Computational Frequentist Inference

* While Monte Carlo techniques are very useful in Bayesian inference, they are less applicable for standard (frequentist) inference in general.

* Consider estimation of a quantity θ by some estimator $\widehat{\theta}$.

* The primary vehicle for frequentist inference is the sampling distribution of $\widehat{\theta}$.

* Unfortunately this distribution is rarely known and even when it is it usually depends on the unknown θ.

* In many situations, the best we have is a limiting distribution as the sample size n gets very large but this may be inaccurate for realistic sample sizes.

The Bootstrap

* Recall that the sampling distribution of $\hat{\theta}$ involves considering how $\hat{\theta}$ varies over repeated sampling.

* If we know the underlying distribution F then we could use simulation to draw repeated samples from F, calculate the estimator for each of these samples and use Monte Carlo methods to approximate quantities such as the bias or sampling variability.

* Of course we generally do not know F since it depends on θ.

* The bootstrap approximates the repeated sampling procedure by taking repeated samples from an estimate of F rather than F itself.

* The method was first proposed by Bradley Efron in the 1978 Weiss Lecture published in the Annals of Statistics in 1979.

The Parametric Bootstrap

* Suppose that $X_1, \ldots, X_n$ is a sample from a population with cdf $F_{\boldsymbol{\psi}}(x) = F(x; \boldsymbol{\psi})$ and we are interested in estimation of $\theta = h(\psi) = t(F_{\boldsymbol{\psi}})$.

* Since $F_{\boldsymbol{\psi}}$ is known up to a finite set of parameters we can use maximum likelihood estimation to estimate $\boldsymbol{\psi}$.

* Then the maximum likelihood estimator of θ is

$$\widehat{\theta} = h(\widehat{\boldsymbol{\psi}}) = t(F_{\widehat{\boldsymbol{\psi}}}).$$

* We are interested in estimating the bias and standard error of $\widehat{\theta}$.

The Parametric Bootstrap

* Suppose that we draw a sample of size n from $F_{\hat{\boldsymbol{\psi}}}$, $\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*)$.

* Then we can use maximum likelihood again to find $\hat{\boldsymbol{\psi}}^*$, the maximum likelihood estimate of $\hat{\boldsymbol{\psi}}$.

* Hence we can get $\hat{\theta}^* = h(\hat{\boldsymbol{\psi}}^*)$ which is the mle of $\hat{\theta}$ based on this simulated sample.

* The bootstrap idea is that under certain conditions and provided that the sampling from $F_{\hat{\boldsymbol{\psi}}}$ mimics that from $F_{\boldsymbol{\psi}}$

$$(\hat{\theta}^* - \hat{\theta}) \mid \boldsymbol{X}^* \overset{iid}{\sim} F_{\hat{\boldsymbol{\psi}}} \xrightarrow{d} (\hat{\theta} - \theta) \mid \boldsymbol{X} \overset{iid}{\sim} F_{\boldsymbol{\psi}}.$$

Bootstrap Bias and Variance

* The implementation of the bootstrap then assumes that the sample size n is sufficiently large that

$$(\widehat{\theta} - \theta) \mid \boldsymbol{X} \overset{iid}{\sim} F_{\boldsymbol{\psi}} \overset{d}{\approx} (\widehat{\theta}^* - \widehat{\theta}) \mid \boldsymbol{X}^* \overset{iid}{\sim} F_{\widehat{\boldsymbol{\psi}}}.$$

* Assuming this distributional approximation holds, we can estimate properties of the sampling distribution of $\widehat{\theta} - \theta$ using those of $\widehat{\theta}^* - \widehat{\theta}$.

