Small sample bias in estimates

sstools Team

2019-12-04

Introduction

What is the issue

The *ssdtools* package uses the method of Maximum Likelihood (ML) to estimate parameters for each distribution that is fit to the data. Statistical theory says that maximum likelihood estimators are asymptotically unbiased, but does not guarantee performance in small samples.

For example, consider the CCME silver data that ships with *ssdtools*.

```
data(ccme_data)
Ag <- ccme_data[ ccme_data$Chemical=="Silver",]
Ag$ecdf <- (rank(Ag$Conc)+.25)/(nrow(Ag)+.5)
Ag
```

```
## # A tibble: 9 x 6
##   Chemical Species                Conc Group       Units  ecdf
##   <chr>    <chr>                 <dbl> <fct>       <chr> <dbl>
## 1 Silver   Oncorhynchus mykiss    0.24 Fish        ug/L  0.132
## 2 Silver   Lemna gibba            0.63 Plant       ug/L  0.237
## 3 Silver   Ceriodaphnia dubia     0.78 Invertebrate ug/L  0.342
## 4 Silver   Pimephales promelas    0.83 Fish        ug/L  0.447
## 5 Silver   Ictalurus punctatus    1.9  Fish        ug/L  0.553
## 6 Silver   Daphnia magna          2.12 Invertebrate ug/L  0.658
## 7 Silver   Hyalella azteca        4    Invertebrate ug/L  0.763
## 8 Silver   Chironomus tentans    13    Invertebrate ug/L  0.868
## 9 Silver   Micropterus salmoides 23    Fish        ug/L  0.974
```

Let us fit a log-normal distribution to the Ag endpoint data and estimate the parameters:

```
fit <- ssd_fit_dists(Ag, dist="lnorm")
fit
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##          estimate Std. Error
## meanlog 0.6840072  0.4640735
## sdlog   1.3922205  0.3281487
```

For most distributions, the MLE must be found numerically by iterative methods, but the log-normal distribution has easily computed estimators.

The *meanlog* parameter shown above represents the mean of the concentrations on the (natural) logarithmic scale and we can easily reproduce this value:

```
mean(log(Ag$Conc))
```

```
## [1] 0.6840072
```

The *sdlog* parameter represents the standard deviation on the logarithmic scale, but the direct computation of the standard deviation gives a slightly different result:

```r
sd(log(Ag$Conc))
```

```
## [1] 1.476673
```

It turns out that in small samples, the MLE of the standard deviation for a log-normal distribution has a negative bias, i.e. the MLE tends to be smaller than the underlying true parameter value. The cause of this bias is found by comparing the formula for the MLE of the standard deviation and the traditional estimator for the standard deviation:

$$\widehat{\sigma}_{MLE} = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n}}$$

$$\widehat{\sigma}_{traditional} = \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{n-1}}$$

where n is the sample size, Y_i are the $\log(concentrations)$, and $\overline{Y}$ is the sample mean concentration again on the logarithmic scale.

