Exercise – Understanding the theory of bootstrap

1. Aim of the exercise

In statistical analysis, assessing the reliability of a computed statistic is crucial. Ideally, this would involve evaluating its behavior across repeated samples from the true population distribution. However, in most practical scenarios, the full population is inaccessible. To address this limitation, we approximate the statistic's sampling distribution using resampling techniques. One powerful method for doing so is the bootstrap, which enables us to estimate uncertainty and variability without relying on strong parametric assumptions. In this exercise, we review the theoretical foundations of the bootstrap method.

2. Theory

The sampling distribution of a statistic can be characterized by its probability density function or cumulative distribution function (CDF). Here we focus on the latter, as the CDF directly expresses probabilities of the form $P(X \leq x)$, which are central to statistical inference procedures such as standard error estimation, hypothesis testing and confidence interval construction.

Let X be a random variable and consider a random sample $X_1,\ X_2, ...,\ X_n$ drawn from the distribution of X. Let F denote the CDF of the random variable X:

$$F(x) \equiv P(X \leq x).$$

Let $T_n = T(X_1, \ldots, X_n)$ be a statistic computed using this sample. Under the true distribution F, denote the finite-sample CDF of this statistic by

$$G_n(\tau, F) \equiv P_F(T_n \leq \tau).$$

A CDF, here $G_n(\tau, F)$, is often unknown because F itself is rarely known, and even when F is specified, it may be difficult to compute or it may lack a closed-form expression.

Conventional asymptotic theory addresses this by approximating $G_n(\tau, F)$ with its limiting form $G_\infty(\tau, F)$, which becomes accurate as the sample size increases. However, since $G_\infty(\tau, F)$ also depends on F, this approximation remains challenging in practice. A common workaround is to substitute F with a consistent estimator, typically a parametric estimate $\hat{F} = F(\cdot, \hat{\theta})$, where $\hat{\theta}$ consistently estimates the true parameter θ_0. However, such plug-in approximations can perform poorly in finite samples, especially when model assumptions are violated or the sample size is small.

To overcome the limitations of conventional asymptotic methods, the bootstrap offers a fully data-driven approach for approximating $G_n(\tau, F)$ without requiring explicit knowledge of the population distribution F. Instead of relying on parametric assumptions or plug-in estimators, the bootstrap replaces F with the empirical distribution function F_n, which is constructed directly from the observed data. Specifically, F_n assigns to each value x the proportion of sample observations less than or equal to x, thereby capturing the structure of the data in a nonparametric way. By repeatedly resampling from F_n, the bootstrap generates an empirical approximation of the sampling distribution, offering a flexible and robust alternative to traditional asymptotic techniques-particularly in settings with small samples or complex estimators.

The Glivenko-Cantelli theorem guarantees that the empirical distribution function F_n, constructed from a sample of size n, converges uniformly to the true population distribution F

as $n \to \infty$. This result justifies using F_n as a consistent, nonparametric estimator of F in statistical procedures.

In the context of bootstrap theory, we are often interested in the sampling distribution of a statistic T_n, evaluated at some threshold τ. Under the true distribution F, we can characterize the sampling distribution by its CDF which we defined above as

$$G_n(\tau, F) = P_F(T_n \leq \tau).$$

However, since F is unknown, we approximate it using the empirical distribution F_n, leading to:

$$G_n(\tau, F_n) = P_{F_n}(T_n \leq \tau).$$

This expression represents the probability – under the empirical distribution – that the statistic T_n does not exceed τ, and serves as the foundation for bootstrap inference.

For each bootstrap sample, the statistic is computed as

$$T_n^b = T(X_1^b, \ldots, X_n^b),$$

where $X_1^b, \ldots, X_n^b$ are drawn with replacement from the original sample. This resampling procedure is repeated B times, producing a collection of bootstrap replicates $\{T_n^b\}_{b=1}^B$. Based on these replicates, the empirical bootstrap CDF is estimated by

$$\hat{G}_n(\tau, F_n) = \frac{1}{B} \sum_{b=1}^{B} I(T_n^b \leq \tau),$$

where $I(T_n^b \leq \tau)$ is the indicator function, equal to 1 if $T_n^b \leq \tau$, and 0 otherwise. This empirical CDF approximates the probability that the statistic falls below the threshold τ, based on the bootstrap distribution.

As the number of bootstrap samples $B \to \infty$, the Law of Large Numbers ensures that the empirical bootstrap CDF converges in probability to the true bootstrap CDF:

$$\hat{G}_n(\tau, F_n) \xrightarrow{p} G_n(\tau, F_n).$$

Furthermore, as the sample size $n \to \infty$, the Continuous Mapping Theorem implies that the true bootstrap CDF - based on the empirical distribution - converges in probability to the sampling distribution CDF under the true population distribution:

$$G_n(\tau, F_n) \xrightarrow{p} G_n(\tau, F).$$

These convergence results hold under the standard assumption of independent and identically distributed (i.i.d.) sampling from the underlying distribution. In practical terms, this means that for sufficiently large n and B, the bootstrap distribution provides a reliable approximation of the true sampling distribution of T_n, thereby ensuring that inference based on the bootstrap is asymptotically valid.