
Appendix

This appendix presents some background on parametric likelihood inference and bootstrap methods. Both methods appeal to asymptotic characterizations of errors and other quantities, and so some asymptotic notions are reviewed first.

A.1 Order and stochastic order notation

The notations $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$ and $o_p(\cdot)$ are used to describe the asymptotic magnitude of statistical quantities.

Consider two sequences of real numbers a_n and b_n where n ranges through integers larger than 1. We say that $a_n = O(b_n)$ if there exists $B < \infty$ with

$$\limsup_{n \rightarrow \infty} \frac{|a_n|}{|b_n|} \leq B. \quad (\text{A.1})$$

Put another way, for some B , there are only finitely many n with $a_n > Bb_n$.

The expression $p_n = q_n + O(b_n)$ means that $p_n - q_n = O(b_n)$. Similarly, when $o(\cdot)$, $O_p(\cdot)$, or $o_p(\cdot)$ as defined below, are added to the right side of an equation, it means that the right side minus the left side is of that order. Here we consider limits indexed by n tending to ∞ , but these notions also apply in other limits, such as quantities indexed by $\epsilon \rightarrow 0^+$.

For (A.1), it is sufficient to have $\lim_{n \rightarrow \infty} |a_n|/|b_n| \leq B$. The more general expression (A.1) also covers cases where $|a_n/b_n|$ is eventually bounded without converging to a limit.

As an example, suppose that $a_n = \gamma^2 n^{-1} + \kappa n^{-2} + n^{-3}$, where γ and κ are finite. Then $a_n = O(n^{-1})$. Here γ and κ are unknowns, such as population skewness and kurtosis, that could vary from one problem instance to another. Suppose that sometimes $\gamma = 0$. Then it is still true that $a_n = O(n^{-1})$, but now it is also true that $a_n = O(n^{-2})$. It is important to remember that a quantity of one order could also be of a smaller order. If $\gamma = \kappa = 0$, then $a_n = O(n^{-3})$, but there are no γ and κ values for which $a_n = O(n^{-4})$.

As a second example, consider $a_n = \mu^2 + n\sigma^2$ where μ and σ are finite. Now $a_n = O(n)$, and when $\sigma = 0$, then $a_n = O(1)$. In this example, a_n is not going to infinity faster than n , if at all, whereas in the previous example a_n is going to zero no slower than n^{-1} .

The notation $a_n = o(b_n)$ means

$$\limsup_{n \rightarrow \infty} \frac{|a_n|}{|b_n|} = 0. \quad (\text{A.2})$$

The quantity $\gamma^2 n^{-1} + \kappa n^{-2} + n^{-3}$ is $o(n^{-1/2})$ and also $O(n^{-1})$, but it is not $o(n^{-1})$ unless $\gamma = 0$.

The o_p and O_p notations are used to describe bounds on a quantity whose magnitude is random. Suppose, for example, that Y_n are independent exponential random variables with mean 1. The quantity $Z_n = n^{-1}Y_n$ seems to be tending to zero like $1/n$, but Z_n is not $O(1/n)$. There is no finite B such that quantity $Z_n/(1/n) = Y_n$ is larger than B at most a finite number of times.

We say that $X_n = O_p(Y_n)$ if for any $\epsilon > 0$ there is a $B = B_\epsilon < \infty$ such that

$$\limsup_{n \rightarrow \infty} \Pr(|X_n| > B|Y_n|) < \epsilon. \quad (\text{A.3})$$

Condition (A.3) allows $\Pr(|X_n| > B|Y_n|) \geq \epsilon$ to hold for at most a finite number of n . In the previous example it is indeed true that $Z_n = O_p(1/n)$. The order $O_p(\cdot)$ applies to quantities that are $O(\cdot)$ apart from exceptions that we can make as improbable as we please for large n .

