

Discrete, Lumpy Data; The Median, and Dislocated Computer Algorithms.

By Ted Micceri

Paper Presented at the *American Statistical Association Conference*, San Antonio, TX, January 1988

Both archaic (Pearson, 1895; Allport, 1934) and modern authors (Bradley, 1977; Walberg, Strykowski, Rovai and Hung, 1984; Micceri, 1989) record the common occurrence of extreme asymmetry among empirical data. When this occurs, parameters (q) such as the arithmetic mean and the median locate at different points and provide somewhat different information about the population. In an extremely asymmetric situation, the median may be considered a distribution's true "center". Moreover, it usually tells us most accurately where the "bulk" of a population locates and provides the least misleading point estimate of location for many purposes. Thus, a precise estimate of this parameter appears critical in application.

Prior to the onset of computers, data analysts usually estimated the median's position (Q_2) using grouped data. Today, computer-based statistical packages almost always use an algorithm assuming ungrouped data in the calculation of Q_2 . This paper points out some differences that occur between the two definitions of Q_2 for a group of large-sample real world data sets and argues that the rationale underlying the grouped, rather than the ungrouped version, is most appropriate for all situations.

In a recent article, Freund & Perles (1987) note that although considerable disagreement occurs when defining quartiles or hinges, there is no disagreement about the "...formula for the position of the median..." for ungrouped data, where $Q_2 = \frac{(n+1)}{2}$. For ordered data this allows an elegant FORTRAN algorithm based on:

$$Q_2 = \frac{(x_h + x_1)}{2} \quad (1.0)$$

where: $x_h = x_{\left(\frac{n+1}{2}\right)}$, $x_1 = x_{\left(\frac{n+2}{2}\right)}$ and n is an integer value (modulus = zero).

This definition of Q_2 derives partly from the unrealistic assumption that real world ungrouped data distributions are smooth and continuous. Numerous authors in a variety of fields note that even when scores are not grouped for purposes of analysis, discreteness frequently characterizes the data of empiricism (Micceri, 1989; Hill and Dixon, 1982; Tapia and Thompson, 1978). Oscar Kempthorne (1979) argues: "...I have often taken the view that real data are necessarily discrete and that the consequences of this are inadequately appreciated...". Unquestionably, discreteness is pervasive among measures in the health sciences, psychology, sociology, business and education, all of which are hotbeds of statistical analysis.

Difficulties arise when more than one case falls into the class interval containing Q_2 , a common occurrence where sample points mass densely among discrete score distributions. This phenomenon receives little attention despite its sometimes substantial influence on estimates. Formula 1.0 defines Q_2 as the mean of two adjacent scores when n is even and as a specific score when n is odd. For discrete data, this

always identifies Q2 as either the class interval midpoint or one of the class interval real limits. For example, using a sample of 40 IQ scores with the ordered 19th through 23rd scores being: 99, 100, 100, 100, 101, formula 1.0 defines Q2 as

$$Q2 = \frac{(x_{20} + x_{21})}{2} = \frac{(100 + 100)}{2} = 100.00.$$

Given the preceding definition, almost any time more than two observations fall in the class interval containing Q2, it becomes the midpoint of that class interval, or a specific scale point for measures such as IQ, dollars and blood pressure. Of course, one may assume that such discrete data are all concentrated at the midpoint of their respective intervals, as we do with \bar{x} , and continue to apply formula 1.0. However, such a premise conflicts with the common measurement assumption that scale points categorize an underlying continuum. Additionally, this definition can cause such absurd results as Q2 locating above \bar{x} for positively skewed distributions. Figure 1 provides an empirical example of such an occurrence.

