Appendix for 'Testing and recommending methods for fitting size spectra to data'

This Appendix gives further mathematical derivations and explanations (Section A.1), and extra numerical results including sensitivity analyses (Section A.2).

In Section A.1 we first give analytical results that can be derived from the individual size distribution (1), including the biomass distribution function, the log-likelihood function, the probability distribution function and the random number generator. We explain the plotting of the abundance size spectrum, including showing Figure 2(h) with a non-logarithmic y-axis, and derive the bin definitions for a given data set as for the `mizer` package. We then demonstrate that the base of the logarithm of the axes does not affect the slope in regression-based binning methods, and explain a further drawback of binning. Finally we derive the likelihood function for data that are only available in binned form (the MLEbin method).

In Section A.2 we provide further numerical results, starting with investigating the estimation of $x_{\max}$ separately for each simulated data set or using one global value across all data sets. We then show the main sensitivity analyses, conducted using $x_{\max} = 10,000$, $b = -2.5$, $b = -1.5$, $b = -0.5$ and $n = 10,000$; for each analysis we show the equivalent results of Figures 2, 3, 4 and Table 2. We explain how the seed for the random number generator can potentially influence results, although we find in practice that this does not happen. Finally, we explain the subsampling of confidence intervals used for Figure 4 and related figures.

A.1 Further analytical results

A.1.1 Derivation of biomass distribution function

The total biomass of all individuals $\leq x$ is, from (1), (3) and (4),

$$\int_{x_{\min}}^{x} B(x)\mathrm{d}x = \int_{x_{\min}}^{x} nxf(x)\mathrm{d}x \tag{A.1}$$

$$= \int_{x_{\min}}^{x} nCx^{b+1}\mathrm{d}x \tag{A.2}$$

$$= nC\left[\frac{x^{b+2}}{b+2}\right]_{x_{\min}}^{x}, \quad b \neq -2 \tag{A.3}$$

$$= nC\frac{x^{b+2} - x_{\min}^{b+2}}{b+2}, \quad b \neq -2. \tag{A.4}$$

For $b = -2$ we have

$$\int_{x_{\min}}^{x} B(x)\mathrm{d}x = \int_{x_{\min}}^{x} nCx^{b+1}\mathrm{d}x \tag{A.5}$$

$$= \int_{x_{\min}}^{x} nCx^{-1}\mathrm{d}x \tag{A.6}$$

$$= nC\left[\log x\right]_{x_{\min}}^{x} \tag{A.7}$$

$$= nC(\log x - \log x_{\min}). \tag{A.8}$$

A.1.2 Log-likelihood function for bounded power-law distribution

For the PLB distribution, the log-likelihood function for $b \neq -1$ is

$$\log[\mathrm{L}(b|\text{data } \mathbf{x})] = \sum_{i=1}^{n} \log f(x_i) \tag{A.9}$$

$$= n\log\left(\frac{b+1}{x_{\max}^{b+1} - x_{\min}^{b+1}}\right) + b\sum_{i=1}^{n}\log x_i, \tag{A.10}$$

where $\mathrm{L}(b|\text{data } \mathbf{x})$ is the likelihood of a particular value of the unknown parameter b given the known data $\mathbf{x} = \{x_1, x_2, x_3, ..., x_n\} = \{x_i\}_{i=1}^{n}$, and $f(\cdot)$ is the probability density function (1); see Appendix A of Edwards (2011) for the derivation, and also Page (1968) and

White *et al.* (2008). The maximum likelihood estimate for $x_{\max}$ is the maximum value of the data. This is shown mathematically by Edwards *et al.* (2012), and basically occurs because there is no evidence that $x_{\max}$ should be greater than the maximum value of the data – the most likely value of $x_{\max}$ is therefore the largest observed value. It cannot be smaller, since this would violate the definition of $x_{\max}$ [see equation (1)]. Similarly, the maximum likelihood estimate for $x_{\min}$ is the minimum value of the data. For $b = -1$ the log-likelihood function is

$$\log[\mathrm{L}(b = -1|\text{data } \mathbf{x})] = -n \log (\log x_{\max} - \log x_{\min}) - \sum_{i=1}^{n} \log x_i. \tag{A.11}$$

A.1.3 Probability distribution function and random number generation

For the bounded power-law distribution (PLB) random numbers are generated using the inverse method. This involves first drawing a random number u from the uniform distribution over the range $[0, 1]$. Then $x = F^{-1}(u)$ is a random number from the PLB distribution, where $F(x)$ is the probability *distribution* function (Grimmett and Stirzaker, 1990), or cumulative distribution function (Bolker, 2008), corresponding to the probability density function (1). By definition, $F(x) = P(X \leq x)$, i.e. the probability that a randomly selected individual has body size $\leq x$.

The calculation of $F(x)$ using $f(x)$ from (1) is, for $b \neq -1$:

$$\begin{aligned} F(x) &= \mathrm{P}(X \leq x) && \text{(A.12)} \\ &= \int_{x_{\min}}^{x} f(x)\mathrm{d}x && \text{(A.13)} \\ &= \int_{x_{\min}}^{x} \frac{b+1}{x_{\max}^{b+1} - x_{\min}^{b+1}}\, x^{b}\mathrm{d}x && \text{(A.14)} \\ &= \frac{\cancel{b+1}}{x_{\max}^{b+1} - x_{\min}^{b+1}} \left[\frac{x^{b+1}}{\cancel{b+1}}\right]_{x_{\min}}^{x} && \text{(A.15)} \\ &= \frac{x^{b+1} - x_{\min}^{b+1}}{x_{\max}^{b+1} - x_{\min}^{b+1}}. && \text{(A.16)} \end{aligned}$$

Then setting $u = F(x)$ and rearranging for x gives

$$\begin{aligned} u &= \frac{x^{b+1} - x_{\min}^{b+1}}{x_{\max}^{b+1} - x_{\min}^{b+1}} && \text{(A.17)} \\ \left(x_{\max}^{b+1} - x_{\min}^{b+1}\right) u &= x^{b+1} - x_{\min}^{b+1} && \text{(A.18)} \\ u x_{\max}^{b+1} + (1-u) x_{\min}^{b+1} &= x^{b+1} && \text{(A.19)} \\ x &= \left[u x_{\max}^{b+1} + (1-u) x_{\min}^{b+1}\right]^{1/(b+1)}. && \text{(A.20)} \end{aligned}$$

For $b = -1$:

$$\begin{aligned} F(x) &= \int_{x_{\min}}^{x} \frac{1}{\log x_{\max} - \log x_{\min}}\, x^{-1}\mathrm{d}x && \text{(A.21)} \\ &= \frac{1}{\log x_{\max} - \log x_{\min}} \left[\log x\right]_{x_{\min}}^{x} && \text{(A.22)} \\ &= \frac{\log x - \log x_{\min}}{\log x_{\max} - \log x_{\min}} && \text{(A.23)} \\ &= \frac{\log(x/x_{\min})}{\log(x_{\max}/x_{\min})}. && \text{(A.24)} \end{aligned}$$

Then setting $u = F(x)$ and rearranging for x gives

$$u = \frac{\log(x/x_{\min})}{\log(x_{\max}/x_{\min})} \tag{A.25}$$

$$u \log\left(\frac{x_{\max}}{x_{\min}}\right) = \log\left(\frac{x}{x_{\min}}\right) \tag{A.26}$$

$$\log\left(\frac{x_{\max}}{x_{\min}}\right)^{u} = \log\frac{x}{x_{\min}} \tag{A.27}$$

$$\left(\frac{x_{\max}}{x_{\min}}\right)^{u} = \frac{x}{x_{\min}} \tag{A.28}$$

$$x = x_{\max}^{u}\, x_{\min}^{1-u}. \tag{A.29}$$

For the LCD method, the estimated fraction of values $\geq x$ for the fitted distribution is $\mathrm{P}(X \geq x)$, which is $1-\mathrm{P}(X < x) = 1-\mathrm{P}(X \leq x) = 1-F(x)$, since $\mathrm{P}(X < x) = \mathrm{P}(X \leq x)$ for continuous distributions. The resulting slope for an unbounded power-law ($x_{\max} \to \infty$ with $b < -1$) is $b+1$, since we have $\log(1 - F(x)) = (b+1)\log x - (b+1)\log x_{\min}$. For the bounded power-law this result is approximately correct where x is small enough relative to $x_{\max}$.

For the MLE method the solid red curve in Figures 2(h) and 6(b) is plotted as $(1 - F(x))n$. This characterises the *abundance size spectrum*, and is more informative than plotting just $1-F(x)$, shown for the LCD method in Figure 2(g), because it includes the sample size, n. It directly shows how abundance varies with body size, and is related to the abundance density function $N(x)$ defined in (3):

$$(1 - F(x))n = n - nF(x) \tag{A.30}$$

$$= n - n\int_{x_{\min}}^{x} f(x)\mathrm{d}x \tag{A.31}$$

$$= n - \int_{x_{\min}}^{x} nf(x)\mathrm{d}x \tag{A.32}$$

$$= n - \int_{x_{\min}}^{x} N(x)\mathrm{d}x. \tag{A.33}$$

For the MLE method, the red curve in Figure 2(h) does not pass through the maximum data point of 399; by definition it cannot. The maximum likelihood estimate for $x_{\max}$ is

this maximum value of 399 (Section A.1.2). The red curve is plotted as $(1-F(x))n$, which at $x = x_{\max}$ this equals 0, because $F(x_{\max}) = 1$ from equation (A.16). The logarithmic scale of the y-axis in Figure 2(h) means that $(1 - F(x_{\max}))n = 0$ cannot be reached (since $\log 0 \to -\infty$), and so the red curve asymptotes to the vertical line $x = x_{\max}$. The logarithmic y-axis is used because the data are plotted in the same way as for the LCD method.

However, the MLE method does not depend on the axes used for plotting, and so in Figure A.1 we re-plot Figure 2(h) but without logging the y-axis. This graphically shows the good fit of the model, including through the $x = 399$ value. The red curve ends at $x = x_{\max} = 399$ because $x_{\max}$ is, by definition, the maximum valid value of the fitted model. The red curve ends at 0 on the y-axis, but the data point has a value of 1 (since there is just the one value ≥ 399, namely 399). This difference does not show up in Figure A.1, but is magnified by the logarithmic y-axis in Figure 2(h).

This shows that the apparent poor fit to the $x = 399$ point in Figure 2(h), which the reader's eye can get drawn to, is an artefact of the logarithmic y-axis. The logarithmic y-axis in Figure 2(h) is recommended because this is how researchers are used to seeing the data when using the LCD method. And the fitted PLB model is straight over most of the plot, which is analogous to the straight lines seen for the other methods that researchers are used to. However, researchers may also wish to plot results in the form of Figure A.1 to aid understanding and interpretation.

A.1.4 Bin definitions in `mizer`

In the R package `mizer` (Scott *et al.*, 2014) the user specifies bins by giving the lower bounds of the smallest and largest bins (`min_w` and `max_w`) and also the number of bins (`no_w`). The lower bounds of the bins are then calculated as

$$\texttt{w} < -10\texttt{\^{}}(\texttt{seq}(\texttt{from} = \texttt{log10}(\texttt{min\_w}), \texttt{to} = \texttt{log10}(\texttt{max\_w}), \texttt{length.out} = \texttt{no\_w})) \quad \text{(A.34)}$$

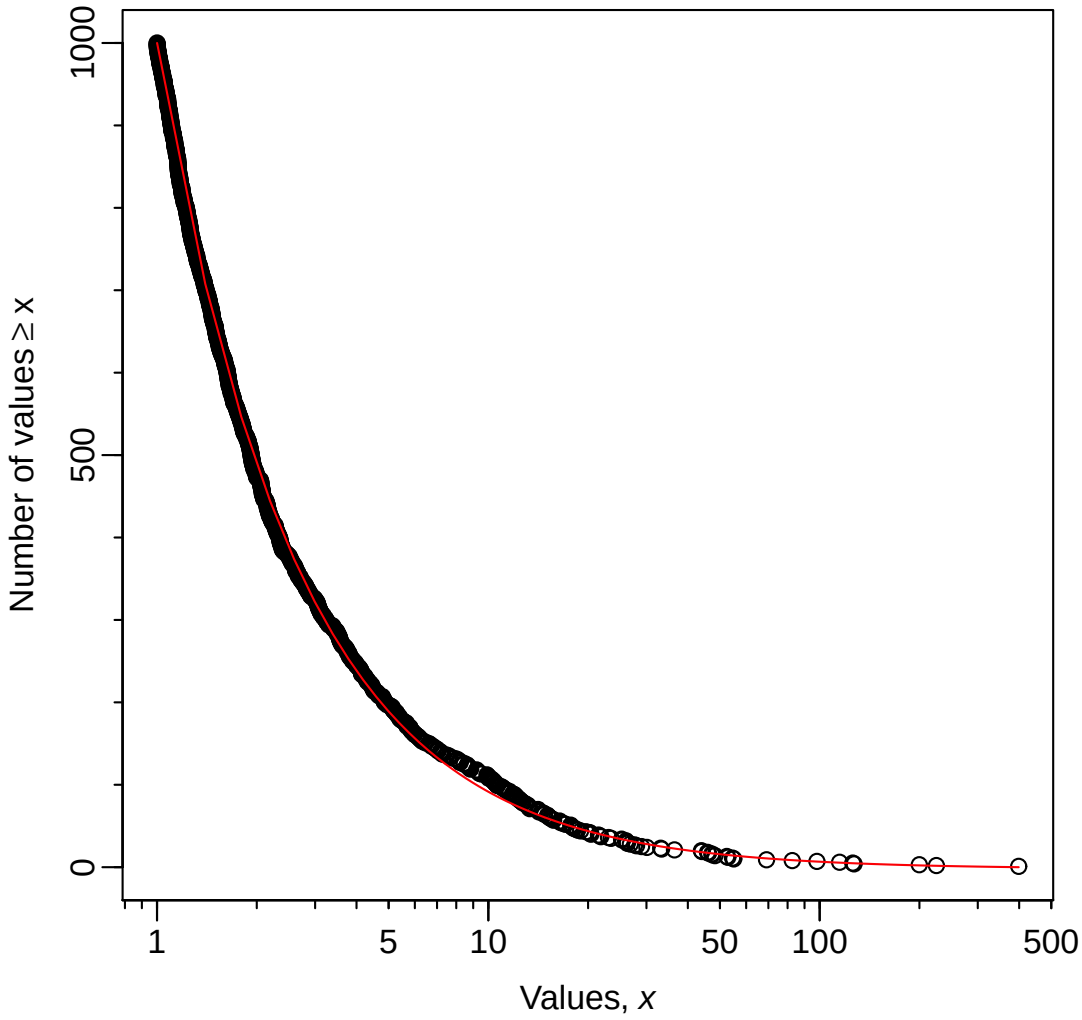

Figure A.1: Plot of Figure 2(h) but with a non-logarithmic y-axis. This shows that the apparent poor fit of the model to the maximum data point at $x = 399$ in Figure 2(h) is an artefact of the logarithmic y-axis, and that this data point is consistent with the model.

and the final bin is then given the same width (on an arithmetic scale) as the penultimate bin (F. Scott, pers. comm). Thus the lower bounds of the bins are equally spaced on a $\log_{10}$ scale, as are the bin widths (except for the final bin).

