Journal of Animal Ecology

British Ecological Society

Journal of Animal Ecology 2015, **84,** 892–897 doi: 10.1111/1365-2656.12382

'HOW TO. . .' PAPER

On the variety of methods for calculating confidence intervals by bootstrapping

Marie-Therese Puth[1], Markus Neuhäuser[1] and Graeme D. Ruxton[2]*

[1]*Fachbereich Mathematik und Technik, RheinAhrCampus, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, 53424 Remagen, Germany; and* [2]*School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, UK*

Summary

1. Researchers often want to place a confidence interval around estimated parameter values calculated from a sample. This is commonly implemented by bootstrapping. There are several different frequently used bootstrapping methods for this purpose.

2. Here we demonstrate that authors of recent papers frequently do not specify the method they have used and that different methods can produce markedly different confidence intervals for the same sample and parameter estimate.

3. We encourage authors to be more explicit about the method they use (and number of bootstrap resamples used).

4. We recommend the bias corrected and accelerated method as giving generally good performance; although researchers should be warned that coverage of bootstrap confidence intervals is characteristically less than the specified nominal level, and confidence interval evaluation by any method can be unreliable for small samples in some situations.

Key-words: parameter estimation, randomization, resampling, statistics

Introduction

It is increasingly common for researchers to offer estimates of (population) parameter values (e.g. mean, effect size, correlation coefficient, etc.) instead of (or as well as) the *P*-values associated with testing of null hypotheses. Evaluation of a sample from the population of interest does not allow us to estimate a population parameter with certainty; rather, it is common to produce a confidence interval in which we aim to enclose the true (population) value on a specified fraction of occasions. For the frequently used 95% confidence interval, the calculated interval is intended to enclose the true value for 95% of samples. Researchers often use *bootstrapping* as a very general means of generating a confidence interval for many commonly used statistics. The essence of bootstrapping is the drawing of pseudo-samples (hereafter called bootstraps) either from the original sample itself (nonparametric bootstrapping) or from a model fitted to the original sample (parametric bootstrapping). The first of these is more commonly used and will be the focus of this paper. Nonparametric bootstrapping makes no assumptions about the nature of the underlying population, and the only information used is the original sample itself. Bootstrapping is a popular method of producing confidence intervals because of its generality: the same method can be used for a very broad range of statistics. Although computationally demanding, various methods have been made available through commonly used statistical software. This means that for almost any situation in animal ecology where you can calculate some statistic, bootstrapping can be used to generate a confidence interval around the point value. Further, the calculation of this can be done with minimal extra effort in commonly used statistical computing software packages.

There are other methods available for calculation of confidence intervals for various parameters. However, almost all alternatives to bootstrapping require assumptions to be made about properties of the underlying distribution from which the sample is drawn. These assumptions are not always made explicit to the reader and may not always be easy to evaluate on the basis of the sample or of other biological knowledge of the system. These methods are also often specific to a particular type of parameter, whereas a bootstrapping technique will work in essentially the same way when applied to different parameters. Finally, especially for uncommon

*Correspondence author. E-mail: gr41@st-andrews.ac.uk

parameters, there may be no available alternatives to general methods such as bootstrapping.

Here we briefly describe a diversity of the different alternative bootstrapping methods easily available to researchers, offer some tentative suggestions on which to use, and encourage researchers to provide more detail on the method they use when reporting results. A survey of the recent literature suggests that such specification is not a widely adopted practice, although it is necessary in order for methodology to be reproducible. We illustrate our results with comparison of the performance of different techniques in a simulation study involving estimation of strength of association.

Different methods of bootstrapping to obtain confidence intervals

Philosophically, given its definition above, the confidence interval should best be constructed by taking many samples of the underlying population. However, often only a single sample is available to us. The essence of nonparametric bootstrapping is that in the absence of any other information about a population, the distribution of values found in a random sample from the population is the best guide to the distribution of values in the population. Therefore to approximate repeated sampling from the population, it is justifiable to resample the sample (with replacement). There are a number of different ways that bootstrapping can be used to obtain the confidence interval for a population parameter, and some of the most popular of these will be described below in the context of two commonly used R packages devoted to bootstrapping: *bootstrap* and *boot*.

