STAT 480 Statistical Computing Applications

Unit 5. Resampling Methods

Lecture 2. Bootstrap & Jackknife

Department of Statistics, Iowa State University

Spring 2019

History of Bootstrap

Bootstrap was introduced by Brad Efron in 1979 to construct confidence intervals for both simple and complex statistics from non-normal data using computer-intensive resampling methods.

- Begin with an observed sample of size n;
- Generate a simulated sample of size n by drawing observations from your observed sample independently and with replacement;
- Compute and save the statistic of interest;
- Repeat this process many times (e.g. 1,000);
- Treat the distribution of your estimated statistics of interest as an estimate of the population distribution of that statistic

Bootstrap in Statistics

- The bootstrap is first of all a way of finding the sampling distribution, at least approximately, from just one sample.
- In statistics, bootstrap is a modern, computer-intensive, general purpose approach to statistical inference, falling within a broader class of resampling methods.
- There are many different types of bootstrap procedure, the one we consider here is called the data re-sampling bootstrap.
- Here is the bootstrap procedure:
  - Step 1: Re-sampling
  - Step 2: Bootstrap distribution

Step 1: Resampling

- A sampling distribution is based on many random samples from the population.
- In place of many samples from the population, create many resamples by repeatedly sampling with replacement from this one random sample. Each resample is the same size as the original random sample.
- Sampling with replacement means that after we randomly draw an observation from the original sample we put it back before drawing the next observation. As a result, any number can be drawn more than once, or not at all.
- If we sampled without replacement, wed get the same set of numbers we started with, though in a different order.

Step 2: Bootstrap Distribution

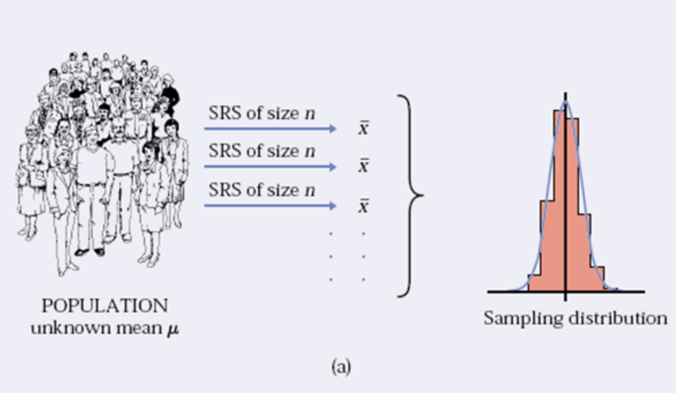
- The sampling distribution of a statistic collects the values of the statistic from many samples.
- The bootstrap distribution of a statistic collects its values from many resamples.
- The bootstrap distribution gives information about the sampling distribution.

Two Kinds of Distributions

Sampling Distribution vs. Bootstrap Distribution

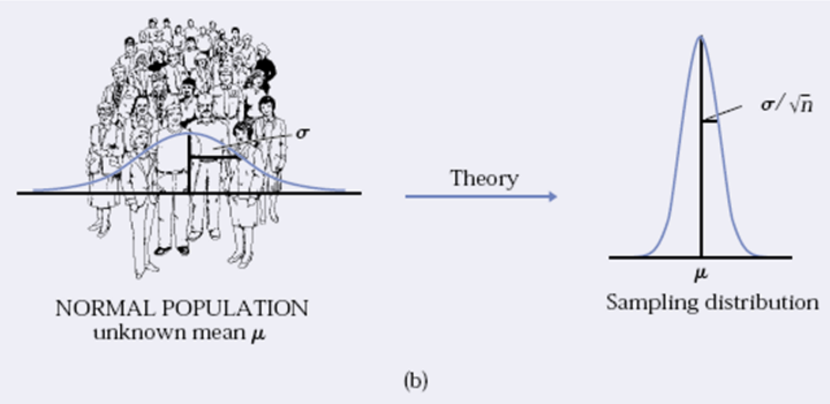
- The population: certain unknown quantities of interest (e.g., mean).
- Multiple sample $\Rightarrow$ sampling distribution
- Bootstrapping:
  - One original sample $\Rightarrow$ B bootstrap samples
  - B bootstrap samples $\Rightarrow$ bootstrap distribution

