The Bootstrap

The bootstrap is a computer-intensive method for estimating measures of uncertainty in problems for which no analytical solution is available.

There are technically two classes of bootstrap methods: parametric and nonparametric.

The nonparametric bootstrap uses two main ideas:

- The empirical CDF
- Monte Carlo integration

Monte Carlo integration is based on the following approximation:

$$\begin{aligned} E[h(Y)] &= \int h(y) dF_Y(y) \\ &\approx \frac{1}{B} \sum_{j=1}^{B} h(Y_j) \end{aligned}$$

where $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$. Note that if $E[|h(Y)|] < \infty$,

$$\frac{1}{B} \sum_{j=1}^{B} h(Y_j) \overset{as}{\to} E[h(Y)]$$

as $B \to \infty$. Typically we have control over B, so we can make the approximation arbitrarily good.

A simple example: Use Monte Carlo integration to approximate

$$\int_{-\infty}^{\infty} \sin^2(x) e^{-x^2} dx$$

Solution: We can write this as $\sqrt{2\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx$, where $f(x)$ is the PDF of a $N(0,1)$ r.v. Therefore, we can

1. Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} N(0,1)$.

```
> B <- 10000; y <- rnorm(B)
```

2. Approximate $\sqrt{2\pi} \int_{-\infty}^{\infty} \sin^2(x) f(x) dx \approx \frac{\sqrt{2\pi}}{B} \sum_{j=1}^{B} \sin^2(Y_j)$.

```
> sqrt(2*pi) * mean(sin(y)^2)
[1] 1.074098
```

A more complicated example: Use Monte Carlo integration to approximate $V_\lambda[median(X_1, \ldots, X_n)]$ when $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$.

This is more complicated in two ways:

1. Unlike an analytical calculation, on the computer we need particular values of n and λ. To see how $V_\lambda[median(X_1, \ldots, X_n)]$ changes with n and λ, we need to use Monte Carlo integration many times for different combinations.

2. For each combination, we need to sample B times from the *sampling distribution* of $median(X_1, \ldots, X_n)$. That is, for each $j = 1, \ldots, B$, we need to sample $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\lambda)$ and calculate the median. Don't confuse n and B: n is the sample size, while B is the number of MC samples.

One combination: Let $n = 10$ and $\lambda = 5$. Then

- Draw $Y_1, \ldots, Y_B \overset{iid}{\sim} F_Y$, where F_Y is the CDF of $median(X_1, \ldots, X_n)$.

```
> n <- 10; lambda <- 5; B <- 10000
> samples <- matrix(rexp(n*B, rate = 1/lambda),
+    nrow = B, ncol = n)
> y <- apply(samples, MARGIN = 1, FUN = median)
```

- Approximate $V_\lambda[median(X_1, \ldots, X_n)] \approx \frac{1}{B}\sum_{j=1}^{B}(Y_j - \bar{Y})^2$.

```
> var(y)
[1] 2.402400
```

Back to the bootstrap...

Suppose we have data $X_1, \ldots, X_n$ and we compute statistic $T_n = g(X_1, \ldots, X_n)$.

It's not always possible to calculate $V_F[T_n]$ analytically, which is where the bootstrap comes in.

If we knew F, we could use MC integration to approximate $V_F[T_n]$. However, we don't in practice, so we make an initial approximation of F with the empirical CDF $\hat{F}_n$.

$$V_F[T_n] \underbrace{\approx}_{\substack{\text{ECDF;}\\ \text{depends on } n}} V_{\hat{F}_n}(T_n) \underbrace{\approx}_{\substack{\text{MC integration;}\\ \text{depends on } B}} \widehat{V}_{\hat{F}_n}(T_n)$$

Sampling from $\hat{F}_n$ is easy: just draw one observation at random from $X_1, \ldots, X_n$. Repeated sampling is "with replacement."

The algorithm:

1. Repeat the following B times to obtain $T^*_{n,1}, \ldots, T^*_{n,B}$, an iid sample from the sampling distribution for T_n implied by $\hat{F}_n$.
   (a) Draw $X^*_1, \ldots, X^*_n \sim \hat{F}_n$.
   (b) Compute $T^*_n = g(X^*_1, \ldots, X^*_n)$.