The R package *bootstrap* offers four different methods taken from Efron & Tibshirani (1993): the *BCa* method, parametric and nonparametric versions of the *ABC* method, and the *studentized* method. The *boot.ci* function of the R package *boot* offers five methods: as well as the *studentized* and *BCa* methods described above, it offers the *first-order normal approximation*, *basic* and *percentile* methods. The implementation of all of these methods is explained fully in Chapter 5 of Davison & Hinkley (1997). From our survey of the literature, another method that seems to commonly be used, but which is not available in either of our focal R packages, is *bias-corrected* (*BC*) percentile confidence limits. Our literature survey suggested that users used the statistical package STATA to access this method. For each of these methods, we describe the underlying rationale behind them, in order of increasing complexity. In each case, we will assume that we want to estimate some population parameter θ, and the estimator of this parameter based on the original sample is $\hat{\theta}$. Under standard bootstrapping, B bootstrap samples are generated and an estimate of the population parameter is calculated for each; these estimates are denoted $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

PERCENTILE METHOD

If the bootstrap estimates $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ are ordered, then the two values that contain the central 100 $(1 - \alpha)$% of these estimates are used as the limits of the 100 $(1 - \alpha)$% confidence interval for the population parameter of interest (commonly $\alpha = 0{\cdot}05$ is used). This method depends upon the fact that the empirical distribution function based on the bootstraps converges to the true distribution function asymptotically in sample size, and the empirical quantiles have a law of large numbers.

BASIC METHOD

For each $\hat{\theta}^*_i$, an error e_i is calculated as $\hat{\theta}^*_i - \hat{\theta}$. It is these e_i that are then ordered, and we identify the lower and upper limits e_l and e_u that enclose the central 100 $(1 - \alpha)$% of these errors. The 100 $(1 - \alpha)$% confidence interval for the population parameter is then $[\hat{\theta} - e_l, \hat{\theta} - e_u]$. This method is based on the assumption that the bootstrap distribution of errors is a good approximation for the real distribution of sampling errors.

FIRST-ORDER NORMAL APPROXIMATION

For this method, if $\hat{\sigma}_B$ is the standard deviation of the B bootstrap samples and $\bar{\theta}$ is the mean, then the 100 $(1 - \alpha)$% confidence interval is given by

$$\bar{\theta} \pm z_{\alpha/2}\, \hat{\sigma}_B,$$

where $z_{\alpha/2}$ is the z-score for a given level of significance α, if $\alpha = 0{\cdot}05$ then $z_{\alpha/2} = 1{\cdot}96$.

This method takes advantage of the observation that bootstraps often approximate a normal distribution.

STUDENTIZED METHOD

Here we work with the t-distribution as a so-called pivot function, rather than the raw bootstraps directly. Specifically we first estimate the standard deviation $(\hat{\sigma}^*_i)$ for each bootstrap i. This can be calculated in the conventional way if the parameter of interest is a mean; otherwise, alternative equations or procedures are required (see Efron & Tibshirani 1993 for details). For example, where the parameter of interest is a correlation coefficient, we can use:

$$\frac{\hat{\sigma}^*_i = 1 - \hat{\theta}^{*2}_i}{\sqrt{n},}$$

and we then calculate a t value for that bootstrap:

$$\frac{t^*_i = \hat{\theta}^*_i - \hat{\theta}}{\hat{\sigma}^*_i.}$$

We then order these t^*_i and find the lower and upper values (t^*_l and t^*_u) that enclose the central 100 $(1 - \alpha)$%

of these t values. The limits of the 100 $(1 - \alpha)$% confidence interval for the population parameter of interest are then $[\hat{\theta} - t_u^*\hat{\sigma}, \hat{\theta} - t_l^*\hat{\sigma}]$, where $\hat{\sigma}$ is the standard deviation of the original sample, calculated in the same way as for the bootstraps. The key attraction of this approach is that the statistic that is bootstrapped is pivotal (i.e. it does not depend on any nuisance parameters). This means that it is possible to make exact probability statements about this statistic, which by a 'pivoting' procedure can be converted to an exact confidence interval for the statistic of interest. Thus, this method should theoretically produce more precise confidence intervals than any of the forgoing.

