Journal of Statistics Education ›

Volume 14, 2006 - Issue 3

Original Articles

Quartiles in Elementary Statistics

Eric Langford

| Published online: 01 Dec 2017

Cite this article https://doi.org/10.1080/10691898.2006.11910589

Abstract

The calculation of the upper and lower quartile values of a data set in an elementary statistics course is done in at least a dozen different ways, depending on the text or computer/calculator package being used (such as SAS, JMP, MINITAB, Excel, and the TI-83 Plus). In this paper, we examine the various methods and offer a suggestion for a new method which is both statistically sound and easy to apply.

Percentiles | Quantiles

1. Introduction

It is well-known among statisticians that there are a number of different definitions in the literature of the upper and lower quartile values of a data set. It should be noted

many repetitions, but to the elementary case where there are a number of different data values with few, if any, repetitions. (I refer to this type of data as “exam data”; I shall discuss the case of data sets with few distinct values but many repetitions in an appendix.) The situation is, I believe, far worse than most realize: In looking at methods which are actually used in elementary statistics textbooks, I have discovered seven distinct methods, together with several others which are apparently different but are actually equivalent to one of the seven. Looking at methods employed by various commonly-used calculator and computer packages yields another five distinct methods, and the literature contains at least six more specific methods. In this paper, I will discuss the various methods, and using a precise definition (Definition 2) of percentile, identify that method which satisfies this definition. Unfortunately this method (the “CDF Method”) is not, in its usual form, the easiest for a student to apply. I will then show how this method can be effected in the spirit of more easily-applied methods, thus providing a new method of calculating quartiles which is both statistically sound and easy to apply. I hope that this new method will enjoy wider use. I also hope that the discussions in this paper will be of interest in the classroom and might provide a basis for a classroom project for ambitious and talented students in an introductory statistics class.

I want to point out that the emphasis in this paper will be on the computation of quartiles and will be at a level suitable for classroom discussion at the level of a first course in statistics. More general definitions of quantiles (percentiles) are given here, but not stressed. With the increasing emphasis on exploratory data analysis (EDA) in the elementary classroom, in particular the ideas of the five-number summary and box-and-whisker plots (boxplots), a thorough understanding of quartiles is mandatory, but a detailed discussion of quantiles may not be necessary for the beginning student.

Elementary discussion of quartiles can be found in Dr. Twe (2002), Freund and Perles (1987), Hayden (1997), Joarder and Firozzaman (2001), John (2000), Journet (1999), and Wessa (2006). Wessa's website also contains a link to an online calculator which will calculate quartiles using eight different methods. For a more complete discussion of quantiles, together with a number of references, see Hyndman and Fan (1996).

It might be thought that with the increasing use of graphing calculators (for example, the TI-83 Plus) and computer packages (MINITAB, SAS, Mathematica, JMP, Microsoft Excel) in the classroom, the need for consistency in the textbook definition of quartiles

consistent definition of the quartiles more necessary, rather than less, since as we shall see later, the TI-83 Plus, MINITAB, SAS, Mathematica, and Microsoft Excel use five different definitions of the quartiles! (The methods used by each of these packages are summarized later in Table 1.) In fact, one recent text (McClave and Sincich (2003)) reproduces results from a TI-83 (p. 46), MINITAB (p. 48), and SAS (pp. 50 and 65), all of which use different methods. What are students to do when they check a MINITAB or SAS or Microsoft Excel calculation on their TI-83 Plus calculator and get a different answer, all of which differ from the answer in the back of the book? This is not an idle concern; a very confused student wrote to the "Ask Dr. Math" section of The Math Forum@Drexel inquiring why his TI-83, Excel, MINITAB, and his own paper-and-pencil calculations all gave different answers for the quartiles of his data set. (See Dr. Twe (2002).)

There is a tendency for statisticians to say, "Why worry? The differences are small so who cares?" Freund and Perles (1987) answer this well:

> "Before we go into any details, let us point out that the numerical differences between answers produced by the different methods are not necessarily large; indeed, they may be very small. Yet if quartiles are used, say to establish criteria for making decisions, the method of their calculation becomes of critical concern. For instance, if sales quotas are established from historical data, and salespersons in the highest quarter of the quota are to receive bonuses, while those in the lowest quarter are to be fired, establishing these boundaries is of interest to both employer and employee. In addition, computer-software users are sometimes unaware of the fact that different methods can provide different answers to their problems, and they may not know which method of calculating quartiles is actually provided by their software."

2. The Methods

Fortunately, all of the books and computer/calculator packages at which I looked were consistent with their definitions of the median: if there are an odd number n of data values, the median value is the middle one when the data are ordered, and if n is even, the median is the average of the two middle values. (Actually, this is not true for some

the paper.) Some imprecision is sometimes found as texts attempt to define the median value as “putting half of the data set above and half below,” trying to emulate the definition of median of a continuous distribution. Suppose for simplicity that the data values are all distinct. If n is even, say n = 2k, then certainly the median does do this. (In fact any number x which satisfies the inequality $x_k < x < x_{k+1}$ will have this property.) If n is odd, say n = 2k + 1, then we cannot have half of the data values greater than (i. e. above) the median and half less than (i. e. below) the median since it is difficult to divide an odd number of data values into equal halves. What can be said is that in this case, there are an equal number (namely k) of the data values greater than the median and less than the median, or, alternatively, an equal number (namely k + 1) of data values greater than or equal to the median and less than or equal to the median.

If there are repeated data values, we must replace “greater than” by “to the right of” and similarly for “less than,” “greater than or equal to,” and “less than or equal to.” But consider the data set (1, 2, 2, 3, 4). No one would disagree that the median is 2. But it is the second “2” in the set and not the first “2” which has the above properties. (All twos are equal, but some are more equal than others!) What we would like to have is a definition of the median (in this case 2) that depends only on its numerical value and not on the particular occurrence of that value. Thus we take the following definition (which is the key to defining percentiles in a precise fashion):

> DEFINITION 1: The median is that number which puts at least half of the data values at that number or below and at least half of the data values at that number or above; if more than one such number exists, there will be an entire interval of such and the median is the midpoint of that interval.

