Week 6: Bootstrap

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-02-10

Recap Quiz:

Assume we have a random sample from the statistical model $X_1, \ldots, X_n \sim F_\psi$ that depends on the unknown parameter ψ. Suppose that $\mathbb{E}[X_1^2] = \psi/4 - 1$. Which of the following is a method of moments estimator $\hat{\psi}_{\text{MoM}}$ for ψ.

a. $\hat{\psi}_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^n X_i^2$.

b. $\hat{\psi}_{\text{MoM}} = 4\bar{X}_n + 1$.

c. $\hat{\psi}_{\text{MoM}} = \frac{4}{n}\left(\sum_{i=1}^n X_i^2 + 1\right)$

d. $\hat{\psi}_{\text{MoM}} = \frac{4}{n}\sum_{i=1}^n (X_i^2 + 1)$.

$\mathbb{E}[X_1^2] \longleftrightarrow \frac{1}{n}\sum_{i=1}^n X_i^2 = m_2$

$m_2 = \psi/4 - 1$

$4(m_2 + 1) = \hat{\psi}_{\text{MoM}}$

$\frac{4}{n}\sum_{i=1}^n X_i^2 + 4$

Recap Quiz:

$\mu_k = \mathbb{E}[X^k]$

$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, $f(\mu) = \theta$

$f(m_k) = \hat{\theta}_{MoM}$ this might be biased for θ

Consider a sample from statistical model

a. $\hat{\psi}_{\text{MoM}}$ is an unbiased estimator but it is not a consistent estimator.

b. $\hat{\psi}_{\text{MoM}}$ is a consistent estimator and an unbiased estimator. m_k as unbiased for μ_k

c. $\hat{\psi}_{\text{MoM}}$ is a consistent estimator but we cannot determine its bias without knowing $\mathbb{E}[X_1]$.

d. $\hat{\psi}_{\text{MoM}}$ is an unbiased estimator, but we cannot use the WLLN to determine its consistency since it only applies to functions of $\bar{X}_n$.

$\frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $Y_i = X_i^k$

Recap Quiz:

Which of the following is a known disadvantage of the Method of Moments (MoM) estimators compared to Maximum Likelihood Estimators (MLE) in large samples?

a. MoM estimators cannot be used for distributions with more than two parameters.

b. MoM equations are always non-linear and harder to solve than the likelihood equations.

c. MoM estimators are usually less efficient, meaning they have higher variance in large samples.

d. MoM estimators are always biased, while MLEs are always unbiased.

Example: Method of Moments

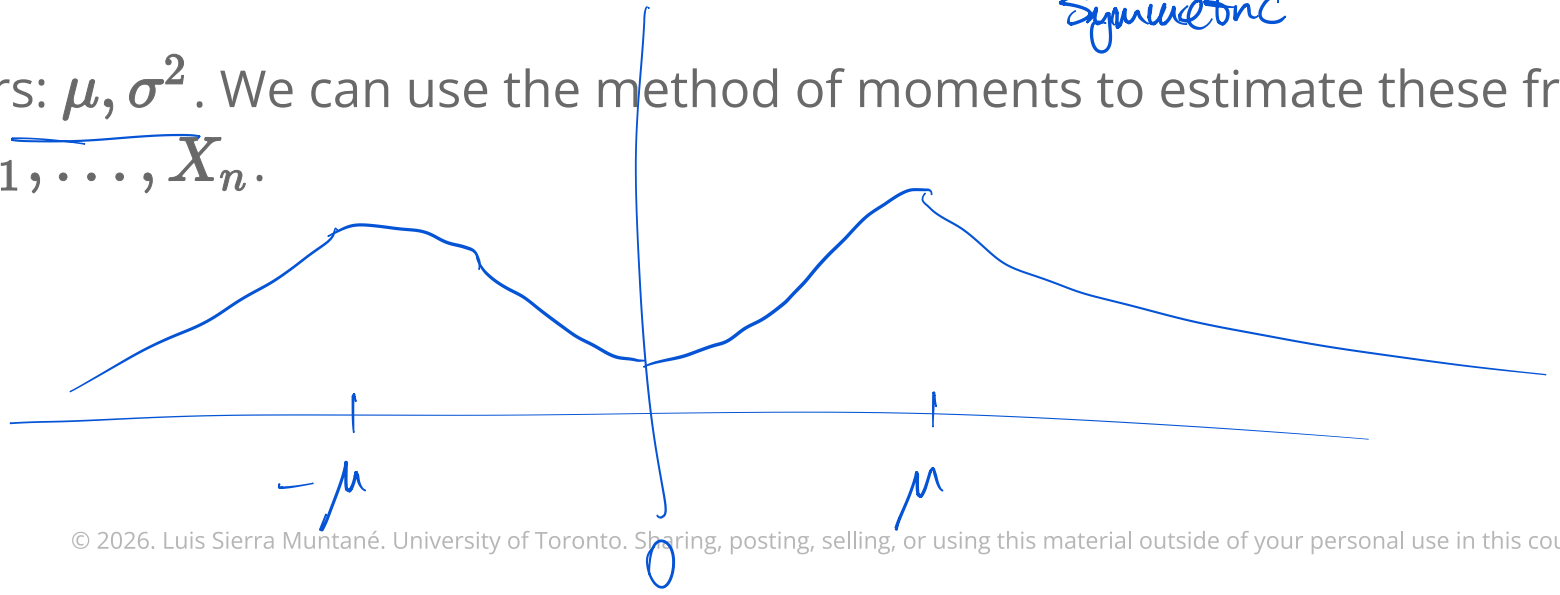
Karl Pearson went to the beach and recorded measurements of crabs, noticing that the distribution he was observing was not unimodal. As such, he chose to model the size of the crabs using a *normal mixture* to model the fact there were two different subpopulations. A simplified version of the model looks like:

$$f(x; \mu, \sigma^2) = \frac{1}{2}\mathcal{N}(\mu, \sigma^2) + \frac{1}{2}\mathcal{N}(-\mu, \sigma^2).$$

$\mu > 0$

Symmetric

Parameters: μ, σ^2. We can use the method of moments to estimate these from a sample $X_1, \ldots, X_n$.



Example: Method of Moments

The population moments are: $\text{Var}(X) = E[X^2] - E[X]^2$

- $\mu_2 = E[X^2] = \sigma^2 + \mu^2,$
- $\mu_4 = E[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$ Kurtosis ("heavy tailedness")

Let $M_2 = \frac{1}{n}\sum X_i^2$ and $M_4 = \frac{1}{n}\sum X_i^4$ be the sample moments. Plug in the values of the sample moments for the population moments:

- $\mu^2 + \sigma^2 = M_2,$ $\longrightarrow$ $\mu^2 = M_2 - \sigma^2$,
- $\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 = M_4.$ $\longrightarrow$ $(M_2 - \sigma^2)^2 + 6(M_2 - \sigma^2)\sigma^2 + 3\sigma^4 = M_4$

Solve for σ^2: $M_2^2 - 2M_2\sigma^2 + \sigma^4 + 6M_2\sigma^2 - 6\sigma^4 + 3\sigma^4 = M_4$

$-2\sigma^4 + 4M_2\sigma^2 + M_2^2 - M_4 = 0$, $\hat{\sigma}^2_{mom} = M_2 \pm \sqrt{\frac{3M_2^2 - M_4}{2}}$

take +

$\hat{\mu}_{mom} = M_2 - \hat{\sigma}^2_{mom} = \sqrt{\frac{3M_2^2 - M_4}{2}}$.

Recap: MLE and Method of Moments

The method of **MLE** works by maximizing the *likelihood function* which depends on the particular model we choose for our samples:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(X;\theta) \iff \hat{\theta} = \arg\max_{\theta} \ell(X;\theta).$$

The **Method of Moments** estimates the moments of the population without making strong distributional assumptions, but recall that for specific parameters we of course need to choose a distribution for our model.

In either case, we need access to aspects of the true underlying distribution.

Recap: Sampling Distribution

How accurate is our estimator $\hat{\theta}$ for θ? The parameter θ defines the distribution of our samples $X_1, \ldots, X_n \sim F_\theta$. Then, $\hat{\theta}$ will follow the *sampling distribution*.

$$\hat{\theta} \sim F_{\hat{\theta}}.$$

The standard deviation of the sampling distribution of a statistic is referred to as the *standard error* of the statistic

$$\text{SE} = \sigma_{\hat{\theta}}.$$

Example: estimating the sample mean

As we have seen many times in this course, from a random sample $X_1, \ldots, X_n \sim F$ the sample mean $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ will have mean $\mu_F = \mathbb{E}_F[X_i]$.