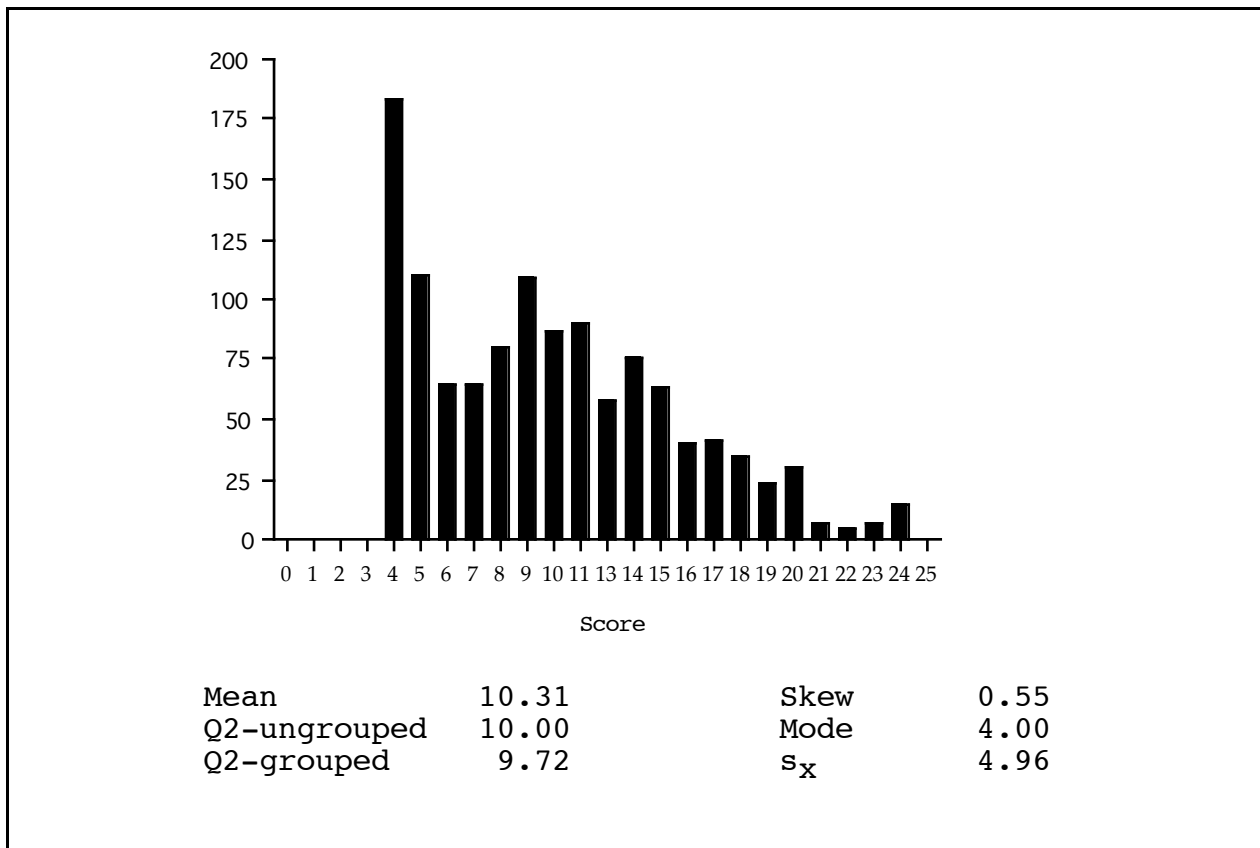


Figure 1: Psychometric Job Satisfaction Scale (N=1258)

An historically accepted alternative is to presume that observations distribute uniformly within each class interval. That is, given five observations within a class, one assumes that each occupies one-fifth of the class interval's width. The appropriate definition of Q2 for ordered data then becomes:

$$Q2 = x_1 + \frac{i(f_1 - f_2)}{(f_3 - f_2)} \quad (2.0)$$

where:

x_1 = lower real limit of interval containing Q2

i = width of the interval

f_1 = cumulative frequency at the sample point defining Q2

f_2 = cumulative frequency at the lower real limit of the interval containing Q2

f_3 = cumulative frequency of the lower real limit at the class interval immediately following Q2

Applying formula 2.0 to the prior example produces: $x_1 = 99.50$, $i = 1.0$ (from 99.50 to 100.49), $f_1 = 20.5$, $f_2 = 20$, $f_3 = 23$, and $Q2 = 99.67$, a value 33 percent of a scale point different from the integer 100 generated by formula 1.0. Formula 2.0 reduces to: $Q2 = x_1 + \frac{i}{2}$ for class intervals containing a single value and becomes equivalent to 1.0.

Thus it only changes the estimate when the discrete or grouped nature of data influences the estimate.

