STAT 390 A
Statistical Methods in Engineering and Science
Week 10 Lectures – Part 1 – Winter 2023
Bootstrap

Alexander Giessing

Department of Statistics

University of Washington

March 6, 2023

Outline

Limitations of "Plug-in" Estimators for the SE

- The "plug-in" principle for estimating the SE of an estimator requires a closed-form expression of the SE.
- So far, we have only discussed cases, in which such a closed-form expression exists, e.g.
  - sample mean,
  - Gaussian error model.
- What can we do if
  - there exists no closed-form expression of the SE, or
  - computing the SE is very complicated (e.g. SE of $\hat{\theta}_{MLE}$ in the example on population genetics)?

Bootstrap Principle

Bootstrap Principle

Let the data set $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ drawn from cdf F. Let $\hat{F}$ be an estimate for F based on the data set $x_1, \ldots, x_n$, and let $X_1^*, \ldots, X_n^*$ be a sample drawn from $\hat{F}$.

Then, the sampling distribution of any statistic $T = h(X_1, \ldots, X_n)$ can be approximated by the sampling distribution of $T^* = h(X_1^*, \ldots, X_n^*)$.

- Suppose that T is an estimator for θ. Sometimes we are not interested in the sampling distribution of just T, but in the sampling distribution of the centered statistic $T - \theta$ or, more generally, the sampling distribution of a (complicated) function $(T, \theta) \mapsto R(T, \theta)$.

- By the bootstrap principle, we can approximate the sampling distribution of $R(T, \theta)$ with the one of $R(T^*, \hat{\theta})$.

Outline

Empirical Bootstrap Procedure

- Recall: The empirical cdf and pmf of a data set $x_1, \dots x_n$ are given by

$$F_n(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq a\} \qquad \text{and} \qquad p_n(a) = \begin{cases} n^{-1} & \text{if } a \in \{x_1, \dots, x_n\}, \\ 0 & 0/w. \end{cases}$$

- Let T be an estimator for θ and $\hat{\theta} = T(x_1, \dots, x_n)$ the estimate.

EMPIRICAL BOOTSTRAP TO ESTIMATE $R(T, \theta)$

Given a data set $x_1, \dots, x_n$ denote its the empirical cdf by F_n.

1. Draw an i.i.d. random sample $X_1^*, \dots, X_n^*$ from F_n. (aka *bootstrap sample*)
2. Compute $T^* = h(X_1^*, \dots X_n^*)$ and $R^* = R(T^*, \hat{\theta})$. (aka *bootstrap statistic*)
3. Repeat Steps 1 and 2 B times to obtain $R_1^*, \dots, R_B^*$.

- "drawing an i.i.d. random sample $X_1^*, \dots, X_n^*$ from F_n" is a fancy way of saying "drawing n elements with replacement from $\{x_1, \dots, x_n\}$".
- Note: The empirical distribution of the bootstrap statistics $R_1^*, \dots R_B^*$ is an approximation of the sampling distribution of $R(T, \theta)$.

Example: Empirical Bootstrap for the SE

Let $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ drawn from F. Let $T = h(X_1, \ldots, X_n)$ be an estimator for θ. Propose an empirical bootstrap procedure for $SE(T)$!

- Since $SE(T) = \mathrm{E}[(T - \mathrm{E}[T])^2]$ is just the standard deviation of T, we decide to bootstrap the distribution of T. The standard deviation of that distribution will be the bootstrap estimate of the SE of T.

Empirical Bootstrap for the SE of T

1. Draw an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from the empirical cdf F_n.
2. Compute $T^* = h(X_1^*, \ldots X_n^*)$.
3. Repeat Steps 1 and 2 B times to obtain $T_1^*, \ldots, T_B^*$.
4. Compute the bootstrap estimate of the SE of T as

$$\widehat{SE}^*(T) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(T_b^* - \frac{1}{B} \sum_{b=1}^{B} T_b^* \right)^2}.$$

Note: *Here, we have taken $R(T, \theta) = T$.*

Example: Empirical Bootstrapping of the Bias

Let $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ drawn from F. Let $T = h(X_1, \ldots, X_n)$ be an estimator for θ and denote the estimate based on $x_1, \ldots, x_n$ by $\hat{\theta} = T(x_1, \ldots, x_n)$. Propose an empirical bootstrap procedure for the bias!

- Since $\mathrm{Bias}(T) = \mathrm{E}[T] - \theta$, we decide to bootstrap the distribution of

$$R(T, F) = T - \theta.$$

  The mean of that distribution will be the bootstrap estimate of the bias.

