Yet Another (but Better) Substitution Method for Datasets having Less-than-Detected Values

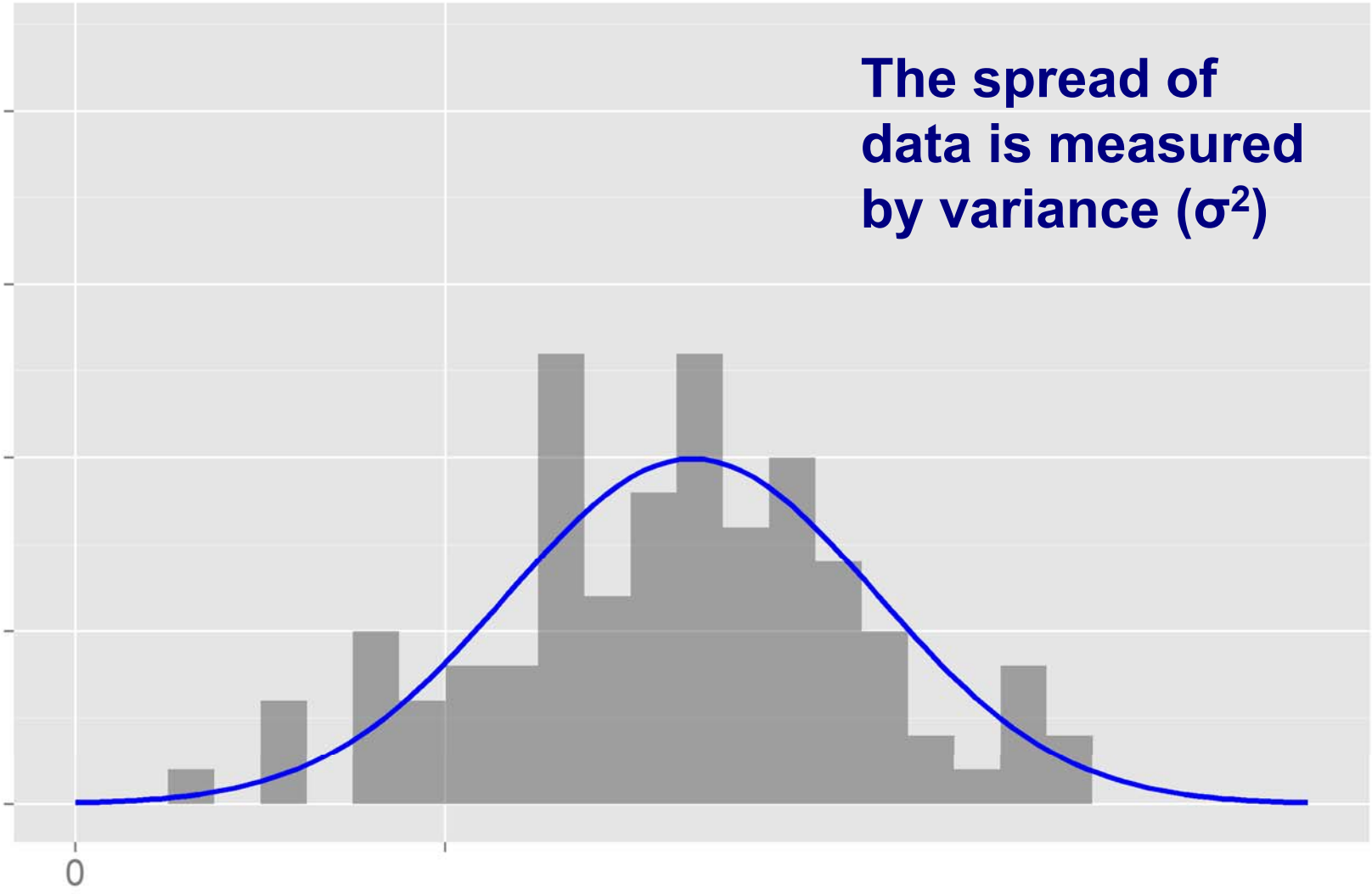
John Warren, ***U.S. EPA***
Kelly Black, ***Neptune & Co.***
Stephanie Fitchett, ***Neptune & Co.***
Adam Sayre, ***Neptune & Co.***