The counts in each bin are calculated, and the slope of an abundance spectra is calculated as the slope of the linear regression of log(counts) against log(`w`). There is an option to calculate a biomass spectra, for which counts in each bin are multiplied by the lower bound of that bin, and then the linear regression is performed.

For a given data set, we wish to apply the method. Let $x_{\min}$ and $x_{\max}$ be the minimum and maximum values in the data; we wish to use these values, respectively, for the lower bound of the lowest bin and the upper bound of the highest bin (so that our bins exactly span the data), for a total of k (= `no_w`) bins. Given $x_{\min}, x_{\max}$ and $k > 1$ we therefore need to calculate the value of `max_w` (the lower bound of the largest bin) to go into the `mizer` bin calculation (A.34).

The first $k-1$ bins are equally spaced on a $\log_{10}$ scale; define $\log_{10}\beta$ to be the constant bin width on the $\log_{10}$ scale, requiring that $\log_{10}\beta > 0$ (and thus $\beta > 1$). Then the first three bin breaks on the $\log_{10}$ scale are:

$$\log_{10} x_{\min} \tag{A.35}$$

$$\log_{10} x_{\min} + \log_{10}\beta = \log_{10}\beta x_{\min} \tag{A.36}$$

$$\log_{10} x_{\min} + 2\log_{10}\beta = \log_{10}\beta^2 x_{\min} \tag{A.37}$$

$$\log_{10} x_{\min} + 3\log_{10}\beta = \log_{10}\beta^3 x_{\min}. \tag{A.38}$$

On the arithmetic scale, all the bin breaks are thus

$$x_{\min} \quad (= \texttt{min\_w}) \tag{A.39}$$

$$\beta x_{\min} \tag{A.40}$$

$$\beta^2 x_{\min} \tag{A.41}$$

$$\beta^3 x_{\min} \tag{A.42}$$

$$... \tag{A.43}$$

$$\beta^{k-2} x_{\min} \tag{A.44}$$

$$\beta^{k-1} x_{\min} \quad (= \texttt{max\_w}) \tag{A.45}$$

$$x_{\max}. \tag{A.46}$$

Thus, the constant $\log_{10} \beta$ bin width on the $\log_{10}$ scale translates to arithmetic bin widths progressively increasing by a multiple of $\beta > 1$. We wish to solve for β (given that we know $x_{\min}$ and $x_{\max}$ from the data and we specify k) such that the final two bins have equal arithmetic width, i.e. such that

$$x_{\max} - \beta^{k-1} x_{\min} = \beta^{k-1} x_{\min} - \beta^{k-2} x_{\min} \tag{A.47}$$

$$0 = 2\beta^{k-1} - \beta^{k-2} - \frac{x_{\max}}{x_{\min}} \tag{A.48}$$

$$0 = \beta^{k-2}(2\beta - 1) - \frac{x_{\max}}{x_{\min}}. \tag{A.49}$$

This cannot be solved algebraically for β, and thus needs to be solved numerically. We use the R function `nlm` to minimise the log of the right-hand side of (A.49) to obtain β. Bin breaks are then given by (A.39)-(A.46).

A.1.5 When determining slope of binned data on logarithmic axes, the base of the logarithm does not matter

A $\log_2$ scale has sometimes been used to bin data in order to increase the number of bins (compared to using $\log_{10}$ bins), but the resulting regressions were fitted based on $\log_{10}$ axes

(presumably because $\log_{10}$ is more intuitive); e.g. Blanchard *et al.* (2005); Jennings *et al.* (2007). The choice of using $\log_2$ or $\log_{10}$ for plotting and regression does not affect the resulting calculated slope of the spectrum. For any quantity X,

$$\log_2 X = \frac{\log_{10} X}{\log_{10} 2} = \frac{\log_{10} X}{0.301}, \tag{A.50}$$

and similarly for any quantity Y,

$$\log_2 Y = \frac{\log_{10} Y}{\log_{10} 2} = \frac{\log_{10} Y}{0.301}. \tag{A.51}$$

The resulting slope of the values plotted on $\log_2$ axes is

$$\frac{\log_2 Y}{\log_2 X} = \frac{\log_{10} Y}{0.301} \cdot \frac{0.301}{\log_{10} X} = \frac{\log_{10} Y}{\log_{10} X}, \tag{A.52}$$

which equals the slope on $\log_{10}$ axes. Thus the choice of logarithmic base to use for the axes does not matter in the calculation of slopes. But note that the choice of logarithmic base for binning of the data will affect the resulting slope, as shown by Vidondo *et al.* (1997).

A.1.6 Binning

Even though we stated the number of bins used, e.g. 8 for the Llin method, this can still give an unambiguous result depending on how the statistical software defines the bin breaks. For example, the range of the simulated data set is $[1, 399]$, and so R, quite reasonably, selected bin breaks of 0, 50, 100, ..., 400, to give 8 bins. However, another choice is having 8 bins that do not extend beyond the data (i.e. bin widths of $(399 - 1)/8 = 49.75$, namely 1, 50.75, 100.5, ..., 399). This will only give a slightly different answer in this case. However, if the simulated data set is restricted to values > 40, the `hist(x, breaks=8)` command in R still selects bin breaks as 0, 50, 100, ..., 400. But in a plot, the first bin will appear to cover values between 0 and 50, but in fact the data only has values > 40. Thus, this bin may appear to be undersampled, but this is really an artefact of the bin-break selection. This is another reason why it is desirable to avoid binning, in particular when estimating parameter values.

A.1.7 Likelihood function for the MLEbin method

Here we derive the likelihood function for the MLEbin method, which is to be used when fitting a bounded power law to data when the data are only available in binned form. We extend and generalise the derivations from Edwards *et al.* (2007) and Edwards (2011) to allow for any type of binning. The aim is to obtain the likelihood functions to calculate the maximum likelihood estimate for the exponent b in (1).

Consider the data to consist of a count (number of individuals) d_j in each bin $j = 1, 2, 3, ..., J$, defining J to be the index of the final bin. Let bin j cover the values of x (weight or length) in the interval $[w_j, w_{j+1})$, such that $w_1, w_2, ..., w_{J+1}$ define the bin breaks. For example, bin $j = 5$ goes from w_5 to w_6. For bin $j = J$ the interval is $[w_J, w_{J+1}]$, which includes the upper bound. The sample size (total number of individuals) is $n = \sum_{j=1}^{J} d_j$, and we assume that the first and last bins each have at least one individual in them (i.e. $d_1, d_J > 0$).

Similar calculations to those by Edwards *et al.* (2012) for unbinned data show that the known w_1 and w_{J+1} are the maximum likelihood estimates for $x_{\min}$ and $x_{\max}$, respectively. So by setting $x_{\min} = w_1$ and $x_{\max} = w_{J+1}$ we now only need to calculate the maximum likelihood estimate of b.

For a single data value, the probability of being in bin j given the parameter b (assume for now that $b \neq -1$) is

$$\begin{aligned}
\mathrm{P}(\text{being in bin } j|b) &= \int_{w_j}^{w_{j+1}} Cx^b \mathrm{d}x && \text{(A.53)}\\
&= \frac{C}{b+1}\left[x^{b+1}\right]_{w_j}^{w_{j+1}} && \text{(A.54)}\\
&= \frac{C}{b+1}\left[w_{j+1}^{b+1} - w_j^{b+1}\right] && \text{(A.55)}\\
&= \frac{w_{j+1}^{b+1} - w_j^{b+1}}{x_{\max}^{b+1} - x_{\min}^{b+1}}, && \text{(A.56)}\\
&= \frac{w_{j+1}^{b+1} - w_j^{b+1}}{w_{J+1}^{b+1} - w_1^{b+1}}, && \text{(A.57)}
\end{aligned}$$

substituting C from (2) to obtain (A.56). Note that these probabilities sum to 1 (because

a data value must be in one of the bins):

$$\sum_{j=1}^{J} \mathrm{P}(\text{being in bin } j|b) = \sum_{j=1}^{J} \frac{w_{j+1}^{b+1} - w_j^{b+1}}{w_{J+1}^{b+1} - w_1^{b+1}} \tag{A.58}$$

$$= \frac{1}{w_{J+1}^{b+1} - w_1^{b+1}} \sum_{j=1}^{J} \left(w_{j+1}^{b+1} - w_j^{b+1}\right) \tag{A.59}$$

$$= \frac{1}{w_{J+1}^{b+1} - w_1^{b+1}} \left(\sum_{j=1}^{J} w_{j+1}^{b+1} - \sum_{j=1}^{J} w_j^{b+1}\right) \tag{A.60}$$

$$= \frac{1}{w_{J+1}^{b+1} - w_1^{b+1}} \left(\sum_{j=1}^{J-1} w_{j+1}^{b+1} + w_{J+1}^{b+1} - \sum_{j=2}^{J} w_j^{b+1} - w_1^{b+1}\right) \tag{A.61}$$

$$= \frac{1}{w_{J+1}^{b+1} - w_1^{b+1}} \left(\cancel{\sum_{j=1}^{J-1} w_{j+1}^{b+1}} + w_{J+1}^{b+1} - \cancel{\sum_{j=1}^{J-1} w_{j+1}^{b+1}} - w_1^{b+1}\right) \tag{A.62}$$

$$= 1. \tag{A.63}$$

Given the counts $\{d_j\}_{j=1}^{J}$ in each bin, the multinomial log-likelihood function (Lawless, 2003) is

$$l(b|d_1, d_2, d_3, ..., d_J) = \sum_{j=1}^{J} d_j \log\left[\mathrm{P}(\text{being in bin } j|b)\right] \tag{A.64}$$

$$= \sum_{j=1}^{J} d_j \log\left(\frac{w_{j+1}^{b+1} - w_j^{b+1}}{w_{J+1}^{b+1} - w_1^{b+1}}\right) \tag{A.65}$$

$$= \sum_{j=1}^{J} d_j \left(\log\left|w_{j+1}^{b+1} - w_j^{b+1}\right| - \log\left|w_{J+1}^{b+1} - w_1^{b+1}\right|\right) \tag{A.66}$$

$$= \sum_{j=1}^{J} d_j \log\left|w_{j+1}^{b+1} - w_j^{b+1}\right| - \sum_{j=1}^{J} d_j \log\left|w_{J+1}^{b+1} - w_1^{b+1}\right| \tag{A.67}$$

$$= \sum_{j=1}^{J} d_j \log\left|w_{j+1}^{b+1} - w_j^{b+1}\right| - \log\left|w_{J+1}^{b+1} - w_1^{b+1}\right| \sum_{j=1}^{J} d_j \tag{A.68}$$

$$= \sum_{j=1}^{J} d_j \log\left|w_{j+1}^{b+1} - w_j^{b+1}\right| - n \log\left|w_{J+1}^{b+1} - w_1^{b+1}\right| \tag{A.69}$$

$$= -n \log\left|w_{J+1}^{b+1} - w_1^{b+1}\right| + \sum_{j=1}^{J} d_j \log\left|w_{j+1}^{b+1} - w_j^{b+1}\right|. \tag{A.70}$$

The two terms inside the absolute symbols $|\cdot|$, i.e. $w_{j+1}^{b+1} - w_j^{b+1}$ and $w_{J+1}^{b+1} - w_1^{b+1}$, are both positive for $b < -1$ and both negative for $b > -1$ (because $w_{j+1} > w_j$ and $w_{J+1} > w_1$ by definition), such that taking their absolute values ensures that (A.65) and (A.66) are equivalent. Equation (A.70) cannot be analytically solved to give the maximum likelihood estimate of b (by differentiating with respect to b and setting to 0), and so numerical methods are required.

For the case where $b = -1$, we have $C = 1/(\log x_{\max} - \log x_{\min}) = 1/(\log w_{J+1} - \log w_1)$, and

$$\begin{aligned} \mathrm{P}(\text{being in bin } j | b = -1) &= \int_{w_j}^{w_{j+1}} C x^{-1} \mathrm{d}x & \text{(A.71)} \\ &= C \left[\log x\right]_{w_j}^{w_{j+1}} & \text{(A.72)} \\ &= \frac{\log w_{j+1} - \log w_j}{\log w_{J+1} - \log w_1}. & \text{(A.73)} \end{aligned}$$

The log-likelihood function is then just

$$\begin{aligned} l(b = -1 | \text{data}) &= \sum_{j=1}^{J} d_j \log \left[\mathrm{P}(\text{being in bin } j | b = -1)\right] & \text{(A.74)} \\ &= \frac{1}{\log w_{J+1} - \log w_1} \sum_{j=1}^{J} d_j \left(\log w_{j+1} - \log w_j\right). & \text{(A.75)} \end{aligned}$$

In Figure 5 we show results from setting bin breaks at 1, 2, 4, 8, ..., 1024 (so we have $w_1 = 1, w_2 = 2, w_3 = 4, w_4 = 8, ..., w_{J+1} = 1024$). Note that for some simulated data sets the final bin break will be 512 or lower if there are no simulated values > 512 (or > 256 or > 128 etc.), as in Figures 1 and 2. But our approach (and R code) is applicable to any set of bin breaks. Thus, it can be used for any data set for which measurements are only available in binned form, including historical data sets.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. 2nd ed., Wiley series in Probability and Statistics, Wiley, New Jersey.

A.2 Further results from the numerical simulations in the main text

A.2.1 Estimating $x_{\max}$ in each of the 10,000 simulated data sets

For each of the 10,000 simulated data sets, for the MLE method we calculated the MLE of $x_{\max}$ separately for each data set (the MLE simply being the maximum value of the data in that data set). Figure A.2 shows that there is no relationship between the MLE of b and the MLE of $x_{\max}$. So, for example, for a simulated data set with a maximum value of 200, the MLE of $x_{\max}$ is 200, yet this seems to have no influence on the MLE method's estimate of b.

A.2.2 The MLEfix method

However, for real data sets we might want to fix $x_{\max}$ across data sets. For example, for body masses that are sampled similarly from year-to-year, we might set $x_{\max}$ to be the largest body mass seen across all years, since we know such a value is attainable, rather than estimate $x_{\max}$ separately for each year. We call this the MLEfix method – instead of estimating $x_{\max}$ as the maximum value of the data set being fitted, we fix it to some value.

