Module 10: Bootstrap

Yaqi Shi

July 22, 2024

Outline

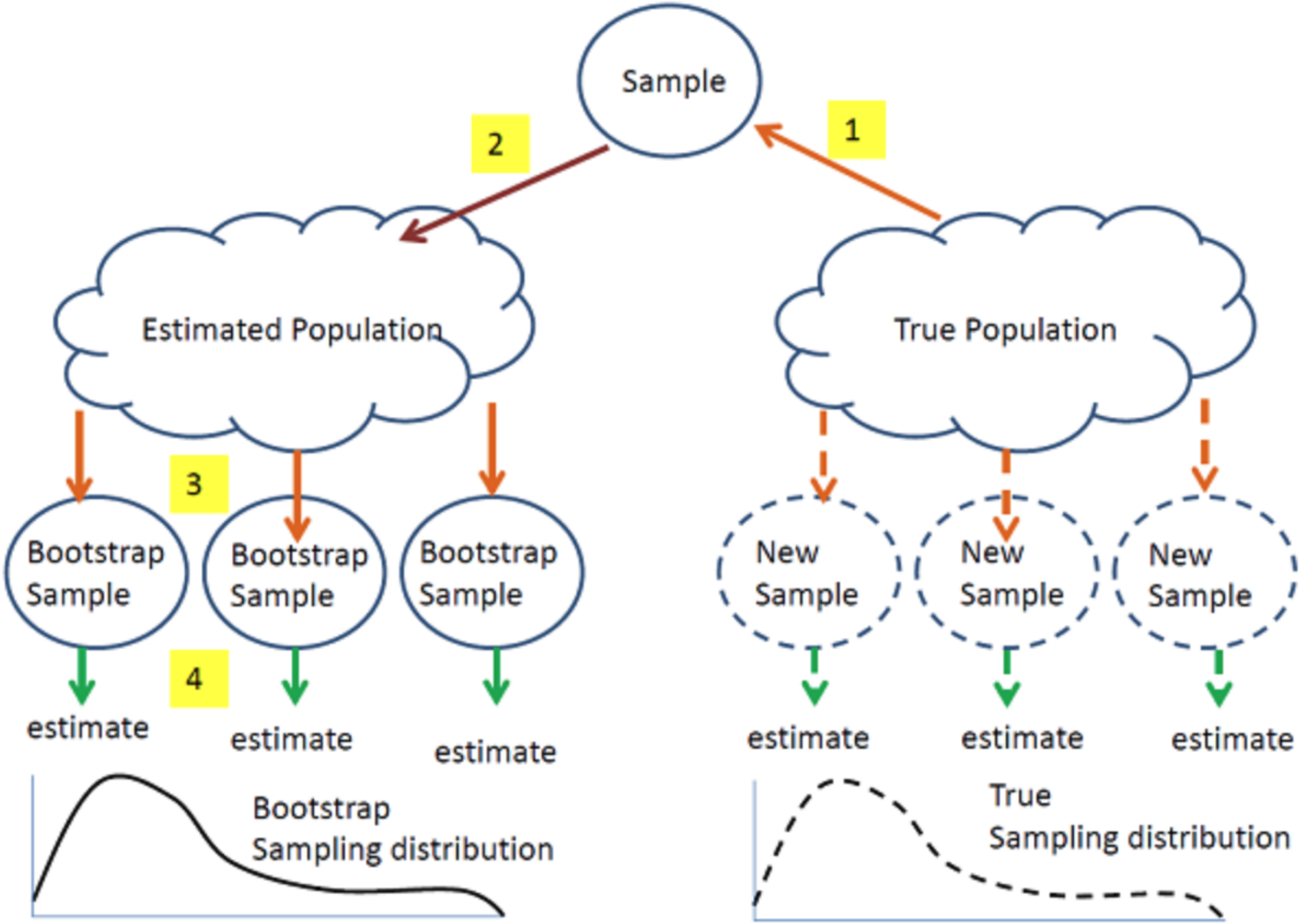
In this module, we will continue our discussion on Bootstrap.

What is bootstrap?

A widely applicable, computer intensive resampling method used to compute standard errors, confidence intervals, and significance tests.

