Percentiles and textbook definitions – confused or what?

Take the following definitions ...

HyperStat Online:
A percentile rank is the proportion of scores in a distribution that a specific score is greater than or equal to. For instance, if you received a score of 95 on a math test and this score was greater than or equal to the scores of 88% of the students taking the test, then your percentile rank would be 88. You would be in the 88th percentile.
http://www.ruf.rice.edu/~lane/hyperstat/A79766.html

Hinkle, D., Wiersma, W., & Jurs, S. (1994). Applied statistics for the behavioral sciences. (3rd ed.). Boston: Houghton Mifflin Company. (p. 49-50)
A percentile is the point in a distribution at or below which a given percentage of scores is found. For example, the 28th percentile of a distribution of scores is the point at or below which 28% of the scores fall.

Monroe County School District, Florida, US
The percentile is a point on a scale of scores at or below which a given percent of the cases falls. For example, a child who scores at the 42 percentile , *is doing as well as, or better than*, 42 percent of the students who took the same test.
http://www.monroe.k12.fl.us/poinciana/what_is_a_percentile_or_percenti.htm

Wisconsin Department of Public Instruction
A percentile is a value on a scale that indicates the percent of a distribution that is equal to it or below it. For example, a score at the 95th percentile is equal to or better than 95 percent of the scores.
http://www.dpi.state.wi.us/dpi/standards/mathglos.html

Moore, D.S. and McCabe, G.P. (1993) Introduction to the Practice of Statistics 2nd Edition. New York: W.H. Freeman and Company (p. 40)
The *pth* percentile of the distribution is the value such that *p* percent of the observations fall at or below it.

Hays, W.L. (1994) Statistics, 5th Edition. Florida: Harcourt Brace. (p. 194)
In any frequency distribution of numerical scores, the percentile rank of any specific value *x* is the percentage of the total cases that fall at or below *x* in value.

Kiess, H.O. (1996) Statistical Concepts for the Behavioral Sciences. London: Allyn and Bacon (p. 46)
A percentile is the score at or below which a specified percentage of scores in a distribution falls.

STATISTICA 6 (Statsoft Inc.)
The *percentile* (this term was first used by Galton, 1885a) of a distribution of values is a number x_p such that a percentage p of the population values are **less than or equal to** x_p. For example, the 25th *percentile* (also referred to as the .25 quantile or lower quartile) of a variable is a value (x_p) such that 25% (p) of the **values of the variable fall below** that value.

Howell, D. (1989) Fundamental Statistics for the Behavioral Sciences. 2nd Edition. Boston: PWS-Kent Publishing. (p. 36)
A percentile is the point on a scale at or below which a given percentage of the scores fall.
*contrast this with the definition by Howell (2002) at the bottom of page 2 overleaf!!

Contrast the above with the following:

Bartram, D. and Lindley, P.A. (1994) BPS Level A Open Learning Training Manual: Scaling Norms and Standardization, Module 2, part 1. London: BPS Publications (p.17)
The proportion of people scoring less than a particular score is called the percentile rank of the score. More commonly we refer to this as just the percentile.

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston. (p. 439)
Loosely speaking, the percentile rank corresponding to a particular raw score is interpreted as the percentage of examinees in the norm group who scored below the score of interest.

Testing And Assessment: An Employer's Guide To Good Practices. A document by the U.S. Department of Labor Employment and Training Administration 1999
Percentile score: The score on a test below which a given percentage of scores fall. For example, a score at the 65th percentile is equal to or higher than the scores obtained by 65% of the people who took the test.
http://www.911dispatch.com/job_file/eta_pub.html#appendixb

Pagano, R.R. (1994) Understanding Statistics in the Behavioral Sciences. 4th Edition. New York: West Publishing Company. (p. 44)
A percentile or percentile point is the value on the measurement scale below which a specified percentage of the scores in a distribution fall.