But what makes this estimate a “better” estimate of the median? Aside from avoiding results such as the absurd but true example shown in Figure 1, it equates the underlying logic of the arithmetic mean and Q2 for all forms of data, not merely for continuous data. In the computation of \bar{x} , the assumption is made that all the observations in a group are concentrated at the midpoint of that group. This assumption also underlies formula 1.0. For \bar{x} this is arithmetically the same as assuming a uniform distribution of observations within the class of interest because the mean equally weights each observation and its location thereby providing proportional or uniform representation. Although appropriate for continuous data when computing Q2, the midpoint assumption is untenable when data are discrete or grouped, because the median’s value depends upon its position within the class. Therefore, to equally weight each observation within that class, as when estimating \bar{x} , the assumption of a uniform distribution of observations is required. This definition has prevailed in the statistical literature for many years, see for example: Mills (1924), Richmond (1957), Hays (1963)

An empirical investigation of differences between Q2 defined as grouped and ungrouped among 440 discrete, large-sample, real world ability and psychometric distributions (Micceri, 1989) found the mean absolute difference for an arbitrary sample of size 100 to be $0.637s_{\bar{x}}$. The maximum absolute difference between estimates using formulae 1.0 and 2.0 was $5.13s_{\bar{x}}$. Actual sample sizes for these data ranged from 190 to 10,893, with almost 70% of the distributions including 1,000 or more cases. Eighty three percent of these distributions had sample spaces of between 10 and 99 scale points, with 12.5% having fewer than ten and 4.3% greater than 99 scale points. Table 1 shows the proportion of these distributions exhibiting differences between the two algorithms greater than $1.00s_{\bar{x}}$ and $2.00s_{\bar{x}}$ respectively for samples of size 100, 25 and 10. Note that even for sample size 10, three percent of the estimates differ by at least one full

standard error.

Table 1
Examples of Possible Differences Between Formulae 1.0 and 2.0

Score	Distribution 1		Distribution 2		Distribution 3	
	Cumulative		Cumulative		Cumulative	
	Frequency	Frequency ^e	Frequency	Frequency ^e	Frequency	Frequency ^e
1	22	22	20	20	10	10
2	10	32	20	40	40	50
3	20	52	22	62	20	70
4	50	102	20	82	10	80
5	0	102	20	102	22	102
Mean		2.96		2.96		2.94
Median 1.0		3.00		3.00		3.00
Median 2.0		3.45		2.95		2.55

The problem becomes particularly acute for measures having extremely small sample spaces, as for example the commonly used psychometric attitude scales. Figure 2

shows a situation in which the distance between \bar{x} and Q2 is substantially different using the two formulae.

