Memo

To: Colleagues
From: Roger Bakeman
Date: May 25, 2020
Re: How to compute percentile scores

Probably most of us think we know how compute percentile scores—including the 50th percentile or median—until we actually try. We are probably lulled by the fact that most statistics in common use are computed without controversy. Their conceptual definition admits to a single algorithm—a single, widely known and used computational formula. Not so percentile scores. There is not one method but many and rarely do computational programs identify which one they use. Worse, most books and web sites I have consulted provide some combination of confusing, contradictory, or incomplete information.

The conceptual definition—a percentile score is a score **below which** that percentage of the scores in a series fall—seems deceptively simple, but admits to many algorithms, each giving varying results. To complicate matters, one conceptual variant defines a percentile score as a score **equal to or below which** that percentage of the scores in a series fall. The first definition is **exclusive** and the variant is **inclusive**.

The input for all algorithms is the same: a series of *N* scores ordered from smallest to largest. Thereafter algorithms are generally of two kinds. The simpler kind of algorithm **identifies a score** in the series, whereas the second kind **interpolates between scores** in the ordered series. When *N* is large (over a hundred at least) and scores are well distributed (few duplicates, gaps between scores don't vary much), differences between the two types of algorithms are slight. But when scores are few, duplicates many, and gaps between scores vary, algorithms that don't interpolate can produce strange and crude results. As a result, most computer programs that produce percentile scores use interpolation.

The devil is in the defaults. When reporting percentiles—most commonly the 50th (aka the median) or the 25th and 75th (aka lower and upper quartiles)—or providing a box-and-whisker plot, most authors (me included) don't state the method used. Chances are that most of us aren't even aware that there are different methods; we just accept the defaults for whatever program we are using.

As a first example, consider the **nearest-rank method**, a method that identifies a score in the series and does not interpolate. It is useful for teaching the concept and often appears in statistical texts and web pages but is a bit crude. It can be either **exclusive** or **inclusive**, although many sources simply present one or the other definition as the only definition. The method is worth understanding because interpolation methods are similar and because it demonstrates nicely the different results exclusive and inclusive definitions give, but I wouldn't recommend it otherwise. For one thing, with an even number of scores, the 50th percentile won't be the mean of the two most central scores, as most of us expect a median to be.

Here's an example of the **nearest-rank method** with an odd number of scores (*N* = 5).

		Exclusive		Inclusive	
Desired percentile (DPC)	**Ordered rank = *N* × DPC**	**R = Rank > ordered rank**	**Percentile score = raw sore at rank R**	**R = Rank ≥ ordered rank**	**Percentile score = raw sore at rank R**
.25	1.25	2	34	2	34
.40	2	3	47	2	34
.50	2.5	3	47	3	47
.60	3	4	54	3	47
.75	3.75	4	54	4	54
.90	4.5	5	81	5	81

Rank	**Raw score**
1	12
2	34
3	47
4	54
5	81

For each desired percentile, compute its **ordered rank** (*N* × desired percentile). Ask what rank is **greater than** the ordered rank (exclusive) or **greater than or equal to** the ordered rank (inclusive). The percentile score is the raw score at that rank.

Exclusive and Inclusive Interpolated Methods

More useful—and far more used—are methods that interpolate between the closest **percentile ranks**, either **inclusively** (the Excel default) or **exclusively** (the SPSS default)—although exclusively is the Excel default for its box-and-whisker plots. The only difference between methods that interpolate exclusively or inclusively is the formula used to compute the **percentile rank**. Thereafter the algorithms are the same.

1. For exclusive methods, percentile rank = desired percentile × (N + 1).
2. For inclusive methods, percentile rank = desired percentile × (N – 1) + 1.

The percentile rank is used to select two scores in the series. The percentile score is then interpolated between those two scores, taking into account the distance between them and the fractional part of the percentile rank. Sometimes only one score is selected, not two, in which case no interpolation is necessary.

