A Resampling Method Called the Bootstrap

Monte Carlo and bootstrap methods are both computer intensive methods used frequently is applied statistics. The bootstrap is a type of Monte Carlo method applied based on observed data (Efron and Tibshirani 1993, Mooney and Duval 1993). The bootstrap was described by Bradley Efron (1979) and he has written much about the method and its generalizations since then. Thousands of papers have been written on the bootstrap in the past 2 decades and it has found very wide use in applied problems. The bootstrap can be used for several purposes, here we we focus on robust estimation of sampling variances or standard errors and (asymmetrical) confidence intervals. It has found use in estimation of model selection frequencies and a variety of other applications.

The bootstrap has enormous potential for the biologist with programming skills; however, its computer intensive nature will continue to hinder its use. We believe that at least 1,000 bootstrap reps are needed in many applications. Often 10,000 reps are needed for some aspects of model selection. In extreme cases, reliable results could take days of computer time to apply the bootstrap to complex data analysis cases.

The fundamental idea of the model-based sampling theory approach to statistical inference is that the data arise as a sample from some conceptual probability distribution, f. Uncertainties of our inferences can be measured if we can estimate f. There are ways to construct a nonparametric estimator of (in essence) f from the sample data. The most fundamental idea of the bootstrap method is that we compute measures of our inference uncertainty from that estimated sampling distribution of f. However, in practical application, the bootstrap means using some form of resampling *with replacement* from the actual data, $\underline{x}$, to generate B bootstrap samples, $\underline{x}^*$. Often, the data (sample) consist of n independent units and it then suffices to take a simple random sample of size n, *with replacement*, from the n units of data, to get one bootstrap sample (i.e. "rep"). However, the nature of the correct bootstrap data re-sampling can be more complex for more complex data structures.

The set of B bootstrap samples is a proxy for a set of B independent real samples from f (in reality we have only one actual sample of data). Properties expected from replicate real samples are inferred from the bootstrap samples by analyzing each bootstrap sample exactly as we first analyzed the real data sample. From the set of results of sample size B we measure our inference uncertainties from sample to (conceptual) population (see figure). The bootstrap can work well for large sample sizes (n), but may not be reliable for small n (say 5, 10 or even 20), regardless of how many bootstrap samples, B, are used.

Estimation of the Sampling Variance

In many cases one can derive an estimator of the sampling variance of an estimator from general likelihood theory. In other cases, an estimator may be difficult to derive or many not exist in closed form. For example, the finite rate of population change (λ) can be derived from a Leslie population projection matrix (a function of age-specific fecundity and age-specific, conditional survival probabilities). The bootstrap is handy for variance estimation in such cases.

Consider a sample of weights of 27 rats ($n = 27$); the data are

57 60 52 49 56 46 51 63 49 57 59 54 56 59 57 52 52 61 59 53 59 51 51 56 58 46 53.

The sample mean of these data = 54.6667, standard deviation = 4.5064 with cv = 0.0824. For illustration, what if we wanted an estimate of the standard error of cv. Clearly, this would be nonstandard; however is represents a way to illustrate the bootstrap. We owe this example to Ken Burnham.

First, we draw a random subsample of size 27 *with replacement*. Thus, while a weight of 63 appears in the actual sample, perhaps it would not appear in the subsample; or is could appear more than once. Similarly, there are 3 occurrences of the weight 57 in the actual sample, perhaps the resample would have, by chance, no values of 57. The point here is that a random sample of size 27 is taken *with replacement* from the original 27 data values. This is the first bootstrap resample (b=1). From this resample, one computes $\hat{\mu}$, the $\hat{se}(\hat{\mu})$ and the cv and stores this in memory.

Second, the whole process is repeated B times (where we will let B = 1,000 reps for this example). Thus, we generate 1000 resample data sets (b = 1, 2, 3, ..., 1000) and from *each* of these we compute $\hat{\mu}$, $\hat{se}(\hat{\mu})$ and the cv and store these values.

Third, we obtain the standard error of the cv by taking the standard deviation of the 1000 cv values (corresponding to the 1000 bootstrap samples). The process is simple. In this case, the standard error is 0.00917.

Just as the analysis of a single data set can have many objectives, the bootstrap can be used to provide insight into a host of questions. For example, for each bootstrap rep one could compute and store the conditional variance-covariance matrix, goodness-of-fit values, the estimated variance inflation factor, model selected, confidence interval width or other quantities.

The illustration of the bootstrap on the rat data is called a nonparametric bootstrap as nothing is assumed (like a parametric distribution) about the underlying process that generated the data. We only assume that the data in the original sample were "representative" and that sample size was moderately large. The parametric bootstrap is frequently used and allows assessment of bias. The use of the parametric bootstrap will be illustrated by the estimation of the variance inflation factor, $\hat{c}$.

Consider an open population capture-recapture study in a setting where the investigators suspect a lack of independence because of the way that family groups were captured in the field. Data analysis reveals $\chi^2_{gof}/df = 3.2$. The investigators suspected some extra-binomial variation, but are surprised by the large estimate of $\hat{c}$. They suspect the estimator is biased high and decide to use a parametric bootstrap to investigate their suspicion. They realize that program *RELEASE* can be used to do Monte Carlo simulations and output a file with the goodness of fit statistics.

They input the MLEs from the real data into *RELEASE* as if they were parameters (ϕ_j and p_j) and use the numbers of new releases in the field data as input. Then the amount of extra binomial variation (called EBV in *RELEASE*) is specified. In this illustration, let EBV $\equiv$ 2. They then run 1000 Monte Carlo reps and obtain the information on the estimated variance inflation factor for each

reps. The average of these 1000 values gives $\hat{E}(\hat{c})$; the question is "does $\hat{E}(\hat{c}) = 2$." It is likely that the estimator $\hat{c}$ is positively biased. This result provides insight to the investigators on what do to about overdispersion in their data.

This bootstrap is parametric in that parameters were specified (in this case, from the MLEs from real data that were available) and used in a (known) model to generate Monte Carlo data. The nonparametric bootstrap does not require parameters nor a model and can address some statistical issues (e.g., precision).

Confidence Intervals

Confidence intervals can be computed in the usual way, using either a parametric or nonparametric bootstrap,

$$cv \pm 2\hat{se}(cv) .$$

This gives a 95% interval of (0.0641, 0.1007) for the rat data. However, the sampling distribution may be non-normal and a more robust interval might be required. Again, the bootstrap is a simple approach. In this case one sorts the B=1000 estimates of the cv in ascending order and selects the values that cut off the lower and upper 2.5 percentiles! Thus, the resulting interval might be asymmetric.

In the rat cv, the bootstrap 95% confidence interval is (0.0627, 0.0981). This interval is about the same width as the traditional approach, but shifted a bit toward 0. Incidentally, the mean of the 1000 bootstrap samples was 0.0802 (compared to the sample mean of 0.0824).

B = 1000 is usually adequate for the estimation of the sampling variance or standard deviation; however, good estimates of confidence intervals often require B = 5000 or more. Program DISTANCE has an option for nonparametric bootstrapping line and point transect data.

The bootstrap has been used to set confidence intervals on the median and mean life span. It is conceptually simple and finding very wide spread use in applied statistics. Biologists planning a career in research or teaching should be familiar with the bootstrap (and the delta method). There is a very large literature on the bootstrap; see Efron and Tibshirani (1993) for a so-called introduction to the subject and a large list of references. In fact, many valid applications of the bootstrap are tricky (even multiple linear regression), so some care is required in more complex settings!

Efron, B, and Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Monographs on Statistics and Applied Probability, No. 57. Chapman and Hall, London. 436pp.