NEPTUNE AND COMPANY

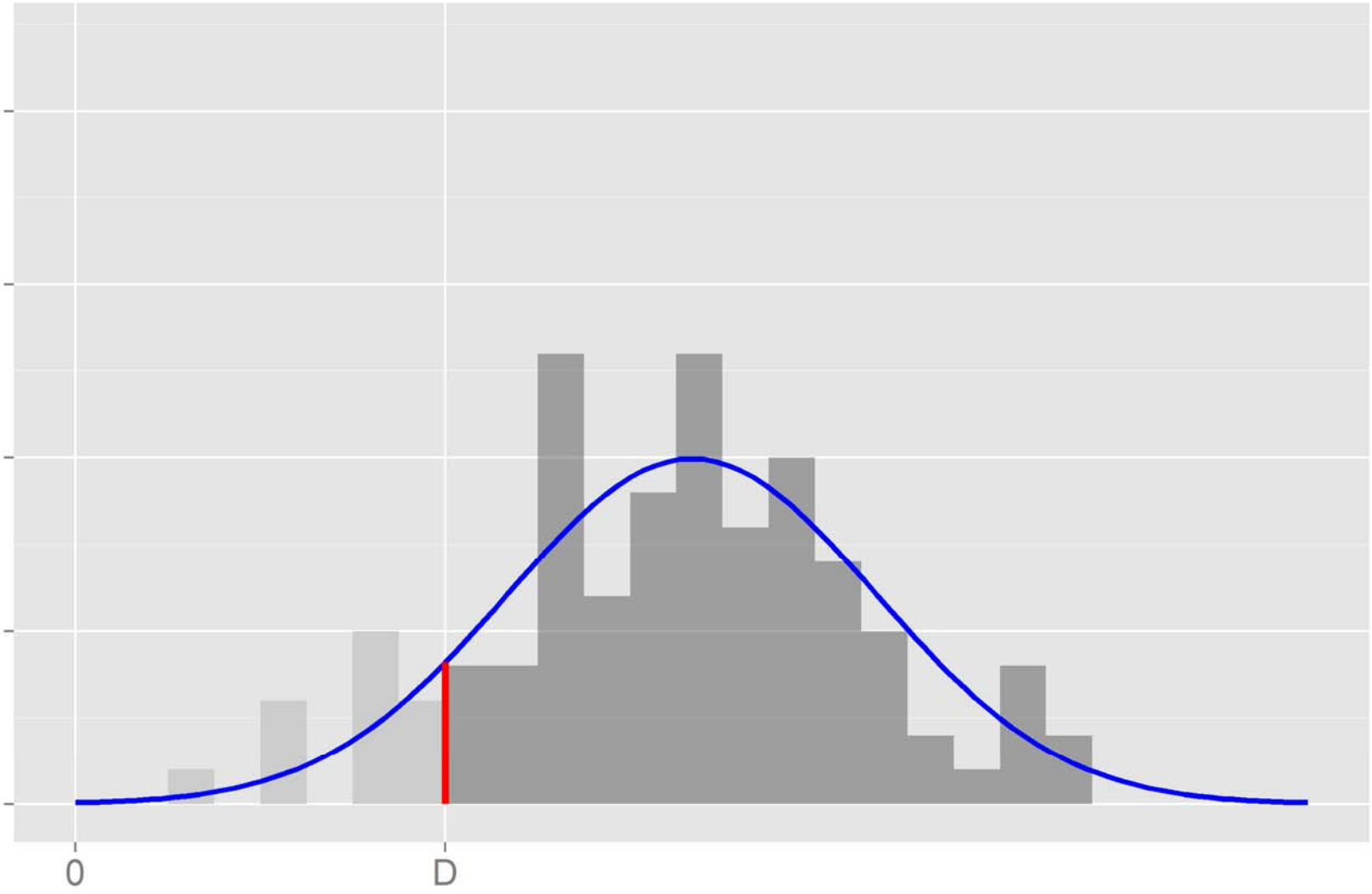
This Presentation

- Background
- The new method
- Simulation results
- Recommendations

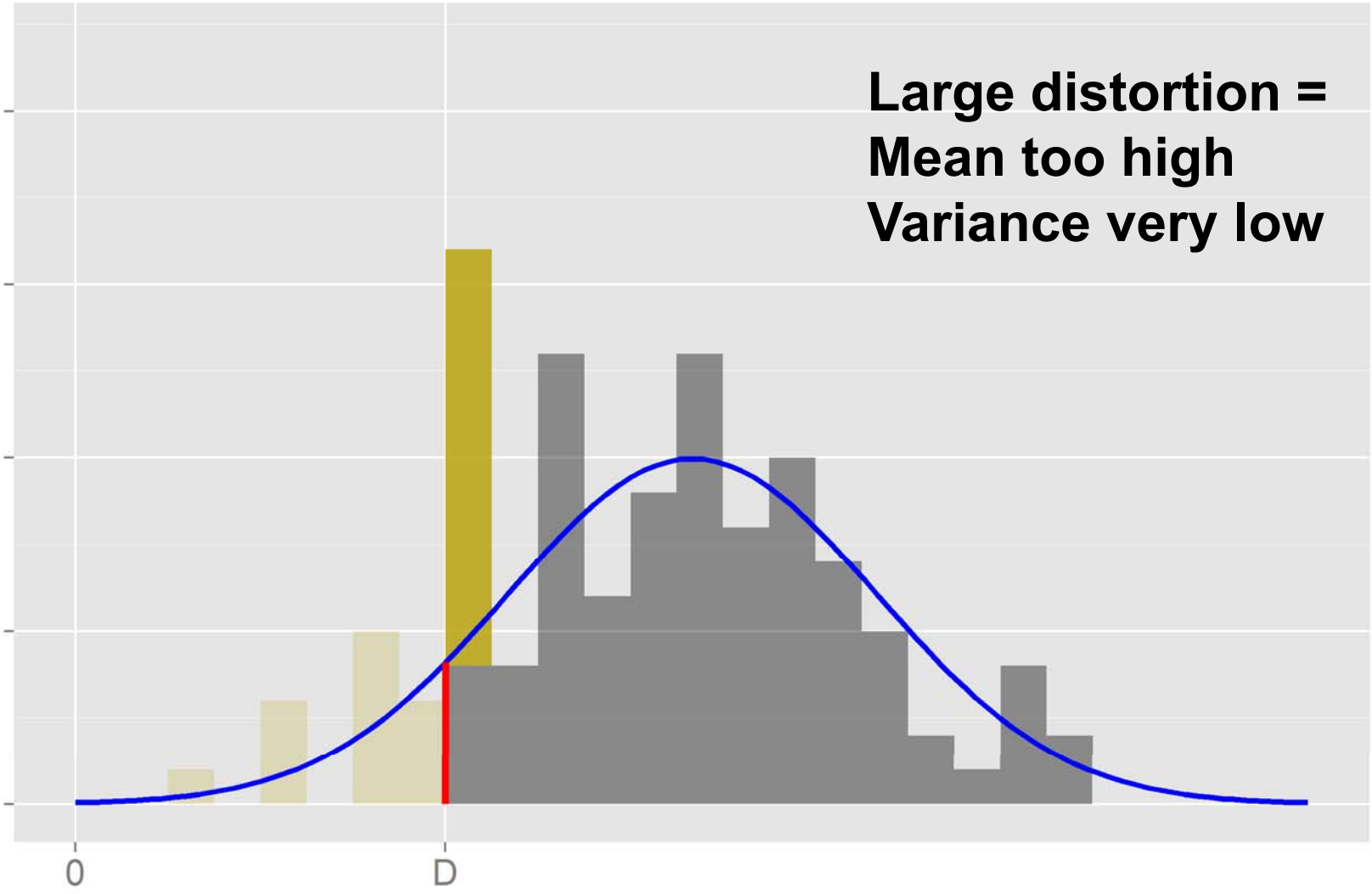
If we had perfect information



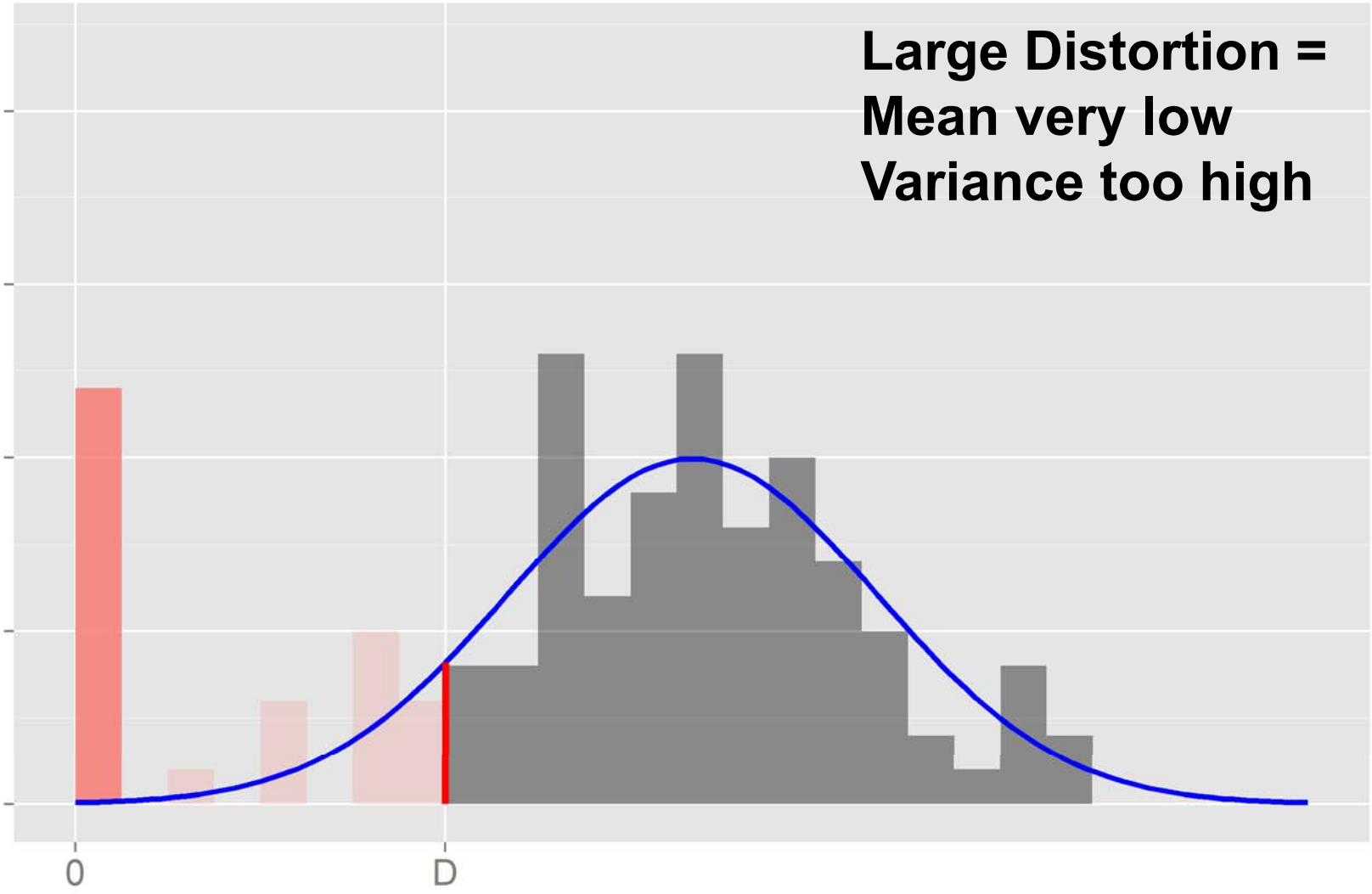
but we can’t see values below detection (D)