The most naive approach in defining quartiles is to think of the median as dividing the data set into halves (“bottom half” and “top half”) and then defining the lower (first) quartile Q_1 to be the median of the bottom half, and the upper (third) quartile Q_3 to be the median of the top half. This makes good sense and is an easy “sell” to students. It works well if n is even, but if it is odd, the question remains: “What do we do with the median value itself?” As you might expect, different authors give different answers. For the remainder of this paper, n will denote the number of data values in the data set.

METHOD 1 (“Inclusive”): Divide the data set into two halves, a bottom half and a top

median of the bottom half and the upper quartile is the median of the top half. As an example, if $S_5 = (1, 2, 3, 4, 5)$, then the inclusive lower half is (1, 2, 3) and hence $Q_1 = 2$. (A summary of all of the methods considered will be given later in Table 2.)

This method is used by Siegel and Morgan (1996) and is equivalent to Method 3 below.

METHOD 2 ("Exclusive"): As above except that in the case of n odd, the median value is excluded from both halves. As an example, if $S_5 = (1, 2, 3, 4, 5)$, then the exclusive lower half is (1, 2) and hence $Q_1 = 1.5$.

This method is used by Moore (2003), Peck, Olsen, and Devore (2001)(p. 117), Brase and Brase (2003), and Moore and McCabe (2003). Because of this last reference, I have seen this method referred to as the "M&M Method." Method 1 of Joarder and Firozzaman (2001) covers both of our Methods 1 and 2.

According to its instruction book (p. 12 – 29) the TI-83 Plus defines the lower quartile as being the "median of the points between the minimum and the median" and the upper quartile similarly. This would lead one to believe that Method 1 is being used. However, in using the TI-83 Plus on the test data sets defined later in this paper, it appears that Method 2 is actually being used. (The TI-84 Plus and TI-89 seem to use the same method.)

Before proceeding further, we will need some notation. To simplify matters, we always assume that the data values are ordered in nondecreasing order: . To say that we take value #(k) where k is an integer is to say we take x_k. If k is not an integer, then x_k will denote the interpolated value between x_j and x_{j+1} where and denotes the "floor function" or "greatest integer function." For example, $x_{1.5}$ is half-way between x_1 and x_2 and $x_{2.25}$ is one-fourth of the way from x_2 to x_3. Note that with this notation, j + 1 can be written as where denotes the "ceiling function," assuming again that k is not an integer. (Here $x_0 = x_1$ and $x_{n+1} = x_n$.)

In his classic book on EDA, Tukey (1977) introduced the concepts of box-and-whisker plot and five-number summary in terms of what he calls the upper and lower hinges (see p. 33). The two hinges form the ends of the box in the box-and-whisker plot and, together with the maximum, minimum, and median values, form the five-number summary. The upward rank of a data value x_k is simply k — the distance one counts upwards to get to that value. (If k is not a whole number, interpolate.) The downward

case is simply n + 1 - k. The depth of a data value is the minimum of its upward and downward ranks. Tukey first defines the median as having depth M where , so that it has equal upward rank and downward rank. It is easy to see that this is equivalent to the usual definition when we interpolate as above when n is even. The depth H of the hinges is then given by , where the lower hinge has upward rank H and the upper hinge has downward rank H. (Sometimes this is called F for “fourths.”) These are often called letter values. One can continue to define (“eighths”) which can be used to form a seven-number summary and so on. (See Hoaglin (1983).)

Tukey is careful to define his box-and-whisker plots and five-number summaries entirely in terms of the hinges, and does not involve quartiles. However, many authors use the quartiles rather than the hinges in their definitions, which is where the confusion arises, because of the many different definitions of the quartiles. We shall formalize the Tukey hinges as Method 3, even though, strictly speaking, Method 3 is used to find hinges not quartiles. In Table 2 later on, we shall see that Tukey hinges are numerically equal to Method 1 quartiles, so we need not worry about what “Tukey quartiles” are.

METHOD 3 (“Tukey”): Let the median be #(M) = #((n + 1)/2) and define . Count H measurements from the bottom and H measurements from the top to get the lower and upper hinges; if H is not an integer, then interpolate; i. e., the lower hinge is #(H) and the upper hinge is #(n + 1 — H). As an example, if S_5 = (1, 2, 3, 4, 5), then the median is #(M) = #(3) = 3 and so H = 2 making the lower hinge also 2.

In addition to Tukey (1977), this approach is used by Milton, McTeer, and Corbet (1997). Also, MINITAB can be used to calculate the hinges by using the EDA option and asking for “letter values.” Curiously enough, MINITAB when asked to draw a box-and-whisker plot will use its own calculation (Method 11) of the quartiles, rather than the Tukey letter values.

In general, unless authors define quartiles using one of the three methods above, they define percentile values and let the lower quartile (25^{th} percentile) and upper quartile (75^{th} percentile) be special cases of that definition. These definitions are usually based on the generalization of the “definition” of the median as being that value which puts “half of the data set above and half of the data set below.” (Recall our previous discussion, which yielded Definition 1.) This generalized “definition” is: “The P^{th} percentile value puts P percent of the data set below and (100 - P) percent of the data

we have already done for the median. (For simplicity of notation, we let p = P/100, so that, for example, the 50^{th} percentile corresponds to p = 0.5.)

One method used is the following. We shall see in the next section that this method, although unwieldy to apply, is the only method that satisfies our precise definition of percentile. We call it the "CDF Method" since it is based on the CDF (cumulative distribution function) of the empirical distribution given by the data set. SAS refers to it as "empirical distribution function with averaging."

METHOD 4 ("CDF"): The P^{th} percentile value is found as follows. Calculate np. If np is an integer, then the P^{th} percentile value is the average of #(np) and #(np + 1). If np is not an integer, the P^{th} percentile value is ; that is, we round up. Alternatively, one can look at #(np + 0.5) and round off unless it is half an odd integer, in which case it is left unrounded. As an example, if S_5 = (1, 2, 3, 4, 5) and p = 0.25, then #(np) = 1.25, which is not an integer so that we take the next largest integer and hence Q_1 = 2. Using the alternative calculation, we would look at #(np + 0.5) = #(1.75) which would again round off to 2. Note that this method can be considered as "Method 10 with rounding."

This method is used by Johnson and Bhattacharyya (1996), Johnson (2000), and Ross (1996). It is Definition 2 of Hyndman and Fan (1996) and Definition 4 of Joarder and Firozzaman (2001), who refer to Smith (1997), p. 36, who uses the alternative calculation. It is the default option PCTLDEF = 5 of the SAS System computer package and is also Method 4 of Wessa (2006).

