8.8 Bootstrapping

The Empirical Distribution Function

In many of the preceding examples, we investigated the sampling distribution of estimators, assuming known population distribution. But in many real-world scenarios, we won't know the exact shape of the population (and we may not even have insight into its approximate form).

For some estimators, we can use the Central Limit Theorem (or stronger results on regularity) to assume the sampling distribution is approximately Normal. But the sample size required for approximate Normalcy depends on the estimator and the underlying population. We'll now outline procedures to estimate the sampling distribution of an estimator, with minimal assumptions on the population.

Def: Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed values of a random sample $\mathbf{X}$, and for each $x \in \mathbb{R}$, define a function $F_n(t)$ to be the proportion of observed values less than or equal to x. That is,

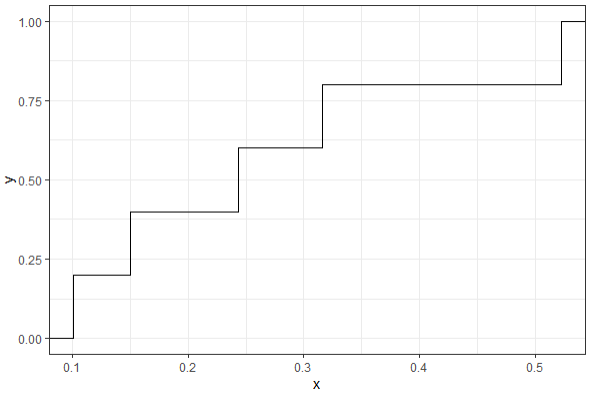
$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{x_i \leq t}$$

The function F_n is called the **sample (or empirical) distribution function**. It can be viewed as the CDF of a discrete distribution that assigns probability $\frac{1}{n}$ to each of the values $x_1, \ldots, x_n$.

Ex 1: Suppose $X_1, \ldots, X_5$ are drawn from an Expo(1) distribution, with

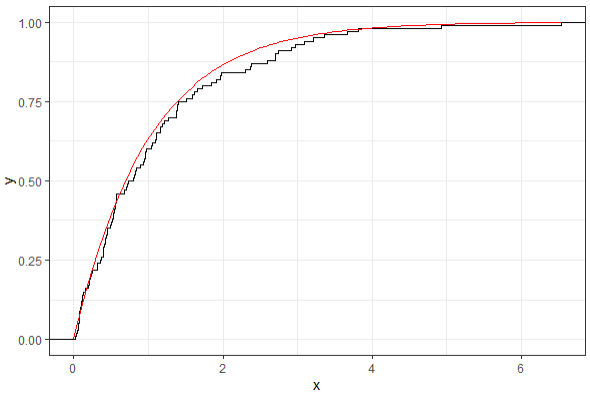
$$\mathbf{x} = (0.52, 0.32, 0.24, 0.15, 0.10)$$

Then the ECDF is



Note that each set of observed values gives rise to a different ECDF. In this way, we can think of the ECDF as a **random function**.

Additionally, observe that as $n \to \infty$, the ECDF converges toward the CDF.



Thm: Let F_n be the ECDF for a sample $\mathbf{X}$, where X_i has CDF F. Then $F_n(t) \to F(t)$ in probability.

Proof. Let $t \in \mathbb{R}$. For each i, let $I_{X_i \leq t}$ be the indicator variable that $X_i \leq t$. Then

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq t}$$

which converges in probability to $E[I_{X_i \leq t}] = P(X_i \leq t) = F(t)$, by the Law of Large Numbers. □

Ex 2: Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed values of a sample $\mathbf{X} = (X_1, \ldots, X_n)$, and let $F_n(t)$ be the corresponding ECDF. Let $\mathbf{X}^*$ be an iid sample with CDF F_n. How does $\mathbf{X}^*$ compare to $\mathbf{X}$?

Bootstrapping Using the ECDF

Using the preceding theorem, we know that the ECDF $\hat{F} = F_n$ is a consistent estimator for F. It also turns out that for each t, $F_n(t)$ is the MLE of $F(t)$. In other words, for relatively large sample sizes, the proportion of observations less than a given value is approximately the probability that a random observation is less than the given value. As such, we can estimate many things about F using F_n.