Interpolated inclusive percentiles are nearer the median than exclusive percentiles; exclusive percentiles are further away. As a result, the **interquartile range** (75^{th} minus 25^{th} percentile, the box in a box-and-whisker plot) is **smaller** when defined with **inclusive** instead of **exclusive** percentile scores. The exclusive definition gives a fatter box and, other things being equal, is the one I prefer.

Here' an example of the **exclusive interpolated** method using the same desired percentiles and five scores used previously.

Desired percentile (DPC)	Percentile rank: PR = DPC × (*N* + 1)	R1 = integer part of PR	If PR has no fractional part, R2 = R1. Otherwise R2 = R1 + 1.	S1 = raw sore at rank R1	S2 = raw sore at rank R2	Difference: Δ = S2 – S1	F = fractional part of PR	Percentile score = S1 + F × Δ
.25	1.5	1	2	12	34	22	0.5	23
.40	2.4	2	3	34	47	13	0.4	39.2
.50	3	3	3	47	47	0	0	47
.60	3.6	3	4	47	54	7	0.6	51.2
.75	4.5	4	5	54	81	27	0.5	67.5
.90	5.4	5	6	81	undefined	undefined	0.4	undefined

With the exclusive method, if scores are few, a high desired percentile may be undefined (Excel returns the #NUM! error). Specifically, it twill be undefined if its value times *N* + 1 is greater than *N* (as here, .90 × 5 = 5.4 > 5). Of course, with few scores it doesn't make much sense to ask for a high percentile score.

And here' an example of the **inclusive interpolated** method using the same desired percentiles and five scores.

Desired percentile (DPC)	Percentile rank: PR = DPC × (*N* – 1) + 1	R1 = integer part of PR	If PR has no fractional part, R2 = R1. Otherwise R2 = R1 + 1.	S1 = raw sore at rank R1	S2 = raw sore at rank R2	Difference: Δ = S2 – S1	F = fractional part of PR	Percentile score = S1 + F × Δ
.25	2	2	2	34	34	0	0	34
.40	2.6	2	3	34	47	13	0.6	41.8
.50	3	3	3	47	47	0	0	47
.60	3.4	3	4	47	54	7	0.4	49.8
.75	4	4	4	54	54	0	0	54
.90	4.6	4	5	54	81	27	0.6	70.2

Excel and SPSS Details

In **Excel**, as noted earlier, the default when using the **percentile** function is interpolated inclusive. Since the 2010 version Excel has included two variants—**percentile.inc** and **percentile.exc**—so you can select which you prefer. However, when using Excel to produce box-and-whisker plots, the default is exclusive, but an option lets you select which you prefer. If you prefer a more compact box you could select inclusive (although I'm not sure when this would be conceptually justified).

The FREQUENCY command (**SPSS**, version 23) computes percentiles (and quartiles) using the interpolated exclusive method. The EXPLORE command likewise uses the exclusive method (called weighted average definition 1) as the default. With syntax, five other methods can be specified (weighted average definition 2, rounded, empirical, and averaged empirical), but none of these are the interpolated inclusive method. SPSS does not seem to have a way to compute inclusive percentiles. At the moment at least, SPSS (Version 23) seems less flexible than Excel.

Bottom Line

I would recommend always using the **exclusive interpolated method**. If not already the default, select it explicitly.

Interpolation allows more differentiated percentile scores, and exclusive fits better our common-sense notion—and the more common conceptual definition—of what a percentile score is: If you score **at** the 90^{th} percentile, 90% of the scores fall **below** yours, not **at or below** yours. Moreover, I can't really think of a solid conceptual reason to prefer the inclusive method—but, reassuringly, with either method, the 50^{th} percentile will always be the central score when the number of scores is odd and the mean of the two most central scores when the number of scores is even.

Most computer programs, whether they tell you or not, use an interpolated method and, as noted, the only choice with SPSS seems to be an exclusive method. But at least now you'll know the difference. And even if it doesn't dramatically affect your professional practice, consider the knowledge satisfying for its own sake. Isn't it intrinsically satisfying to know something most of your colleagues don't?