergo 100 percent rectification was 46 percent. The error rate as determined by that complete rectification of these same work lots was 22 percent. In one of the other coding operations, the respective error estimates were 52 percent and 4 percent. Rectification is a dependent process and rectifiers are influenced by the codes they see.

4. CONCLUSIONS

If we are willing to assume that an independent verification scheme produces a reliable estimate of incoming error rate (process average), we must infer that a dependent verification scheme produces an unreliable estimate. The two estimates have been seen to have insignificant correlation at the α=.05 level. The miss rate associated with dependent schemes causes them to be not only biased or scaled-down estimators as reported by Minton (1969) but broadly unreliable. An estimate of a coder's error rate for time period t cannot be safely used to predict an error rate for period t + 1. Perhaps the only utility of a dependent verification plan is psychological effect on those being controlled; they are probably not aware that the estimates are not accurate. Finally, rectification schemes are not effective in that they too are dependent in nature and have an attendant miss rate. If these conclusions are too broadly inferred, we hope they will provoke research toward counter-examples.

REFERENCES

Dalenius, T. and Frank, O. (1968): Control of Classification. Review of the International Statistical Institute, Vol. 36:6, pp. 279-295.

Grant, E. L. (1946): Statistical Quality Control. McGraw-Hill, New York.

Hansen, M. H., Fasteau, H. H., Ingram, J. J., and Minton, G. (1962) Quality Control in the 1960 Census. Proceedings of the 1962 Middle Atlantic Conference. American Society for Quality Control, Milwaukee, 1962, pp. 311-339.

Lavin, M. (1946): Inspection Efficiency and Sampling Inspection Plans. Journal of the American Statistical Association, Vol. 41, pp. 432-438.

Lyberg, L. (1981): Control of the Coding Operation in Statistical Investigations -- Some Contributions. Ph.D. Thesis, Statistics Sweden, Urval No. 13.

Lyberg, L. (1983): The Development of Procedures for Industry and Occupation Coding at Statistics Sweden. Statistical Review, 1983:5, pp. 139-156.

Mersch, M. L., and Dyke, T. C. (1978) Construction and Use of Quality Control Tables. ASQC Technical Conference Transactions. American Society for Quality Control, Chicago 1978, pp. 360-364.

Minton, G. (1969): Inspection and Correction Error in Data Processing. Journal of the American Statistical Association. Vol. 64, pp. 1256-1275.

Minton, G. (1970a): Comments on Quality Control and Research in Data Processing Programs. Paper presented at the Conference of the American Society for Quality Control, March 12-13, Arlington, Virginia.

Minton, G. (1970b): Some Desicion Rules for Administrative Applications of Quality Control. Journal of Quality Technology, Vol. 2, pp. 86-98.

Minton, G. (1972): Verification Error in Single Sampling Inspection Plans for Processing Survey Data. Journal of the American Statistical Association, Vol. 67, pp. 46-54.

U. S. Bureau of the Census (1972): Effects of Coders. Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960. Series ER60, No. 9.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author(s) and do not necessarily reflect those of the Census Bureau.