The Korean Journal of Applied Statistics (2010)
23(2), 325–333

Investigations into Coarsening Continuous Variables

Dong-myeong Jeong[1] · Jay J. Kim[2]

[1]Statistics Korea; [2]National Center for Health Statistics

(Received November 2009; accepted January 2010)

Abstract

Protection against disclosure of survey respondents' identifiable and/or sensitive information is a prerequisite for statistical agencies that release microdata files from their sample surveys. Coarsening is one of popular methods for protecting the confidentiality of the data. Grouped data can be released in the form of microdata or tabular data. Instead of releasing the data in a tabular form only, having microdata available to the public with interval codes with their representative values greatly enhances the utility of the data. It allows the researchers to compute covariance between the variables and build statistical models or to run a variety of statistical tests on the data. It may be conjectured that the variance of the interval data is lower that of the ungrouped data in the sense that the coarsened data do not have the within interval variance. This conjecture will be investigated using the uniform and triangular distributions. Traditionally, midpoint is used to represent all the values in an interval. This approach implicitly assumes that the data is uniformly distributed within each interval. However, this assumption may not hold, especially in the last interval of the economic data. In this paper, we will use three distributional assumptions - uniform, Pareto and lognormal distribution - in the last interval and use either midpoint or median for other intervals for wage and food costs of the Statistics Korea's 2006 Household Income and Expenditure Survey(HIES) data and compare these approaches in terms of the first two moments.

Keywords: Coarsening, lognormal distribution, Pareto distribution, triangular distribution.

1. Introduction

Protection against disclosure of respondents' identifiable and/or sensitive information is a necessary function of statistical agencies that release microdata files from their sample surveys. Federal laws require that the data collection agencies protect respondents' identifiable or confidential information in public use files(PUFs). Any federal employee who violates these laws is subject to criminal and civil penalties. Also, the potential for disclosure risk could increase refusals in surveys. Thus, many agencies make special efforts to mask their data before releasing PUFs. The types of masking schemes available are: rounding, coarsening(grouping), micro-aggregation, additive noise, multiplicative noise, sampling, multiple imputation, swapping, broadening categories, top- and/or bottom coding, suppression, randomized response techniques, post-randomization method, *etc.*

[1]Corresponding author: Director, Policy Support Division of Statistics Korea, Government Complex-Daejeon, Korea. E-mail: jedomy@korea.kr

Grouping is one of the widely used methods for masking continuous data. Grouped data can be released in the form of tabular data or microdata. Instead of releasing the data in a tabular form only, having microdata available to the public with interval codes in it greatly enhances the utility of the data. It allows the researchers to compute covariance between the variables and build statistical models or to run a variety of statistical tests on the data.

It is well-known that by grouping, we lose precision of estimates. It is also known that, as the number of intervals increases, the variance of the interval data gets closer to that of the ungrouped data. It may be conjectured that the variance of the interval data is lower than that of the ungrouped data in the sense that the coarsened data do not have within interval variance. It is because within each interval, only one value, *i.e.*, midpoint is used for the computation of the variance. However, in some situations, the variance of the coarsened data can be greater than that of the ungrouped data.

When grouping the data, the choice of a measure of centrality in the last interval can cause large differences. Currently, midpoints for all intervals are used to compute the first two moments of the variable. However, the midpoint of the last interval could be much larger than the rest, which could inflate the mean and variance. Instead, we can fit some probability density function to the data and calculate the median of the last interval data, which in turn can be used for computing the first and second moments of the variable. Pareto and lognormal distribution are known to fit economic data well. In this paper, we will demonstrate the above facts using interval data whose underlying distributions are uniform and triangular.

We can take different approaches for developing a representative value for each of the intervals between the last interval and the rest of the intervals. For the last interval, we will try the midpoint and the median assuming uniform, Pareto or lognormal distribution. For the rest, we tried the midpoint and median in the intervals. In this paper we will empirically compare 6 different approaches for representative values of intervals using the Statistics Korea's 2006 Household Income and Expenditure Survey(HIES) data.

2. Effects of Coarsening on Mean and Variance in Theoretical Settings

Continuous data can be grouped into intervals and their interval codes are placed on the micro data file. The interval codes on the micro data file can be used for computing covariance and thus for building models. To use this type of data in any statistical investigations, analysts almost without exception use midpoints of intervals. If the data are grouped into intervals for protecting confidentiality, wider intervals are preferred. However, wider intervals have adverse effects on the data utility.

