Effects of Grouping Data on First and Second Distribution Moments

Jay J. Kim, Myron Katzoff, Joe Fred Gonzalez, Jr. and Lawrence H. Cox
National Center for Health Statistics

Data such as income are often grouped and released as interval data. Interval data, which is considered to be one of the best ways of summarizing data, can reduce disclosure risk as well. Class marks (midpoints) of intervals are then often used to calculate the mean and variance of the grouped data. In most situations, using midpoints to represent each observation in the interval smoothes the data, thereby reducing the variance. It can be shown, as in analysis of variance, that, by using midpoints we lose the within-interval variance component if within-interval data have a symmetric distribution. However, if distributions within some intervals are peaked or skewed, use of the midpoints of the interval data can result in higher variance estimates than would be obtained with the raw data. Moreover, for those data, the mean of the grouped data based on the use of midpoints is biased. If class (conditional) means are used for calculating overall mean and variance, the mean of the raw data can be recaptured and the variance will be lower. We report some initial results from our study of the impact of accepted practices for approximating moments with summarized data.

Key Words : Interval data, Class Mark, Midpoint, Variance, Data Summarization, Disclosure Risk

I Introduction

Quite often data are grouped into intervals when released to the general public. The reasons are two-fold. First, this is regarded as one of the best ways of summarizing the data. Second, this can reduce the disclosure risk on a micro data file. Interval data can be in a tabular form or in a form of a micro data file. A micro data file can have information on individual persons, families, households, establishments, companies or governmental units. In the latter case, group codes associated with intervals are placed on the micro data file. If the data are grouped into intervals for protecting confidentiality of the units mentioned above, it would be better to use wide intervals. However, this type of grouping has impact on the data utility. Having the interval codes on the micro data file can provide more data utility, since the field can be used for computing correlation and thus for building models. To use that type of data in any statistical investigations, analysts almost without exception use midpoints of intervals, unless the interval means are available.

One can easily suggest that forming interval data from the continuous or multiple category raw data may result in lost accuracy of some statistics such as variance. More specifically, using midpoints for every observation in the interval essentially smoothes the data. That is, every observation in the interval is now represented by its midpoint or interval mean. In the context of analysis of variance (ANOVA), we can see that, using midpoints in the case of uniformly distributed within interval data, or interval means for any within-interval distribution, we will lose the within-group (or within-interval) variance component. In other words, from the interval data, we can capture the between variance component alone. In this paper we will investigate how much variance we lose if we use interval data in place of the continuous or multiple category raw data, assuming, for convenience, discrete uniform distribution on 0, 1, 2, 3, $\ldots, 10^m - 1$, m = 1, 2, 3, . . . We will also compare the variance of the grouped data with that of the raw data when the underlying distribution is not uniform. We will show that the amount of variance lost depends upon the number of intervals. If only a few intervals are used, the reduction in variance can be large; but, as the number of intervals increases, the variance reduction decreases. We will also show the cases where the variance of the grouped data exceeds that of the ungrouped data.

II Uniformly distributed data

Suppose we have a data set consisting of continuous integers, 0, 1, 2, 3, ... $10^m - 1$. We denote $n = 10^m - 1$. Suppose we form k intervals of equal size. We assume n+1 is a multiple of k.

The midpoint of the i^{th} interval is

$$\frac{n+1-k}{2k} + (i-1)\frac{n+1}{k}, \quad i = 1, 2, \ldots k$$

The mean of this interval data is

$$\bar{x} = \frac{1}{k}\left[\frac{n+1-k}{2k} + \left(\frac{n+1-k}{2k} + \frac{n+1}{k} \right) \right.$$

$$\left. + \left(\frac{n+1-k}{2k} + 2\frac{n+1}{k} \right) + \ldots + \left(\frac{n+1-k}{2k} + (k-1)\frac{n+1}{k} \right) \right]$$

$$= \frac{1}{k}\left(k\frac{n+1-k}{2k} + \{1+2+3+ \ \ldots \ (k-1)\}\frac{n+1}{k} \right)$$

$$= \frac{n}{2}.$$

This mean is the same as the mean of the raw data.