Yet another method is found in Mendenhall and Sincich (1995).

METHOD 5 ("M&S"): For the lower and upper quartile values take #((n + 1)p) with p = 0.25 for the lower quartile and p = 0.75 for the upper quartile. Then round to the nearest integer. If (n + 1)p is half an odd integer, round up for the lower quartile and down for the upper quartile. For example, if S_5 = (1, 2, 3, 4, 5) and p = 0.75, then #((n + 1)p) = #(4.5) and hence Q_3 = 4. Note that this can be considered as "Method 11 with complete rounding," in the same way that Method 4 can be considered as "Method 10 with rounding." For general percentiles, the authors say to "take #(n + 1)p and "round to the nearest integer," perhaps implying the same kind of rounding as for the quartiles when (n + 1)p is half an odd integer.

METHOD 6 (“Lohninger”): This method is the same as the previous method except in the case of (n + 1)p equal to half an odd integer we always round up. Using the same example as above, we would round up rather than down and obtain $Q_3 = 5$.

Joarder and Firozzaman (2001) refer to a method of Vining (1998), p. 44:

METHOD 7 (“Vining”): Define Q_1 to be #((n + 3)/4) if n is odd and #((n + 2)/4) if n is even and define Q_3 to be #((3n + 1)/4) if n is odd and #((3n + 2)/4) if n is even. For example, if $S_5 = (1, 2, 3, 4, 5)$, then we take $Q_1 = \#(8/4) = 2$. (We shall see from Table 2 that this is equivalent to Method 1.)

Joarder and Firozzaman (2001) also propose formulas which they call the “Remainder Rule.” In terms of our notation, it looks like the following: First write n = 4m + k, where k = 0, 1, 2, or 3. If k = 0 or 1, let Q_1 be #(m + 0.5) and Q_3 be #(n — m + 0.5). If k = 2 or 3, let Q_1 be #(m + 1) and Q_3 be #(n — m). After a little algebra, this rule can be seen to be equivalent to the following:

METHOD 8 (“J&F”): Define Q_1 to be #((n + 1)/4) if n is odd and #((n + 2)/4) if n is even and define Q_3 to be #((3n + 3)/4) if n is odd and #((3n + 2)/4) if n is even. For example, if $S_5 = (1, 2, 3, 4, 5)$, then we take $Q_1 = \#(6/4) = 1.5$. (We shall see from Table 2 that this is equivalent to Method 2.)

Still another method is used by Hogg and Ledolter (1992).

METHOD 9 (“H&L”): The P^{th} percentile value is found by taking that value with #(np + 0.5). If this is not an integer, take the average (not the weighted average) of and . As an example, if $S_5 = (1, 2, 3, 4, 5)$ and p = 0.25, then #(np + 0.5) = #(1.75) and so we average #(1) and #(2) implying that $Q_1 = 1.5$.

These authors observe (p. 21, bottom) “alternatively, one could interpolate using the weighted averages … [but that the] differences, however, will usually be quite small.” This provides still another method, distinct from all of the others, since it gives a value of 1.75 for Q_1 when applied to S_5. Even though this method was not actually used by any of the texts that I have examined, it is referred to in the literature and is used by Mathematica. Note that it makes a nice complement to Methods 11 and 12.

METHOD 10 (“H&L-2”): The P^{th} percentile value is found by taking that value with #(np + 0.5). If this is not an integer, take the interpolated value between and . As an

example, if $S_5 = (1, 2, 3, 4, 5)$ and $p = 0.25$, then $\#(np + 0.5) = \#(1.75)$ and so $Q_1 = 1.75$.

This method is Method 5 of Hyndman and Fan (1996) who refer to it as "a very old definition, proposed by Hazen (1914) and popular among hydrologists … ." It is used by Mathematica in calculating "Quartiles" or "InterpolatedQuantiles."

Other texts use a method which is used by MINITAB.

METHOD 11 ("MINITAB"): The P^{th} percentile value is found by taking that value with # ((n + 1)p). If (n + 1)p is not an integer, then interpolate between and as explained previously. For example, if $S_5 = (1, 2, 3, 4, 5)$ and $p = 0.25$, then $\#((n + 1)p) = \#(1.5)$ and hence $Q_1 = 1.5$.

This method is used by Mendenhall, Beaver and Beaver (2003), Hogg and Tanis (1997), and by Khazanie (1996), as well as by MINITAB and JMP (See JMP® User's Guide (1994), p. 159). It is also Definition 6 of Hyndman and Fan (1996) who refer to Weibull (1939) and Gumbel (1939). It is Method 5 of Joarder and Firozzaman (2001), Method 2 of Wessa (2006), and it can also be found in Snedecor (1946), p. 51. It is also the PCTLDEF = 4 option of the SAS System computer package. Method 7 of Wessa, which he calls the "TrueBasic" method is similar to this except it uses a "backwards interpolation"; for example, $x_{2.25}$ is calculated as one quarter of the way from x_3 back to x_2.

Microsoft Excel has a built-in quartile and percentile routine. Under its "Help Topics," Excel states that "If k is not a multiple of PERCENTILE interpolates to determine the value at the k"th percentile." This implies that the method is given by the following:

METHOD 12 ("Excel"): To calculate the P^{th} percentile take $\#((n — 1)p + 1)$, with interpolation. As an example, if $S_5 = (1, 2, 3, 4, 5)$ and $p = 0.25$, then $\#((n — 1)p + 1) = \#(2)$ and hence $Q_1 = 2$.

I have not seen this method used by any textbook, but it is Method 7 of Hyndman and Fan (1996) who refer to Gumbel (1939). It can also be found in Freund and Perles (1987) and is Method 5 of Wessa (2006).

Note that all of the first twelve methods with the exception of the Lohninger Method 6 are what I call symmetric. That is, the two quartiles Q_1 and Q_3 have equal depth in the sense of Tukey. Symbolically, if $Q_1 = \#(q_1)$ and $Q_3 = \#(q_3)$ then $q_1 + q_3 = n + 1$. You

The SAS System, in its univariate procedures, offers the user five different options for computing percentiles, using its “PCTLDEF =” option. (See SAS® Procedures Guide (1990), p. 625.) As noted before, the default option, PCTLDEF = 5 (“empirical distribution function with averaging”), is the same as our Method 4 (“CDF”) and the PCTLDEF = 4 option is the same as our Method 11 (“MINITAB”). The first three options, PCTLDEF = 1, 2, and 3, in certain circumstances give values for the median that are not consistent with the usual definition. We present them here for completeness, but we shall not consider them further.