By creating interval data from the continuous data, we can lose precision of estimates. Using the midpoint for every observation in an interval can be seen as smoothing the data, thus reducing the variance. In the context of analysis of variance(ANOVA), we can see that using midpoints, we can lose the within-group(within-interval) variance component. That is, from the interval data, we can capture the between-variance component only. However, if distributions within some intervals are peaked or skewed, use of the midpoints of the interval data can result in increase in variance compared to the variance of raw data. As the number of intervals increases, the reduction or increase in variance decreases.

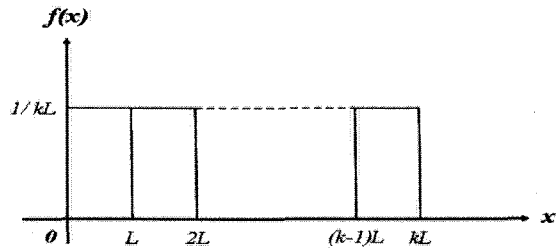

Figure 2.1. Interval data for uniform distribution

2.1. Interval data whose underlying distribution is continuous uniform distribution

Suppose we have a dataset denoted by x that follows a continuous uniform distribution. Suppose we coarsen the data resulting in k intervals of length L. It can be expressed as follows (see, Figure 2.1)(see, Johnson and Kotz, 1970) .

Thus, overall distribution of the data, as well as the distribution of within-interval data is uniform. The probability that x falls in any interval is $1/k$. The first two moments of the raw data are:

$$E(x) = \int_0^{kL} \frac{1}{kL} x\, dx = \frac{kL}{2}, \tag{2.1}$$

$$\mathrm{Var}(x) = \frac{(kL)^2}{12}. \tag{2.2}$$

The corresponding first two moments of the interval data are as follows. Note class marks, *i.e.*, midpoints are used for these derivations.

$$\bar{x} = \frac{1}{k}\left[\frac{L}{2} + \frac{3L}{2} + \cdots + \frac{(2k-1)L}{2}\right] = \frac{kL}{2}. \tag{2.3}$$

The mean of the coarsened data in Equation (2.3) is the same as that for the original data in Equation (2.1)

$$\sum_{i=1}^{k} x_i^2 f(x_i) = \frac{1}{k}\left[\left(\frac{L}{2}\right)^2 + \left(\frac{3L}{2}\right)^2 + \left(\frac{5L}{2}\right)^2 + \cdots + \left(\frac{(2k-1)L}{2}\right)^2\right] = \frac{L^2(4k^2-1)}{12}.$$

Thus

$$\mathrm{Var}(x) = \frac{(kL)^2 - L^2}{12}. \tag{2.4}$$

Comparing the expression in (2.4) with that in (2.2), one can notice that the interval data loses variance by $L^2/12$. $L^2/12$ is indeed the within-interval variance of the data which follows the continuous uniform distribution, where minimum and maximum of x in the interval are 0 and is L, respectively. What the variance in Equation (2.4) suggests is that as the interval length gets shorter, the decrease in variance due to coarsening decreases. Or as the number of intervals increases, the variance of the grouped data gets closer to the original variance (see, Kim *et al.*, 2004; Kim, 2008).

2.2. Interval data whose underlying distribution is triangular distribution

Suppose the data set at hand follows a triangular distribution. The triangular distribution is a very flexible distribution, as it can approximate normal, gamma or beta distribution. The triangular

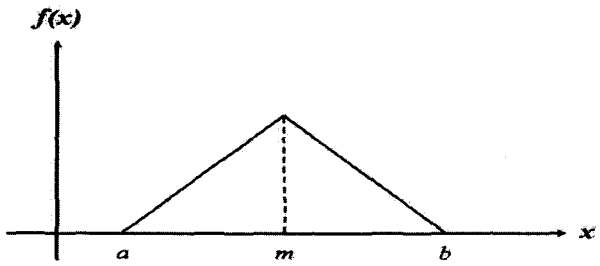

Figure 2.2. Triangular distribution

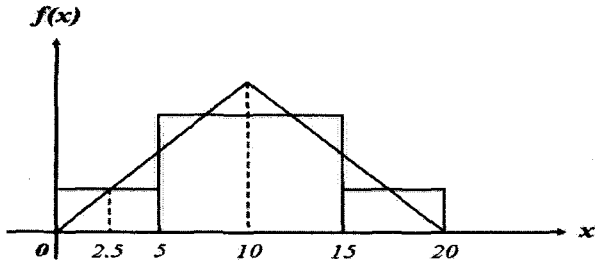

Figure 2.3. Interval data for triangular distribution

Table 2.1. Midpoint and probability of 4 intervals

Interval	Midpoint	Probability
1: [0, 5)	2.5	0.125
2: [5, 10)	7.5	0.375
3: [10, 15)	12.5	0.375
4: [15, 20]	17.5	0.125

distribution in general has the following form (Fiure 2.2)(see, Jeong, 2008).