Finally, $X_n = o_p(Y_n)$ if for any $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \Pr(|X_n| > \epsilon|Y_n|) = 0. \quad (\text{A.4})$$

For example, let $X_n = Y_n/n^{1+\delta}$ for $\delta > 0$ and Y_n independent exponential random variables with mean 1. Then $X_n = o_p(n^{-1})$.

A.2 Parametric models

When we make some calculations on data, we usually know that there is some uncertainty in our answer. The data could have come out differently, and then our answer almost certainly would have been different. But we also usually feel that there are reasonable limits to how different the answer might have been.

Probability models are widely used to provide a notion of the true value of a quantity that can be different from the value we have computed. Such models also allow us to quantify the uncertainty in our answers, help us to decide what to compute from the data, and even help us decide how to gather our data.

In a probability model we suppose that our data are the observed values, say $x_1, \dots, x_n \in \mathbb{R}^d$ of corresponding random variables $X_1, \dots, X_n \in \mathbb{R}^d$. In a parametric probability model, the joint distribution of X_1, \dots, X_n takes a known form, involving the data and a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$. By contrast non-parametric models do not assume that a finite dimensional parameter indexes the distribution of the data.

Here we will consider only models in which X_i are independent and identically distributed and the sample size n is not random. Then specifying the distribution

of X_1 specifies the distribution of the whole sample. Parametric models and likelihood methods extend to more general settings.

As an example, if X_i take nonnegative integer values, then one such model, the Poisson distribution, has

$$p(x; \theta) = \Pr(X = x; \theta) = \frac{e^{-\theta} \theta^x}{x!}, \quad x = 0, 1, \dots \quad (\text{A.5})$$

for $\theta > 0$, while for X_i taking nonnegative real values, the exponential distribution has

$$\Pr(X \leq x; \theta) = \begin{cases} 1 - \exp(-x\theta), & 0 \leq x < \infty \\ 0, & x < 0, \end{cases}$$

for $\theta > 0$, so the probability density function of X is $f(x; \theta) = \theta \exp(-x\theta) 1_{x \geq 0}$.

When we compute a value from the sample, we can usually identify it with a corresponding feature of the distribution of X_i . Perhaps that feature is the value we would get as $n \rightarrow \infty$, or perhaps it is the average of the values we would get in a large number of independently generated samples of size n . This feature is necessarily a function of θ and so interest switches to learning θ from X_1, \dots, X_n .

Any function of X_1, \dots, X_n that we compute is also a random variable with a distribution parameterized by θ . This allows us to quantify the uncertainty of our estimates within the parametric model. We may then seek a method of estimation that minimizes some measure of this uncertainty.

A good model can be a wonderfully effective tool for organizing statistical inference, but choosing a parametric model for applied work is a subtle task. Sometimes a parametric model is thought to be a faithful description of the mechanism generating the data. More often, a model is adopted because it is mathematically convenient and is thought to capture the important features of the problem.

One of the most vexing issues with parametric models is testing whether a given model is compatible with our data. When a goodness of fit test fails to reject our model, it may simply mean that the test was not powerful enough, perhaps because the sample size was too small. Conversely, when a test does reject our model, it may have identified a flaw with a negligible effect on the conclusions we would have drawn using the model. Finally, there are those cases in which a test shows that our model fits badly in a way that affects our answers. Then we may have to seek a correction term, or look for a new model.

One of the best blessings of nonparametric methods is that they reduce the need for goodness of fit testing. There can be a cost in using a nonparametric model for a problem that fits a parametric description. Some nonparametric methods make less efficient use of the data than do parametric methods. When using a nonparametric estimator or test, we should consider its efficiency or power, respectively.

A.3 Likelihood

Given a parametric model and some data, the likelihood function is the probability of getting the observed values from the assumed model, taken as a function of the parameter. The likelihood is thus

$$L(\theta; x_1, \dots, x_n) = L(\theta) = \prod_{i=1}^n p(x_i; \theta)$$

for discrete data and

$$L(\theta; x_1, \dots, x_n) = L(\theta) = \prod_{i=1}^n f(x_i; \theta) dx_i$$

for continuous data, where X_i has been observed to lie in a small set of volume dx_i near the value x_i . We also write $L(\theta; X_1, \dots, X_n)$ for a random likelihood taking the value above when $X_i = x_i$.