Suppose we have a random sample $\mathbf{X}$ from an unknown distribution F, and want to investigate a quantity that involves both parameters of F and $\mathbf{X}$. For example, suppose we want to estimate the variance of a statistic $g(\mathbf{X})$ as an estimator for the maximum value of F.

The main idea behind the **bootstrap** is to estimate an unknown distribution F using a known distribution $\hat{F}$ (which may be based on a sample $\mathbf{X}$ from F), then compute the quantity of interest using $\hat{F}$ and a sample $\mathbf{X}^*$ from $\hat{F}$.

Ex 3: Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ are a random sample from a population with unknown mean μ and variance σ^2. Assume that we know nothing more about the population distribution F other than that it has finite mean and variance.

Suppose we are interested in estimating the variance of the sample mean $\bar{X}$: $\theta = \text{Var}(\bar{X})$. In theory, the variance of the sample mean is $\frac{\sigma^2}{n}$. But we don't actually know σ^2, so we need to estimate it. **Why can't we just use the MLE estimator for σ^2?** (The MLE can only be computed with a known distribution F)

Let $\hat{F}$ be an estimate for F; in particular, the function $\hat{F}$ depends on the observed data $\mathbf{x}$. This CDF has some mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, both of which depend on $\mathbf{x}$. Let $\mathbf{X}^*$ be a random sample from $\hat{F}$. Then $\bar{X}^*$ has mean $\hat{\mu}$ and variance $\hat{\sigma}^2/n$. Since the distribution $\hat{F}$ is known, we should be able to calculate $\hat{\sigma}^2/n$, which can be used as an estimate for σ^2/n. **Will it be a good estimator?** (That depends on how we choose $\hat{F}$)

Suppose we have observed data $\mathbf{x} = x_1, \ldots, x_n$, and let

$$\bar{x} = \frac{1}{n} \sum x_i \qquad s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

1. If we take $\hat{F} = F_n$, then $X_i^* \sim \text{DUnif}(x_1, \ldots, x_n)$ and

$$\hat{\sigma}^2(\mathbf{x}) = \text{Var}(X_i^*|\mathbf{x}) = E[(X_i^* - E[X_i^*])^2|\mathbf{x}] = \frac{1}{n}\sum(x_i - \bar{x})^2 = s^2$$

   Then the variance of $\bar{X}^*$, given $\mathbf{x}$ is $\frac{s^2}{n}$, and so our estimate for $\theta = \frac{\sigma^2}{n}$ is $\hat{\theta} = \frac{s^2}{n}$.

2. If we take $\hat{F}$ be Normal and choose parameters equal to the MLE estimates from $\bar{X}$, then $\hat{F} \sim N(\bar{x}, s^2)$, and the variance of $\bar{X}^*$ given $\mathbf{x}$ is also $\frac{s^2}{n}$. Once again, our estimator for $\theta = \frac{\sigma^2}{n}$ is $\hat{\theta} = \frac{s^2}{n}$.

In the preceding example, it was relatively easy to calculate the quantity of interest. But in many cases, finding the desired quantity may be computationally intractable.

For example, if we want to calculate the 2.5% and 97.5% percentiles for the bootstrap distribution, we'll need an explicit formula for the joint PDF of $(X_1^*, X_2^*, \ldots, X_n^*)$ which requires n parameters (the values $x_1, \ldots, x_n$). And then an explicit formula for CDF for $\bar{X}^*$ which is discrete and requires approximately n^n parameters.