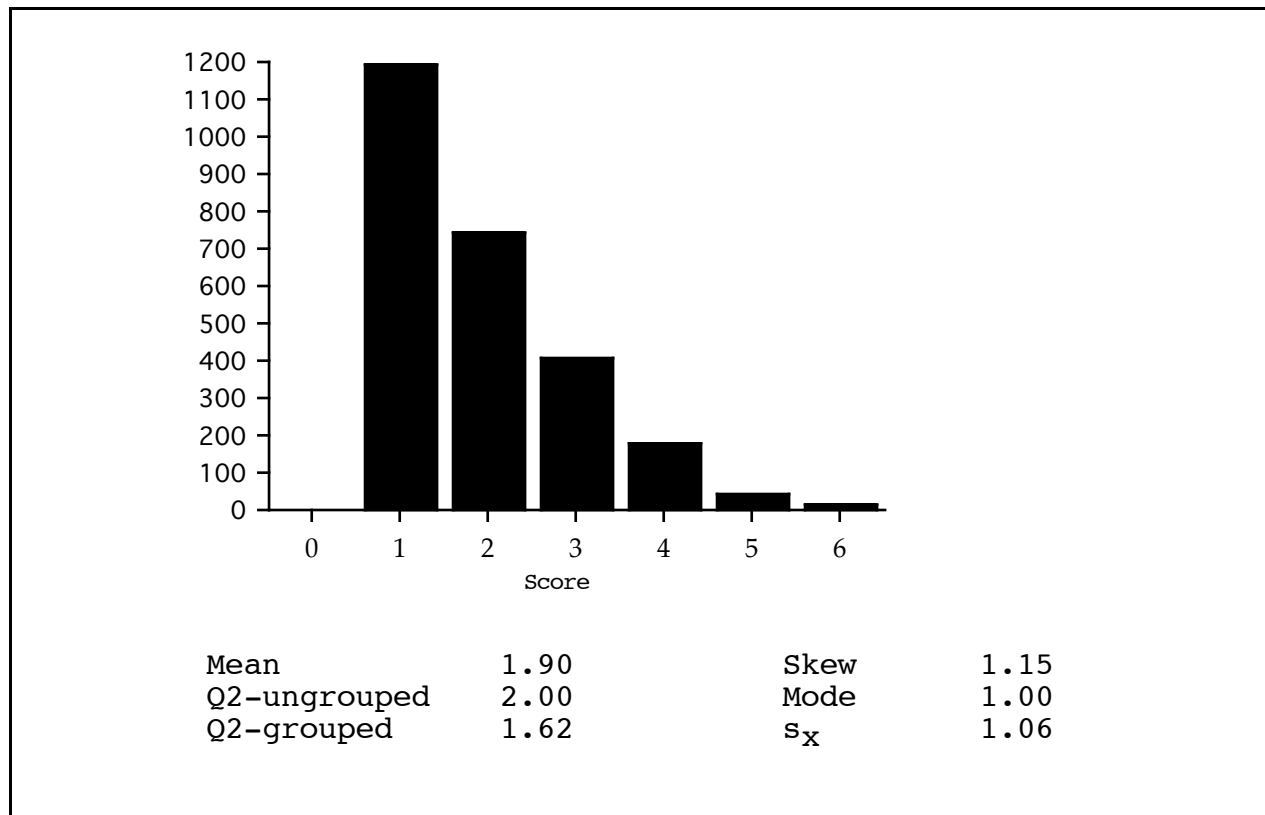


Figure 2: Third Grade Mathematics Test (N=2576)

Of course, it is tempting to assume ungrouped data and program a simple six line algorithm in place of the far longer one required by formula 2.0. Unfortunately, many major statistical software vendors (e.g. SAS, SPSSX, BMDP, IMSL, SYSTAT, MINITAB) fall prey to this over-simplification and compute the median for ungrouped data. Old SPSS subroutine FREQUENCIES (version 9.0) did compute Q2 for grouped data; however, for purposes of algorithmic simplicity, the class interval was assumed to be 1.00. This produced absurd results when either (1) a gap occurred between the median and the scale point immediately below, or (2) the class interval differed from 1.0, as for the distribution in Figure 3.