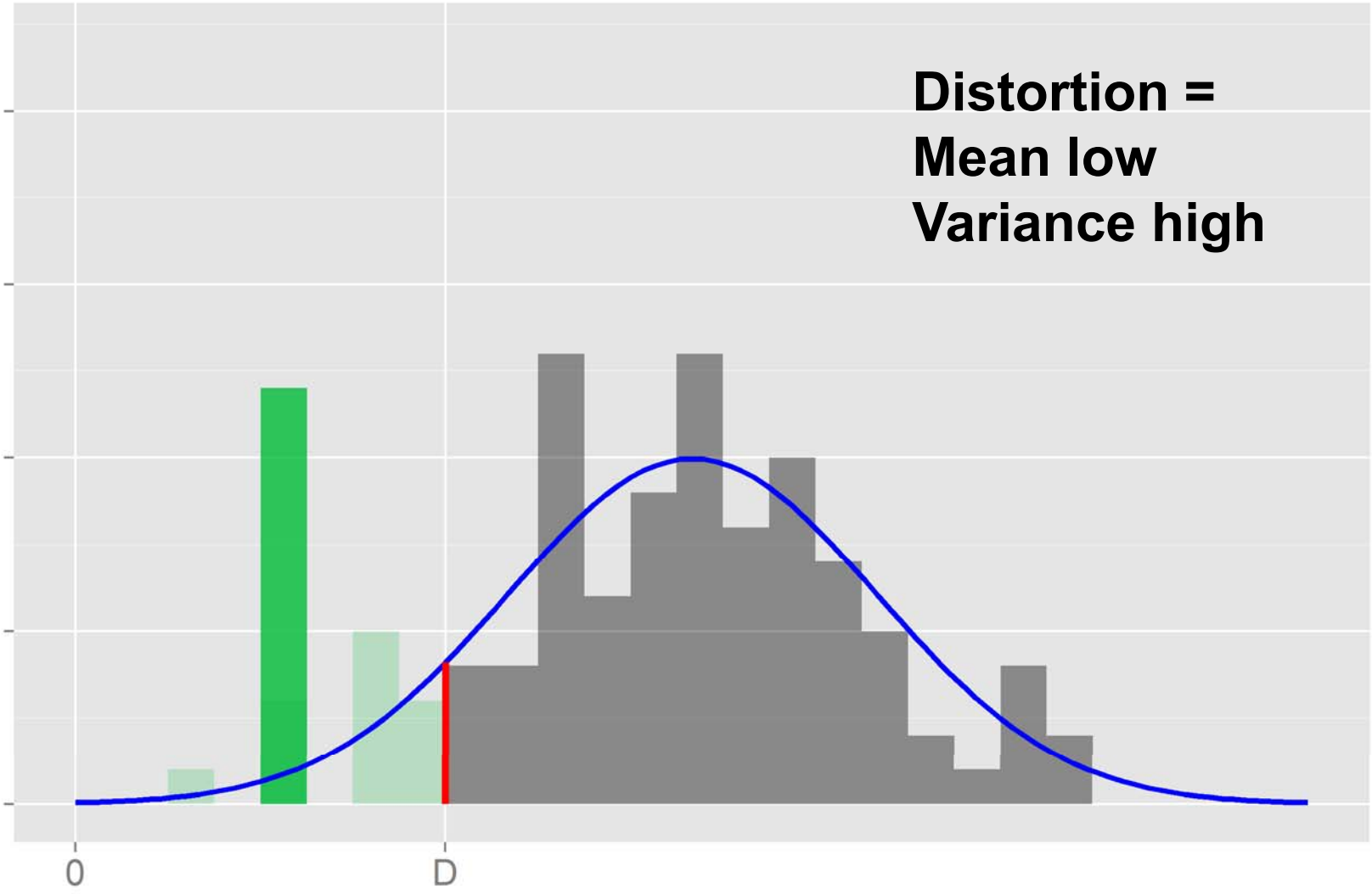
Replace all non-detects by D



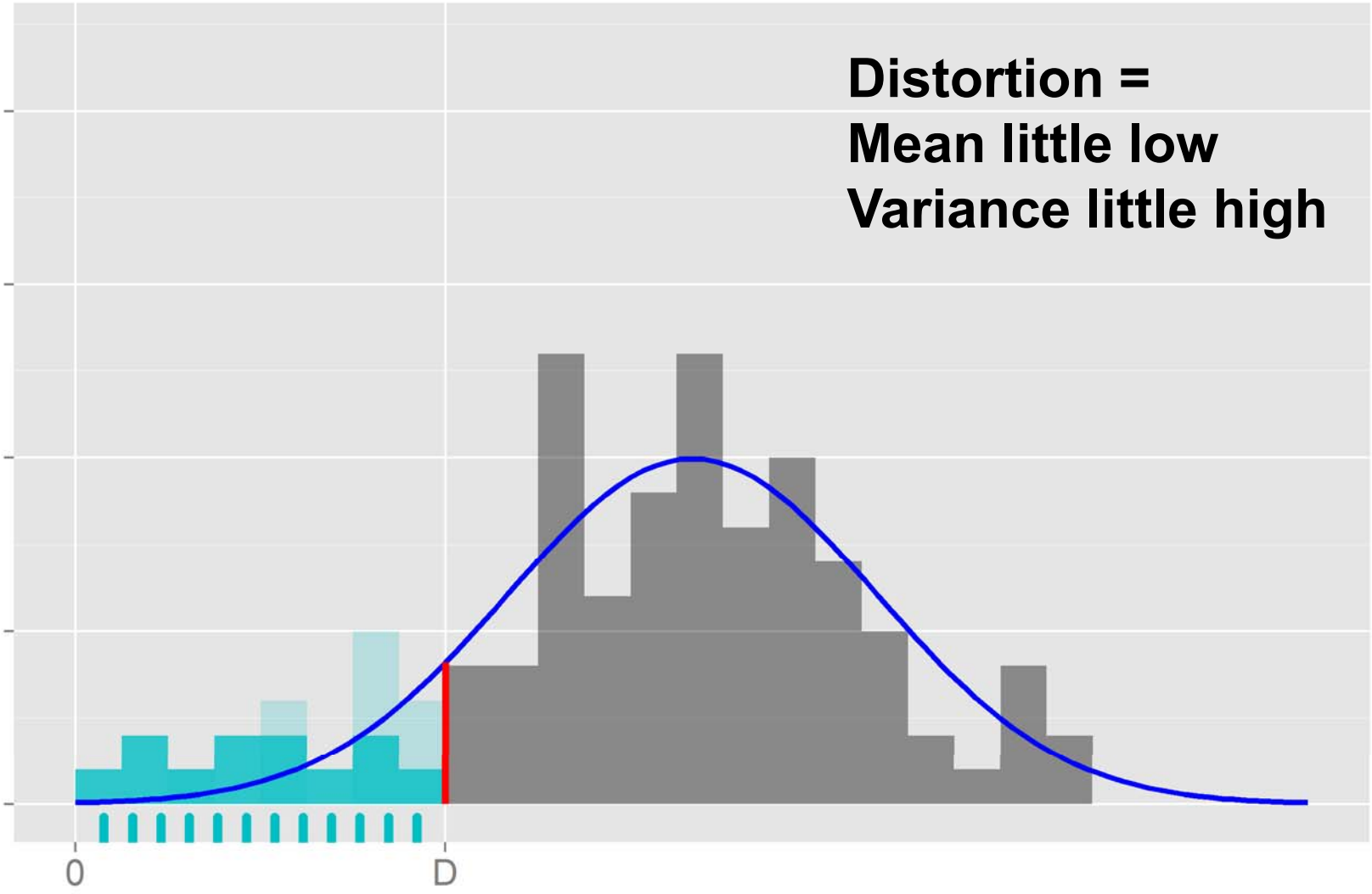
Replace all non-detects by 0



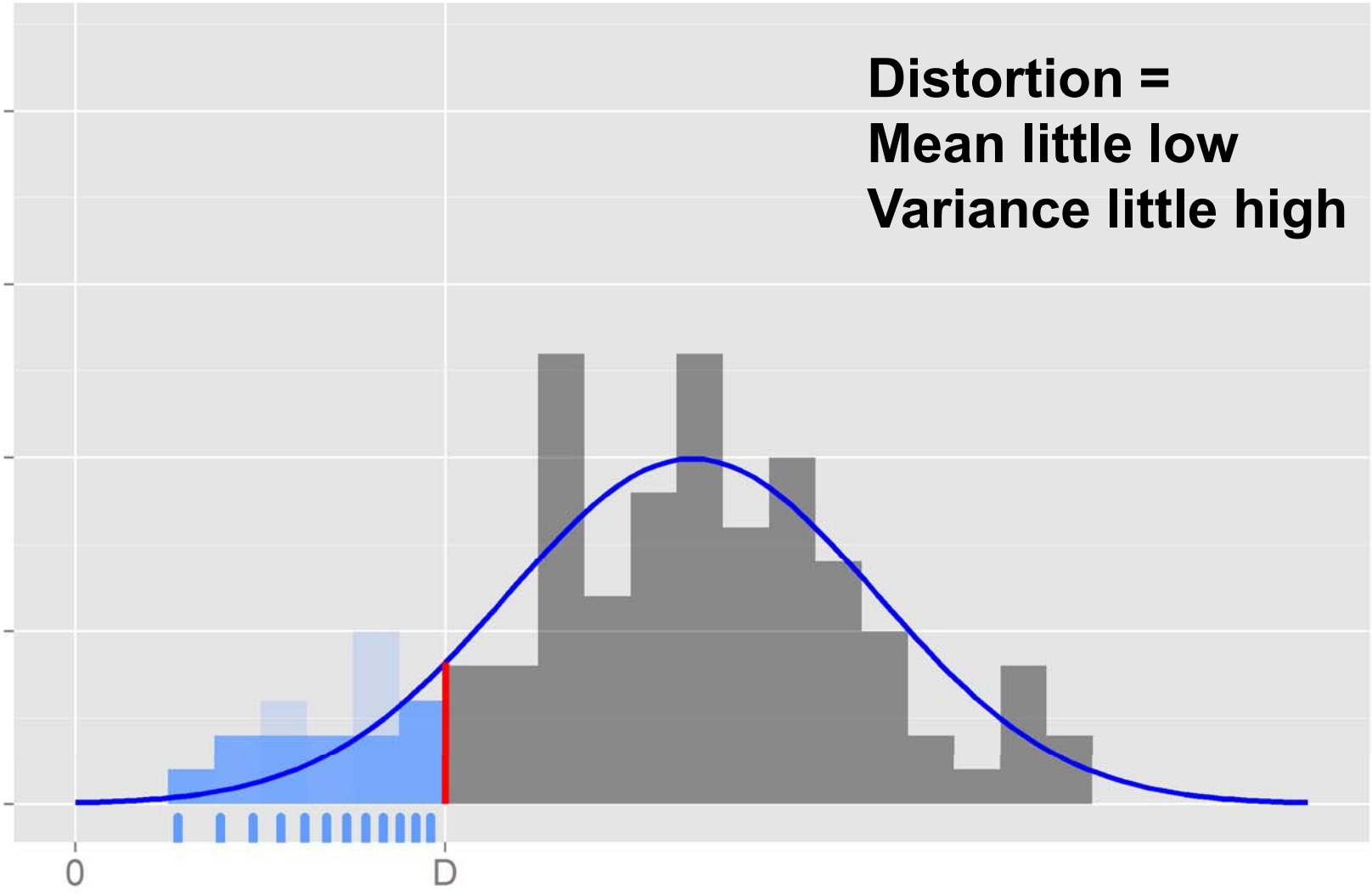
Replace all non-detects by D/2



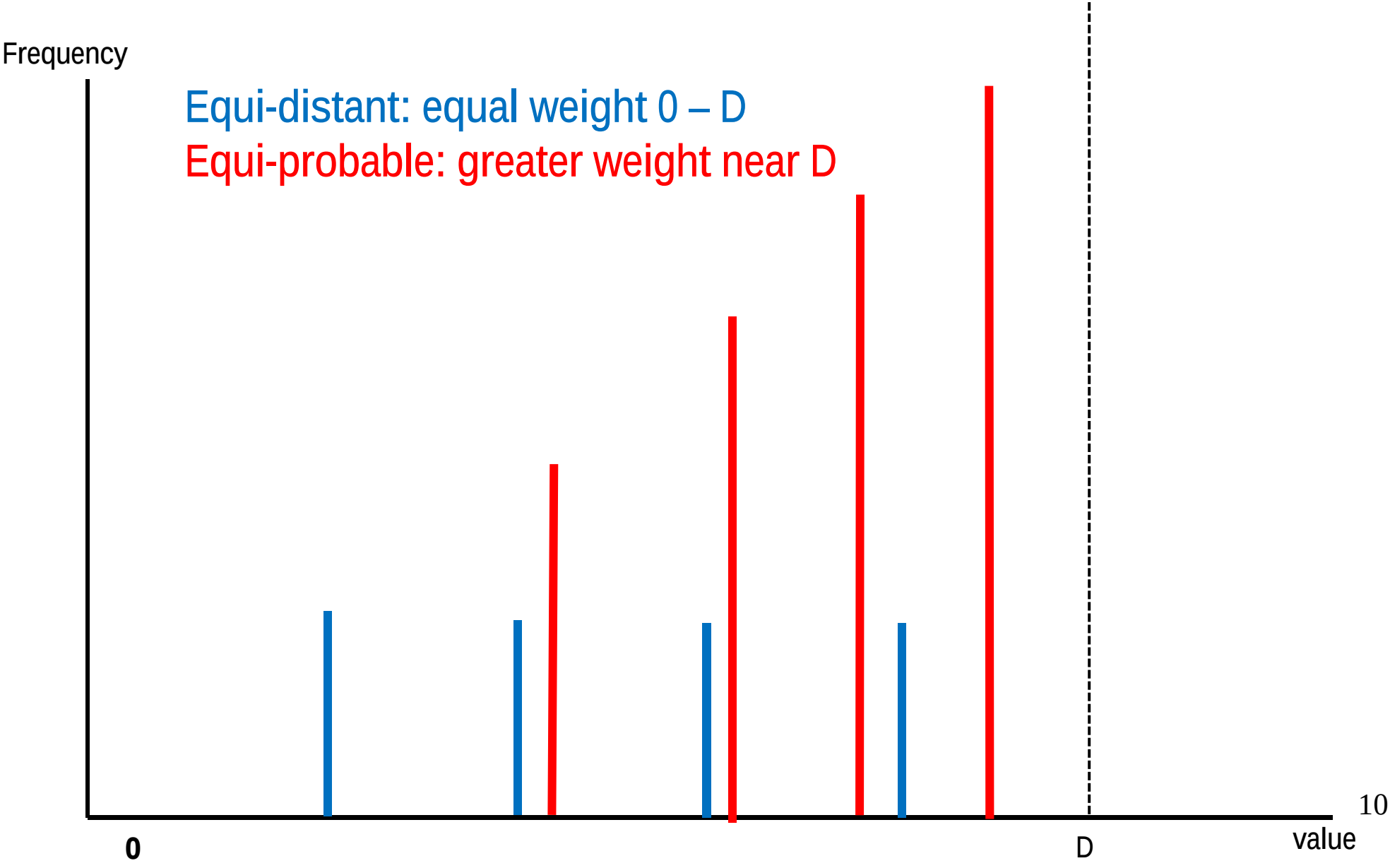
Replace by equal spacing 0 - D



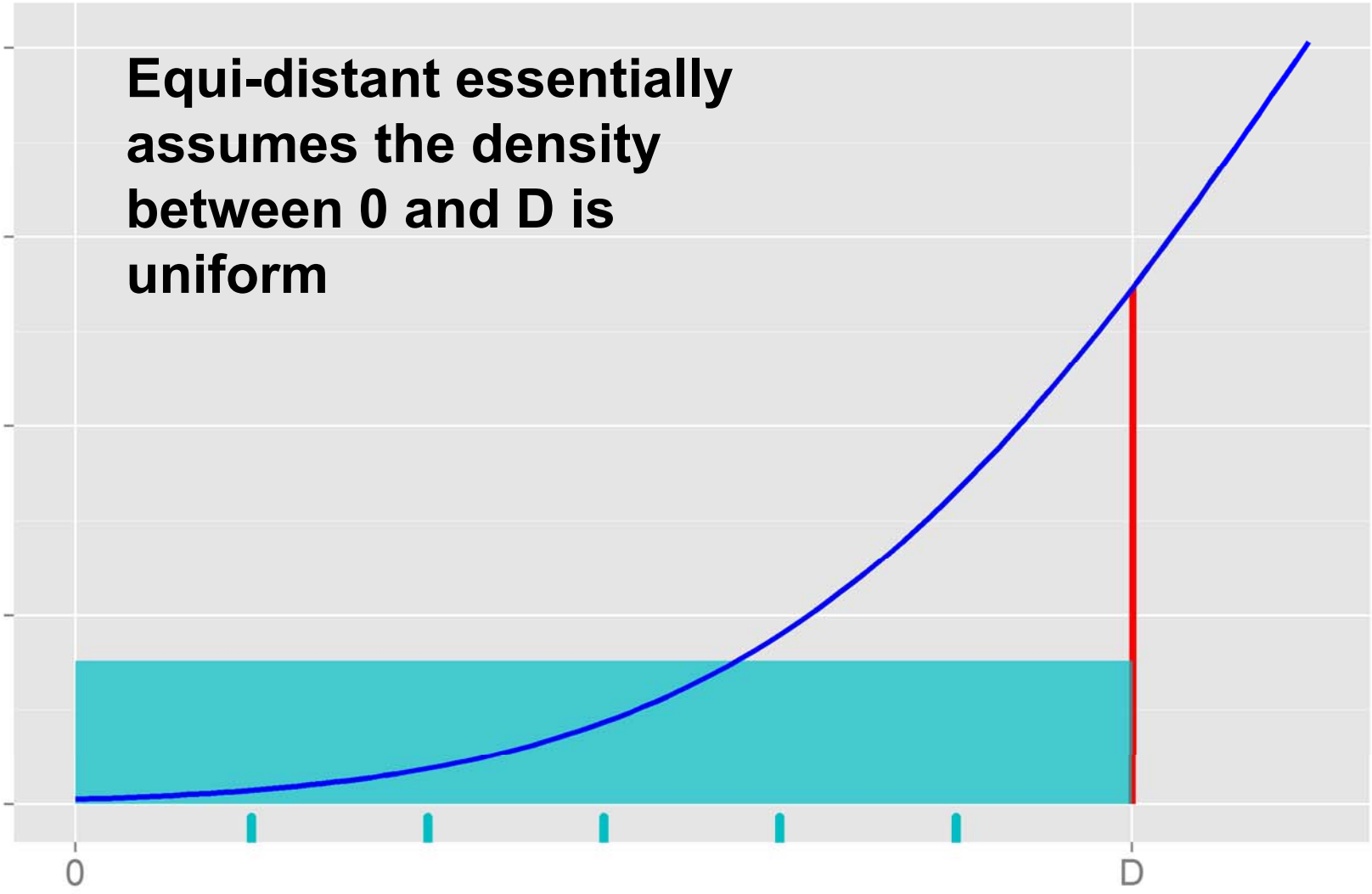
Replace by equi-probable 0 - D



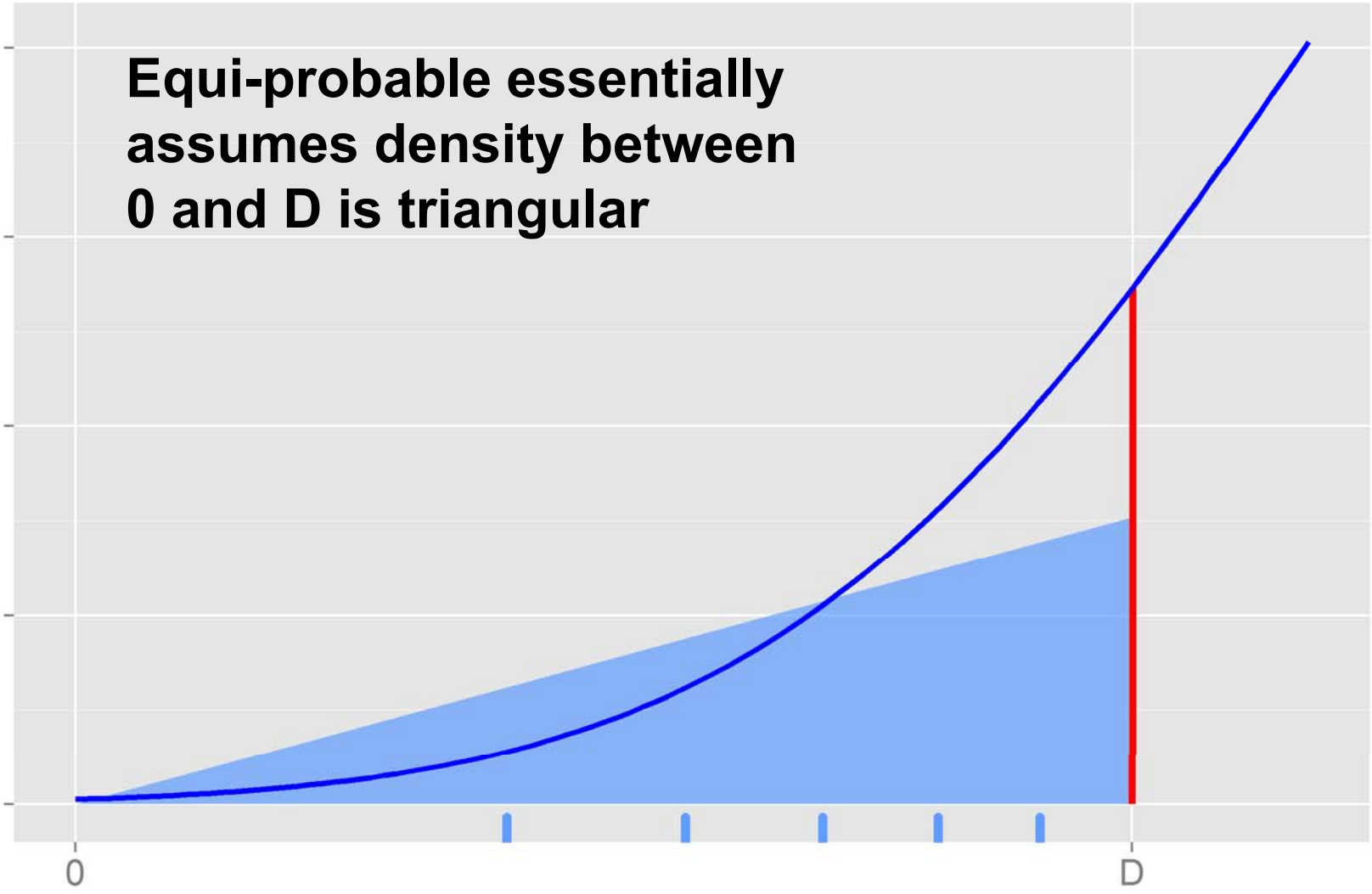
Closer look: equi-distant vs equi-probable