METHOD 13 (“SAS-1”): To calculate the P^{th} percentile take #(np) with interpolation. SAS refers to this as “PCTLDEF = 1.” This method gives in every case values for the median which are not the same as the usual values. For example, if $S_3 = (1, 2, 3)$, this method would give the median as 1.5 rather than 2.

This method is Definition 4 of Hyndman and Fan (1996) who refer to Parzen (1979) and is Method 1 of Wessa (2006). It is also used by Mathematica in calculating “AsymmetricQuartiles.”

METHOD 14 (“SAS-2”): To calculate the P^{th} percentile take x_k where k is the closest integer to np, rounding to the even value if np is half an odd integer. SAS refers to this as “PCTLDEF = 2.” This method gives values for the median which are not the same as the usual values unless n is of the form 4k + 3. For example, if $S_5 = (1, 2, 3, 4, 5)$ and p = 0.5, then np = 2.5, so rounding to the even, 2, would give the median as 2 rather than 3.

This method is Definition 3 of Hyndman and Fan (1996). A similar method is Method 6 of Wessa (2006), which he refers to as the “closest observation” method. Wessa's method is: To calculate the P^{th} percentile take x_k where k is the closest integer to np, rounding up if np is half an odd integer. It can be seen that this is equivalent to taking and that this gives the usual value for the median if n is odd, but not if n is even.

METHOD 15 (“SAS-3”): To calculate the P^{th} percentile take . SAS refers to this as “PCTLDEF = 3,” the “empirical distribution function” method. It is not hard to see that this gives the usual value for the median if n is odd, but not if n is even.

This method is Definition 1 of Hyndman and Fan (1996) and Method 3 of Wessa (2006). It is also used by Mathematica in calculating “Quantiles.”

For the convenience of the user of calculator/computer statistical packages, we now give a table which gives the method each such package uses.

Table 1. Methods Used in Statistical Packages

Download CSV Display Table

A little thought will show that if we are considering just quartiles, then the results that the various methods give depend only on the congruence class (mod 4) in which n falls, that is, on the remainder that occurs when n is divided by 4. It is also possible to show by taking the four cases of n= 4k, n= 4k + 1, n= 4k + 2, n= 4k + 3 that we need look at only four "canonical" data sets: S_4, S_5, S_6, S_7, consisting of (1, 2, 3, 4), (1, 2, 3, 4, 5), (1, 2, 3, 4, 5, 6), and (1, 2, 3, 4, 5, 6, 7) respectively. (In a sense we are simply looking at the position of the data value in the data set, rather than its actual numerical value.) As was observed by Peck, Olsen, and Devore (2001), two methods are the same if and only if they agree on these four data sets. (With one exception: Method 14. However we are not considering this method.) Here is a table (Table 2) comparing the lower and upper quartile values (Q_1, Q_3) given by each of the methods for each of the four canonical data sets, together with the interquartile range (IQR).

Table 2. Comparing the Various Methods on the Canonical Data Sets

Download CSV Display Table

We can make several observations from the table. The Tukey Method 3 and the Vining Method 7 are seen to be the same as the Inclusive Method 1, whereas the J&F Method 8 is seen to be the same as the Exclusive Method 2. Henceforth, we shall not consider these to be separate methods. The first nine methods can be thought of as "averaging" methods, since their quartile (indeed, percentile values in the cases of the CDF Method 4 and the H&L Method 9) are always individual data values or halfway between two successive data values. The last three methods can be thought of as "interpolation" methods, since their quartile (and percentile) values may lie elsewhere between successive data values.

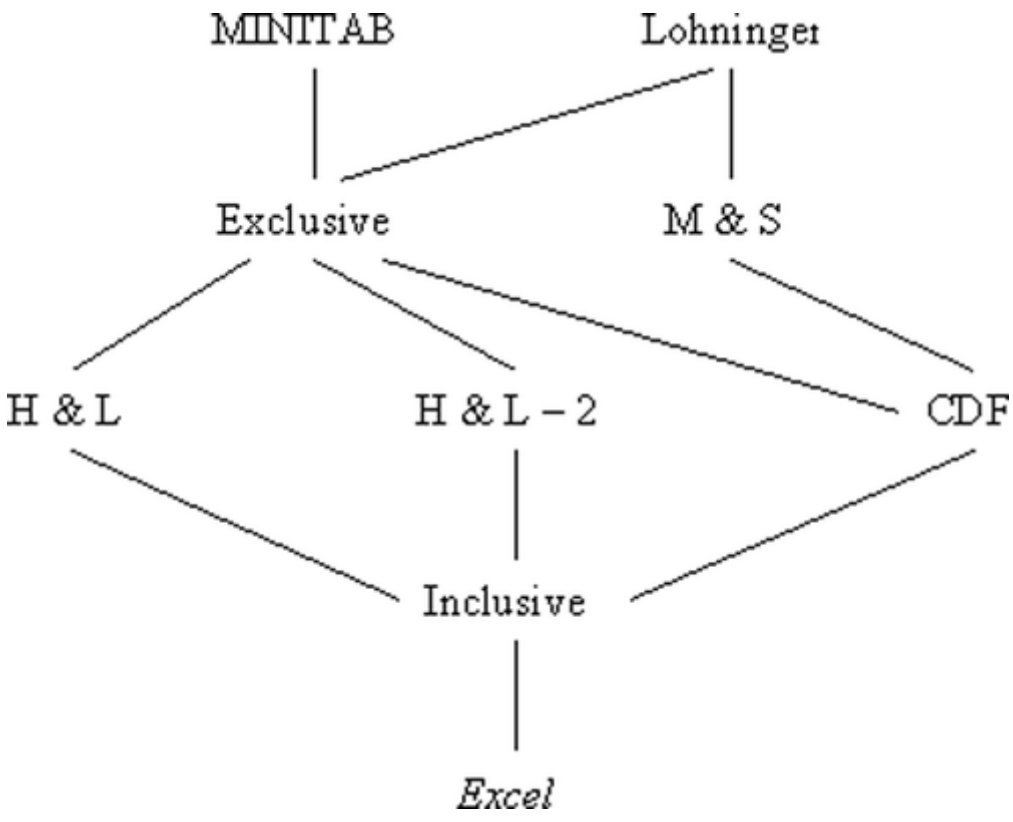
The M&S Method 5 and the Lohninger Method 6 are unique in the sense that they give only values which are data values themselves. The other averaging methods all agree if n is even, whereas if n is odd, then the CDF Method 4 agrees with the Inclusive Method 1 if n is of the form 4k + 1 and with the Exclusive Method 2 if n is of the form 4k + 3, whereas exactly the opposite is true for the H&L Method 9. Therefore these four methods (remember that Methods 3, 7, and 8 are redundant) exhaust all possibilities for the inclusion and exclusion of the median value in the “top-half, bottom-half” idea. More precisely, the Inclusive Method 1 includes the median (in both halves) in both of the cases 4k + 1 and 4k + 3; the Exclusive Method 2 excludes it in both of the cases; the CDF Method 4 includes it in the case 4k + 1 and excludes it in the case 4k + 3; and the H&L Method 9 excludes it in the case 4k + 1 and includes it in the case 4k + 3.