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(m-a)(b-a)}, & a \le x \le m, \\ \dfrac{2(b-x)}{(b-m)(b-a)}, & m \le x \le b. \end{cases} \tag{2.5}$$

The expected value and variance of x are

$$E(x) = \frac{m+a+b}{3}, \tag{2.6}$$

$$\mathrm{Var}(x) = \frac{m^2 - m(a+b) + a^2 - ab + b^2}{18}. \tag{2.7}$$

Suppose $a = 0$, $m = 10$, $b = 20$. It is a symmetric distribution around m. The mean and variance of the ungrouped data are 10 and 16.67, respectively.

Suppose the data is grouped into four intervals: $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20]$. The probability for x to fall in the first, second, third and last interval is 0.125, 0.375, 0.375 and 0.125, respectively. Thus the interval data can be summarized as follows (Table 2.1).

The mean and variance of this interval data are 10 and 18.75. The mean of this data is the same as that for the ungrouped data. This is since the distribution of the original data is symmetric and

Table 2.2. Sum of squares contributed by each interval

Interval	Sum of squares	
	Original data	Interval data
1: [0, 4)	0.64	0.32
2: [4, 8)	9.60	8.64
3: [8, 12)	36.45	36.00
4: [12, 16)	45.87	47.04
5: [16, 20]	24.11	25.92
total	116.67	117.92

grouping is also symmetric around the mean. The variance of the grouped data is greater than that of the original data.

Suppose the data is grouped into five intervals: $[0, 4)$, $[4, 8)$, $[8, 12)$, $[12, 16)$, $[16, 20]$. The probability that falls in each interval above is 0.08, 0.24, 0.36, 0.24 and 0.08, respectively. The mean and variance of this interval data are 10 and 17.92. Again, the mean of this data is the same as that for the ungrouped data. The variance is still greater than that of the original data, but is lower than that of the four intervals data.

Finally, suppose the data is grouped into ten intervals: $[0, 2)$, $[2, 4)$, $[4, 6)$, $[6, 8)$, $[8, 10)$, $[10, 12)$, $[12, 14)$, $[14, 16)$, $[16, 18)$, $[18, 20]$. The probability that x falls in each of the intervals above is $0.02, 0.06, 0.10, 0.14, 0.18, 0.18, 0.14, 0.10, 0.06$ and 0.02, respectively. The mean and variance of this interval data are 10 and 17. Thus, the mean of this data is again the same as that for the ungrouped data. The variance is greater than that of the original data, but is lower than that of the data with five intervals. In short, the variance of the interval data gets closer and closer to that of the original data as the number of intervals increases or the length of the interval gets shorter.

With the uniform distribution, we observed that coarsening reduces variance. However, with the triangular distribution, an opposite phenomenon was observed. We investigated a reason for this phenomenon. With the uniform distribution, the distribution in each interval was continuous uniform, but with the triangular distribution, no distribution is uniform in any interval. When the within interval distribution is non-uniform, the interval data can have a higher variance than the original data. Sum of squares each interval contributes to the variance is shown in Table 2.2 when 5 intervals are created from the triangular distribution above.

The sums of squares (around zero) for the first interval in the above table are computed as follows: For the original data, $\int_0^4 x^2(x/100)dx = 0.64$ and for the interval data, $(\text{midpoint})^2 \times f(x) = 0.32$.

Each sum of squares for the first three intervals from the top is higher for the original data than for the interval data. However, for the last two intervals, the sum of squares for the interval data is higher than that for the original data. Because the sum of the differences between the two sums of squares for the last two intervals is greater than the sum of the differences for the first three intervals, the total sum of squares for the interval data is higher than that for the raw data. This causes the higher variance for the grouped data.

3. Comparisons of Different Approaches for Developing Representative Values of Intervals

For coarsening, we took different approaches for developing a representative value of each of the intervals between the last interval and the rest of the intervals. For the last interval, we tried the

Table 3.1. Approaches taken for representative values of intervals

Method	Intervals other than last	Last interval
1	midpoint	midpoint
2	midpoint	median(Pareto dist.)
3	midpoint	median(Lognormal dist.)
4	median	median(Uniform dist.)
5	median	median(Pareto dist.)
6	median	median(Lognormal dist.