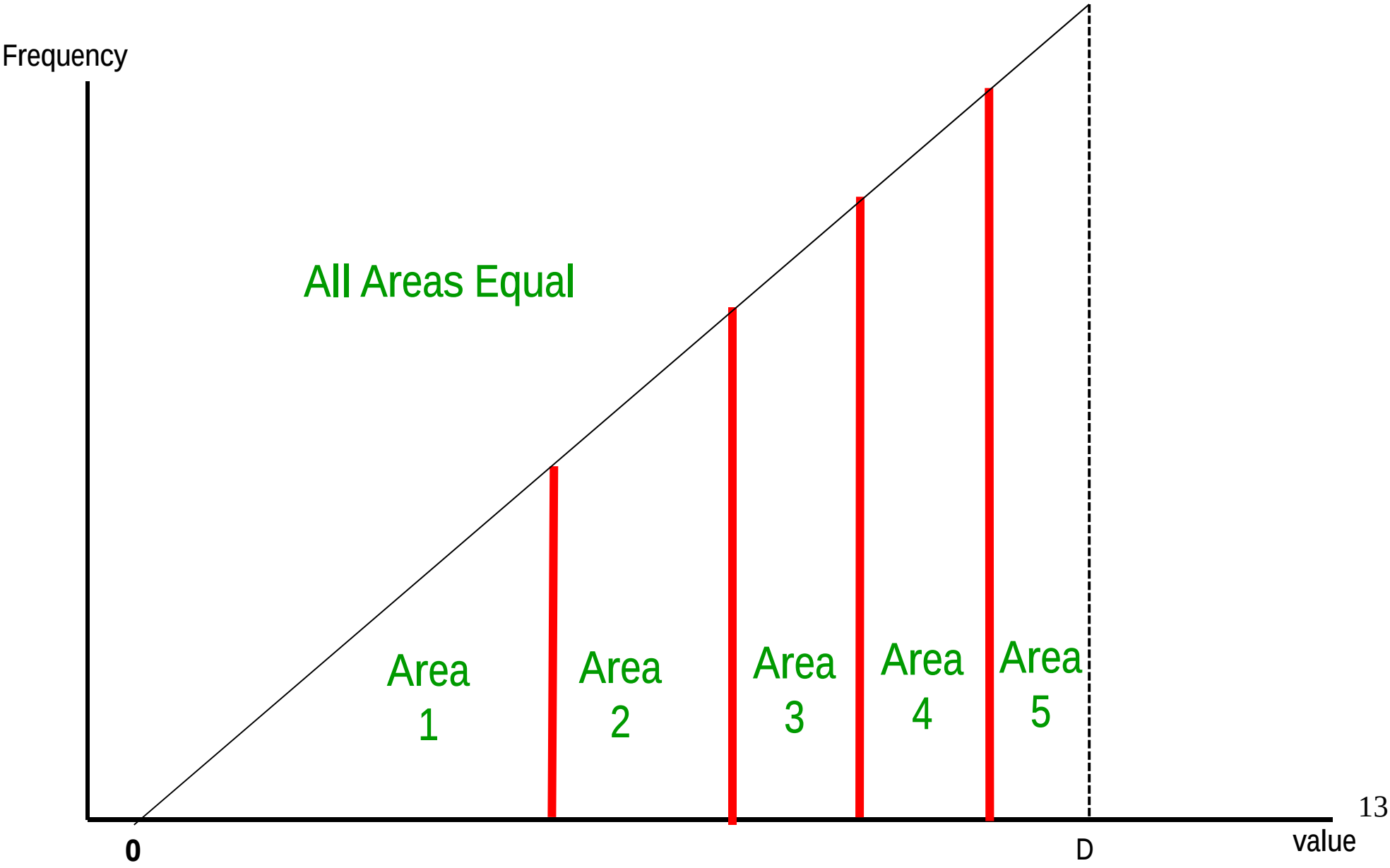
Closer look: equi-distant



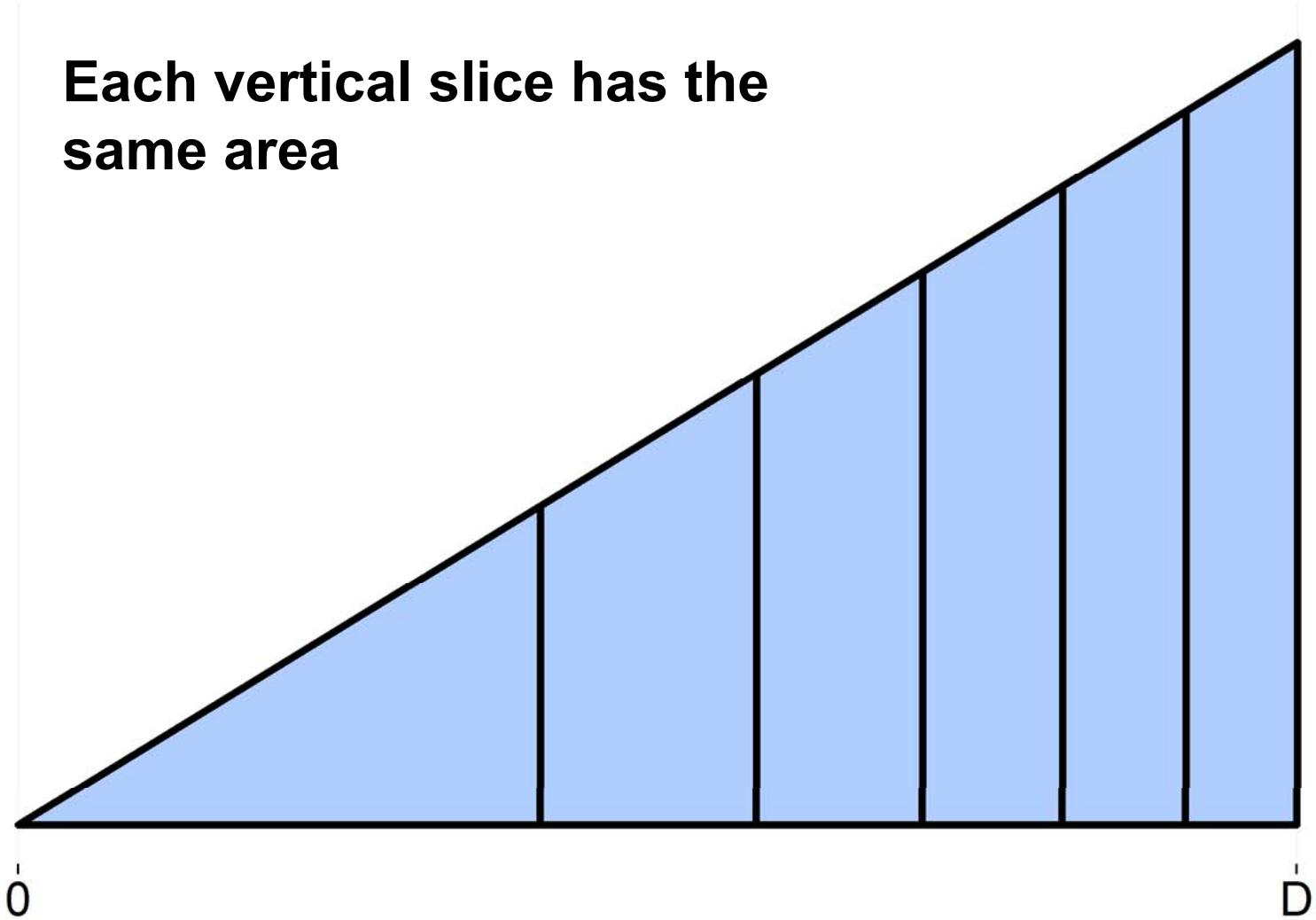
Closer look: equi-probable



How does equi-probable work?



How does equi-probable work?



How do you achieve this?

If there are "m" non-detects, each non-detect is replaced according to formula:

$$\left(\frac{k}{m+1}\right)^{1/2} D \qquad \text{for } k = 1, 2, 3, \ldots\ldots m$$

For example, suppose m = 4, then each one is replaced by:

$$(1/5)^{1/2}D,\ (2/5)^{1/2}D,\ (3/5)^{1/2}D,\ (4/5)^{1/2}D$$

i.e. 0.447D, 0.632D, 0.774D, 0.894D

What practical applications are there?

Many applications require estimating the mean and Upper Confidence Limit of the mean (UCL) using the "n" detected values and "m" non-detects.

The mean level is simply the average, X-bar, of the (n + m) values.

The UCL is X-bar + $t_{(1-\alpha)}$ s

where $t_{(1-\alpha)}$ is Student's *t* for (m + n – 1) degrees of freedom with confidence of 1 – α, and "s" is the standard deviation of the (n + m) values.

Example: Newquay lead measurements

This is essentially a random sample of 16 measurements from the Newquay site arranged smallest to largest (ppm):

<5, <5, <5, <5, 5, 8, 8, 9, 9, 9, 9, 10, 10, 10, 15, 16

The instrumentation had a detection limit of 5ppm.

There are 4 non-detects (m), and 12 detects (n).

Different substitution methods will be tried.

Comparison of substitution methods

Substitution Method	Mean	Std dev	95% UCL
0	7.38	5.08	9.60
DL	8.63	3.34	10.09
D/2	8.00	4.16	9.82
Equi-spaced	8.00	4.20	9.84
Equi-probable	8.23	3.86	9.72

Discussion

The results are pretty much as expected, setting the non-detects to 0 gave the lowest mean with the highest standard deviation; setting the non-detects to D (5ppm) gave the reverse.

All the estimates are reasonable, but because this is a real site we have no way of comparing these answers to the true value as it is unknown.

Theoretical investigation proves to be impossible and so recourse to computer simulation studies must be made.

Design of the simulation study

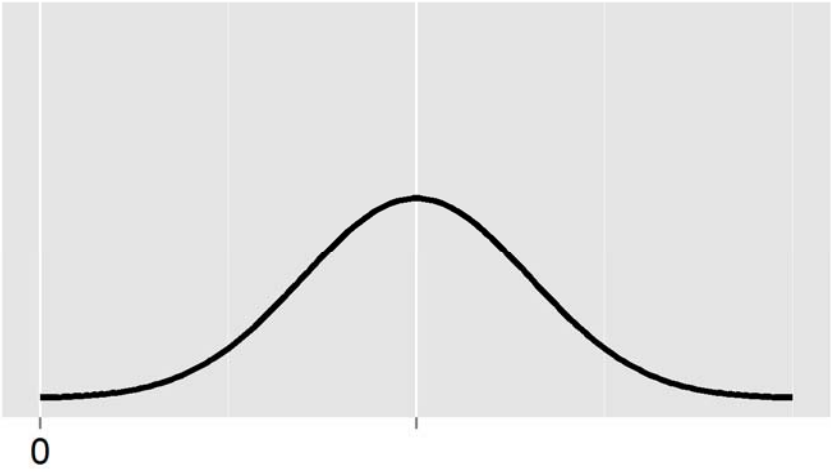
- 3 Commonly encountered distributions with common mean (10.0), but different variability:
  - Normal: std dev 1.0, and 3.0
  - Lognormal: std dev 3.0, and 5.0
  - Gamma: std dev 3.0, and 5.0
- 3 sample sizes (10, 20, and 40)
- 5 percents NDs (10%, 20%, 30%, 40%, and 50%)

- Total of 90 different combinations

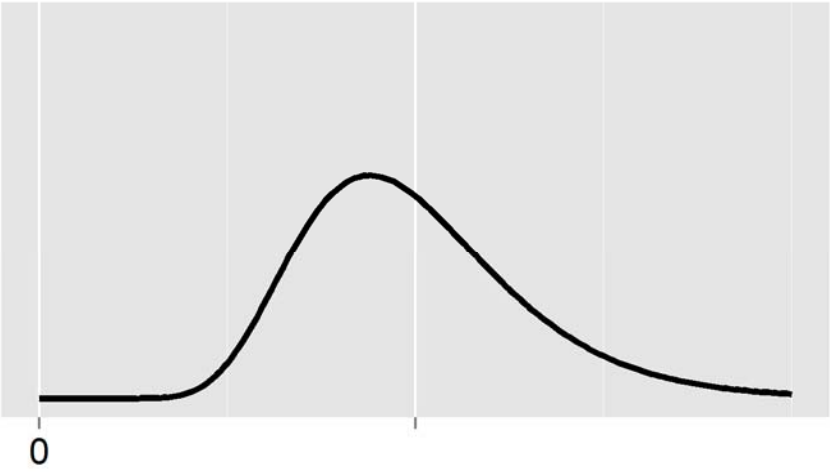
- Compare D/2, equi-spaced, equi-probable, and Kaplan-Meier (a nonparametric method)

The three distributions

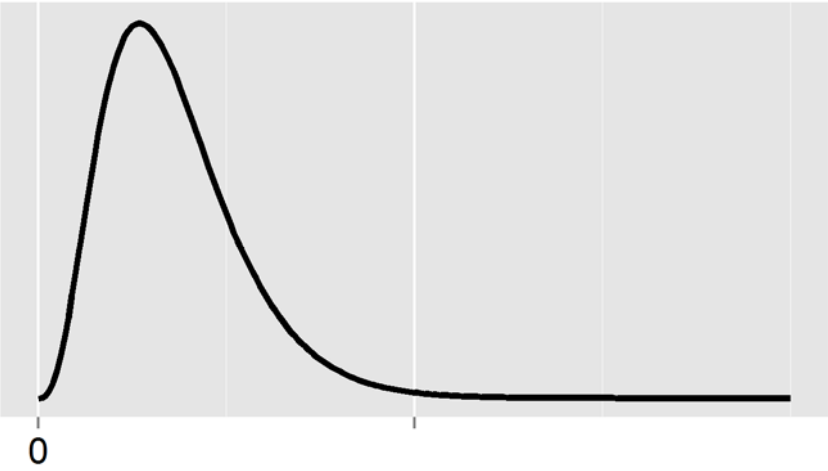
Normal:



Lognormal:



Gamma:



How simulation works

- For each combination of factors: 1000 samples taken
- For each sample the mean and 95% UCL calculated, first on the full sample (to be used for reference), then on each of the four substituting methods
- Kaplan-Meier is a method in ProUCL that may be very effective for datasets containing multiple detection limits (the subject of future studies)

How results are compared

For every combination of factors and for each of the 1000 samples taken, the following metric was calculated:

$$\text{Result} = \frac{(\text{Approximation Answer}) - (\text{Full Sample Answer})}{(\text{Full Sample Answer})}$$

and then turned into a percentage by multiplying by 100.