* In particular we have the bias and variance approximations

$$b(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta \mid \boldsymbol{X} \overset{iid}{\sim} F_{\boldsymbol{\psi}}) \approx \mathsf{E}(\widehat{\theta}^* - \widehat{\theta} \mid \boldsymbol{X}^* \overset{iid}{\sim} F_{\widehat{\boldsymbol{\psi}}})$$

$$\mathsf{Var}(\widehat{\theta}) = \mathsf{Var}(\widehat{\theta} - \theta \mid \boldsymbol{X} \overset{iid}{\sim} F_{\boldsymbol{\psi}}) \approx \mathsf{Var}(\widehat{\theta}^* - \widehat{\theta} \mid \boldsymbol{X}^* \overset{iid}{\sim} F_{\widehat{\boldsymbol{\psi}}})$$

Monte Carlo Parametric Bootstrap

* Since we know $F_{\hat{\psi}}$, we can sometimes find these bias and variances analytically. Usually, however, we use Monte Carlo techniques

* Generate R independent samples each of size n from $F_{\hat{\psi}}$.

* For each generated sample calculate the estimate $\hat{\theta}^*$ to produce the R repeated estimates of $\hat{\theta}$: $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_R$.

* Use the empirical bias and variance of the replicates to estimate the bias and variance of $\hat{\theta}$.

$$\hat{b}_{\text{boot}}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^*_r - \hat{\theta})$$

$$\widehat{\text{Var}}_{\text{boot}}(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}^*_r - \overline{\hat{\theta}^*})^2$$

Alternatives to the Parametric Bootstrap

* Since we are in a parametric family there are always alternatives to the parametric bootstrap.

* When the true sampling distribution of $\widehat{\theta} - \theta$ is completely known then using this distribution will always give better results since the bootstrap is an asymptotic procedure.

* In the case of the mle it can be shown that

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \text{Normal}\Big(0, v(\theta)\Big)$$

where $v(\theta)$ is the Cramér-Rao lower bound.

* Generally $v(\theta)$ is unknown since it depends on θ so we use $v(\widehat{\theta})$ instead.

Comparison

* A major advantage of the asymptotic likelihood inference is that there is no need for Monte Carlo simulation.

* This is clearly a computational saving but it also improves the accuracy since the bootstrap accuracy is a function of the simulation size R.

* However it is impossible to estimate the bias or any asymmetry in the distribution of $\widehat{\theta} - \theta$ for finite n.

* The parametric bootstrap alleviates these problems somewhat since it does not impose unbiasedness or symmetry.

* Both results are asymptotic in the sample size n so neither is truly correct for finite n. The bootstrap, however, may be closer to accurate for finite n and very large R.

An Example

* Let us compare the parametric bootstrap and the asymptotic likelihood inference for a specific example.

* Suppose that $X_1, \ldots, X_n$ are exponential with rate parameter λ.

* The mle of λ is $\hat{\lambda} = 1/\overline{X}$.

* The asymptotic variance of $\sqrt{n}(\hat{\lambda} - \lambda)$ is

$$v(\lambda) \;=\; \lambda^2 \;\Rightarrow\; \text{se}(\hat{\lambda}) \;=\; \hat{\lambda}/\sqrt{n}$$

* For the parametric bootstrap we generate samples from the exponential distribution with rate equal to $\hat{\lambda}$, and use R replicates of $\hat{\lambda}^* \;=\; 1/\overline{X}^*$ to find the bias and standard error.

Example Continued

* In this case the exact distribution of $\hat{\lambda}$ is known since

$$\overline{X} \sim \text{Gamma}(\alpha = n, \beta = 1/n\lambda)$$

* From this we find the bias and variance

$$b(\hat{\lambda}) = \frac{\lambda}{n-1} \qquad \text{Var}(\hat{\lambda}) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}$$

* Since these depend on the unknown λ we replace λ by $\hat{\lambda}$ to get the estimated bias and standard error.

* It is interesting to note that these results are the same as the theoretical parametric bootstrap results (assuming an infinite Monte Carlo size).

Simulation Study

* Based on the previous example we can conduct a simulation study.

* We will use $\lambda = 0.1$ and $n = 10$.

