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 met

Percentiles

1. I

It is well ... finitions in the litera ... be noted that I am ... set has 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 ... dent.

Element ... nd Perles (1987), ... (1999), and Wes ... tor which will cal ... discussion of quar ... 96).

It might ... example, the TI-83 ... Microsoft Excel) in ... of quartiles would be ... or a 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
> ..., and they
> ...rovided by

2. Th...

Fortuna... ...ooked were
consiste... ...n of data
values, t... ... if n is even,
the med... ...ue for some
of the S... ...s later in

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
> rval.

The mos ividing the
data set wer (first)
quartile ile Q_3 to be
the med students. It
work o with the
medi answers. For
the rema data set.

METHOD and a top
half. If n artile is the

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... is half way between x_1 and x_2

and $x_{2.2}$... , j + 1 can

be writte... is not an

integer.

In his cla... -whisker

plot a... er hinges

(see ... plot and,

together ... umber

summar... e counts

upwards ... downward

rank of t... which in this

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 ... e Tukey letter va...

In gener... above, they define p... er quartile (75th ... ually based on th... which puts "half of ... ous discussi... e Pth percenti... of the data set abov... e precise as we have already done for the median. (For simplicity of notation, we let p = P/100, so that, for example, the 50th 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 Pth percentile value is found as follows. Calculate np. If np is an integer, then the Pth percentile value is the average of #(np) and #(np + 1). If np is not an integer, the Pth 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 anot

METHOD ... p) with p =
0.25 for ... to the
nearest ... artile and
down for ... , then #((n
+ 1) ... thod 11 with
comp ... Method 10
with rou ... and “round
to the ne ... the quartiles
when (n

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 exam… and so we average…

These a… using the weighte… ite small.” This pro… ves a value of 1.… y used by any of t… s used by Mathem…

METHOD … e with #(np + 0.5). I… s 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 Pth 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 … with
interpola … — 1)p + 1)
= #(2) a …

I have n … ndman and
Fan (… Perles
(1987) a …

Note tha … r Method 6
are what … epth in the
sense of … + 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 met … Method 6
of Wessa … Wessa's
method … ger to np,
rounding … t to taking
and that … even.

MET … is as
“PCTL … to see that
this give …

This met … essa (2006).
It is also …

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

Downlo

We can … the Vining
Method … &F Method 8
is se… consider
these … "averaging"
methods … CDF Method
4 and th… een two
successi… polation"
methods … ween

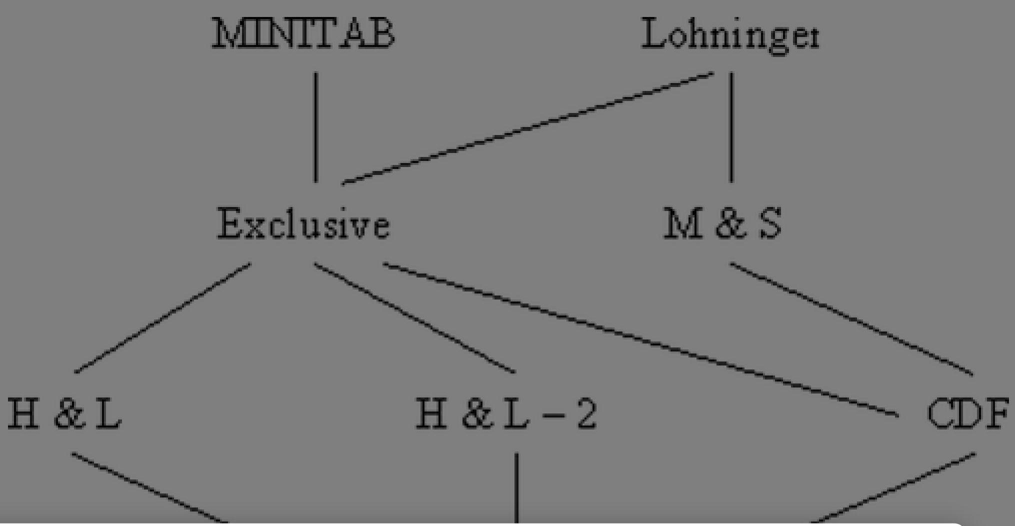
The M&S Method 5 and the Lohninger Method 6 are unique in the sense that they give only values which are data values themselves. The other averaging methods all agree if n is even, whereas if n is odd, then the CDF Method 4 agrees with the Inclusive Method 1 if n is of the form 4k + 1 and with the Exclusive 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 inte ” the data
values s n from
which th ibution. (C.
f. the me 0) for
graphs ly
strai ple, one
would e uld be the
same as for Methods
10 and 1 od 12 gives
different 2, 3, 3, 4,

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-leg

Display

3. Eva

What ma ... calculating

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.

