Method for absolute status based on CDF scoring functions

Elgin Perry

4/15/2010

Presented to TMAW 10/07/2010
reviewed 5/4/2011

Idea is like grading on a curve.

Curve could based on remainder of population (relative status)

Curve could based on ideal population (absolute status)

Based on the mathematical methods of
CBP - relative status used from 1997-200?

CBP relative status

- **reference = 1st 6 years of CBP data**
- **For each parameter a CDF was estimated for the reference.**
- **Log-Logistic CDF was used**
- **Data were transformed to a quantile scale (0,1) using CDF**
- **Assessment endpoint was 3-year median of quantile observations**
- **Medians were placed in 1/3 categories based on Beta distribution**

(note- these 1/3 categories were sample size dependent)

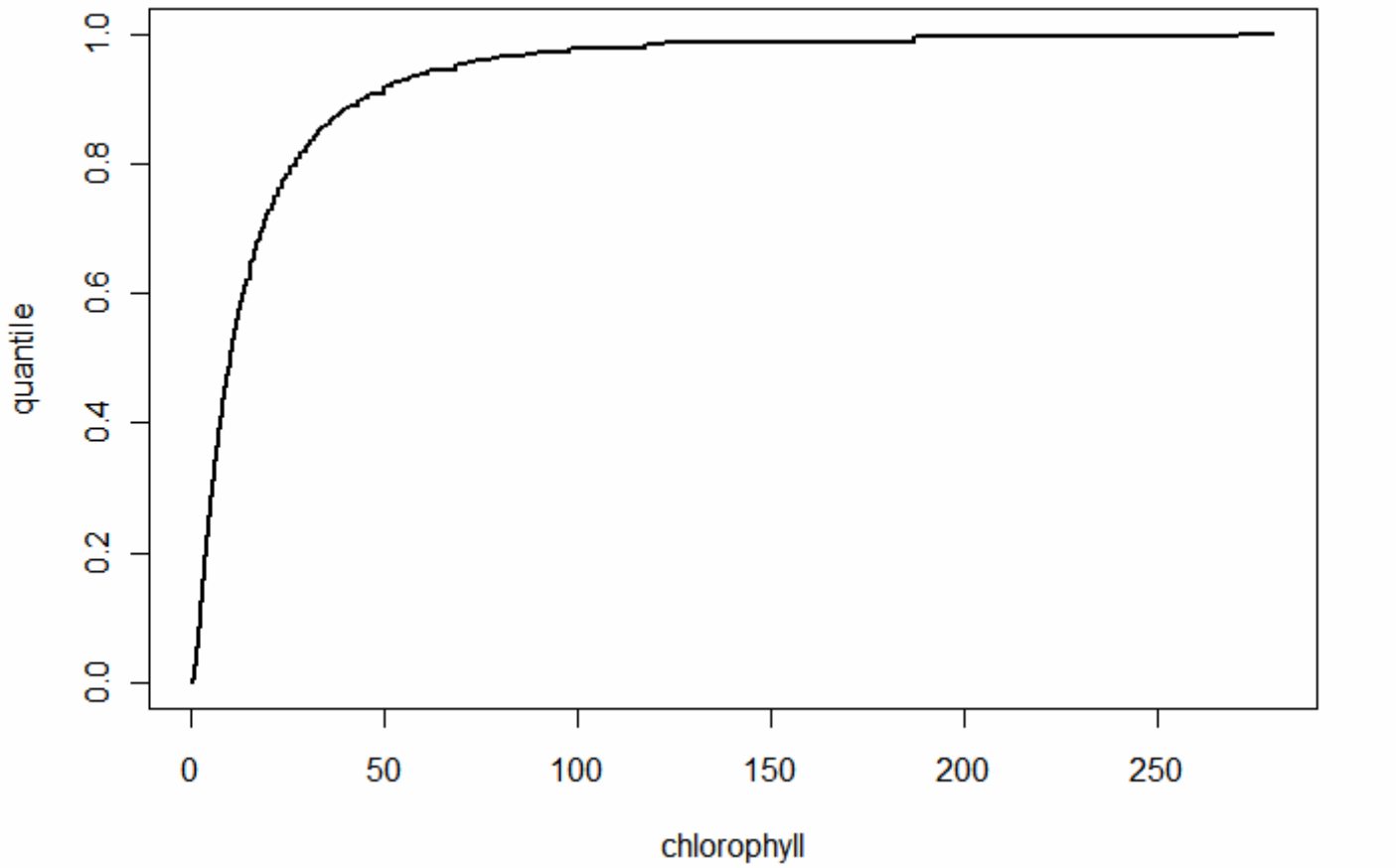


Figure 1. ECDF for reference chlorophyll.

