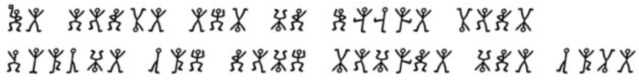
Lecture 6. Bias Correction and Estimation Efficiency

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. Iterative Bias Correction



Let's begin with a biased estimator $\hat{\theta}$ of θ in the problem of estimating θ of the distribution family $\mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$ using data $X = \{X_1, \cdots, X_n\}$, where we assume Θ is compact and our estimator can be standardized by some constant σ, i.e., we have

$$\frac{\sqrt{n}(\hat{\theta} - \mathbb{E}\hat{\theta})}{\sigma} \xrightarrow{d} N(0,1). \tag{1.1}$$

Now, if we look at the bias,

$$Bias(\theta, n) = \mathbb{E}\hat{\theta} - \theta$$

is nothing but a function of the true parameter θ and the sample size n (or think it as a function of θ and $1/n$). If, for a moment, we boldly treat $Bias(\theta, n)$ as a bivariate function and pretend that it is defined on "$\Theta \times \mathbb{R}$" instead of restricting n to be an integer, then by conduct Taylor expansion for the second coordinate with respect to 0, we have

$$Bias(\theta, n) = B_0(\theta) + \sum_{k=1}^{\infty} \frac{B_k(\theta)}{n^k}, \quad \text{with} \max_{\theta \in \Theta, k \geq 0} |B_k(\theta)| < \infty, \tag{1.2}$$

where $\{B_k\}_{k \geq 0}$ are functions of θ who does not depend on the sample size n.

- If $B_0(\theta)$ does not always equals to 0 when θ varies within Θ, i.e., $\hat{\theta}$ is not asymptotically unbiased. According to (1.1) and (1.2) and Slutsky's theorem, we have

$$\sqrt{n}\left(\frac{\hat{\theta} - (\theta + B_0(\theta))}{\sigma}\right) \xrightarrow{d} N(0,1). \tag{1.3}$$

If we denote

$$f(\theta) = \theta + B_0(\theta), \text{ and } g(f(\theta)) = \theta \text{ being the inverse of } f(\cdot),$$

and further assume g' exists and is continuous with $g'(\theta) \neq 0$ for $\theta \in \Theta$, then by applying Delta's method to (1.3), we have

$$\sqrt{n}\left(\frac{g(\hat{\theta}) - g(f(\theta))}{|g'(\theta)|\sigma}\right) = \sqrt{n}\left(\frac{g(\hat{\theta}) - \theta}{|g'(\theta)|\sigma}\right) \xrightarrow{d} N(0,1). \qquad (1.4)$$

Since $|g'(\theta)|\sigma$ is bounded (because Θ is compact and g' is continuous), so (1.4) implies that $g(\hat{\theta}) - \theta \xrightarrow{p} 0$. If we further ask $\{g(\hat{\theta}), n \geq 1\}$ to be uniformly integrable, then

$$\mathbb{E}g(\hat{\theta}) - \theta \to 0, \quad \text{i.e., } g(\hat{\theta}) \text{ is an asymptotically unbiased estimator of } \theta.$$

Meaning we have improved our estimator from a non-asymptotically unbiased one $\hat{\theta}$, to an asymptotically unbiased estimator $g(\hat{\theta})$.

- If $B_0(\theta) \equiv 0$ for $\forall\theta \in \Theta$, i.e., $\hat{\theta}$ is an asymptotically unbiased estimator of θ. Then we have

$$Bias(\theta, n) = \sum_{k=1}^{\infty} \frac{B_k(\theta)}{n^k} = \frac{B_1(\theta)}{n} + O\left(\frac{1}{n^2}\right). \qquad (1.5)$$

and (1.3) change to $\sqrt{n}(\hat{\theta} - \theta)/\sigma \xrightarrow{d} N(0,1)$. Now, if we define

$$\hat{\theta}^{(1)} = \hat{\theta} - \frac{B_1(\hat{\theta})}{n},$$

Since $B_1(\cdot)$ is continuously differentiable, and Θ is compact, so $B_1'(\cdot)$ and $B_1''(\cdot)$ both are bounded, so by Taylor's expansion

$$\mathbb{E}B_1(\hat{\theta}) - B_1(\theta) = B_1'(\theta)\big(\mathbb{E}\hat{\theta} - \theta\big) + B_1''(\tilde{\theta}) \cdot \mathbb{E}\Big[\big(\hat{\theta} - \theta\big)^2\Big] \qquad (1.6)$$

where $\tilde{\theta}$ lays in between of θ and $\hat{\theta}$. Since $\{\hat{\theta}, n \geq 1\}$ are uniformly integrable, therefore, (1.3) implies that

$$\mathbb{E}\Big[n\big(\hat{\theta} - \theta\big)^2/\sigma^2\Big] \to 1, \quad i.e., \ \mathbb{E}\Big[\big(\hat{\theta} - \theta\big)^2\Big] = O\left(\frac{1}{n}\right).$$

Meanwhile, we have

$$\mathbb{E}\hat{\theta} - \theta = \frac{B_1(\theta)}{n} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n}\right),$$

hence (1.6) implies

$$\mathbb{E}B_1(\hat{\theta}) - B_1(\theta) = B_1'(\theta) \cdot O\left(\frac{1}{n}\right) + B_1''(\tilde{\theta}) \cdot O\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right),$$

which further leads to

$$\mathbb{E}\hat{\theta}^{(1)} - \theta = \mathbb{E}\hat{\theta} - \frac{\mathbb{E}B_1(\hat{\theta})}{n} - \theta = -\frac{\mathbb{E}B_1(\hat{\theta}) - B_1(\theta)}{n} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right).$$

Meaning we have improved our estimator from having bias of order $O\left(\frac{1}{n}\right)$, to another estimator whose bias is of order $O\left(\frac{1}{n^2}\right)$.

- We may further repeat this process, by define

$$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} - \frac{B_k(\hat{\theta}^{(k-1)})}{n^k},$$

and we would have

$$\mathbb{E}\hat{\theta}^{(k)} - \theta = O\left(\frac{1}{n^{k+1}}\right),$$

for arbitrary fixed k.

Therefore, for an arbitrary estimator (biased or unbiased), as long as we have the regularity condition holds (marked as blue, which is also not that restrict, for example, MLE would usually fulfill those conditions), we can **iteratively improve our estimator** through the above procedure. But due to computational complicity, we often stop when we have $\hat{\theta}^{(1)}$, which is already asymptotically unbiased with bias having order $O\left(\frac{1}{n^2}\right)$, and it is called the **bias corrected estimator**.