Kline, P. (2000) A Psychometrics Primer. London; Free Association Books. (p. 41) and Kline, P. (2000) A Handbook of Psychological Testing. London: Routledge. (p. 59)
A percentile is defined as the score below which a given proportion of the normative group falls.

Ferguson, G.A. and Takane, Y. (1989) Statistical Analysis in Psychology and Education 6th Edition. New York: McGraw-Hill (p. 482)
If k percent of the members of a sample have scores less than a particular value, that value is the k^{th} percentile point.

Rosenthal, R. and Rosnow, R. (1991) Essentials of Behavioral research: Methods and Data Analysis 2nd Edition. New York: McGraw-Hill. (p. 625)
A percentile is the location of a score in a distribution defining the point below which a given percentage of the cases fall. E.g. a score at the 90th percentile falls at a point such that 90 percent of the scores fall at or below that score.

Cronbach, L.J. (1990) Essentials of Psychological Testing 5th Edition. New York: Harper Collins. (p. 109-110).
"Tony stands third out of 40 on Test A, tenth on test B". Because ranks depend upon the number of persons in the group, we have difficulty when group size changes. Therefore ranks are changed to percentile scores. A percentile rank tells what proportion of the group falls below this person.

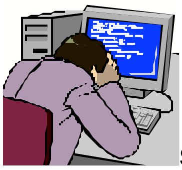
Howell, D.C. (2002) Statistical Methods for Psychology 5th Edition. Duxbury Press. (p. 62)
Finally, most of you have had experience with percentiles, which are values that divide the distribution into hundredths. Thus the 81st percentile is that point on the distribution below which 81% of the scores lie.

Glass, G.V. and Hopkins, K.D. (1996) Statistical Methods in Education and Psychology, 3rd Edition. London: Allyn and Bacon. (p. 25)
Percentiles are points in a distribution below which a given p percent of the cases lie.

Fisher, L.D. and van Belle, G. (1993) Biostatistics: a methodology for the Health Sciences. New York: Wiley. (Wiley Series in Probability and Mathematical Statistics) (p. 51)
The 25th percentile is that value of a variable such that 25% of the observations are less than that value, and 75% of the observations are greater.

Armitage, P. and Berry, G. (1994) Statistical Methods in Medical Research, 3rd edition. London: Blackwell Science. (p. 34)
The value below which P% of the values fall is called the P^{th} percentile

SPSS Inc. (version 10.05)
Percentiles are values that divide cases according to values below which certain percentages of cases fall. For example, the median is the 50% percentile, the value below which 50% of the cases fall.

**So, what exactly is it?**

A percentile is the point in a distribution **at or below** which a given percentage of scores is found

-or-

The value **below** which P% of the values fall is called the P^{th} percentile

Answer:
In fact, both definitions are correct. What is at fault is the lack of clarity in some cases over what constitutes a "score". Let's use the median to exemplify what's going on.

All authors invariable refer to an observed frequency distribution which is referred to a continuous value, real-number distribution like the Normal Distribution. Further, examples will be given in terms of the median value for a set of scores, which is that number above and below which 50% of the scores in a distribution lie. In short, the 50^{th} percentile. If you recall, the calculation for the median for an odd-numbered set of *ordered* scores is the middle value. So, if there are 5 *ordered* scores, the median is the 3^{rd} score in the series. If it is an equal number of scores (say 4), then the median is the average of the 2^{nd} and 3^{rd} score. Note carefully, this score is sometimes not defined when using integer test scores e.g. take four scores on a test which is scored out of 10, in integer units … 2, 4, 5, 9. The median of these scores is (4+5)/2 = 4.5. This is the 50^{th} percentile score – yet no-one can ever obtain it as the test scores are always 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
So, the **most correct** definition for a percentile is, given this example is:
The value **below which** P% of the values fall is called the P^{th} percentile
as this is the score below which 50% of the observations will lie. And nobody can equal it.

But, now take the scores 2, 4, 5, 8, 9. The median is 5. This is an attainable score. What do we say if someone scores a 5? You guessed it … *the person scores **at** the 50^{th} percentile – attaining a median score*. So the definition that now looks most appropriate in this case is:
A percentile is the point in a distribution **at or below which** a given percentage of scores is found.