In Bayesian analysis, our model takes θ to be a random variable with a prior distribution π_0 , and then the posterior distribution of θ is

$$\pi_1(\theta | X_1, \dots, X_n) \propto L(\theta)\pi_0(\theta).$$

Specialized numerical and sampling techniques are available to compute conclusions from $L(\theta)\pi_0(\theta)$. Just as with parametric probability models for the data, judgment is required to make a good choice of π_0 .

A frequentist analysis postulates an unknown true value of θ . When we need to distinguish a generic value from the true value, we denote the latter by θ_0 . The method of maximum likelihood estimates θ_0 by finding the value (or sometimes set of values) $\hat{\theta}$ that maximize L :

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; X_1, \dots, X_n).$$

A maximizer $\hat{\theta}$ is called a maximum likelihood estimate (MLE) of θ_0 . Intuitively, it gives the best explanation of the data we got, by making that data most probable.

Usually it is easier to work with the log likelihood function

$$\ell(\theta) = c + \sum_{i=1}^n \log f(x_i; \theta),$$

where c depends on the dx_i but not, we assume, on θ . We will use the notation of the continuous data case; the discrete data case is very similar, starting with $\ell(\theta) = \sum_{i=1}^n \log p(x_i; \theta)$. For the most widely applied likelihoods, $\hat{\theta}$ is found by solving the score equation $\partial \ell(\theta) / \partial \theta = 0$, which may be written as solving estimating equations $\sum_{i=1}^n m(x_i, \theta) = 0$, where

$$m(x, \theta) = \frac{\partial}{\partial \theta} \frac{f(x; \theta)}{f(x; \theta)}.$$

For the Poisson distribution (A.5), $\hat{\theta}$ equals \bar{X} , so the sample mean is used

to estimate the population mean. For the double exponential distribution, with density $f(x; \theta) = \exp(-|x - \theta|)/2$, the population mean and median coincide at θ , and we find that the median of X_i is the MLE of θ . Maximum likelihood estimation is also widely used to construct estimates of quantities that we might otherwise not know how to estimate.

The value $\hat{\theta}$ is unlikely to match θ_0 exactly, and furthermore, a value $\tilde{\theta}$ with $L(\tilde{\theta})$ very close to $L(\hat{\theta})$ would seem to be nearly as good as $\hat{\theta}$. We define the likelihood ratio function by $R(\theta) = L(\theta)/L(\hat{\theta})$. One way to separate reasonable from unreasonable values of θ is to order them by $R(\theta)$ and consider $C = \{\theta \mid R(\theta) > r\}$ to fit the data better than other values of θ .

The usual way to pick r is to aim for a given probability that $\theta_0 \in C$. Wilks's theorem has that $-2 \log R(\theta_0) \rightarrow \chi_{(p)}^2$ in distribution as $n \rightarrow \infty$, where p is the dimension of Θ . This result holds in considerable generality, though there are some exceptions. When it holds, we can use $C^{1-\alpha} = \{\theta \mid R(\theta) \geq r_0\}$ as an approximate $1 - \alpha$ confidence region for θ_0 with $r_0 = \exp(-\chi_{(p)}^{2, 1-\alpha}/2)$. A Bartlett correction replaces r_0 by $\exp(-(1 + a/n)\chi_{(p)}^{2, 1-\alpha}/2)$, where a judiciously chosen scalar a can improve the rate at which $\Pr(\theta_0 \in C^{1-\alpha})$ tends to $1 - \alpha$ as $n \rightarrow \infty$.

