Quality and Quantity, 16 (1982) 19–27
Elsevier Scientific Publishing Company, Amsterdam—Printed in The Netherlands

Note

Calculating minimum and maximum possible variances from *n*-tile grouped data

STEPHEN ROSSKAMM SHALOM
Department of Political Science, William Paterson College of New Jersey

GEORGE MANDEVILLE
Department of Physics, William Paterson College of New Jersey

Summary social science data are frequently encountered for which the raw data are unavailable and for which the standard deviations are unreported. This often precludes further analysis or even meaningful use of the summary data. At times, however, the published data include the means for different quartile, quintile, decile, or, in general, *n*-tile groups. This is particularly the case for income distribution figures, which will often report the average income of the bottom 20% of income units, of the second 20%, etc. We demonstrate here that when such submeans are known it is possible to calculate minimum and maximum possible values for the standard deviation or variance, which will then permit tests of statistical significance to be employed.

Organization of the Data

Let N observations be divided into n mutually exclusive groups, with x_{ij} being the jth observation of the ith group. We shall call these "n-tile groups" when $x_{ij} \leqslant x_{i+1,k}$ for all i, j, and k. Let the ith n-tile group contain $w_i N$ observations, where $0 < w_i \leqslant 1$ and $\Sigma_i w_i = 1$. (Note that each n-tile group need not contain the same number of observations.) Let the mean of the $w_i N$ observations in the ith n-tile group be denoted m_i.

Sometimes data are encountered in which the observations of one n-tile group are a subset of another. (For example, the average income of each income quintile might be given as well as that of the top 5%.) Such cases can readily be made compatible with the definition of mutually exclusive n-tile groups above, while still retaining the maximum amount of information.

Assume that n-tile group A is a subset of n-tile group B. Then an n-tile group C can be defined containing only those observations in B that are not in A. C will contain $(w_B - w_A)N$ observations and will have a mean given by $(w_B m_B - w_A m_A)/(w_B - w_A)$. Then the n-tile groups C and A will be mutually exclusive and between them will include all the observations in B. In a like manner, all overlapping n-tile groups can be eliminated and the maximum amount of information derived from them. Hereafter, we will assume that all data are in a form consistent with the definitions of the previous paragraph.

In all of the derivations that follow, we make no assumptions about the nature of the distribution of the observations, such as normality, symmetry, unimodality, etc. In other words, the derivations are entirely distribution-free.

Given only the mean for a variable, the minimum possible variance will equal zero, which occurs when all of the observations fall precisely on the mean. But as soon as two distinct submeans – for example, the mean for the lower 50% of observations and the mean for the upper 50% of observations – are given, then the variance must be greater than zero, since there must be some dispersion around the grand mean.

Similarly, given only the mean for a set of data, the maximum possible variance is infinite, since there is no limit upon how far the observations may be located from the mean. When two submeans are given, however, the maximum variance becomes circumscribed. This is because the lower group of observations must balance about the lower submean and none may exceed the upper submean. The latter requirement follows from the fact that we are employing n-tile group means which include as part of their definition the condition that all observations in the ith n-tile group are less than or equal to all observations in the ith $+ 1$ n-tile group. If we were using means of other kinds of subgroups – say, geographic regions – each subgroup could in principle have unlimited dispersion and the overall maximum possible variance would be infinite.

In addition, knowledge of an upper or lower limit on the observations – such as the fact that many variables must be non-negative – will further reduce the magnitude of the maximum possible variance.

The Minimum Variance

The derivation of an expression for the minimum possible variance will be familiar to those acquainted with analysis of variance (e.g., Hays, 1973, pp. 465–67).

Let x_{ij} represent the jth observation in the ith n-tile group. Let m_i be the mean of the ith n-tile group, $w_i N$ the number of observations in the ith n-tile

group, and N the total number of observations. The grand mean then is $M = \Sigma w_i m_i$. The variance is defined as

$$\sigma^2 = \sum_i \sum_j (M - x_{ij})^2 / N \tag{1}$$

Hence

$$\begin{aligned} N\sigma^2 &= \sum_i \sum_j (M - x_{ij})^2 \\ &= \sum_i \sum_j \left[(M - m_i) + (m_i - x_{ij})\right]^2 \\ &= \sum_i \sum_j (m_i - x_{ij})^2 + 2\sum_i (M - m_i) \sum_j (m_i - x_{ij}) + \sum_i \sum_j (M - m_i)^2 \end{aligned} \tag{2}$$

But each of the sums $\sum_i (M - m_i)$ and $\sum_j (m_i - x_{ij})$ must be zero, since they are both sums of deviations about a mean. The last term on the right-hand side of eqn. (2) can be written

$$\sum_i \sum_j (M - m_i)^2 = \sum_i w_i N (M - m_i)^2 = N \sum_i w_i (M - m_i)^2$$