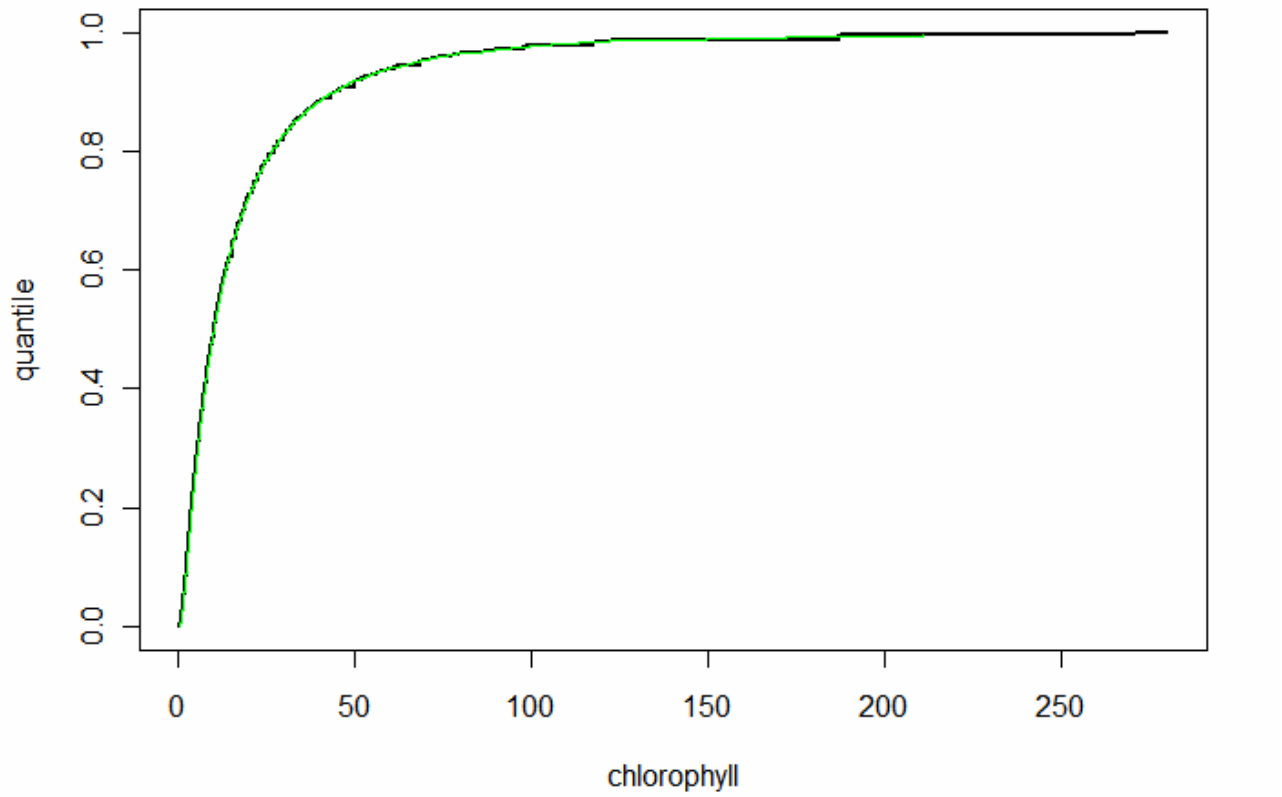


Figure 2. Adding the estimated CDF scoring function (green).

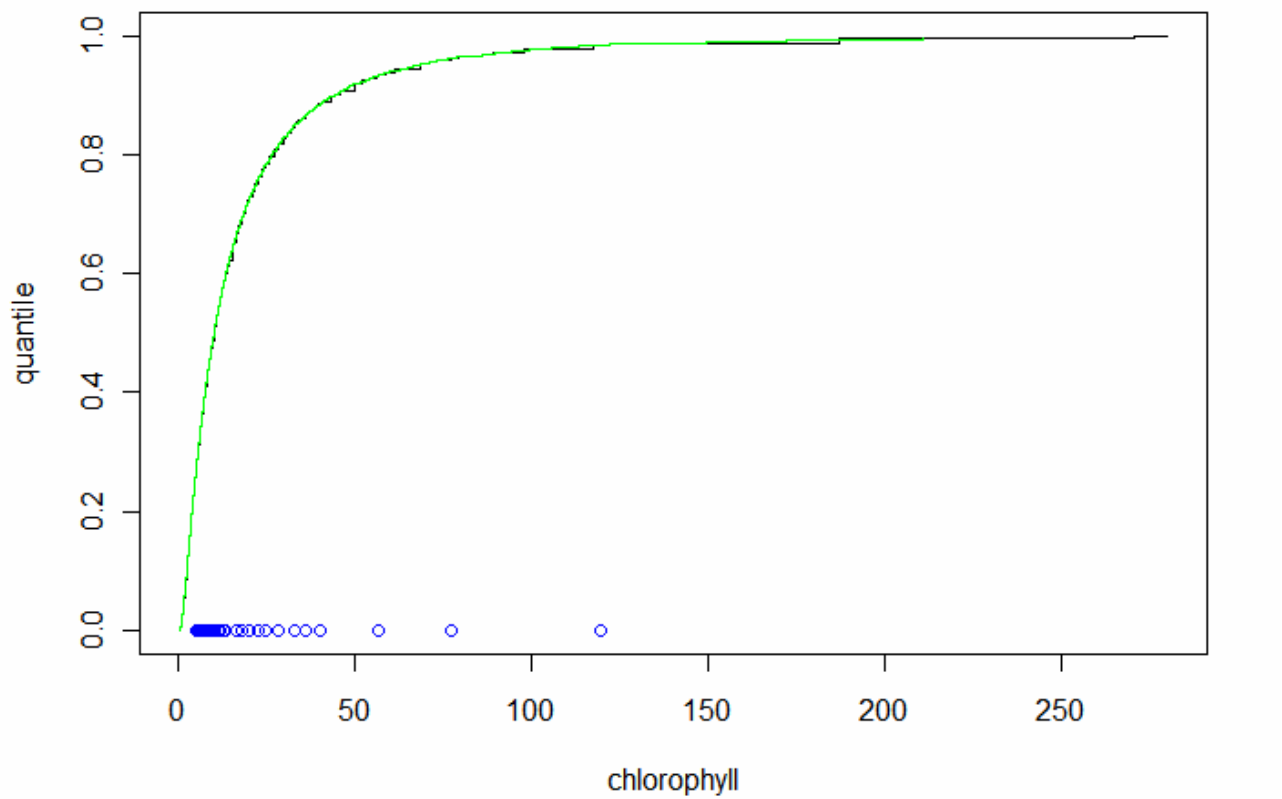


Figure 3. Adding 36 assessment data points (blue)