Computational shortcuts to $C^{1-\alpha}$ are widely used. For large n the log likelihood ratio is very nearly quadratic around $\hat{\theta}$, under regularity conditions. A Taylor approximation gives $\ell(\theta) \doteq \ell(\hat{\theta}) - (1/2)(\theta - \hat{\theta})' \hat{I}(\theta - \hat{\theta})$ near $\hat{\theta}$, where \hat{I} is the Hessian of ℓ at $\hat{\theta}$. This approximation motivates the use of the ellipsoid

$$\left\{ \theta \mid (\theta - \hat{\theta})' \hat{I}(\theta - \hat{\theta}) \leq \chi_{(p)}^{2, 1-\alpha} \right\} \quad (\text{A.6})$$

as a confidence region, and under standard assumptions, the set (A.6) has asymptotic probability $1 - \alpha$ of containing θ_0 .

Equation (A.6) is one form of the Wald confidence region. The Fisher information in X_1 is defined as

$$I_1(\theta) = \int \left(\frac{\partial}{\partial \theta} \log f(x; \theta) \right) \left(\frac{\partial}{\partial \theta} \log f(x; \theta) \right)' f(x; \theta) dx.$$

As $n \rightarrow \infty$ we usually have $n^{-1} \hat{I} \rightarrow I_1(\theta_0)$, and another version of the Wald region uses $nI_1(\hat{\theta})$ instead of \hat{I} . Another important large sample confidence region due to Rao works with a Taylor approximation around a hypothesized value of θ , not the MLE.

Sometimes we are interested in some, but not all of the parameters. Let us retain θ for the parameters of interest, and suppose that there is an additional parameter ν , possibly a vector, used to complete the definition of the probability of the data. Then the likelihood is $L(\theta, \nu)$ involving θ and the nuisance parameter ν . The MLE of θ is found by maximizing L over θ and ν jointly yielding $\hat{\theta}$ and $\hat{\nu}$.

Likelihood ratios for θ alone are complicated by the presence of ν . If we knew the true value ν_0 , then it would be natural to use $R_{\nu_0}(\theta) = L(\theta, \nu_0)/L(\hat{\theta}, \nu_0)$ as a likelihood ratio function. There is no practical difference between a likelihood

with a known nuisance parameter and a likelihood without a nuisance parameter. But the idea of a known nuisance parameter suggests the following: we could plug in $\hat{\nu}$ for ν_0 and use $R_{\hat{\nu}}(\theta) = L(\theta, \hat{\nu})/L(\hat{\theta}, \hat{\nu})$ as a likelihood ratio function for θ . To see the problem with plugging in a value like $\hat{\nu}$, consider a parameter pair $(\tilde{\theta}, \tilde{\nu})$ with $L(\tilde{\theta}, \tilde{\nu})$ just slightly smaller than $L(\hat{\theta}, \hat{\nu})$, and for which $L(\tilde{\theta}, \tilde{\nu}) \ll L(\hat{\theta}, \hat{\nu})$. A confidence region based on $R_{\hat{\nu}}$ might fail to include $\tilde{\theta}$ even though it belongs to a parameter pair that fits the data nearly as well as the MLE does. The profile likelihood function

$$R_{\text{PrO}}(\theta) = \frac{\max_{\nu} L(\theta, \nu)}{L(\hat{\theta}, \hat{\nu})},$$

remedies this flaw in $R_{\hat{\nu}}$, and is the most widely used way to construct a likelihood ratio function for an individual parameter in the presence of nuisance parameters. Letting $\hat{\nu}(\theta) = \arg \max_{\nu} L(\theta, \nu)$, we may write the profile likelihood as $R_{\text{PrO}}(\theta) = L(\theta, \hat{\nu}(\theta))$. A version of Wilks's theorem holds for profile likelihoods.

For well-behaved parametric models, the MLE is a particularly good estimator. Typically there is an information inequality, one form of which shows that any unbiased estimator T of θ has a variance matrix at least as large as the Cramér-Rao bound $(nI_1(\theta))^{-1}$. Under the usual assumptions, $\sqrt{n}(\hat{\theta} - \theta_0)$ has asymptotic distribution $N(0, (nI_1(\theta_0))^{-1})$. The MLE $\hat{\theta}$ typically achieves the lower bound on variance, with an asymptotically negligible bias of order $1/n$.