$$\frac{1}{k}\sum x^2 = \frac{1}{k}\left[\left(\frac{n+1-k}{2k} \right)^2 + \left(\frac{n+1-k}{2k} + \frac{n+1}{k} \right)^2 + \ldots \right.$$

$$\left. \ldots + \left(\frac{n+1-k}{2k} + (k-1)\frac{n+1}{k} \right)^2 \right]$$

After some algebra, the above reduces to

$$\frac{1}{12k^2}\left[n^2(4k^2-1) + 2n(k^2-1) + (k^2-1) \right]$$

Hence

$$V(x) = \frac{1}{12k^2}\left[n^2(4k^2-1) + 2n(k^2-1) + (k^2-1) \right] - \frac{n^2}{4}$$

$$= \frac{k^2-1}{12k^2}(n+1)^2.$$

Note the variance formula for the raw data is

$$V(x) = \frac{(n+1)^2-1}{12}.$$

The ratio of the variance of the interval data to that of the raw data, ignoring 1/12 in the raw data formula, is

$$\frac{k^2-1}{k^2}.$$

What the above formula suggests is that, as the number of intervals, k, gets larger, the ratio gets closer to 1, or the variance of interval data gets close to the variance of raw data. For example, if k = 10, the above ratio is .99. That is, if 10 intervals are used, the loss in variance is about one percent of the variance of raw data.

III Micro-aggregation

Micro-aggregation is a type of interval data. As mentioned before, interval data can be shown in two forms. One is in typical tabular form. The other is in micro-data in which a code associated with each interval is given. By aggregating the values of the codes, data users can come up with a table. Micro-aggregation, broadly speaking, belongs to the latter. It is used for micro-data, in which the mean of each interval, rather than a code, is shown for each observation in an interval. For example, when a field can have values 0, 1, 2, 3, . . 9, some U.S. federal agencies use micro-aggregation as follows. Keep zero (0) as zero and form three intervals 1-3, 4-6 and 7-9. Mean (midpoint) of the intervals above are 2, 5 and 8, respectively. Replace each original value by the interval mean, depending on which interval the original value belongs to. In this case, the variance of the micro-aggregated variable does not capture the total variance of raw data. For the example given above, we compute the mean and variance.

Table 1. Intervals and midpoints

Interval	Midpoint
0	0
1 - 3	2
4 - 6	5
7 - 9	8

We assume again the uniform distribution of the raw data. Thus, the probability that 2, 5 or 8 appears is 3/10. The mean of the this micro-aggregated data is 4.5, which is the same as the mean of the raw data. However, the variance of the micro-aggregated data is 7.65, while that of the raw data is 8.25. Thus the relative loss of variance due to this micro-aggregation is

$$\frac{8.25-7.65}{8.25} \text{ x } 100 = 7.27\%$$

IV Within-interval variance

Suppose we have observations 0, 1,2,3, . . , 10. We will assume different distributions for the observations and check on the within-interval variance.

1. The uniform distribution

Suppose the data above follows a uniform distribution. Then the variance of the data is

$$V(x) = \frac{11^2 - 1}{12} = 10 .$$

2. Quadratic distribution

Suppose the data follow $p(x) = \frac{(x-5)^2}{\sum_x (x-5)^2}$. We denote $(x-5)^2$ by g(x).

Table 2. Probability Mass

x	g(x)	p(x)
0	25	25/110
1	16	16/110
2	9	9/110
3	4	4/110
4	1	1/110
5	0	0
6	1	1/110
7	4	4/110
8	9	9/110
9	16	16/110
10	25	25/110
Sum	110	1

The mean of the above data is 5 and its variance is 17.8. In comparison with the variance of the uniformly distributed data, its variance is 78 % times larger. The reason why the data following the quadratic distribution has higher variance is that the relative mass of x sharply increases as x departs further from its mean.