In Figure A.3(b) we show the equivalent results to Figure 3(h) for the MLEfix method (with the original MLE method's results in Figure A.3(a) for comparison). The MLEfix method fixes $x_{\max}$ to the true value of 1,000, and consequently performs marginally better (51% rather than 44% of the estimated values lie below the true value of -2, though the 5 and 95% quantiles are slightly worse), but otherwise both methods perform well. A similar conclusion is reached from the confidence intervals in Figure A.4.

Figure A.5 repeats Figure A.2 but for the MLEfix method. The x-axis is now labelled as the maximum of x from each data set (although the values are the same as in Figure A.2

because these equal the MLE of $x_{\max}$). There is a statistically significant trend in the MLE of b with respect to the maximum of the data set. This make sense – the MLEfix method is being told that $x_{\max} = 1,000$. But for a particular data set that has a maximum x of, say, only 200, there are no values between 200 and 1,000, and so the method will tend to produce a slightly steeper power law than if higher values had been observed. Similar results were found for real data in Supplementary Tables 1 and 2 of Edwards *et al.* (2007). However, here, while the estimated trend in the MLE of b with respect to the maximum x is significantly different to zero, it is small. Its value of 2.4×10^{-5} equates to an overall increase in b of only 0.024 for an 1,000-fold increase in the maximum of x. Such a minor change is smaller than the general variability of the estimated b in Figure A.5, and smaller than the uncertainty accounted for in the confidence intervals in Figure A.4 (which have a minimum width of 0.11 across both methods). Thus, we conclude that while the trend in Figure A.5 is statistically significant, it is not ecologically significant.

Thus, for the simulated data sets it appears that the MLE and MLEfix methods yield only minor numerical differences in results, which would not translate into meaningful changes in ecological interpretation. For real data such sensitivity could be tested, and the final choice of method justified depending on the type of data.

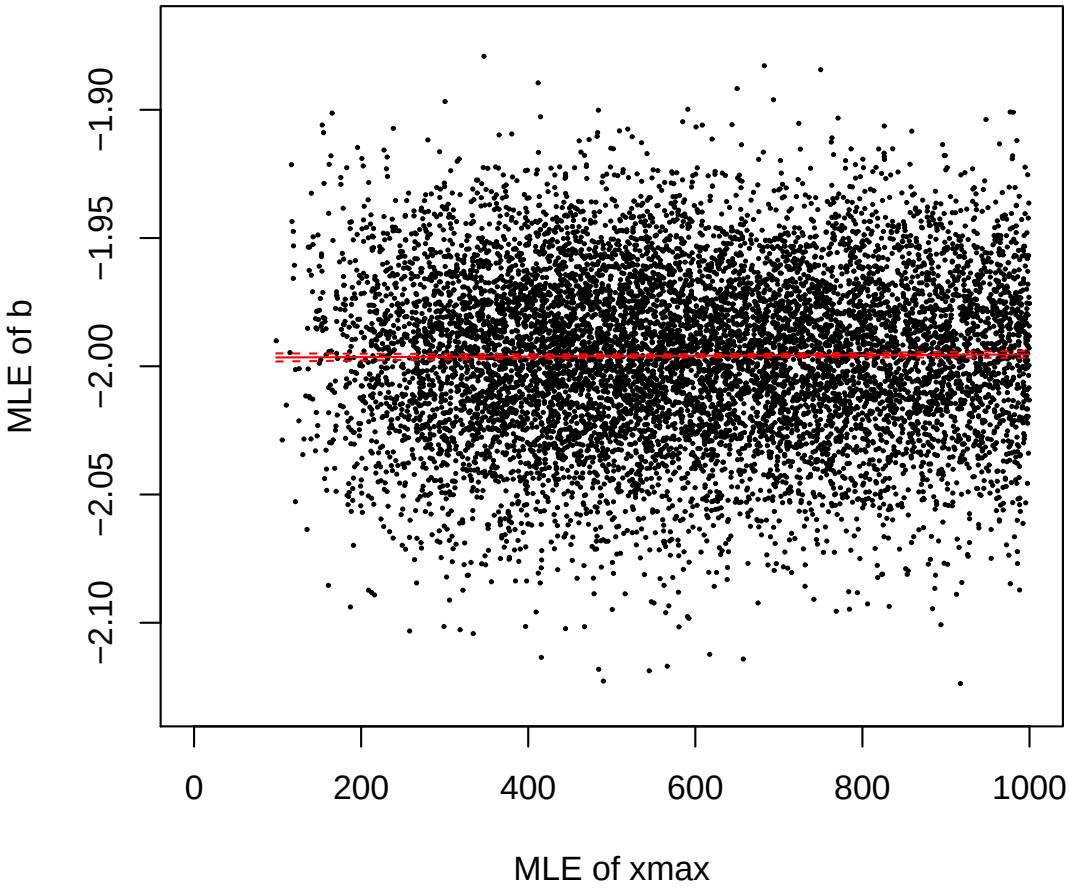

Figure A.2: Relationship between maximum likelihood estimate (MLE) of b and MLE of $x_{\max}$ for each of the 10,000 simulated data sets used in Figure 3. The red line is a fitted linear regression (with confidence intervals), and the fitted slope of 1.3×10^{-6} (standard error 1.5×10^{-6}) is not significantly different from zero ($p = 0.39, R^2 < 10^{-4}$). The MLE of $x_{\max}$ for each data set is simply the maximum value in that data set. It appears here that a lower maximum value does not influence the estimation of b.

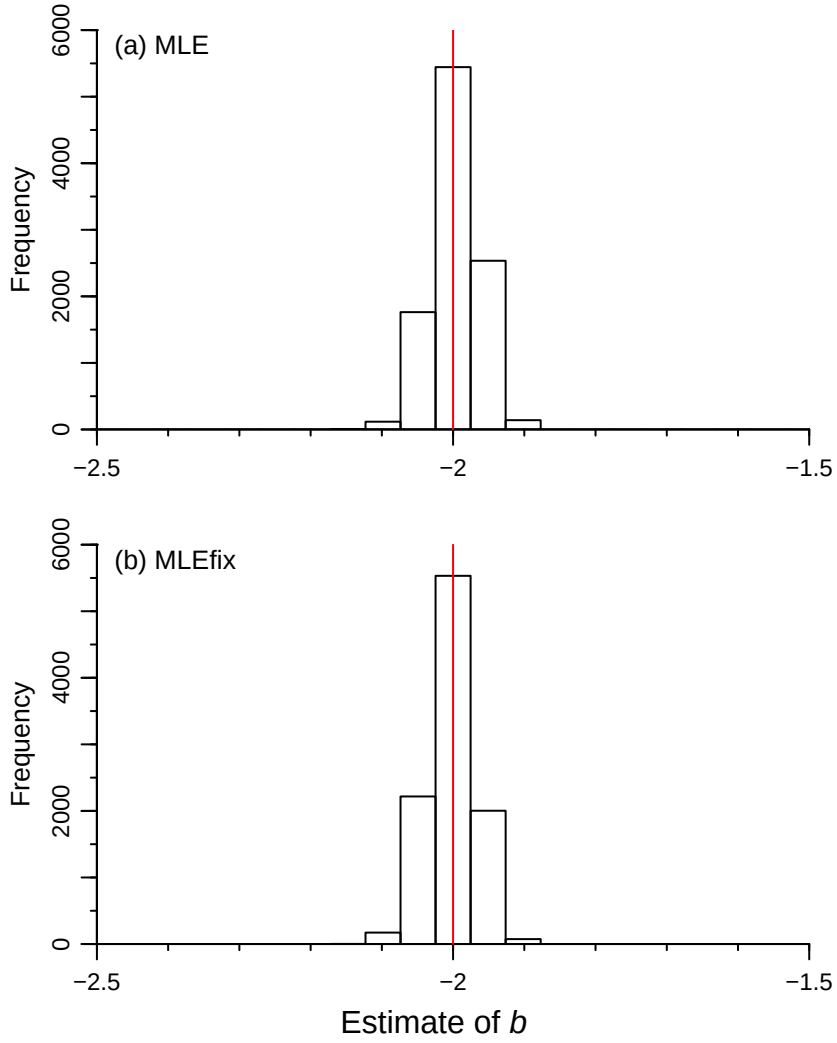

Figure A.3: As for Figure 3 but for just the MLE and MLEfix methods. The MLEfix method fixes $x_{\max} = 1,000$ rather than estimating it separately for each of the 10,000 simulated data sets. The histograms looks similar, and the statistics for the estimated b (as in Table 2) for the MLEfix method are: 5% quantile is -2.06, median and mean are both -2.00, 95% quantile is -1.95, and 51% of the values lie below the true value of -2. So the statistics are very similar to the MLEfix method, except for the final one which is closer to the desired 50% than all methods.

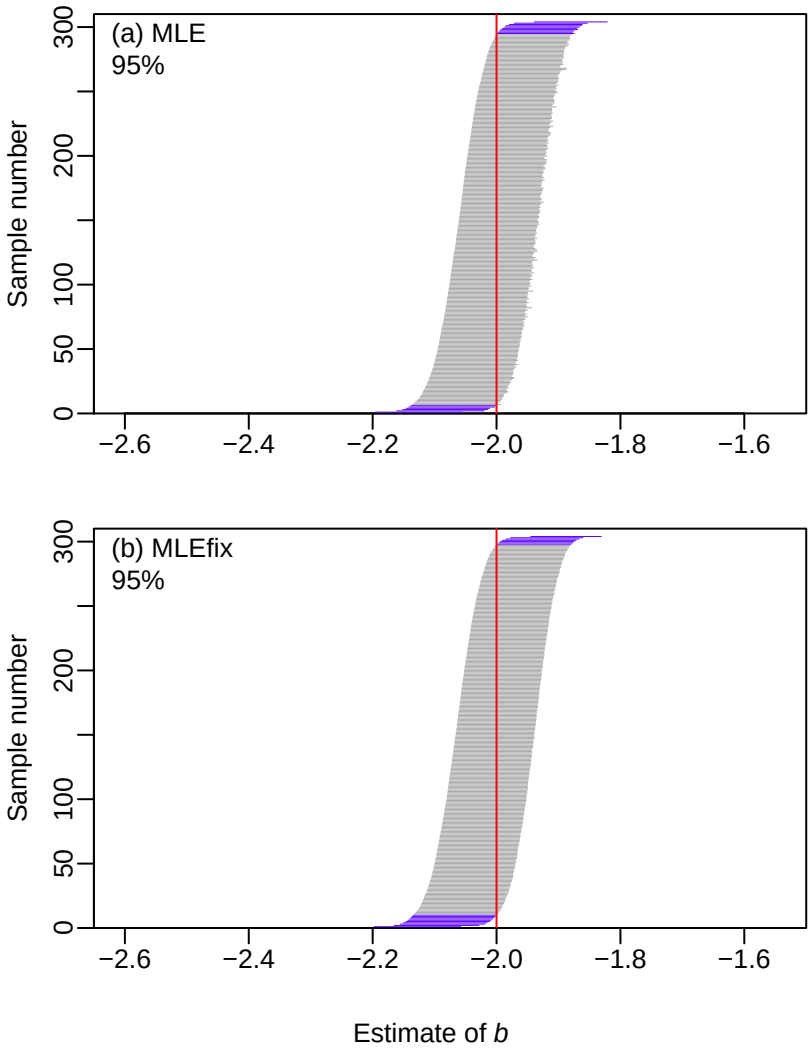

Figure A.4: As for Figure 4 but for just the MLE and MLEfix methods. The confidence intervals for both methods demonstrate the ideal observed coverage of 95%. Thus the MLEfix method performs just as well as the MLE method.

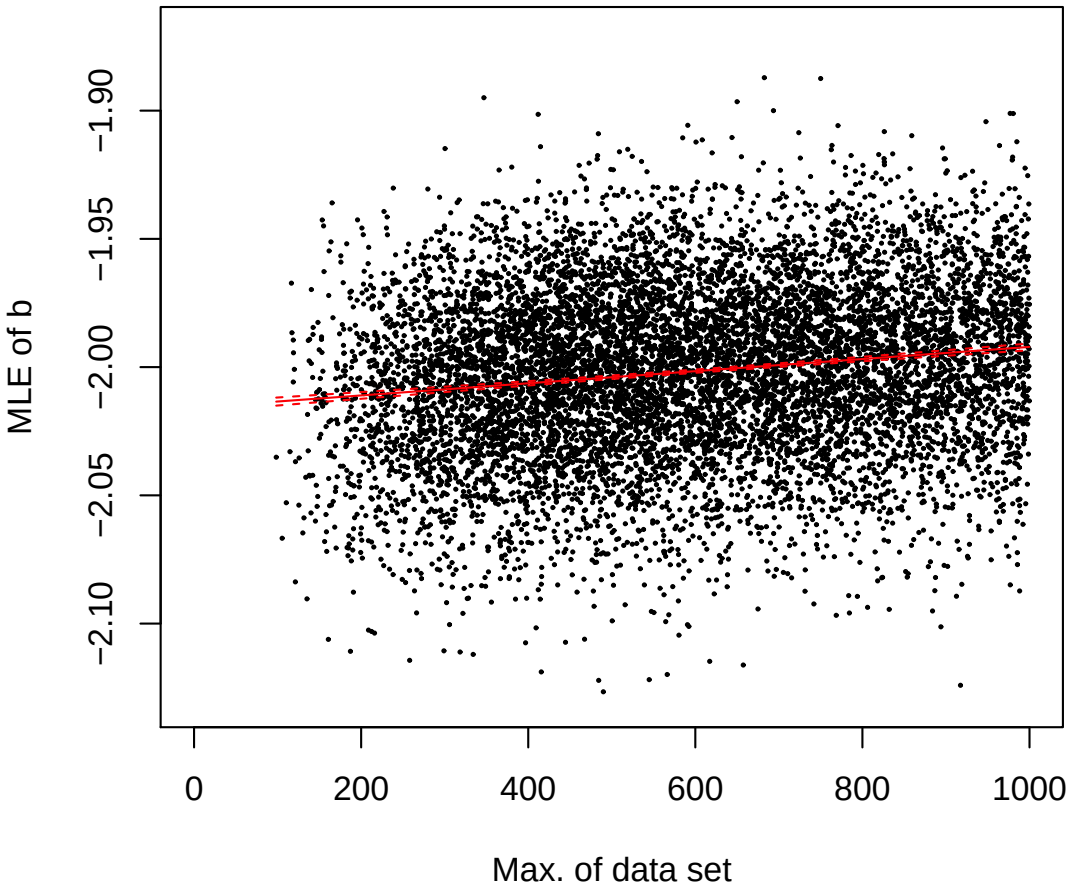

Figure A.5: As for Figure A.2, using the same 10,000 simulated data sets, but for the MLEfix method. The estimated slope of 2.4×10^{-5} (s.e. 1.5×10^{-6}) is now significantly different from zero ($p =< 10^{-15}$, $R^2 = 0.025$). So now a lower maximum realised value for a data set does statistically influence the estimation of b, as expected, although ecologically such a small slope is not important.