How accurate an estimator is $\bar{X}$ for μ_F? If F has a defined second moment $\mathbb{E}_F[X_i^2]$, then the *standard error* can be defined as the standard deviation of $\bar{X}$ under the sampling distribution of size n:

$$\sigma_{\bar{X}} = \sqrt{\mathbf{Var}_F(\bar{X})}.$$ → variance of sampling distribution

But note that in practice, we don't know $\mathbf{Var}_F(\bar{X})$!

Instead, we can use the *estimated standard error*. How do we estimate it?

Estimation considerations

The CLT applies for sums of iid random variables. We cannot apply it directly to:

- Correlation coefficient ρ
- Quotients $\bar{X}/\bar{Y}$
- Quantiles of a distribution q_p (including the median, or other quartiles)
- Trimmed mean, median of means
- The eigenvalues of a matrix

Therefore, we cannot find the sampling distributions for estimators of such quantities with the tools we have.

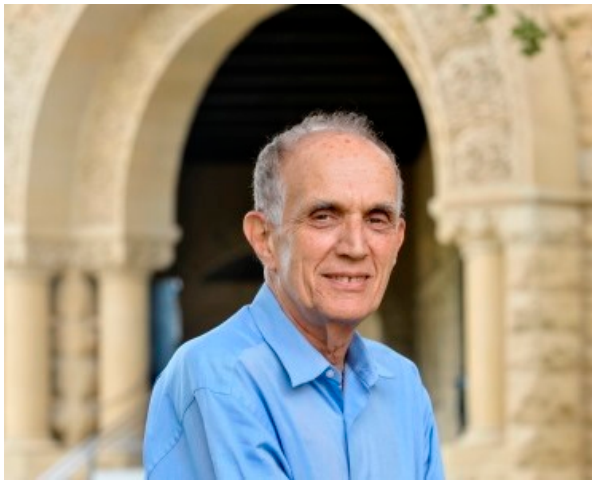
We also cannot find the standard error of estimators of these quantities.

The Bootstrap

Replace F by the eCDF $\hat{F}_n$ so that

$$\mathrm{SE} = \sqrt{\mathrm{Var}_{\hat{F}_n}(\hat{\theta})}.$$

Re-sample our data from our available samples.



Bradley Efron

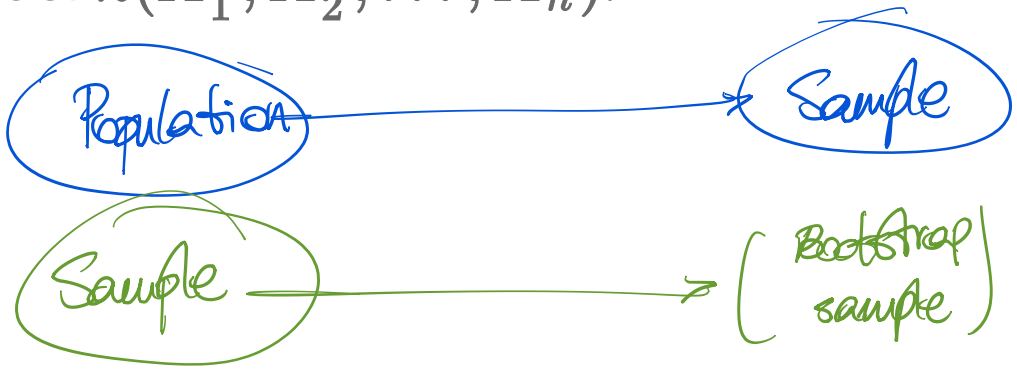
1979

Bootstrap principle

Bootstrap principle. Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}_n$ for the population distribution function F.

Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}_n$.

We can then approximate the probability distribution of the sample statistic $h(X_1, X_2, \ldots, X_n)$ by that of $h(X_1^*, X_2^*, \ldots, X_n^*)$.