Time to Pick your Pony

Flame

Zero

Whiz

DL/2

Blaze

Equi-Probable

Streak

Equi-Spaced

Lightning

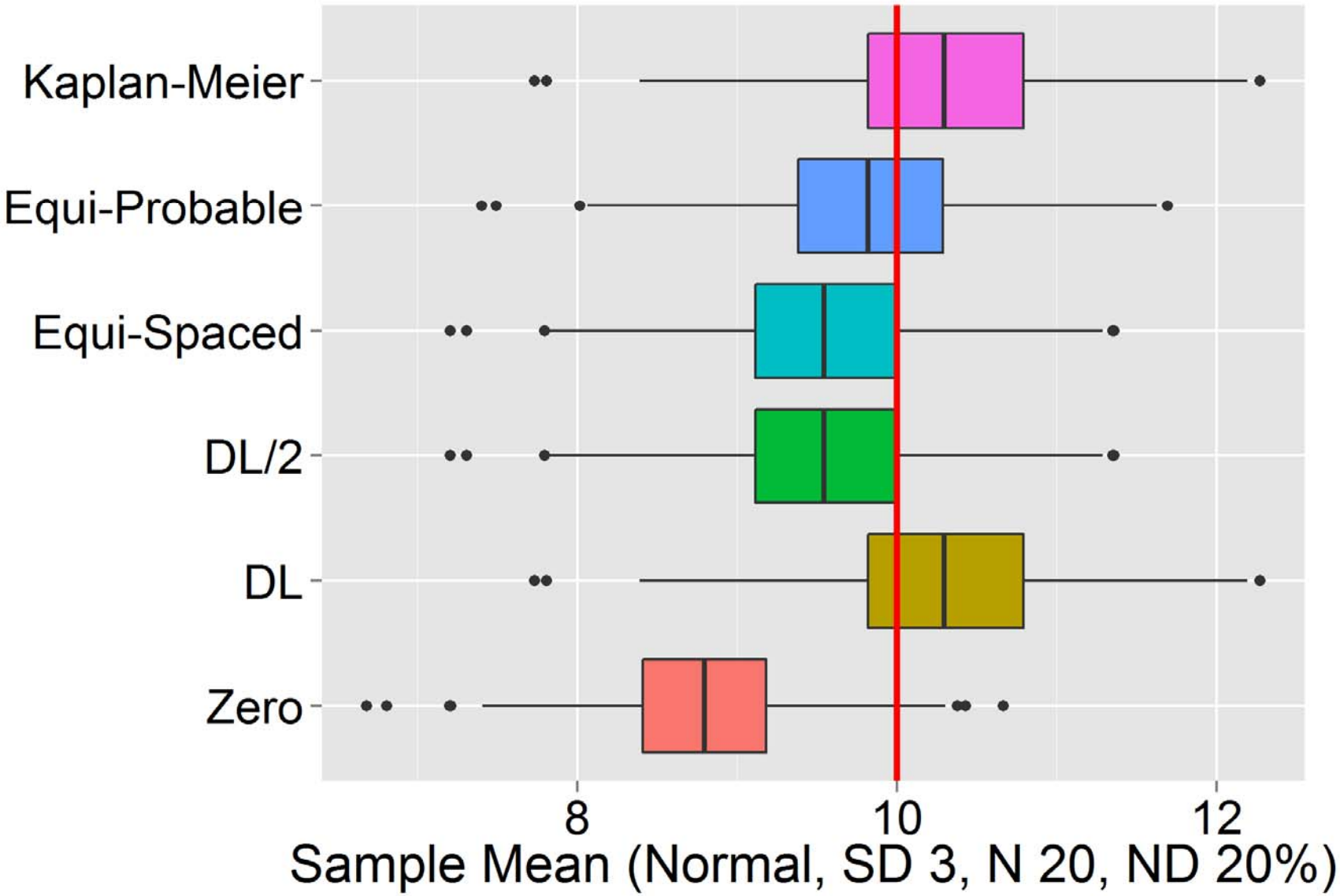
DL

Flash

Kaplan-Meier

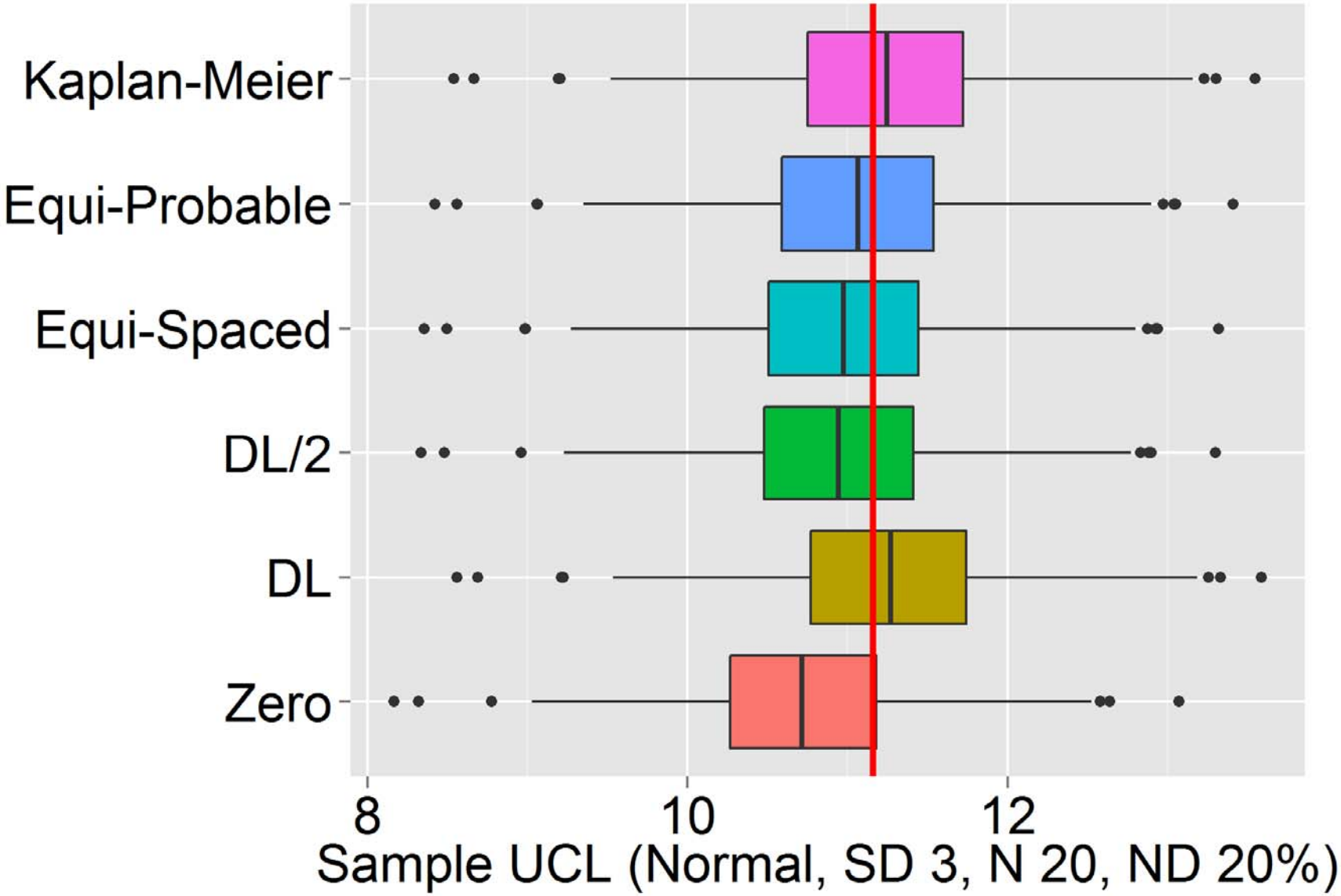
Normal: mean 10, std dev 3, 20% non-detects

Estimation of the Mean



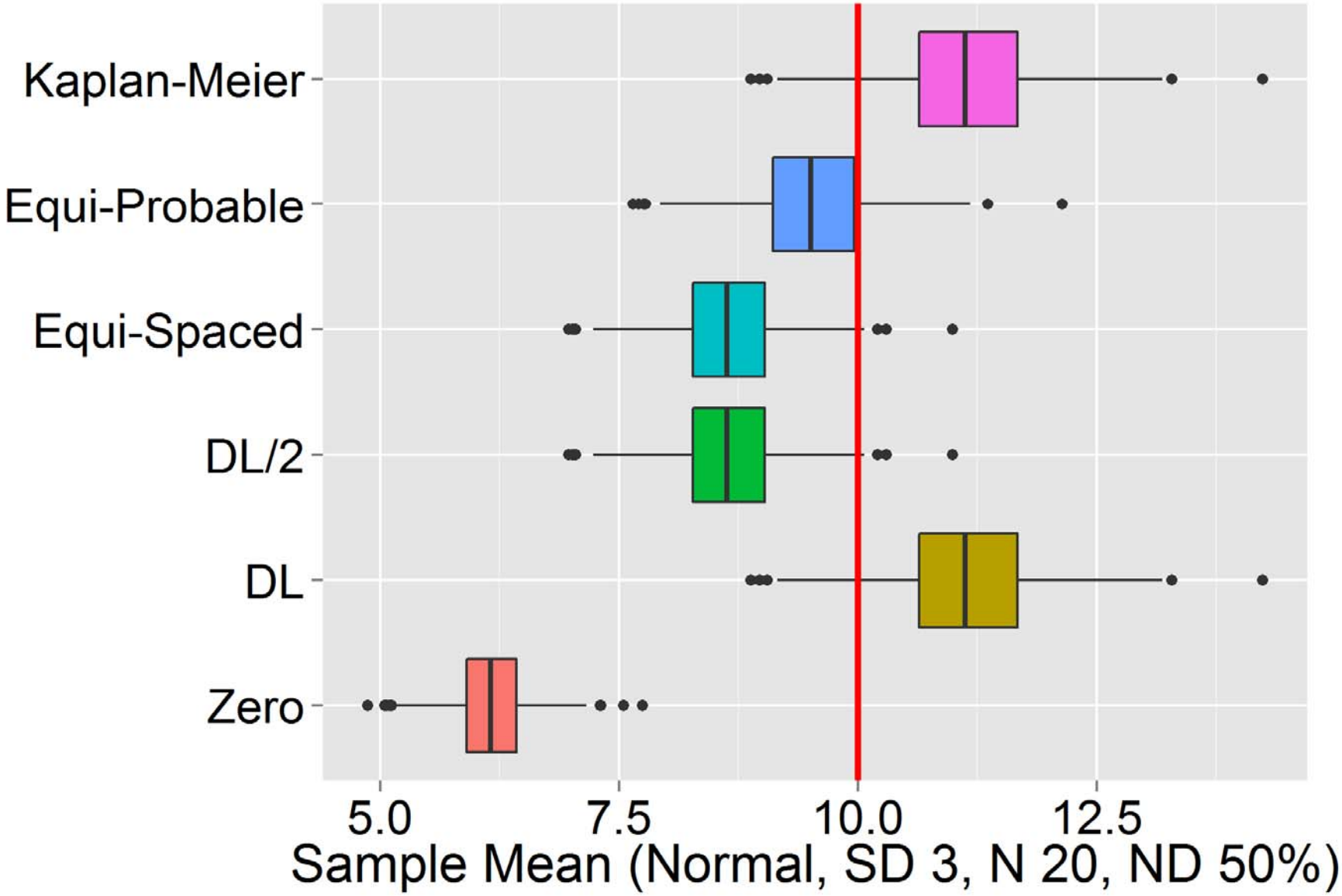
Normal: mean 10, std dev 3, 20% non-detects