A.2.3 Increasing $x_{\max}$ to 10,000 in the simulated data sets

The results in the main text (except for the gold histograms in Figure 3) used simulated data sets with $x_{\max} = 1,000$. Here we test the sensitivity to that choice, and show the equivalent results for $x_{\max} = 10,000$, still with $x_{\min} = 1$.

For $x_{\max} = 10,000$, Figure A.6 shows the standard histogram. Figure A.7 shows the resulting estimates of slopes and/or b for a single data set, and Figure 3 already shows the estimated values of b for 10,000 randomly generated data sets (the statistical results are given in Table A.1). The histograms of estimated b in Figure 3 for the LBmiz, LBbiom and LBNbiom methods have drifted to the right compared to Figure 3, showing that they are less accurate with the increase in $x_{\max}$. The LCD method remains fairly accurate but with 40%, rather than 59%, of estimated values of b being < -2. The confidence intervals in Figure A.8 show worse observed coverage for the LBmiz, LBbiom and LBNbiom methods, with the MLE method again showing the desired 95% observed coverage.

For the MLEfix method and $x_{\max} = 10,000$, the histogram of estimates of b and the plot of confidence intervals are essentially identical to those in Figures A.3 and A.4 for $x_{\max} = 1,000$ and are not shown.

The equivalent figures to A.2 and A.5 are shown in Figure A.9, to investigate the effects of estimating $x_{\max}$ as the maximum data value for each simulated data set (the MLE method), or fixing it to $x_{\max} = 10,000$ (the MLEfix method). Although the regression slopes are significantly different to zero for both methods, the magnitudes of the change in b across the range 1-10,000 are only 0.0051 and 0.024, respectively. So, as with $x_{\max} = 1,000$, the trends are statistically but not ecologically significant.

Table A.1: As for Table 2 but for simulations with $x_{\max} = 10,000$, corresponding to the gold histograms in Figure 3.

Method	Slope represents	5% quantile	Median	Mean	95% quantile	Percentage below -2
Llin	–	-0.02	0.00	-0.01	0.00	0
LT	b	-3.02	-2.44	-2.48	-2.06	98
LTplus1	b	-2.74	-2.18	-2.21	-1.77	72
LBmiz	$b+1$	-2.08	-1.94	-1.94	-1.82	24
LBbiom	$b+2$	-2.07	-1.93	-1.93	-1.80	20
LBNbiom	$b+1$	-2.07	-1.93	-1.93	-1.80	20
LCD	$b+1$	-2.06	-1.99	-1.99	-1.92	40
MLE	b	-2.05	-1.99	-2.00	-1.94	43

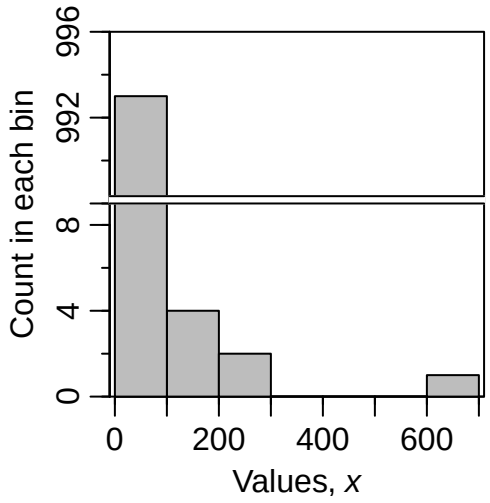

Figure A.6: Standard histogram of a random sample of 1,000 values from a bounded power-law distribution with known exponent $b = -2$, $x_{\min} = 1$ and $x_{\max} = 10,000$. So as for Figure 1 but with $x_{\max}$ increased ten-fold. Note that we specified eight bins again for the histogram, but the `hist()` command in `R` selected only seven for this data set (to have widths of 100), of which three are empty.

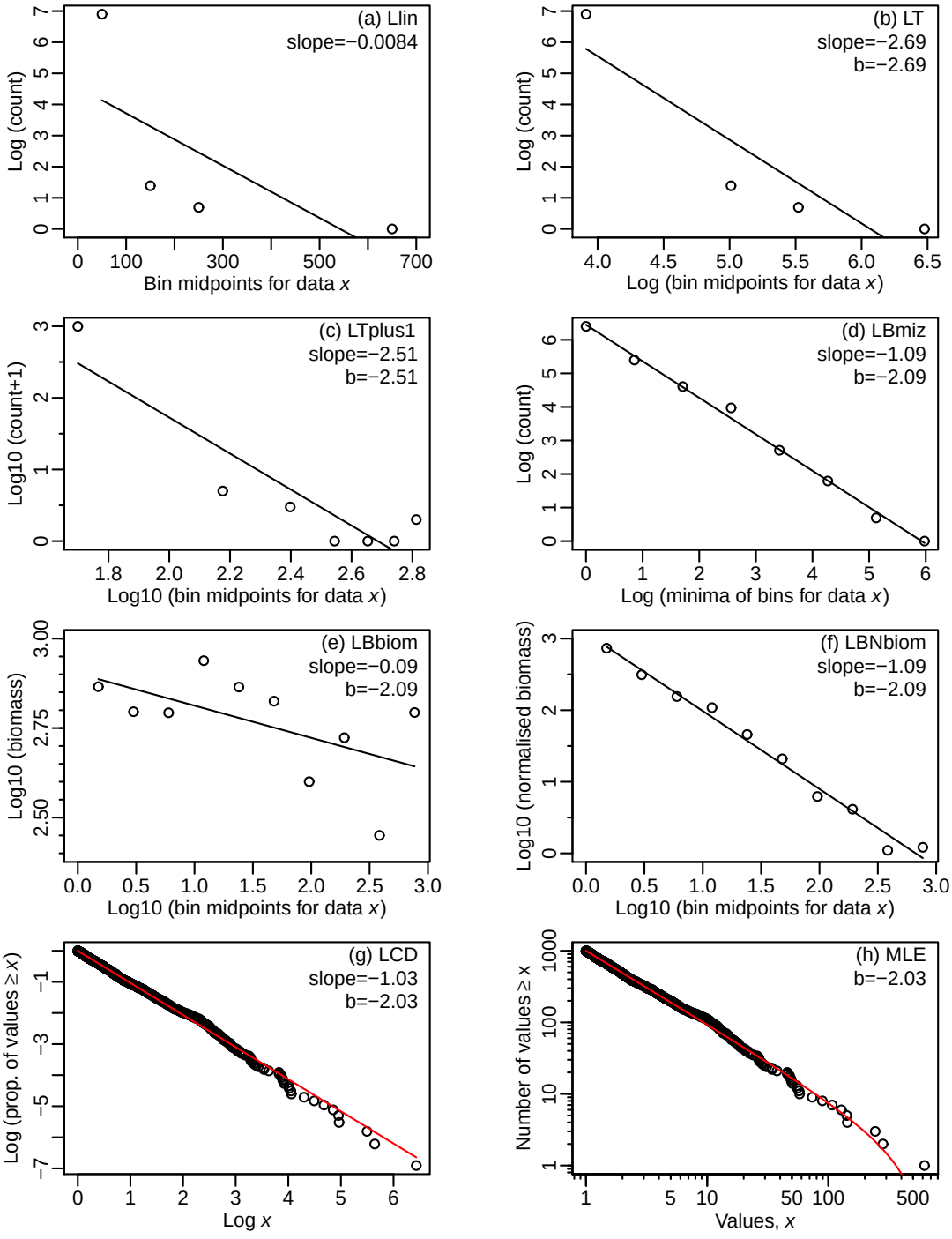

Figure A.7: As for Figure 2 but with $x_{\max} = 10,000$.

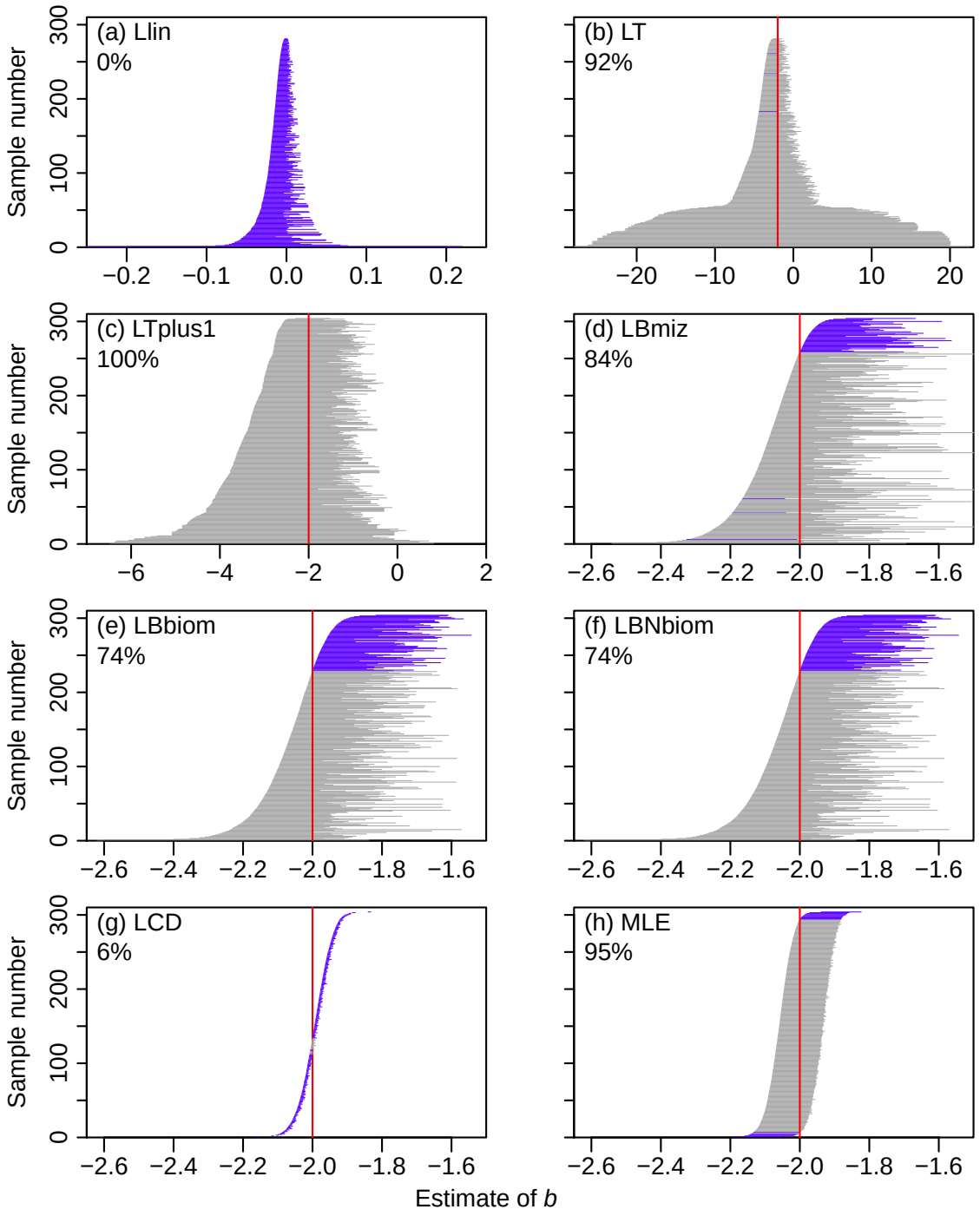

Figure A.8: As for Figure 4 (with the same axes) but with $x_{\max} = 10,000$, showing the confidence intervals for each method.

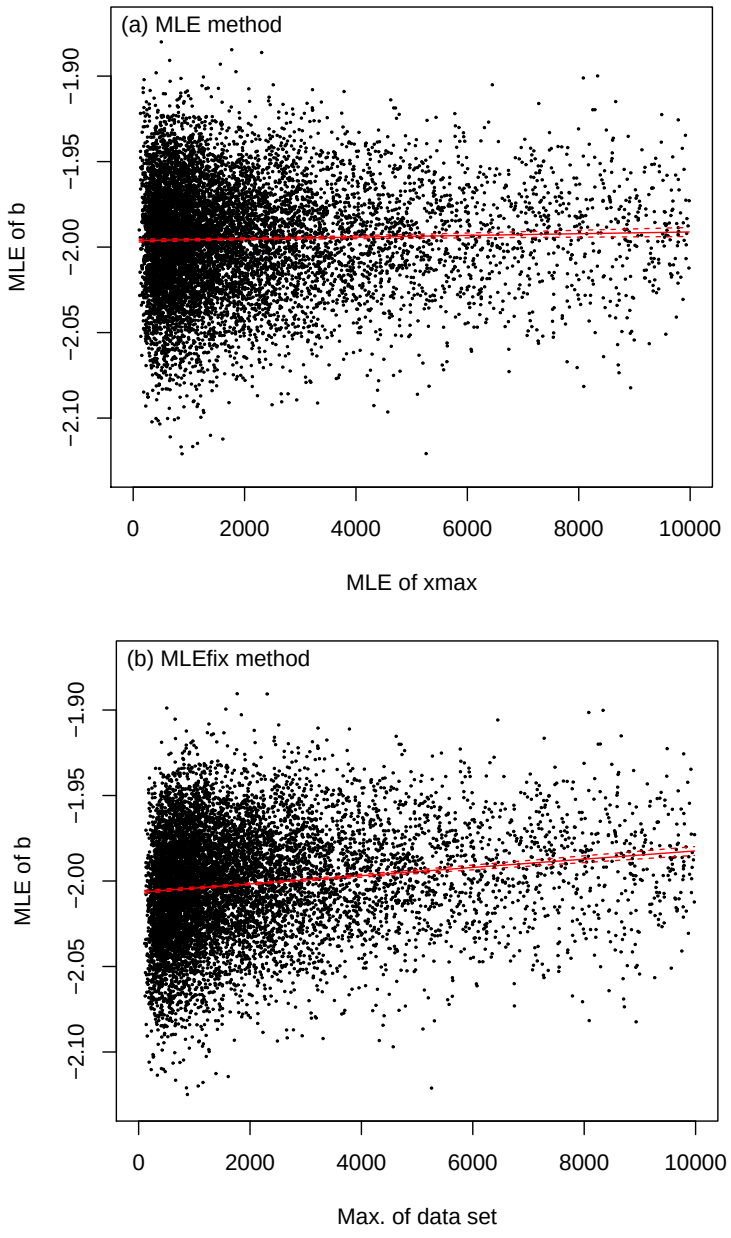

Figure A.9: As for Figures A.2 and A.5 but with $x_{\max} = 10,000$. For both cases the slope of the fitted regression is significantly different from 0: (a) slope is 5.1×10^{-7} (s.e. 1.6×10^{-7}), $p = 0.002$, $R^2 = 0.001$; (b) slope is 2.4×10^{-6} (s.e. 1.6×10^{-7}), $p < 10^{-15}$, $R^2 = 0.022$.