- *Example* 1.1. Let $X = \{X_1, \cdots, X_n\}$ be i.i.d $N(\mu, \sigma^2)$ random variables.

(i) Please give the bias corrected MLE of σ^2.
(ii) Please give the bias corrected MLE of μ^2 when σ^2 is known.
(iii) Please give the bias corrected MLE of μ^2 when σ^2 is not known.
(iv) Compare the MSE (risk function corresponding to the quadratic loss) of the MLE of σ^2 to the MSE of the bias corrected MLE of σ^2.

Answer. Notice that we are focusing on correcting bias for the MLE, which as mentioned earlier, usually fulfill the regularity conditions. Now,

(i) Since the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

thus, we have

$$\mathbb{E}\hat{\sigma}^2 - \sigma^2 = -\frac{1}{n}\sigma^2,$$

therefore,

$$\left(\hat{\sigma}^2\right)^{(1)} = \hat{\sigma}^2 - \left(-\frac{\hat{\sigma}^2}{n}\right) = \frac{n+1}{n}\hat{\sigma}^2 = \frac{n+1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is the bias corrected MLE of σ^2, and

$$Bias\left[\left(\hat{\sigma}^2\right)^{(1)}\right] = \mathbb{E}\left[\frac{n+1}{n}\hat{\sigma}^2\right] - \sigma^2 = \frac{(n+1)(n-1)-n^2}{n^2}\sigma^2 = -\frac{1}{n^2}\sigma^2.$$

(ii) When σ^2 is known, we know the MLE of μ^2 is $\bar{X}^2$ according to the invariance principle of MLE. Notice that $\bar{X} \sim N(\mu, \sigma^2/n)$, so

$$\mathbb{E}\left(\bar{X}\right)^2 - \mu^2 = \mu^2 + \frac{\sigma^2}{n} - \mu^2 = \frac{\sigma^2}{n},$$

therefore,

$$\left(\bar{X}^2\right)^{(1)} = \bar{X}^2 - \frac{\sigma^2}{n}$$

is the bias corrected MLE of μ^2 when σ^2 is known, and

$$Bias\left[\left(\bar{X}^2\right)^{(1)}\right] = \mathbb{E}\left[\bar{X}^2 - \frac{\sigma^2}{n}\right] - \mu^2 = 0.$$

(iii) When σ^2 is not known, we know the MLE of (μ^2, σ^2) are $(\bar{X}^2, \hat{\sigma}^2)$ according to the invariance principle of MLE. Since $\bar{X} \sim N(\mu, \sigma^2/n)$ and

$$\mathbb{E}\left(\bar{X}\right)^2 - \mu^2 = \mu^2 + \frac{\sigma^2}{n} - \mu^2 = \frac{\sigma^2}{n},$$

therefore,

$$\left(\bar{X}^2\right)^{(1)} = \bar{X}^2 - \frac{\hat{\sigma}^2}{n} = \bar{X}^2 - \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

is the bias corrected MLE of μ^2 when σ^2 is not known, and

$$Bias\left[\left(\bar{X}^2\right)^{(1)}\right] = \mathbb{E}\left[\bar{X}^2 - \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] - \mu^2 = \mu^2 + \frac{\sigma^2}{n} - \frac{(n-1)\sigma^2}{n^2} - \mu^2 = \frac{\sigma^2}{n^2}.$$

(iv) For the MSE of MLE of σ^2, we have $MSE(\hat{\sigma}^2) = (2n-1)\sigma^4/n^2$. Meanwhile, for the bias corrected MLE of σ^2 derived in (i), we have

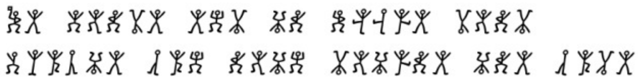
$$\begin{aligned} MSE\left[\left(\hat{\sigma}^2\right)^{(1)}\right] &= \frac{(n+1)^2}{n^2}\operatorname{Var}\left(\hat{\sigma}^2\right) + \left(Bias\left[\left(\hat{\sigma}^2\right)^{(1)}\right]\right)^2 \\ &= \frac{2(n+1)^2(n-1)\sigma^4}{n^4} + \frac{\sigma^4}{n^4} > MSE(\hat{\sigma}^2) = (2n-1)\sigma^4/n^2 \end{aligned}$$

when $n > 1$. Meaning even though our bias corrected estimator reduces bias, it causing the overall risk to increase for arbitrary $\sigma^2 > 0$, so it is actually inadmissible when we use MSE as our risk function.

□

Except for creating inadmissible estimator, another limitation of this iterative bias correction is, it requires the calculation of exact functions $\{B_k\}_{k\geq 0}$, which can be troublesome when we facing complicated distribution functions.

2. Jackknife Estimator



As an alternative to the iterative bias correction methods, the jackknife estimator introduced in this section forms a general method for bias reduction which was initiated by Quenouille (1949), Quenouille (1956) and later named the jackknife by Tukey (1958).

Definition 2.1 (♣ **Jackknife Estimator**). Let $T(X)$ be an estimator of a parameter $\tau(\theta)$ based on a sample $X = \{X_1, \cdots, X_n\}$ and satisfying

$$\mathbb{E}T(X) = \tau(\theta) + \frac{B_1(\theta)}{n} + \frac{B_2(\theta)}{n^2} + O\left(\frac{1}{n^3}\right),$$

where B_1, B_2 are functions of θ who does not depend on the sample size n. Define X_{-i} to be the sample excluding X_i, or in other words,

$$X_{-i} = \{X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n\}.$$

It's often called the **left one out sample** (left i out sample) or **delete one sample** (delete i sample). Then, **the jackknife estimator of $\tau(\theta)$ based on $T(X)$**, or sometimes called **the jackknife version of $T(X)$**, is defined as

$$\begin{aligned} T_J(X) &= nT(X) - \frac{n-1}{n}\sum_{i=1}^{n} T(X_{-i}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[T(X) + (n-1)\big(T(X) - T(X_{-i})\big)\Big] \triangleq \frac{1}{n}\sum_{i=1}^{n} T_{Ji}(X) \end{aligned} \tag{2.1}$$

Now, for arbitrary $1 \leq i \leq n$ and large n, notice that

$$\mathbb{E}T(X_{-i}) = \tau(\theta) + \frac{B_1(\theta)}{n-1} + \frac{B_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Therefore,

$$\begin{aligned} \mathbb{E}T_J(X) = n\tau(\theta) + B_1(\theta) + \frac{B_2(\theta)}{n} + O\left(\frac{1}{n^2}\right) \\ - (n-1)\cdot\left[\tau(\theta) + \frac{B_1(\theta)}{n-1} + \frac{B_2(\theta)}{(n-1)^2} + O\left(\frac{1}{n^3}\right)\right] \\ = \tau(\theta) + \frac{B_2(\theta)}{n(n-1)} + O\left(\frac{1}{n^2}\right) = \tau(\theta) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Meaning we have improved our estimator from having bias of order $O\left(\frac{1}{n}\right)$, to the jackknife estimator whose bias is of order $O\left(\frac{1}{n^2}\right)$. While most importantly, we dont have to know the exact form of B_1 and B_2.