Substituting these results in eqn. (2), and solving for the variance, gives

$$\sigma^2 = \left[\sum_i \sum_j (m_i - x_{ij})^2 / N \right] + \sum_i w_i (M - m_i)^2 \tag{3}$$

The first term on the right-hand side represents the deviations of the observations about their submeans; the second term represents the deviations of the submeans about the grand mean. For a given set of submeans, the latter term will remain constant, since it is strictly a function of the parameters m_i and w_i (recall that $M = \sum_i w_i m_i$). Since $\sum_i \sum_j (m_i - x_{ij})^2 / N$ is a sum of squares divided by a positive number, it must be non-negative. The minimum value of σ^2 will thus occur when $\sum_i \sum_j (m_i - x_{ij})^2 / N$ equals zero. Therefore

$$\sigma^2_{\min} = \sum_i w_i (M - m_i)^2 \tag{4}$$

That is, the variance will be a minimum when the deviations of the observations about the submeans are zero, i.e. when all the observations fall on the submeans.

The Maximum Variance

Let N observations x_i have mean M and be bounded on both sides such that $L \leqslant x_i \leqslant H$ for all i. We call these bounds "primary restrictions" when it

is consistent with N and M that there can be at least one x_i equal to L and one x_i equal to H.

If the value of M is free to vary, then the maximum variance of the x_i's can be simply expressed as a function of the range $R = H - L$: $\sigma^2_{\max} = R^2/4$ (see Hammond and Householder, 1962, pp. 130–31). But when M is fixed, an alternative expression must be derived.

The x_i's will be maximally dispersed about M when all are at L and H [1]. Let K out of the N observations be located at L, and $(N-K)$ at H. By definition of the mean,

$$K(M-L) = (N-K)(H-M)$$

Solving for K gives

$$K = N(H-M)/(H-L)$$

The variance for this maximal dispersion will be

$$\sigma^2_{\max} = \left[K(M-L)^2 + (N-K)(H-M)^2\right]/N$$

Substituting for K and simplifying, the variance is obtained as

$$\sigma^2_{\max} = (H-M)(M-L) \tag{5}$$

Now consider the case of N x_{ij}'s divided into n n-tile groups, each containing $w_i N$ observations and having known means m_i, $i = 1,\ldots,n$. Let L and H be known values such that $L \leqslant x_{ij} \leqslant H$ for all x_{ij} and assume that these bounds are primary restrictions as defined above. We now introduce variables which represent the partitions (or dividing lines) between the n-tile groups. The partitions p_i have the following properties:

$$x_{ij} \leqslant p_i \qquad \text{for all } j; \tag{6a}$$

$$x_{i+1,j} \geqslant p_i \qquad \text{for all } j; \tag{6b}$$

$$m_i \leqslant p_i \leqslant m_{i+1} \tag{6c}$$

In terms of the p_i, each n-tile group is now bounded and has a maximum variance within that constraint as defined by eqn. (5), i.e.

$$\sigma^2_{i_{\max}} = (p_i - m_i)(m_{i-1}) \tag{7}$$

where L and H have been set equal to p_0 and p_n, respectively. Since the variance for the ith n-tile group is

$$\sigma_i^2 = \left[\sum_j (m_i - x_{ij})^2\right]/w_i N \tag{8}$$

multiplying by w_i and summing over i will produce the first term on the right-hand side of eqn. (3). Therefore, substituting into (3) gives

$$\sigma^2 = \sum w_i \sigma_i^2 + \sigma^2_{\min} \tag{9}$$

Substituting eqn. (7) into (9) gives the maximum with respect to variations of the data between, but not across, the partitions:

$$\sigma^2 = \sigma^2_{\min} + \sum_i w_i(p_i - m_i)(m_i - p_{i-1}) \tag{10}$$

This, however, is not the maximum variance of the data, but the maximum variance for a given set of partitions p_i. Expansion of the sum in eqn. (10) in terms of the p_i shows that σ^2 depends linearly on each p_i as each of the remaining $p_j, j \neq i$, is held fixed. Consequently, σ^2 must increase or decrease monotonically with each p_i [2]. Beginning with any set of p_i's, at least one of which is not at the limit expressed by eqn. (6c), σ^2 can always be increased by moving this p_i toward one or the other of its limits. Therefore, σ^2 cannot be at a maximum unless all p_i are at their limits, which places each p_i coincident with an m_i. In addition, since all the data of each n-tile group are concentrated at the bounds of the n-tile group, i.e., at the p_i's, then all the data must exist at m_i's as well. At first glance this may appear to be identical to the minimum variance condition; the difference is that for $\sigma^2_{\max}$ all data are at m_i's, but not all m_i's have data.