Table 3.2. Means of grouped data

		Original	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Wage	N	1,967,254	2,112,772	1,944,156	1,961,531	1,949,609	1,944,656	1,962,031
	S	2,133,225	2,182,667	2,103,509	2,122,965	2,105,207	2,107,763	2,127,219
	O	1,941,942	2,102,113	1,919,854	1,936,911	1,925,880	1,919,782	1,936,839
Food costs	N	175,216	228,515	172,236	174,189	172,048	171,880	173,833
	S	194,296	234,354	190,450	192,565	190,227	190,182	192,298
	O	172,460	227,671	169,605	171,534	169,423	169,236	171,166

midpoint and the median assuming uniform, Pareto or lognormal distribution. For the rest, we tried the midpoint and median in the interval. This results in six combinations as shown in Table 3.1. Note that Method 1 is the traditional approach. To determine the number of intervals, we used the following approach. We first assigned the largest 5 percent of observations to the last interval. Then, for the remainder of the data, let k be the number of intervals and n be the sample size. We pick k which is the minimum integer that satisfies $2^k \geq n$. Interval size is then determined by $(\max x - \min x)/k$ (see, Sturges, 1926), where $\max x$ is the maximum x excluding the largest 5percent of observations.

For the last interval, the following maximum likelihood estimator(MLE) of the median of the Pareto distribution was used for Methods 2 and 5.

Suppose that x follows Pareto distribution of the first kind. Then we have

$$P(X \geq x) = \left(\frac{k}{x}\right)^a,$$

where $k = \min x_i$ and a is Pareto constant. Then MLE of the median(m) is $\hat{m} = \hat{k} \times 2^{1/\hat{a}}$, where $\hat{k}$ and $\hat{a}$ are MLE of k and a, respectively, and $\hat{a} = n/\sum_i (\ln x_i - \hat{k})$.

The following maximum likelihood estimator(MLE) of the median of the lognormal distribution was used for Methods 3 and 6.

Suppose that x follows lognormal distribution. Then

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right],$$

where μ and σ^2 are mean and variance of $\ln x$. Then the MLE of the median of x is $\exp(\hat{\mu})$, where $\hat{\mu}$ is MLE of μ, or the sample mean of $\ln x$. Table 3.2 shows the mean of wage and food costs for the six different coarsening approaches.

In the above Table 3.2, N stands for nation, S for Seoul and O for the rest of the country. To help compare the methods, absolute relative differences are computed and shown in Table 3.3.

Table 3.3. Relative difference in percent between means for original and grouped data

		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Wage	N	7.40	1.17	0.29	0.90	1.15	0.27
	S	2.32	1.39	0.48	1.31	1.19	0.28
	O	8.25	1.14	0.26	0.83	1.14	0.26
Food costs	N	30.42	1.70	0.59	1.81	1.90	0.79
	S	20.62	1.98	0.89	2.09	2.12	1.03
	O	32.01	1.66	0.54	1.76	1.87	0.75

Table 3.4. Lower and upper limits of the last interval(in 1,000)

		Lower limit	Upper limit
Wage	N	4,192	20,328
	S	4,700	10,000
	O	4,075	20,328
Food costs	N	424	7,540
	S	438	6,972
	O	421	7,540

Table 3.5. Representative values for the last interval(in 1,000)

		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Wage	N	12,260	4,854	5,179	4,950	4,854	5,179
	S	7,350	5,468	5,841	5,435	5,468	5,841
	O	12,201	4,729	5,051	4,852	4,729	5,051
Food costs	N	3,982	500	538	501	500	538
	S	3,705	520	560	523	520	560
	O	3,981	497	534	498	497	534

Methods 3 and 6 consistently produce means closest to that of the original data. Both assume lognormal distribution for the last interval to get a representative value for the interval. Methods 4, 5 and 2 perform similarly. This suggests that using Pareto distribution is second best for developing a representative value for the last interval. Method 1 is invariably the worst.

The information for the last interval is provided in Tables 3.4~3.5. Table 3.4 provides the lower and upper limits of the last interval for wage and food costs. Table 3.5 provides the representative values of the last interval for wage and food costs.

Note in the above table, the lower limit for Seoul's wage is much higher than other areas. The upper limit for wage for the nation and non-Seoul area is more than twice Seoul's upper limit. Also note that the upper limit for wage is 5 times the lower limit for the nation and non-Seoul area. However, the upper limit for wage for Seoul is just over 2 times the lower limit.

The upper limit for food costs is at least 15.9 times the lower limit. The difference between the lower and upper limit for food costs is much higher than that for wage. This causes much bigger mean food costs than mean wage for Method 1 in comparison with the original data. That is, in Table 3.3, for Method 1 the relative difference for wage is 2.32 to 8.25 percent, while the difference for food costs is 20.32 to 32.01 percent. Also in Table 3.5, Method 1's representative values for wage and food costs are much higher than other methods'.