As one may notice, we can also repeating this bias-correcting procedure, such as the jackknife estimator based on $T(X)$ defined in (2.1) is called the first-order jackknife. We may obtain a second-order jackknife estimator by jackknifing $T_J(X)$ in (2.1). Practically, it will reduce bias in exchange of the increment of variance.

2.1. Delete-d jackknife

The jackknife estimator $T_J(X)$ is often been treated as sum of "weak dependent" random variables as in (2.1). Therefore its variance can often be estimated by a "pseudo-variance" given by

$$V_J^2(X) \triangleq \frac{1}{n-1}\sum_{i=1}^{n}\big(T_{Ji}(X) - T_J(X)\big)^2.$$

But this impression of $\{T_{Ji}\}_{1\le i\le n}$ being "weakly dependent" is not always true. For instance, when n is large, one would suspect that terms in $\{T_{Ji}\}_{1\le i\le n}$ might be too close to each other, causing the pseudo-variance $V_J^2(X)$ to be mostly sampling error. In that case,we can define a delete-d jackknife instead of leaving out one observation at a time.

Specifically, in a delete-d jackknife, define

$$X_{-\beta} = X_{-(\beta_1,\cdots,\beta_k)} = X - \{X_{\beta_1},\cdots,X_{\beta_k}\},$$

then the delete-d jackknife is of a statistic $T(X)$ is defined as

$$\begin{aligned} T_{J(d)}(X) &= \frac{n}{d}\cdot T(X) - \frac{n-d}{d}\frac{1}{\binom{n}{d}}\sum_{\beta} T(X_{-\beta}) \\ =&\frac{1}{d\binom{n}{d}}\sum_{\beta}\Big[d\cdot T(X) + (n-d)\big(T(X) - T(X_{-\beta})\big)\Big] \triangleq \frac{1}{d\binom{n}{d}}\sum_{\beta} T_{J(d),\beta}(X) \end{aligned}$$

where $\sum_\beta$ represents summation over all possible value of β, and $\beta = (\beta_1,\cdots,\beta_k)$ take values as arbitrary d terms of $\{1,\cdots,n\}$ without repetition.

2.2. The Block Jackknife

Similar in the case of delete-d Jackknife, when our sample forms a time series observations instead of i.i.d random variables, we would worry about $\{T_{Ji}\}_{1\le i\le n}$ being too close to each other. Therefore, to create "distance" between $\{T_{Ji}\}_{1\le i\le n}$, we may use the **block technique**.

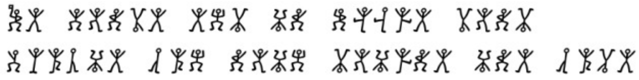
Specifically speaking, assume $n = n_b k$, where k is the block size and n_b is the number of blocks. Therefore, we may leaving out a whole block of data at a time. Define

$$X_{-b(i)} = X_{-((i-1)k+1,\cdots,ik)} = X - \{X_{(i-1)k+1},\cdots,X_{ik}\}$$

then the block jackknife with block size k is defined as

$$\begin{aligned}T_{BJ}(X) &= \frac{n}{k}\cdot T(X) - \frac{n-k}{n}\sum_{i=1}^{n_b} T(X_{-b(i)}) \\ &= \frac{1}{n}\sum_{i=1}^{n_b}\Big[k\cdot T(X) + (n-k)\big(T(X) - T(X_{-b(i)})\big)\Big] \triangleq \frac{1}{n}\sum_{i=1}^{n_b} T_{BJ,i}(X).\end{aligned}$$

3. Bootstrap Estimator



3.1. Prelude Example of Bootstrap Resampling and Simulation: Estimating the Variance of a Median Estimator

Consider a simple problem start from a random sample $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$, which is a collection of n realizations of the r.v. X, been drawn independently and identically from an unknown population distribution function F_0. We are interested in finding an estimator for the median of F_0, which is defined as any real number m such that

$$\mathbb{P}(X \le m) \ge \frac{1}{2}\,, \quad \text{and} \quad \mathbb{P}(X \ge m) \ge \frac{1}{2}\,,$$

and we are interested in the variance of our median estimator.

- For the first task: finding an estimator for the median.

  Notice that we can rewrite the definition of median, which is any real number m such that

$$\begin{aligned} \mathbb{E}_{F_0}\mathbb{1}\big(x \le m\big) &= \int \mathbb{1}\big(x \le m\big)\, dF_0(x) \ge \frac{1}{2}, \\ \text{and}\quad \mathbb{E}_{F_0}\mathbb{1}\big(x \ge m\big) &= \int \mathbb{1}\big(x \ge m\big)\, dF_0(x) \ge \frac{1}{2}. \end{aligned} \tag{3.1}$$

  Or more specifically, we say m is the solution of a "population equation" given above. But since F_0 is unknown, so we think about estimating F_0 by the empirical distribution

$$\hat{F}_0(x) = F_1(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big(X_i \le x\big).$$

  Now, from Glivenko-Cantelli theorem we know that the empirical distribution function is an good estimator of the population distribution function, and by using $\hat{F}_0 = F_1$ to estimate F_0, we can find an estimator for the meadian by looking at the following equations of m,

$$\begin{aligned} &\mathbb{E}_{F_1} \mathbb{1}\big(x \le m\big) = \int \mathbb{1}\big(x \le m\big)\, d\hat{F}_0(x) \ge \frac{1}{2}, \\ \text{and} \quad &\mathbb{E}_{F_1} \mathbb{1}\big(x \ge m\big) = \int \mathbb{1}\big(x \ge m\big)\, d\hat{F}_0(x) \ge \frac{1}{2}. \end{aligned} \tag{3.2}$$

As one may notice, we just simply replace the F_0 in the population equation (3.1) to $\hat{F}_0 = F_1$, which the new equation (3.2) is called a "sample equation". And the solution of this sample equation, which is any real number m s.t.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(X_i \le m\big) \ge \frac{1}{2}\ , \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(X_i \ge m\big) \ge \frac{1}{2}\ ,$$

and the solution can be further simplified to the sample median

$$\hat{\xi}_{1/2} = X_{(\lfloor \frac{n}{2} \rfloor)}.$$

We call this estimator $\hat{\xi}_{1/2}$ as the bootstrap estimate.