Estimation of the 95% UCL



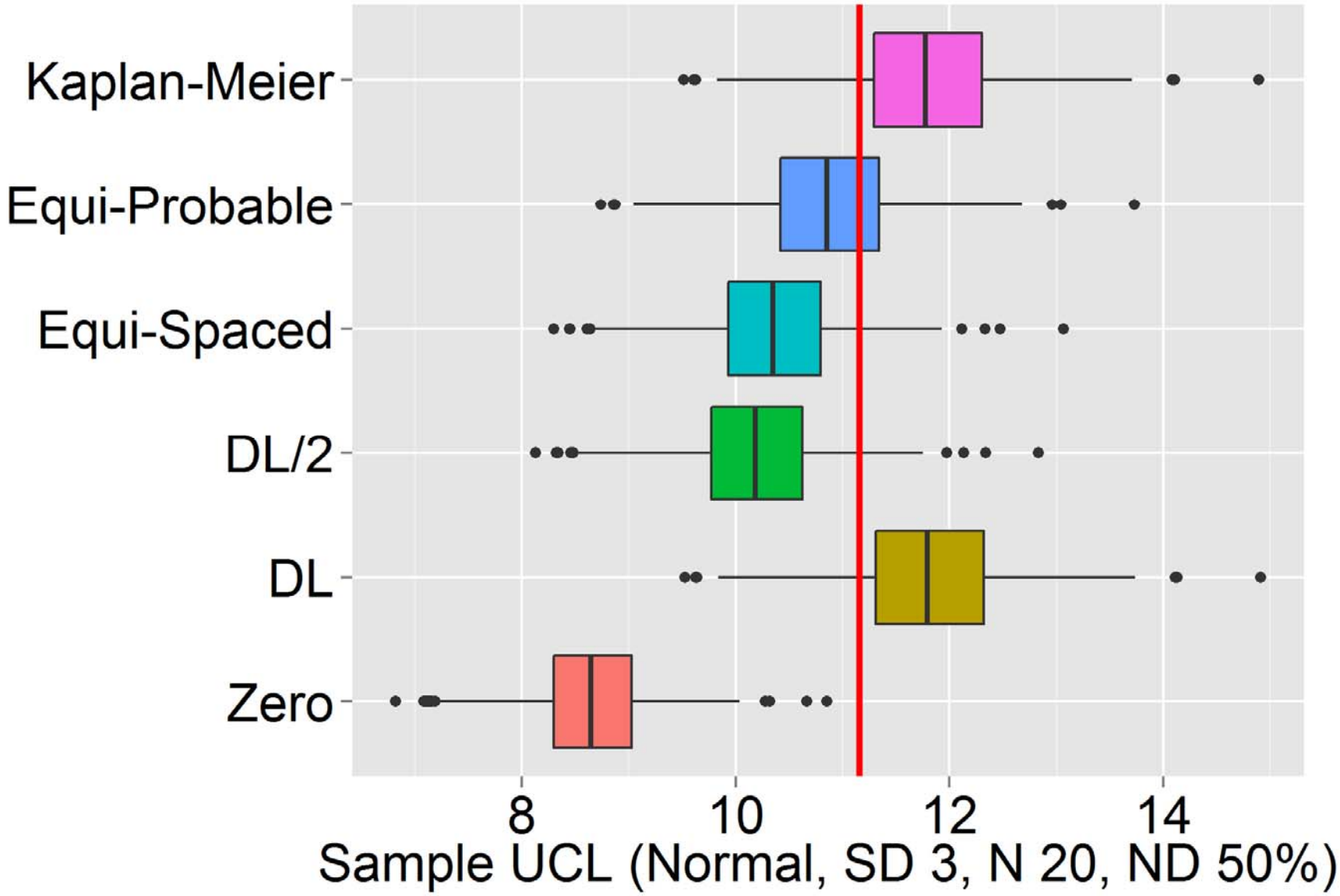
Normal: mean 10, std dev 3, 50% non-detects

Estimation of the Mean



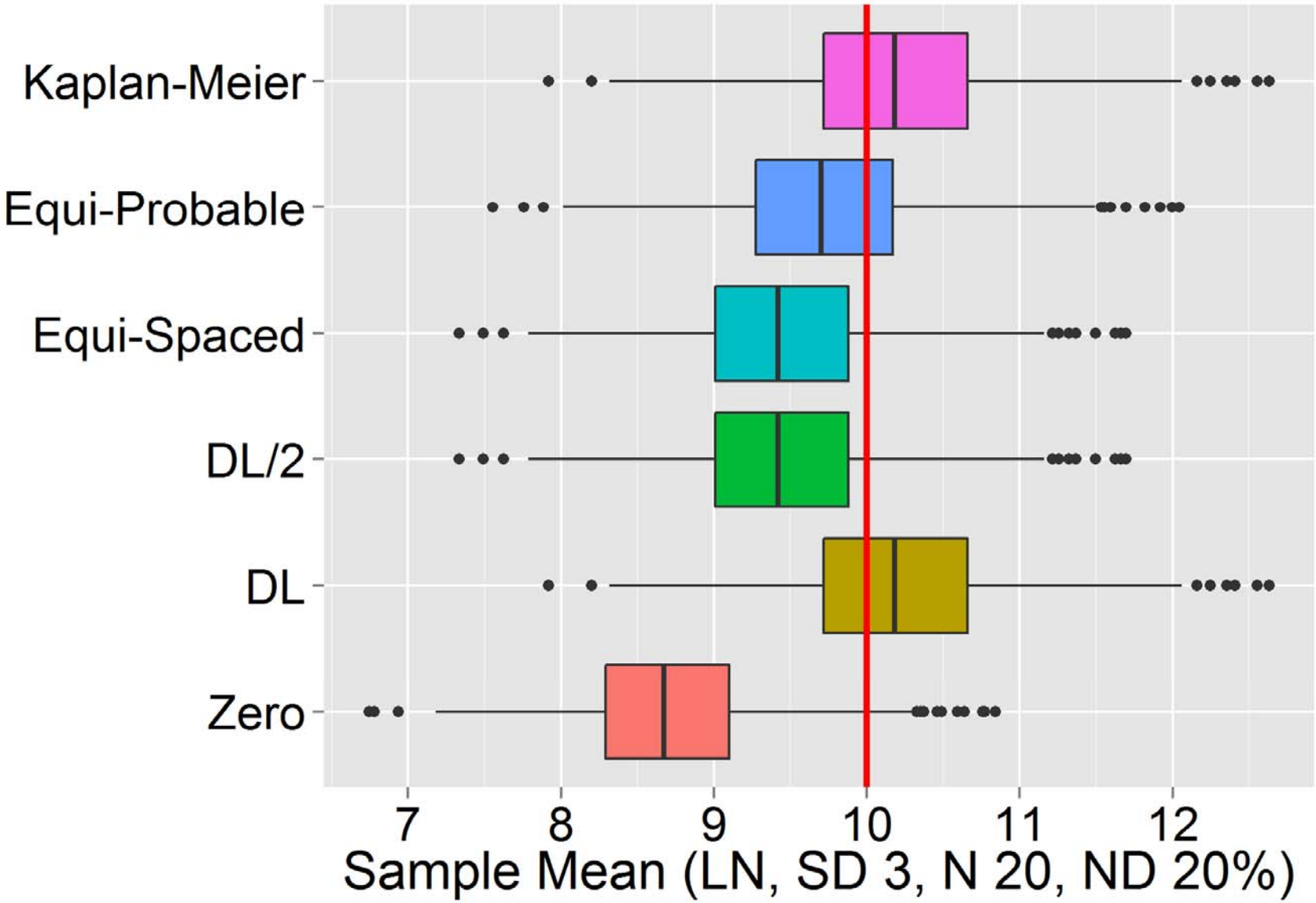
Normal: mean 10, std dev 3, 50% non-detects

Estimation of the 95% UCL



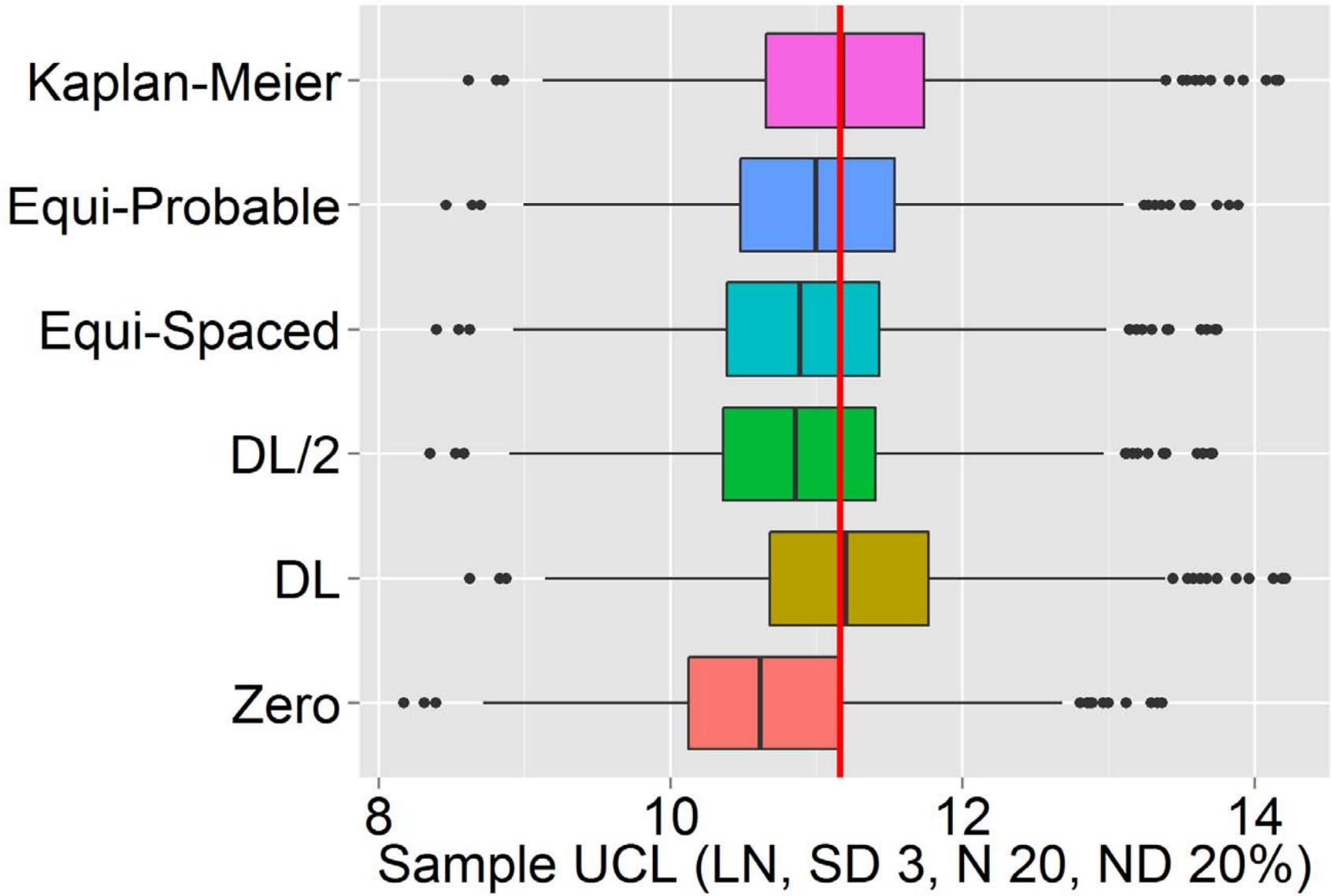
Lognormal: mean 10, std dev 3, 20% non-det