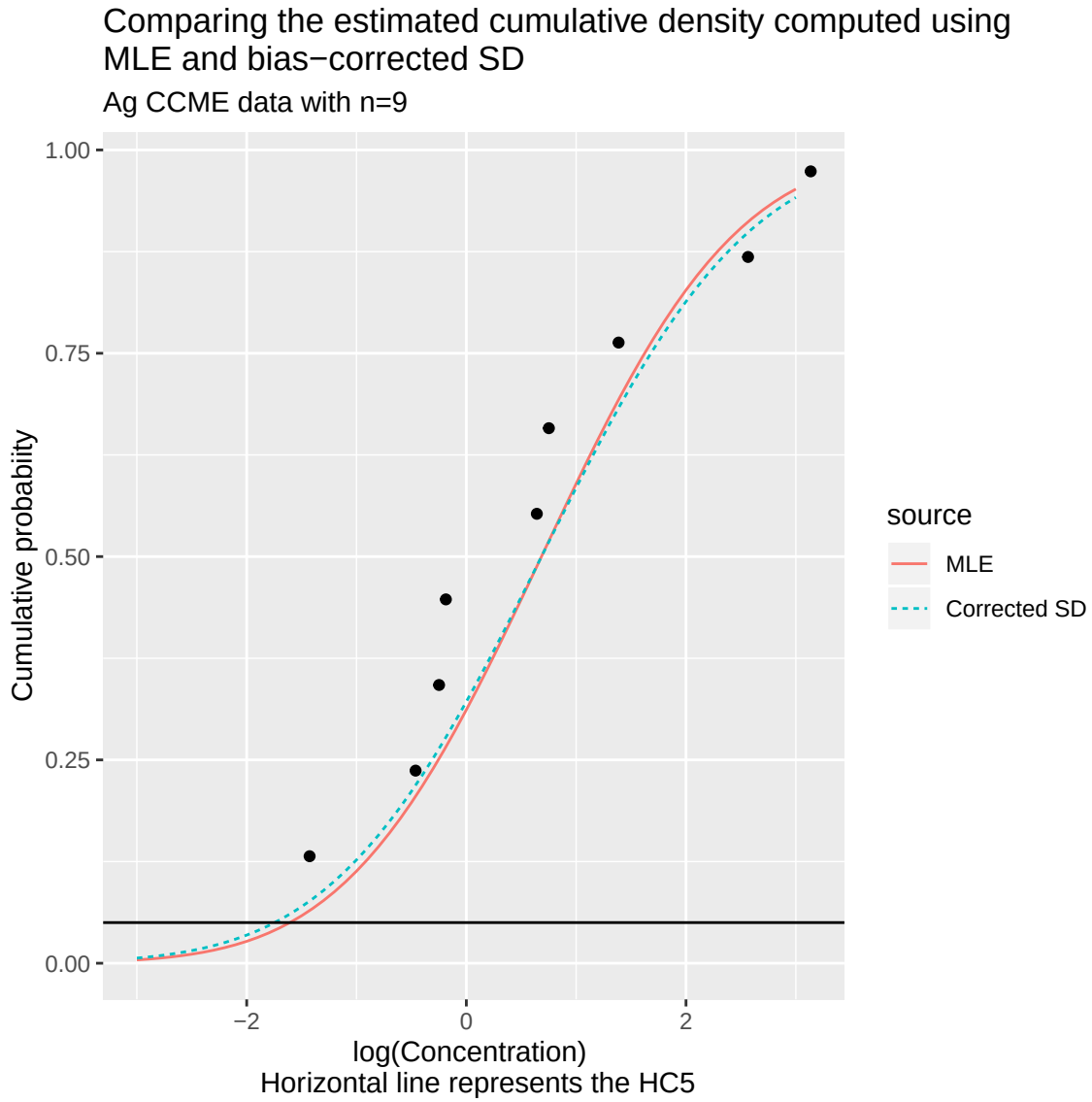
We notice that the MLE uses a divisor of n while the traditional method uses a divisor of $n-1$. Hence the MLE has a negative bias and its value is 0.94x the usual estimator for σ which is $\sqrt{\frac{n-1}{n}} = 0.94$ evaluated at $n = 9$.

As the sample size increases the absolute size of the bias will get smaller and smaller, i.e., if $n = 20$, then the MLE estimator is 0.97x the traditional estimator for σ which is neglible given the uncertainties in the actual end points.

Conversly, as the sample size decreases, the absolute size of the bias could become quite large, i.e., if $n = 4$, then the MLE is 0.87x the traditional estimator. But if you are fitting a species sensitivity distribution to only 4 data values, perhaps concern about bias in MLE is misplaced.

What is the impact on the HCx value?

If the standard deviation is underestimated, then the tails of the distribution will be pulled inwards and the HCx values will tend to be larger compared to the case where the standard deviation is not deflated as shown in the following plot:



The HC5 estimated from the MLE fit is -1.61 on the logarithmic concentrations scale or 0.201 on the concentration scale. The HC5 estimated after correcting the standard deviation for small sample bias is -1.74 on the logarithmic concentrations scale or 0.175 on the concentration scale. The ratio of HC5 values is 0.87x on the concentration scale.

The differences between the HCx computed from the MLE fit and using the corrected standard deviation will become more pronounced for small HCx values. For example, the HC1 estimated from the MLE fit is 0.078 and the HC1 estimated using the corrected standard deviation is 0.064 on the concentration scale. The ratio of the HC1 values is 0.82x on the concentrations scale

What can be done?