3. The triangular distribution

The triangular distribution can be defined as follows:

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

For the same data set as in Table 2, we can assume the triangular distribution. Then we have m = 5, a = 0 and b = 10. In this case we have mean = 5 and variance = 4. Variance of 4 is much less than 10 for the uniform distribution.

What we can glean from these observations is that if the data points are concentrated around mean, the variance of the data can be smaller than that of data that follow the uniform distribution. However, if the data points are heavily concentrated around the ends of the interval, the variance of the data is greater than that of the uniform distribution.

V Whole data set does not follow the uniform distribution

Example 1.

Suppose that we have a dataset in Table 3 in the Appendix, where p(x) is the probability mass of x. Graph 1 in the Appendix shows this distribution, which has an extremely high peak and flat tails. This (raw) data set has mean of 9.5 and variance of 8.25. Mean sum of square is 98.5. Its kurtosis is 6.03, which is twice that of the normal distribution. Skewness is zero.

Suppose we use four intervals of equal size. Then we can have the following table.

Table 4. Intervals and Midpoints

Interval	Midpoint	Probability
0 - 4	2	.05
5 - 9	7	.45
10 -14	12	.45
15 -19	17	.05

The mean and variance of the above interval data, using mid-points, are 9.5 and 11.25, respectively. The mean sum of squares using mid-points is 101.5. Note this interval data has higher variance (11.25) than the raw data (8.25). This is opposite to what we have observed with the data following the uniform distribution.

To investigate the reason for this phenomenon, the mean sum of squares was computed within each interval for both the interval and the raw data. They are as follows.

Table 5. Mean Sum of Squares

Interval	Interval Data	Raw Data
0 - 4	0.20	0.30
5 - 9	22.05	30.90
10 -14	64.80	52.75
15 -19	14.45	14.55
Sum	101.05	98.50

Comparing the sum of squares between the interval and

the raw data within each interval, one can notice that in the "interval 10-14", the interval data have a higher mean sum of squares than the raw data. This is caused by the fact that, in that interval, 86.7 percent of the frequency is on 10 and 11, but the midpoint is 12. Consequently, that high frequency is applied to 12, which is greater than 10 and 11, resulting in the phenomenon. Note that in interval 5-9, 86.7 percent of the frequency is on 8 and 9 and the midpoint is 7. This time the high frequency is applied to 7, which is lower than 8 and 9, making the sum of squares for the interval lower than that for the raw data. When the interval is flat, interval data has lower sum of squares.

If we use the mean for each interval for computing overall mean and variance, then the resulting variance is lower than that of the raw data. For example,

Table 6. Intervals and Interval Means

Interval	Interval mean	Probability [f(x)]
0 - 4	2	.05
5 - 9	8.22	.45
10 - 14	10.78	.45
15 - 19	17	.05

In the second and third intervals of the table, the interval mean was calculated as $\frac{\sum xf(x)}{\sum f(x)}$, since $\sum f(x) \neq 1$ within each interval. The data in the first and last interval follow uniform distributions. Using the interval means, we have mean = 9.5 and variance = 7.09. Note this variance is smaller than 8.25. Analogy to the ANOVA applies to the variance based on these interval means.

Example 2.

The following example is cited from Neil Weiss' Introductory Statistics (p 53). A pediatrician who tested the cholesterol levels of several young patients was alarmed to find that many had levels over 200 mg per 100mL. The readings of 20 patients with high cholesterol levels are presented below.

Table 7. Cholesterol levels for 20 high-level patients

210	209	212	208
217	207	210	203
208	210	210	199
215	221	213	218
202	218	200	214

The mean and variance of the above data is 210.2 and 35.16.

From the above data set, the author created a frequency table using a class width of five and starting at 195. The result is as follows.

Table 8. Classes and frequencies for the cholesterol level data in Table 7.

Cholesterol level	Frequency
195 - 199	1
200 - 204	3
205 - 209	4
210 - 214	7
215 - 219	4
220 - 224	1
Sum	20

The mean and the variance of the above grouped data is 210.25 and 38.19. Note there is almost no difference in the mean between the grouped and the raw data. However, the variance of the grouped data is 8.62 percent higher than the raw data. Thus if a user tries to do a hypotheses testing using the grouped data, he might get a different conclusion than that from the raw data.