The three interpolation methods can be thought of as different generalizations of the median value as . The Excel Method 12 looks at the first form, the H&L-2 Method 10 looks at the second, and the MINITAB Method 11 looks at the third. As was noted by Freund and Perles (1987), these three methods when applied to the quartiles Q_i (i = 1, 2, 3) yield, respectively, , , and , and that these can be viewed as the special cases = 0, 0.5, 1 of the general formula . The generalizations of these to arbitrary quantiles are #((n — 1)p + 1), #(np + 0.5), #((n + 1)p), and . Other values of are used in the literature and provide still more methods. Method 8 of Hyndman and Fan (1996) uses = 2/3, Benard and Bos-Levenbach (1953) use = 7/10, and Method 9 of Hyndman and Fan (1996) uses = 5/8, referring to Blom (1958). Blom considers essentially the formula which if = reduces to when we let = = 1 — . See Hyndman and Fan (1996) for a more complete discussion.

The interpolation methods can be viewed as various methods of “smearing” the data values so that the “stair-step” CDF is replaced by a piecewise linear function from which the percentiles are calculated as they would be for a continuous distribution. (C. f. the method discussed in the appendix.) See Journet (1999) and John (2000) for graphs of some of these functions. If the data values are distinct, this is fairly straightforward, but if there are repeated values, difficulties arise. For example, one would expect that the quartile values for the data set S_5 = (1, 2, 3, 4, 5) would be the same as for the data set $2S_5$ = (1, 1, 2, 2, 3, 3, 4, 4, 5, 5), but they are not for Methods 10 and 11 as can be seen by comparing Table 2 and Table 3. Similarly Method 12 gives different results on the data sets S_7 = (1, 2, 3, 4, 5, 6, 7) and $2S_7$ = (1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7).

We see that we now have an entire infinite family of possible interpolation methods! For each of these, we can obtain other possible methods by “rounding” (i. e., by rounding to the nearest integer except when we get a value which is half an odd integer as in the CDF Method 4) and by “complete rounding” (i. e., by rounding to the nearest integer, with some rule as to what to do when we get a value which is half an odd integer as in Methods 5 and 6). For example, the CDF Method 4 is the case of = 1/2 with rounding, and Method 6 of Wessa (2006) is the same case with complete rounding. Method 8 of Wessa (2006) is the case of = 1 with rounding, whereas the M&S Method 5 and the Lohninger Method 6 are the same case with two different kinds of complete rounding.

Finally, looking at the IQRs, we can see, for example, that in every case, the Excel Method gives IQR values which are no larger than those given by any other method. We can summarize all such relationships in the following diagram (Figure 1) where if Method A lies above Method B in the figure, then the IQR values of Method A are at least as large as those of Method B in every case.



short-legendFigure 1

Display full size

3. Evaluation of the Methods

What makes for a “good” quartile method? A lot depends on the purpose of calculating the quartiles. Are we dealing with the data set in and of itself, or are we thinking of the data set as a sample from some population and trying to use it to estimate parameters of the underlying population? We shall take the first approach. One criterion is that the first quartile should divide the data set so that “approximately” 25% of the data values are to the left and “approximately” 75% are to the right, and vice versa for the third quartile. Another criterion is that the two quartiles and the median should divide the data set into four “approximately” equal pieces. As we saw with the median, these ideas can slippery, especially when the data set may contain repeated values. These criteria have been investigated for various methods by Freund and Perles (1987), Hyndman and Fan (1996), and Joarder and Firozzaman (2001). We shall use the first of the two criteria as it generalizes most easily to other percentiles. Based on our precise definition of the median stated earlier, we take for our generalization of the P^{th} percentile value the following (see, for example, Bain and Englehardt (1992)):

> DEFINITION 2: A P^{th} percentile value is a number which puts at least P percent of the data values at that number or below and at least (100 - P) percent of the data values at that number or above. If more than one such number exists, there will be an entire interval of such and we choose the P^{th} percentile value to be the midpoint of that interval.

The question remains, how are such values to be found? We claim that it is the CDF Method 4 which does the job. That the CDF Method meets the definition for all percentiles is not totally obvious and we include a proof for completeness.

THEOREM: The CDF Method 4 provides the P^{th} percentile value for all possible values of P.