So how can both be correct – yet seem to be more appropriate under different conditions? The clue is spread throughout the various texts quoted above. The test score, although in many cases an integer value, is in fact deemed a point-estimate of a hypothetical interval of continuous real-value number scores. So, a test score of 4 is actually considered to be a point-estimate of scores that can range from 3.5 through to 4.4999999999999999999999999999999999999999. Therefore, when computing the median of 2, 4, 5, 9 as (4+5)/2 = 4.5, we are in fact computing an average of 4.4999999999999999999999999999999999999999 + 4.5 = 4.5 (rounded). The first number is the upper bound of the point-estimate 4.0. The second number is the lower bound of the point-estimate 5.0.

Now take the example 2, 4, 5, 8, 9. The median is 5. But, the upper bound of this number is 5.49999999999999999999999999999999. It is a verbal "shorthand" that states that 5 is the median – in fact the upper bound of the median is 5.49999 etc (note it could also be as low as 4.5 given the definition of a point-estimate number).

So, we have to be very careful with our terminology of what a "score" is actually said to represent. If we are referring to observed, integer-value scores, without any regard to the hypothetical score intervals, then to find the percentile of a distribution of scores requires finding that single observed score which cleanly separates the scores above and below it into an integer percentile. i.e the score value below which 33% of the scores lie, and above which 67% of the score lie. This one score will be the 33^{rd} percentile. However, unless we have extremely large samples of scores (in the thousands), **and** a test score range of exactly 0 to 100 in unit (=1) steps, this is never likely to happen. So, the most efficient way of always being able to compute an exact percentile score is by using a standard formula to calculate any required percentile for any frequency distribution of scores. What this requires however is that we taken into account the upper and lower bound for every integer score – assuming that each exact integer score is actually the middle score of an interval extending 0.5 either side … in which an infinity of continuous, real-valued scores can be theoretically “observed” (which begs the question “how!!?!”).

The formula is:

$$P_i = ll + \left(\frac{np - cf}{f_i} \right) \cdot w$$

where

$P_i =$ the i^{th} percentile

$ll =$ the exact lower limit of the interval containing the percentile point

$n =$ the total number of scores

$p =$ the proportion corresponding to the desired percentile

$cf =$ the cumulative frequency of scores below the interval containing the percentile point

$f_i =$ the frequency of scores in the interval containing the i^{th} percentile point

$w =$ the width of the class interval

Let’s take an example of some test scores … the EPQR Extraversion scale, with a 0-21 test score range…

	Frequency table: LONG_E (EPQR100M.STA)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	9	9	1.475410	1.4754
1	12	21	1.967213	3.4426
2	13	34	2.131148	5.5738
3	17	51	2.786885	8.3607
4	16	67	2.622951	10.9836
5	12	79	1.967213	12.9508
6	15	94	2.459016	15.4098
7	16	110	2.622951	18.0328
8	22	132	3.606557	21.6393
9	26	158	4.262295	25.9016
10	32	190	5.245902	31.1475
11	31	221	5.081967	36.2295
12	36	257	5.901639	42.1311
13	31	288	5.081967	47.2131
14	29	317	4.754098	51.9672
15	33	350	5.409836	57.3770
16	39	389	6.393443	63.7705
17	35	424	5.737705	69.5082
18	29	453	4.754098	74.2623
19	31	484	5.081967	79.3443
20	34	518	5.573770	84.9180
21	39	557	6.393443	91.3115
22	33	590	5.409836	96.7213
23	20	610	3.278689	100.0000
Missing	0	610	0.000000	100.0000

What would be the 75th percentile score – that score below which 75% of the sample score? Well, we can see from the above table that it must be between 18 and 19 … as this is where between 74.26% and 79.34% of the sample scores are found. Applying the formula …

Our scores in this case are single values – no range at all. So, our class intervals are in fact the scores themselves. E.g. 0-0, 1-1, 2-2, 3-3 etc. The exact limits however correspond to ±0.5 around each class interval boundary score – the 0, 1, 2, 3, 4, 5, 6 etc. So, our exact limits are:
0 = -0.5 to +0.5
1 = +0.5 to +1.5
2 = +1.5 to +2.5
3 = +2.5 to +3.5
etc.

