STAT 641: Bootstrapping Methods

Jiyoun Myung

Spring 2021, Day 1

Intro to Bootstrapping

Department of Statistics and Biostatistics, CSU, East Bay

Introduction to Bootstrapping

Things to know/recall for the course

1. Population/Parameter/Sample/Statistic
2. Normal distribution
3. MLE
4. Sampling distribution of statistics
5. Standard error
6. Sampling in R
7. Creating functions in R

Three Basic Questions in Statistical Theory

Statistical theory attempts to answer three basic questions:

1. How should I collect my data? (Data Collection)
2. How should I analyze and summarize the data that I've collected? (Summary)
3. How accurate are my data summaries? (Statistical Inference)

What is Statistics?
Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation, and presentation of data. Statistics deals with all aspects of data including the planning of data collection in terms of the design of experiments.
(From Wikipedia: http://en.wikipedia.org/wiki/Statistics)

The Question

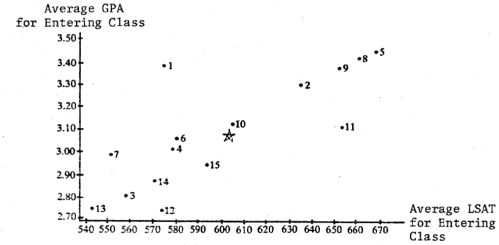
Suppose we are interested in the estimation of an unknown parameter θ. How do we answer these two questions?

1. What estimator of θ should be used?
2. Having chosen an estimator, how accurate is it?

Example

From *"Computer Intensive Methods In Statistics"* by Diaconis and Efron (1983)

We have data describing the 1973 class of 15 American Law schools. The plot shows the average LSAT score on the x-axis and average undergraduate GPA on the y-axis.



Example (Continued..)

The (sample) correlation coefficient was $r = 0.776$ implying strong positive correlation.

Note: r is an estimated correlation coefficient, based on a random sample of 15 law schools and not the true population correlation for the entire population of law schools.

Questions for you:

How accurate is this estimate of $r = 0.776$?
The sample size is big enough?
What would we do ideally?

Bootstrapping

- The key idea is to perform computations on the data itself to estimate the variation of statistics that are themselves computed from the same data. That is, the data is ‘pulling itself up by its own bootstrap.’ (In Germany one calls the bootstrap method “die Munchhausen Methode,” named after Baron von Munchhausen, a fictional character who is supposed to have saved his life by pulling himself out of a swamp by holding on to the straps on his boots).
- The empirical bootstrap is a statistical technique popularized by Bradley Efron in 1979. Though remarkably simple to implement, the bootstrap would not be feasible without modern computing power.
- Such techniques existed before 1979, but Efron widened their applicability and demonstrated how to implement the bootstrap effectively using computers. He also coined the term ‘bootstrap’

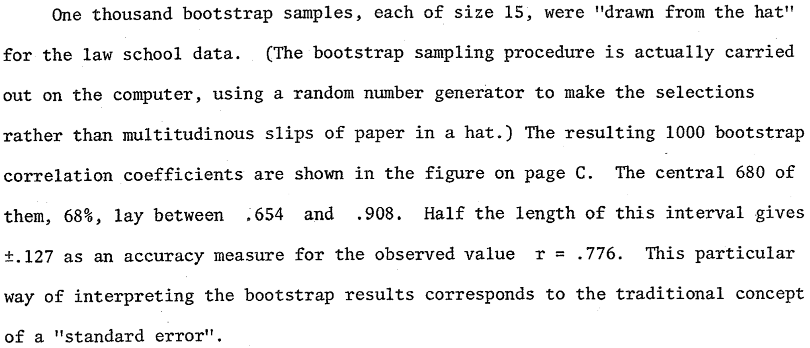
Bootstrapping

- To understand bootstrap, suppose it were possible to draw repeated samples (of the same size) from the population of interest, a large number of times. Then, one would get a fairly good idea about the sampling distribution of a particular statistic from the collection of its values arising from these repeated samples.
    - But, that does not make sense as it would be too expensive and defeat the purpose of a sample study. The purpose of a sample study is to gather information cheaply in a timely fashion

Bootstrapping

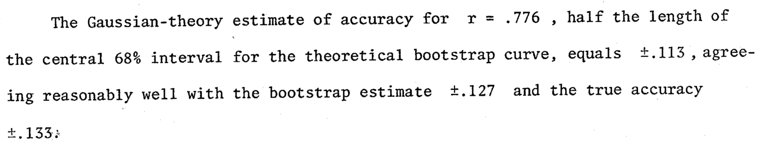
- The idea behind bootstrap is to use the data of a sample study at hand as a “surrogate population”, for the purpose of approximating the sampling distribution of a statistic; i.e. to **resample (with replacement)** from the sample data at hand and create a large number of “phantom samples” known as **bootstrap samples.**
- The sample summary is then computed on each of the bootstrap samples (usually a few thousand). A histogram of the set of these computed values is referred to as the **bootstrap distribution** of the statistic.