Before the second task, we first introduce the concept of resampling.

Definition 3.1 (♣ **Resampling**). In nonparametric problems, when given a random sample $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$, a resample

$$\mathcal{X}^* = \{X_1^*, \cdots, X_n^*\}$$

is an unordered collection of n items drawn randomly from $\mathcal{X}$ with replacement, i.e., $\mathcal{X}^*$ forms a random sample drawn from the empirical distribution $\hat{F}_0 = F_1$ of the given sample $\mathcal{X}$,

$$\mathbb{P}\big(X_i^* = X_j | \mathcal{X}\big) = n^{-1}, \quad 1 \le i, j \le n.$$

and we further denote the empirical distribution of $\mathcal{X}^*$ as

$$\hat{F}_1(x) = F_2(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i^* \le x).$$

Remark 3.2. One thing worth emphasizing is, based on definition, $\mathcal{X}^*$ is likely to contain repeats, all of which must be listed in the collection $\mathcal{X}^*$. One should distinguish this with the regular set notation.

- For the second task: estimating the variance of our median estimator.

  Despite that we are stating the obvious, it's of great importance to notice that the parameter of interests of a distribution F_0 is simply a function of

F_0, i.e, as in this case, we have the median $m = g(F_0)$ for some function g, and we can see that

$$g(F_0) = \arg\max_m \mathbb{1}\left(F_0(m) \geq \frac{1}{2}, 1 - F_0(m-) \geq \frac{1}{2}\right).$$

Therefore, the variance of our bootstrap estimator $\hat{\xi}_{1/2} = g(F_1)$ of the median $m = g(F_0)$, can be expressed as

$$\operatorname{Var}(\hat{\xi}_{1/2}) \triangleq \sigma^2 = \mathbb{E}\left[\left(g(F_1) - \mathbb{E}\big[g(F_1)|F_0\big]\right)^2 \Big| F_0\right]. \tag{3.3}$$

However, it's obviously difficult to calculate the exact variance of our bootstrap median estimator even after we derived such a population equation expression (3.3), so instead, we can look for an estimator of the variance of $\hat{\xi}_{1/2}$ using again a similar approach we used in obtaining our bootstrap median estimator. Instead of solving the "population equation" (3.3), we solve $\hat{\sigma}^2$ for the following "sample equation",

$$\hat{\sigma}^2 = \mathbb{E}\left[\left(g(F_2) - \hat{\mu}\right)^2 \Big| F_1\right], \quad \text{where} \quad \hat{\mu} = \mathbb{E}\big[g(F_2)\big|F_1\big], \tag{3.4}$$

and $g(F_2)$ is simply the median of $\mathcal{X}^*$. Unfortunately, even with the form of F_1, the calculation of above sample equation (3.4) is still too complicated, so one practical way is to do an approximation using law of large numbers, i.e., instead of just resampling once, we resample from $\mathcal{X}$ independently for B times and obtain $\{\mathcal{X}_1^*, \mathcal{X}_2^*, \cdots, \mathcal{X}_B^*\}$, and for each i between 1 to B, we have one resample $\mathcal{X}_i^* = \{X_{i1}^*, \cdots, X_{in}^*\}$ with empirical distribution F_{2i}. Thus, by law of large numbers and conditional on F_1, we have

$$\tilde{\mu} \triangleq \frac{1}{B}\sum_{i=1}^{B} g(F_{2i}) \xrightarrow{p} \hat{\mu} = \mathbb{E}\big[g(F_2)\big|F_1\big],$$

$$\text{and} \quad \tilde{\sigma}^2 \triangleq \frac{1}{B}\sum_{i=1}^{B} \left(g(F_{2i}) - \tilde{\mu}\right)^2 \xrightarrow{p} \hat{\sigma}^2 = \mathbb{E}\left[\left(g(F_2) - \hat{\mu}\right)^2 \Big| F_1\right],$$

should be close to $\hat{\mu}$ and $\hat{\sigma}^2$ respectively when B is sufficiently large, provided that $\mathbb{E}\tilde{\sigma}^4 < +\infty$.

As a summary of the above content, our purpose is to calculate the variance of our bootstrap median estimator, which satisfy a population equation but unable to be solved (Because F_0 is unknown). So, instead of solving the population equation, we samply solve the corresponding sample equation, which the solution $\hat{\sigma}^2$ is called the **bootstrap variance estimator of the estimator $\hat{\xi}_{1/2}$**. But again, solving the sample equation is not feasible due to heavy complixity, so we use law of large numbers to calculated an approximation $\tilde{\sigma}^2$, which is called the **bootstrap variance approximation of the estimator $\hat{\xi}_{1/2}$**.

The above is the rigorous definition, but due to the simplicity of calculation of the approximation, people sometimes directly using $\tilde{\sigma}^2$ even when $\hat{\sigma}^2$ is solvable in certain cases. Therefore, in tremendous amount of books or papers, when people mention bootstrap variance estimator, they are refering to $\tilde{\sigma}^2$ for most of time.

- Coding-wise: Sudocode for finding a numerical approximation of our bootstrap variance estimator of the median estimator $\hat{\xi}_{1/2}$.