* Taking $N = 1000$ samples we can compare the following quantities
  1. The true bias and standard deviation.
  2. The Monte Carlo repeated sampling bias and standard deviation.
  3. The average bias and standard error based on the exact distribution.
  4. The average asymptotic maximum likelihood standard error (bias is not available).
  5. The average parametric bootstrap bias and standard error.

Results of Simulation Study

* The true bias and standard deviation are 0.0111 and 0.0393.
* Monte Carlo estimates from the simulation are 0.0111 and 0.0400.
* The bias and standard error from the exact distribution have means 0.0123 and 0.0436.
* The asymptotic standard error has mean 0.0351 so is anti-conservative.
* The bootstrap bias and standard error have means 0.0124 and 0.0436.
* The bootstrap results and those from the exact distribution are quite similar for all samples.
* The asymptotic method cannot estimate bias and badly underestimates the standard error.

The Nonparametric Bootstrap

* The bootstrap is probably most powerful when we do not wish to make any distributional assumptions.

* In such cases there are rarely any asymptotic results such as those for the likelihood and the sampling distribution of an estimator is almost always unknown.

* The bootstrap can still be applied, however.

* The idea is still the same: estimate the population distribution F and consider estimates based on sampling from this estimated distribution.

Nonparametric Estimation of F

* To apply the bootstrap we still need to estimate the unknown F.

* The most common estimator is the empirical distribution function.

Definition 9

Suppose that $X_1, \ldots, X_n$ is a sample from a population with cdf $F(x)$, then the empirical distribution function is defined by

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leqslant x)$$

Properties of the EDF

* The empirical distribution function is a valid distribution function.

* It corresponds to a discrete distribution putting equal mass on each of the observed values $x_1, \ldots, x_n$.

Theorem 12

Suppose that F is a distribution function and $\widehat{F}$ is the empirical distribution function based on $X_1, \ldots, X_n$ then for any point $x \in \mathbb{R}$

$$\mathsf{E}\Big(\widehat{F}(x)\Big) = F(x)$$

$$\mathsf{Var}\Big(\widehat{F}(x)\Big) = \frac{F(x)\Big(1 - F(x)\Big)}{n}$$

Statistical Functionals

* In nonparametric inference, there are no parameters to estimate.

* Instead we concentrate on estimation of quantities called statistical functionals.

* A statistical functional is a (usually scalar) quantity which can be defined by $\theta = t(F)$.

* Some statistical functional that might be of interest are
  - $t(F) = F(x_0) = \mathrm{P}_F(X \leqslant x_0)$, a tail probability
  - $t(F) = F^{-1}(q)$, a quantile.
  - $t(F) = F^{-1}(0.5)$, the median.
  - $t(F) = \int x dF(x) = \mathrm{E}_F(X)$, the mean.

The Plug-in Principle

* Suppose that $\theta = t(F)$ is a statistical functional.

* The plug-in principle says that to find an estimator of θ we should replace F with an estimator of F.

* In the parametric setting we have $F(x) = F_{\boldsymbol{\psi}}(x) = F(x; \boldsymbol{\psi})$ then we would estimate $\boldsymbol{\psi}$ by $\widehat{\boldsymbol{\psi}}$ and use the estimator

$$\widehat{\theta} = t\left(F_{\widehat{\boldsymbol{\psi}}}\right).$$

* For nonparametric inference we use the empirical distribution function and so get the estimator

$$\widehat{\theta} = t\left(\widehat{F}\right).$$

The Nonparametric Bootstrap

* As our estimator of the population distribution function we will take the empirical distribution function $\widehat{F}$.

* We will generally be interested in some scalar functional $\theta = t(F)$ which we will estimate using the plug-in estimator $\widehat{\theta} = t(\widehat{F})$.

