OUTLIER EDITING: A CASE STUDY FROM A SURVEY OF BANKS

Carl Ramirez, U.S. General Accounting Office[1]
U.S. GAO, 441 G St., NW, Room 2921, Washington DC 20548

Key Words: Outliers, Editing, Frame

Outliers in survey measurements arising in complex probability samples can have major effects on estimates in terms of both precision and validity. They may arise in a number of ways, and automated or manual procedures may be employed to detect and edit them or adjust estimates to compensate for them (Lee, 1995; Granquist, 1995).

This paper summarizes a mechanism in an establishment survey resulting in one type of outlier and the treatment used to resolve it. In this case study, an outlier in a survey of U.S. banks and thrifts estimating the number of Automated Teller Machines (ATMs) resulted primarily from a mismatch between the definition of the sample unit and the reporting unit. The case was representative, in that the extreme and influential ATM value reported was a valid observation and could not be regarded as unique (Chambers, 1986). The treatment used was not to edit outlier values themselves but to reassign them across related sample elements within corporate families. This type of edit preserved the multivariate relationships between other survey items unaffected by the outlier report. This suggests an approach for detecting outliers and identifying likely edits needed, using information on the relationship between sample elements within larger corporate hierarchies.

The GAO Bank Survey

The second wave of the GAO Survey of Banks and Thrifts on ATM Surcharges was conducted in February and March of 1998. The study population was defined as all independently chartered active U.S. banks and thrifts. The sampling frame was derived from a database of the September 1998 Statements of Financial Condition ("Call Reports") that federal regulators require all insured depository institutions to file quarterly.

A mixed panel stratified probability sample was used -- all 1997 sample elements still eligible for the survey were resampled in 1998, along with a supplemental sample of banks not in the 1997 survey. The 501 sample elements were allocated across 11 strata defined by total assets as of September 1998; sampling rates varied across strata, with proportionally more of the larger banks being selected into the sample.

The self-administered mail out/fax back questionnaire requested the following data from banks: the number of ATMs operated, an enumeration of these ATMs by the level of per-transaction fees ("surcharges") levied on ATM users who did not have accounts at that bank ("noncustomers"), and the number of withdrawals completed by customers and noncustomers at all of the sampled bank's ATMs in a reference month. The number of ATMs operated were to be broken down by respondents into those "on-premises" (at branch locations of the sampled banks) and those "off-premises" (at other locations away from bank properties, such as shopping centers).

All of these statistics were to be reported for the sampled bank, not individual branches of that bank, nor any parent companies or at the entire enterprise level.

A telephone precontact was made to prescreen establishments for the presence of ATMs, and for those banks with ATMs, to then identify the best qualified respondent. Up to 3 mailout waves were conducted, each consisting of a cover letter and a replacement questionnaire. Limited telephone nonresponse followup was conducted after the mailout period. A final response rate of 89% was achieved.

Sampling and Reporting Units

The sampling frame comprised all active, independently chartered depository institutions that were federally insured. Each individual bank or thrift is chartered by one of several federal regulators. These entities may have branches or other business locations (establishments) under them, but they do not figure in any aspect of data collection for this survey. However, it is common in the banking industry for a number of chartered banks to be fully-owned subsidiaries of even larger enterprises -- bank holding companies. Such umbrella organizations file different reports of financial condition, and are not considered in the specification of the sampling frame.

Until recently, a family of banks wishing to operate across state lines would be required to charter at least one separate bank in each state. Therefore, several banks with identical names, under the control of the same corporation, could exist in different states and have independent probabilities of selection. Operationally, these banks acted as no more than

"branches" under the control of a parent or lead bank (typically the largest and oldest flagship institution, usually in the home state and most closely associated with or synonymous with the holding company). However, these banks were considered separate entities in terms of regulatory reporting and the resulting sampling frame.

In this and other bank surveys, reporting problems often arise when a parent organization rolls up answers for its independently chartered banks. If administrative control and program management is located solely in the holding company, and the organization's information system does not distinguish member banks' programs and operations as unique, a survey request to report on individual banks may meet with unexpected responses.

The Case of the Bank X Outlier

The following case illustrates a typical form of misreporting in this and similar bank surveys that results in outliers: The Bank X Corporation[2], a family of independently chartered banks operating under a holding company, had 6 of its member banks drawn into the 1998 survey sample. These 6 banks, all of which responded to the survey, were each located in a different state in the U.S.