A.4 The bootstrap idea

The bootstrap is a powerful method for estimating statistical uncertainties. Suppose that we have computed a statistic $\hat{\theta} = T(F_n)$ for which the corresponding true value is $T(F_0)$. The variance of $\hat{\theta}$ is yet another property of the unknown distribution F_0 . Call it $\text{VT}(F_0)$. If we knew F_0 , then we might compute $\text{VT}(F_0)$, by calculus, or by a numerical method such as Monte Carlo simulation.

Of course, if we knew F_0 we might instead use simulations to find $T(F_0)$, so the interesting case is what to do when we do not know F_0 . The bootstrap is based on a plug-in principle: make your best guess \hat{F} for F_0 , and plug it in, using $\text{VT}(\hat{F})$ as the estimate of $\text{VT}(F_0)$. Very often, the guess we use for F_0 is the ECDF F_n . Because this ECDF is a nonparametric maximum likelihood estimate (NPMLE), we find that the bootstrap estimate of VT is the NPMLE $\widehat{\text{VT}}$.

To compute $\text{VT}(F_n)$ by simulation, we sample with replacement, as follows. Let $J(i, b)$ be independent uniform random draws from the set $\{1, 2, \dots, n\}$, for $i = 1, \dots, n$ and $b = 1, \dots, B$. Then put $X_i^{*b} = X_{J(i,b)}$, let F_n^{*b} be the empirical distribution of $X_1^{*b}, \dots, X_n^{*b}$, and take $T^{*b} = T(F_n^{*b})$, the statistic T applied to

resampled data $X_1^{*b}, \dots, X_n^{*b}$. Now

$$\text{VT}(F_n) \doteq \frac{1}{B} \sum_{b=1}^B (T^{*b} - \bar{T}^*)^2 \quad (\text{A.7})$$

where $\bar{T}^* = (1/B) \sum_{b=1}^B T^{*b}$. This computation may well be easier than that required to find $\text{VT}(F)$ for an arbitrary known distribution F .

It is quite surprising that we can sample our original data over and over and obtain a useful error estimate. The reason that the bootstrap can work is that, for large enough n , F_n becomes close to F_0 and, if VT is continuous near F_0 , then $\text{VT}(F_n)$ is close to $\text{VT}(F_0)$. The flatter VT is, as a function of F , the better we could expect the bootstrap to work. As an extreme case, if VT were constant in F we could compute the variance of $\hat{\theta}$ exactly.

The approximation in (A.7) converges to $\text{VT}(F_n)$ as $B \rightarrow \infty$, by the law of large numbers, under the very mild assumption that $\text{VT}(F_n)$ is finite. If, under sampling from F_n , the fourth moment of T is finite, then the error in (A.7) is $O_p(B^{-1/2})$. In practice, we might take B large enough that we feel safe that the error in (A.7) is negligible, either absolutely, or relative to $\text{VT}(F_n) - \text{VT}(F_0)$.

A.5 Bootstrap confidence intervals

Consider a scalar statistic $T(F_n) \in \mathbb{R}$ used to estimate $T(F_0)$. A $100(1 - \alpha)\%$ confidence interval for T is a pair of random numbers $L = L(F_n)$ and $U = U(F_n)$ such that

$$\Pr(L(F_n) \leq T(F_0) \leq U(F_n)) = 1 - \alpha. \quad (\text{A.8})$$

Notice that equation (A.8) describes the probability that a random interval $[L, U]$ contains a nonrandom value $T(F_0)$. By contrast, in a prediction interval, a random quantity lies between two fixed endpoints with a given probability.

Equation (A.8) is supposed to hold for all F_0 . Exact nonparametric confidence intervals do not exist outside of a few special cases. In particular, they do not exist for the mean. In practice, we use asymptotic confidence intervals with

$$\Pr(L(F_n) \leq T(F_0) \leq U(F_n)) = 1 - \alpha + o(1) \quad (\text{A.9})$$

as $n \rightarrow \infty$.

