

Simulations in statistical inference

Recall the confidence interval for the population mean:

$$\bar{x} \pm z \text{SE}(\bar{x})$$

The idea was that \bar{x} follows approximately the normal curve.

What if we are interested in an estimator $\hat{\theta}$ for some parameter θ and the normal approximation is not valid for the estimator $\hat{\theta}$? What if there is no formula for $\text{SE}(\hat{\theta})$?

In such situations, simulations can often be used to estimate these quantities quite well. In fact, simulations may result in better estimates even in cases where the normal approximation is applicable!

The Monte Carlo Method

What is the average height of all people living in the United States?

This is difficult to determine exactly but can easily be estimated quite well:

Sample $n = 100$ (say) people at random. Then use the average height of these n people as an estimate of the average height of all people in the US.

This is an example of the general problem where we are interested in a unknown **parameter** θ of a population.

We estimate θ with a **statistic (estimator)** $\hat{\theta}$ which is based on a sample of n observations X_1, \dots, X_n drawn at random from the population:

$$\hat{\theta} = \text{average of the sample} = \frac{1}{n} \sum_{i=1}^n X_i$$

The Monte Carlo Method

$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$ tends to be close to the uncomputable population mean θ , even for moderate sample sizes such as $n = 100$.

This example is a special case of the **Monte Carlo Method** or **Simulation**:

- ▶ We approximate a fixed quantity θ by the average of independent random variables that have expected value θ .
- ▶ By the law of large numbers, the approximation error can be made arbitrarily small by using a large enough sample size.

The Monte Carlo Method

The Monte Carlo Method can also be used for more involved quantities. For example, we can use it to compute the standard error (SE) of a statistic $\hat{\theta}$.

Recall that the standard error tells roughly how far off the statistic will be from its expected value. The precise definition is

$$\text{SE}(\hat{\theta}) = \sqrt{E(\hat{\theta} - E(\hat{\theta}))^2}.$$

- ▶ Get many (say 1,000) samples of 100 observations each.
- ▶ Compute $\hat{\theta}$ for each sample, resulting in 1,000 estimates $\hat{\theta}_1, \dots, \hat{\theta}_{1000}$.
- ▶ Compute the standard deviation of these 1,000 estimates:

$$s(\hat{\theta}_1, \dots, \hat{\theta}_{1000}) = \sqrt{\frac{1}{999} \sum_{i=1}^{1000} (\hat{\theta}_i - \text{ave}(\hat{\theta}_i))^2}$$

Note that this is not an average of independent random variables. But it can be shown that the law of large numbers still applies and Monte Carlo works:

$$s(\hat{\theta}_1, \dots, \hat{\theta}_{1000}) \approx \text{SE}(\hat{\theta}).$$

We can use Monte Carlo only if we can draw many samples of size 100!

The Bootstrap principle

We have an estimate $\hat{\theta}$ for a parameter θ and want to know how accurate $\hat{\theta}$ is: we would like to find $SE(\hat{\theta})$ or give a confidence interval for θ .

The bootstrap can do this in quite general settings.

Example: θ = average height of all people in the US.

θ is unknown but can be estimated by the average height $\hat{\theta}$ of 100 randomly selected people.

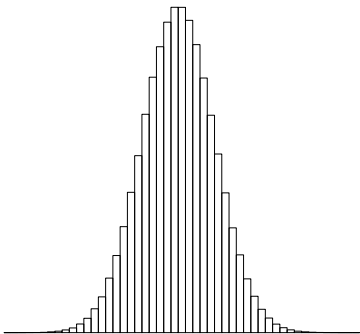
This illustrates the **plug-in principle**:

We can't compute the population mean because we can't access the whole population. So we 'plug in' the sample in place of the population and compute the mean of the sample instead.

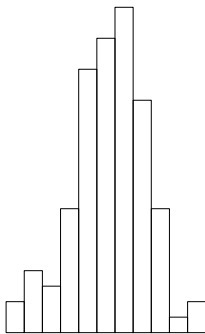
The Bootstrap principle

The rationale for the plug-in principle is that the sample mean $\hat{\theta}$ will be close to the population mean θ because the sample histogram is close to the population histogram.