The parent or lead bank ("Bank A") was the largest in the family, and was in the take-all stratum of the 81 banks with the highest total assets. "Bank F," the smallest bank in the family (in the 6th size stratum, out of 11), reported a disproportionately large number of off-premise ATMs and withdrawal transactions by noncustomers. See Table 1 below.

Table 1: Pre-Edit Characteristics of 6 Bank X Elements

Bank	Size Stratum (of 11)	Final Weight†	Total on-premise ATMs	Total off-premise ATMs	Customer Transactions	Non-Customer Transactions
A (parent)	1	1.00	371	109	1,692,880	560,796
B	1	1.20	325	48	1,955,734	526,564
C	1	1.20	240	44	1,292,293	402,877
D	2	1.29	115	46	534,921	191,357
E	4	1.64	96	22	372,970	130,125
F (outlier)	6	8.98	29	**5,464**	167,048	**1,691,541**

† Reflects non-response weights applied within adjustment cells (not equivalent to strata).

This 6th stratum observation, and others like it, were flagged for review after informal examination of univariate frequencies; this survey did not use any automated statistical editing or outlier detection procedures. This outlier no doubt would have been identified at any practical tolerance level using typical outlier detection methods, such as those based on the interquartile range, regardless of any masking that the skewness of the distributions might have caused.

Impact of outlier on estimates

This one observation had a major impact on the estimates made for the totals of off-premise ATMs and noncustomer transactions for the 6th stratum, and resulted in large variances for estimates of this variable for the population, within that stratum, and in an analytical grouping of several "mid-sized" strata. If this observation had been in a donor class for reweighting other cells with missing data, the impact of the outlier would have been even higher.

Note, however, that the total number of transactions by *customers* for the outlier (Bank F) in Table 1 seems to be unaffected -- this report is much smaller and more consistent with the size of the bank. This unusual relationship between customer and non-customer transactions, which usually should correlate highly, is another indicator of a data problem. The mechanism behind the outlier, when it became known, also explained this unexpected relationship.

Explanation of Outlier Source

Followup interviewer contact with the reporters for these banks revealed that the large number of off-premise ATMs in question had been recently acquired from a non-bank operator of ATMs. The interviewer was told that a large, but undetermined portion of the 5,464 ATMs had been purchased in this way. The ATMs were not all in the same state as Bank F (the outlier case), nor were they truly "owned and operated" (the questionnaire specification) by Bank F. The

operational control of fees charged and other characteristics of these ATMs resided in a unit at the corporate headquarters – in the same state as, and co-located with the offices of the lead bank, Bank A – which directed electronic banking services for the entire holding company.

Corporate officials stated that the reporter who completed most of the questionnaires sent to sampled family members had chosen to associate these ATMs with the relatively small Bank F because that type of ATM made up a large proportion of Bank F's ATMs, and perhaps the acquired institution had been headquartered in that state. However, it was largely an arbitrary decision because the new ATMs did not fall under the control of any one entity in the family, and the administrative systems of Bank X did not classify "ownership and operation" in a way that would translate to a clear assignment.

We also learned of two possible reasons why the number of customer transactions was relatively small compared to the unusually large number off-premise ATMs and non-customer transactions. Not only were the ATMs carried on the books of Bank X for a time before they were "re-labeled" and integrated into the existing ATM network (during which time customers would not tend to use them), but Bank X may also have reported some proportion of transactions these ATMs had processed during the reporting period but before the ATMs officially came under the ownership of Bank X, making all of those transactions necessarily "non-customer" transactions.

Ultimately, this type of outlier occurs because the administrative data forming the sample frame reflects the regulatory view of the organizational structure of the banking industry (where "establishment level" chartered banks are the units of analysis), but in reality the corporate governance of banking institutions is often organized differently (control is at the "enterprise," or holding company level). The bank characteristic of interest (ATM operations) is not well structured by the sample frame, and information on the individual sample elements is often not maintained at the enterprise level, or even by those elements themselves.

Also, rapid consolidation of this industry and changing laws on interstate banking make for many births and deaths (at the individual bank level) and other changes in structure, making it difficult for banks to produce retrospective reports for units which no longer correspond to a firm's structure.