Let's re-label the table to correspond with our notation in the formula …

	Frequency table: LONG_E (EPQR100M.STA)					
Score	Exact Limits	Midpoint	f	cf Count	Percent	Cumulative Percent
0	-0.5 to 0.5	0	9	9	1.475410	1.4754
1	0.5 to 1.5	1	12	21	1.967213	3.4426
2	1.5 to 2.5	2	13	34	2.131148	5.5738
3	2.5 to 3.5	3	17	51	2.786885	8.3607
4	3.5 to 4.5	4	16	67	2.622951	10.9836
5	4.5 to 5.5	5	12	79	1.967213	12.9508
6	5.5 to 6.5	6	15	94	2.459016	15.4098
7	6.5 to 7.5	7	16	110	2.622951	18.0328
8	7.5 to 8.5	8	22	132	3.606557	21.6393
9	8.5 to 9.5	9	26	158	4.262295	25.9016
10	9.5 to 10.5	10	32	190	5.245902	31.1475
11	10.5 to 11.5	11	31	221	5.081967	36.2295
12	11.5 to 12.5	12	36	257	5.901639	42.1311
13	12.5 to 13.5	13	31	288	5.081967	47.2131
14	13.5 to 14.5	14	29	317	4.754098	51.9672
15	14.5 to 15.5	15	33	350	5.409836	57.3770
16	15.5 to 16.5	16	39	389	6.393443	63.7705
17	16.5 to 17.5	17	35	424	5.737705	69.5082
18	17.5 to 18.5	18	29	453	4.754098	74.2623
19	18.5 to 19.5	19	31	484	5.081967	79.3443
20	19.5 to 20.5	20	34	518	5.573770	84.9180
21	20.5 to 21.5	21	39	557	6.393443	91.3115
22	21.5 to 22.5	22	33	590	5.409836	96.7213
23	22.5 to 23.5	23	20	610	3.278689	100.0000
Missing			0	610	0.000000	100.0000

where

P_{75} = the 75^{th} percentile

ll = 18.5 = the exact lower limit of the interval containing the percentile point

n = 610 = the total number of scores

p = 0.75 = the proportion corresponding to the desired percentile (note this is nothing more than the percentile expressed as a proportion ($75 \div 100$)

cf = 453 = the cumulative frequency of scores below the interval containing the percentile point

f_i = 31 = the frequency of scores in the interval containing the i^{th} percentile point

w = 1.0 = the width of the class interval

feeding these values into the formula we obtain …

$$P_i = ll + \left(\frac{np - cf}{f_i} \right) \cdot w$$

$$P_{75} = 18.5 + \left(\frac{610 \cdot 0.75 - 453}{31} \right) \cdot 1.0$$

$$P_{75} = 18.5 + \left(\frac{457.5 - 453}{31} \right) \cdot 1.0$$

$$P_{75} = 18.5 + 0.1452 \cdot 1.0 = 18.645$$

So, the 75th percentile is a score of 18.645. This is the score at which 75% of observations will be observed to be below this score. BUT – the score is unattainable as this is an integer scored test. What we actually observe is that 74.26% scores will lie at or below 18, with 79.34% of scores at 19 or below. IF we want to use exact percentiles – then we have to accept that our scores are estimates of hypothetical real-valued continuous numbers, hence a score of 18.645 is perfectly valid under these conditions, and the definition of a percentile is most correctly defined as **the value below which P% of the values fall**.