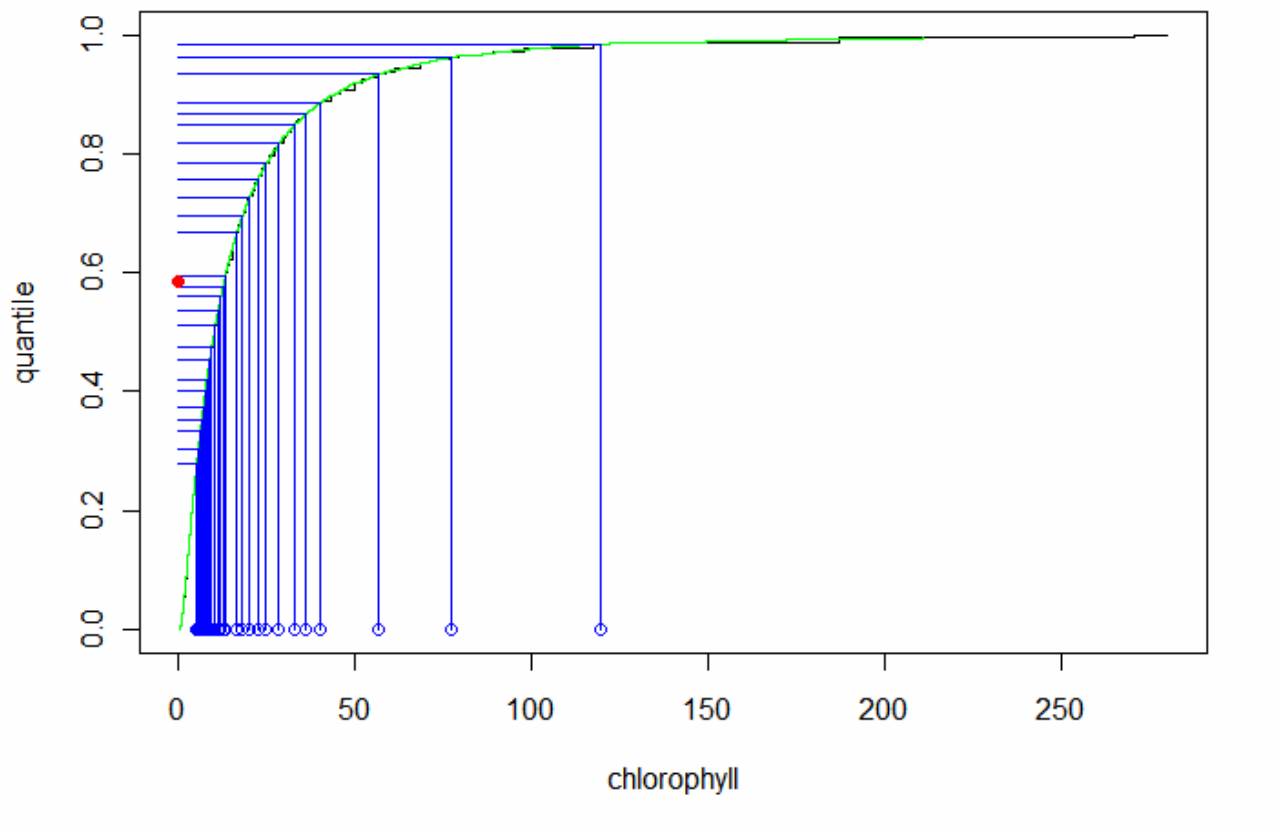


Figure 4. Using green CDF to score the assessment data.