Histogram of population



Histogram of sample



The bootstrap principle

The bootstrap uses the plug-in principle and the Monte Carlo Method to approximate quantities such as $SE(\hat{\theta})$.

Here is the reasoning behind the bootstrap:

Suppose we can draw as many samples from the population as we wish. Then we can approximate $SE(\hat{\theta})$ with Monte Carlo:

- ▶ Draw a sample X_1, \dots, X_n and use it to compute $\hat{\theta}$.
- ▶ Repeat B times (say $B=1,000$) to get $\hat{\theta}_1, \dots, \hat{\theta}_B$.
- ▶ The standard deviation of these B estimates is close to $SE(\hat{\theta})$ if B is large, by the law of large numbers.

However, we have only one sample X_1, \dots, X_n and we can't simulate more because the population is not accessible.

The bootstrap uses the plug-in principle to get around this: It simulates from the sample instead of from the population.

The bootstrap principle

The bootstrap pretends that the sample histogram is the population histogram and then uses Monte Carlo to simulate the quantity of interest.

Simulating a bootstrap sample X_1^*, \dots, X_n^* means that we draw n times with replacement from X_1, \dots, X_n .

The bootstrap consists of two steps:

- ▶ Draw B bootstrap samples and compute $\hat{\theta}^*$ for each bootstrap sample:

$$X_1^{*1}, \dots, X_n^{*1} \rightarrow \hat{\theta}_1^*$$

$$\vdots$$

$$X_1^{*B}, \dots, X_n^{*B} \rightarrow \hat{\theta}_B^*$$

- ▶ Use $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ to approximate the quantity of interest.

For example, we approximate $\text{SE}(\hat{\theta})$ by the standard deviation of $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.

More about the bootstrap

The **nonparametric bootstrap** simulates a bootstrap sample X_1^*, \dots, X_n^* by drawing with replacement from X_1, \dots, X_n .

Sometimes a parametric model is appropriate for the data, e.g. a normal distribution with unknown mean and standard deviation. Then one may be better off with the **parametric bootstrap**, which simulates the bootstrap samples from this model, using estimates for the unknown parameters.

So far, the bootstrap samples were drawn independently. If there is dependence in the data (time series), then this needs to be incorporated, e.g. with the **block bootstrap**.

Bootstrap confidence intervals

If the sampling distribution of $\hat{\theta}$ is approximately normal, then

$$\left[\hat{\theta} - z_{\alpha/2} \text{SE}(\hat{\theta}), \hat{\theta} + z_{\alpha/2} \text{SE}(\hat{\theta}) \right]$$

is an approximate $(1 - \alpha)$ -confidence interval for θ .

$\text{SE}(\hat{\theta})$ can be estimated by the bootstrap.

If $\hat{\theta}$ is far from normal, then we have to use the bootstrap to estimate the whole sampling distribution of $\hat{\theta}$, not just $\text{SE}(\hat{\theta})$.

The sampling distribution of $\hat{\theta}$ can be approximated by that of $\hat{\theta}^*$, which in turn can be approximated by the histogram of $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.

This gives the **bootstrap percentile interval**

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

where $\hat{\theta}_{(\alpha/2)}^*$ is the $\alpha/2$ percentile of the $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.

Bootstrap confidence intervals

An alternative to bootstrapping the distribution of $\hat{\theta}$ is to do so for $\hat{\theta} - \theta$.

The hope is that this approach is less sensitive to θ and therefore produces a more accurate confidence interval.

This results in the **bootstrap pivotal interval**

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

Bootstrapping for regression

We have data $(X_1, Y_1), \dots, (X_n, Y_n)$ from the simple linear regression model

$$Y_i = a + bX_i + e_i$$

From the data we can compute estimates \hat{a}, \hat{b} . How can we use the bootstrap to get standard errors and confidence intervals?

- ▶ Compute the residuals $\hat{e}_i = Y_i - \hat{a} - \hat{b}X_i$
- ▶ Resample from those residuals to get e_1^*, \dots, e_n^*
- ▶ Compute the bootstrapped responses $Y_i^* = \hat{a} + \hat{b}X_i + e_i^*$

This gives a bootstrap sample $(X_1, Y_1^*), \dots, (X_n, Y_n^*)$, from which we can estimate the parameters \hat{a}^* and \hat{b}^* in the usual way.