Algorithm 1 : Approximation of Bootstrap Variance Estimator

```
Require: $\mathcal{X}$, $B$,
Ensure: sigmasquare
Function ApproxSigma($\mathcal{X}$,$B$){
    $M \leftarrow rep(0, B)$
    for $i$ in $1 : B$ do
        M[i]=median(sample($\mathcal{X}$,length($\mathcal{X}$),TRUE))
    end for
    return sigmasquare = $(B - 1) * var(M)/B$
end Function
```

3.2. Generally definition of Bootstrap Principle

Comparing (3.1) to (3.2), (3.3) to (3.4), we see many statistical problems may be posed in terms of a solution to a "population equation", involving integration with respect to the population distribution function F_0. Formally, given a functional f_t from a class $\{f_t : t \in \tau\}$, we wanna determine that value t_0 of t which solves an equation

$$\mathbb{E}\left\{ f_t(F_1, F_0) \middle| F_0 \right\} = 0 \, .$$

This is called the **"population equation"** because we need properties of the population if we are to solve this equation exactly, and conditioning on F_0 here serves the role of emphasising that the expectation is taken with respect to the distribution F_0.

But since F_0 is unknown, so solve the population equation is not feasible here. In order to obtain an approximation solution of the population equation, we follow the same resampling procedure and argument, that to be said, we sample from F_1 to obtain a F_2, then just simply replace the pair (F_0, F_1) by (F_1, F_2), thereby, transforming the above population equation into a **"sample equation"**

$$\mathbb{E}\left\{ f_t(F_2, F_1) \middle| F_1 \right\} = 0 \, .$$

The solution $\hat{t}_0$ to sample equation is an estimator for the true solution t_0 of the population equation, which, $\hat{t}_0$ **is called the bootstrap estimate and we refer this procedure as "the bootstrap principle"**.

Sometimes even the sample equation and $\hat{t}_0$ is still hard to solve, so we go one step further by approaximate the sample equation using law of large numbers. By independently resample F_{2i} from F_1 for $i = 1, \cdots, B$, and we solve the **"approximated sample equation"**

$$\frac{1}{B}\sum_{i=1}^{B} f_t(F_{2i}, F_1) = 0 \ .$$

The solution $\tilde{t}_0$ to the approximated sample equation is nothing but an estimator for the true solution t_0 of the population equation, which, **$\tilde{t}_0$ is called the bootstrap approximation estimator**.

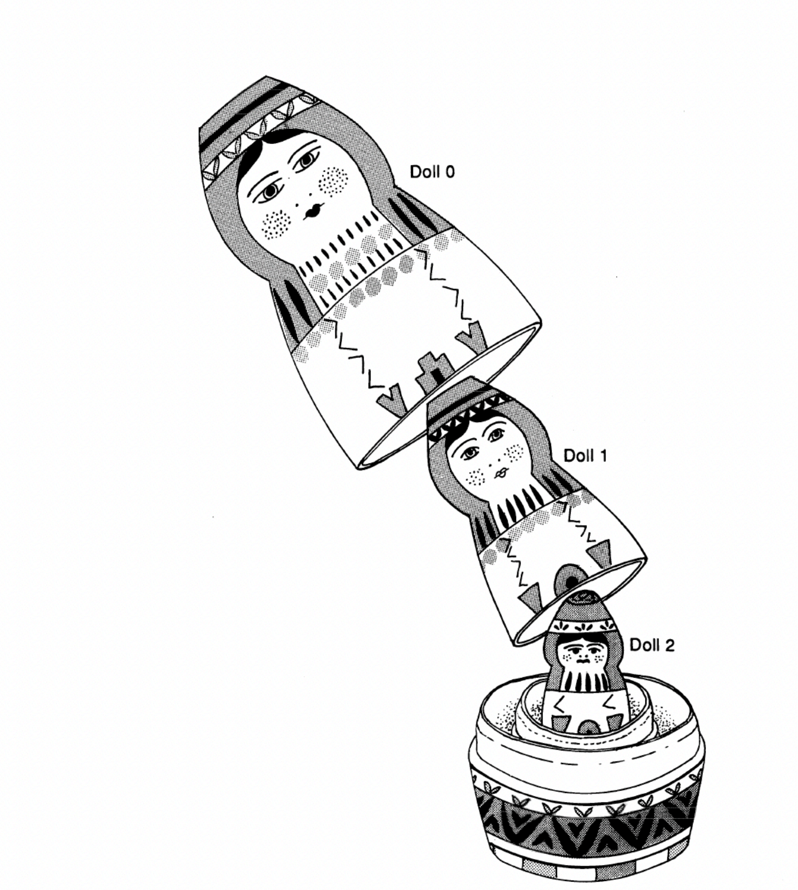


Figure 1: Figure is from the Book "The Bootstrap and Edgeworth Expansion" written by Peter Hall. A metaphorically description of the main principle of Bootstrap: Russian Matryoshka nesting dolls.

Now, after we obtained a bootstrap estimator $\hat{t}_0$ of t_0, and we have $(\hat{t}_0 - t_0)/\sigma_n \xrightarrow{d}$

$N(0,1)$, a natural followup question is to obtain a confidence interval of t_0 with confidence level $1-\alpha$. If σ_n is known, then

$$\left[\hat{t}_0 - \mu_{1-\alpha/2}\cdot\sigma_n, \hat{t}_0 + \mu_{1-\alpha/2}\cdot\sigma_n\right] \tag{3.5}$$

fulfills our requirements. However, if σ_n is not known, we may have to further estimate σ_n using bootstrap method and replace σ_n with $\hat{\sigma}_n$ in (3.5), and as a consequence, we resample from F_2 for a F_3. Apparently, bootstrap resampling procedure may keep going on and it has been metaphorically described as a Russian Matryoshka nesting dolls in Hall (2013).

Practically, in order to prevent resampling for too much layers and for avoiding other critical disadvantage of bootstrap, Horowitz (2001) made several practical suggestions,

- Don't use the bootstrap to estimate the probability distribution of a nonasymptotic pivotal quantity such as a regression slope coefficient or standard error if an asymptotic pivotal quantity is available.
- Do recenter the residuals of an overidentified model before applying the bootstrap to the model.
- Don't apply the bootstrap to models for dependent data, semi or nonparametric estimators, or non-smooth estimators without careful justification.

3.3. Bootstrap in Bias Reduction

We shall take several examples for the illustration of the powerfulness of bootstrap. Without other specification, we denote $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$ a random sample drawn from the r.v. $X \sim F_0$, and denote the empirical distribution of $\mathcal{X}$ as F_1 in the remaining chapters.