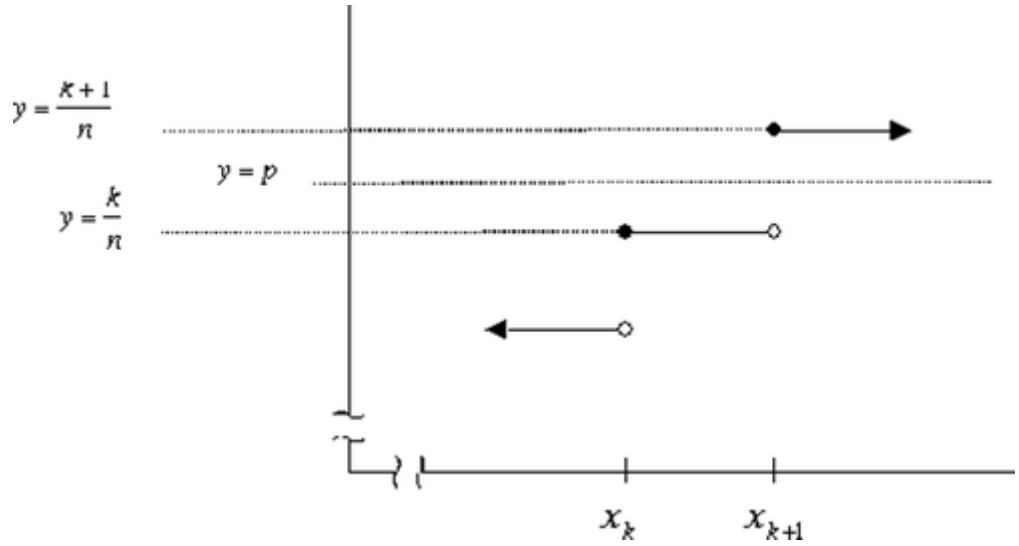
PROOF: We first assume for the sake of simplicity that the data values are all distinct and are ordered. Consider the random variable X which puts probability 1/n at each data value and let be its cumulative distribution function (CDF). In terms of the CDF, a number x is a P^{th} percentile value (note the article) if and only if and . But where so we have that a necessary and sufficient condition that x be a P^{th} percentile value is that

(1)

We see then that we have two cases:

Case 1: The line y = p does not intersect the graph of y = F(x); it passes through a jump at $x = x_{k+1}$. This occurs if and only if

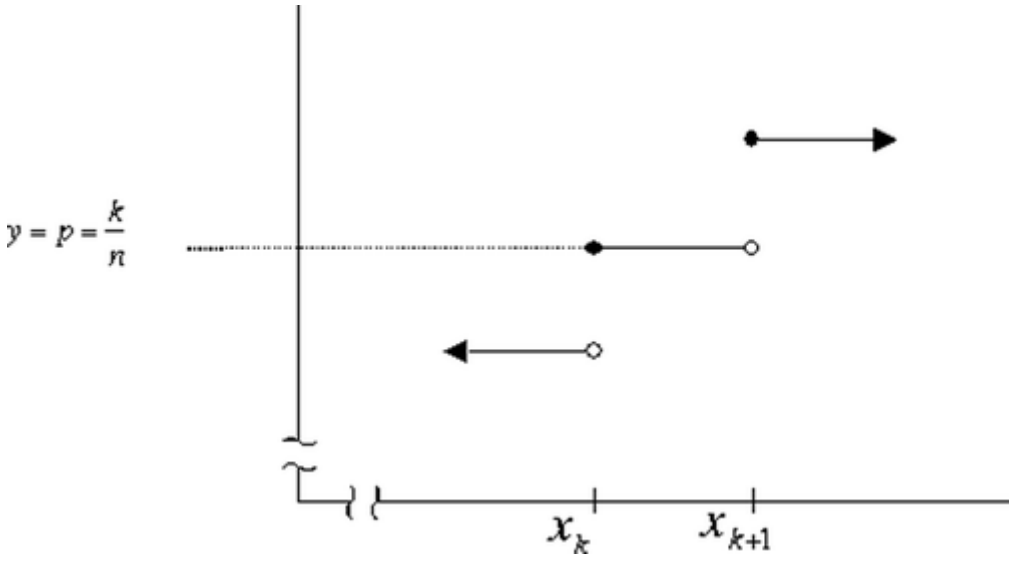
That is, this occurs if and only if np is not an integer and lies between k and k + 1. It is easy to see that $x = x_{k+1}$ is the only value of x which satisfies (1) since if $x > x_{k+1}$ then whereas if $x < x_{k+1}$ then . Therefore x_{k+1} is the P^{th} percentile value. See Figure 2 below.



short-legendFigure 2

Display full size

Case 2: The line y = p does intersect the graph of y = F(x). Since the graph of the CDF has a “stair-step” shape, the line must intersect the graph along an entire interval, say the interval $[x_k, x_{k+1})$. In this case, obviously $p = F(x_k) = k/n$ so that np = k, an integer. Evidently every x satisfying $x_k < x < x_{k+1}$ is a P^{th} percentile value since for every such x, . Moreover, x_k is such a value since $F(x_k) = p$ and . In the same way, x_{k+1} is such a value since and . That there are no other such values is shown as in Case 1. Hence the interval of P^{th} percentile values is $[x_k, x_{k+1}]$ and we select the midpoint and call it the P^{th} percentile value. See Figure 3 below.



short-legendFigure 3

Display full size

If there are repeated values, the argument is similar. Suppose, for example that $x_{k-1} < x_k = x_{k+1} < x_{k+2}$. Then if $np = k$, the line $y = p$ does not intersect the graph, so that we actually have Case 1 in this situation and the argument given in that case shows that we should take x_{k+1} for our P^{th} percentile value. The CDF Method however thinks of this as Case 2 and tells us to average x_k and x_{k+1}; since $x_k = x_{k+1}$ there is no problem. □

A little thought will show that if we are talking only about quartiles, then to meet Definition 2, the first quartile values Q_1 for S_1, S_2, S_3, S_4 would have to be 1.5, 2, 2, and 2 respectively, as any number between 1 and 2 inclusive would serve as a 25^{th} percentile value for S_1. The Lohninger Method 6 does not even provide a 75^{th} percentile value in the case of S_5, but it appears that the M&S Method 5 gives quartile values consistent with the first part of Definition 2 anyway. This is true, but the M&S Method fails to give values which meet even the first part of Definition 2 for other quantiles. As an example, consider finding the second decile value D_2 (i. e. the first quintile) of S_6. Then $(n + 1)p = 7/5 = 1.4$ which rounds to 1, implying that $D_2 = 1$. But this puts only 1/6 = 17% of the data values at or below D_2, rather than the required 20%. Looking at Table 2 we can see that the CDF Method 4 is the only method that provides quartile values consistent with the complete Definition 2.