Estimation of the Mean



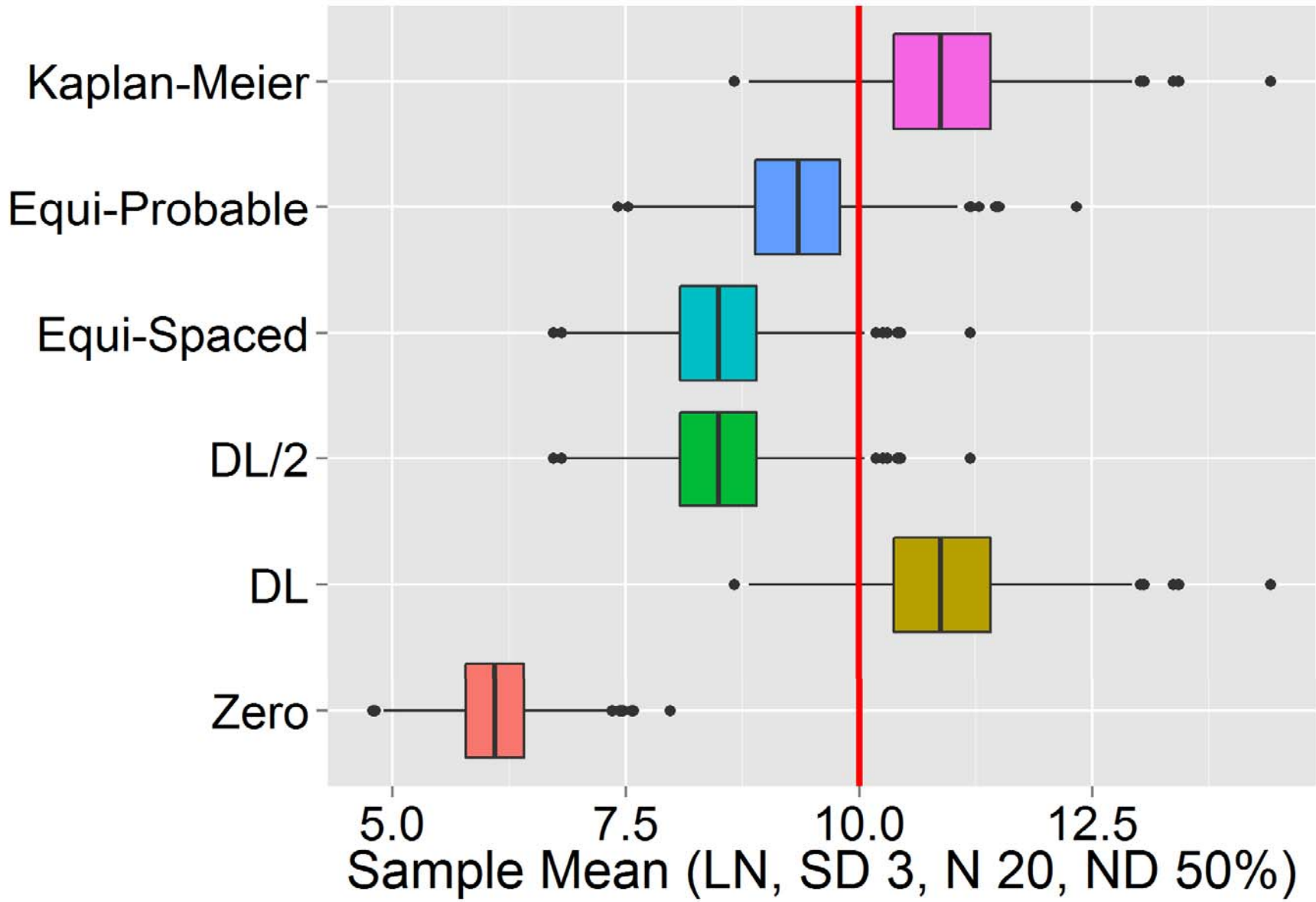
Lognormal: mean 10, std dev 3, 20% non-det

Estimation of the 95% UCL



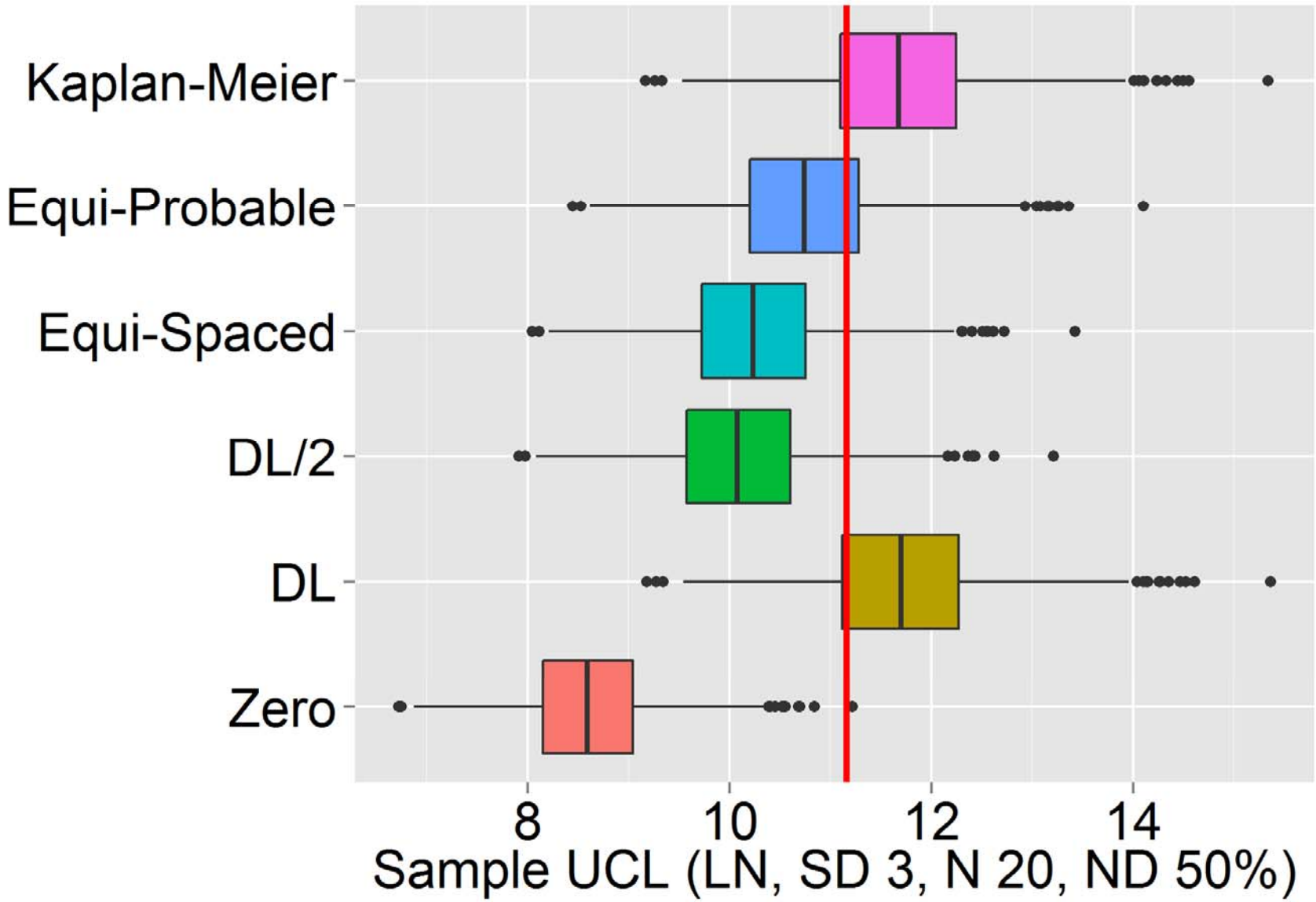
Lognormal: mean 10, std dev 3, 50% non-det

Estimation of the Mean



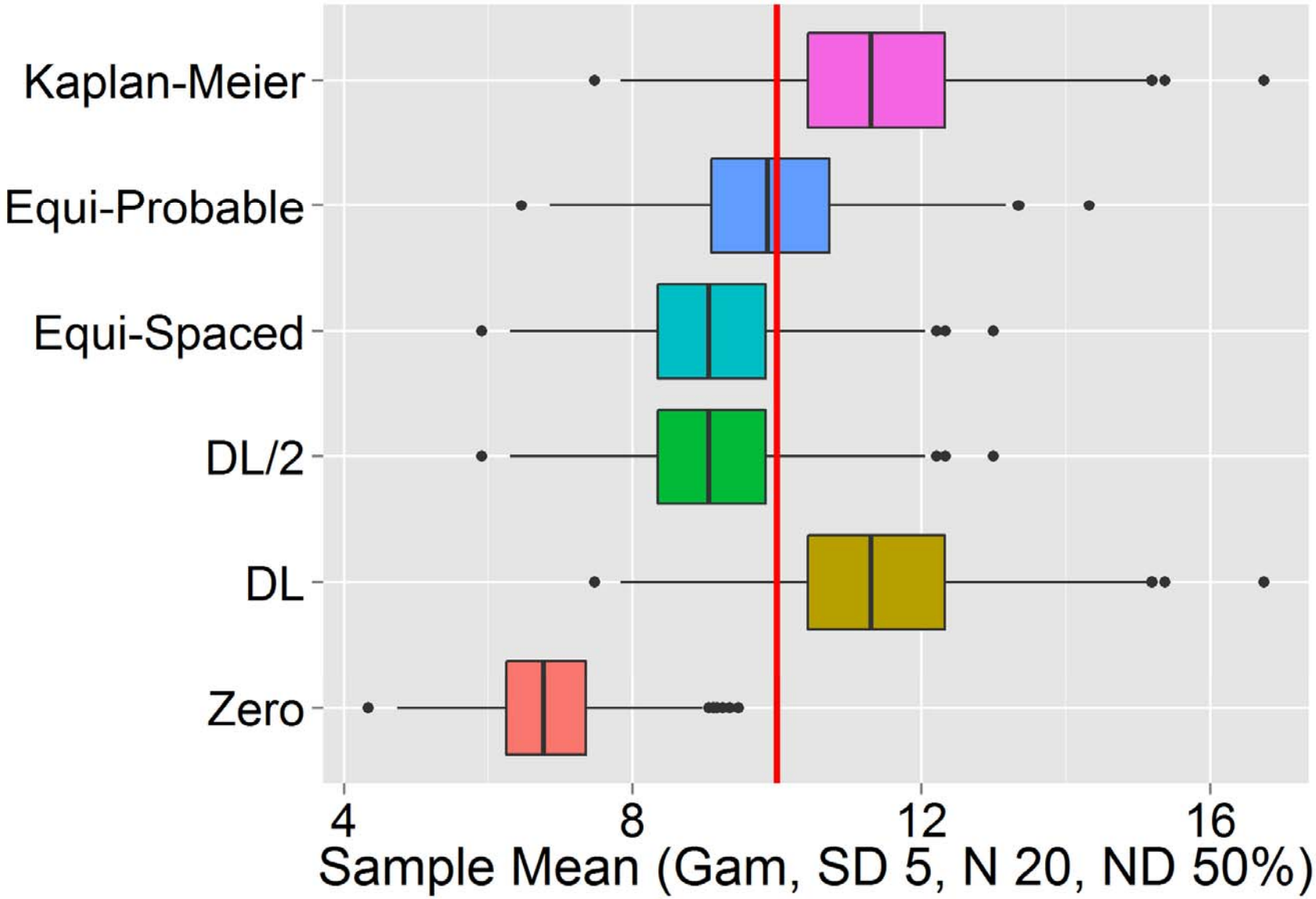
Lognormal: mean 10, std dev 3, 50% non-det