A similar concernalso occurs with other distributions. Howeve, except for a few distributions, such as the normal distribution, analytical expressions for the MLE and for unbiased estimators do not exist. The *mle.tools* package from CRAN provides a method that numerically corrects the bias after the fit is completed.

Bias correction using Cox-Snell method - log-normal distribution

For example, again using the Ag log-normal fit we have:

```
# apply the Cox and Snell (1968) bias correction using mle.tools.
# what is the density function
norm.pdf <- quote(1 / (sqrt(2 * pi) * sigma) * exp(-0.5 / sigma ^ 2 * (x - mu) ^ 2))
norm.pdf
```

```
## 1/(sqrt(2 * pi) * sigma) * exp(-0.5/sigma^2 * (x - mu)^2)
```

```
# what is the log(density) function (ignoring constants)
log.norm.pdf <- quote(- log(sigma) - 0.5 / sigma ^ 2 * (x - mu) ^ 2)
log.norm.pdf
```

```
## -log(sigma) - 0.5/sigma^2 * (x - mu)^2
```

```
bias.correct <- coxsnell.bc(density = norm.pdf,
            logdensity = log.norm.pdf,
            n = length(Ag$Conc),
            parms = c("mu", "sigma"),
            mle = c(fit$lnorm$estimate["meanlog"],
                    fit$lnorm$estimate["sdlog" ]),
            lower = '-Inf', upper = 'Inf')
bias.correct
```

```
## $mle
##        mu     sigma
## 0.6840072 1.3922205
##
## $varcov
##                   mu         sigma
## mu     2.153642e-01 -4.458357e-14
## sigma -4.458357e-14  1.076821e-01
##
## $mle.bc
##        mu     sigma
## 0.6840072 1.5082388
##
## $varcov.bc
##                   mu         sigma
## mu     2.527538e-01 -2.018812e-13
## sigma -2.018812e-13  1.263769e-01
##
## $bias
##            mu         sigma
## -2.863140e-12 -1.160184e-01
```

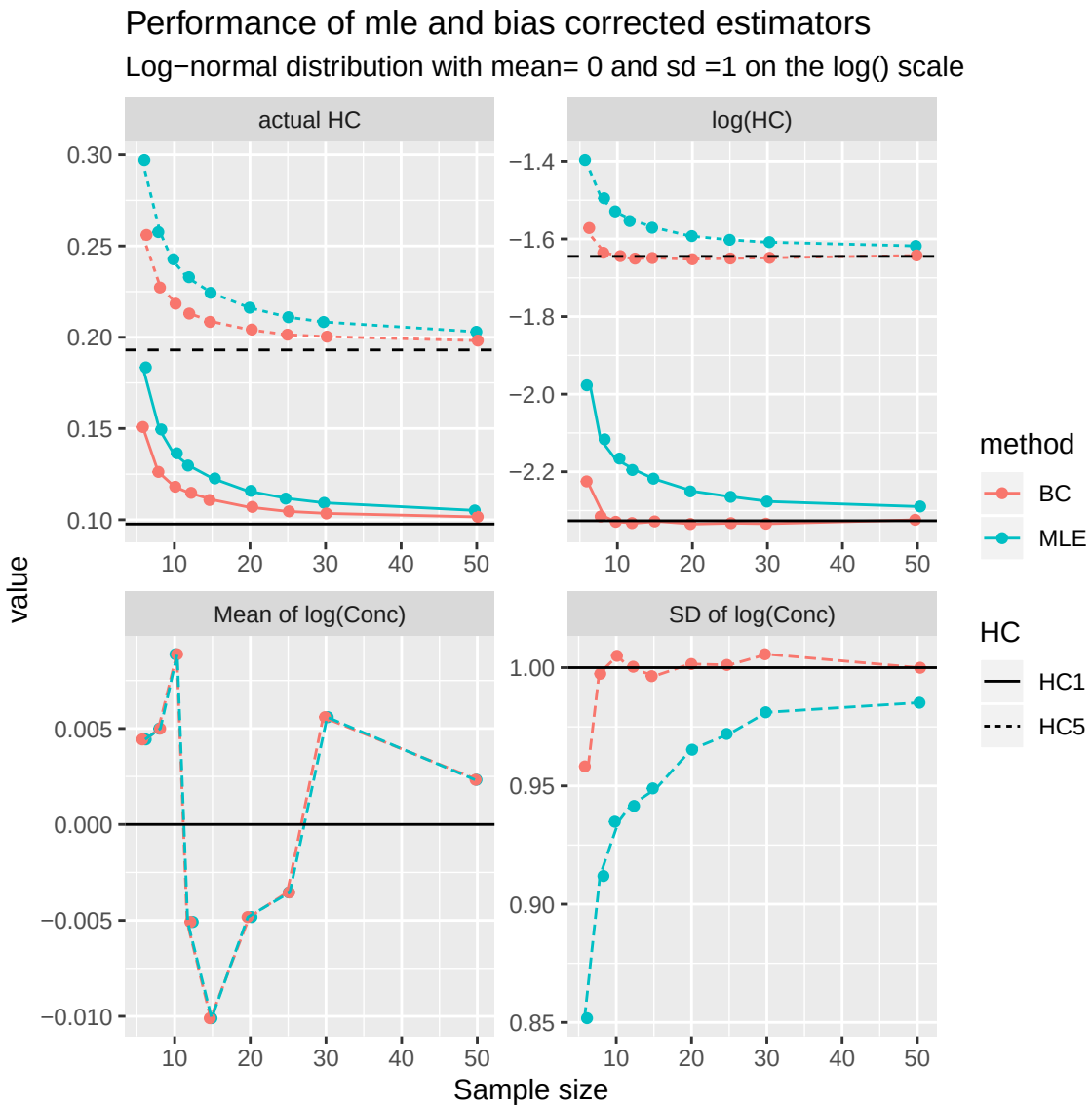
The biased corrected value for the standard deviation is 1.51 which is comparable to the standard deviation of the log(concentration) found earlier of 1.48.