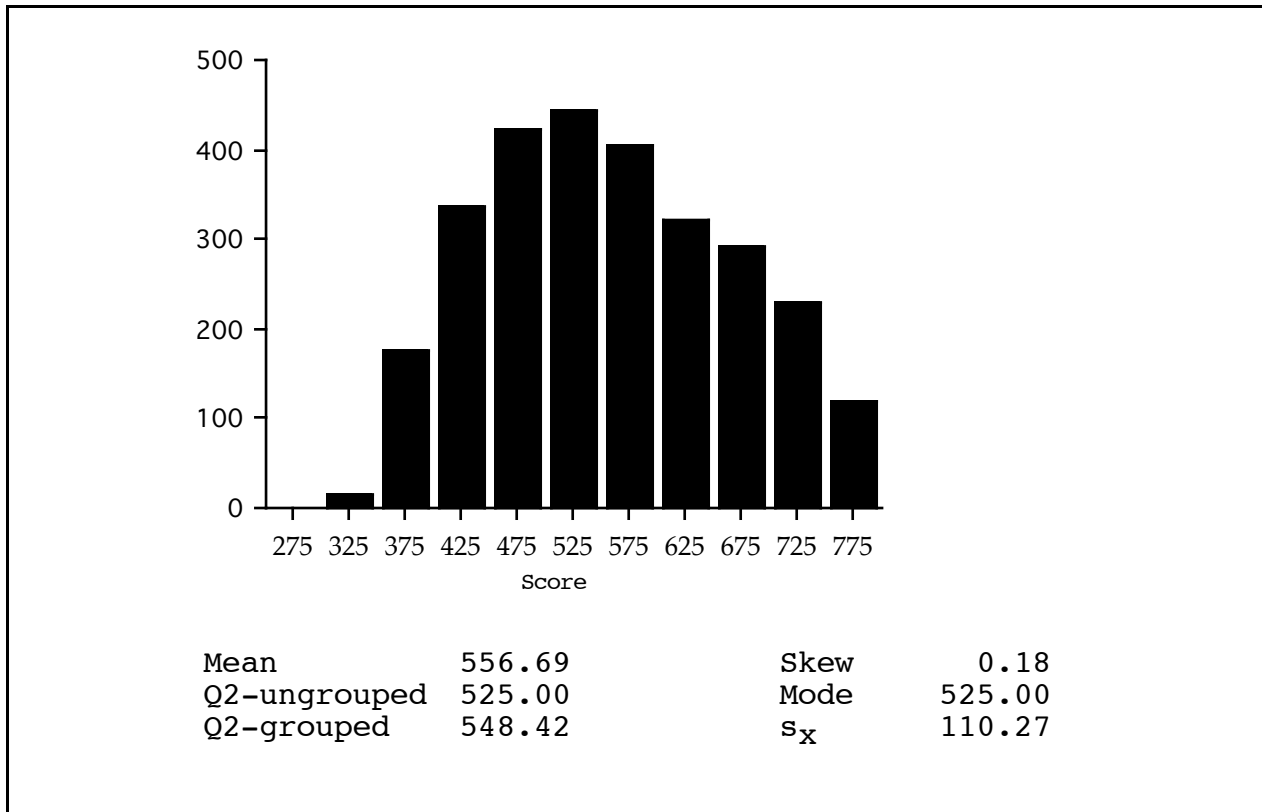


Figure 3: College Board Latin Exam (N=2755)

In conclusion, the particulars presented here may be condensed into five reasons for the universal application of computer algorithms for grouped rather than ungrouped data:

1. Discreteness is common among measures of interest to those conducting statistical analyses.
2. Arithmetically, assuming a uniform distribution of observations across a class interval for Q2 is the same as the assumption for \bar{x} of observations concentrating at the class mid-point. Additionally, from a measurement perspective, it appears reasonable to adopt the uniform distribution assumption.
3. Empirical findings show that different estimates frequently result from the two formulae and that sometimes disparities between the two definitions are substantial.

4. Q2 for ungrouped data sometimes produces absurd results such as that shown in Figure 1.
5. It is simple to round or truncate an estimate for grouped data to produce the ungrouped estimate; however, the reverse is far from simple.

Thus, the use of formula 1.0 in the name of either elegance or cost containment appears incomplete. Although appropriate for ungrouped or continuous data, it can produce results that could embarrass a statistician or researcher if reported without careful scrutiny. Since most of us concentrate on the mean in our presentations, this is not unforeseeable. At the least, a grouped version of Q2 should be available as an option in any descriptive subroutine. Similar algorithms are applied in the computation of all percentiles, however, as one goes farther toward a distribution's tail (e.g. 75th, 95th, 97.5th percentiles) the effects lessen with density. Hopefully, this discourse will stimulate awareness of some questions one might wish to address to the output of canned statistical packages in the computation of Q2 or any other percentile. For those inclined, a stamped, self enclosed envelope to the author will bring forth an inelegant FORTRAN 77 algorithm capable of computing Q2 and its surrounding class interval for grouped data.

REFERENCES

- Allport, F. M. (1934). The J-curve hypothesis of conforming behavior. Journal of Social Psychology, 5, 141-183.
- Bradley, J. W. (1977). A common situation conducive to bizarre distribution shapes. The American Statistician, 31, 147-150.
- Freund, J. E. & Perles, B. M. (1987). A new look at quartiles of ungrouped data. The American Statistician, 41:3, 200-203.
- Hays, W. L. (1963). Statistics for Psychologists. New York: Holt, Rinehart and Winston, Inc..
- Hill, M. & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. Biometrics, 38,377-396.
- Kempthorne, O. (1978). Some aspects of statistics, Sampling and randomization (pp. 11-28). In H. A. David (Ed.) Contributions to Survey Sampling and Applied Statistics. New York: Academic Press.
- Mills, F. C. (1924). Statistical Methods. New York: Henry Holt and Company.
- Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. Psychological Bulletin, 105:1, 156-166.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution - II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society, A186, 343-414.
- Richmond, S. B. (1957). Statistical Analysis. New York: The Ronald Press Company.
- Tapia, R. A. & Thompson, J. R. (1978). Nonparametric Probability Density Estimation. Baltimore: Johns Hopkins University Press.
- Walberg, Herbert J., Strykowski, B.F., Rovai E., & Hung S.S. (1984). Exceptional performance. Review of Educational Research, 54, 87-112.