A.2.4 Setting $b = -2.5$ in the simulated data sets

We now set $b = -2.5$ (with other values as in the main text), which represents a steeper size spectrum slope than for our default of $b = -2$. Analogous results to those in the main text are in Figures A.10, A.11, A.12 and Table A.2.

For the single simulated data set (Figure A.10) there are no sample values > 100, even though $x_{\max} = 1,000$. This is because with $b = -2.5$ there is very little chance of obtaining values in the tail (i.e. very few big fish, with 'big' defined as > 100 times larger than the smallest fish of size 1). To be precise, $\mathrm{P}(X \leq 100) = 0.9990316$ [using our R code: `pPLB(100, b=-2.5, xmin=1, xmax=1000)`]. Raising this to the power 1,000 (for 1,000 fish) gives 0.38, which is the probability that all 1,000 random fish sizes are < 100. Thus, we would expect to see at least one fish size > 100 in only 62% of random samples of 1,000 sizes with $b = -2.5$. For $b = -2$ the equivalent percentage is 99.99%, demonstrating the dramatic influence of the value of b (and further demonstrating the need for accurate estimation of b).

For the 10,000 simulations, the resulting distributions of estimated b are wider (Figure A.11 and Table A.2) than for when $b = -2$. For the LT method the distribution is less biased (more centered around the true value of b), but for the LTplus1, LBmiz, LBbiom and LBNbiom the distributions are more biased. In particular, for the LBmiz, LBbiom and LBNbiom methods, for $b = -2$ the medians and means were within 0.01 of $b = -2$, but for $b = -2.5$ they are ≥ 0.09 away from the true value. The distributions for the LCD and MLE methods remain centered around the true value of b.

Compared to the $b = -2$ results, the observed coverage of the 95% confidence intervals is slightly better (closer to the desired 95%) for the LT and LTplus1 methods (Figure A.12). But it is worse for the LBmiz, LBbiom and LBNbiom methods, and remains the same for the LCD (6%) and MLE (95%) methods. Thus, as for $b = -2$, the MLE method is the only one for which the confidence intervals exhibit the desired 95% observed coverage.

Table A.2: As for Table 2 but for simulations with $b = -2.5$, corresponding to the histograms in Figure A.11.

Method	Slope represents	5% quantile	Median	Mean	95% quantile	Percentage below -2.5
Llin	-	-0.13	-0.05	-0.05	-0.01	0
LT	b	-2.98	-2.49	-2.51	-2.15	48
LTplus1	b	-2.74	-2.27	-2.30	-1.91	21
LBmiz	$b+1$	-2.61	-2.40	-2.40	-2.17	22
LBbiom	$b+2$	-2.64	-2.41	-2.41	-2.15	26
LBNbiom	$b+1$	-2.64	-2.41	-2.41	-2.15	26
LCD	$b+1$	-2.59	-2.48	-2.48	-2.38	39
MLE	b	-2.57	-2.49	-2.49	-2.42	43

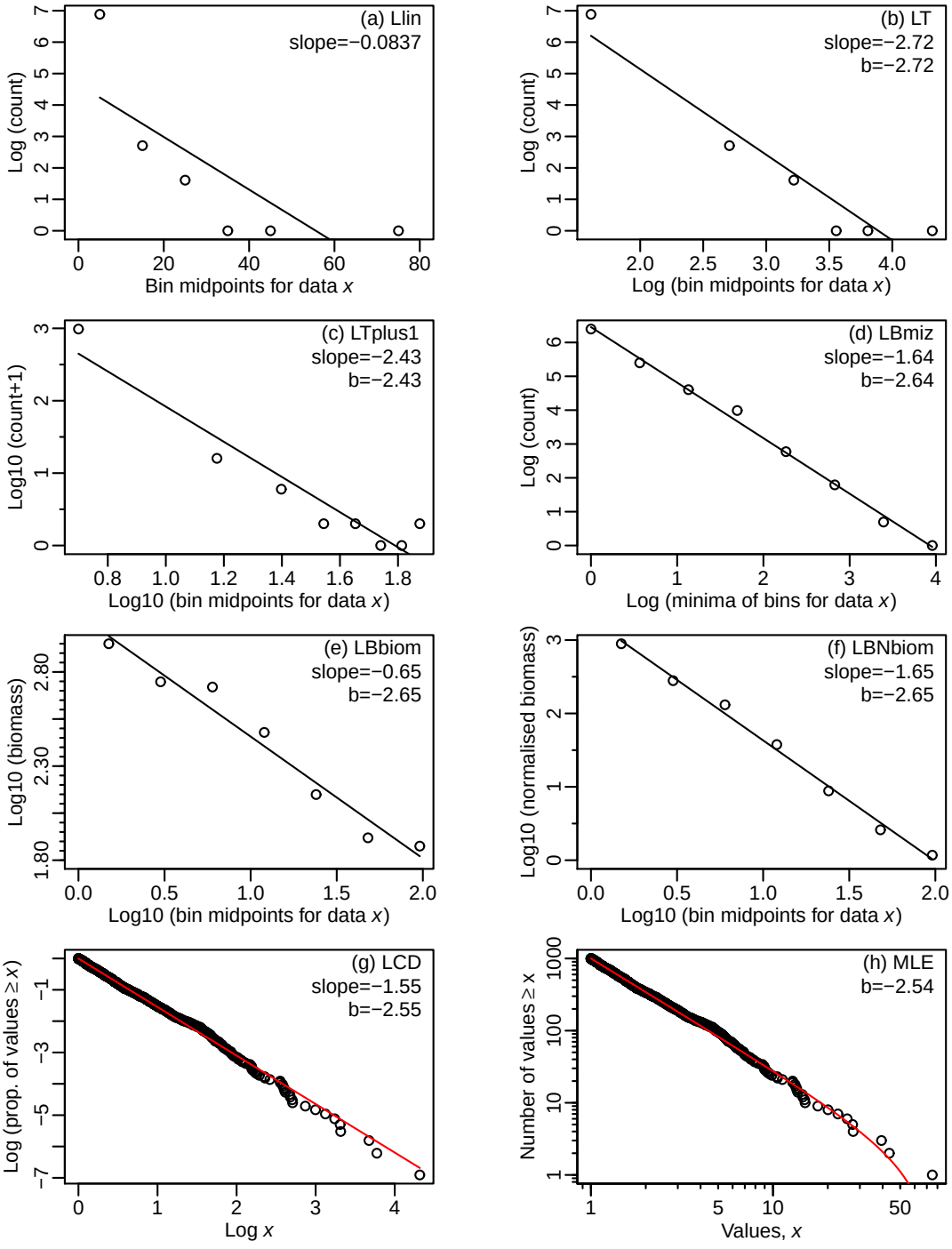

Figure A.10: As for Figure 2 but with $b = -2.5$.

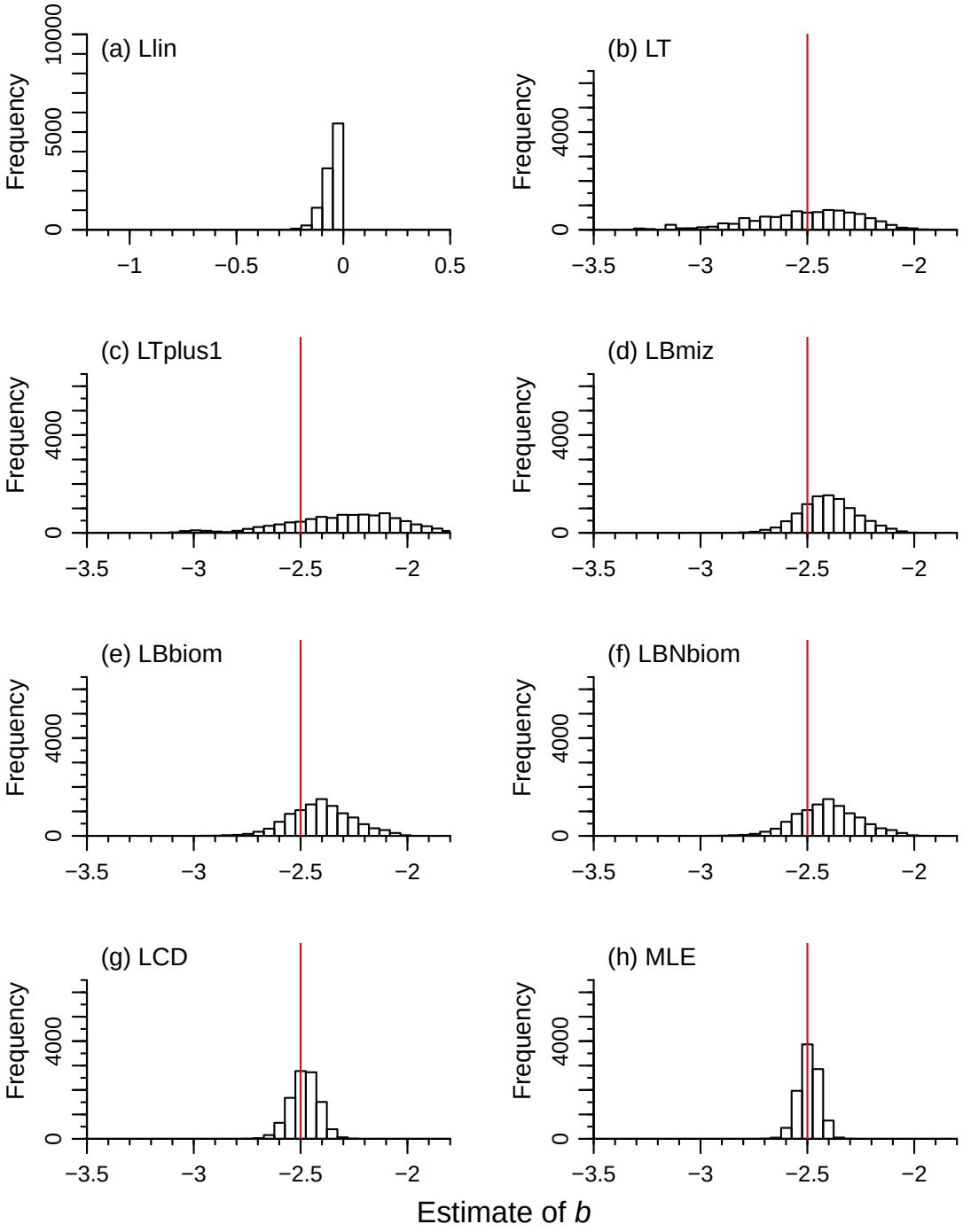

Figure A.11: As for Figure 3 but with $b = -2.5$.

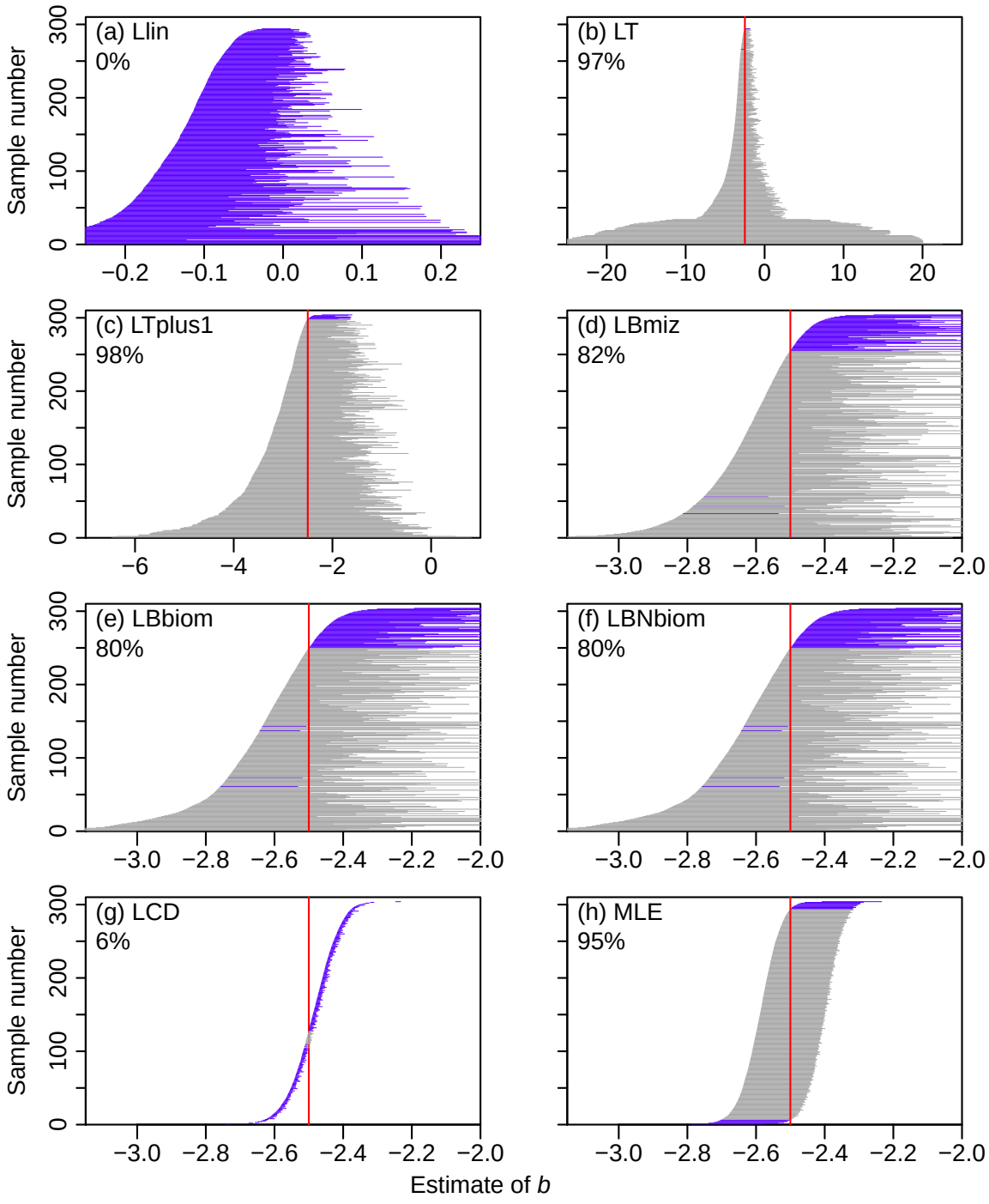

Figure A.12: As for Figure 4 but with $b = -2.5$, showing the confidence intervals for each method.