BIAS-CORRECTED (BC) PERCENTILE CONFIDENCE LIMITS

The percentile method could be improved if any bias that arises because the true parameter value is not the median of the distribution of bootstrap estimates could be estimated and corrected for. Essentially, this revised method is based on the assumption that there is a monotonic increasing transformation $f()$ of the estimator $\hat{\theta}$ such that the transformed values $f(\hat{\theta})$ are normally distributed with mean $f(\theta) - z_o$ and standard deviation one. The central challenge of implementing this method is to provide an estimator for z_o.

ACCELERATED BIAS-CORRECTED PERCENTILE LIMITS (BCA)

This method is based on the assumption that a transformation $f()$ of the estimator $\hat{\theta}$ exists such that $f(\hat{\theta})$ has a normal distribution with mean $f(\theta) - z_o\ (1 + af\ (\theta))$ and standard deviation $1 + af\ (\theta)$, where z_o and a are constants that we must estimate to use this function. This method seeks to correct for skewness as well as bias in the bootstrap distribution.

ABC METHODS

These were developed as approximations to the BCa method that require much less computational effort. Increasing availability of computing power reduces concerns about computational effort, and so we do not explore these methods in detail. Further details of these methods can be found in Efron & Tibshirani (1993) or Manly (2007).

This is by no means an exhaustive list of possible methods for calculating confidence intervals by bootstrapping. For instance, Chernick & LaBudde (2011) describe further methods called *iterated bootstrap* and *tilted bootstrap*. Manly (2007) provides references to the literature on a number of other methods. However, as a generality, the methods described above are certainly the easiest to implement, and also those most widely used.

Survey of usage of this technique in the literature

In order to survey the use of bootstrap methods to calculate confidence intervals, we examined 132 recent papers (spanning the publication dates October 2010–September 2014) from the journals *Journal of Animal Ecology*, *Ecology* and *Behavioural Ecology* (chosen because they allow searching of the full text of papers for key terms) that use bootstrapping to generate confidence intervals. We found papers by using the search term 'bootstrap* AND confidence interval*'. Twenty-seven papers (20% of our survey) used the *percentile* method; 21 (16%) used *BC,* and 7 (5%) used *BCa*. None of these papers offered an explanation for why the authors chose to use a particular method. In the remaining 77 (58%) of papers, the method used was not stated and could not be inferred from information provided in the manuscript. Of these 132 papers, 51 used bootstrapping to generate a confidence interval for a measure of or difference in central tendency (e.g. mean or median), 19 for a measure of dispersion (e.g. variance of interquartile range), 27 for a measure of association (e.g. correlation coefficient, gradient of a line), and 35 for more complex composite parameters (e.g. species diversity index).

An example to compare the different techniques

As an example comparison of the different methods, we evaluate the different methods in terms of the actual frequency of occasions in which calculated 95% confidence intervals include the true population value of correlation coefficient for 10 000 bivariate random samples on the basis of 10 000 bootstraps in each case. For each random sample, we generate two samples (denoted by U_1 and U_2) from populations that are Normal with mean zero and standard deviation one, with the population value of the Pearson product–moment correlation coefficient between them specified by a parameter ρ. We chose a sample size of 20 since Taborsky (2010) found that that the average sample size per treatment in laboratory experiments in the study of behaviour was approximately 18, and the mean sample size per treatment in our sample of papers describe above was 23. For comparison, we repeated our analyses with sample sizes of 10 and 100. Pearson's correlation coefficient was chosen for illustrative purposes only, and we are not seeking to make any inferences about the properties of that measure in particular. For bivariate normal data, there is a commonly used (non-bootstrapping) method for calculating confidence intervals (Fisher's transformation, which is used by the most popular R function for measuring association *cor.test*), and we can use this as a comparator for the different bootstrapping methods. In our sample of 132 papers, correlation coefficients were the subject of confidence interval construction by bootstrapping in 21 (16%) of them, so it is also an

example that seems relevant to current research practice in our field.