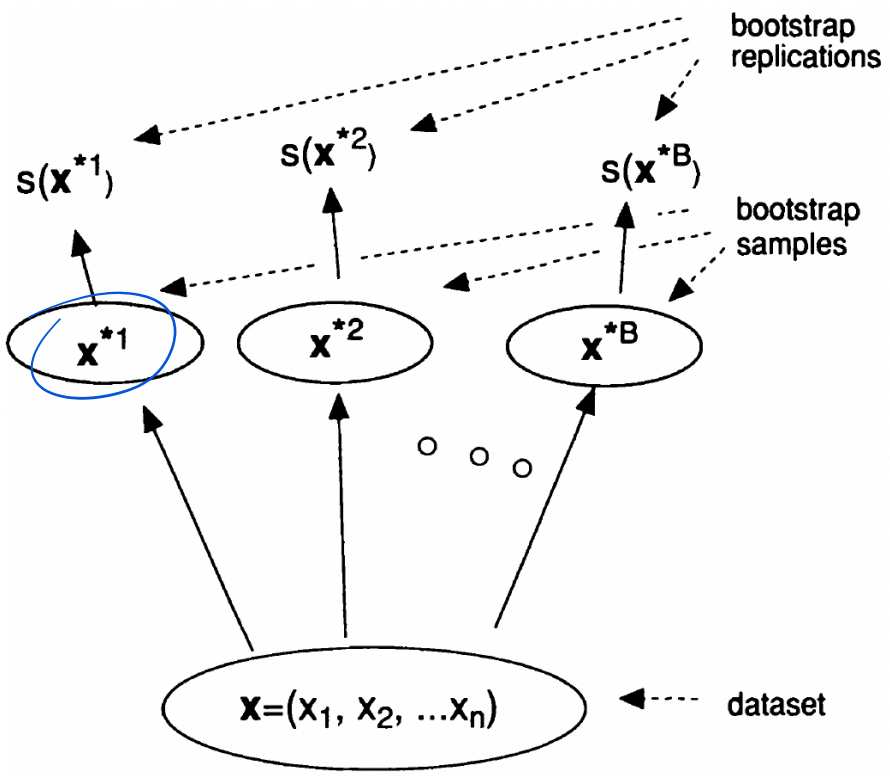
Bootstrap

X_3, X_1, X_3, X_4

A bootstrap sample $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ is obtained by sampling n times, *with replacement*, form the original sample $X_1, \ldots, X_n$.

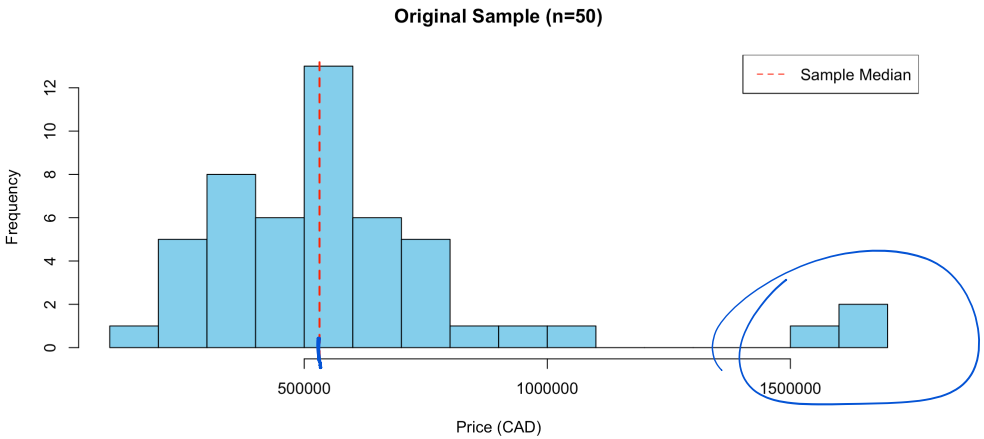
Repeat this B times and we essentially have B surrogate samples from the distribution $\hat{F}_n$.