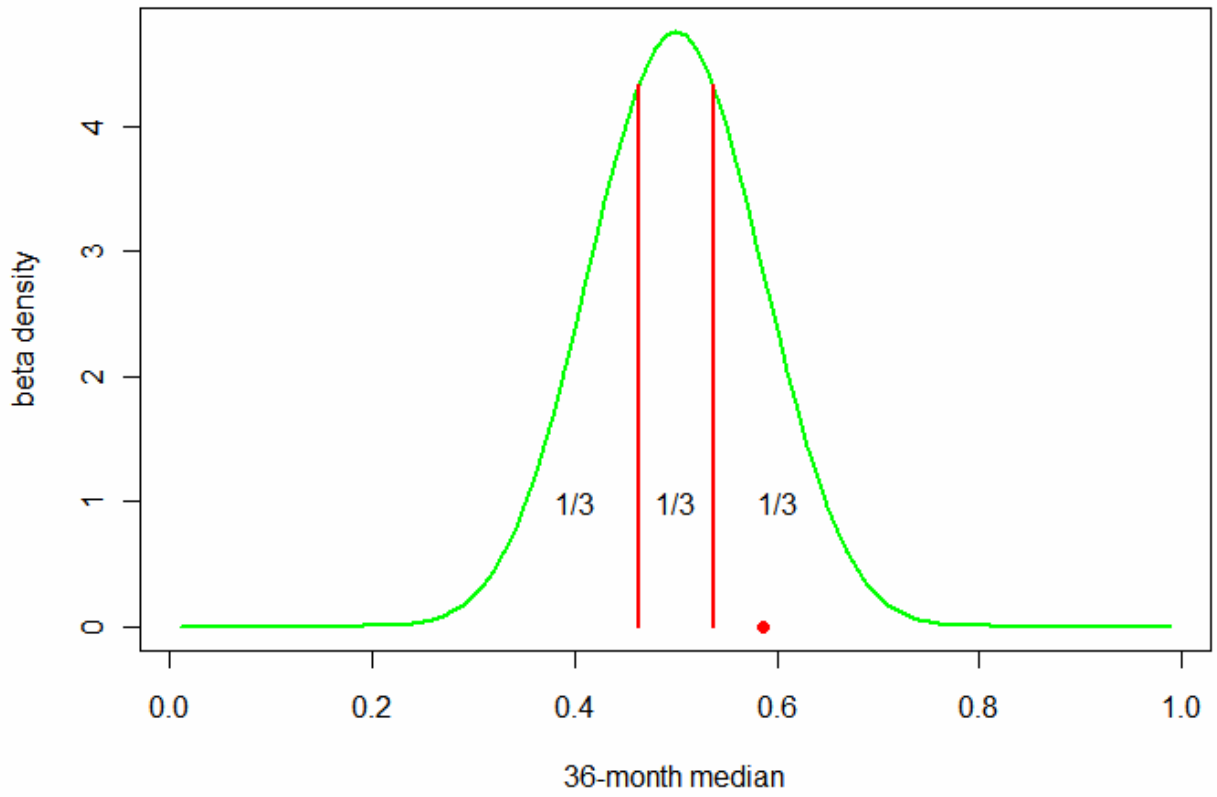


Figure 5. Beta density (B,18,18) showing 1/3 cutpoints for status categories. Red dot is median from previous figure.

To change this method to absolute status, just need to change the reference distribution:

- reference data,
- the water quality model,
- values taken from literature review,
- best professional judgment,
- shifted distribution

minimum requirements are central tendency and measure of variance.

Chlorophyll Example

1. **Reference = "Better-Best" tidal fresh - summer data (Buchanan et al. 2005)**
2. **Use log-normal CDF / scoring function**
4. **transform data to quantile scale (0-1)**
5. **compute summary statistics**
6. **implement cutpoints pass/fail**
7. **statistical inference**

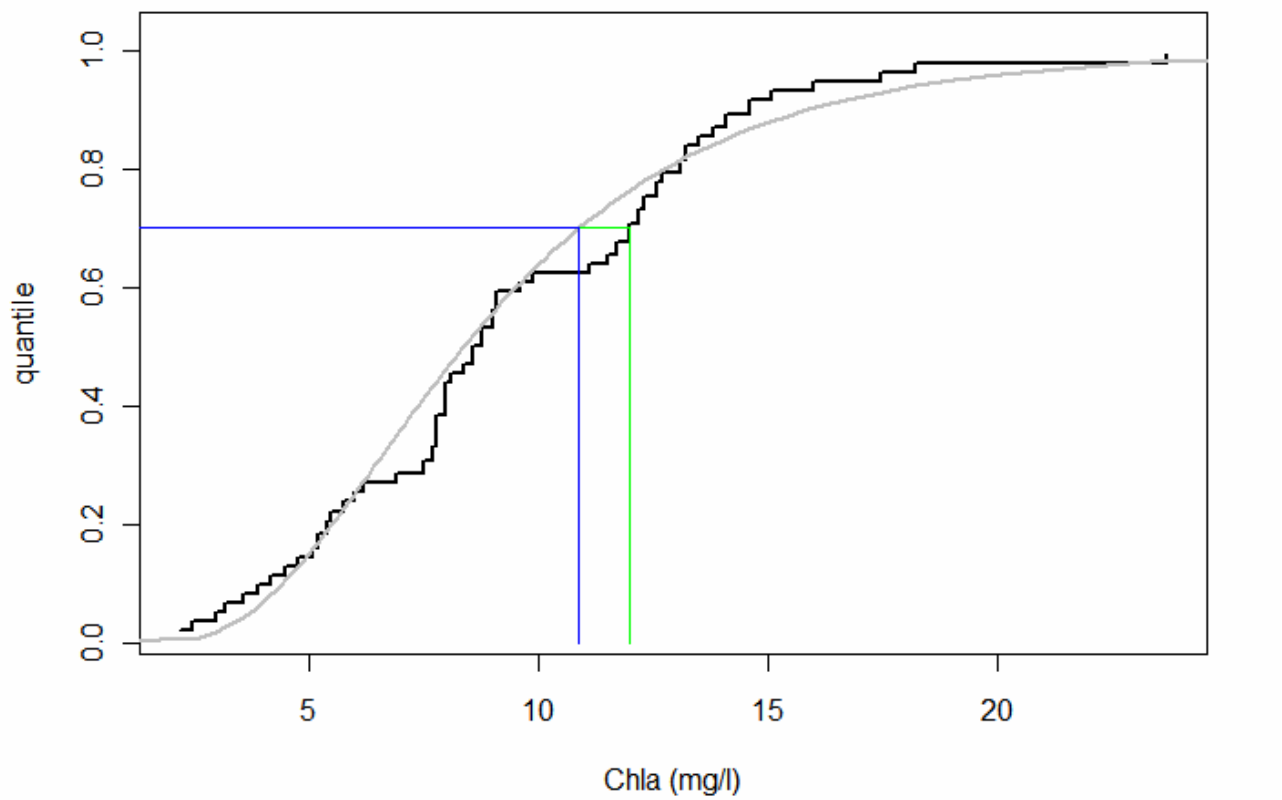


Figure 6. ECDF of reference data (step function), estimated CDF (grey curve), 70th percentile of data (green), 70th percentile estimated from CDF.

