The Bootstrap

Ugarte Chapter 10 (Section 9)
Other Notes on Canvas

Photo by Clay Banks on Unsplash

non-normal
σ-unknown
$n > 30$
$\Rightarrow$ Z-interval (CLT)

Example: Suppose wait times for a particular bank teller are exponentially distributed with rate parameter λ minutes. A random sample of $n = 100$ customer wait times were selected, $\bar{x} = 1.2$ and $s^2 = 2.1$. Compute an approximate 90% confidence interval (two-sided) for the population mean $\mu = 1/\lambda$.

$z_{\alpha/2} = z_{0.05} = 1.64$

90% CI : $(0.96, 1.44)$

Photo by Ben Weber on Unsplash

Recall that we derived confidence intervals and test statistics for several scenarios.

But! There were some special cases that we didn’t consider.

μ

	n >= 30	n < 30
Underlying normal distribution	σ known	σ known
	σ unknown	σ unknown
Underlying non-normal distribution	σ known	σ known
	σ unknown	σ unknown

Special Cases


Photo by Ben Weber on Unsplash

For “non-normal” cases, e.g., when $n < 30$ and the underlying distribution is unknown, we have to:

1. Make a specific assumption about the form of the population distribution and derive a CI based on that assumption.

2. Use other methods (such as **bootstrapping**) to make reasonable confidence intervals.

The bootstrap comes in two flavors:

1. *Parametric bootstrap*: We assume that the sample is a realization from a **known** distribution (e.g., exponential) but with **unknown** parameter(s) (e.g., rate λ).

2. *Nonparametric bootstrap*. We assume that the sample is a realization from an **unknown** distribution with **unknown** parameter values.

In class, we will focus on the nonparametric bootstrap. A homework question will walk you through the parametric bootstrap.


Photo by Ben Weber on Unsplash

The **bootstrap** is an alternative method for estimating standard errors of estimators, computing confidence intervals, conducting hypothesis tests, etc.

Let $f(\mathbf{x}; \theta)$ be an **unknown** probability distribution (density) for a given population. We want to estimate θ using $\hat{\theta}$. What is the distribution and standard error of $\hat{\theta}$?

In some cases, we know the distribution (e.g., by using the CLT, of $\hat{\theta} = \bar{X}$). But in other cases, we won't (it depends, in part, on the population distribution, $f(\mathbf{x}; \theta)$).


Photo by Ben Weber on Unsplash

We can estimate the distribution of our estimator $\hat{\theta}$ using the bootstrap!

1. First, we gather a **random sample** of size n from the unknown distribution $f(\mathbf{x}; \theta)$ and use it to find a point estimate $\hat{\theta}$ of θ.

$$\begin{bmatrix} 6 \\ 4 \\ 2 \\ 6 \\ 1 \end{bmatrix}$$

2. Then draw B (where B is large) new random samples of size n, **with replacement**, from the original sample (we can do this in R using the sample() function). We denote these new *bootstrap samples* as $\mathbf{x}_1^*, \ldots \mathbf{x}_B^*$. Each bootstrap sample comes from a distribution $\hat{f}$, which approximates f.

3. We can now compute $\hat{\theta}_j^*$, for $j = 1,...,B$.

4. Since $\hat{f}$ approximates f, the unknown distribution of $\hat{\theta}$, based on $\hat{f}$, is approximated by the distribution of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. This approximating distribution is used to estimate standard errors, CIs, etc.

your "real" data

$\hat{\theta}^*_1$ $\hat{\theta}^*_B$

$$f(\mathbf{X};\theta) \rightarrow \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \Longrightarrow \hat{f}(\mathbf{X};\theta) \rightarrow \begin{bmatrix} X^*_{1,1} \\ X^*_{2,1} \\ \vdots \\ X^*_{n,1} \end{bmatrix}, \ldots, \begin{bmatrix} X^*_{1,B} \\ X^*_{2,B} \\ \vdots \\ X^*_{n,B} \end{bmatrix}$$