A small simulation study was conducted to investigate the effect of sample size on the bias correction and effects of the small-sample bias in the estimates of HC5 and HC1. For this simulation study, it was assumed that a log-normal distribution represented the distribution of endpoints among species with a mean of 0 and a standard deviation of 1 (on the logarithmic scale). These values are arbitrary, but any log-normal distribution can be rescaled (e.g. by changing units) to have this mean and standard deviation.

Simulated data sets at various sample sizes were generated, the MLE and bias-corrected estimates were

obtained and these were used to estimate the HC5 and HC1 on the log() and anti-log scales. The average value of each response was then computed and plotted vs. the actual parameter values based on the known mean and standard deviation (shown in the plot below as a black horizontal line). For example, for a log-normal distribution with a mean of 0 and a standard deviation of 1 on the log-scale, the $log(HC5)$ is the 0.05 quantile of the normal distribution or -1.645.

A plot of the results is:



The MLE is unbiased for the mean of the log-normal distribution (bottom left plot) - the apparent deviations from the true value of 0 are very small (note the scale on the Y axis) and simply simulation artefacts.

The MLE for the standard deviation is biased downwards (lower right plot) and the bias become smaller with increasing sample size (the curve for the mean of the MLE estimate of the standard deviation increases and approaches the true value of 0). The bias-correction for the standard deviation is effective for all but the smallest sample sizes.

The estimated log HC (upper right plot) based on the MLE is biased upwards (i.e. larger) than the true values but the bias declines with sample size (as expected). The estimate of the log HC based on the bias-corrected estimates performs well (close to the true value) except at very small sample sizes.

Finally, the estimated $HC1$ and $HC5$ values are again biased upwards (upper left plot). This bias consists of two parts

1. bias in the underlying estimates of the parameters of the distribution
2. non-linear tranformation bias, i.e. the mean of a function of the parameter values is not equal to the function evaluated at the mean of the parameter values. For example, the HC5 is found as the anti-log of the 5^{th} percentile of the normal distribution. Suppose we have two simulation results where the estimated 5^{th} percentile of the fitted normal distribution were -1.8 and -1.5. The mean of the estimated $5th$ percentile is $\frac{-1.8+(-1.5)}{2} = -1.65$ and is unbiased for the actual percentile value of -1.645. However, the actual HC5 is found as the anti-log of the two individual estimates, i.e. $\exp(-1.8) = 0.165$ and $\exp(-1.6) = .223$ whose mean is 0.194, but the anti-log of the average, $exp(-1.65) = .192$ which is not the same value.