Scoring Function

Normal CDF

$$F(x;\mu,\sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz$$

Logistic CDF

$$F(x;\alpha,\beta) = \frac{1}{(1+e^{-(x-\alpha)/\beta})^2}$$

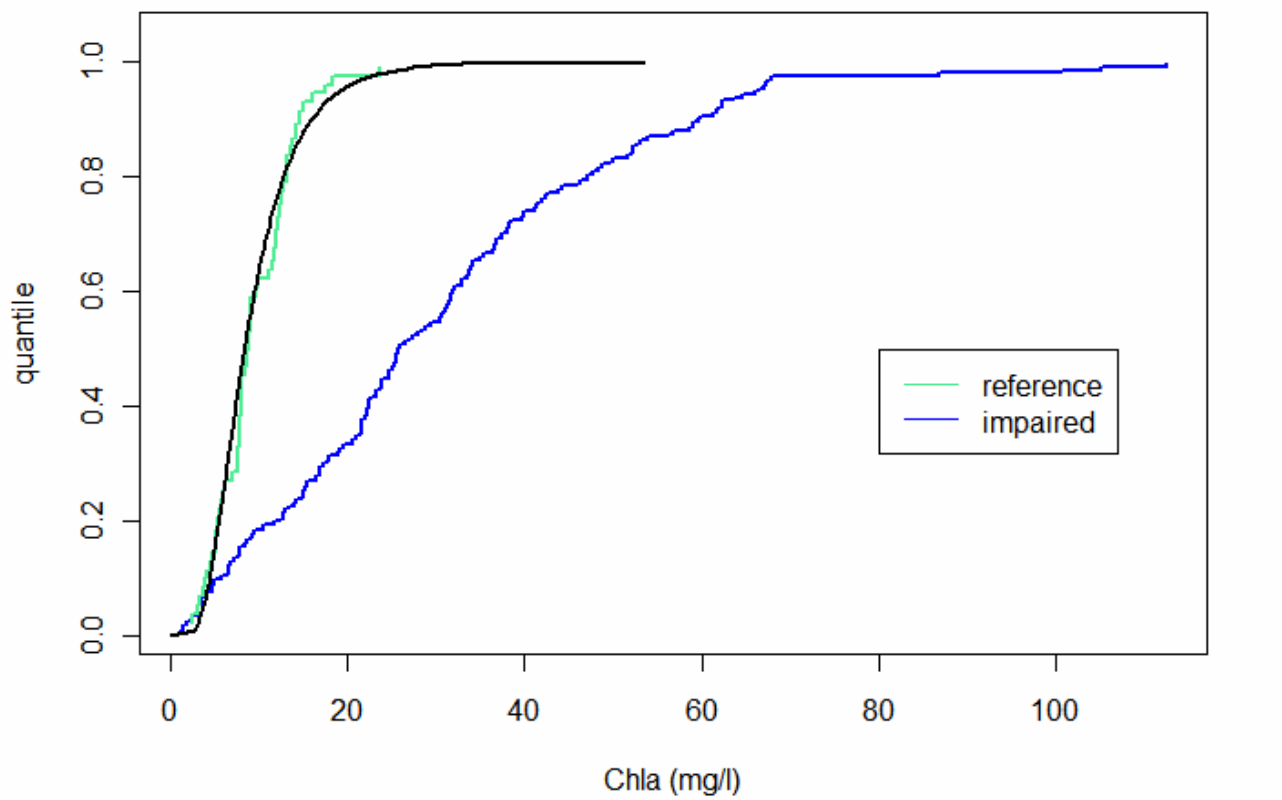


Figure 7. The empirical distribution function of chlorophyll a from 'good' regions (green step function) with an overlay of the log-normal scoring function (smooth black curve) and the ECDF of chlorophyll a from 'bad' regions (blue step function).

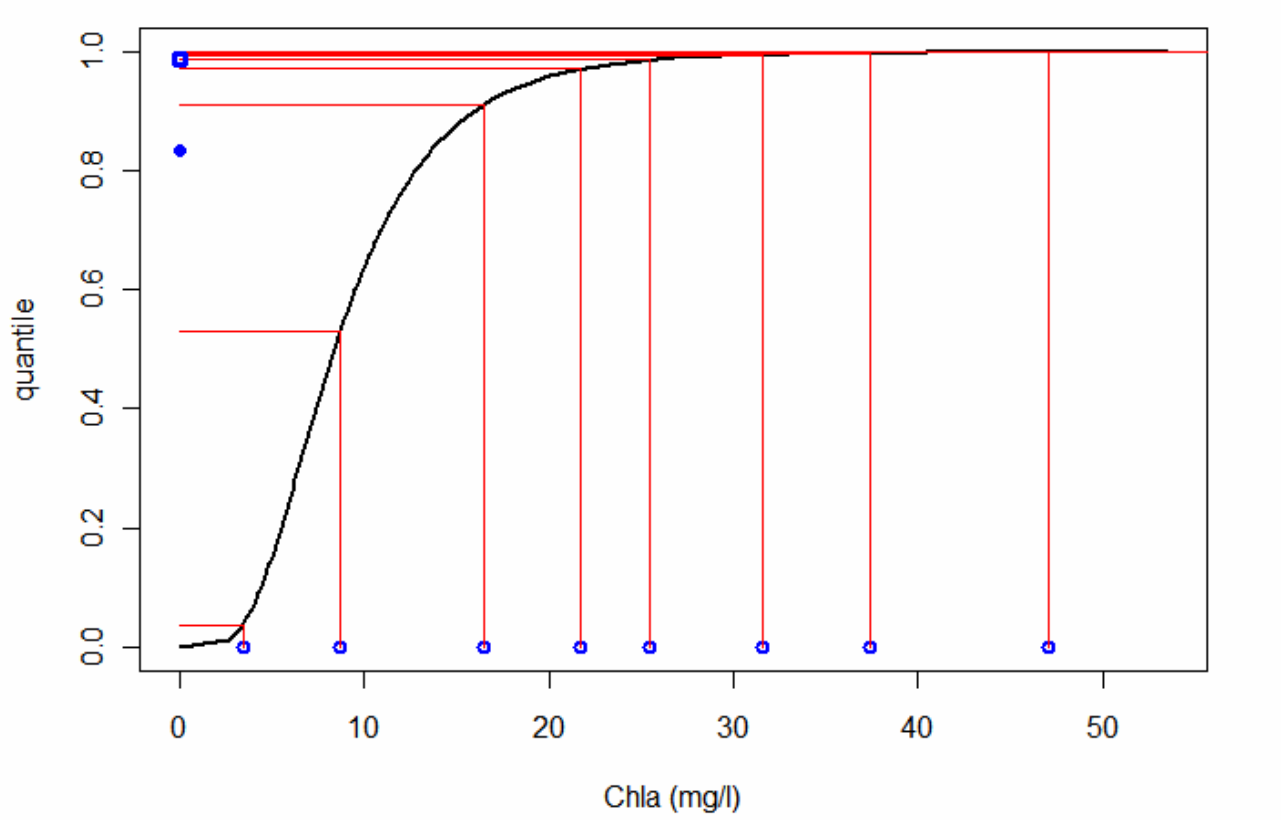


Figure 8. The estimated CDF or the log-normal scoring function (smooth black curve) based on reference with a subset of chlorophyll observations from 'bad' conditions (blue circles). The red lines show quantile scoring of 9 observations.