Why bootstrap?

- The exact sampling distribution of an estimator can be **difficult** to obtain
- Asymptotic expansions are sometimes easier but expressions for standard errors based on large sample theory may not perform well in finite samples



https://online.stat.psu.edu/stat555/node/119/

Example: estimate the variance of an estimator

Now let $\text{Var}_F(T_n)$ denote the variance of T_n. The subscript F means the variance is a function of F where F is the CDF of X. If we know F, we can directly compute the variance. For example, if then $x_1, \cdots, x_w$ iid

$$T_n = \bar{X}_n, = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Var}_F(T_n) = n^{-1}\left(\int x^2 dF(x) - \left(\int x dF(x)\right)^2\right)$$

which is a function of F.

Example: estimate the variance of an estimator

When F is unknown, one can use an estimate of F, e.g., the empirical CDF $\hat{F}_n$, i.e.,

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)}{n}.$$

We then use a plug-in estimator for $\mathrm{Var}_{\hat{F}_n}(T_n)$:

$$\mathrm{Var}_{\hat{F}_n}(T_n) = n^{-1}\left(\int x^2 d\hat{F}_n(x) - \left(\int x d\hat{F}_n(x)\right)^2\right).$$

Example: estimate the variance of an estimator

However, the plug-in estimator above can be hard to compute, we can then approximate it with a simulation estimate, denoted by V_{boot} .

The following algorithm illustrate how one can do this through bootstrap.

Algorithm 1: Bootstrap variance estimation algorithm

Input data $(X_1, \ldots, X_n)$; Number of iteration B;
for $i \leftarrow 0$ to B do
    Draw $X^*_{1,i}, \ldots, X^*_{n,i} \sim \hat{F}_n$;
    Compute $T^*_{n,i} = g\left(X^*_{1,i}, \ldots, X^*_{n,i}\right)$;
end
$V_{\text{boot}} = \frac{1}{B}\sum_{i=1}^{B}\left(T^*_{n,i} - \frac{1}{B}\sum_{j=1}^{B} T^*_{n,j}\right)^2$.

By law of large numbers, we have $V_{\text{boot}} \xrightarrow{\text{a.s.}} \text{Var}_{\hat{F}_n}(T_n)$ as $B \to \infty$.

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

- If we knew P,
  - we could obtain $\left\{X^b \mid b = 1, \ldots B\right\}$ from P and use Monte-Carlo to estimate the sampling distribution of $\hat{\theta}$

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

- If we knew P,
  - we could obtain $\left\{X^b \mid b = 1, \ldots B\right\}$ from P and use Monte-Carlo to estimate the sampling distribution of $\hat{\theta}$
- However, we don't,
  - we do the next best thing and resample from original sample, i.e. the empirical distribution, $\hat{P}$
  - we expect the empirical distribution to estimate the underlying distribution well by the *Glivenko-Cantelli* Theorem

Why bootstrap works

Definition 2.1. Let F, G be two CDF's on a sample space X. Let $\rho(F, G)$ be a metric on the space of CDF's on X. We say G_n^* is weakly ρ-consistent if

$$\rho\left(G_n^*, G_n\right) \xrightarrow{P} 0$$

as $n \to \infty$. Similarly, G_n^* is strongly ρ-consistent if

$$\rho\left(G_n^*, G_n\right) \xrightarrow{\text{a.s.}} 0.$$

Why bootstrap works

Now the measure of closeness between two CDF's depends on the metric ρ. The following two metrics are commonly used for the CDF's:

(i) Kolmogorov metric:

$$K(F, G) = \sup_{x \in R} |F(x) - G(x)|.$$

(ii) Mallows-Wasserstein metric:

$$l_2(F, G) = \inf_{T_{F,G}} \left(E|Y - X|^2 \right)^{1/2},$$

where $T_{F,G}$ is the collection of all possible joint distribution of the pair (X, Y) whose marginal distributions are F, G. The following theorem illustrates a special case of the bootstrap theory.