The total bias does not appear to be large except in the case of very small sample sizes.

Bias correction using Cox-Snell method - gamma distribution

We can also apply this to other distributions such as the gamma distribution. If we fit a gamma distribution to the Ag data we obtain:

```
fit.gamma <- ssd_fit_dists(Ag, dist="gamma")
fit.gamma
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters:
##        estimate Std. Error
## scale 8.0858115  4.6620883
## shape 0.6389626  0.2544429
```

The bias corrected estimates are:

```
# apply the Cox and Snell (1968) bias correction using mle.tools.
# what is the density function
gamma.pdf <- quote(1 /(scale ^ shape * gamma(shape)) * x ^ (shape - 1) * exp(-x / scale))
gamma.pdf
```

```
## 1/(scale^shape * gamma(shape)) * x^(shape - 1) * exp(-x/scale)
```

```
# what is the log(density) functiong ingoring constants
log.gamma.pdf <- quote(-shape * log(scale) - lgamma(shape) + shape * log(x) -
 x / scale)
log.gamma.pdf
```

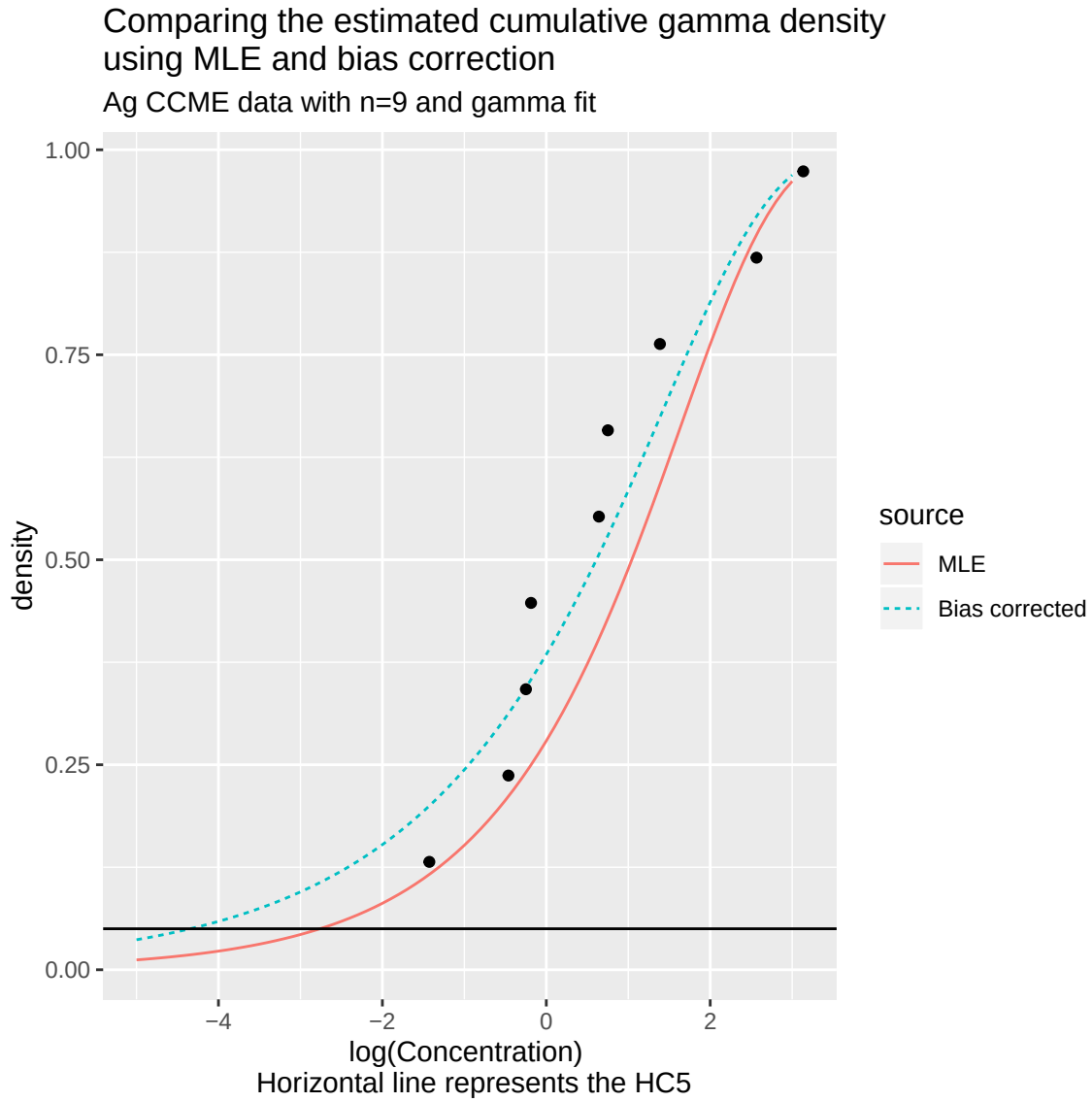
```
## -shape * log(scale) - lgamma(shape) + shape * log(x) - x/scale
```

```
bias.correct.gamma <- coxsnell.bc(density = gamma.pdf,
            logdensity = log.gamma.pdf,
            n = length(Ag$Conc),
            parms = c("shape", "scale"),
            mle = c(fit.gamma$gamma$estimate["shape"],
                    fit.gamma$gamma$estimate["scale" ]),
            lower = 0, upper = 'Inf')
bias.correct.gamma
```

```
## $mle
##     shape     scale
## 0.6389626 8.0858115
##
```

```
## $varcov
##             shape      scale
## shape  0.06474778 -0.8193984
## scale -0.81939841 21.7394342
##
## $mle.bc
##     shape     scale
## 0.4776978 8.8397972
##
## $varcov.bc
##             shape      scale
## shape  0.03426721 -0.6341122
## scale -0.63411218 29.9097308
##
## $bias
##      shape      scale
##  0.1612648 -0.7539858
```

