

OUTLIER DETECTION IN ASYMMETRIC SAMPLES: A COMPARISON OF AN INTER-QUARTILE RANGE METHOD AND A VARIATION OF A SIGMA GAP METHOD.

Julie Bernier and Karla Nobrega¹

ABSTRACT

This paper presents the results of a simulation that compares the number of observations flagged as outlying by an inter-quartile range method and a variation of a sigma gap method. This study was undertaken for the new Unified Enterprise Statistical Program (UESP), which will use donor imputation to impute missing data while excluding outliers (units that would not be good potential donors) from the donor pool. The outlier detection method used should be robust against asymmetry and small sample size and detect only those observations that are very dissimilar to other observations in the sample. The simulation using administrative data shows that the sigma gap method consistently identifies an acceptably low percentage of units as outlying observations. In large samples, less than one percent are identified. Unlike the inter-quartile outlier detection method, it is possible that no observations at all are selected as outliers if the distribution of gaps in the data is consistent across all observations. A similar outlier detection method will be used to identify units for review after imputation, using less restrictive parameters.

KEY WORDS: Outliers; Sigma gap; Inter-quartile range.

RÉSUMÉ

Cette article exposera les résultats de simulations comparant le nombre d'observations aberrantes détectées selon la méthode de l'écart inter-quartile et une variante de la règle de l'écart-sigma. Cette étude a été élaborée dans le cadre de la nouvelle enquête unifiée sur les entreprises. Cette enquête utilisera l'imputation par donneur pour imputer les données manquantes, en excluant au préalable les valeurs aberrantes i.e. celles qui ne seraient pas de bons donneurs potentiels. On recherche une méthode de détection des données aberrantes robuste à un échantillon de petite taille de distribution asymétrique et capable de détecter les unités différentes du reste des unités de l'échantillon. Les résultats provenant de données simulées et de données administratives ont montré que la règle de l'écart-sigma ne détecte pas trop de données aberrantes soit moins de un pourcent. De plus, contrairement aux méthodes traditionnelles de détection de données aberrantes, il est possible qu'aucune observation ne soit identifiée aberrante si les écarts entre les données ordonnées sont plutôt constants. Une méthode similaire de détection de données aberrantes sera utilisée pour identifier des unités à réviser après imputation, en utilisant des paramètres moins restrictifs.

MOTS CLÉS: Observations aberrantes; écart-sigma; écart inter-quartile.

1. INTRODUCTION

Outliers are often defined as observations that are not consistent with the majority of the data set; they are observations, weighted or unweighted, that are improbable or rarely occurring (Hampel, 1986). Outliers can occur because of the underlying distribution of the data, sampling designs, response errors and capture errors. When the data are from asymmetric or heavy tailed symmetric distributions the

presence of apparent outliers increases. Unlike household surveys, the populations of business surveys are asymmetric. Often, only a handful of enterprises represent a large proportion of the market share.

The edit and imputation strategy for the Unified Enterprise Statistics (UES) pilot survey will be used to decide which records are inconsistent and which ones require donor imputation. Records that have a high impact on the estimates should be verified in order to

¹ Julie Bernier and Karla Nobrega, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6, bernjul@statcan.ca, nobrkar@statcan.ca

ensure the integrity of the estimates. These critical units should be reviewed manually.

Donor imputation will be used in the pilot survey of the UES. Data from all of the strata will be combined in the edit and imputation process. The data could be from a single establishment or from a collection of many establishments. To ensure that a donor is similar to its recipient the data are divided into 98 data groups. These data groups are broken down by 6-digit North American Industry Classification (NAIC) code and by geographic region. The data groups range in size from 43 to 624 with a median size of 134 observations.

The edit and imputation system will be automated, in that the data will be entered into the system and will proceed without manual intervention through the process. Outlier detection is performed prior to edit and imputation on each data group. These outliers are not used as donors for imputation. If any outliers require imputation, they will be reviewed and manually imputed after the edit and imputation process.

