Calculating Percentiles

[Ian Robertson, January 09, 2004]

Percentiles are very handy for exploring the *distribution* of number sets using various EDA graphs, including the well-known (and still underused) boxplot.

The meaning of *percentile* can be captured by stating that the *p*th percentile of a distribution is a number such that approximately *p* percent (*p*%) of the values in the distribution are equal to or less than that number. So, if ‘28’ is the 80th percentile of a larger batch of numbers, 80% of those numbers are less than or equal to 28.

A percentile can be (1) calculated directly for values that actually exist in the distribution, or (2) interpolated for values that don’t exist (but which you may want to use to plot specific kinds of graphs, for example).

To calculate percentiles, sort the data so that x_1 is the smallest value, and x_n is the largest, with n = total number of observations.

x_i is the p_ith percentile of the data set where:

$$p_i = 100\frac{i - 0.5}{n} \qquad [1]$$

For example:

(original data)

5	1	9	3	14	9	7

(sorted data)

x_i	1	3	5	7	9	9	14
I	1	2	3	4	5	6	7
p_i	(calculate, using equation [1], as shown below…)						

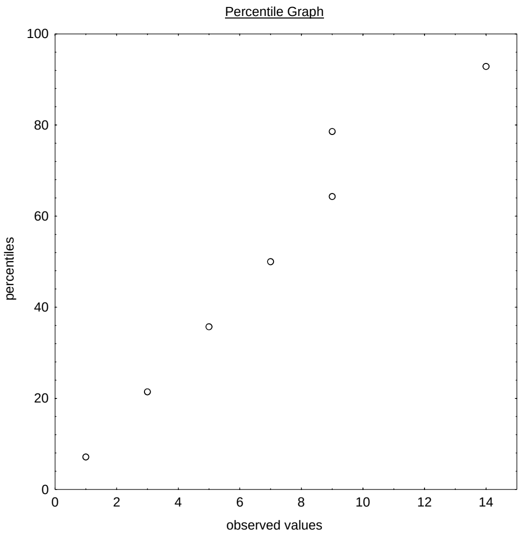
p_1 = 100(1 - 0.5) / 7 = 7.1
p_2 = 100(2 - 0.5) / 7 = 21.4
p_3 = 100(3 - 0.5) / 7 = 35.7
p_4 = 100(4 - 0.5) / 7 = 50
etc…

(filling in the final row, we get)

x_i	1	3	5	7	9	9	14
i	1	2	3	4	5	6	7
p_i	7.1	21.4	35.7	50.0	64.3	78.6	92.9

To take a single example, 7 is the 50th percentile of the distribution, and about half of the values in the distribution are equal to or less than 7. In EDA jargon, 50 is sometimes referred to as the *p-value* of 7.

Among many other things (such as constructing boxplots) percentiles can be plotted against observed values to produce a ‘percentile plot,’ a graphic and often highly useful depiction of the distribution of the data. Here is a percentile plot based on the previous example:



If we want to calculate a specific percentile (25%, 50%, etc.) we have to turn equation [1] around:

$$i = \frac{np_i}{100} + 0.5 \qquad [2]$$

If i is an integer, x_i is simply the pth percentile. (You can easily confirm this by calculating the 50th percentile, which matches perfectly the observed x_4 value, 7). If i is not an integer, we can interpolate as follows:

let k = the integer part of i
let f = the fractional part of i
(i.e., if k = 10.375, then k = 10 and f = 0.375)
let x_{int} = the value we want to interpolate between x_k and x_{k+1}:

$$x_{int} = (1-f)x_k + fx_{k+1} \qquad [3]$$

ex: to find the 25th percentile:

i = (7*25)/100+0.5 (from equation [2])
i = 2.25

k = 2; f = 0.25

x_{int} = (1-0.25)*3 + 0.25*5 (from equation [3])
x_{int} = 3.5

[Note: I describe the outcome of equation [3] as x_{int} rather than p. I think that using p is more common in EDA circles, but I find it needlessly confusing. I hope that using x_{int} (the subscripted ‘*int*’ is meant to make you think *interpolation*) adds clarity…]