Estimation of the 95% UCL



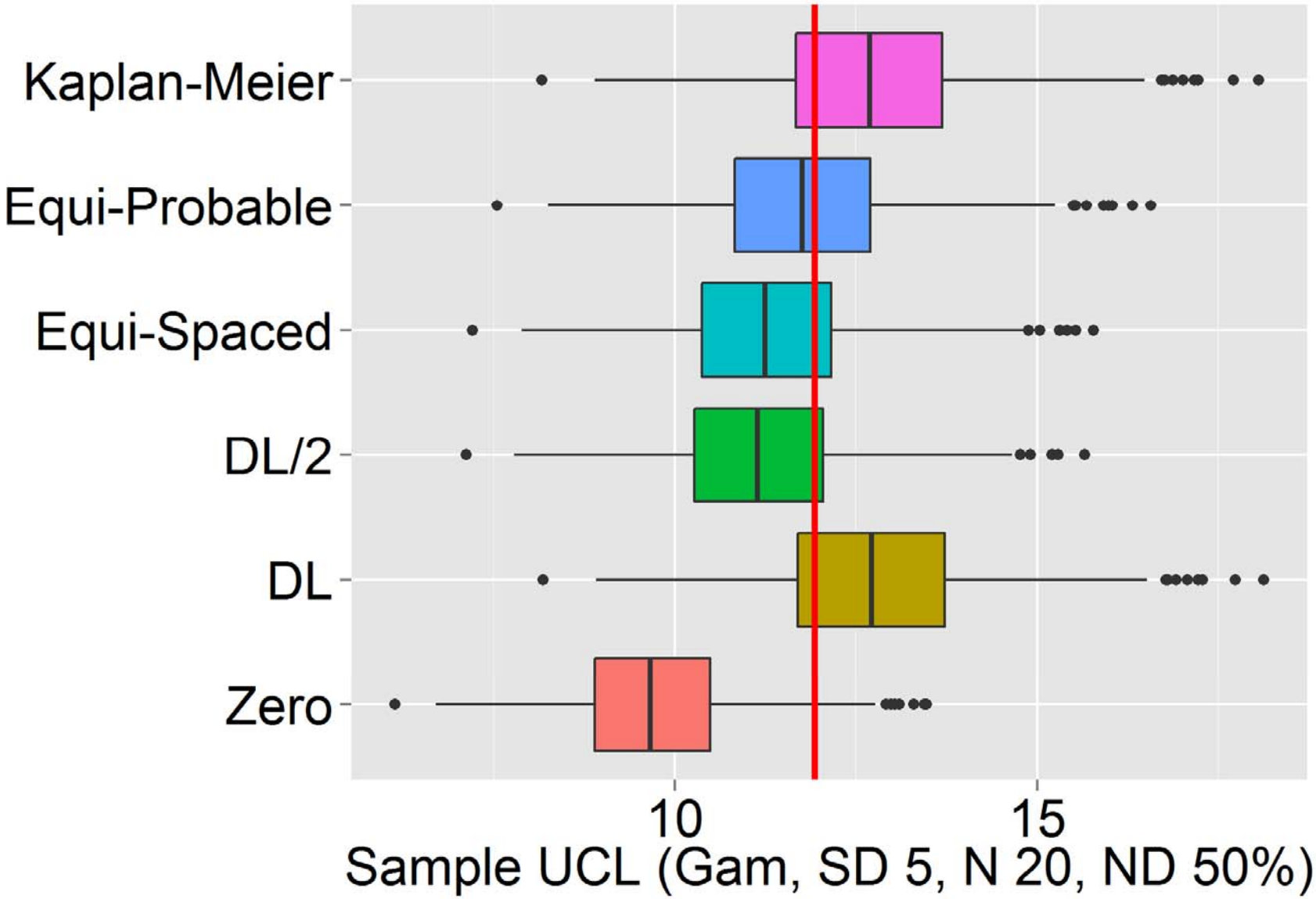
Gamma: mean 10, std dev 3, 50% non-det

Estimation of the Mean



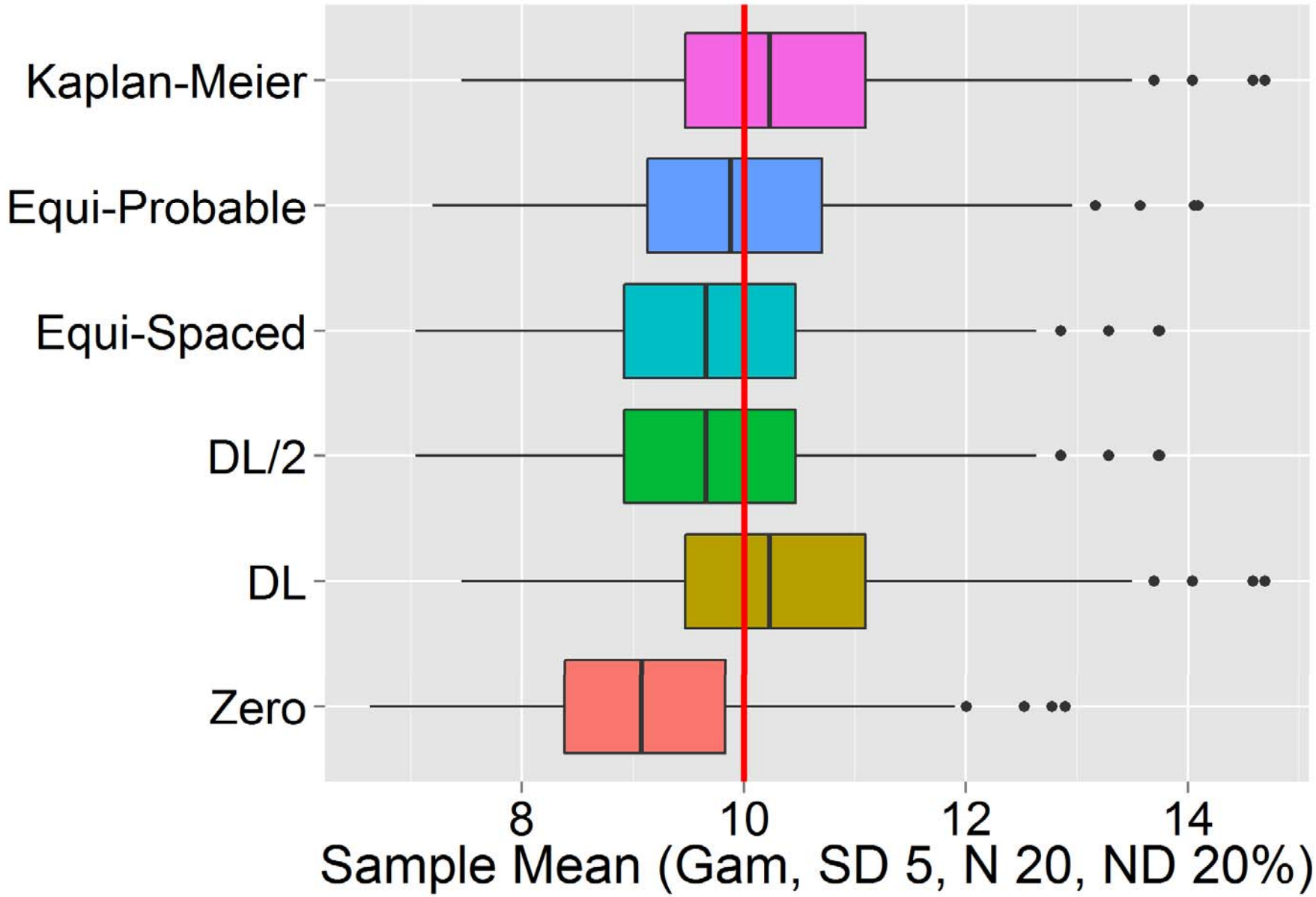
Gamma: mean 10, std dev 3, 50% non-det

Estimation of the 95% UCL



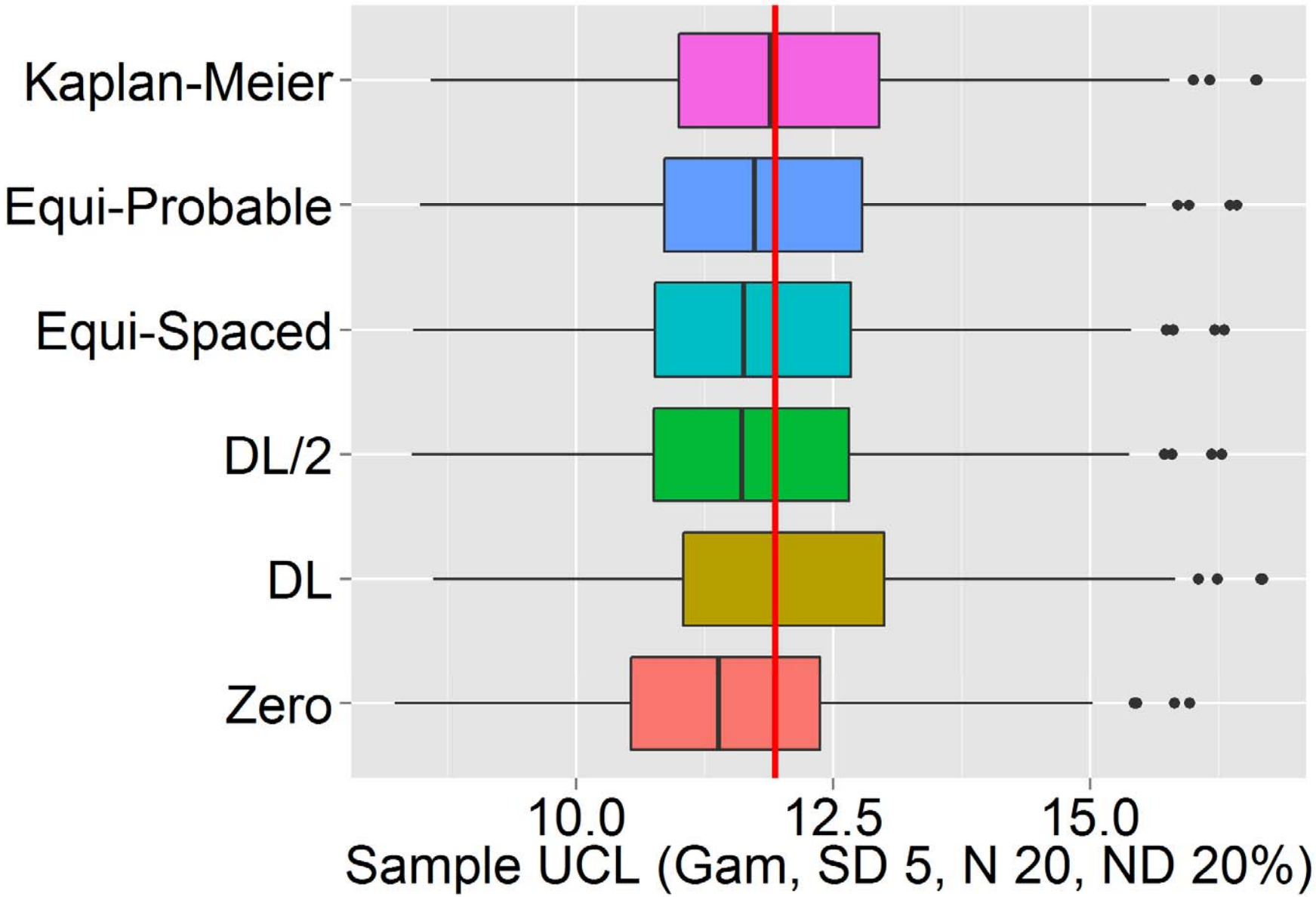
Gamma: mean 10, std dev 5, 20% non-det

Estimation of the Mean



Gamma: mean 10, std dev 5, 20% non-det

Estimation of the 95% UCL



And the winner* is….

Blaze

Equi-Probable

*At least for the distributions tested in this study.

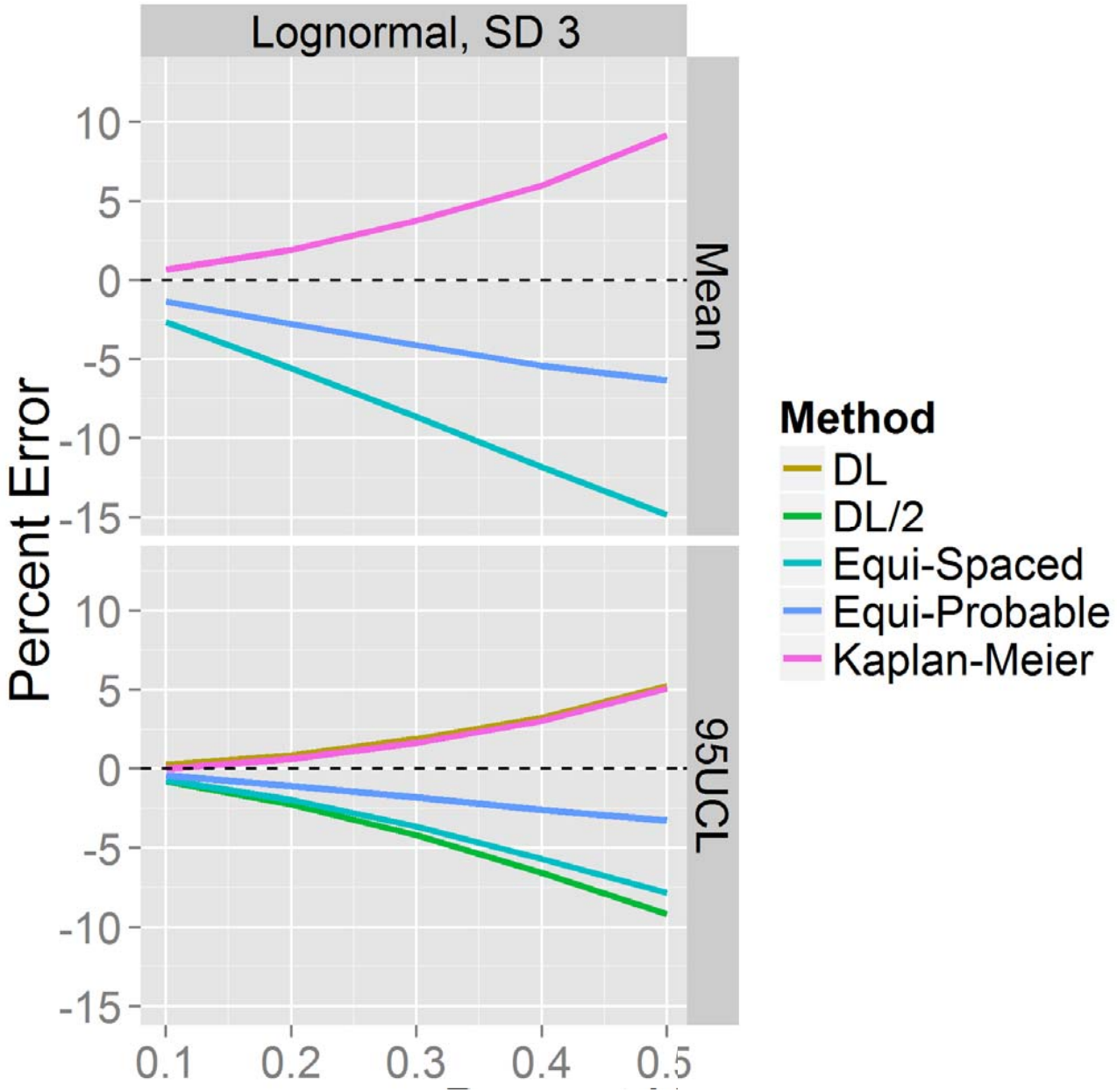
Why does Equi-Probable work well?

https://neptuneshiny

What works best?

Mean

95% UCL



Percent Non-detect

General Remarks

- Distributional Family (Normal, Lognormal, Gamma) is not the driving factor

- Efficacy depends more on:
  - Relative magnitude of mean, standard deviation, and zero
  - Percentage of non-detects

- More investigation is needed

Future Investigation

- How does Equi-probable compare with Kaplan-Meier when the datasets have multiple detection limits?
- How do the methods compare when estimating tolerance intervals (UTL -- very different than UCL)?
- Field-testing these methods
- Creating algorithms for Equi-probable to encourage use