A.2.5 Setting $b = -1.5$ in the simulated data sets

We now set $b = -1.5$ (with other values as in the main text), which represents a shallower size spectrum slope than for our default of $b = -2$. Analogous results to those in the main text are in Figures A.13, A.14, A.15 and Table A.3.

Compared to the results with $b = -2$, the LCD method has performed noticeably worse (Figure A.14(g)), with all 100% of the estimates of b lying below the true value of $b = -1.5$, compared to 59% for $b = -2$. The higher value of b gives more random values in the tail of the distribution (Figure A.13), close to the upper bound of $x_{\max} = 1,000$. The LCD cannot fit these values because it implicitly assumes an unbounded power-law rather than a bounded one, but there are no values $> x_{\max} = 1,000$. Such values would be expected for an unbounded power law: $\mathrm{P}(X \leq 1,000)$ for a single value from an unbounded power-law is, using our R code, `pPL(1000, b=-1.5, xmin=1)` giving 0.968, which when raised to 1,000 is 10^{-14}; i.e. essentially zero probability that all values are $< 1,000$. The LBmiz method performs somewhat worse than for $b = -2$ (Table A.3), and the distributions for the LBbiom and LBNbiom are shifted slightly away from being centered around the true value of b. The ranges of distributions for all methods are narrower than for $b = -2$, but only the distribution for the MLE method remains centered around the true value of b. The observed coverage of the 95% confidence intervals is slightly closer to the desired 95% for the LBbiom and LBNbiom methods, and for the MLE method remains at the desired 95% level (Figure A.15).

Table A.3: As for Table 2 but for simulations with $b = -1.5$, corresponding to the histograms in Figure A.14.

Method	Slope represents	5% quantile	Median	Mean	95% quantile	Percentage below -1.5
Llin	-	-0.01	-0.01	-0.01	0.00	0
LT	b	-2.24	-2.03	-2.04	-1.83	100
LTplus1	b	-2.07	-1.89	-1.89	-1.73	100
LBmiz	$b+1$	-1.61	-1.55	-1.55	-1.50	93
LBbiom	$b+2$	-1.56	-1.51	-1.51	-1.46	60
LBNbiom	$b+1$	-1.56	-1.51	-1.51	-1.46	60
LCD	$b+1$	-1.66	-1.63	-1.63	-1.60	100
MLE	b	-1.53	-1.50	-1.50	-1.47	47

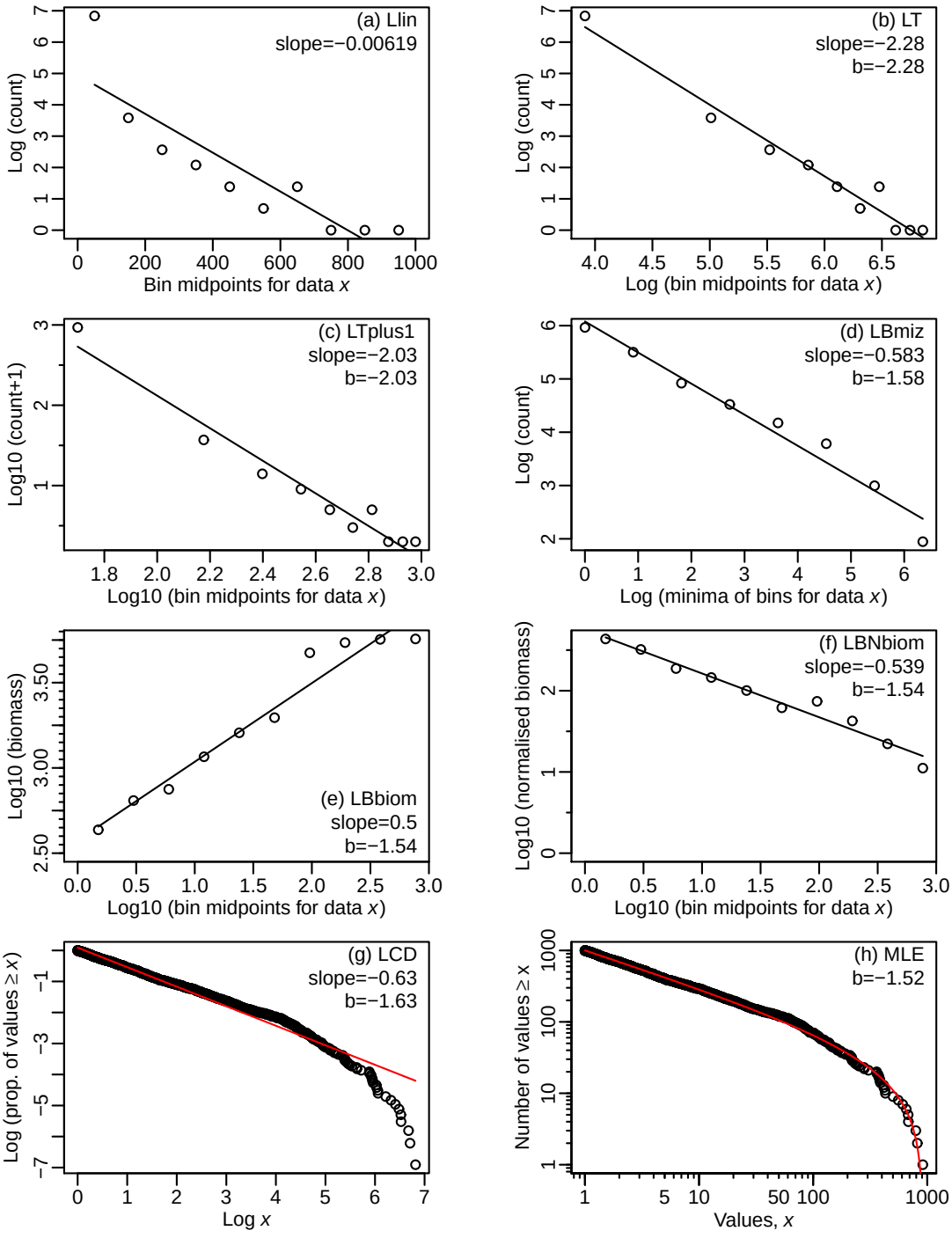

Figure A.13: As for Figure 2 but with $b = -1.5$.

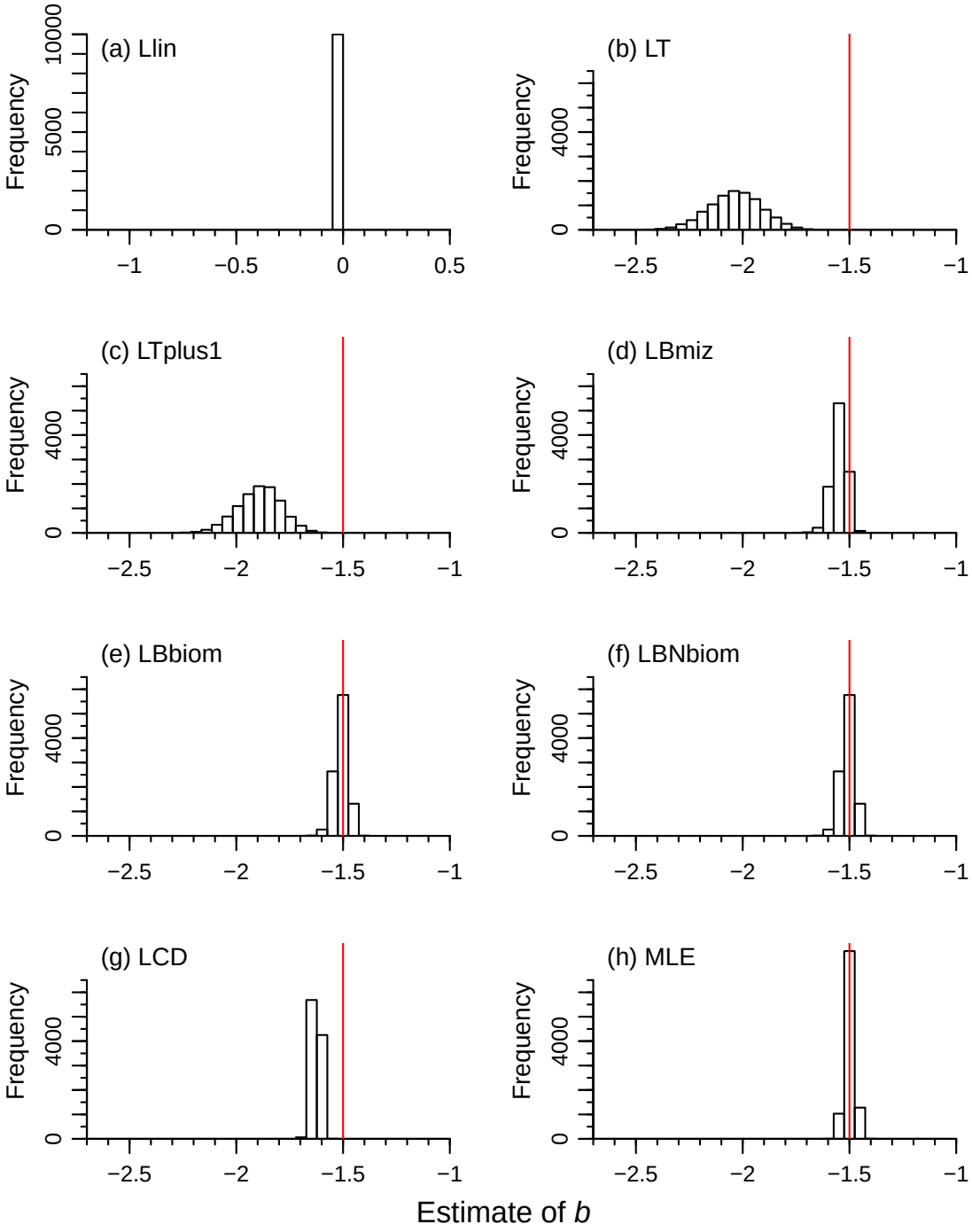

Figure A.14: As for Figure 3 but with $b = -1.5$.

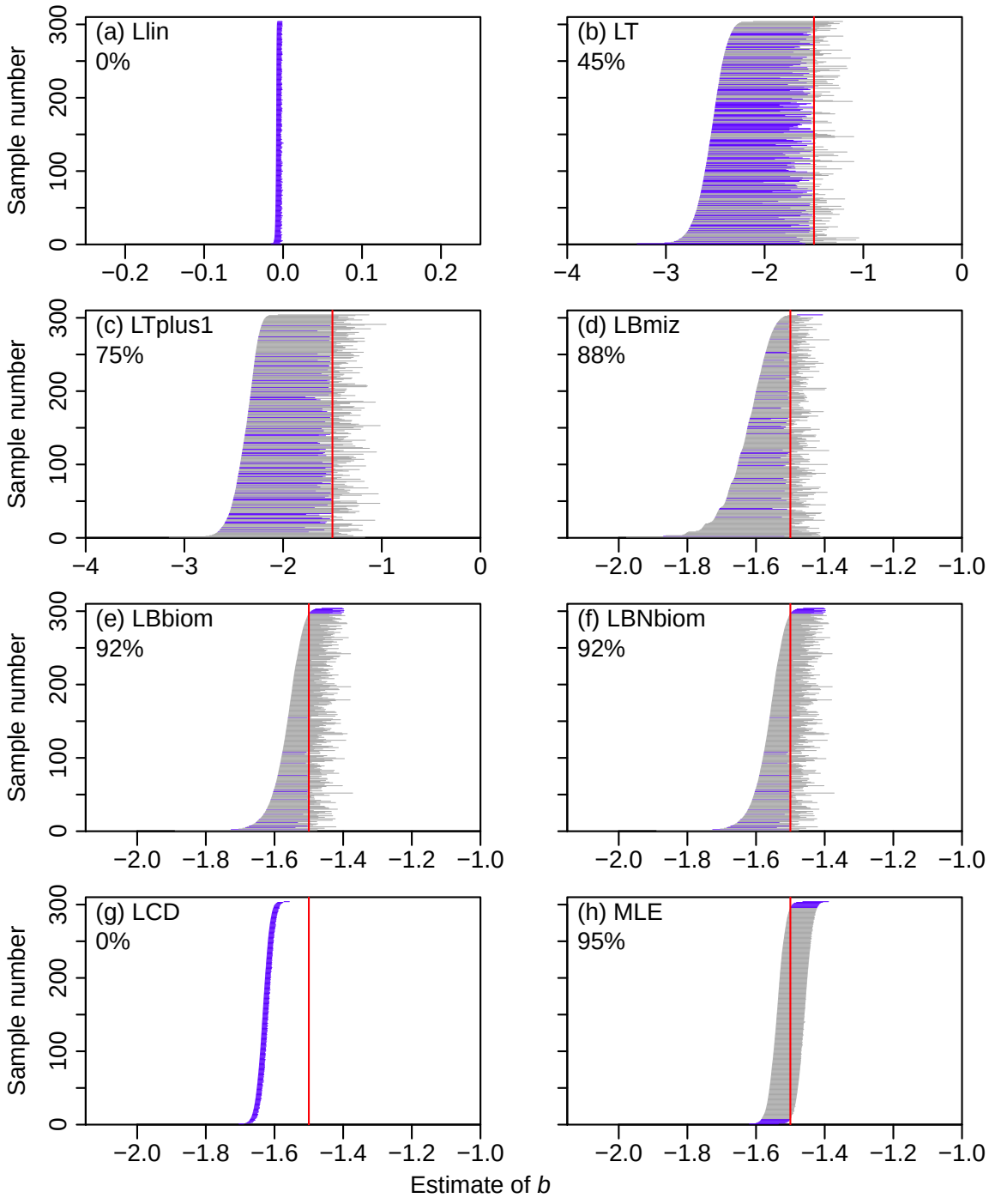

Figure A.15: As for Figure 4 but with $b = -1.5$, showing the confidence intervals for each method.

A.2.6 Setting $b = -0.5$ in the simulated data sets

We now set $b = -0.5$, which represents an even shallower size spectrum slope than the previous $b = -1.5$, and is in the vicinity of the values of around -0.22 estimated by Graham *et al.* (2005) using the LTplus1 method. Analogous results to those in the main text are in Figures A.16, A.17, A.18 and Table A.4.