Sampling Distribution



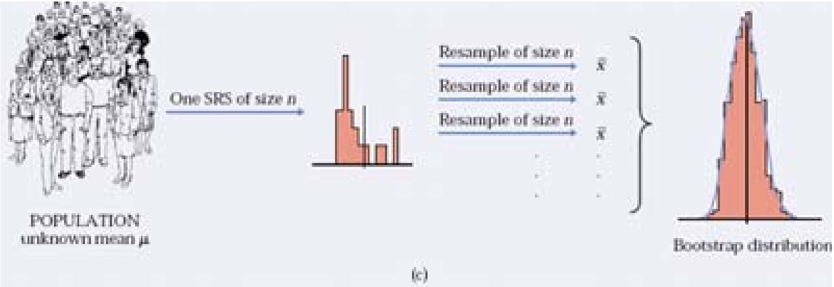
The idea of the sampling distribution of the sample mean $\bar{x}$: take very many samples, collect the $\bar{x}$-values from each, and look at the distribution of these values.

Normality



If we know that the population values follow a normal distribution, theory tells us that the sampling distribution of $\bar{x}$ is also normal.

Resampling & Normality



The bootstrap idea: when theory fails and we can afford only one sample, that sample stands in for the population, and the distribution of x in many resamples stands in for the sampling distribution.

Two Kinds of Distributions

Sampling Distribution vs. Bootstrap Distribution

- Bootstrap distributions usually approximate the shape, spread, and bias of the actual sampling distribution.
- Bootstrap distributions are centered at the value of the statistic from the original sample plus any bias.
- The sampling distribution is centered at the value of the parameter in the population, plus any bias.

Pros and Cons of Bootstrap

- Advantage of bootstrap over analytical methods
  - Great simplicity.
  - Straightforward to apply.
  - Under some conditions, it is asymptotically consistent.
- Cases where bootstrap does not apply
  - Small data sets: the original sample is not a good approximation of the population
  - Dirty data: outliers add variability in our estimates.
  - Dependence structures (e.g., time series, spatial problems): Bootstrap is based on the assumption of independence.

Jackknife Method

- Jackknife, which is similar to bootstrap, is used in statistical inference to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it.
- History: invented in 1958 by the statistician John Tukey "a boy scouts jackknife is symbolic of a rough and ready instrument capable of being utilized in all contingencies and emergencies".

Idea: Jackknife Method

- Systematically recompute the statistic estimate leaving out one observation at a time from the sample set.
- From this new set of “observations” for the statistic, compute, for example, the estimate for the variance of the statistic.
- NOTE: You can leave out groups rather than individual observations if the sampling/data structure is complex (e.g. clustered data).

Jackknife Variance

- Let $\hat{\theta}$ be an estimator of θ based on $x = (x_1, \cdots, x_n)$.
- For $i = 1, \ldots, n$
  1. generate a jackknife sample $x_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ by leaving out the ith observation.
  2. Calculate $\hat{\theta}_{-i}$ by applying the estimation process to the jackknife sample.
- Calculate the jackknifed estimate

$$\hat{\theta}_* = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i}$$

and the jackknife estimate of variance

$$\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{-i} - \hat{\theta}_*)^2$$

Bootstrap vs. Jackknife

- Both methods estimate the variability of a statistic from the variability of that statistic between subsamples, rather than from parametric assumptions.
- Jackknife is less general than the bootstrap, and explores the sample variation differently.
- Jackknife does not perform well if the statistic under consideration does not change "smoothly" across simulated samples.
- Jackknife does not perform well in small samples because you dont end up generating many resamples.
- Jackknife is easier to apply to complex sampling schemes, such as multi-stage sampling with varying sampling weights, than the bootstrap.

Bootstrap vs. Jackknife

- However, Jackknife is good at detecting outliers/infuential cases. Those sub-sample estimates that differ most from the rest indicate those cases that has the most influence on those estimates in the original full sample analysis.
- Jackknife and bootstrap may in many situations yield similar results. But when used to estimate the variance of a statistic, bootstrap gives slightly different results when repeated on the same data, whereas the jackknife gives exactly the same result each time.