* As $\widehat{\theta}$ is an estimator, we can always write

$$\widehat{\theta} \;=\; t(\widehat{F}) \;=\; \widehat{\theta}(X_1, \ldots, X_n).$$

* We then consider samples $X_1^*, \ldots, X_n^*$ drawn from $\widehat{F}$ and the bootstrap replicates

$$\widehat{\theta}^* \;=\; \widehat{\theta}(X_1^*, \ldots, X_n^*)$$

* The procedure is then identical to the parametric bootstrap.

The Nonparametric Bootstrap

* As with the parametric bootstrap we can sometimes find the nonparametric bootstrap bias and standard error exactly.

* We do this by recalling that $\widehat{F}$ corresponds to a discrete distribution with equal probability at each of the n datapoints.

* For example we can show

$$\mathrm{E}^*\left(X^*\right) = \overline{x} \qquad \mathrm{Var}^*\left(X^*\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

* From these results we can get the bootstrap bias and standard error of the sample mean.

* Usually, however, the quantity of interest does not allow us to use analytical calculations and so we use Monte Carlo methods instead.

Sampling From $\hat{F}$

* Recall that $\hat{F}$ corresponds to a probability mass function with probability $1/n$ on each of the observed datapoints $x_1, \ldots, x_n$.

* Sampling from this discrete distribution is done by generating $U \sim \text{uniform}(0, 1)$ and then setting

$$X^* = x_{(i)} \quad \text{if} \quad \frac{i-1}{n} < U \leqslant \frac{i}{n}; \quad i = 1, \ldots, n.$$

  where $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the ordered values of $x_1, \ldots, x_n$.

* This is exactly the same as drawing one of the $x_1, \ldots, x_n$ at random.

* Getting a sample from $\hat{F}$ therefore is equivalent to sampling from $x_1, \ldots, x_n$ with replacement.

Monte Carlo Nonparametric Bootstrap

* To implement the nonparametric bootstrap using Monte Carlo we proceed as follows:
  1. Let $\widehat{\theta} = t(\widehat{F}) = h(X_1, \ldots, X_n)$ be the estimate of the quantity of interest.
  2. Generate $X_1^*, \ldots, X_n^* \overset{iid}{\sim} \widehat{F}$ by sampling with replacement from $X_1, \ldots, X_n$.
  3. Calculate $\widehat{\theta}^* = h(X_1^*, \ldots, X_n^*)$.
  4. Repeat Steps 2 and 3 R times to get $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_R^*$.
  5. The bootstrap bias and variance estimates are then exactly as for the parametric bootstrap.

$$b_{boot}(\widehat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\theta}_r^* - \widehat{\theta}) = \overline{\widehat{\theta}^*} - \widehat{\theta}$$

$$v_{boot}(\widehat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}(\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2$$

Drawbacks of the nonparametric bootstrap

* The nonparametric bootstrap does not always work.

* One of the major problems is due to the discreteness of bootstrap distribution.

* For the sample mean, there are only $\binom{2n-1}{n-1}$ possible values that $\overline{X}^*$ can take on. Fortunately as $n \to \infty$ this number increases quite rapidly and so the bootstrap distribution becomes like a continuous distribution.

* For some other estimators, such as the sample median, the problem is even more severe.

* An estimator for which the nonparametric bootstrap does not work at all is the sample maximum (or minimum).

Other Nonparametric Estimates of F

* Although it is most commonly used and has some very nice asymptotic properties, there is no reason why we have to use the empirical distribution function to estimate F.

* Since discreteness causes problems, maybe using a continuous estimate of F would be better.

* The most common continuous estimate of a density function is the kernel density estimator

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x - x_i}{h}\right)$$

where w is a continuous pdf symmetric about 0 with variance 1.

Sampling From a Kernel Estimator

* The kernel density estimator can be written as a finite mixture density

$$\hat{f}(x) = \sum_{i=1}^{n} p_i g_i(x)$$

where $p_i = 1/n$ and g_i is a member of the location-scale family with mean x_i, variance h^2 and standard density w.

* Sampling and observation X from g_i is equivalent to sampling $Y \sim w$ and setting $X = x_i + hY$.