We generate samples of bivariate data (U_1, U_2) with specified (population level) correlation ρ using the method of Berry & Mielke (2000): specifically

$$U_1 = X\sqrt{(1-\rho^2)} + Y\rho \quad \text{and} \quad U_2 = Y$$

where X and Y are independent identically distributed univariate random variables drawn from N(0,1), and ρ is the specified (population) correlation coefficient. We performed this for ρ = 0·0, 0·2 and 0·7 (Code valuable through Dryad – see Puth 2015). Figure 1 provides example confidence intervals for the correlation coefficient calculated for the same sample of twenty pairs of values by the six different methods and the fractions of confidence intervals from 10 000 replicate samples that included the specified population value ρ. We obtained confidence intervals by the following six methods: BCa, BC, basic method, percentile method, normal approximation and studentized method.

We first focus on the middle row of Fig. 1 for sample sizes of 20. This shows that for our example case with a small sample size, all of the bootstrap techniques excepted *studentized* are liberal (producing confidence intervals that are on average too narrow to maintain the specified 95% nominal level of coverage); however, this effect is relatively minor for the BCa and BC methods. Our example also illustrates that for a single sample, the six methods can produce confidence intervals that are quite different from each other: both in size and level of asymmetry about the point estimate from that sample. All these effects can also be seen on the top row for sample sizes of ten. It is not surprising that the smaller samples size leads to generally larger confidence intervals, but notice that the confidence interval for the studentized method is now very strongly conservative (the confidence interval is much too large). Indeed, this confidence interval (and those for the normal and basic methods) can extend beyond the range of values that are possible for Pearson's correlation coefficient [−1, 1]. Finally turning to the bottom row (n = 100), the same effects can still be seen; all methods tend to be liberal, although this effect is now relatively minor, as are differences between methods.

Discussion

Our example shows that for the same sample, the six different methods for calculating confidence intervals can produce very substantially different confidence intervals. Hence, we would strongly recommend that when researchers present such confidence intervals obtained by bootstrapping, they explicitly state the method used. Our survey of the literature suggests that this is not common practice currently. Researchers should also state the number of bootstrap samples used in their calculations. We generally use 10 000. Our previous research on randomization techniques more generally suggests that authors