In the UES pilot survey, very large outlying observations are not similar enough to other units ranked beneath them to be used as donors, even if all units are in the same industry and province. If a large unit is used as a donor for a smaller record that has even a medium weight, the overall estimates, as well as the micro data, can be seriously affected. In the case that the smaller recipient record has a small weight, the overall estimates are not as seriously affected but the micro data will not be consistent. These large firms, which are unlike others in the sample, should be identified and subsequently excluded as donors during imputation.

The economic data collected from Statistics Canada's UES pilot survey are often small, asymmetric samples, with missing values imputed by donors. The outlier detection method used should be robust against asymmetry and small sample size and detect only those observations that would not make good donors. This paper presents the results of a simulation that compares the number of observations flagged as outlying by an inter-quartile range method and a variation of sigma gap method.

2. DONOR IMPUTATION IN STATISTICS CANADA'S GENERALIZED EDIT AND IMPUTATION SYSTEM

In Statistics Canada's Generalized Edit and Imputation System (GEIS) records with fields requiring

imputation are first identified using Fellegi and Holt's (1976) minimum change rule that was later developed by Sande (1979). The donor imputation in this system is a nearest neighbour approach. The donor is selected by a minimax distance using the rank of the observation (Kovar et.al, 1991). The rank is a useful transformation of the data that allows the user to combine continuous units with different scales.

Continuous data from distributions with no large gaps are generally suitable donors in imputation modules. There exists the possibility that a record could be an outlier and if used as a donor, this would create more outliers. In this case since the donors are matched by rank if there is a large distance between two ordered adjacent observations then each of the observations on either side has a chance of being selected as a donor for the middle observation. Also, it is possible that a donor the imputation procedure creates new outliers. Although this problem can be lessened by specifying post-imputation rules, or by specifying more that one matching field, if the user knows in advance that a small portion of the data set is not similar to others in the data group then removing these units is preferred.

3. INTER-QUARTILE RANGE METHOD

The inter-quartile range or quartile method is one method used in the Generalized Edit and Imputation System at Statistics Canada to detect outliers (Kovar, et al. 1991). An observation is considered an outlier if $x_i > \text{median}(x_i) + c_U(q_3 - q_2)$ or $x_i < \text{median}(x_i) + c_L(q_2 - q_1)$, where q_3 is the upper quartile, q_2 is the median and q_1 is the lower quartile. This method is advantageous because it deals directly with the data. Both the upper and lower quartiles will reflect asymmetries in the data. The constants can also be chosen independently of one another and can be fine tuned to fit the data.

A general result can be shown for order statistics that the expected value of $F(Y_j) - F(Y_i)$, $i < j$ is $(j - i) / (n+1)$ (Hogg and Craig, 1978). Therefore, it is expected that $[3/4(n+1) - 1/2(n+1)] / (n+1) = 1/4$ of the data will be between the $(q_3 - q_2)$. Often $c_U = 3$ is used as a bound to identify extreme observations. This bound is based on normality assumptions. The data in most business surveys are not normally and not symmetrically distributed.

The main disadvantage to this method is that the data must be available in order both to define and fine tune the constants. This process will need to be repeated through several iterations to make certain that only a small percentage of outlying observations are selected.

The constants must be re-determined for each data group. It is possible to select outliers even when no outliers are present in the data. For example, if the data set does not have large gaps in its distribution then in fact all the data will make reasonable donors and none should be selected as outliers. To be able to determine if this is the case all of the data sets will need to be visually inspected in addition to the fine tuning.

4. SIGMA GAP METHOD

Sigma gap is a method developed at Statistics Canada to select 'take all' strata in the redesign of the National Farm Survey in 1983 (Ingram and Davidson, 1983). The data set is first ordered, then the distances between adjacent pairs are calculated. Starting after the median of the ordered data set, if $x_{i+1} - x_i > \sigma_x$ all observations greater than x_i are considered part of the take-all strata. For our case, observations beyond this gap are considered outliers.