* Hence we can sample X^* from $\hat{f}$ by
  1. Sample Z^* from the pmf with probability $1/n$ on each of $x_1, \ldots, x_n$.
  2. Sample $Y^* \sim w$.
  3. Set $X^* = Z^* + hY^*$.

Sampling From a Kernel Estimator

* If X^* is sampled from the kernel estimator then

$$\mathrm{E}^*(X^*) = \bar{x} \qquad \mathrm{Var}^*(X^*) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + h^2$$

* Hence the variance of the bootstrap observations is greater than the variance of the original observations.

* This is not a big problem provided $h \to 0$ as $n \to \infty$.

* However, for any finite n, the variance of observations generated from the smoothed distribution will be greater than those from $\widehat{F}$ and so standard error estimates will generally also be larger.

The Shrunken Smoothed Bootstrap

* One way to correct this is the shrunken smoothed bootstrap which uses the continuous density estimator

$$\widehat{f}(x) \;=\; \frac{1}{nbh}\sum_{i=1}^{n} w\left(\frac{x-a-bx_i}{hb}\right)$$

where the quantities a and b are

$$b \;=\; \left[1+\frac{nh^2}{\sum(x_i-\overline{x})^2}\right]^{-1/2} \qquad a \;=\; (1-b)\overline{x}$$

* Sampling from this distribution is also very easy and is

$$X^* \;=\; a + bZ^* + hbY^*$$

where Z^* and Y^* are as before.

Extension to Multiple Sample Problems

* In many situations, we actually have 2 or more samples from different populations

$$X_{k,1}, \ldots, X_{k,n_k} \overset{iid}{\sim} F_k \qquad k = 1, \ldots, K$$

* The jackknife and bootstrap extend easily to such situations.

* For the bootstrap we estimate each of the F_k separately and sample

$$X^*_{k,1}, \ldots, X^*_{k,n_k} \overset{iid}{\sim} \hat{F}_k \qquad k = 1, \ldots, K$$

* Estimates of bias and variance are as before.

Bootstrap Confidence Intervals

* Suppose that $\widehat{\theta}$ is an estimator of θ and Let $a_{\alpha/2}$ be the point such that

$$\mathrm{P}(\widehat{\theta} - \theta < a_{\alpha/2}) = \alpha/2$$

* A $100(1-\alpha)\%$ equi-tailed confidence interval for θ can then be found from

$$\mathrm{P}(a_{\alpha/2} < \widehat{\theta} - \theta < a_{1-\alpha/2}) = 1 - \alpha$$

to be the interval

$$\left[\widehat{\theta} - a_{1-\alpha/2},\ \widehat{\theta} - a_{\alpha/2}\right]$$

* To derive confidence intervals we need to find the quantiles of the distribution of $\widehat{\theta} - \theta$.

Bootstrap Normal Intervals

* Assume that $\widehat{\theta} - \theta \dot{\sim}$ Normal(b, v)
* Approximate b and v using the bootstrap estimates $\widehat{b}_{\text{boot}}$ and $\widehat{v}_{\text{boot}}$
* $100(1-\alpha)\%$ confidence interval is then

$$\left[\widehat{\theta} - \widehat{b}_{\text{boot}} - z_{\alpha/2}\sqrt{\widehat{v}_{\text{boot}}}, \quad \widehat{\theta} - \widehat{b}_{\text{boot}} + z_{\alpha/2}\sqrt{\widehat{v}_{\text{boot}}}\right]$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal.

Basic Bootstrap Interval

* In a Monte Carlo bootstrap we can use a normal quantile-quantile plot of the bootstrap replicates to assess approximate normality.

* If this plot is badly non-linear then we should not rely on the intervals which assume normality.

* We can apply the bootstrap technique to say that

$$\hat{\theta}^* - \hat{\theta} \stackrel{d}{\approx} \hat{\theta} - \theta$$

and so we can approximate $a_{\alpha/2}$ by $a_{\text{boot},\alpha/2}$ such that

$$\mathrm{P}^*(\hat{\theta}^* - \hat{\theta} < a_{\text{boot},\alpha/2}) = \alpha/2$$

* Typically we cannot do this analytically.