In Table 3.5, for the nation and non-Seoul area, Method 1's representative values for wage are 2.37 and 2.42 times, respectively, those of Methods 3 and 6. For Seoul, the representative value for Method 1 is 1.26 times those for Methods 3 and 6. For the nation and non-Seoul area, Method 1's

Table 3.6. Standard deviations of grouped data

		Original	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Wage	N	1,238,689	1,729,843	1,150,619	1,194,896	1,155,391	1,144,076	1,188,589
	S	1,371,615	1,504,904	1,270,077	1,321,745	1,251,458	1,258,945	1,310,988
	O	1,215,169	1,761,396	1,129,335	1,172,412	1,138,176	1,123,465	1,166,759
Food costs	N	149,618	392,873	124,863	130,085	125,116	124,743	129,976
	S	169,632	367,032	125,298	130,983	124,924	124,987	130,690
	O	146,297	396,462	124,581	129,741	124,926	124,487	129,656

Table 3.7. Standard deviations of grouped data

		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Wage	N	39.65	7.11	3.54	6.73	7.64	4.05
	S	9.72	7.40	3.64	8.76	8.21	4.42
	O	44.95	7.06	3.52	6.34	7.55	3.98
Food costs	N	162.58	16.55	13.06	16.38	16.63	13.13
	S	116.37	26.14	22.78	26.36	26.32	22.94
	O	171.00	14.84	11.31	14.61	14.91	11.38

representative values for food costs are 7.4 and 7.46 times, respectively, those of Methods 3 and 6. For Seoul, the representative value for Method 1 is 6.62 times those for Methods 3 and 6. In short, Method 1's midpoint in the last interval is much larger than the representative values for other methods which results in artificial bloating of the mean for Method 1, especially for the nation and non-Seoul area.

In Tables 3.6 and 3.7, Method 1 overestimates the standard error of wage by 10 to 45 percent and the standard error of food costs by 116 to 171 percent, which is the worst among all methods. As mentioned before, this overestimation is mainly due to the large midpoint for the last interval. The standard error of food costs is much worse than that for wage, just as we observed with the means, because overestimation of the midpoint of the last interval for the food costs is much worse than that for wage. Methods 3 and 6 consistently outperform the rest of the methods. The standard deviation of Method 3 is slightly closer to that of the original data without exception. Again the performance of Methods 2, 4 and 5 are very similar.

4. Concluding Remarks

Thus far, we have investigated the effects of grouping on the data quality both theoretically and empirically. We investigated grouping for two data sets: one following a continuous uniform distribution and the other triangular distribution. We observed that the within-interval variance gets lost due to grouping of the first dataset whose underlying distribution is a uniform distribution. However, grouping increases the variance for the latter dataset whose underlying distribution is a triangular distribution. The variance of the grouped data approaches that of the ungrouped data if we increase the number of intervals. Traditionally, midpoints of the intervals are used for computing mean and variance for the grouped data. However, the last interval that contains the largest observation can affect the precision of the estimates. We thus compared the traditional approach with other alternatives using the Statistics Korea's 2006 Household Income and Expenditure Survey data. By assuming that the data in the last interval follows the lognormal, Pareto or uniform distribution, we estimated interval median and used it for computing the moments.

In short, the traditional approach which uses the midpoint for the last interval is worst in estimating mean and variance and the approach employing the assumption of the lognormal distribution performs best. Note that our estimates ignored sampling weights. We also ignored the complex sample design. That is, we assumed a simple random sampling for the computations. This is our first attempt for this type of research and we plan to extend this approach to the more complex situations such as weighted estimates and estimates based on the complex sample design.

References

Jeong, D. M. (2008). Schemes for masking the household income and expenditures survey data, Internal Memorandum, Statistics Korea.

Johnson, N. L. and Kotz, S. (1970). *Distribution in Statistics, Continuous Univariate Distributions-1*, John Wiley and Sons.

Kim, J. J. (2008). Probability of Falling in Intervals and Sum of Squares. U.S. National Center for Health Statistics Internal Memorandum.

Kim, J. J., Katzoff, M., Gonzalez, Jr. J. F. and Cox, L. H. (2004). Effects of grouping on first and second distribution moments, *2004 Proceedings of the Survey Research Methods Section*, American Statistical Association, 3808–3815.

Statistics Korea (2006). Household Income and Expenditure Survey.

Sturges, H. A. (1926). The choice of a class interval, *Journal of the American Statistical Association*, **21**, 65–66.