4. The Doubling Idea

The Inclusive and Exclusive Methods 1 and 2 have the advantage of being easy to comprehend by students and easy to apply, but they have the obvious flaw that in the case of odd n, the median measurement is used twice to compute the upper and lower quartiles (in the case of the Inclusive Method 1) and not at all (in the case of the Exclusive Method 2). A clever student might ask, “why not put half of the median in the top half of the data set and half of the median in the bottom half of the data set?” Of course, we cannot cut a measurement in half, but we can instead, repeat each measurement twice. (For example, look at the data sets $S_4 = (1, 2, 3, 4)$ and $2S_4 = (1, 1, 2, 2, 3, 3, 4, 4)$.) The Inclusive and Exclusive Methods 1 and 2 will then agree on the doubled set, and we can with some justification say that this common value would be what would result if we could cut the middle measurement in half. In Table 3 below we compare the various methods on the doubled data sets. Note that Methods 1, 2, 4, 9, and 10 all agree; in particular they now agree with the CDF Method 4. But comparing Tables 2 and 3 we see that the CDF Method 4 has the same values on both the original set and the doubled set. This makes sense intuitively since it is based on the CDF of the data set considered as a random variable, and from this point of view, the two data sets are the same. But as can be seen from Table 3, of all of the methods, with the exception of the M&S Method 5, this is the only one with this property. This seems to me to be another reason why the CDF Method 4 should be considered “best.” In fact, the CDF Method 4 will satisfy the doubling property for any quantile, whereas the M&S Method 5 will not. Recall the example above of the second decile D_2 applied to S_6, which gave a value of $D_2 = 1$. If we apply the M&S Method to the doubled set $2S_6$, we get $\#(13/5) = \#(2.6)$, so that, rounding off, $D_2 = \#(3) = 2$. Table 3 below compares the lower and upper quartile values (Q_1, Q_3) given by each of the methods for each of the four doubled canonical data sets, together with the interquartile range (IQR).

Table 3. Comparing the Various Methods on the Doubled Data Sets

Download CSV Display Table

5. Summary

In summary, I hope that I have convinced you that the CDF Method 4 is to be preferred as it most closely follows the idea that the lower quartile “puts 25% below and 75% above” and similarly for the upper quartile. In addition, as noted above, except for the M&S Method 6, this is the only one of the nine methods that remains unchanged when the data sets are doubled. The M&S Method has the weakness as described above that it does not fit the Definition 2 idea for other percentiles than the quartiles and fails to have the doubling property in general; two other weaknesses are that its IQR value for S_4 is actually larger than that for S_5, and that one must make a special definition to get the usual definition of the median, as the M&S Method always rounds to a data value. The only drawback to the CDF Method is that, for the average student, it is difficult to motivate and to apply. The Inclusive and Exclusive Methods 1 and 2 are much easier for the average student to grasp. But as I will now show, the CDF Method 4 can be restated in a form similar to these two methods!

I offer the following proposal for classroom use: Define the quartiles by using the “25% below, 75% above” idea and present the Inclusive and Exclusive Methods 1 and 2, discussing the problem of the “middle measurement.” Then tell the students that if they could split the middle measurement in half (one might discuss the doubling idea), they would get quartile values that meet the definition. Then use the following method to calculate the quartiles. As noted before, the CDF Method 4 includes the middle measurement in the case of n= 4k + 1 and excludes it in the case of n= 4k + 3. But in each of these cases, we end up with an odd number of data values in both of the top and bottom halves. Thus the following method is equivalent to the CDF Method 4, yet has the flavor of the Inclusive and Exclusive Methods 1 and 2 and thus should be more accessible to students.

SUGGESTED METHOD: Divide the data set into two halves, a bottom half and a top half. If n is odd, include or exclude the median in the halves so that each half has an odd number of elements. The lower and upper quartiles are then the medians of the bottom and top halves respectively.

I have not yet had the opportunity to test this method in the classroom, but in a statistics class I recently taught, I used Hogg and Ledolter (1992). Not wishing to change the definition of quartiles given in the book, I used the equivalent form which says: Divide the data set into two halves, a bottom half and a top half. If n is odd, include or exclude the median in the halves so that each half has an even number of

halves respectively. The class had no trouble using this definition and thought that it was much easier to apply than the form given in the book. I expect that the situation will be the same in using the suggested method.

Acknowledgements

I would like to thank the Associate Editor and several reviewers for their careful and helpful comments and for references to the literature. I would also like to thank my daughter, Rebecca, for preparing Figures 1, 2, and 3 and for performing the Mathematica calculations. Further thanks are due to Professor Russel John for helpful e-mails.

References

1. Bain, L. J. and Englehardt, M. (1992), Introduction to Probability and Mathematical Statistics (2nd ed.), Belmont, CA: Duxbury Press.

   Google Scholar

2. Benard, A. and Bos-Levenbach, E. C. (1953), “Het Uitzetten van Waarnemingen op Waarschijnlijkheitspapier,” Statistica, 7, 163–173.

   Google Scholar

3. Blom, G. (1958), Statistical Estimates and Transformed Beta-Variables, New York: John Wiley & Sons.

   Google Scholar

4. Brase, C. H. and Brase, C. P. (2003), Understandable Statistics (Concepts and Methods) (7th ed.), Lexington, MA: D. C. Heath and Company.

   Google Scholar

5. Dr. Twe (2002), Reply to “Tom” about quartiles, online at mathforum.org/library/drmath/view/60969.html.

Google Scholar

6. Freund, J. E. and Perles, B. M. (1987), “A new look at quartiles of ungrouped data,” The American Statistician, 41 (3), 200–203.

Web of Science ® | Google Scholar

7. Freund, J. E. and Perles, B. M. (2004), Statistics a First Course (8th ed.), Upper Saddle River, NJ: Pearson Prentice Hall.

Google Scholar

8. Gumbel, E. J. (1939), “La Probabilit” des Hypoth-ses,” Comptes Rendus de l'Académie des Sciences (Paris), 209, 645–647.

Google Scholar

9. Hayden. R. (1997), “Ticky-Tacky Boxes,” online at either exploringdata.cqu.edu.au/docs/tt_box2.docorexploringdata.cqu.edu.au/ticktack.htm

Google Scholar

10. Hazen, A. (1914), “Storage to be Provided in Impounding Reservoirs for Municipal Water Supply,” (with discussion), Transactions of the American Society of Civil Engineers, 77, 1539–1669.

Google Scholar

11. Hoaglin, D. C. (1983), “Letter Values: A Set of Selected Order Statistics” in Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (Editors), Understanding Robust and Exploratory Data Analysis, New York: John Wiley & Sons.