• *Example* 3.3 (♣ **Bootstrap Mean Estimator**). Suppose we want to estimate the population mean θ,

$$\theta = \mathbb{E}(X_1) = \int x dF_0(x) = \theta(F_0)$$

so we can see θ is the solution t_0 to the population equation with

$$\mathbb{E}\left\{ f_t(F_1, F_0) \middle| F_0 \right\} = \int (x-t) dF_0(x) = 0,$$

hence our sample equation should be

$$\mathbb{E}\left\{ \int x dF_1(x) - t \middle| F_1 \right\} = 0$$

which the solution $\hat{t}_0 = \hat{\theta} = \theta(F_1)$ is

$$\hat{t}_0 = \int x dF_1(x) = \bar{X},$$

hence the bootstrap mean estimate for θ is $\bar{X}$. As a matter of fact, for any estimand admit the form $\theta(F_0)$, we can see through the above argument, the bootstrap estimate for $\theta(F_0)$ is $\theta(F_1)$.

- *Example* 3.4 (♣ **Bootstrap MSE Estimator for the MSE of a Bootstrap estimator of θ_0**). Suppose we want to estimate the MSE of a Bootstrap estimator of θ, assume that θ_0 admits the form $\theta_0 = \theta(F_0)$, so based on the early example, we have the bootstrap estimator of $\theta_0 = \theta(F_0)$ is $\hat{\theta}_0 = \theta(F_1)$, and the MSE we wanna estimate here is τ^2 defined as

$$\tau^2 = \mathbb{E}(\hat{\theta}_0 - \theta_0)^2 = \mathbb{E}\left\{ [\theta(F_1) - \theta(F_0)]^2 \middle| F_0 \right\}.$$

Accordingly, τ^2 is the solution t_0 to the population equation with

$$f_t(F_1, F_0) = \big(\theta(F_1) - \theta(F_0)\big)^2 - t,$$

hence our sample equation should be

$$\mathbb{E}\left\{ \big(\theta(F_2) - \theta(F_1)\big)^2 \middle| F_1 \right\} = t \ ,$$

which leads to the bootstrap MSE estimate for the MSE of $\hat{\theta}_0 = \theta(F_1)$ is given by the solution

$$\hat{t}_0 = \hat{\tau}^2 = \mathbb{E}\left\{ \big(\theta(F_2) - \theta(F_1)\big)^2 \middle| F_1 \right\}.$$

Further, if the above expectation within the definition of $\hat{t}_0$ with respect to F_1 is infeasible to compute, we can use the law of large numbers to obtain the approximation of $\hat{\tau}^2$, which is given by resampling B times, obtain $\{\mathcal{X}_1^*, \cdots, \mathcal{X}_B^*\}$, such that each $\mathcal{X}_i^*$ has empirical distribution F_{2i}, for $1 \le i \le B$, and

Algorithm 2 : Approximation of Bootstrap Variance Estimator

```
Require: $\mathcal{X}$, $B$,
Ensure: tausquare
Function ApproxMSE($\mathcal{X}$,$B$){
    $M \leftarrow rep(0, B)$
    for $i$ in $1 : B$ do
        M[i]=$\theta$(sample($\mathcal{X}$,length($\mathcal{X}$),TRUE))
    end for
    return tausquare $= mean\Big(\big(M - rep(\theta(\mathcal{X}), B)\big)^2\Big)$
end Function
```

$$\tilde{\tau}^2 = \frac{1}{B}\sum_{i=1}^{B} \big(\theta(F_{2i}) - \theta(F_1)\big)^2 = \frac{1}{B}\sum_{i=1}^{B} \big(\theta(\mathcal{X}_i^*) - \theta(\mathcal{X})\big)^2 .$$

• *Example* 3.5 (♣ **Bootstrap Bias-Corrected Estimate**). For an estimand θ_0 admits the form $\theta_0 = \theta(F_0)$, we know the bootstrap estimator of θ_0 is $\hat{\theta}_0 = \theta(F_1)$ from Example 3.3. But this bootstrap estimate can be an biased estimator, whose bias is given by

$$Bias(\hat{\theta}_0) = \mathbb{E}\big[\theta(F_1) - \theta_0 | F_0\big] = \mathbb{E}\big[\theta(F_1)|F_0\big] - \theta(F_0).$$

So the bias is also a function of F_0. Natually, one would think about give a bootstrap estimator for this bias, i.e., define $f_t(F_1, F_0)$ as

$$f_t(F_1, F_0) = \theta(F_1) - \theta(F_0) - t \ ,$$

this $Bias(\hat{\theta}_0)$ is the solution t_0 of the population equation (equation about t) given by

$$E\left\{ f_t(F_1, F_0) \middle| F_0 \right\} = 0 \ .$$

By applying the bootstrap principle, we replace (F_1, F_0) by (F_2, F_1) and solve the solution $\hat{t}_0$ of the corresponding sample equation (equation about t),

$$\mathbb{E}\left\{ f_t(F_2, F_1) \middle| F_1 \right\} = \mathbb{E}\left\{ \theta(F_2) - \theta(F_1) - t \middle| F_1 \right\} = \mathbb{E}\left\{ \theta(F_2) \middle| F_1 \right\} - \theta(F_1) - t = 0$$

which leads to the solution $\hat{t}_0$

$$\widehat{Bias(\hat{\theta}_0)} \triangleq \hat{t}_0 = \mathbb{E}\left\{ \theta(F_2) \middle| F_1 \right\} - \theta(F_1) \ ,$$

and we have $\hat{t}_0$ is our bootstrap bias estimator. If the expectation respect to F_1 is too complicated to compute, we can do one step further by using the law of large numbers, and obtain an approximation bootstrap bias estimate with form

$$\widetilde{Bias(\hat{\theta}_0)} \triangleq \left(\frac{1}{B} \sum_{i=1}^{B} \theta(F_{2i}) \right) - \theta(F_1) = \left(\frac{1}{B} \sum_{i=1}^{B} \theta(\mathcal{X}_i^*) \right) - \theta(\mathcal{X})$$

where we resampled for B times and have the resample $\{\mathcal{X}_1^*, \cdots, \mathcal{X}_B^*\}$ with F_{2i} being the empirical distribution of $\mathcal{X}_i^*$. So instead of our original estimator $\hat{\theta}_0$ for θ_0, we may use the new bias corrected estimator defined as

$$T \triangleq \hat{\theta}_0 - \widehat{Bias(\hat{\theta}_0)} = 2\theta(\mathcal{X}) - \mathbb{E}\left\{ \theta(\mathcal{X}^*) \middle| \mathcal{X} \right\}$$

or its approximation form

$$\tilde{T} \triangleq \hat{\theta}_0 - \widetilde{Bias(\hat{\theta}_0)} = 2\theta(\mathcal{X}) - \left(\frac{1}{B} \sum_{i=1}^{B} \theta(\mathcal{X}_i^*) \right) .$$