In Section A.4 an unknown variance under sampling from F_0 was estimated by the corresponding variance under sampling from F_n . Bootstrap confidence intervals can be constructed by estimating the distribution of an approximately pivotal quantity under F_0 by its distribution under F_n .

Let us write $\mathcal{L}(T(F_n) \mid F_0)$ for the distribution of $T(F_n)$ when X_1, \dots, X_n have distribution F_0 . Similarly let $\mathcal{L}(T(F_n^*) \mid F_n)$ be the common distribution of each T^{*b} drawn on bootstrap samples from F_n .

The percentile method takes the resampled statistics at face value, using the

approximation

$$\mathcal{L}(T(F_n) | F_0) \doteq \mathcal{L}(T(F_n^*) | F_n).$$

Let us order the resampled T^{*b} values getting $T^{*(1)} \leq T^{*(2)} \leq \dots \leq T^{*(B)}$. For large B essentially 95% of the resampled values are between $L_{\text{perc}} = T^{*(.025B)}$ and $U_{\text{perc}} = T^{*(.975B)}$, where for noninteger values $0.025B$ and $0.975B$, some rounding or interpolation is applied.

Having seen that

$$\Pr(L_{\text{perc}} \leq T(F_n^*) \leq U_{\text{perc}} | F_n) = 0.95 \quad (\text{A.10})$$

we could estimate that

$$\Pr(L_{\text{perc}} \leq T(F_n) \leq U_{\text{perc}} | F_0) = 0.95. \quad (\text{A.11})$$

In equation (A.10) we have found by simulation a prediction interval for $T(F_n^*)$ in sampling from $T(F_n)$. Equation (A.11) uses this prediction interval as an estimate of a prediction interval for $T(F_n)$ when sampling from F_0 .

Because the endpoints L_{perc} and U_{perc} are computed from the data, they are indeed random. And it turns out that in reasonable generality equation (A.9) holds for L_{perc} and U_{perc} , so that the percentile interval can be used for confidence statements. We return to this point below.

The bias-corrected percentile method is based on the approximation

$$\mathcal{L}(T(F_n) - T(F_0) | F_0) \doteq \mathcal{L}(T(F_n^*) - T(F_n) | F_n). \quad (\text{A.12})$$

Equation (A.12) describes an unknown quantity $T(F_n) - T(F_0)$ whose distribution we can approximate through simulation of known quantities $T(F_n^*) - T(F_n)$. Quantities containing the unknown, but having a known distribution, are called pivots. Where the distribution is approximately known, the quantities are approximate pivots.

The 0.025 and 0.975 quantiles of $T(F_n^*) - T(F_n)$ are, of course, $T^{*(.025B)} - T(F_n)$ and $T^{*(.975B)} - T(F_n)$. Therefore, the probability is approximately 0.95 that

$$T^{*(.025B)} - T(F_n) \leq T(F_n) - T(F_0) \leq T^{*(.975B)} - T(F_n), \quad (\text{A.13})$$

and rearranging to get $T(F_0)$ in the middle, we get that

$$2T(F_n) - T^{*(.975B)} \leq T(F_0) \leq 2T(F_n) - T^{*(.025B)} \quad (\text{A.14})$$

holds with approximate probability 0.95. The bias-corrected method has endpoints $L_{\text{bc}} = 2T(F_n) - U_{\text{perc}}$ and $U_{\text{bc}} = 2T(F_n) - L_{\text{perc}}$. It is the percentile interval flipped around $T(F_n)$.

We can use the bias-corrected interval to provide an explanation of the percentile interval. It commonly happens that for large n , the approximate pivot $T(F_n) - T(F_0)$ has a nearly symmetric distribution. Then the bias-corrected confidence interval is nearly unchanged by flipping it around $T(F_n)$. The percentile interval was derived using a more general symmetry assumption. Suppose that $\mathcal{L}(\phi(T(F_n)) - \phi(T(F_0)) | F_0)$ has a symmetric distribution, for some possibly

unknown monotone transformation ϕ . Then the plug-in idea yields the percentile interval.