Example 3.

Suppose we have observations 0,1, 2, 3, 4, . . . 20, which follows a triangular distribution. See Graph 2 in the Appendix. The mean and variance of the above are 10 and 16.5. We can have the following grouped data.

Table 9. Grouped data from the above

Interval	Frequency
0 - 6	.21
7 - 13	.58
14 - 20	.21

From this data we get mean = 10 and variance = 20.58. The relative increase in variance due to grouping is

$$\frac{20.58 - 16.5}{16.5} = 22.9\ \%\ .$$

If we use more than 3 intervals, the increase in variance would be decreased. That is, the estimate would be more precise. But the disclosure risk will increase. It is interesting to find the optimum number of intervals in light of variance and disclosure risk, which is our future

research topic.

Example 4. From the data which follow the quadratic distribution, we create interval data as follows.

Table 10. Grouped data that follow quadratic distribution

Interval	Frequency
0 - 3	.49
4 - 6	.02
7 - 10	.49

This data set has mean = 5 and variance of 12.005, which is 20 % higher than that of the raw data.

VI Comparison of variance of grouped data with that of the raw data – Normal distribution

Spruill and Gastwirth imply that the within-group variance of the grouped data is close to zero. Assuming that the underlying distribution of the raw data is normal, we will investigate whether the claim is valid. Note, in their case, the within-group variance is based on the group (interval) means.

Suppose x follows the standard normal distribution. Spruill and Gastwirth used equal size in terms of probability for each interval. Thus we will have the probability of .10 for each interval. As the standard normal distribution is symmetric around zero, we will deal with the positive side only and the final results are multiplied by two. Note that $\Phi^{-1}(.50) = 0$, $\Phi^{-1}(.60) = .25335$, $\Phi^{-1}(.70) = .5244$, $\Phi^{-1}(.80) = .84162$ and $\Phi^{-1}(.90) = 1.28155$, where Φ is the cumulative normal probability. We assume $\Phi^{-1}(1.0) = 3.50$. The five intervals are (1) 0 - .25335, (2) .25335 - .5244, (3) .5244 - .84162, (4) .84162 – 1.28155 and (5) 1.28155 – 3.5.

VI.1 Variance based on the group means

The interval means are calculated using the following formula.

$$\frac{\int_a^b x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}{\Phi(b) - \Phi(a)} \quad .$$

That is, we are in the truncated normal situation. Means for the intervals are, in ascending order, (1) .126, (2) .3865, (3) .67731, (4) 1.04464 and (5) 1.74626, respectively. The variance of the grouped data using the group means is

$$V(x) = .95294.$$

Note, the raw data has variance of 1.0. In terms of the analysis of variance, the above variance is the between-group variance or the variance conditional on the group means. Thus we are missing the within-group variance of .04706. The ratio of the within-group variance to the between-group variance is .04939.

VI.2 Variance based on the class marks (midpoints)

If the group means are not released by the data collectors, users must use the class marks for computing the mean and the variance of the grouped data. Here we assume the group means are not available. The midpoints of the five groups are (1) .12667, (2) .38887, (3) .68301, (4) 1.06159 and (5) 2.39078, respectively. The mean of the grouped data based on the midpoints is zero (0). The variance of the grouped data based on the midpoints is 1.49531. This means the grouped data has 49.5 percent higher variance than the raw data, if the midpoints are used for computing the variance.

Lemma. The underestimation of the variance of 4.7 percent, when computation is based on the group means, and the overestimation of the variance of 49.5 percent, when the midpoints are used for the calculation can be observed for all normal distributions, when 10 intervals of equal size in probability are used.

The variance of a variable is invariant under the scale and location change (or linear transformation), and thus the ratio of two variances remain the same throughout the linear transformation.

VI.3 Intervals are equally spaced except two extreme ones.