$X_1^*, \ldots, X_n^* \sim \hat{F}_n$ eCDF

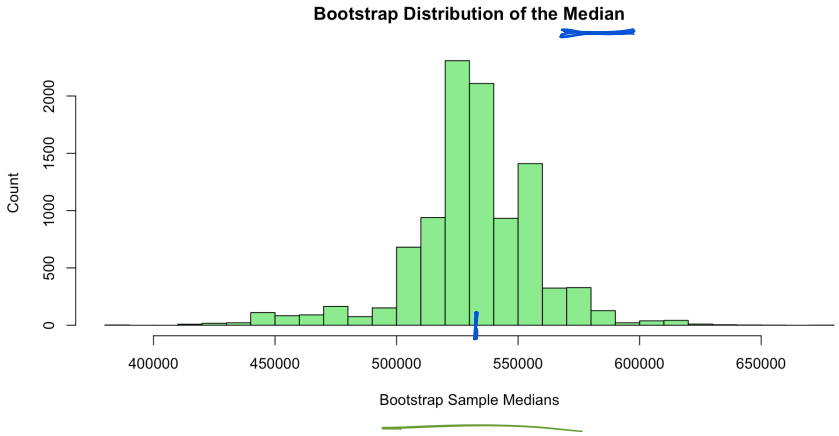


Bootstrap Example

Lognormal



$\text{Median}(X_1^*, \ldots, X_{50}^*)$



$X_1, X_2, X_3, \ldots, X_{49}, X_{50}$

Select randomly
Repeat 50 times → independent

Computationally
Expensive

Bootstrap intuition

The bootstrap works because the Empirical Distribution Function (eCDF) is a consistent estimator of the true Cumulative Distribution Function (CDF).

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty, X_i]}(x), \quad \hat{F}_n \xrightarrow{p} F,$$

$$\mathbb{E}_F[\hat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}_{(-\infty, X_i]}(x)] = \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x).$$

> So we can say that $\hat{F}_n$ is unbiased for F.

The WLLN can then help us establish the consistency of $\hat{F}_n$ for F.

Exercise: Bootstrap I

18.1 in MIPS: we generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_6^*$ from the empirical distribution function of the dataset

$$2 \quad 1 \quad 1 \quad 4 \quad 6 \quad 3,$$

→ 5 distinct values

i.e., we draw (with replacement) six values from these numbers with equal probability 1/6. How many different bootstrap datasets are possible? Are they all equally likely to occur?

5^6, taking $B > 5^6$ would result in definite repetitions.

accounting for different reorderings: $\binom{6+6-1}{6} = \binom{11}{6} = 462$.

Exercise: Bootstrap II

18.3 in MIPS: We generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_{10}^*$ from the empirical distribution function of the dataset

sample = {0.39, 0.41, 0.38, 0.44, 0.40, 0.36, 0.34, 0.46, 0.35, 0.37.}

pretend like this is a population

a. Compute the probability that the bootstrap dataset has exactly three elements equal to 0.35.

b. Compute the probability that the bootstrap dataset has at most two elements less than or equal to 0.38.

c. Compute the probability that the bootstrap dataset has exactly two elements less than or equal to 0.38 and all other elements greater than 0.42.

a) $P(X_1^* = 0.35) = \frac{1}{10}$. $\quad X_1^*, \underline{X_2^*}, \dots, \underline{X_9^*}, \underline{X_{10}^*}$

$$\binom{10}{3} p^3 (1-p)^7 = \frac{10!}{3!\,7!} \frac{1}{10^3} \left(\frac{9}{10}\right)^7$$

b) $P(X_i^* \le 0.38) = \frac{1}{2} = (1-p)$

$$P(|\{X_i^* : X_i^* \le 0.38\}| \le 2)$$

$$= p^{10} + 10\, p^9 (1-p) + \binom{10}{2} p^8 (1-p)^2$$

c) $P(X_i^* \le 0.38) = 1/2, \quad P(X_i^* \ge 0.42) = \frac{1}{5}$

$$\longrightarrow \quad X_1^*, \dots, X_{10}^*$$

Multinomial probability:

$$\binom{10}{2,0,8} p_1^2 p_2^8 = \frac{10!}{2!\,8!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{5}\right)^8.$$

Summary

- We have seen an example of how the Method of Moments requires some distributional assumptions to be applied.
- There are important limitations to distributional-based modelling.
- This motivates the use of the eCDF as our primary tool to obtain information.
- We introduced the concept of the bootstrap and seen how it can help in recovering information of our population