STAT 200B 2019 Week03

Soyeon Ahn

1 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 $\hat{F}_n$.

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

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.

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

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

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

2 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 the maximum likelihood estimation.

2.1 The method of moments

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** $\hat{\theta}_n = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is defined to be the value of θ s.t.

$$\begin{aligned}
\alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^n X_i \\
\alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \\
\vdots \quad &\vdots \quad \vdots \\
\alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k == \frac{1}{n}\sum_{i=1}^n X_i^k
\end{aligned}$$

2.2 The maximum likelihood estimation

Let $X^n = (X_1, \dots, X_n)$ be iid with pdf $f(x;\theta)$, where $\theta \in \Theta$.

The **likelihood function** is

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

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

The **log-likelihood function** is defined by

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

The likelihood function is analytically the same as the joint density (probability distribution) of the data, it is a function of the parameter θ for a given set of data points. Thus, it is not a probability density function. The likelihood function can evaluate the plausibility of parameter values.

The **maximum likelihood estimator** (MLE) θ_n is obtained by maximizing the likelihood function Since the maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, since log is a strictly increasing function.

It's often easier to work with the log-likelihood function. If the log-likelihood is differentiable with respect to θ, possible candidates for the MLE are those in the interior of Θ that solve

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

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.

Example Let $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(p). The probability function is $f(x;p) = p^x(1-p)^{1-x}$.

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(x_i;p) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

Therefore,

$$\ell_n(p) = n\bar{x}\log(p) + n(1-\bar{x})\log(1-p)$$

Take the derivative of $\ell_n(p)$ set it equal to 0 to find that the MLE is

$$\hat{p}_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The total number of successes observed in the n trials over the total number of trials, i.e., the relative frequency of the observation 'success', is our estimator.

global maximum Note that the log likelihood function is strictly concave (i.e. $\ell_n(\theta)'' < 0$), $\hat{p}$ is a global maximum. In general, we take the second derivatives of $\ell_n(\theta)$ with respect to the (vector-valued) parameter. The matrix is called the Hessian. If the Hessian of the log likelihood at $\hat{\theta}$ is negative semi-definite, then $\ell_n(\theta)$ is concave, and it will be a global maximum.

In the above example with n Bernoulli samples,

$$\ell_n(p)' = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{(1-p)}$$

$$\ell_n(p)'' = -\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2}$$

with $p \in (0,1)$, the second derivative is always < 0.

Example Let Y follow Binomial(n, p), where n is known and p is the parameter, to be estimated. The likelihood function is

$$\mathcal{L}(p) = \frac{n!}{y!(n-y)!} \cdot p^y (1-p)^{n-y}$$

(I don't use the notation n since there is only single binomial observation.) Note that $p^y(1-p)^{n-y}$ is identical to the above n Bernoulli trials. Since the MLE is a function of the parameter p,

$$\hat{p}_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

, the sample proportion of successes

We can represent $Y = \sum X_i$, where $X_i \stackrel{iid}{\sim}$ Bernoulli(p). The MLE based on n independent Bernoulli random variables and the MLE based on a single binomial random variable is the same.

In general, whenever we have repeated, independent Bernoulli trials with the same probability of success p for each trial, the MLE will always be the sample proportion of successes.

sufficient statistic The sample proportion of successes (a statistic $T(X_1, \ldots, X_n)$) contains all information about θ (the parameter).

Example Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$. Find the MLE for $\theta = \mu$.

$$\begin{aligned}\mathcal{L}_n(\mu, 1) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(x_i-\mu)^2}{2}\Big) \\ &= \frac{1}{(\sqrt{2\pi})^n} \exp\Big(-\frac{\sum_i^n (x_i-\mu)^2}{2}\Big)\end{aligned}$$

$$\log \mathcal{L}_n(\mu, 1) = -\frac{n}{2}\log(2\pi) - \frac{\sum_i^n (x_i-\mu)^2}{2}$$

$$\frac{\partial}{\partial \mu} \log \mathcal{L}_n(\mu, 1) = \sum_i^n (x_i - \mu) = n(\bar{x} - \mu)$$

$$\hat{\mu} = \bar{x}$$

Example What if $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$? $\theta = (\mu, \sigma^2)$ is a vector valued parameter.

$$\log \mathcal{L}_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_i^n (x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \mu} \log \mathcal{L}_n(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_i^n (x_i - \mu) = \frac{1}{\sigma^2} n(\bar{x} - \mu)$$

$$\begin{aligned}\frac{\partial}{\partial \sigma^2} \log \mathcal{L}_n(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i^n (x_i-\mu)^2 \\ &= \frac{n}{2(\sigma^2)^2}\Big(\sigma^2 - \frac{1}{n}\sum_i^n (x_i-\mu)^2\Big)\end{aligned}$$

Hence,

$$\hat{\mu} = \bar{x}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_i^n (x_i - \hat{\mu})^2 = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2$$

Note that we differentiate with respect to σ.

$$\frac{\partial}{\partial \sigma} \log \mathcal{L}_n(\mu, \sigma^2) = -\frac{n}{\sigma} + \frac{1}{(\sigma^3)}\sum_i^n (x_i-\mu)^2$$

Note that the MLE for σ^2 is a biased estimator.