In this example, we use equally spaced intervals along the x axis, except for the far right and left intervals. We use 10 intervals. The length of each interval is .4 on the x axis, except for the two mentioned above. Five intervals on the positive side of the normal curve are: (1) 0 - .4, (2) .4 - .8, (3) .8 – 1.2, (4) 1.2 – 1.6 and (5) 1.6 - 3.5, respectively. The probability of each interval above in the corresponding order is (1) .1554, (2) .1327, (3) .0968, (4) .0603 and (5) .0548, respectively.
The group means based on the truncated normal distribution for each interval are: (1) .19738, (2) .59215,

(3) .98663, (4) 1.3809 and (5) 2.008.

The mean of x is zero and the variance of x based on the group means is .96557. The within-group variance is .03443. Note this is lower than the one which is based on the intervals that have the same probabilities.

The variance of x based on the midpoints is 1.251. In other words, using the grouped data, the users will have 25.1 percent higher variance. Again this variance is lower than the case where each interval has a probability of .10.

VII Concluding remarks

When disclosure risk issue arises for a micro data file, one of the most frequently suggested methods to limit the risk is to form intervals for continuous or multiple category data. For example, when a file has a field which can have 0, 1, . . . 9, three intervals 1 - 3, 4 – 6, and 7 – 9 are formed and midpoints of the three intervals, 2, 5 and 8, respectively, are given for each observation in each interval. Note zero (0) remains zero. This is practiced by some U.S. federal government agencies. As we observed before, data users will get a variance which is underestimated by 7.27 % from this grouped data. Note in the above calculation, we assumed that the observations follow the uniform distribution.

Again assuming the uniform distribution for the observations, if k intervals are used, around $\frac{1}{k^2}$ x 100 percent of the variance of the raw data will be lost. This suggests that care should be taken in including these interval data in a statistical model. Or any statistical analysis involving interval or micro-aggregated data should be performed or interpreted with special care.

In the context of the analysis of variance (ANOVA), the total sum of squares (or variance) can be decomposed into the within- and between-group sum of squares (variance). Note in the calculation of within- and between-variance, group means are used. On the other hand, in computing variance of the interval data, most often class marks, or midpoints are used. In the case of the uniform distribution within the intervals, the midpoints are the same as the group means. That is why the variance of the grouped data is always lower than that of the raw data.

We compared the variance of the grouped data to that of the raw data, when the raw data do not follow uniform distribution. The distributions considered include the triangular distribution and the quadratic distribution. We also considered a distribution which has a very high top and is symmetric. For all the cases we considered, the variance of the grouped data was greater than that of the raw data, when the mid-points are used for the computation. In the case where the data have a high top (kurtosis of 6.03), the interval data have 36.4-% higher variance, when four intervals are used. In the case of the triangular distribution which has 21 data points only, the variance of the grouped data is 22.9-% higher when three intervals are used. But when seven intervals are used, the increase in variance decreases to 4.73-%. This confirms that if we increase the number of intervals, we can lower the difference in variance between the grouped and the raw data, disregarding the underlying distribution. Using the normal distribution, the variance of the grouped data and that of the raw data were compared, when 10 intervals were used. Two different approaches are employed to form 10 intervals. One is based on the probability. That is, intervals are made such that each of them has a probability of .10. The other is based on the equal size (.4 on the standard normal distributed variable) on the x axis, except the two extreme intervals. The variance of the interval data was calculated using the interval means and class marks, respectively. The grouped data, based on the equal probability and the group mean, provided around 5 % lower variance than the raw data. However, when midpoints are used, the grouped data provided around 50 % higher variance than the raw data.

When the variance of the grouped data is calculated based on the same interval size on the x-axis except for the two extreme intervals, the loss in variance is around 3.4-percent, when the group means are used, but the gain in variance is around 25-percent, when the class marks are used.

Based on the observations above, data disseminating agencies could consider computing the variance before and after grouping or aggregation and informing the users of any difference in variance. In the above, we used a simple random sampling approach, but the data collected are usually from a complex multi-stage cluster sample. Thus computing and releasing the design effects for the raw data should be considered. Some data users use sample weights, but others not. Thus agencies could repeat the calculations for both weighted and unweighted data.