Compared to the results with $b = -1.5$, the LBbiom and LBNbiom methods have actually improved in accuracy (Figure A.17 and Table A.4), whereas they had slightly worsened from $b = -2$ to $b = -1.5$. The MLE method retains the accuracy it had for $b = -1.5$ while the remaining methods remain poor, almost always over- or under-estimating the value of b (Table A.4). Compared to $b = -1.5$, the observed coverage of the 95% confidence intervals is the same (at 92%) for the LBbiom and LBNbiom methods, and remains at the desired 95% level for the MLE method (Figure A.18).

Table A.4: As for Table 2 but for simulations with $b = -0.5$, corresponding to the histograms in Figure A.17.

Method	Slope represents	5% quantile	Median	Mean	95% quantile	Percentage below -0.5
Llin	-	0.00	0.00	0.00	0.00	0
LT	b	-0.62	-0.57	-0.57	-0.51	98
LTplus1	b	-0.62	-0.56	-0.56	-0.51	97
LBmiz	$b+1$	-0.61	-0.57	-0.57	-0.53	99
LBbiom	$b+2$	-0.55	-0.50	-0.50	-0.45	51
LBNbiom	$b+1$	-0.55	-0.50	-0.50	-0.45	51
LCD	$b+1$	-1.48	-1.46	-1.46	-1.44	100
MLE	b	-0.53	-0.50	-0.50	-0.47	53

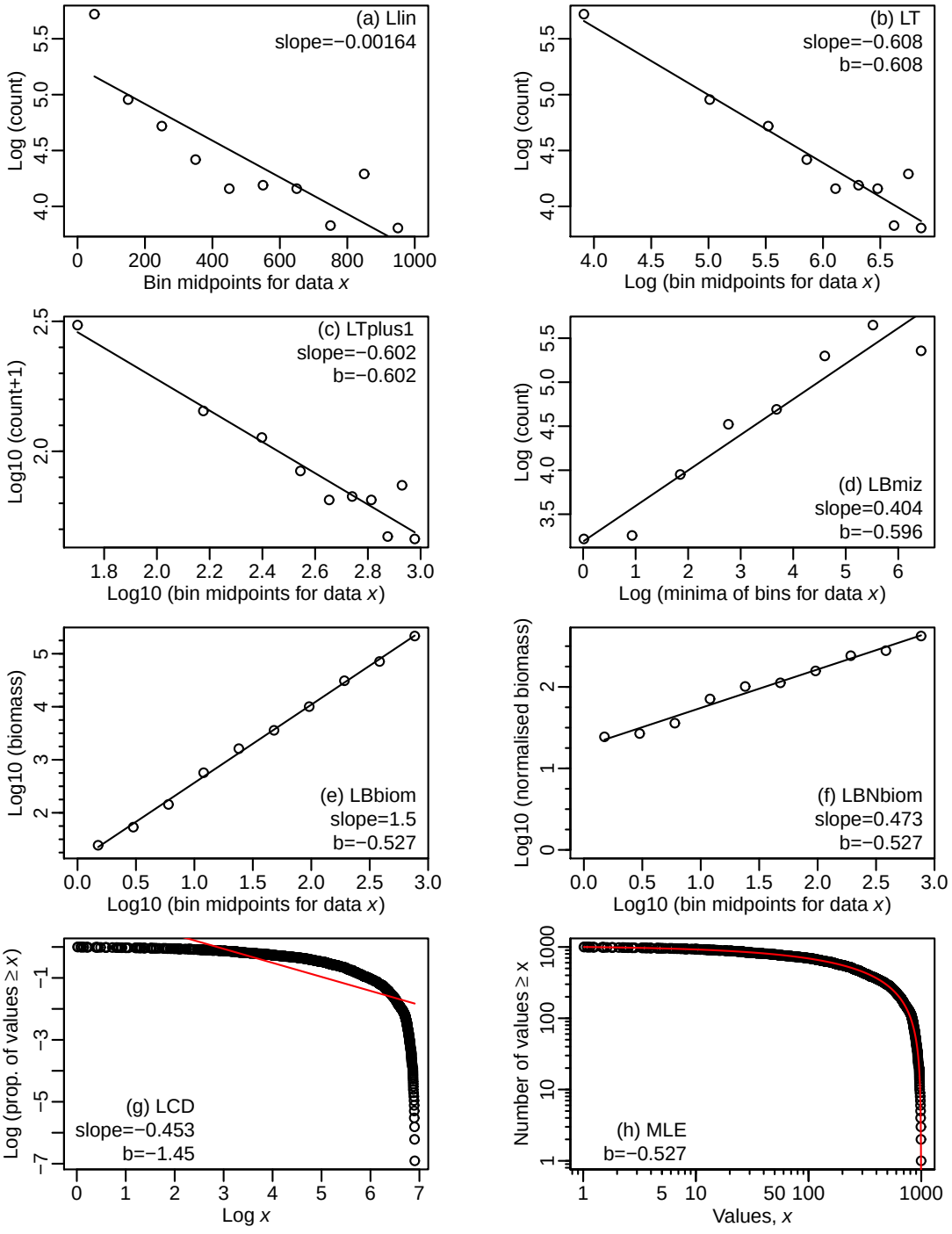

Figure A.16: As for Figure 2 but with $b = -0.5$.

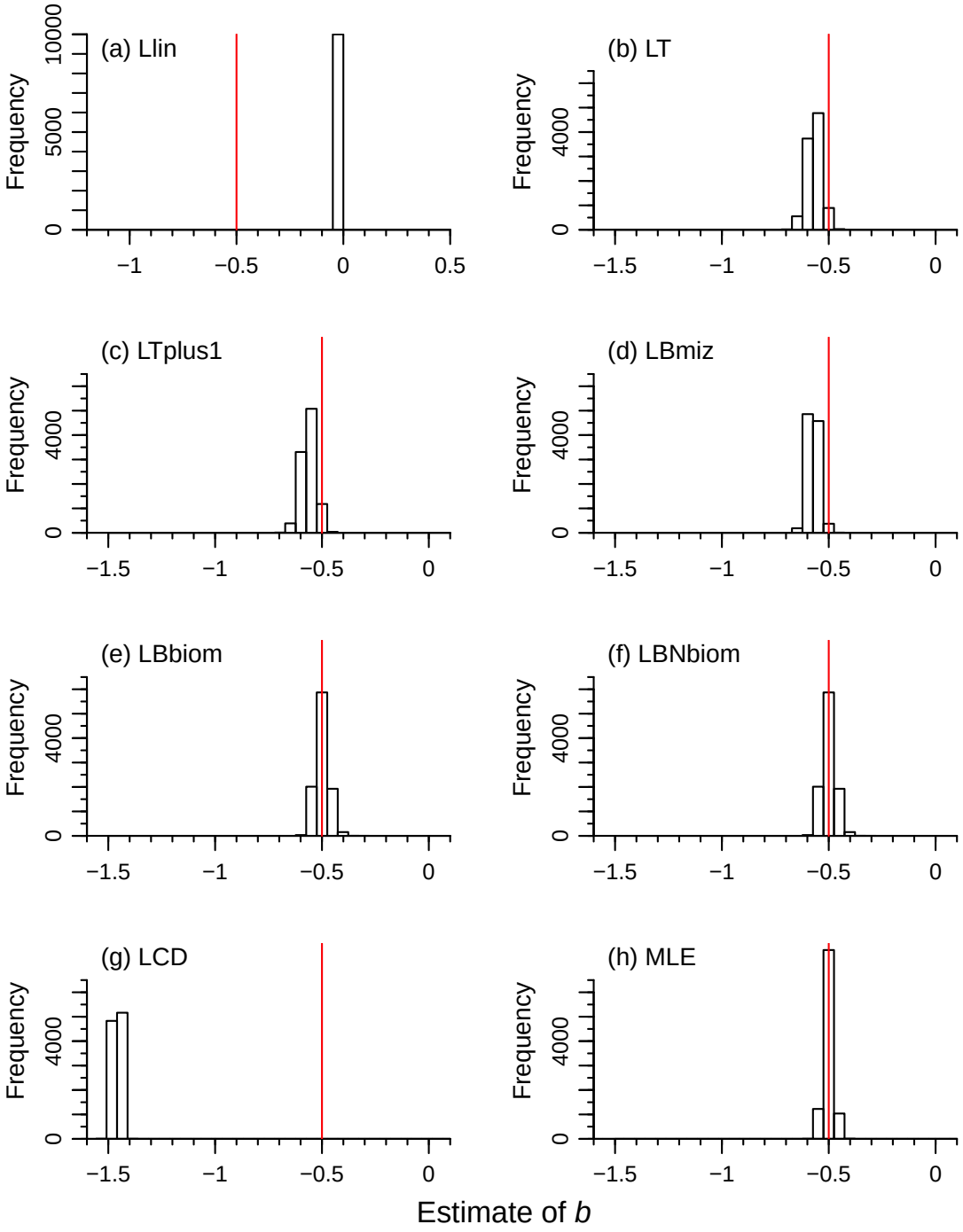

Figure A.17: As for Figure 3 but with $b = -0.5$.

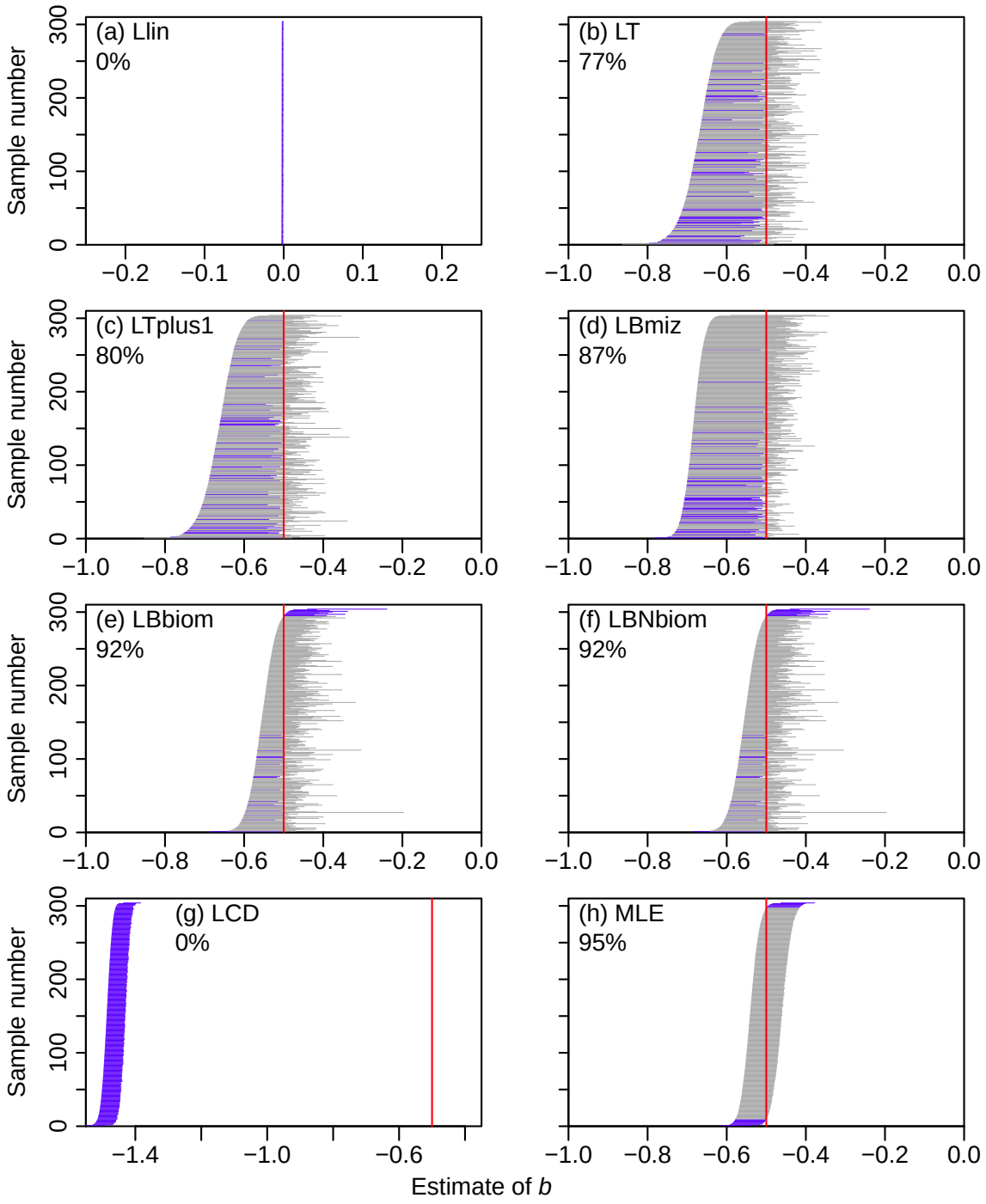

Figure A.18: As for Figure 4 but with $b = -0.5$, showing the confidence intervals for each method.

A.2.7 Setting $n = 10,000$ in the simulated data sets

We now set the sample size $n = 10,000$ (with other values as in the main text) to test the effects of a ten-fold increase in sample size. Analogous results to those in the main text are in Figures A.19, A.20, A.21 and Table A.5.

Compared to the original results with $n = 1,000$, for $n = 10,000$ the range of estimates of b is tighter around the true $b = -2$ for the LBmiz, LBbiom, LBNbiom, LCD and MLE methods (Figure A.20 and Table A.5). However, all except the MLE method now have at least 59% of the estimates being below the true value. The observed coverage of the 95% confidence intervals has actually worsened (further away from the desired 95% value) for all methods, though only to 94% for the MLE method (Figure A.21 compared to Figure 4). The confidence intervals have become narrower for all methods with the increase in sample size.

Table A.5: As for Table 2 but for simulations with $n = 10,000$, corresponding to the histograms in Figure A.20.