For the ith n-tile group either (i) all of the data are at m_i ($x_{ij} = m_i$, for all j) or (ii) all of the data are at neighboring means ($x_{ij} = m_{i-1}$ or m_{i+1}, for all j). The former case (i) contributes nothing to the sum in eqn. (10) and does not impose any restrictions on the distributions in the other n-tile groups. The latter case (ii) contributes a variance equal to $w_i(m_{i+1} - m_i)(m_i - m_{i-1})$ to the sum in eqn. (10) and forces all data in the adjacent n-tile groups to be located at the corresponding means: i.e., the data are distributed as in the former case. Maximum variance occurs when a particular set of non-consecutive n-tile groups contributes maximum variance (case (ii)) and the remaining n-tile groups contribute zero variance (case (i)).

If known values of L and H are unavailable the data can still be made to conform to this analysis by computing effective lower and upper bounds based only upon limitations imposed by the number of observations in the first and last n-tile groups and by the means. In the first n-tile group, the maximum possible value is m_2 (the mean of the second n-tile group) [3]. The minimum value will occur when all of the observations in the first group except one are located at m_2. Let the remaining observation be x_{1a}. Then

$$m_1 = \sum_j x_{ij}/w_1 N = \left[(w_1 N - 1)m_2 + x_{1a}\right]/w_1 N$$

Solving for x_{1a}, which can be used as the value of p_0 ($= m_0 = L$) in eqn. (10), gives

$$x_{1a} = p_0 = m_0 = L = w_1 N m_1 - (w_1 N - 1)m_2 \tag{11a}$$

Likewise,

$$H = p_n = m_{n+1} = w_n N m_n - (w_n N - 1) m_{n-1} \tag{11b}$$

The task of determining $\sigma^2_{\max}$ then becomes that of selecting which n-tile groups should contribute to the variance and which should not. For the general case of n n-tile groups, the contribution of any n-tile group to the variance will be either zero or

$$s_i^2 = w_i (m_{i+1} - m_i)(m_i - m_{i-1})$$

which can be abbreviated somewhat by defining $\Delta_i = m_i - m_{i-1}$, etc., from which

$$s_i^2 = w_i \Delta_{i+1} \Delta_i$$

The maximum variance $\sigma^2_{\max}$ can then be found by forming a table of Δ_i's and s_i's and selecting – by trial and error, intuition, or computer program – the non-consecutive s_i's that produce the largest sum. By using non-consecutive s_i's, no Δ_i is used twice. The sum of the s_i's is then added to $\sigma^2_{\min}$ to obtain $\sigma^2_{\max}$.

Decision Rules with Minimum and Maximum Variances

The computed minimum and maximum variances of a given set of n-tile grouped data can be used in any standard test of statistical significance. A result that is statistically significant using the *maximum* variance will be significant for any variance. A result that is not statistically significant using the *minimum* variance will not be significant for any variance. When a result is significant using the minimum but not when using the maximum variance, it is not possible to decide whether the actual variance will yield a significant result or not.

Application

The Philippine Bureau of the Census and Statistics (1973, p. xxi) has published data on Philippine average family income by deciles [4]. The data are based on sample surveys and the sample size is given (p. ix). No standard deviations are reported. Table I shows the data for 1961 and 1965, converted to constant 1965 pesos (using a consumer price index; Central Bank, 1972, p. 372).

We wish to test whether there has been a significant change from 1961 to 1965 in the real income of (a) all families, (b) the bottom 90% of families, (c) the bottom 40% of families, (d) the bottom 30%, and (e) the bottom 20%. To

TABLE I

Philippines, Average Family Income, Constant 1965 Pesos

		Income	
		1961	1965
Lowest	10% of families	338	293
2nd	10% of families	607	601
3rd	10% of families	764	880
4th	10% of families	1012	1161
5th	10% of families	1237	1458
6th	10% of families	1484	1804
7th	10% of families	1866	2272
8th	10% of families	2473	2851
9th	10% of families	3484	3910
top	10% of families	9218	10178
All families		2249	2541
Sample size		6977	4747

this end we perform conventional, large-sample, two-tailed tests of significance for a difference between means with a 0.001 level of significance. Note that although no assumptions are made about the distribution of the population data, by the central-limit theorem the sampling distribution of the means will be normally distributed and thus the z statistic is appropriate.

