Announcement of methodology change in "Provisional Accident & Emergency Quality Indicators for England. Experimental Statistics by provider"

1 Background

From July 2012 the Health and Social Care Information Centre have taken responsibility for the production of the Provisional Accident & Emergency Quality Indicators for England which were previously supplied by the Department of Health. The source data and definition of indicators remain unchanged but due to a change to a specialist data handling and statistical software package (SAS) the methodology for calculating percentiles within the publication has changed slightly. This document describes this change.

2 Percentiles

A percentile is a value in a set of values below which a certain percent of observations fall. For example, the 20th percentile is the value which 20 percent of the observations are less than or equal to.
Percentiles, when used in conjunction with sample size and averages, are a useful way to describe the distribution of the data without having to explore every value.
So, a dataset with 1,000 records may have a mean value of 50 and a max of 150, knowing that the 50th percentile (or median) is 55 and that the 95th percentile is 107 tells us a little more.
i.e. 55 is the value of the middle number when all are sorted in order, and that 95% of all the values are less than or equal to 107.

3 Interpolation

The percentile in a distribution does not always correspond to a value in the distribution. For example, consider the list of values: 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, and 10.
There are 35 values, so we cannot simply select the value at the 95th percentile, instead we must calculate the percentile value by interpolating between the discrete values in the distribution.

There is no universally agreed method of doing this, and different software offer different methods.

By default Microsoft Excel uses the formula $1 + (n-1)p = j + g$
Where n = the number of observations, p = the percentile expressed as a fraction, j = integer part of the result, and g = the decimal part.

$1 + (35 – 1) * 0.95 = 33.3$
i.e. $j = 33$ and $g = 0.3$

The percentile is then calculated as
$y = (1 – g)x_j + gx_{j+1}$

where x_j is the value in the distribution at the j-th position, so where j=33 x_j in our list of values is 8.

y = (1-0.3)*8 + 0.3 * 9
y = 8.3

Microsoft Excel calculates the 95th percentile to be 8.3

SAS (using the selected calculation method) uses the same calculation for the percentile but has a different starting point in that j and g are calculated differently.

SAS option 4 uses the formula (n+1)p = j + g

(35 + 1) * 0.95 = 34.2
i.e j = 34 and g = 0.2

using $y = (1 – g)x_j + gx_{j+1}$ as before.

y = (1 – 0.2)*9 + 0.2 * 10
y = 9.2

SAS calculates the 95th percentile to be 9.2

These two methods can be described as inclusive (Excel) of the 0th and 100th percentile, or exclusive (SAS). Excel 2010 introduced the option to use the exclusive method; however, the Department of Health used Microsoft Excel 2002 in order to generate the percentile values in the publication. Therefore slightly different percentile values are generated even though the input data is the same. This is particularly true where there are a small number of records being evaluated, and can have a marked effect on the 95th percentile where the largest and second largest values are very different.

Percentiles derived from distributions with fewer than 100 members have been highlighted in italics to note the fact that they're particularly likely to have been affected by this change of calculation method.