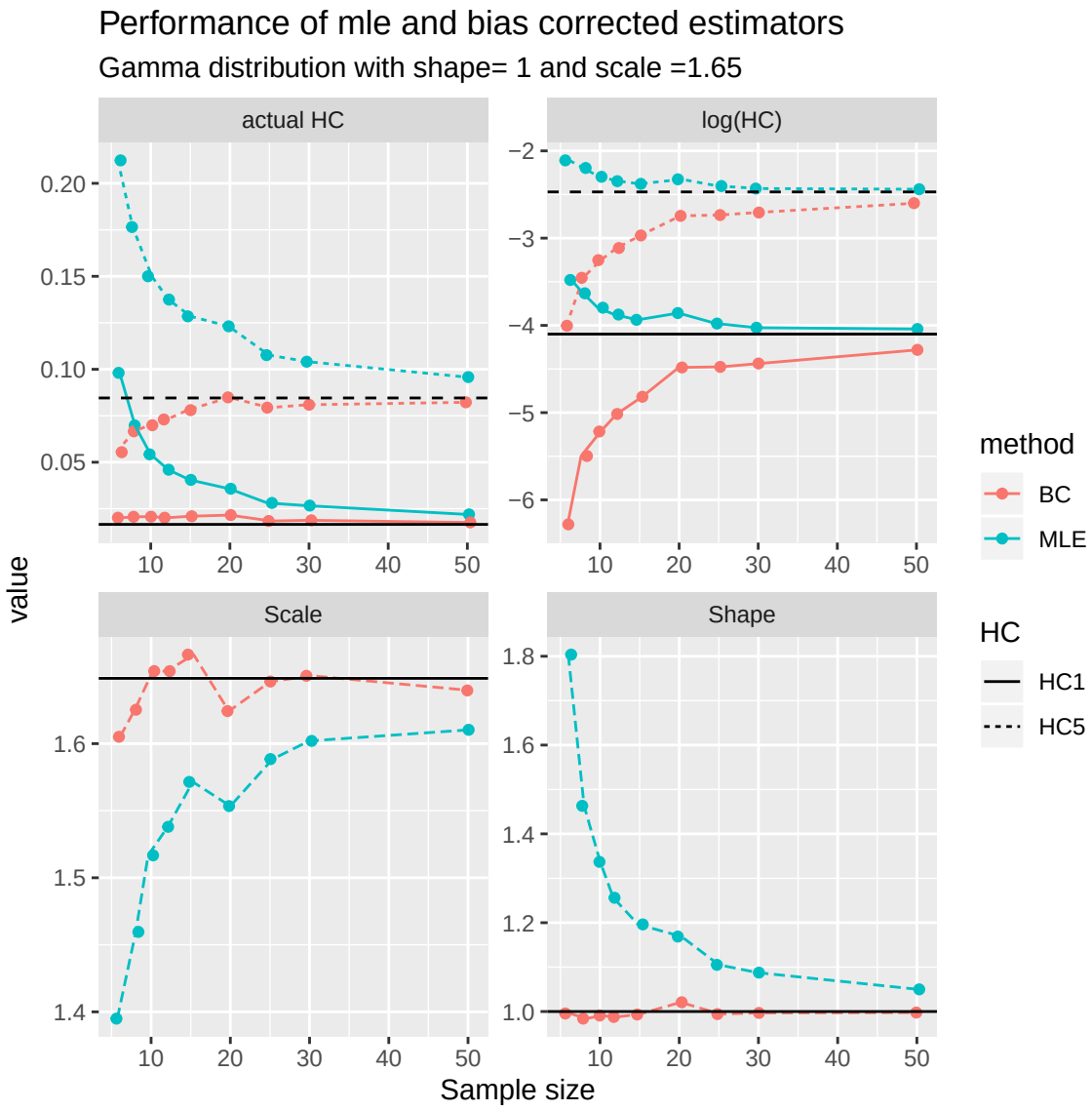
The two cumulative density functions are:



The HC5 estimated from the MLE.gamma fit is -2.76 on the logarithmic concentrations scale or 0.063 on the concentration scale. The HC5 estimated after correcting for small sample bias is -4.34 on the logarithmic concentrations scale or 0.013 on the concentration scale. The ratio of these two HC5 values is 0.205x on the concentration scale.

The differences between the HCx computed from the MLE and for the bias corrected estimates will become more pronounced for small HCx values. For example, the HC1 estimated from the MLE.gamma fit is 0.005 and the HC1 estimated using the biased corrected estimates is 0.000446 on the concentration scale. The ratio of these two values is now 0.088x.

We repeated a similar simulation study with the gamma distribution. The shape and scale parameters were chosen to match the mean and variance of the log-normal distribution used in the previous simulation study.



The MLEs are biased in small-samples for both the shape and scale (bottom row of plots) but the small-sample bias declines as sample size increases (as expected). The biases of the two parameters are in opposite directions (i.e. one bias is positive and one bias is negative). The bias corrected estimates are unbiased (as expected).

The estimated $\log HC$ (upper right plot) based on the MLE is slightly biased upwards (i.e. larger) than the true values but the bias rapidly declines with sample size (as expected). Rather surprisingly, the estimated HC5 and HC1 values using the bias-corrected estimates are biased downwards, likely an artefact of the non-linear transformation from scale and shape to the HCx.

Finally, the estimated $HC1$ and $HC5$ values are again biased upwards (upper left plot) based on the MLEs, but the estimated HCx values based on the bias-corrected estimates appear to exhibit less bias despite the bias in the $\log(HCx)$ values.

Recommendations

In cases with reasonably large sample sizes (around 15+), the small sample bias is unlikely to be of concern given the uncertainty in the endpoints actually used for the fit, and the uncertainty generated for the HCx

from the model averaging process.

The small sample bias in the estimates is expected to affect the smaller HCx values (e.g. HC1) more than larger HCx values (e.g. HC5). This is not unexpected because you are trying to extrapolate out to the extreme tails of the distribution where there is typically no data available and small changes to parameter values can have large impacts on the extreme tails.

For smaller sample sizes, a similar exercise as above can be used to estimate the impact of the small sample bias. Howeverr, for small sample sizes, this exercise may be akin to "fiddling while Rome burns", i.e, this does not change the basic problems with small sample sizes including (a) most distributions will have adequate fits and it is unlikely be possible to discriminate between distributions; and (b) extrapoloting even to a moderate tail fraction (e.g. HC5) is very, very dependent on the chosen distribution; (c) there is no data available to support even moderate extrapolation to tail proportions. Higher certainty in the estimates can only be obtained by increasing sample sizes.