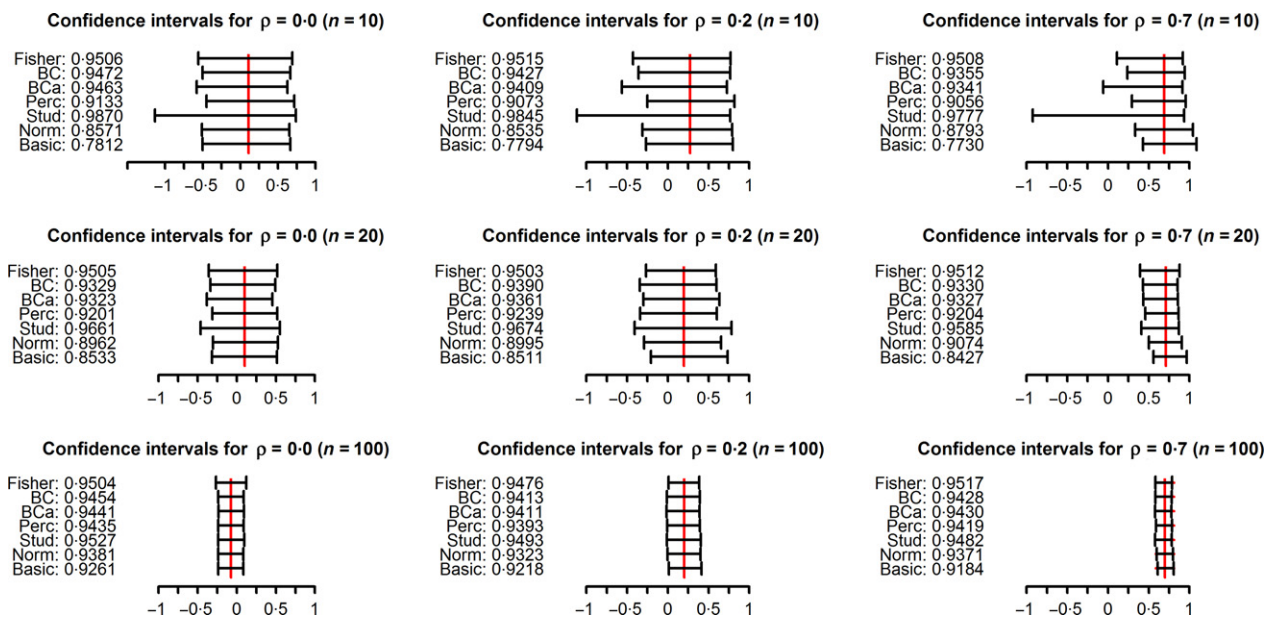

Fig. 1. Bootstrap estimates of the confidence intervals for the Pearson correlation coefficient between two samples of size n = 10, 20 and 100, drawn from a population with fixed correlation coefficient ρ. For each of three values of ρ (0·0, 0·2 or 0·7), we first drew a single sample of twenty pairs of values then calculated and plotted the estimated correlation coefficient and the 95% confidence interval based on that single sample, using six different bootstrap methods. We then drew 10 000 replicate samples and calculated 95% confidence intervals for each by all six methods (*BC*: bias corrected; *BCa*: bias corrected and accelerated, *Perc*: percentile method, *Stud*: studentized method, *Norm*: normal method and *Basic*: basic method). For each method and value of ρ, we quote the fraction of these confidence intervals that enclosed the true population value (ρ) to the left of the associated single-sample example confidence interval. We also include for comparison confidence intervals calculated by the non-bootstrapping method of Fisher's transformation, which is the method used by the most popular R function for measuring association (*cor.test*). The code used is provided as an electronic supplement.

often do not specify the number of bootstrap samples used, and amongst those that do, practice can be quite varied. Specifically in a sample of 20 recent papers using randomization techniques, we found seven used 10 000, six used 1000, one each used 500, 250 and 100, and for four papers insufficient information was given to determine the number used (Ruxton & Neuhäuser 2013). Further discussion of selection of an appropriate number of bootstraps for calculating confidence intervals can be found on pp. 101–103 of Manly (2007) and pp. 50–53 of Efron & Tibshirani (1993).

In our example, almost all the confidence intervals had coverage that was lower than the specified nominal 95%; this is a general trait of bootstrap confidence intervals from relatively small sample sizes (Chernick & LaBudde 2010), and researchers should also bear this in mind when interpreting calculated confidence intervals. Carpenter & Bithell (2000) emphasize that this problem may be more acute for 99% confidence intervals; hence, we would caution against using 99% confidence intervals calculated by bootstrapping unless the sample size is substantial ($n = 50$ as a minimum).

As regards which method to adopt, it is impossible to identify a method that gives best performance in all situations, but we recommend BCa as the method that we choose by default. BCa was specifically designed to give reasonable performance across a broader range of situations than BC but to give similar performance where BC does relatively well (Efron & Tibshirani 1993) and thus can be favoured over BC for this reason. Efron (1987) argues that BCa should be favoured over the studentized methods because both have good accuracy (differences from nominal coverage shrinking to zero quickly with increasing sample size) but BCa tends to have shorter interval lengths. In our example summarized in Fig. 1, the studentized method can provide coverage that is higher than the desired 0·95 and the confidence intervals can be very long. [DiCiccio & Efron (1996) warn that the '*bootstrap-t algorithm can be numerically unstable, resulting in very long confidence intervals*', see p. 199.] Karavarsamis *et al.* (2013) compare the basic, normal, studentized and percentile methods for the generation of confidence intervals for estimation of site occupancy and species detection probabilities in species monitoring studies. They found that the studentized method offered the most consistent performance. Inspection of their data shows that this method, as expected from our discussion above, consistently and often substantially produced coverage higher than the desired 0·95. Carpenter & Bithell (2000) provide more extensive guidance on how to select the most appropriate method, but they highlight BCa as the most consistently effective technique. BCa is also the method recommended by Crawley (2007).

Regarding software implementation, we would caution against selecting the 'all' option in the R function *boot.ci* which calculates and presents the confidence interval by all five methods available in that package. Such an approach risks (even unconsciously) selecting the confidence interval that fits the researcher's expectation.