EMPIRICAL BOOTSTRAP FOR THE BIAS OF T

1. Draw an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from the empirical cdf F_n.
2. Compute $T^* = h(X_1^*, \ldots X_n^*)$ and $R^* = R(T^*, \hat{\theta}) = T^* - \hat{\theta}$.
3. Repeat Steps 1 and 2 B times to obtain $R_1^*, \ldots, R_B^*$.
4. Compute the bootstrap estimate of the bias of T as

$$\widehat{\mathrm{Bias}}^*(T) = \frac{1}{B}\sum_{b=1}^{B} R_b^* = \left(\frac{1}{B}\sum_{b=1}^{B} T_b^* - \hat{\theta}\right).$$

Outline

Parametric Bootstrap

- Suppose we know that the data set $x_1, \dots x_n$ is a realization of a random sample $X_1, \dots X_n$ from $F = F(\cdot, \eta)$ but the parameter η unknown.
- How can we incorporate this information in our bootstrap principle?
- Note: The more information we have about the data, the "better" estimators we can construct, i.e. less biased and more efficient.
- Let T be an estimator of $\theta(\eta)$ and the form of θ (as a function of η) is known. Let $\hat{\eta}$ be an estimate of η. Then, we approximate the sampling distribution of $R(T, \theta)$ by the sampling distribution of $R\big(T^*, \theta(\hat{\eta})\big)$.

Parametric Bootstrap Procedure

- We construct an estimate of F via the "plug-in" principle, i.e. given an estimate $\hat{\eta}$ based on $x_1, \ldots, x_n$ we have $\hat{F} := F(\cdot, \hat{\eta})$.
- Let T be an estimator for θ and construct the "plug-in" estimate $\theta(\hat{\eta})$.

Parametric Bootstrap to estimate $R(T, \theta)$

Given a data set $x_1, \ldots, x_n$ construct estimates $\hat{\eta}$ and $\theta(\hat{\eta})$.

1. Draw an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from $F(\cdot, \hat{\eta})$.
2. Compute $T^* = h(X_1^*, \ldots X_n^*)$ and $R^* = R(T^*, \theta(\hat{\eta}))$.
3. Repeat Steps 1 and 2 B times to obtain $R_1^*, \ldots, R_B^*$.

- "drawing an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from $F(\cdot, \hat{\eta})$" means that we use the computer to simulate random variable with cdf $F(\cdot, \hat{\eta})$.
- The empirical distribution of the bootstrap statistics $R_1^*, \ldots R_B^*$ is an approximation of the sampling distribution of $R(T, \theta(\eta))$.

Example: Parametric Bootstrap for the SE

Let $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ drawn from $F(\cdot, \eta)$, where the parameter η is unknown. Suppose that we are interested in the feature $\theta \equiv \theta(\eta)$. Let $T = h(X_1, \ldots, X_n)$ be an estimator for θ. Let $\hat{\eta}$ be an estimate of η. Propose a parametric bootstrap procedure for the $SE(T)$!

PARAMETRIC BOOTSTRAP FOR THE SE OF T

1. Draw an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from the cdf $F(\cdot, \hat{\eta})$.
2. Compute $T^* = h(X_1^*, \ldots X_n^*)$.
3. Repeat Steps 1 and 2 B times to obtain $T_1^*, \ldots, T_B^*$.
4. Compute the bootstrap estimate of the SE of T as

$$\widehat{SE}^*(T) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(T_b^* - \frac{1}{B} \sum_{b=1}^{B} T_b^* \right)^2}.$$

Outline

Example with R-code: Breakdown voltage

Reconsider the data set on 20 measurements of dielectric breakdown voltage for pieces of epoxy resin (Lecture Week 7, Part 1):

$$\begin{array}{cccccccccc} 24.46 & 25.61 & 26.25 & 26.42 & 26.66 & 27.15 & 27.31 & 27.54 & 27.74 & 27.94 \\ 27.98 & 28.04 & 28.28 & 28.49 & 28.50 & 28.87 & 29.11 & 29.13 & 29.50 & 30.88. \end{array}$$

- Let's consider the following estimators of the center of the distribution:
  - $T_1(X_1, \ldots, X_n) = \bar{X}_n \implies t_1 = 555.86/20 = 27.793.$
  - $T_3(X_1, \ldots, X_n) = X_{\text{median}} \implies t_3 = (27.94 + 27.98)/2 = 27.960.$
- The sample standard deviation is $s_n = 1.462$; hence, we estimate the SE of T_1 as

$$\widehat{SE}(T_1) = 1.462/\sqrt{20} = 0.327.$$

- In the following, we discuss parametric and empirical bootstrap estimates of the SE of T_1. We use above number 0.327 is a reference point.

Example with R-code: Breakdown voltage (Cont.)

- Let's assume that the dielectric breakdown voltage for pieces of epoxy resin is known to be normally distributed with unknown mean μ and variance σ^2.

- Then, e can also use a parametric bootstrap procedure to estimate the SE of $T_1(X_1, \ldots, X_n) = \bar{X}_n$.

```
> ### Parametric Boostrap of SE for the Mean
> B <- 1000  # No. of Bootstrap samples
> n <-  20 # sample size
> means <- matrix(NA, nrow=B, ncol=1)
# Compute bootstrap estimates of the mean
> for (b in 1:B) {
+   X.star <- rnorm(n, 27.793, 1.462) # draw bootstrap samples
+   means[b] <- mean(X.star) # compute bootstrap statistic
+ }
> sd(means) # bootstrap estimate of SE
[1] 0.3283883 # very close to .327, what we got from the formula
> mean((means-27.793)^2) # bootstrap estimate of MSE
[1] 0.1077754
```

Example with R-code: Breakdown voltage (Cont.)