Why bootstrap works

Theorem. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ and $E\left(X_i^2\right) < \infty$. Let

$$T_n = g\left(X_1, \ldots, X_n\right) = \sqrt{n}(\bar{X} - \mu).$$

Then

$$K\left(G_n^*, G_n\right) \xrightarrow{\text{a.s.}} 0, l_2\left(G_n^*, G_n\right) \xrightarrow{\text{a.s.}} 0.$$

Continued example

Besides from estimating the variance of T_n, bootstrap can also be used to approximate the CDF of T_n. Suppose

$$G_n(t) = P\left(T_n \leq t\right)$$

be the CDF of T_n. Then the bootstrap approximate to G_n is

$$G_n^*(t) = \frac{1}{B}\sum_{i=1}^{B} 1\left(T_{n,b}^* \leq t\right)$$

Continued example: bootstrap confidence intervals

Assuming $(1-\alpha)\mathrm{CI}$, the bootstrap normal CI is then

$$T_n \pm z_{\alpha/2}\hat{\mathrm{se}}_{\mathrm{boot}} \,.$$

The above interval is not accurate unless T_n is close to normal. There is also the bootstrap pivotal CI: $1-\alpha$ bootstrap.

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

where T^*_β denotes the β sample quantile of $(T^*n, 1, \ldots, T^*n, B)$. Last but not least, there is the percentile CI:

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

Forms of bootstrap

Based on how the population is estimated,

1. Nonparametric bootstrap
2. Parametric bootstrap
3. Empirical Bootstrap (Paired Bootstrap)
4. Residual Bootstrap
5. Wild Bootstrap

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

- Step 3: Estimate the standard error se $(\hat{\theta})$ by the sample standard deviation of the B replications of $\hat{\theta}^*(b)$

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

- Step 3: Estimate the standard error se $(\hat{\theta})$ by the sample standard deviation of the B replications of $\hat{\theta}^*(b)$
- Step 4: Estimate the confidence interval by finding the $100(1-\alpha)$ percentile bootstrap CI,

$$\left(\hat{\theta}_L, \hat{\theta}_U\right) = \left(\hat{\theta^*}^{\alpha/2}, \hat{\theta^*}^{1-\alpha/2}\right)$$

Parametric bootstrap

- Assumes the data comes from a known distribution with unknown parameters
- First estimate the parameters from the data and then use the estimated distribution to simulate the samples

Bootstrap for Regression

Bootstrap can also be used to conduct statistical inference for regression model. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the observed data and

$$E(Y_i \mid X_i = x) = \beta_0 + \beta_1 x,$$

i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

There are several variants to the bootstrap under regression setting.

Empirical Bootstrap (Paired Bootstrap)

We generate a new sets of i.i.d. observations $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ such that for each I,

$$P(X_I^* = X_i, Y_I^* = Y_i) = \frac{1}{n}$$

for all i. Similar to bootstrap estimation of variance, we repeat this procedure B times and fit a linear regression model for each of the generated paired data:

$$\left(X_1^{*(1)}, Y_1^{*(1)}\right), \ldots, \left(X_n^{*(1)}, Y_n^{*(1)}\right) \xrightarrow{\text{fit linear regression}} \hat{\beta}_0^{*(1)}, \hat{\beta}_1^{*(1)}$$

$$\ldots$$

$$\left(X_1^{*(B)}, Y_1^{*(B)}\right), \ldots, \left(X_n^{*(B)}, Y_n^{*(B)}\right) \xrightarrow{\text{fit linear regression}} \hat{\beta}_0^{*(B)}, \hat{\beta}_1^{*(B)}.$$

Empirical Bootstrap (Paired Bootstrap)

We then estimate the bootstrap variance

$$\hat{\text{Var}}\left(\hat{\beta}_0\right) = \frac{1}{B}\sum_{l=1}^{B}\left(\hat{\beta}_0^{*(l)} - \bar{\beta}_0^*\right)^2, \bar{\beta}_0^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}_0^{*(l)},$$

and

$$\hat{\text{Var}}\left(\hat{\beta}_1\right) = \frac{1}{B}\sum_{l=1}^{B}\left(\hat{\beta}_1^{*(l)} - \bar{\beta}_1^*\right)^2, \bar{\beta}_1^* = \frac{1}{B}\sum_{l=1}^{B}\hat{\beta}_1^{*(l)}.$$

We can also obtain the bootstrap CI as

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\sqrt{\hat{\text{Var}}\left(\hat{\beta}_0\right)}$$

and

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{\hat{\text{Var}}\left(\hat{\beta}_1\right)}$$

Empirical Bootstrap (Paired Bootstrap)

Although empirical bootstrap works well in practice, it may lead to a bad results, especially in the presence of influential observations (some X_i 's are very far away from the others).

