Werner Wosniok

University of Bremen
Institute of Statistics
Am Fallturm 1
D 28334 Bremen, Germany
wwosniok@math.uni-bremen.de

TECHNICAL NOTE

On the treatment of values below the limit of detection

When calculating statistics from data sets that contain values below the limit of detection (LD), there are essentially two ways to handle these:

1. using a calculation method for truncated data (the elegant and most appropriate, but sophisticated way), or
2. replacing all values below the LD by a surrogate value and using these as if they were regular observations (the simple, though less accurate way).

The first way is particularly advisable if the proportion of data below the LD is large (40% or more). For smaller proportions the second way is accurately enough in most cases. In the paper by Kammann *et al*. (2005) on EROD activity in dab, the second way was used because the number of values below the LD was sufficiently small.

The rationale for the choice of a surrogate value for continuous quantities like EROD and many other, though not all, biomarkers is as follows:

We assume that the distribution of the measured quantity between zero and the LD can reasonably well be approximated by a triangular distribution with density = 0 at values ≤ 0. This is the case for standard distributions like log normal, log logistic, Weibull distributions and many more, and also for mixtures of these. Also the EROD data in Kammann *et al.* (2005) has a shape of this kind. Fig. 1 shows an example for such an approximation.

Under this assumption, the expected value (mean value) of measurements in the range below the LD is (2/3) * LD, as the following calculation shows:

Let x denote the measured quantity and f(x) its density function. Then the general definition of the expected value for measurements between 0 and LD is

$$E(X) = \int_0^{LD} x \cdot f(x)\, dx \qquad [1]$$

This equation could be solved exactly, if necessary by numerical methods, for a given density function f(x). However, in many cases, the part of f(x) below x = LD can well be approximated by a triangular distribution with end points (0, 0) and (LD, f(LD)). This triangular distribution has the density

$$f_T(x) = \frac{2}{LD^2} x \qquad \text{for } 0 \le x \le \text{LD, and zero otherwise.} \qquad [2].$$

Replacing f(x) in [1] by $f_T(x)$ leads to

$$E(X) = \int_0^{LD} x \cdot \frac{2}{LD^2} \cdot x\, dx = \frac{2}{LD^2} \int_0^{LD} x^2\, dx\,, \qquad [3]$$

which resolves to

$$E(X) = \frac{2}{LD^2} \left[\frac{x^3}{3} \right]_0^{LD} = \frac{2}{LD^2} \frac{LD^3}{3} = \frac{2}{3} LD\,, \qquad [4]$$

which completes the proof. Hence, under the assumptions from above the expected value for measurements below the LD is (2/3) * LD. This recommends (2/3) * LD as a surrogate value.

The frequently seen recommendation to use (½)*LD as a surrogate value is justified if f(x) is symmetric around x = LD/2. The most interesting special case of this condition is f(x) = constant > 0 for x ≤ LD. Fig. 2 shows this

(here inappropriate) kind of approximation for the same curve that was used in Fig. 1. Each positive constant leads to the same surrogate value of (½)*LD. But for a data distribution as in Fig. 1 an approximation of f(x) by a constant for x < LD is always worse than the approximation by a triangular distribution, no matter which constant is chosen. f(x) being constant below LD implies that values of x = 0 occur with a certain probability larger than zero. This may be the case for, e.g., pollutants which may have a concentration equal or close to zero in some samples. It is certainly not reasonable for a quantity like EROD and other biomarkers, which must be larger than zero to keep the organism alive.

In general, the choice of a reasonable surrogate value depends on the characteristics of the quantity under study. A situation in which neither (1/2)LD nor (2/3)LD are reasonable choices can easily be imagined: consider the case that the quantity measured may be either zero (non-responder subjects) or > zero (responders). If there is a substantial number of non-responders, neither the triangular nor the uniform distribution describe the data distribution for x < LD with sufficient accuracy, and a reasonable surrogate value is likely to depend on the proportion of responders (and to differ from (1/2L)D and (2/3)LD).

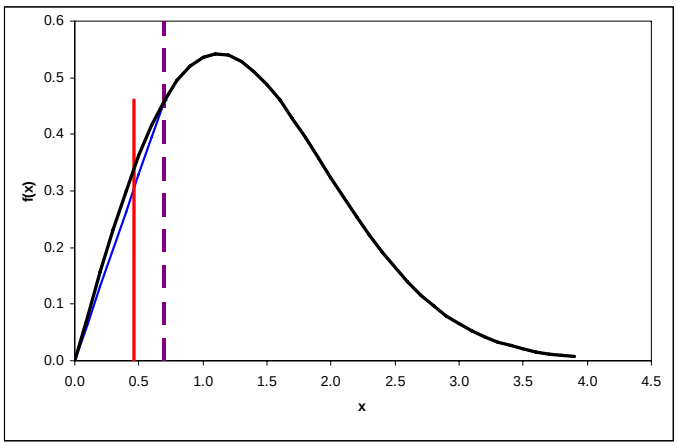

Fig. 1: Approximation of a density f(x) (black) for x < LD (dashed violet) by a triangular density (blue) and the resulting surrogate value (red) of 2/3 LD.

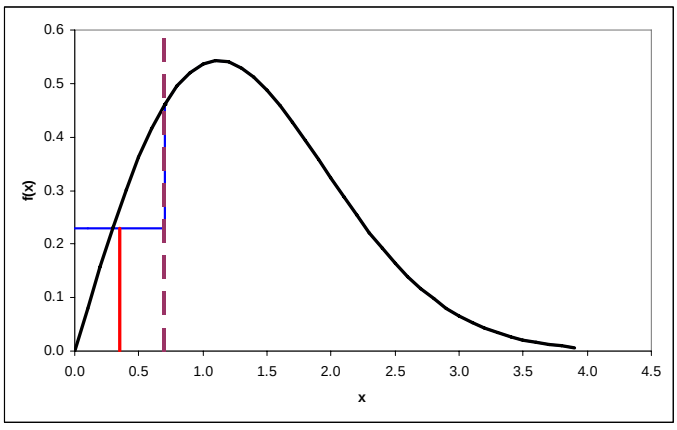

Fig.2: Approximation of a density f(x) (black) for x < LD (dashed violet) by a uniform (=constant) density (blue) and the resulting surrogate value (red) of ½ LD.

Reference

U. Kammann, T. Lang, M. Vobach, W. Wosniok (2005): Ethoxyresorufin-O-deethylase (EROD) activity in dab (Limanda limanda) as biomarker for marine monitoring. Env Sci & Pollut Res 12(3) 140-145.

Last modified: April 18, 2007