Monte Carlo Basic Bootstrap Interval

* Use Monte Carlo methods to get the bootstrap replicates

$$\widehat{\theta}_1^* - \widehat{\theta}, \ldots, \widehat{\theta}_R^* - \widehat{\theta}$$

* Order the bootstrap replicates and use them to estimate the correct bootstrap quantile

$$\widehat{a}_{\text{boot},\alpha/2} = \widehat{\theta}^*_{(R\alpha/2)} - \widehat{\theta}$$

where $\widehat{\theta}^*_{(R\alpha/2)}$ is the $R\alpha/2$ ordered bootstrap replicate.

* In that case the interval becomes

$$\left[2\widehat{\theta} - \widehat{\theta}^*_{(R(1-\alpha/2))},\ 2\widehat{\theta} - \widehat{\theta}^*_{(R\alpha/2)}\right]$$

Monte Carlo Basic Bootstrap Interval

* Accurate Monte Carlo estimation of tail quantiles is more challenging than Monte Carlo estimation of means.

* We generally need to take R to be quite large to get stable confidence interval limits

* We should generally take R such that $R\alpha/2 > 50$.

* This means we need $R \geqslant 2000$ for 95% intervals and $R \geqslant 5000$ for 99% intervals.

* Of course increasing the simulation size will reduce simulation variability and so improve the stability of the intervals.

* The actual coverage of the intervals, however, is determined by the sample size n.

Parameter Transformations

* Intervals are not transformation invariant.

* Performance can be improved using by working on an appropriate scale.

* Suppose that h is a one-to-one function and let us define $\phi = h(\theta)$.

* We will sometimes find it useful to construct a confidence interval for ϕ and then apply the inverse transformation to the endpoints to get a confidence interval for $\theta = h^{-1}(\phi)$.

* We can base our intervals on bootstrap replicates

$$\widehat{\phi}_r^* - \widehat{\phi} \;=\; h(\widehat{\theta}_r^*) - h(\widehat{\theta})$$

Percentile Intervals

* If we can assume that the distribution of $\widehat{\theta} - \theta$ is symmetric about 0 then we have that $-a_{\alpha/2} = a_{1-\alpha/2}$

* This results in the interval

$$\left[\widehat{\theta} + a_{\alpha/2},\ \widehat{\theta} + a_{1-\alpha/2}\right]$$

* Now suppose we use a Monte Carlo bootstrap so that

$$\widehat{a}_{\text{boot},\alpha/2} = \widehat{\theta}^*_{(R\alpha/2)} - \widehat{\theta} \qquad \widehat{a}_{\text{boot},1-\alpha/2} = \widehat{\theta}^*_{(R(1-\alpha/2))} - \widehat{\theta}$$

* Then the bootstrap interval becomes

$$\left[\widehat{\theta}^*_{(R\alpha/2)},\ \widehat{\theta}^*_{(R(1-\alpha/2))}\right]$$

Percentile Intervals on Transformed Scale

* Suppose that h is a one-to-one transformation and $\phi = h(\theta)$.

* If we assume that the distribution of $\widehat{\phi} - \phi$ is symmetric about 0 then we get the percentile interval for ϕ

$$\left[\widehat{\phi}^*_{(R\alpha/2)},\ \widehat{\phi}^*_{(R(1-\alpha/2))}\right]$$

* If we know apply h^{-1} to get an interval for θ we see it is

$$\left[\widehat{\theta}^*_{(R\alpha/2)},\ \widehat{\theta}^*_{(R(1-\alpha/2))}\right]$$

* The percentile interval is transformation invariant.

* The interval is valid as long as some h exists such that the distribution of $h(\widehat{\theta}) - h(\theta)$ is symmetric about 0.