For a more concrete example, if we define μ to be the population mean, and say we are interested in $\theta_0 = \theta(F_0) = \mu^3$, which corresponds to the population equation

$$f_t(F_1, F_0) = \left(\int x dF_0\right)^3 - t.$$

and the bootstrap estimator of this $\theta_0 = \theta(F_0) = \mu^3$ is given by

$$\hat{\theta}_0 = \theta(F_1) = \left(\int x dF_1\right)^3 = \big(\bar{X}\big)^3.$$

On the other hand, since

$$\theta(F_2) = \left(\int x dF_2\right)^3 = \big(\bar{X}^*\big)^3,$$

so the bootstrap bias estimator is given by

$$\begin{aligned}
\widehat{Bias(\hat{\theta}_0)} &= \mathbb{E}\left\{\theta(F_2)\middle|F_1\right\} - \theta(F_1) = \mathbb{E}\left\{\Big(\frac{1}{n}\sum_{i=1}^{n} X_i^*\Big)^3\middle|\mathcal{X}\right\} - \big(\bar{X}\big)^3 \\
&= \frac{1}{n^3}\Big\{n\mathbb{E}\Big[\big(X_i^*\big)^3\Big|\mathcal{X}\Big] + 3n(n-1)\mathbb{E}\Big[\big(X_i^*\big)^2\Big|\mathcal{X}\Big]\mathbb{E}\big[X_j^*\big|\mathcal{X}\big] \\
&\qquad + n(n-1)(n-2)\Big[\mathbb{E}\big[X_j^*\big|\mathcal{X}\big]\Big]^3\Big\} - \big(\bar{X}\big)^3 \\
&= 3n^{-1}\bar{X}\hat{\sigma}^2 + n^{-2}\hat{\gamma},
\end{aligned}$$

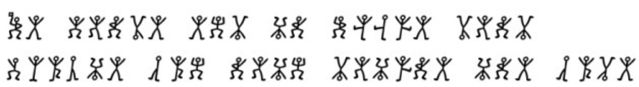
where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\ , \text{ and } \hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3\ .$$

Therefore, the bootstrap bias-corrected estimator is given by

$$T = \bar{X}^3 - 3n^{-1}\bar{X}\hat{\sigma}^2 - n^{-2}\hat{\gamma}.$$

4. Efficiency, Asymptotic Efficiency and Relative Efficiency



Let us denote $\delta(X)$ to be an estimator of $g(\theta)$. If $\delta(X)$ is unbiased, then according to Cramér-Rao lower bound, we know

$$\mathrm{Var}(\delta) \geq \frac{\big(g'(\theta)\big)^2}{I_X(\theta)}.$$

where $I_X(\theta)$ is the information contained in the whole sample, if we have $X = \{X_1, \cdots, X_n\}$ being a random sample, then we have $I_X(\theta) = nI_{X_1}(\theta)$.

Since variance is without doubt one of the most crucial quantity we are genuinely cared about, which measures the variation of our estimator. So it's natural to look at the ratio of the variance to this lower bound, formally, we define the **unbiased in the limit, asymptotic unbiased, efficiency, asymptotic efficiency, and relative efficiency**, as following,

Definition 4.1. For $\{\delta_n\}_{n\geq 1}$ and $\{T_n\}_{n\geq 1}$ being two sequence of estimators for $g(\theta)$. Assume $k_{n1}(\delta_n - g(\theta)) \overset{L_2}{\to} H_1$ and $k_{n2}(T_n - g(\theta)) \overset{L_2}{\to} H_2$ for some sequence of numbers $\{k_{n1}\}_{n\geq 1}$ and $\{k_{n1}\}_{n\geq 2}$.

- We say δ_n (and T_n) is **unbiased in the limit** if $\lim_{n\to\infty} \mathbb{E}\delta_n = g(\theta)$ (similarly, if $\lim_{n\to\infty} \mathbb{E}T_n = g(\theta)$).
- We say δ_n (and T_n) is **asymptotic unbiased** if $\mathbb{E}H_1 = 0$ (and $\mathbb{E}H_2 = 0$).
- The **efficiency** of the estimator δ_n (and $\text{eff}(T_n, g(\theta))$ can be similarly defined) is defined as

$$\text{eff}(\delta_n, g(\theta)) = \frac{\left(g'(\theta)\right)^2}{I_X(\theta)} \Bigg/ \text{Var}(\delta_n).$$

For the **asymptotic efficiency, and relative efficiency**, we usually only discuss these two terms in the sense of asymptotic normality holds for our estimator δ_n and T_n obtained using a random sample $X = \{X_1, \cdots, X_n\}$, i.e., we have $\sqrt{n}(\delta_n - g(\theta)) \overset{L_2}{\to} N(0, \nu_1^2(\theta))$ and $\sqrt{n}(T_n - g(\theta)) \overset{L_2}{\to} N(0, \nu_2^2(\theta))$ instead of having some non-normal limit distributions H_1 and H_2, accordingly,

- The **asymptotic efficiency** of the estimator δ_n (and $\text{AE}(T_n, g(\theta))$ can be similarly defined) is defined as

$$\text{AE}(\delta_n, g(\theta)) = \frac{\left(g'(\theta)\right)^2}{I_{X_1}(\theta)} \Bigg/ \nu_1^2(\theta).$$

- The **relative efficiency** (or **asymptotic relative efficiency**) of the estimator δ_n with respect to T_n is defind as

$$\text{ARE}(\delta_n, T_n) = \nu_1^2(\theta) \Big/ \nu_2^2(\theta).$$

* **Remark 4.2.** *Here in the definition, we will always assume the condition required for Cramér-Ra0 lower bound and the existence of Fisher information holds, which the core one is the exchange ability of the differential operator and the integration operator.*

Notice that, when the estimator δ_n are unbiased, we have the $\text{eff}(\delta_n, g(\theta))$ and $\text{AE}(\delta_n, g(\theta))$ all falls in between 0 and 1. But when came to some biased estimator δ_n, even if we still have δ_n to be asymptotic unbiased, do there exist δ_n

such that the asymptotic efficiency great than 1? The answer to this question is affirmative, we look at the following famous counterexample called the Hodge-Le Cam estimator,

- *Example* 4.3 (♣ **Superefficiency and Hodge-Le Cam estimator**). Assune we have a random sample $X = \{X_1, \cdots, X_n\}$ drawn from the $N(\theta, 1)$ and we are interested in estimating the Gaussian mean. Now, define our estimator as

$$\delta_n = \bar{X} \cdot \mathbb{1}(|\bar{X}| \geq n^{-1/4}).$$

Please show that δ_n is asymptotic unbiased and $\mathrm{AE}(\delta_n, g(\theta)) > 1$ for $\theta = 0$.