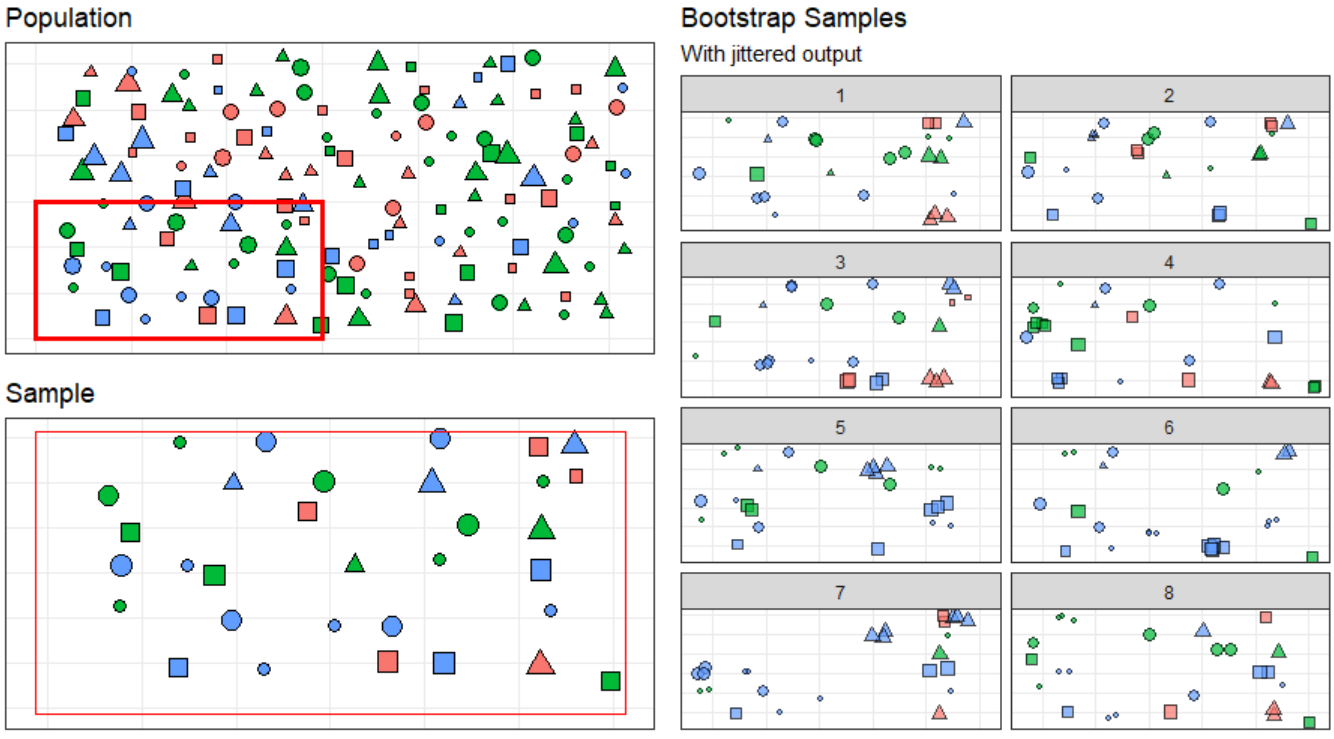
Thus, we approximate the bootstrap distribution using simulation.

The idea: We **resample** from the bootstrap distribution to represent new samples from the population.

Given a sample $\mathbf{X}$ of size n from F and a bootstrap estimator $\hat{F}$. Let B be a large number. To obtain a bootstrap approximation:

1. Draw bootstrap samples $\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(B)}$ from $\hat{F}$.
2. Calculate the statistic of interest $\hat{\theta}^{(i)}$ for each bootstrap sample $\mathbf{X}^{*(i)}$
3. Estimate features of the bootstrap distribution using the distribution of $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$.

The intuition for why this works:



What does the bootstrap distribution give us? It allows us to approximate the shape, spread, bias and symmetry of the actual sampling distribution. That is, it makes use of the 3rd and higher moments of the sample to make estimates. However, the bootstrap distribution **does not** provide accurate estimate of the center of the sampling distribution.

How accurate is the bootstrap?

1. How accurate is the theoretical bootstrap?
2. How accurate is the simulation-based approximation to the theoretical bootstrap?

Bootstrap distributions have two sources of random variation:

1. The original sample is chosen at random from the population
2. Bootstrap resamples are chosen at random from the original sample

Generally, we have control over the latter source of randomness. The number of bootstrap samples is only limited by computational resources. Brad Efron, the original inventor of the bootstrap technique, suggested that $B = 200$ (or even as few as $B = 25$) suffices for estimating standard error, and that $B = 1000$ is enough for confidence intervals.

But with widespread access to high speed computing, more resamples are often appropriate. For quick calculations, at least $B = 1000$ should be used. For routine, but robust analysis, at least $B = 10000$ should be used, and for calculations where accuracy is essential, use between $B = 15000$ and $B = 50000$.

Demo in RStudio.