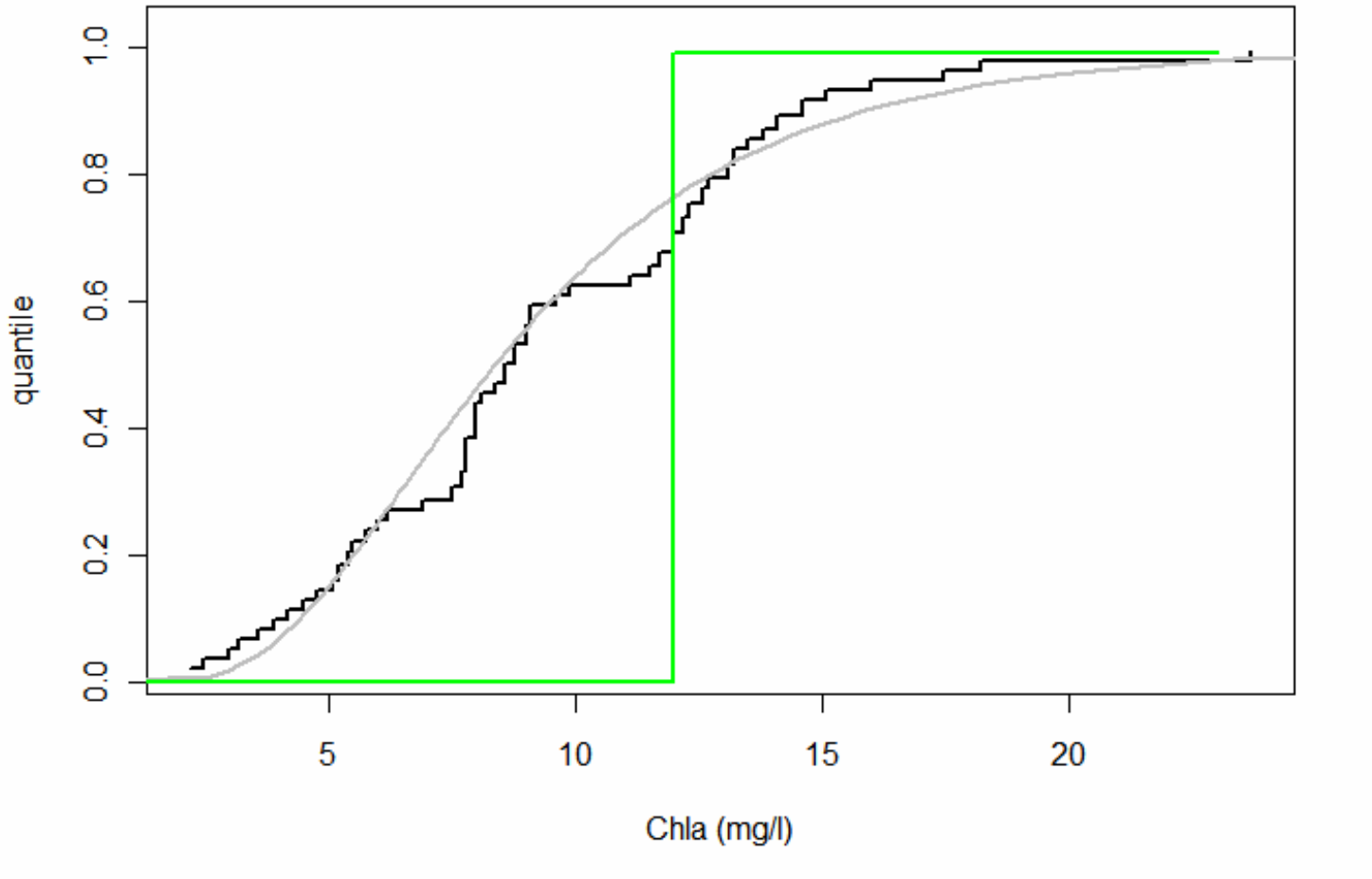
Square = mean, dot = median.

Assessment Endpoints and inference:

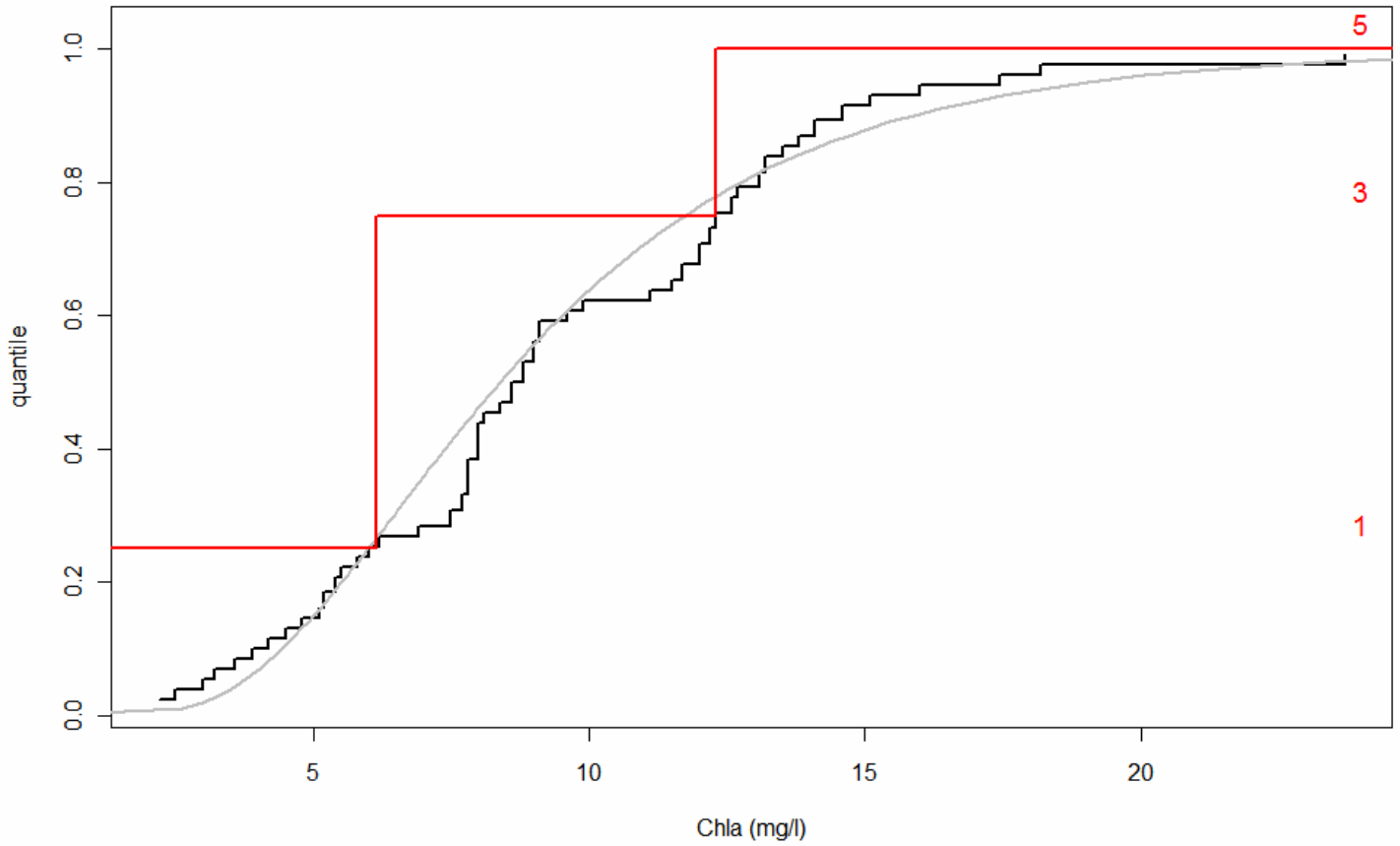
- median quantile score / beta distribution
- mean quantile score / ? distribution
- distribution tests, e.g. Kolmogorov-Smirnov, Shapiro-Wilks
- bright line criteria with scientific basis

Comparison to other approaches:

Single number criteria:

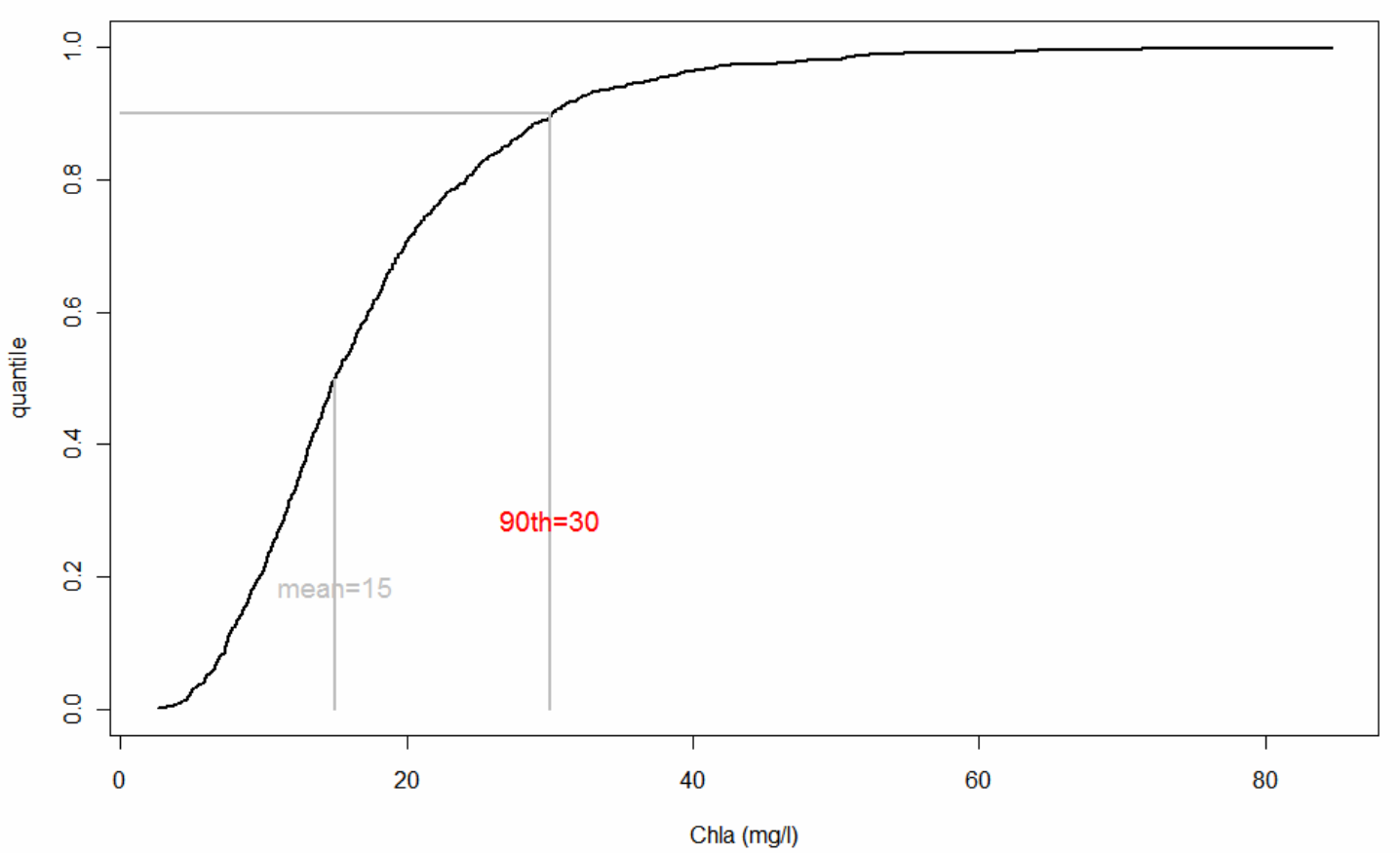


IBI scoring approach:

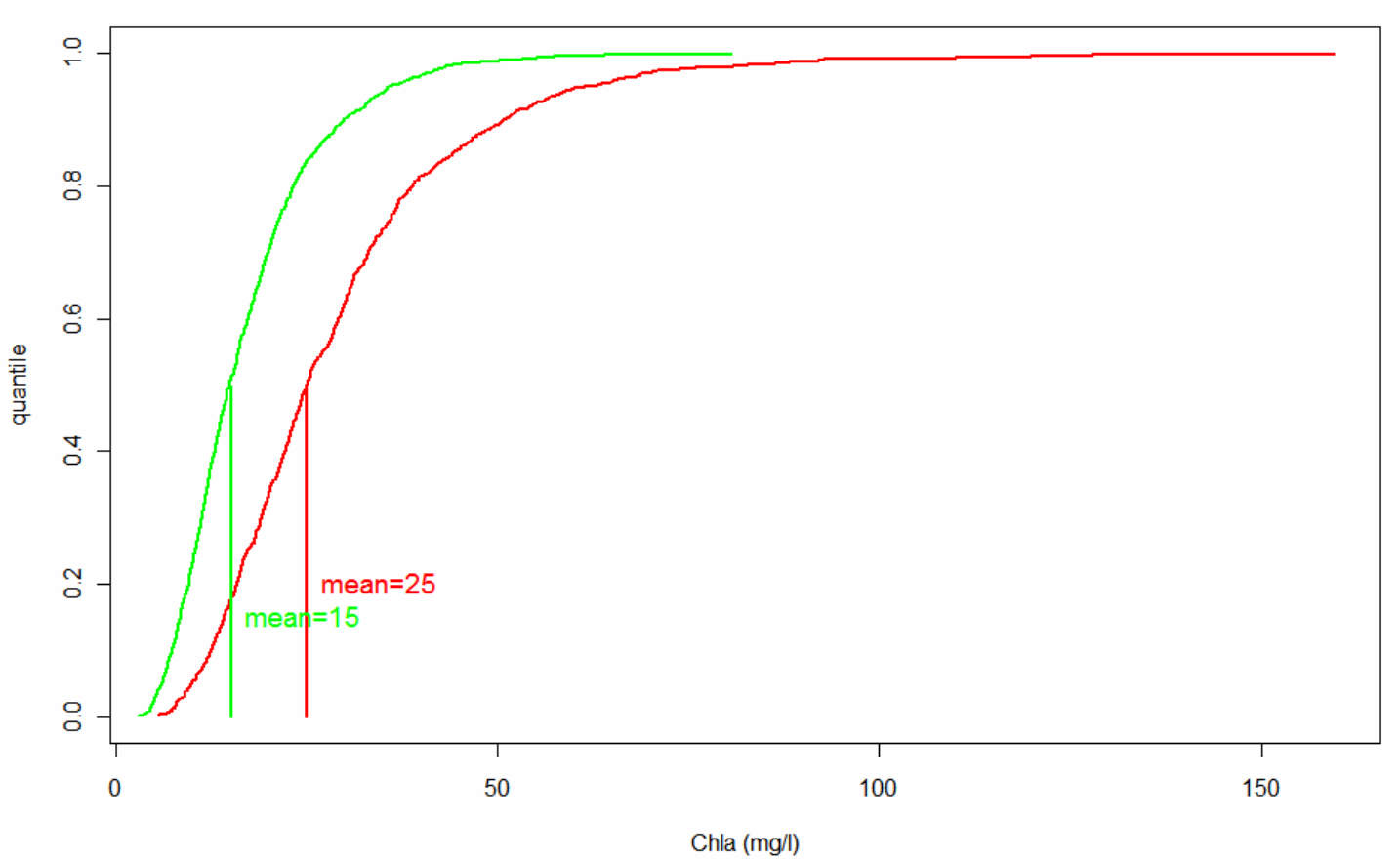


Examples of specifying Reference Distribution:

Mean and upper bound: mean = 15, exceed 30 not more than 10%



Distribution Shift:


quantile
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Chla (mg/l)
mean=25
mean=15

Advantages:

- **preserves continuity and ranking of original data**
- **scores data to universal 0-1 scale to facilitate averaging over dissimilar scales**
- **scored data have uniform 0-1 statistical distribution under probability integral transform**
- **medians of scored data follow beta distribution**
- **easy to explain through grading on a curve analogy**
- **minimal requirements for specifying reference distribution**
- **has the potential to accommodate context sensitive criteria**

Outstanding issues:

- Choice of scoring function
- Nonparametric density scoring
- Risk if density if mis-specified
- space - time issues
- potential for context sensitive status assessment