To test whether there has been a change in overall family income, it is necessary first to determine upper and lower limits. The upper limits are computed using eqn. (13b) and come to 4.00×10^6 for 1961 and 2.98×10^6 for 1965; the lower limits are set equal to zero because negative incomes are assumed to be impossible. Now eqn. (4) and the procedures outlined above give

	1961	*1965*
$\sigma^2_{\min}$	6.20×10^6	7.56×10^6
$\sigma^2_{\max}$	2.29×10^{10}	1.86×10^{10}

The test statistic may then be calculated as

$$z = (2541 - 2249) / \left[\left(\sigma^2_{1961} / N_{1961} \right) + \left(\sigma^2_{1965} / N_{1965} \right) \right]^{1/2}$$

$$= \begin{cases} 5.86 \text{ with minimum variances} \\ 0.11 \text{ with maximum variances} \end{cases}$$

Since one value of z is greater than 3.29 and the other is less than 3.29, in

this case it is not possible to decide whether the actual variance will give a z value above or below 3.29. Thus, no conclusion can be drawn without making additional assumptions about the distribution of the data. This ambiguous result will often occur when one of the limits (here the upper limit) has such an extreme value.

For the situation of the bottom 90% of families, there is a lower limit of 0 and an upper limit equal to the mean for the tenth income group (since no observation in the ninth decile can exceed the mean for the tenth decile). $\sigma^2_{\min}$ and $\sigma^2_{\max}$ are now, for 1961, 8.90×10^5 and 7.10×10^6, and for 1965, 1.20×10^6 and 8.38×10^6, respectively. Calculating z values (remembering to use 0.9 times the sample size) gives

$$z = \begin{cases} 10.57 \text{ with minimum variances} \\ 3.90 \text{ with maximum variances.} \end{cases}$$

Since both values are greater than 3.29 we may conclude that z will be greater than 3.29 for all possible variances and therefore that there has been a significant change in the mean income of the bottom 90% of families.

A similar procedure for the bottom 40% of families gives

$$z = \begin{cases} 6.19 \text{ with minimum variances} \\ 3.64 \text{ with maximum variances} \end{cases}$$

Again, since both values are above 3.29 we may conclude that there has been a significant change from 1961 to 1965 for the bottom 40%.

For the bottom 30%, $z = 2.83$ with minimum variances and $z = 1.37$ with maximum variances. Here both values of z are less than 3.29 and hence we may conclude that the difference between the means (570 in 1961 and 591 in 1965) may be attributable solely to sampling error.

Finally, for the bottom 20% of families the mean income declined from 473 to 447. The z values, of 4.13 with minimum variances and 1.81 with maximum variances, leave us in the ambiguous situation where no conclusion can be drawn.

The results may be summarized as follows. Although no standard deviations were reported for the Philippine income data, we can conclude that there were definite changes in the incomes of the bottom 90% and the bottom 40% of families from 1961 to 1965; the apparent change in income for the bottom 30% may be attributable to sampling error; and for the entire sample of families and for the bottom 20% the method of calculating minimum and maximum variances does not permit any conclusions to be drawn [5].

Notes

1 For discrete variables, when N does not divide evenly into $(H-L)$, one x_i will fall between H and L. The maximum variance will then be smaller than in the calculation that follows. For large N, this difference is negligible.

2 One exception to this exists. If the coefficient of p_i is zero, σ^2 is completely independent of p_i and the value of p_i is arbitrary.

3 An argument can legitimately be raised that the upper limit on the first n-tile group is not m_2, but in fact p_1. That p_1 is likely to coincide with m_2 neglects the fact that the above analysis is based on the linear dependence of σ^2 on the p_i's. If p_0 is expressed in terms of p_1 in eqn. (12), the dependence on p_1 becomes quadratic. We state without proof that while such dependence significantly complicates the analysis, the results are unaffected. The partial derivative of σ^2 with respect to p_i vanishes at $p_1 = m_1$, but this extremum has already been considered as a limit of p_1. The same argument can be applied to dependence on p_n.

4 There are problems with these data in terms of the accuracy of respondents' self-reported income and the coverage of the lowest income groups. For a discussion, see Shalom (1981, pp. 146–50, 238–43).

5 Choosing a different significance level, of course, will lead to different conclusions, but the procedures and logic will be the same.

References

Bureau of the Census and Statistics, Republic of the Philippines (1973). *BCS Survey of Households Bulletin: Family Income and Expenditures, 1971.* Series No. 34. Manila.

Central Bank, Republic of the Philippines (1972). *Statistical Bulletin* XXIV (December 1972): 372.

Hammond, Kenneth R. and Householder, James E. (1962). *Introduction to the Statistical Method: Foundations and Use in the Behavioral Sciences.* New York: Knopf.

Hays, William L. (1973). *Statistics for the Social Sciences,* 2nd edition. New York, Holt, Rinehart and Winston.

Shalom, Stephen Rosskamm (1981). *The United States and the Philippines: A Study of Neocolonialism.* Philadelphia: Institute for the Study of Human Issues.