Example (Continued..)



One thousand bootstrap samples, each of size 15, were "drawn from the hat" for the law school data. (The bootstrap sampling procedure is actually carried out on the computer, using a random number generator to make the selections rather than multitudinous slips of paper in a hat.) The resulting 1000 bootstrap correlation coefficients are shown in the figure on page C. The central 680 of them, 68%, lay between .654 and .908. Half the length of this interval gives ±.127 as an accuracy measure for the observed value r = .776. This particular way of interpreting the bootstrap results corresponds to the traditional concept of a "standard error".

Example (Continued..)



The Gaussian-theory estimate of accuracy for r = .776 , half the length of the central 68% interval for the theoretical bootstrap curve, equals ±.113 , agreeing reasonably well with the bootstrap estimate ±.127 and the true accuracy ±.133.

Drawbacks of parametric assumptions

- No easy way to check the assumption that the 15 points are from a Normal population.
- Here working with correlation coefficient is okay but what is the statistics of interest is something non-traditional. For examples see Diaconis and Efron (1983)

Bootstrapping method?

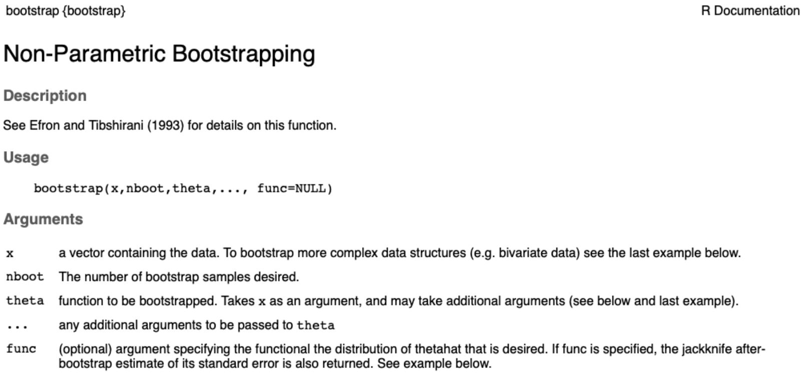
- Requires no theoretical calculations.
- Not based on asymptotic results.
- Available no matter how complicated the estimator $\hat{\theta}$ is.

PAUSE ... 15 minutes

- Read the Diaconis and Efron (1983) paper.
- Highlight or make memos anything that you have questions about or appears interesting.

The BOOTSTRAP PACKAGE

Install the **bootsrap** package in R.



bootstrap {bootstrap}   R Documentation

Non-Parametric Bootstrapping

Description

See Efron and Tibshirani (1993) for details on this function.

Usage

```
    bootstrap(x,nboot,theta,..., func=NULL)
```

Arguments

`x` a vector containing the data. To bootstrap more complex data structures (e.g. bivariate data) see the last example below.

`nboot` The number of bootstrap samples desired.

`theta` function to be bootstrapped. Takes `x` as an argument, and may take additional arguments (see below and last example).

`...` any additional arguments to be passed to `theta`

`func` (optional) argument specifying the functional the distribution of thetahat that is desired. If func is specified, the jackknife after-bootstrap estimate of its standard error is also returned. See example below.

Examples

```
# 100 bootstraps of the sample mean
# (this is for illustration;  since "mean" is  a
# built in function, bootstrap(x,100,mean)
#  would be simpler!)
    x <- rnorm(20)
    theta <- function(x){mean(x)}

    results <- bootstrap(x,100,theta)
```

Examples

```
# To bootstrap functions of more complex data structures,
# write theta so that its argument x
#  is the set of observation numbers
#  and simply  pass as data to bootstrap the vector 1,2,..n.
# For example, to bootstrap
# the correlation coefficient from a set of 15 data pairs:

   xdata <- matrix(rnorm(30),ncol=2)
   n <- 15
   theta <- function(x,xdata){ cor(xdata[x,1],xdata[x,2]) }
   results <- bootstrap(1:n,20,theta,xdata)
```

Homework

- Re-read the Diaconis and Efron paper.
- Read Chapter1 to Chapter 4.

See you next week!