Finally, the variation of ATM characteristics within strata was already high because the measure of size used to stratify (total assets under management) was not as highly correlated (.705) with total number of ATMs as other potential measures such as number of branches (.908) or total deposits (.877).

Resolution

It was concluded that the family of Bank X did actually operate these ATMs, and that the collection of bank elements did have the fee and transaction characteristics ascribed to them by survey reporters. Therefore, this was a representative outlier, in that the case did not require value editing or imputation. However, the association of the large body of ATMs in question to a specific sample unit (Bank F) was not realistic and introduced unwanted variability.

The 5,464 ATMs were kept within the family, but most were reassigned to the lead bank (Bank A), which was most closely associated with the holding company level which actually managed the ATMs. The lead bank was also the largest sampled element from the holding company, and so this allocation would be most consistent with the distribution of ATMs at other banks and thus cause the least increase in variance. We had auxiliary information to help us apportion the ATMs. Respondents had been asked to itemize the number of ATMs by surcharge fee level. Bank X officials told us that all of the acquired ATMs had the same fee level. Of the 5,464 off-premise ATMs at Bank F, there were 5,339 at the $1.50 fee level, 97 at the $1.00 level, and 28 at the $0 (no surcharge) level. We assumed the 5,339 represented the acquired ATMs, and added them to the 109 already operated by Bank A. We then reassigned a corresponding proportion of the noncustomer transactions from Bank F to Bank A, after making the assumption that all of the off-premise ATMs involved generated approximately the same number of transactions.

Implications for Outlier Detection/Edit Procedures

For this survey, a more rigorous system for identifying this type of outlier mechanism should be developed. The use of corporate family relationship information could help detect less obvious misreports (that would not be caught by ad hoc inspection) due to divergence between reporting and sample units. For example, the following three diagnostics could be performed:

1. Identify cases in which a subsidiary family member reports a survey value larger than a preset multiple of the next largest family member. This may be used to identify outlier conditions similar to that found in Bank X.

2. Identify cases in which the parent or lead bank has a high value, and subsidiaries report very few or none of the characteristic. This is the opposite of the case described in this paper, and indicates that no attempt has been made to report for the sampled elements; all activity has been rolled up to the holding company level.

3. Identify cases in which all the subsidiaries have equal numbers reported. This indicates an attempt to satisfice by allocating a value obtained for the family equally over its members in the absence of information.

4. Identify cases in which the parent's total on some characteristic equals the total for all subsidiaries, indicating possible double counting.

Discussion

Reporting units chosen by the respondent can differ from sample units, and the distribution of the survey characteristic within a complex, multi-divisional enterprise may not be well represented by a frame developed for other official purposes. This type of edit – the reclassification of reported values across related sample elements – may be a form of value editing but is uniquely associated with a specific frame and reporting problem. In another sense, this is essentially an adjustment to the weight applied to a value, not achieved by changing the weight itself, but shifting the value itself to an element in another weight class. This outlier-causing mechanism (mismatch of sample and reporting unit) suggests outlier detection/edit strategies that use corporate family information, in addition to other measures of distance and the multivariate relationships between characteristics within individual cases.

References

Chambers, Raymond L (1986) "Outlier Robust Finite Population Estimation," *JASA*, Vol. 81, pp. 1063-1069.

Franklin, Sarah, and Marie Brodeur (1997) "A Practical Application of a Robust Multivariate Outlier Detection Method," *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA: ASA, pp. 186-191.

Granquist, Leopold (1995) "Improving the Traditional Editing Process" in B.G. Cox, et al. (eds.), *Business Survey Methods*, New York: Wiley, pp. 385-401.

Lee, Hyunshik (1995) "Outliers in Business Surveys" in B.G. Cox, et al. (eds.), *Business Survey Methods*, New York: Wiley, pp. 503-526.

Struijs, Peter and Ad Willeboordse (1995) "Changes in Populations of Statistical Units," in B.G. Cox, et al. (Eds.), *Business Survey Methods*. New York: Wiley, pp. 65-84.

---

[1] The views expressed are the author's own and do not necessarily reflect the position of the U.S. General Accounting Office.

[2] Because non-public bank information of a potentially sensitive nature is discussed, identities and certain bank characteristics are suppressed or made intentionally vague.