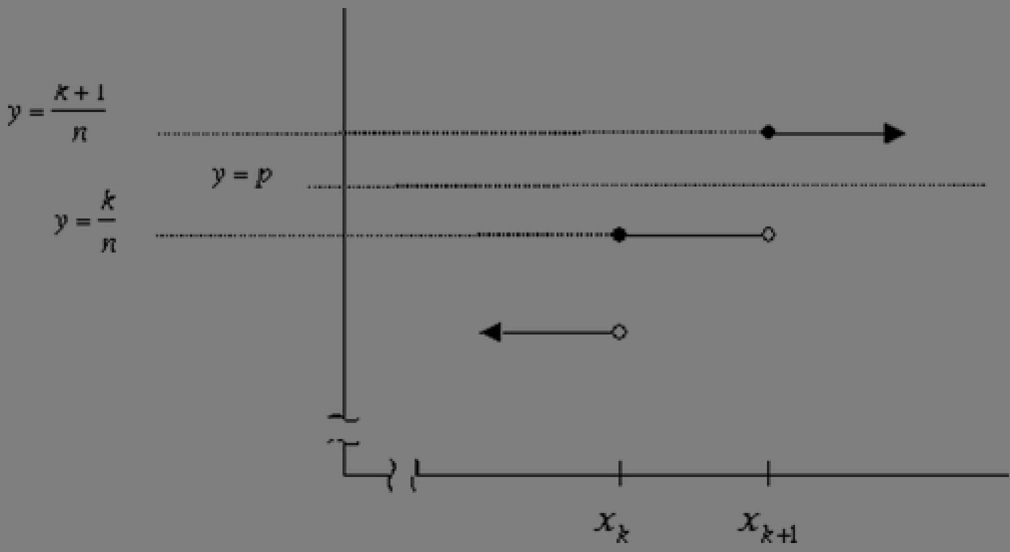
PROOF: ... ll distinct
and are ... at each
data val ... the CDF, a
number ... where so
we ha ... value is that

We see t ...

Case 1: ... rough a
jump at ...

In this article

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 … d call it the P^{th} perce…

short-leg…

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

compreh … that in the

case of … r and lower

quartiles … f the

Exclusiv … edian in the

top half … a set?" Of

cours … ach

meas … d $2S_4$ = (1,

1, 2, 2, … gree on the

doubled … would be

what wo … 3 below we

compare … 1, 2, 4, 9,

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

5. Summary

In summary, I hope that I have convinced you that the CDF Method 4 is to be preferred
as it mos … and 75%
above” … ept for the
M&S Me… nged when
the data … above that
it does … nd fails to
have … R value for
S_4 is … nition to get
the usu… data value.
The only … difficult to
motivate … ch easier for
the aver… be restated

In this article

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 elements. The lower and upper quartiles are then the medians of the bottom and top halves respectively. The class had no trouble using this definition and thought that it was muc situation will be t

Acknov

I wou eful and helpfu ank my daughte Mathem or helpful e-mails.

In this article

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 ... ed data,"
The A...

7. Freund ... pper Saddle
Ri...
Go...

8. Gumb ... l'Académie
des Sc...
Goog...

In this article

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.
Google Scholar

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.
Goog

15. Hyndr es,” The
Ameri

16. Joa aching
Statist

17. John, earlier

In this article

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.haiweb.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. McCla ... Upper
Saddl
Goog

24. Mende ... obability
and S
G

25. Mende ... Sciences
(4th e
Goog

In this article

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.
Google Scholar

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, ata Analysis, Pacific
Goog

33. Ros
G

34. SAS Ir ary, NC: SAS Institu
Goog

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 Engineering and the Sciences, Berlin; New York: Springer.
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.
Goog

42. Weibu enskaps
Akade
Goog

43. We nt and
Educa
Goog

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 for the first set and that of 4 for the second set. This method of calculating the median 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.)

How Significant is a Boxplot Outlier?

Robert Dawson
Journal of Statistics Education
Published online: 29 Aug 2017

View more