It should be remembered that for small sample sizes, sometimes none of the methods work well. For example, Manly (2007) demonstrated this for estimation of the variance for samples of size 20 drawn from an exponential distribution, and this is true generally for estimation of higher moments from small samples drawn from very skewed distributions. It is difficult to offer clear guidance on precisely when confidence intervals based on bootstrap resamples become unreliable in this way, but in truth this is not an especial weakness of bootstrapping. For such heavily skewed distributions, small samples can often be quite unrepresentative of the underlying population and any methodology will then struggle to accurately characterize the population based on a single small sample. Statistics textbooks often suggest that $n > 30$ is sufficient for the mean to have a reasonably normal sampling distribution. Other statistics, such as the correlation coefficient used in our example, may require larger sample sizes, but few, if any, require smaller. Chernick & LaBudde (2011) warn about the potential for unreliability of bootstrapping methods when sample sizes are less than $n = 20$. Figure 1 also presents an alternative non-bootstrapping method especially designed for calculating the confidence interval for the product–moment correlation coefficient. For our example, this Fisher method (Fisher 1921; also the method implemented by the most commonly used R function for calculating correlation coefficients *cor.test*) provides coverage that is closer to the nominal level that any of the bootstrapping methods considered. This occurs because our simulated data are drawn from a bivariate normal distribution; for other distributions, bootstrapping might well be superior to the Fisher method. It should also be remembered that even for large sample sizes, there are some statistics for which bootstrapping does not provide a reliable way of estimating a confidence interval. Andrews (2000) provides a list of examples. These generally have in common estimation of a value that is on or near the boundary of allowed values, for example a situation where the true value of the statistic concerned is zero or very close to zero, but negative values are impossible. If estimating confidence intervals in such a situation, or if estimated bootstrap intervals appear odd, then Andrews (2000) offers advice on alternative techniques.

Acknowledgement

We thank Matthew Spencer and another anonymous reviewer for very helpful comments.

Data accessibility

The R code used to generate our results is available from the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.r390f (Puth, Neuhäuser & Ruxton 2015).

References

Andrews, D.W.K. (2000) Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, **68**, 399–405.

Berry, K. & Mielke, P. (2000) A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions. *Psychological Reports*, **87**, 1101–1114.

Carpenter, J. & Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, **19**, 1141–1164.

Chernick, M.R. & LaBudde, L.A. (2010) Revisiting qualms about bootstrap confidence intervals. *American Journal of Mathematical and Management Sciences*, **29**, 437–456.

Chernick, M.R. & LaBudde, L.A. (2011) *Bootstrap Methods with Applications to R*. Wiley, New York.

Crawley, M.J. (2007) *The R Book*. Wiley, New York.

Davison, A.C. & Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

DiCiccio, T.J. & Efron, B. (1996) Bootstrap confidence intervals (with discussion). *Statistical Science*, **11**, 189–228.

Efron, B. (1987) Better bootstrapping confidence intervals (with discussion). *Journal of the American Statistical Society*, **82**, 171–200.

Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fisher, R.A. (1921) On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.

Karavarsamis, N., Robinson, A.P., Hepworth, G., Hamilton, A.J. & Heard, G.W. (2013) Comparison of four bootstrap-based interval estimators of species occupancy and detection probabilities. *Australian & New Zealand Journal of Statistics*, **55**, 235–252.

Manly, B.F.J. (2007) *Randomisation, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Chapman & Hall, New York.

Puth, M.-T., Neuhäuser, M. & Ruxton, G.D. (2015) Data from: On the variety of methods for calculating confidence intervals by bootstrapping. *Dryad Digital Repository*, doi:10.5061/dryad.r390f.

Ruxton, G.D. & Neuhäuser, M. (2013) Improving the reporting of *P*-values by randomisation methods. *Methods in Ecology & Evolution*, **4**, 1033–1036.

Taborsky, M. (2010) Sample size in the study of behaviour. *Ethology*, **116**, 185–202.

Received 10 October 2014; accepted 3 April 2015
Handling Editor: Dylan Childs

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Comparison of different bootstrap methods.