There are other choices for the pivot. In some settings it may be more natural to use

$$\mathcal{L}\left(\frac{T(F_n)}{T(F_0)} \mid F_0\right) \doteq \mathcal{L}\left(\frac{T(F_n^*)}{T(F_n)} \mid F_n\right) \quad (\text{A.15})$$

or

$$\mathcal{L}\left(\frac{T(F_n) - T(F_0)}{S(F_n)} \mid F_0\right) \doteq \mathcal{L}\left(\frac{T(F_n^*) - T(F_n)}{S(F_n^*)} \mid F_n\right), \quad (\text{A.16})$$

where $S(F_n)$ is an estimate, such as a standard error, of how large $T(F_n) - T(F_0)$ might be.

Inverting the pivot in equation (A.16) gives the bootstrap-t (or percentile-t) method. The bootstrap-t method provides especially accurate confidence intervals for the univariate mean. In that context $T(F_0) = E(X)$, $T(F_n) = \bar{X}$,

$$S(F_n)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad \text{and}$$

$$S(F_n^*)^2 = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2.$$

The explanation is that Student's t statistic has a distribution that depends relatively weakly on F , making it a good approximate pivot.

The bootstrap-t interval can be applied in other settings, but the limiting factor is the availability of a good denominator statistic S . If one can afford to bootstrap the bootstrap, then $S(F_n)$ can be estimated as the square root of $\text{VT}(F_n)$ and each $S(F_n^{*b})$ can be estimated as a bootstrap variance $\text{VT}(F_n^{*b})$.

Bootstrap calibration of empirical log likelihood ratios works so well because the distribution of those ratios is only weakly dependent on the underlying data distribution.

A.6 Better bootstrap confidence intervals

A major focus of bootstrap research has been the construction of confidence intervals with one-sided coverage errors that are $O(n^{-1})$. The percentile and bias-corrected percentile methods are typically $O(n^{-1/2})$ accurate on one-sided inferences, though their two-sided inferences give $O(n^{-1})$ errors.

The bootstrap-t intervals are this accurate, but they do not respect transformations. The endpoints of the interval for $\exp(\theta)$ do not equal exponentiated endpoints for the bootstrap-t interval for θ .

The $100(1 - \alpha)\%$ bias-corrected accelerated, or BC_a , interval is approximately transformation respecting and has one-sided coverage errors that are $O(n^{-1})$. The derivation is based on a complicated approximately normal pivot. See the references in Section A.7. The BC_a interval has endpoints $L_{\text{bca}} = T^{*(\alpha_1 B)}$ and

$U_{\text{bca}} = T^{*(\alpha_2 B)}$ where

$$\alpha_1 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z^{\alpha/2})} \right) \quad (\text{A.17})$$

$$\alpha_2 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z^{1-\alpha/2})} \right), \quad (\text{A.18})$$

wherein \hat{z}_0 and \hat{a} are the bias correction and acceleration constant (defined below), Φ is the standard normal cumulative distribution function, and $z^p = \Phi^{-1}(p)$. For a 95% confidence interval $z^{\alpha/2} = -1.96$ and $z^{1-\alpha/2} = 1.96$.

The bias correction factor is

$$\hat{z}_0 = \Phi^{-1} \left(\frac{1}{B} \sum_{b=1}^B 1_{T^{*b} < T(F_n)} \right).$$

It vanishes if exactly half of the bootstrap samples were smaller than $T(F_n)$, otherwise it shifts the interval.