Method	Slope represents	5% quantile	Median	Mean	95% quantile	Percentage below -2
Llin	-	-0.01	-0.01	-0.01	-0.01	0
LT	b	-3.12	-2.89	-2.89	-2.68	100
LTplus1	b	-2.94	-2.71	-2.72	-2.55	100
LBmiz	$b+1$	-2.10	-2.03	-2.04	-1.98	85
LBbiom	$b+2$	-2.07	-2.01	-2.01	-1.96	59
LBNbiom	$b+1$	-2.07	-2.01	-2.01	-1.96	59
LCD	$b+1$	-2.04	-2.02	-2.02	-2.00	95
MLE	b	-2.02	-2.00	-2.00	-1.98	48

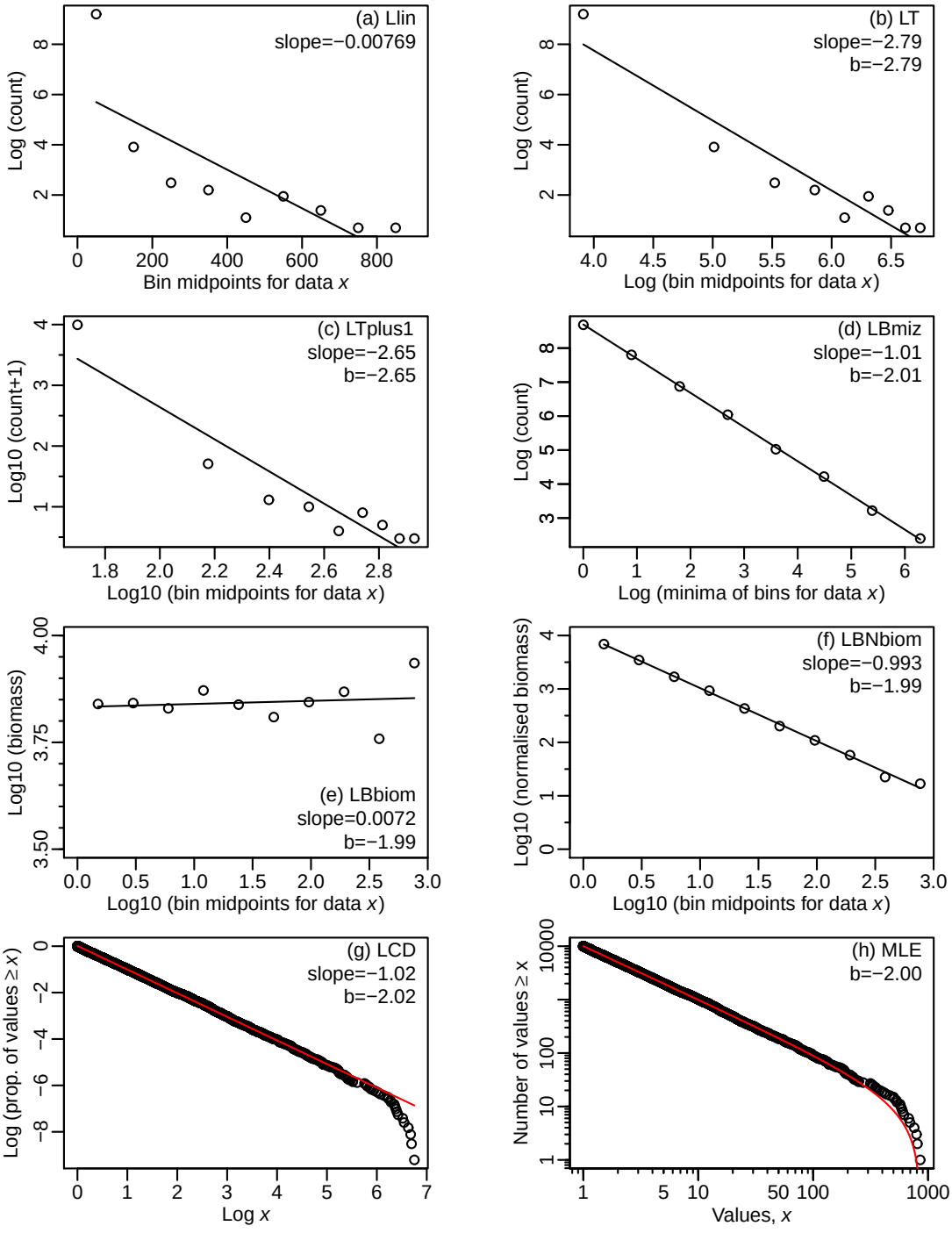

Figure A.19: As for Figure 2 but with $n = 10,000$.

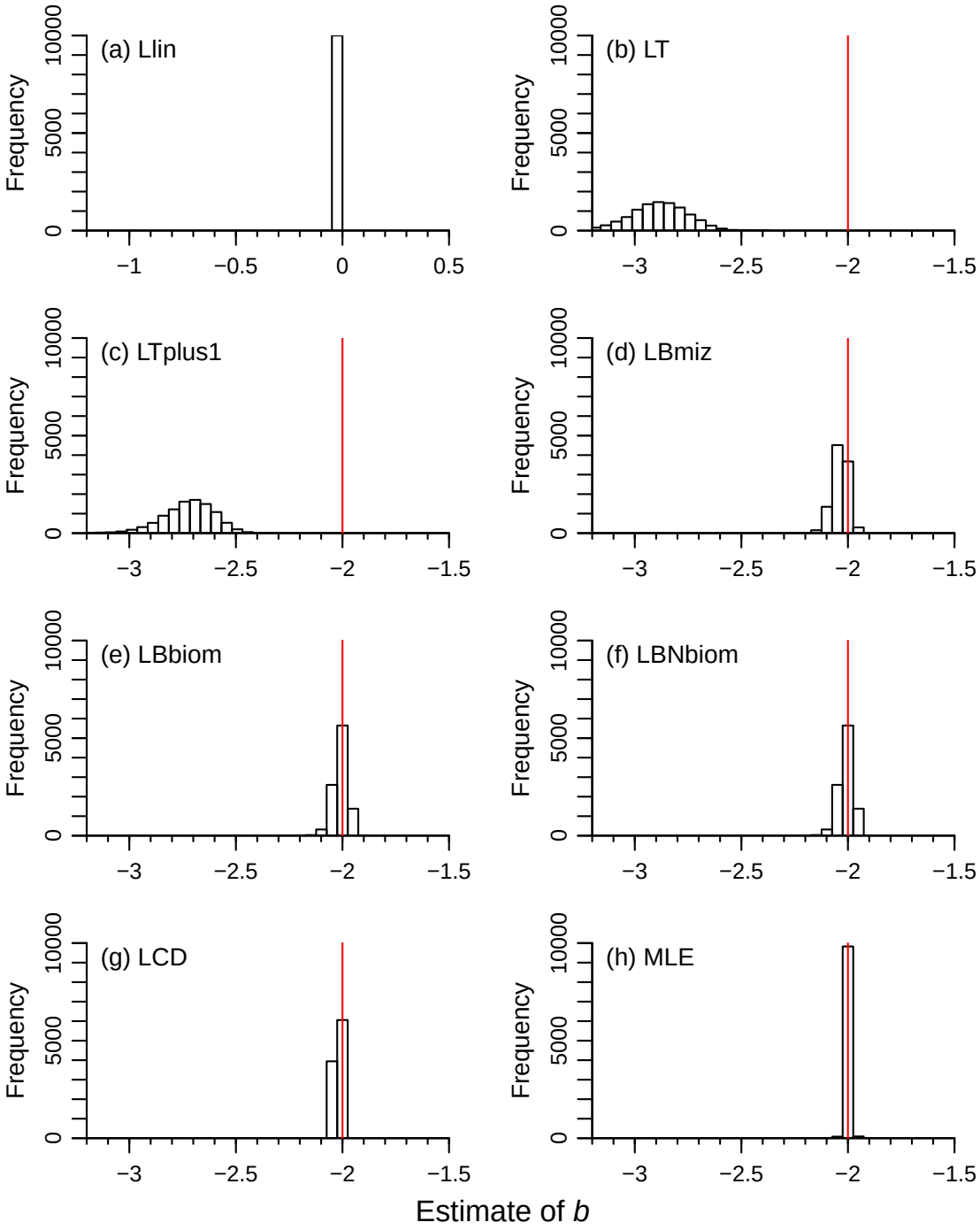

Figure A.20: As for Figure 3 but with $n = 10,000$. Note that there are 180 estimated b values below the minimum shown for the LT method, and 15 for the LTplus1 method; this is to keep the x-axes the same as in Figure 3. The bin widths change slightly from Figure 3, but still ensure that $b = -2$ is in the centre of a bin. The y-axes are the same for all panels here.

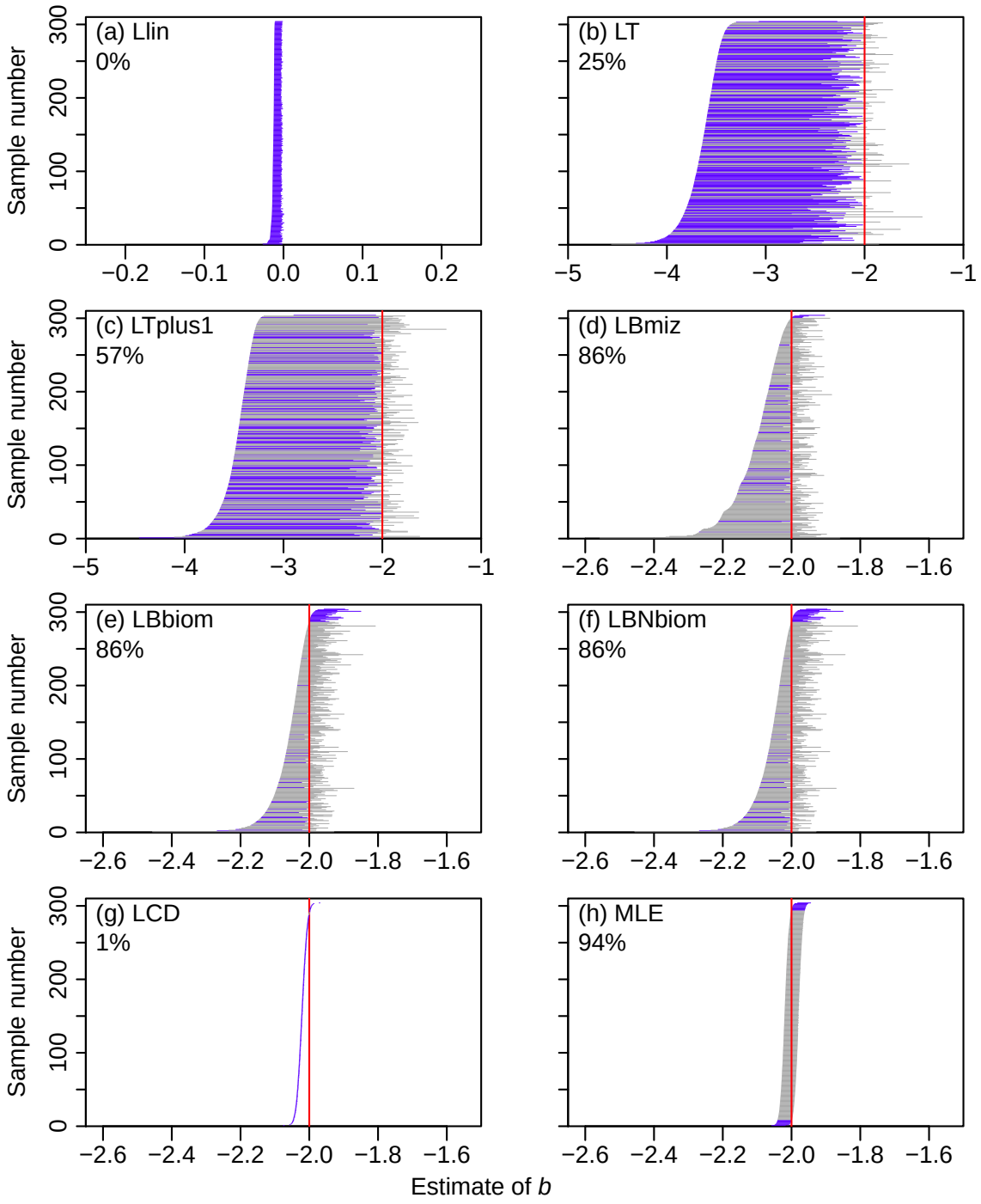

Figure A.21: As for Figure 4 but with $n = 10,000$, showing the confidence intervals for each method.

A.2.8 Re-running with a different seed

We fixed the seed of the random-number generator to 42, to enable reproduction of all results. However, because of the way that random numbers are generated, care has to be taken to ensure results are not dependent on the seed. For example, Figures 1 and A.6 have $x_{\max} = 1,000$ and $x_{\max} = 10,000$, respectively, with everything else the same (seed set to 42, $n = 1,000$, $x_{\min} = 1$ and $b = -2$). However, because the seed generates the same sequence of 1,000 random uniform numbers (which are then used to generated the bounded power-law numbers), the resulting samples of 1,000 power-law numbers are very similar. Taking the element-wise ratio of the two samples, we find that 91.5% of the $x_{\max} = 1,000$ samples are $> 99\%$ of their respective $x_{\max} = 10,000$ values (e.g. the first element in each sample is 11.61322 and 11.72533, which is a ratio of $> 99\%$). This suggests that the similarities seen between the black and gold histograms in Figure 3 could be partly due to the same seed being used.

As seen in Figure A.6, the maximum sample value for $x_{\max} = 10,000$ is < 700. Even though we allow values up to 10,000, the nature of power-law distributions is that values $> 1,000$ will be rare. From (A.16) with $x_{\max} = 10,000$, we calculate $\mathrm{P}(X \leq 1000) = 0.9990999$ [using our R code: `pPLB(1000, b=-2, xmin=1, xmax=10000)`]. Raising this to the power 1,000 gives 0.41, which is the probability that all 1,000 random numbers are $< 1,000$. Thus, in only 59% of random samples of 1,000 numbers with $x_{\max} = 10,000$ would be expect to see a value $> 1,000$. Changing $x_{\max}$ from 1,000 to 10,000 does not guarantee we will obtain any values $> 1,000$, and because of the potential influence of occasional large numbers (a consequence of power-law distributions) on the methods, we need to be careful about the seed.

To ensure that our conclusions are not dependent on the seed, we have re-run all main code with a different seed. However, the results and general conclusions do not change (or any changes are minimal). For example, with the seed set to 43, Table 2 is identical,

except that five of the statistics related to b change by ≤ 0.01 and three of the percentages change by 1% (although the actual changes will be even less because the reported values are rounded). And all observed coverage values in Figure 4 and Figure A.8 are unchanged. Therefore, our results and conclusions appear robust to the choice of seed.

A.2.9 Subsampling of confidence intervals

For the confidence interval plots such as Figure 4 we present subsamples of the 10,000 calculated confidence intervals, since plotting all 10,000 intervals is not feasible. For each method, the 10,000 confidence intervals are ranked in increasing value of the lower bound of the interval, and then the subsample is taken as the samples with ranks 1, 34, 64, ..., 9,967 and 10,000, so as to include samples 1 and 10,000 (i.e. the intervals with the smallest and largest lower bounds). This yields 304 intervals for each method, giving adequate resolution when the intervals are plotted as horizontal lines. The left endpoints (lower bounds) create a smooth monotonically increasing curve as the sample number increases because the samples are ranked by the lower bounds; the right endpoints (upper bounds) do not have to be monotonically increasing.