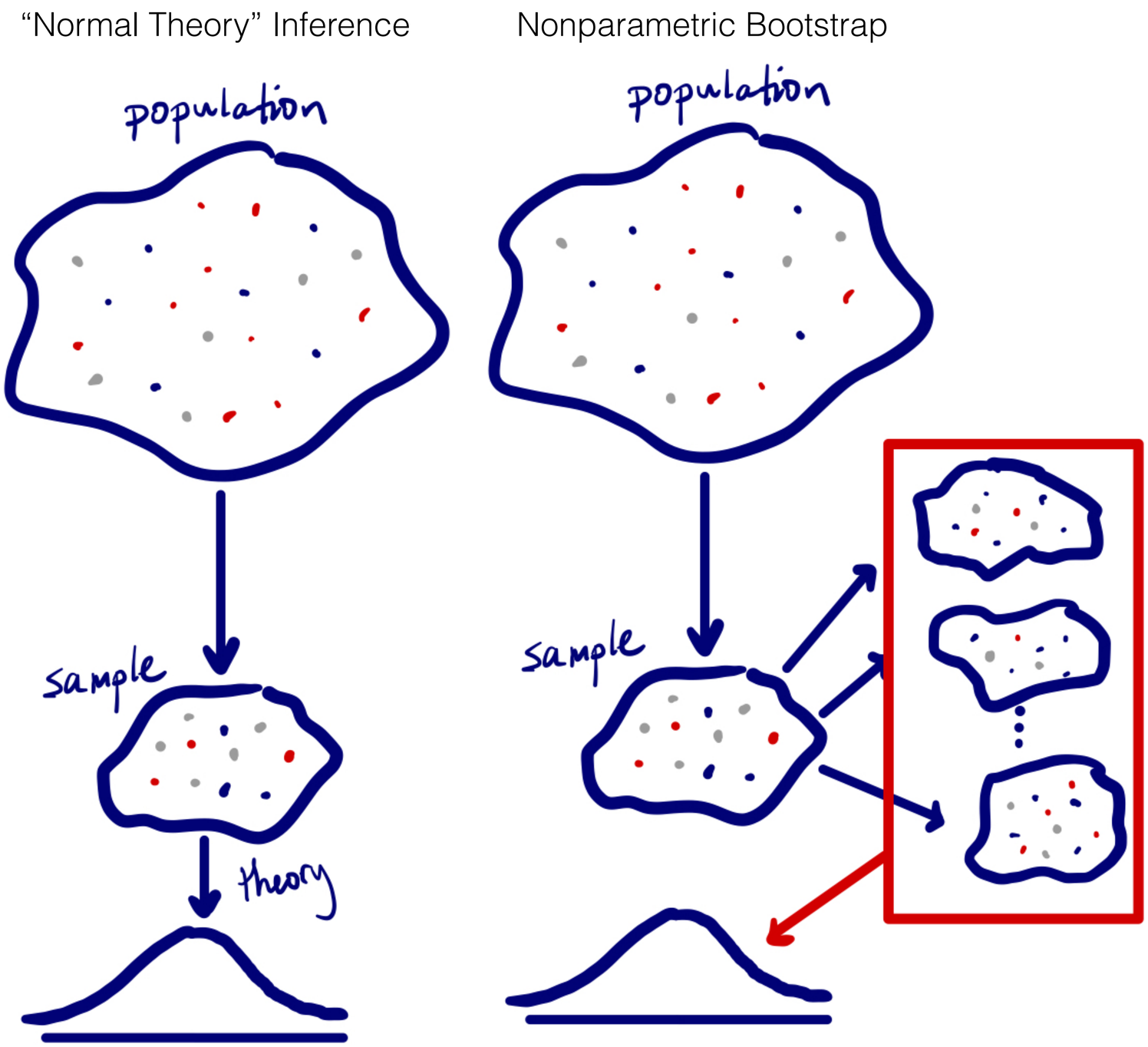
unknown pop. dist.

B of these

$\bar{X}^*_1$ = $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ is a (large) sample of size B from the (approximate) distribution of $\hat{\theta}$.

= $\bar{X}$


Photo by Ben Weber on Unsplash


"Normal Theory" Inference
Nonparametric Bootstrap
population
sample
theory
population
sample

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^* \qquad \text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Using the bootstrap samples, we can:

1. Estimate the variance or standard error of $\hat{\theta}$:

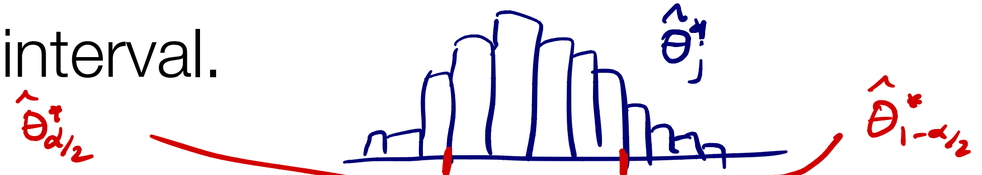
$$\widehat{Var}(\hat{\theta}) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{\theta}_j^* - \bar{\hat{\theta}}^*\right)^2, \quad \widehat{s.e.}(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$$

2. Estimate the bias of $\hat{\theta}$:

$$\widehat{Bias}(\hat{\theta}) = \underbrace{\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^*}_{\bar{\hat{\theta}}^*} - \hat{\theta}$$

$$\widehat{Bias}(\hat{\theta}) \approx Bias(\hat{\theta})$$

Now that we can estimate the variance (and thus standard error) of an estimator $\hat{\theta}$, we can construct bootstrap CIs. There are several types. We'll look at two: the percentile confidence interval and the pivot confidence interval.



The *bootstrap percentile confidence interval* is based on the quantiles of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. Let $\hat{\theta}^*_{\alpha/2}$ be the $\alpha/2$ quantile of the sample $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. Then the $(1-\alpha) \times 100\,\%$ confidence interval is:

$$\left(\hat{\theta}^*_{\alpha/2} \;,\; \hat{\theta}^*_{1-\alpha/2} \right)$$

quantile(thetaStar, $\alpha/2$)

$\hat{\theta}$: estimator from sample

$\hat{\theta}^*$: random values from est. bootstrap sampling dist.

$\hat{\theta}^*$ = estimator across b.s. samples

The **bootstrap pivot confidence interval** is another possibility. It assumes that $\theta - \hat{\theta}$ has roughly the same distribution as $\hat{\theta} - \hat{\theta}*$. Using this, let's derive the bounds of this CI. Again, let $\hat{\theta}^*_{\alpha/2}$ be the $\alpha/2$ quantile of the sample $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. Then:

Approx $(1-\alpha)\times 100$ CI:

$$\left(2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2},\ 2\hat{\theta} - \hat{\theta}^*_{\alpha/2}\right)$$