This intuitive method is easy to program and computes quickly. Although it works well when selecting take all strata and selecting outlying observations, it has no bound on the number of gaps it will select. As well, because it does not compare the datapoints themselves but rather considers the distance between them, the upper bound on the percentage of data selected is 50% if the first large gap is located directly after the median.

5. MODIFIED SIGMA GAP METHOD

For any data set the Chebyshev inequality guarantees that the probability that an observation falls within k standard deviations of the mean is $(1-1/k^2)$. However the mean is not a robust estimator of central tendency. One outlying observation can cause the mean no longer to be a useful estimator. The percentage of contamination that can exist in a data set before an estimator loses meaning is called the breakdown point of the estimator (Rousseeuw and Leroy, 1987). The

median is less affected by outliers and has a higher breakdown point: one half of the data can be outlying observations and the median does not change (Barnett and Lewis, 1984).

The bound on the probability of an observation being within k standard deviations of the median is given by $(1 - (1/k^2 + (\text{median}(X) - \mu)^2 / (k^2 \sigma^2)))$, which as the median approaches the mean or as the standard deviation approaches infinity, tends to the same bound as Chebyshev.

With the modified sigma gap method, no more than $(1/3^2 + (\text{median}(X) - \mu)^2 / (3^2 \sigma^2))$ of the distances fall outside three standard deviations of the median, whereas, with the original sigma gap method there is no upper bound on the number of distances that could be selected as outlying. However, with the modified sigma gap method, there is still the same problem as the original sigma gap method. If one begins looking from the median up, there is the possibility of selecting up to 50% of the observations as outlying, if the first large gap is directly after the median.

6. RESULTS

Since survey data were not available for the testing, 1995 tax data from the seven industries surveyed in the UES pilot were used. The construction industry contained the largest number of observations and so this industry was used to compare the methods with different sample sizes. The results show that both the sigma gap method and the modified sigma gap method select fewer outliers than an inter-quartile range regardless of sample size (Table 1). The modified sigma gap method identified only a few more observations than the original sigma gap method as outliers. As the sample size increased the number of observations selected as outlying in both the sigma gap method and the modified sigma gap method decreased. This occurred because as the sample size increased the number of gaps decreased when you have a finite population.

Table 1
Mean Percentage of Outliers Detected (50 replicates in each sample size)

Outlier Detection Method	Sample Size			
	50	100	150	500
Modified Sigma Gap	4.2	2.8	1.7	0.7
Sigma Gap	4.6	2.6	1.5	0.7
Inter-Quartile Range [†]	11.7	12.0	12.2	12.0

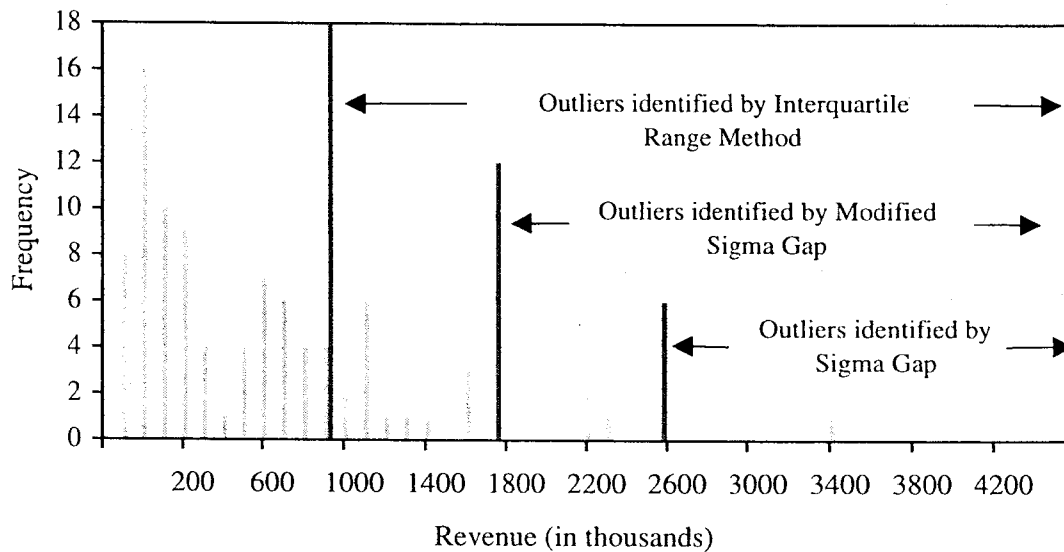
[†]Median + 3*(q4-q3)