Answer. First notice that $I_{X_1}(\theta) = 1$, so we have the Cramér-Rao lower bound for the unbiased estimator of θ is

$$\frac{1}{nI_{X_1}(\theta)} = \frac{1}{n}, \quad \text{and} \quad \frac{1}{I_{X_1}(\theta)} = 1.$$

We separate the cases into $\theta = 0$ and $\theta \neq 0$ for further discussion.

(i) When $\theta = 0$, then by central limit theorem, $\sqrt{n}\bar{X} \overset{d}{\to} N(0, 1)$, hence

$$\mathbb{P}\big(\delta_n = 0\big) = \mathbb{P}\big(\sqrt{n}|\bar{X}| < n^{1/4}\big) \to 1,$$

or in other words, $\sqrt{n}(\delta_n - \theta) = \sqrt{n}\delta_n \overset{d}{\to} 0$. In this case, we obviously have $\mathrm{AE}(\delta_n, \theta) > 1$ and δ_n is asymptotic unbiased.

(ii) When $\theta \neq 0$, then by central limit theorem, $\sqrt{n}\big(\bar{X} - \theta\big) \overset{d}{\to} N(0, 1)$, hence

$$\mathbb{P}\big(\sqrt{n}|\bar{X}| < n^{1/4}\big) = \mathbb{P}\big(-n^{1/4} - \sqrt{n}\theta \leq \sqrt{n}\big(\bar{X} - \theta\big) < n^{1/4} - \sqrt{n}\theta\big) \to 0,$$

which implies that

$$\mathbb{1}(|\bar{X}| \geq n^{-1/4}) \overset{p}{\to} 1, \quad \text{and} \quad \sqrt{n}\theta \cdot \mathbb{1}(|\bar{X}| < n^{-1/4}) \overset{p}{\to} 0.$$

Therefore, by Slutsky's theorem, we have

$$\sqrt{n}(\delta_n - \theta) = \sqrt{n}(\bar{X} - \theta) \cdot \mathbb{1}(|\bar{X}| \geq n^{-1/4}) + \sqrt{n}\theta \cdot \mathbb{1}(|\bar{X}| < n^{-1/4}) \overset{d}{\to} N(0, 1).$$

Hence in this case, we also have δ_n been asymptotic unbiased.

Overall, we conclude that δ_n is asymptotic unbiased and $\mathrm{AE}(\delta_n, g(\theta)) > 1$ for $\theta = 0$, which is said to be **having a superefficiency point at $\theta = 0$**. □

* **Remark 4.4.** *Here, after we obtain the converge in distribution, we further need to apply the Skorokhod's representation theorem and use the uniform integrabity of $\sqrt{n}(\delta_n - \theta)$ to get the L_2 convergence, just in order to discuss the asymptotic unbiasedness. But as it falls outside the scope of this class, so we omit this part of the proof.*

Accordingly, one may wonder when can we make an assessment for a general estimator without worrying about the superefficiency. Luckily, the following theorem states,

Theorem 4.5 (♣ **Range of Asymptotic Efficiency**)**.** *Let* $X = \{X_1, \cdots, X_n\}$ *be a random sample drawn from* $f(x|\theta)$ *with* θ *being a real-valued parameter. If the regularity conditions* $\mho$ *listed in the supplement hold, then for any estimator* δ_n *satisfy* $\sqrt{n}(\delta_n - \theta) \xrightarrow{L_2} N(0, \nu_1^2(\theta))$*, we have*

$$0 < \mathrm{AE}(\delta_n, g(\theta)) \leq 1$$

except on a set of Lebesgue measure zero.

• *Example* 4.6. Consider $X = \{X_1, \cdots, X_n\}$ be a random sample drawn from $f(x|\theta)$ with θ being a real-valued parameter. Assume the condition for the CAN property of MLE holds here for $\hat{\theta}$, which is the MLE of θ, then $\mathrm{AE}(\hat{\theta}, \theta) = 1$.

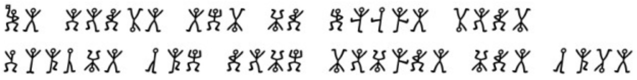
Answer. Apparently, the CAN property of MLE states that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, I_{X_1}(\theta)^{-1}\right),$$

so by the definition of asymptotic efficiency, we have $\mathrm{AE}(\hat{\theta}, \theta) = 1$. □

5. *Supplement

Here we list the conditions.



Condition 5.1 (Condition $\mho$ for Theorem.4.5).

(i) The parameter space Θ is an open interval (not necessarily finite).

(ii) The distribution family $\mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$ has common support, i.e., the set $A = \{x : f(x|\theta) > 0\}$ is independent of θ.

(iii) For every $x \in A$, the density $f(x|\theta)$ is twice differentiable with respect to θ, and the second derivative is continuous in θ.

(iv) The integral $\int f(x|\theta)dx$ can be twice differentiated under the integral sign, i.e.,

$$\frac{\partial^i}{\partial\theta^i}\int f(x|\theta)dx = \int \frac{\partial^i}{\partial\theta^i} f(x|\theta)dx, \quad \text{for } i = 1, 2.$$

(v) The Fisher information $0 < I_{X_1}(\theta) < \infty$ is well defined and not degenerate.

(vi) For any given $\theta_0 \in \Theta$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right| \le M(x), \quad \text{for } \forall x \in A, \text{ and } |\theta_0 - \theta| < c,$$

while $E_{\theta_0}\big(M(X)\big) < \infty$.

□

References

Hall, P. (2013). The bootstrap and Edgeworth expansion. Springer Science & Business Media.

Horowitz, J. L. (2001). The bootstrap. In Handbook of econometrics (Vol. 5, pp. 3159-3228). Elsevier.

Quenouille, M. H. (1949, July). Approximate tests of correlation in time-series 3. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 45, No. 3, pp. 483-484). Cambridge University Press.

Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika, 43(3/4), 353-360.

Tukey, J. (1958). Bias and confidence in not quite large samples. Ann. Math. Statist., 29, 614.