

Meeting the Median, Quartiles and Percentiles: Measurements of Position

Dov Chelst

The general goal of descriptive statistics is to make sense of a set of data values however large or small that set may be. We have already learned to measure the center and the spread of a data set by computing its mean (average) and standard deviation. Now, let's explore another technique.

Most of you are familiar with standardized aptitude tests—SAT, ACT, etc. Imagine that after taking the SAT, you are told that you scored a 640 out of 800 in mathematics. At the same time, this score is accompanied by a percentile value such as 83. This means that that 83% of those examined received a score that was 640 or lower while the rest (100-83=17%) scored a 640 or better. The percentile value adds another level of meaning to the raw score of 640, stating its *position* within the set of all scores.

Now, it's our turn to apply this technique to our own data sets. Given some data and a specific *percentile value* P from 0% to 100%, how can I locate the *data value* that best represents it? Here's my recipe.

Dov's Percentile Recipe (Yum!) ¹

Ingredients: A set of N data values, a percentile value P and a calculator.

Step #1 *Sort* your data from lowest to highest.

Many students neglect this important step.

(If working by hand, a stem-and-leaf plot is often useful.)

Step #2 *Math:* Convert your number into a rank A as follows.

Multiply N times P divided by 100. $A = \frac{NP}{100}$.

Step #3 *Adjust:* If the result A is an integer, then add 0.5 to it.

Otherwise, round up.

Step #4 *Locate:* Go to the specific rank A in your sorted data counting up from the lowest value. By the way, a half location is just the average of the nearest two values.

The 50th percentile, P_{50} , is known as the **median** \tilde{x} and it represents the middle of a data set. The 25th and 75th percentile values, P_{25} and P_{75} , are called the first and third **quartile** values and usually written Q_1 and Q_3 respectively. In fact, to get a reasonable picture of a data set, we might quickly compute a **five number summary** listing in order the minimum, Q_1 , median, Q_3 and maximum values.

No good recipe is complete without a picture of the final result. A year ago, I surveyed the students in my undergraduate statistics class. Here is a list of twenty of their ages, sorted for our convenience. Let's practice locating the median, quartile values and also the 83rd percentile P_{83} value.

Student Ages

18 18 18 19 19 19 20 21 21 23
24 28 30 30 31 31 33 33 35 38

Now, let's locate our percentile values. Here $N = 20$

- Median (P_{50}): $\frac{20(50)}{100} = 10 \Rightarrow 10.5$; the average of the 10th and 11th values is $\frac{23+24}{2} = 23.5$
- First Quartile (P_{25}): $\frac{20(25)}{100} = 5 \Rightarrow 5.5$; the average of the 5th and 6th values is $\frac{19+19}{2} = 19$
- Third Quartile (P_{75}): $\frac{20(75)}{100} = 15 \Rightarrow 15.5$; the average of the 15th and 16th values is $\frac{31+31}{2} = 31$
- 83rd Percentile (P_{83}): $\frac{20(83)}{100} = 16.6 \Rightarrow 17$; the 17th value in our data set is 33
- To summarize: Median=23.5, $Q_1=19$, $Q_3=31$ and $P_{83}=33$

Now, imagine that a new student enrolls after the end of the first week of class. This student is an 80-year old senior citizen who has always wanted to study statistics. With this new set of age data ($N=21$), whose maximum value is now 80, how are our percentile calculations affected?

- Median: $\frac{50(21)}{100} = 10.5 \Rightarrow 11$. The 11th value is 24.
- First Quartile: $\frac{25(21)}{100} = 5.25 \Rightarrow 6$. The 6th value is 19.
- Third Quartile: $\frac{75(21)}{100} = 15.75 \Rightarrow 16$. The 16th value is 31.
- 83rd Percentile: $\frac{83(21)}{100} = 17.43 \Rightarrow 18$. The 18th value is 33.
- To summarize: Median=24, Quartiles are 19 and 31, $P_{83} = 33$.

Notice that the addition of our atypical student, or *outlier*, has little or no effect on any of the calculated percentile values. In contrast, if we were to compare the original and final averages and standard deviations, we would find a larger effect. The average changes from 25.45 to 28.05. The standard deviation changes from 6.65 to 13.56.

Just to make sure that we all have the right idea, let's calculate the five number summary for another set of data. Here are the final exam grades for students taking College Algebra during the Spring 2003 semester. A total of 38 students took the exam and their scores are listed below.

Final Exam Grades - Spring 2003

12 31 41 43 52 53 53 56 57 58
62 66 66 67 68 73 76 76 77 78
79 81 81 82 83 85 87 87 89 91
91 92 99 101 101 102 103 106

Now we compute the five number summary. I guarantee that the first and last steps will be effortless!

- Minimum: **12**
- First Quartile: $A = \frac{25(38)}{100} = 9.5 \Rightarrow 10$; the 10th value is **58**
- Median: $A = \frac{50(38)}{100} = 19 \Rightarrow 19.5$; the average of the 19th and 20th values is $\frac{77+78}{2} = \mathbf{77.5}$
- Third Quartile: $A = \frac{75(38)}{100} = 28.5 \Rightarrow 29$; the 29th value is **89**
- Maximum: **106**

That's it. If you follow my recipe and look at the examples, you should be able to calculate your own percentiles in no time. Who needs those SAT people anyway? However, let me leave you with one last bit of information and a warning.

Note that the **interquartile distance**, or $Q_3 - Q_1$ can be used to measure the spread of your data.

Warning! Finally, I should warn you that percentile recipes may vary. The results won't differ much, but they won't match exactly. In particular, another popular procedure for computing quartiles will not equal the values listed above. Moreover, the quartile function shipped with some versions of Microsoft Excel uses a formula that I cannot determine. So, don't rely on Excel's quartile function to yield accurate values while you're in this class. By the way, you will find that the larger the set of data values, the less these different recipes will disagree.