Table 2
Mean Percentage of Outliers Detected by Industry (sample size 100)

Outlier Detection Method	INDUSTRY						
	Aquaculture	Construction	Couriers	Food	Lessors	Real Estate	Taxis
Modified Sigma Gap	3.2	2.8	1.9	3.2	2.0	3.1	2.8
Sigma Gap	1.9	2.6	1.9	1.7	1.9	2.5	2.2
Inter-Quartile Range [†]	17.9	12.0	11.9	11.1	11.0	14.8	14.9

† Cu = 3

Figure 1



Using tax data, with a sample size of 100, sigma gap and the modified sigma gap methods selected between 1.7 % and 3.2 % of the observations as outlying across all industries. This variation reflects differences in the asymmetry and variability of each of the industries. The inter-quartile range consistently selected more than 11 % as outlying (Table 2).

Typically, the inter-quartile range selects more outliers without discriminating whether or not they fall after a large gap in the data. The two sigma gap methods select observations after a gap in the data. The modified sigma-gap selects more outliers when there are more than one group of observations that are far from the rest of the observed observations (Fig.1). Multiple gaps can often occur with very asymmetrical data, very small data sets, or data that come from more than one underlying distribution.

7. DISCUSSION

The detection and treatment of outliers is an important step in the survey process. The strategy to detect and treat outliers should be related to the overall survey design and built alongside the other survey processes. It is not desirable to manually check every record at the micro level for inconsistencies (Granquist and Kovar, 1997). The edit, imputation and outlier detection systems should be designed to minimize the number of records that will be removed, thus allowing a homogeneous donor pool that is as large as possible. As well, the system should minimize the number of observations selected as outlying because if they need to be imputed these units require manual imputation.

By eliminating from the donor pool very large firms and/or very large collection entities (a group of smaller

establishments that report as a single unit), the possible mismatch of a small recipient with a large donor is eliminated.

Results from simulated data and administrative data show that the sigma gap method does not over-produce outlying observations. With larger donor pools it can select less than 1 % as outliers. When the results are prepared visually, observations that lie far from the majority of the data are the ones correctly selected as outlying. Across all industries the modified sigma gap method performs consistently. There is no need with either the sigma gap method or the modified sigma gap method to iterate or change constants as with the inter-quartile range method.

REFERENCES

- Barnett, V., and Lewis, T. (1984). *Outliers in Statistical Data*. New York: Wiley.
- Fellegi, I.P., and Holt, D. (1976). A systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 17-35.
- Granquist, L., and Kovar, J.G. (1997). Editing of Survey Data: How much is Enough? *Monograph*

of the International Conference on Survey Measurement and Process Quality, 1995.

- Hampel, F.R. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
- Hogg, R., V., and Craig, A.,T. (1978). *Introduction to Mathematical Statistics*. New York: Macmillan Publishing Co., Inc.
- Ingram, S., and Davidson, G. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 220-225.
- Kovar, J.G., MacMillan, J.H., and Whitridge, P. (1991). Overview and Strategy for the Generalized Edit and Imputation System. Working Paper No. BSMD-88-007E/F
- Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- Sande, G. (1979). Numerical Edit and Imputation. Presented at the 42nd International Statistical Institute Meeting, Manila, Philippines.