Bootstrap – resampling from the data

August 29, 2019

Motto

> ”It is conjectured that Mr. Murphee will now be enabled to hand himself over the Cumberland river or a barn yard fence by the straps of his boots.”

Workingman’s Advocate (1834)

There is no more in data, than the data

If one had unlimited access to the data, no statistical inference would be needed.

- The bootstrap is a general tool for assessing statistical accuracy by ‘creating’ data from the data.
- It is based on sampling randomly from data to study how a quantity of interest behaves when observed in this process
- There are mathematical (mostly asymptotic – large sample size) results that justify using this form of data analysis
- It is a simple form of data mining, since it samples indiscriminately from the data to discover some properties
- Most often it is used to assess the variability of a certain characteristics

General bootstrap scheme

- Let $\mathbf{Z} = (z_1, \ldots, z_N)$ be a certain data set
- Randomly draw 'new' datasets with replacement from $\mathbf{Z}$
- Each new sample has the same size as the original set
- This is done B times ($B = 100$ say), producing B **bootstrap datasets**

$$\mathbf{Z}_1^*, \ldots, \mathbf{Z}_B^*$$

- $S(\mathbf{Z})$ is any quantity computed from the data $\mathbf{Z}$ For example, it can be an estimator, or a prediction at some time point, etc.
- From the bootstrap sampling we can estimate any aspect of the distribution of $S(\mathbf{Z})$ by taking its equivalent in the bootstrap samples $S(\mathbf{Z}_1^*), \ldots, S(\mathbf{Z}_B^*)$
- This can be, for example, the variance of $S(\mathbf{Z})$ taken as the **bootstrap sample variance**

Memory flashes from the past: What is sample variance?

Estimating variance using bootstrap

- Our interest is in assessing the variance of $S(\mathbf{Z})$, i.e. $\text{Var}[S(\mathbf{Z})]$
- **Bootstrap estimate of the variance** is given by

$$\widehat{\text{Var}}[S(\mathbf{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}(S(\mathbf{Z}^{*b}) - \bar{S}^{*})^2,$$

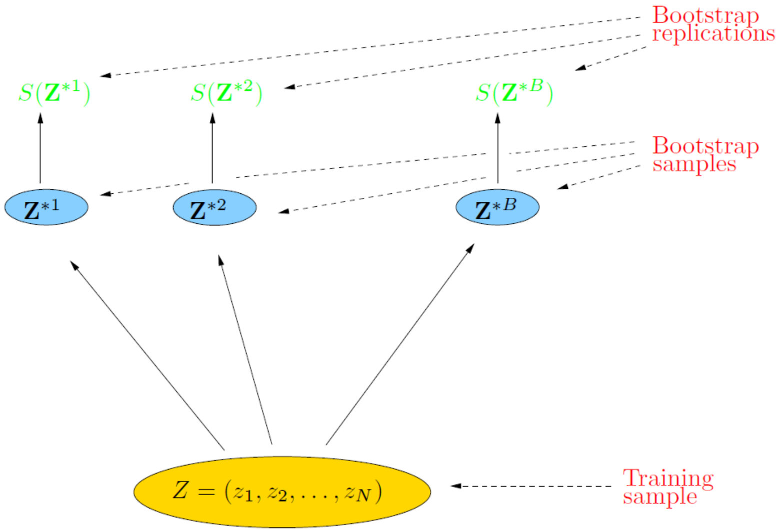
where $\bar{S}^{*}$ is the **bootstrap sample mean**, i.e.

$$\bar{S}^{*} = \frac{1}{B}\sum_{b=1}^{B} S(\mathbf{Z}^{*b}).$$

- We note that the bootstrap sample size B is only limited by our computational power.
- We call the method a non-parametric bootstrap because it does not assume any parametric model about the data – the scheme does not require any model specification.

Memory flashes from the past: Variance $Var(X)$ of a random variable X is...

Schematic illustration of bootstrapping



How bootstrap works? – calibrating a measurement device

A small numerical experiment is using measurements of concentration of a certain chemical used in the chemical device calibration experiment that are given in `Table2_1.txt`.

```
x=scan("Table2_1.txt")
n=length(x)
mean(x)
sd(x)
```

Questions of interest:

- What would be a **good estimate of the concentration**?
- What is the **standard deviation of such an estimate** (accuracy of estimation)?
- What could be an **estimate of the variance of the concentration measurements**?

Memory flashes from the past: What is standard deviation?

Bootstraping means – example continued

- Estimating standard deviation

  `sd(x)/sqrt(n)`

- Bootstraping means

```
help(sample)
B=100
Bmean=vector('numeric',100)
for(i in 1:B)
{
Bmean[i]=mean(sample(x,n,rep=T))
}
sd(Bmean)
```

- Compare the two obtained values. Conclusions?

Bootstraping variances – example continued

- Bootstraping variances
  - What is the estimate of the variance or standard deviation of the concentration measurements?
  - What is the standard error of this estimate?

```
B=1000
Bvar=vector('numeric',B)
Bsd=Bvar
for(i in 1:B)
{
Bvar[i]=var(sample(x,n,rep=T))
Bsd[i]=sqrt(Bvar[i])
}
sd(Bvar)

sd(Bsd)
```

Is bootstrap working?

- There are mathematical results showing that asymptotically (large sample size) bootstrap is working.
- One can also examine the bootstrap by the **Monte Carlo method**

```
#Data
n=100
x=rnorm(n,5,10)
```

- ```
  #Bootstrap
  B=10000
  Bvar=vector('numeric',B)
  for(i in 1:B){
  Bvar[i]=var(sample(x,n,rep=T))}
  mean(Bvar)
  sd(Bvar)
  hist(Bvar,nclass=10)
  ```

- ```
  #Monte Carlo
  N=15000
  MCvar=vector('numeric',N)
  for(i in 1:N){MCx=rnorm(n,5,10)
  MCvar[i]=var(MCx)}
  mean(MCvar)
  sd(MCvar)
  quartz() #Mac graphical window
           #windows() in Windows
           #X11() in Linux
  hist(MCvar,nclass=10)
  ```

The Monte Carlo method

- The Monte Carlo method is a technique to study models based on the probability theory by **simulating independent random data from the model** and statistically analyze the obtained data.
- Bootstrap resembles the Monte Carlo method.
- The difference is that the Monte Carlo simulates independently **from the model**, while the bootstrap samples independently **from the data**.
- The combination of the Monte Carlo with estimation from the data is often called **the parametric bootstrap**:
    - first, estimate parameters
    - then do Monte Carlo on the model with the estimated parameters.

Why bootstrap works?

Here a heuristic argument for the bootstrap

- Probabilistic modeling assumes that $\mathbf{Z}$ is drawn from a certain distribution, say F.
- Bootstrap samples $\mathbf{Z}^{*b}$'s are drawn from the empirical distribution F_n (based on $\mathbf{Z}$).
- Probabilistic modeling makes sense only because the empirical distribution F_n approximates the true one F.
- Thus sampling from F_n (bootstrap) should have similar properties as sampling from the true distribution F.
- In short, bootstrap is a method that subscribes under the directive:

  "There is no more in data than the data themselves"

Bootstrap and Estimation

Bootstrap is a data mining technique

Estimation of the parameters based on the original data only belongs to the classical theory of statistics

Parametric bootstrap combines the classical theory based on a model with data mining through independent sampling from the estimated model