However, what would the 76th percentile look like …

$$P_i = ll + \left(\frac{np - cf}{f_i} \right) \cdot w$$

$$P_{76} = 18.5 + \left(\frac{610 \cdot 0.76 - 453}{31} \right) \cdot 1.0$$

$$P_{76} = 18.5 + \left(\frac{463.6 - 453}{31} \right) \cdot 1.0$$

$$P_{76} = 18.5 + 0.3419 \cdot 1.0 = 18.84$$

note, all figures remain the same except for the proportion – which changes from 0.75 to 0.76.

So, when an individual scores 19 on a test, what do we conclude?

Here we need to compute the **percentile rank** of the score – which is just the reverse of computing the score for a particular percentile. Now we know the score (=19), but need to compute the percentile for it …

The formula is:

$$PR_x = \left[\frac{\left(cf + \left(\frac{x - ll}{w} \right) \cdot f_i \right)}{n} \right] \cdot 100.0$$

where

$PR_x =$ the percentile rank of score x

$ll =$ the exact lower limit of the interval containing the score x

$n =$ the total number of scores

$cf =$ the cumulative frequency of scores below the interval containing the score x

$f_i =$ the frequency of scores in the interval containing x

$w =$ the width of the class interval

So, for a score of 19, the exact percentile rank is:

$$PR_x = \left[\frac{\left(cf + \left(\frac{x - ll}{w} \right) \cdot f_i \right)}{n} \right] \cdot 100.0$$

$$PR_{19} = \left[\frac{\left(453 + \left(\frac{19 - 18.5}{1.0} \right) \cdot 31 \right)}{610} \right] \cdot 100.0$$

$$PR_{19} = \left[\frac{(453 + 0.5 \cdot 31)}{610} \right] \cdot 100.0$$

$$PR_{19} = 76.80\%$$

a score of 19 is at the 76.8th percentile – the score at which 76.80% of scores will be found to be below this score.

<u>BUT</u> ...

All the above is standard fare – and is highly confusing given that only integer value scores can ever be observed. What we know from our observed frequency distribution table is that 79.3443% of individuals scored 19 or below. By invoking the "exact limits" property around each possible score, the values of observed frequencies no longer tally with the hypothetical "score" frequencies. So, the definition ...

A percentile is the point in a distribution **at or below** which a given percentage of scores is found now seems more sensible – when using actual scores vs exact-limit interval scores. But, using actual scores means that only certain % values can be provided – based upon the exact number of frequencies observed for each score. So, there can be no 75th percentile for our observed frequency distribution – only a 74.26th or 79.34th percentile. Whereas, if we invoke the "scores are sampled from a hypothetical distribution of real-valued, continuous scores, we can compute exact percentiles – and must express our scores accordingly:

Observed frequency data percentile for score 19 = 79.34%
Assumed continuous score percentile for score 19 = 76.80%

Conclusions

1. If you want to assign exact percentile ranks to scores, then you must use the formulae above and assume each integer score is actually a point-estimate from an interval of possible scores. Here, the definition of a percentile is the value **below** which P% of the values fall.

2. If you are happy with simply stating the frequency of people who score at or below an observed test score, then you use the actual frequencies of scores in your normative data. Here, the percentile is the point **at or below** which a given percentage of scores is observed.

Hopefully, you can see the contrast that is being made, why, and why the definitions when taken as simple statements seem so contradictory! They are not – because they are based upon different conceptualisations of the frequency distribution of scores. The first assumes that scores are continuous but observed as integer, the second assumes scores are as they are represented – integers, and no more.

Which now begs the question – which assumption is most valid? Well, when you compute a mean score, invariably this will be a real number (say 12.57), yet the scores are integer. Likewise the plethora of statistics computed using conventional quantitative techniques. It is only with the class of statistics known as non-parametric or order-statistics will the integers be preserved as unique entities. All other techniques for numeric manipulation will treat the numbers as continuous (by definition of the arithmetic operations permitted). So, if you want to remain consistent with probably almost every other way you treat test scores, then adopt conclusion #1. If you are a purist – and do not want to treat your scores as anything but ordinal rank values (integers), then adopt #2. Your definition of percentile will change accordingly to suit.