This will also strongly affect the empirical bootstrap in the sense that if an empirical bootstrap does not select these influential points, the regression coefficients can be very different.

Residual Bootstrap

To solve the problem of empirical bootstrap illustrated previously, we may use the residual bootstrap. We first fit the original data to obtain the OLS estimate $\hat{\beta}_0, \hat{\beta}_1$. Define

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

e_i here are the fitted residuals and can be regarded as a good approximation of ϵ_i. The residual bootstrap generate i.i.d. $\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*$ such that for each $\hat{\epsilon}_l^*$,

$$P\left(\hat{\epsilon}_l^* = e_i\right) = \frac{1}{n}.$$

Then we generate a new bootstrap sample

$$\left(X_1^*, Y_1^*\right), \ldots, \left(X_n^*, Y_n^*\right)$$

via

$$X_i^* = X_i, Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i^*.$$

Wild Bootstrap

Another issue usually occur in linear regression is the so called phenomenon of heteroskedasticity, i.e., $\text{Var}(\epsilon_i \mid X_i)$ depends on the value of X_i.

Now residual bootstrap cannot deal with this issue. In particular, the residual bootstrap will be unstable because the residual bootstrap will swap all the residuals regardless of the value of the covariate. The solution to this is through the wild bootstrap.

The wild bootstrap first generate i.i.d. random variables $V_1, \ldots, V_n \sim \mathcal{N}(0, 1)$ and then generate the bootstrap sample

$$(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$$

by

$$Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + V_i e_i, X_i^* = X_i.$$

Bootstrap hypothesis testing

```r
boot_dat <- function(x, n, mu, std) {
  newx <- sample(x, n, replace = T)
  return((mean(newx) - mu)/(std/sqrt(n)))
}


pvalues_1 <- rep(NA, nsim)
pvalues_2 <- rep(NA, nsim)

for (i in 1:nsim){
  x <- rnorm(n, mean = mu, sd = std)
  stat <- abs((mean(x) - mu)/(std/sqrt(n)))
  pvalues_1[i] <- 2*(1 - pnorm(stat))


  newx <- x - mean(x) + mu
  boot_scores <- sapply(1:nboot,
                        function(index) boot_dat(newx, n, mu, std))
  boot_scores <- as.vector(boot_scores)
  # hist(boot_scores)
  # mean(boot_scores)
  pvalues_2[i] <- length(which(boot_scores > stat))/length(boot_scores) +
    length(which(boot_scores < (-stat)))/length(boot_scores)
}
```

Failures of Bootstrap

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{unif}\,(0,1)$ and $M_n = \min(X_1, \ldots, X_n)$ be the minimum of the sample. One can show that

$$nM_n \xrightarrow{\mathcal{D}} \exp(1)$$

However, let $M_n^* = \min(X_1^*, \ldots, X_n^*)$ be the minimum of a bootstrap sample, then

$$P(\text{ none of } X_1^*, \ldots, X_n^* \text{ select } M_n) = \left(1 - \frac{1}{n}\right)^n \approx e^{-1},$$

which implies that

$$P(M_n^* = M_n) \approx 1 - e^{-1}.$$

Therefore, M_n^* has a huge probability mass at M_n. This means the distribution of M_n^* is very different from the distribution of M_n. The detailed proof is left as homework.

StatQuest videos

Check out these videos made by Josh Starmer with vivid illustration for the boostrap!

- Bootstrapping Main Ideas [link]
- Using Bootstrapping to Calculate p-values [link]

Resources

This tutorial is based on

- PennState STAT555 Statistical Analysis of Genomics Data [links].
- Harvard's Biostatistics Preparatory Course Methods [links].