Google Scholar

12. Hoel, P. G. (1966), Elementary Statistics (2nd ed.), New York: John Wiley & Sons.

13. Hogg, R. V. and Ledolter, J. (1992), Applied Statistics for Engineers and Physical Scientists, New York: Macmillan.

Google Scholar

14. Hogg, R. V. and Tanis, E. A. (1997), Probability and Statistical Inference (5th ed.), Upper Saddle River, NJ: Prentice Hall.

Google Scholar

15. Hyndman, R. J. and Fan, Y. (1996), “Sample quantiles in statistical packages,” The American Statistician, 50 (4), 361–365.

Web of Science ® | Google Scholar

16. Joarder, A. H. and Firozzaman, M. (2001), “Quartiles for discrete data,” Teaching Statistics, 23, 86–89.

Google Scholar

17. John, R. (2000), “How Statistics Packages Calculate Sample Quartiles”; an earlier version of this paper entitled “How to Calculate a Quartile (If You Must),” can be found online at www.maths.murdoch.edu.au/units/statsnotes/samplestats/quartilesmore.html

Google Scholar

18. Johnson, R. A. (2000), Miller and Freund's Probability and Statistics for Engineers (6th ed.), Upper Saddle River, NJ: Prentice Hall.

Google Scholar

19. Johnson, R. A. and Bhattacharyya, G. K. (1996), Statistics - Principles and Methods (3rd ed.), New York: John Wiley & Sons.

Google Scholar

20. Journet, D. (1999), “Quartiles: How to Calculate Them?” online at www.hajweb.org/medicineprices/manual/quartiles_iTSS.pdf

Google Scholar

21. Khazanie, R. (1996), Statistics in a World of Applications (4th ed.), New York: HarperCollins.

Google Scholar

22. Lohninger, H. (1999), Teach/Me Data Analysis, Berlin-New York-Tokyo: Springer-Verlag.

Google Scholar

23. McClave, J. T. and Sincich, J. (2003), A First Course in Statistics (8th ed.), Upper Saddle River, NJ: Prentice Hall.

Google Scholar

24. Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2003), Introduction to Probability and Statistics (11th ed.), Pacific Grove, CA: Brooks/Cole-Thompson.

Google Scholar

25. Mendenhall, W. and Sincich, T. (1995), Statistics for Engineering and the Sciences (4th ed.), Upper Saddle River, NJ: Prentice Hall.

Google Scholar

26. Milton, J. S., McTeer, P. M., and Corbet, J. J. (1997), Introduction to Statistics, New York: McGraw-Hill.

Google Scholar

27. Moore, D. S. (1996), Statistics - Concepts and Controversies (4th ed.), New York: W. H. Freeman and Co.

Google Scholar

28. Moore, D. S. (2003), The Basic Practice of Statistics (3rd ed.), New York: W. H. Freeman and Co.

29. Moore, D. S. and McCabe, G. P. (2003), Introduction to the Practice of Statistics (4th ed.), New York: W. H. Freeman and Company.

Google Scholar

30. Parrish, R. S. (1990). “Comparison of quantile estimators in normal sampling,” Biometrics, 46, 247–257.

Web of Science ® | Google Scholar

31. Parzen, E. (1979), “Nonparametric Statistical Data Modeling” (with discussion), Journal of the American Statistical Association, 74, 105–131.

Web of Science ® | Google Scholar

32. Peck, R., Olsen, C., and Devore, J. (2001), Introduction to Statistics and Data Analysis, Pacific Grove, CA: Duxbury Press.

Google Scholar

33. Ross, S. M. (1996), Introductory Statistics, New York: McGraw-Hill.

Google Scholar

34. SAS Institute, Inc. (1990), SAS® Procedures Guide, Version 6 (3rd ed.), Cary, NC: SAS Institute, Inc.

Google Scholar

35. SAS Institute, Inc. (1994), JMP® User's Guide, Version 3, Cary, NC: SAS Institute, Inc.

Google Scholar

36. Siegel, A. F. and Morgan, C. J. (1996), Statistics and Data Analysis - An Introduction (2nd ed.), New York: John Wiley & Sons.

Google Scholar

37. Smith, P. J. (1997), Into Statistics: A Guide to Understanding Statistical Concepts in

Google Scholar

38. Snedecor, G. W. (1946), Statistical Methods Applied to Experiments in Agriculture and Biology (4th ed.), Ames, IA: Iowa State College Press.

Google Scholar

39. TI-83 Plus Graphing Calculator Guidebook, Texas Instruments Inc. (1999)

Google Scholar

40. Tukey, J. W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

Google Scholar

41. Vining, G. G. (1998), Statistical Methods for Engineers, Pacific Grove, CA: Duxbury Press.

Google Scholar

42. Weibull, W. (1939), “The Phenomenon of Rupture in Solids,” Ingeni-rs Vetenskaps Akademien Handlingar, 153, 17.

Google Scholar

43. Wessa, P. (2006), Free Statistics Software, Office for Research Development and Education, version 1.1.18, online at www.wessa.net

Google Scholar

Appendix:

Data Sets with Many Repetitions

If there are many repetitions of a few distinct data values the definitions of this paper are not appropriate, even for the median. (This situation might occur, for example, in student evaluations where students are asked to rate their instructor on a 1 to 5 scale.) As an example, consider the two data sets (3, 3, 3, 3, 4, 4, 4) and (3, 3, 3, 4, 4, 4, 4). Using the definition of the median given in this paper would lead to a median value of 3

gives a misleading impression of the data. A solution is to change our definition of the median (and other percentiles) by considering the data to be pooled data in a histogram. For example, the values of “3” are to be considered to be uniformly smeared over the interval from 2.5 to 3.5. This makes the discrete distribution continuous, and we then simply divide the histogram into two equal areas to find the median. Using this method, we find that the median of the (3, 3, 3, 3, 4, 4, 4) data set would occur .875 of the way through the “3” class interval so that it would be equal to 2.5 + 0.875 = 3.375. The median of the (3, 3, 3, 4, 4, 4, 4) data set would occur 0.125 of the way through the “4” class interval so that it would be equal to 3.5 + 0.125 = 3.675. These values provide a much more meaningful comparison of the two data sets. (See, for example, Freund and Perles (2004) or Hoel (1966), p. 37.)