Example In a simple linear regression,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

, where the ϵ_i are independent $N(0, \sigma^2)$. The outcome variables are continuous and the mean vector can be specified by a given predictor x_i.

In a simple linear regression, it is sufficient to specify that y_i is $N(\mu_i, \sigma^2)$ and $\mu_i = \beta_0 + \beta_1 x_i$

Traditionally the model is written in matrix form as

$$y = X\beta + \epsilon$$

, where X is an $n\times$ (number of coefficients) design matrix.

In Week 04

1. Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.
2. Consistency: $\hat{\theta}_n \xrightarrow{P} \theta_*$, where θ_* is the true value of the parameter.
3. Asymptotic normality: $(\hat{\theta}_n - \theta_*)/se(\hat{\theta}_n) \xrightarrow{D} N(0, 1)$.
4. Asymptotic efficiency: The MLE has the smallest asymptotic variance among asymptotically normal estimators.

2.3 Bayesian inference

Thomas Bayes (1701 – 1761) an English minister and mathematician. None of his work was published during his lifetime.

The conditional probability of an event is a probability obtained with the additional information that some other event has already occurred. The conditional probability of event B occurring, given that event A has already occurred.

If A is a binary event,

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Frequentist statistics

- Interprets probability in terms of long-run frequencies of events.
- Treats parameters as unknown, fixed constants.
- Focuses on point estimation, confidence intervals, and hypothesis tests.

Bayesian statistics

- Interprets probability as representing degree of belief.
- Makes probability statements about parameters, reflecting beliefs.
- Bases all inference on the posterior distribution, which we can summarize in various ways.

The conditional density of $p(\theta|y)$ is

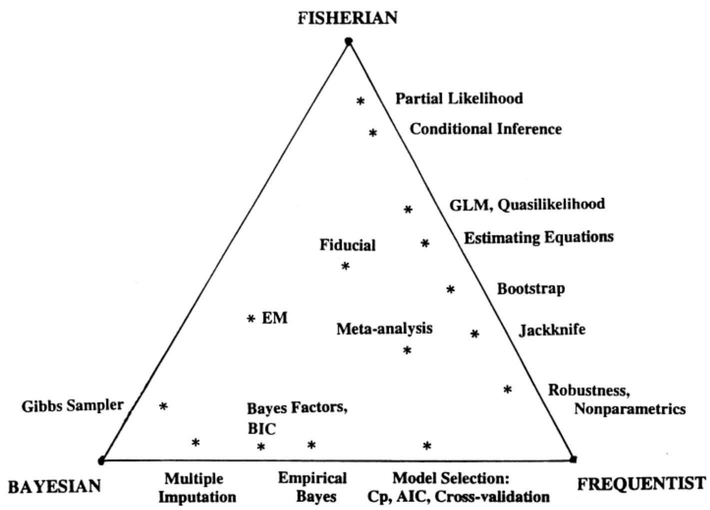
$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

Note that the denominator, $p(y)$ is a function of the data.

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

, where $p(\theta)$ is the prior density of θ, and $p(y|\theta)$ is the likelihood, and $p(\theta|y)$ is the posterior density of θ. The posterior is proportional to the prior times the likelihood.

BAYES	FISHER	FREQUENTIST
1. Individual (personal decisions)	***	Universal (world of science)
2. Coherent (correct)	*************	Optimal (accurate)
3. Synthetic (combination)	****	Analytic (separation)
4. Optimistic (aggressive)	*****	Pessimistic (defensive)


FISHERIAN
Partial Likelihood
Conditional Inference
GLM, Quasilikelihood
Estimating Equations
Fiducial
Bootstrap
EM
Meta-analysis
Jackknife
Robustness,
Nonparametrics
Gibbs Sampler
Bayes Factors,
BIC
BAYESIAN
Multiple
Imputation
Empirical
Bayes
Model Selection:
Cp, AIC, Cross-validation
FREQUENTIST