To define the acceleration constant, consider data sets $F_{n,-i}$ consisting of the $n - 1$ original observations other than observation i . Let $T_{-i} = T(F_{n,-i})$ and $\bar{T}_{-\bullet} = (1/n) \sum_{i=1}^n T_{-i}$. Then the acceleration constant is

$$\hat{a} = \frac{\sum_{i=1}^n (T_{-i} - \bar{T}_{-\bullet})^3}{6 \left[\sum_{i=1}^n (T_{-i} - \bar{T}_{-\bullet})^2 \right]^{3/2}},$$

which makes a skewness adjustment.

The ABC method constructs approximate endpoints for the BC_a method. Instead of resampling, it makes a Taylor expansion based on the behavior of $T(F)$ for distributions F that are close to F_n and put all their probability on the sample. For $\epsilon \in (-1/n, 1)$, let $F_{i,\epsilon}$ be the distribution $(1 - \epsilon)F_n + \epsilon\delta_{X_i}$. The α confidence limit is

$$T_{\text{ABC}}^\alpha = T \left(F_n + \frac{\tilde{z}_\alpha}{(1 - a\tilde{z}_\alpha)^2} \sum_{i=1}^n k_i \delta_{X_i} \right), \quad (\text{A.19})$$

as described below. It is the value of T at a specially chosen reweighting of the data. For an asymptotic $100(1 - \alpha)\%$ central confidence interval, take $L_{\text{ABC}} = T_{\text{ABC}}^{\alpha/2}$ and $U_{\text{ABC}} = T_{\text{ABC}}^{1-\alpha/2}$. The necessary quantities are defined through

$$l_i = \frac{d}{d\epsilon} T(F_{i,\epsilon})|_{\epsilon=0}, \quad q_{ii} = \frac{d^2}{d\epsilon^2} T(F_{i,\epsilon})|_{\epsilon=0}, \quad v_L = \frac{1}{n^2} \sum_{i=1}^n l_i^2,$$

$$a = \frac{1}{6} \frac{\sum_{i=1}^n l_i^3}{\left(\sum_{i=1}^n l_i^2 \right)^{3/2}}, \quad b = \frac{1}{2n^2} \sum_{i=1}^n q_{ii}, \quad k_i = n^{-2} v_L^{-1/2} l_i,$$

and

$$\tilde{z}_\alpha = z_\alpha + a + c - bv_L^{-1/2}, \quad c = \frac{1}{2v_L^{1/2}} \frac{d^2}{d\epsilon^2} T(F_n + \epsilon k)|_{\epsilon=0},$$

taking $F_n + \epsilon k$ as a shorthand for $F_n + \epsilon \sum_{i=1}^n k_i \delta_{X_i}$. The derivatives may be computed numerically as divided differences, and z_α is defined through $\Pr(Z \leq z_\alpha) = \alpha$ for $Z \sim N(0, 1)$. It is possible for $F_n + k\tilde{z}_\alpha / (1 - a\tilde{z}_\alpha)^2$ to put negative probability weight on some observations.

Both the BC_a and ABC methods achieve one-sided coverage errors that are $O(n^{-1})$. The formulas are less intuitive than those based on simple pivoting arguments. Similarly, adjustments to empirical or parametric likelihoods to attain $O(n^{-1})$ one-sided coverage errors give rise to more complicated expressions than for the corresponding unadjusted versions.

A.7 Bibliographic notes

The stochastic order notation, O_p and o_p , is due to Mann & Wald (1943).

Cox & Hinkley (1974) and Bickel & Doksum (2000) provide good coverage of the mathematics of parametric likelihood models. In higher order asymptotics, the profile likelihood ratio function does not have all the properties of an ordinary log likelihood function. There has been considerable work on modifying profile likelihoods to be more like likelihoods, starting with Barndorff-Nielsen (1983) and surveyed in Mukerjee & Reid (1999).

Standard references on the bootstrap are Efron & Tibshirani (1993), Hall (1992), and Davison & Hinkley (1997). The first two, as their titles indicate, are introductory and mathematical, respectively. The third has good coverage of computational issues. The discussion of the BC_a method is based on Efron & Tibshirani (1993), while that of the ABC method is based on Davison & Hinkley (1997).