VIII References

Cramer, J.S. (1964). Efficient Grouping, Regression and Correlation in Engel Curve Analysis. The Journal of the American Statistical Association, Vol. 59, pp 233-250.

Johnson, David (1997), The triangular distribution as a proxy for the beta distribution in risk analysis, The Statistician, 46, 387-398.

Johnson, Normal L. and Kotz, Samuel (1969). Distributions in statistics. Discrete distributions. Houghton Mifflin Company.

Mood, Alexander M, Graybill, Franklin A. and Boes, Duane C. (1974). Introduction to the theory of statistics. 3rd Ed. McGraw-Hill Book Company, 538.

Spruill, Nancy L. and Gastwirth, Joseph L. (1982). On the estimation of the Correlation Coefficient from Grouped Data, The Journal of the American Statistical Association, Vol. 77, No. 379, pp 614-620.

Weiss, Neil A. (1997). Introductory Statistics. Addison-Wesley Publishing Company, Inc. 4th Ed.

VIII Appendix

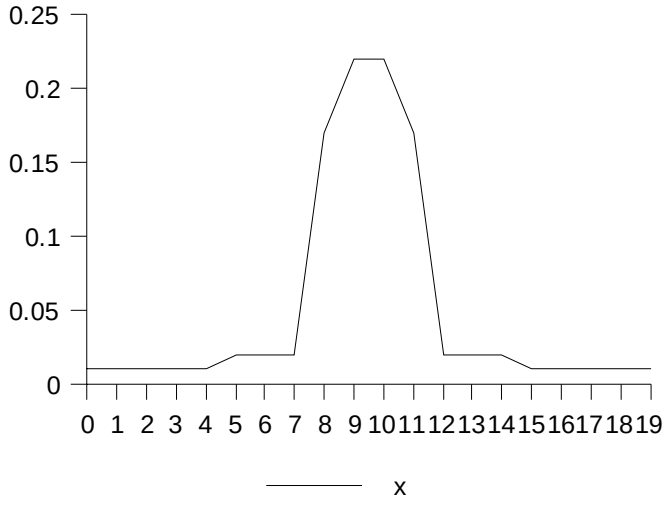
Table 3. Probability masses

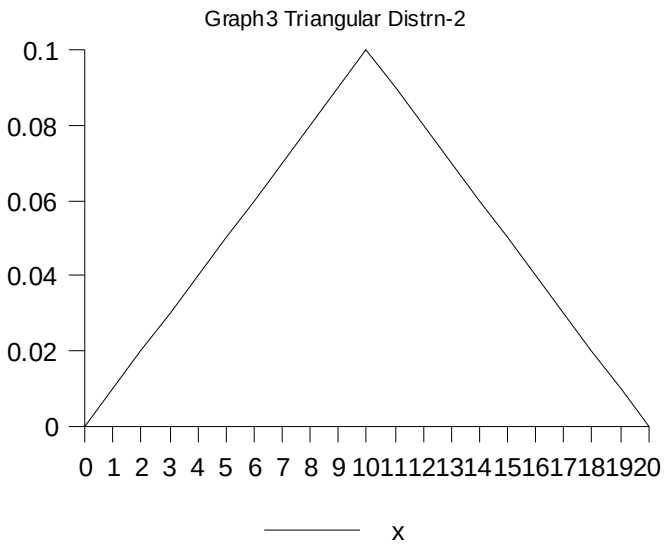
x	0	1	2	3	4	5	6	7	8	9
P(x)	.01	.01	.01	.01	.01	.02	.02	.02	.17	.22

Continuation of Table 3

x	10	11	12	13	14	15	16	17	18	19
P(x)	.22	.17	.02	.02	.02	.01	.01	.01	.01	.01

Graph 1. Frequency Distribution




Graph3 Triangular Distrn-2
0.1
0.08
0.06
0.04
0.02
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
x