- Now, let's assume that we do not know the distribution of the dielectric breakdown voltage for pieces of epoxy resin.
- Therefore, we use the empirical bootstrap procedure to estimate the SE of $T_1(X_1, \ldots, X_n) = \bar{X}_n$.

```
> ### Empirical Boostrap of SE for the Mean
> X <- c(24.46, 25.61, 26.25, 26.42, 26.66, 27.15, 27.31, 27.54,
+  27.74, 27.94, 27.98, 28.04, 28.28, 28.49, 28.50, 28.87,
+  29.11, 29.13, 29.50, 30.88)
> B <- 1000  # No. of Bootstrap samples
> n <-  20 # sample size
> means <- matrix(NA, nrow=B, ncol=1)
# Compute bootstrap estimates of the mean
> for (b in 1:B) {
+   X.star <- sample(X, n, replace=T) # draw bootstrap samples
+   means[b] <- mean(X.star) # compute bootstrap statistic
+ }
> sd(means) # bootstrap estimate of SE
[1] 0.3192261 # very close to .327, what we got from the formula
> mean((means-27.793)^2) # bootstrap estimate of MSE
[1] 0.09750167
```

Example with R-code: Breakdown voltage (Cont.)

- One can show that the variance of the sample median X_{median} of a random sample $X_1, \ldots, X_n$ from a distribution F with pdf f is

$$\text{Var}(X_{\text{median}}) = \frac{1}{4nf(q_{0.5})^2},$$

  where $q_{0.5}$ is the 50%-percentile (aka median) of F.

- Since the cdf F and pdf f are unknown (otherwise, no need to estimate the median!), this formula is not helpful for estimating the SE of X_{Median}. However, we can use the following empirical bootstrap procedure.

```
> ### Empirical Boostrap of SE for the Median
> meds <- matrix(NA, nrow=B, ncol=1)
# Compute bootstrap estimates of the mean
> for (b in 1:B) {
+   X.star <- sample(X, n, replace=T) # draw bootstrap samples
+   meds[b] <- median(X.star) # compute bootstrap statistic
+ }
> sd(meds) # bootstrap estimate of SE
[1] 0.3193039
```

Outline

Bootstrap confidence intervals for the mean

How can we construct CIs if we have a small sample and the data is from an unknown (not normal) distribution F?

- Recall the approach to small sample CIs for normal data:
  - If we can find numbers $c_l < c_u$ such that

    $$P\left(c_l < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < c_u\right) = 1 - \alpha,$$

    we can construct a $100(1-\alpha)\%$ CI as

    $$\left(\bar{x}_n - c_u \frac{s_n}{\sqrt{n}},\ \bar{x}_n - c_l \frac{s_n}{\sqrt{n}}\right),$$

    where $\bar{x}_n$ and s_n sample average and sd of the data set $x_1, \ldots, x_n$.
  - To find the numbers c_l and c_u we need to know the distribution of

    $$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

  - Since $X_1, \ldots, X_n \sim_{iid} N(\mu, \sigma^2)$, we know that $T \sim t(n-1)$.

Bootstrap confidence intervals for the mean (Cont.)

- Idea: Use the bootstrap principle to approximate the distribution of

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

  - Given a data set $x_1, \ldots, x_n$ determine an estimate $\hat{F}$ of F.
  - Let $X_1^*, \ldots, X_n^*$ be a random sample from $\hat{F}$ and define

$$T^* = \frac{\bar{X}_n^* - \bar{X}_n}{S_n^*/\sqrt{n}}.$$

  - The distribution of T^* can be used to approximate the distribution of T.

Bootstrap confidence intervals for the mean (Cont.)

Empirical bootstrap CI for the mean

Given a data set $x_1, \ldots x_n$ denote its empirical cdf by F_n.

1. Draw an i.i.d. random sample $X_1^*, \ldots, X_n^*$ from F_n.
2. Compute the studentized sample average for the bootstrap data set:
$$T^* = \frac{\bar{X}_n^* - \bar{X}_n}{S_n^*/\sqrt{n}},$$
where $\bar{X}_n^*$ and S_n^* are sample mean and sd of the bootstrap data set $X_1^*, \ldots, X_n^*$.
3. Repeat Steps 1 and 2 B times to obtain $T_1^*, \ldots, T_B^*$.
4. Compute the critical values as the $\alpha/2$ and $1-\alpha/2$ order statistics of $T_1^*, \ldots, T_B^*$, i.e.
$$c_l^* = T^*_{(B\alpha/2)} \qquad \text{and} \qquad c_u^* = T^*_{(B(1-\alpha/2))}.$$
5. A $100(1-\alpha)\%$ empirical bootstrap CI for the mean is
$$\left(\bar{x}_n - c_u^* \frac{s_n}{\sqrt{n}},\, \bar{x}_n - c_l^* \frac{s_n}{\sqrt{n}}\right).$$