2. Use this sample to approximate $V_{\hat{F}_n}(T_n)$ by MC integration. That is, let

$$v_{boot} = \widehat{V}_{\hat{F}_n}(T_n) = \frac{1}{B}\sum_{j=1}^{B}\left(T^*_{n,j} - \frac{1}{B}\sum_{k=1}^{B} T^*_{n,k}\right)^2$$

Confidence intervals can also be constructed from the bootstrap samples.

Method 1: Normal-based interval

$$C_n = T_n \pm z_{\alpha/2}\widehat{se}_{boot}$$

where $\widehat{se}_{boot} = \sqrt{v_{boot}}$; this only works well if the distribution of T_n is close to Normal. Note that asymptotic normality of T_n is a property involving n, not B.

Method 2: Quantile intervals

$$C_n = \left(T^*_{\alpha/2}, T^*_{1-\alpha/2}\right)$$

where T^*_β is the β quantile of the bootstrap sample $T^*_{n,1}, \ldots, T^*_{n,B}$.

Method 3: Pivotal intervals

In parametric statistics, a pivot is a function $R(X_1, \ldots, X_n, \theta)$ whose distribution doesn't depend on θ. This is useful because we can construct a confidence interval for $R_n = R(X_1, \ldots, X_n, \theta)$ without knowing θ and then manipulate it to construct a confidence interval for θ.

In nonparametric statistics, we typically can't find a quantity that is exactly pivotal, i.e., whose distribution doesn't depend on the unknown F.

If $\theta = T(F)$ is a location parameter, then $R_n = \hat{\theta}_n - \theta$ is approximately pivotal. If we knew the CDF H of R_n, we could construct an exact $1 - \alpha$ confidence interval for θ of (a, b), where

$$\begin{aligned} a &= \hat{\theta}_n - H^{-1}(1 - \alpha/2) \\ b &= \hat{\theta}_n - H^{-1}(\alpha/2) \end{aligned}$$

Since we don't know H, we estimate it using the bootstrap samples.

$$\hat{H}(r) = \frac{1}{B}\sum_{j=1}^{B} I(R^*_{n,j} \leq r)$$

where $R^*_{n,j} = \hat{\theta}^*_{n,j} - \hat{\theta}_n$. In other words, we form the empirical CDF of H using the bootstrap samples of the pivot. Therefore, the plug-in estimates of $H^{-1}(1-\alpha/2)$ and $H^{-1}(\alpha/2)$ are just the $1-\alpha/2$ and $\alpha/2$ sample quantiles of these samples.

This gives a $1-\alpha$ bootstrap pivotal interval of

$$C_n = \left(2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2}\right)$$

Parametric Inference

A parametric model has the form

$$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$$

where $\Theta \subseteq \mathbb{R}^k$ is the parameter space.

We typically choose a class $\mathcal{F}$ based on knowledge about the particular problem. We might say we're making certain assumptions about the data generating mechanism. It's good practice when using a parametric model to look for violations of these assumptions.

We'll begin with two methods for constructing estimators of θ: the method of moments and maximum likelihood estimation.

Suppose $\theta = (\theta_1, \ldots, \theta_k)$. For $j = 1, \ldots, k$, define the j^{th} moment

$$\alpha_j \equiv \alpha_j(\theta) = E_\theta[X^j] = \int x^j dF_\theta(x)$$

and the j^{th} sample moment $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$.

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ s.t.

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ \vdots \quad &\vdots \quad \vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

The maximum likelihood estimator (MLE) is obtained by maximizing the likelihood function

$$\begin{aligned} \mathcal{L}_n(\theta) &= f(X_1, \ldots, X_n; \theta) \\ &= \prod_{i=1}^{n} f(X_i; \theta) \quad \text{if the data are independent} \end{aligned}$$

That is, the likelihood is just the joint density of the data, but viewed as a function of θ.

It's often easier to work with the log-likelihood function

$$\ell_n(\theta) = \log \mathcal{L}_n(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

If the log-likelihood is differentiable with respect to θ, possible candidates for the MLE are those in the interior of Θ that solve

$$\frac{\partial}{\partial \theta_j} \ell_n(\theta) = 0, \quad j = 1, \ldots, k$$

We still need to check that we've found the global maximum. Also note that if the maximum occurs on the boundary of Θ, the first derivative may not be zero.

It's not always possible to maximize the likelihood analytically, and in these cases we turn to numerical maximization methods.

Examples:

- Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. Find the MLE for θ.
- Now solve the same problem, but with the restriction $\Theta = [0, \infty)$.
- Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(0, \theta)$. Find the MLE for θ.