Sane's STATS Academy of Statistics

Bootstrapping

An Introduction with Example

What is Bootstrapping?

- Bootstrapping is a method which relies on random sampling with replacement
- An Example:
  - We find multiple estimates on the various samples with replacement in this technique
  - Consider a set of some 10 values namely, 7, 23, 45, 67, 42, 56, 90, 12, 58, 79. We can find the mean of these values.
  - In bootstrapping we consider a sample with replacement from the same sample with same size. This method is called resampling.

Sane's STATS Academy of Statistics

Resampling

- As seen below, we have generated 5 resamples for the same sample

```
> vect <- c(7, 23, 45, 67, 42, 56, 90, 12, 58, 79)
> resample1 <- sample(vect , size = 10 , replace=TRUE)
> resample1
 [1] 45 79 58 42  7 23 56 79 56 42
> resample2 <- sample(vect , size = 10 , replace=TRUE)
> resample2
 [1] 67 12 23 79 56 79 42 42 67 23
> resample3 <- sample(vect , size = 10 , replace=TRUE)
> resample3
 [1]  7 58 12 79 56 23 56 67 90 58
> resample4 <- sample(vect , size = 10 , replace=TRUE)
> resample4
 [1] 67  7 45 45 56 12 45 12  7  7
> resample5 <- sample(vect , size = 10 , replace=TRUE)
> resample5
 [1] 23  7 12  7 42 67 58 90 79 67
```


Sane's STATS Academy of Statistics

Calculation of 95% CI by t-test formula

```
> vect <- c(7, 23, 45, 67, 42, 56, 90, 12, 58, 79)
> mean(vect)
[1] 47.9
>
> tt <- t.test(vect)
> tt$conf.int
[1] 28.09711 67.70289
attr(,"conf.level")
[1] 0.95
```

- By t-test formula, 95% CI is (28.1, 67.7)

Sane's STATS Academy of Statistics

Bootstrapping Example

- Suppose that we want to calculate the 95% confidence interval for population mean.
- We have got the sample of size 10.
- The following are steps we can perform for bootstrapping:
  1. We draw a resample from the sample
  2. For the resample, we compute the mean
  3. We repeat steps 1 and 2 for about 1000 (say) times
  4. Hence , now we have 1000 different means calculated
  5. We take 2.5^{th} Percentile as the lower confidence limit and 97.5^{th} percentile as upper confidence limit

Sane's STATS Academy of Statistics

Illustration

```
> vect <- c(7, 23, 45, 67, 42, 56, 90, 12, 58, 79)
> mean(vect)
[1] 47.9
>
> resamples <- array(dim=c(1000,10))
>
> for(i in 1:1000){
+   resamples[i,] <- sample(vect, size = 10 , replace = T)
+ }
>
> remeans <- apply(resamples, 1 , mean)
> Conf.Limits <- quantile(remeans, c(0.025,0.975))
> Conf.Limits
   2.5%   97.5%
31.1950 64.2025
```

- Hence, we find that 95% CI by bootstrap method, is (31.2, 64.2)

Sane's STATS Academy of Statistics

Bootstrapping using package *boot*

1. Define a function that returns a statistic or multiple statistics. In case of single statistic, let the function return a value and in case of multiple statistic let it return a vector. Function must have an argument named `indices` that the function boot( ) can use to select observations while repeated processing
2. Call this function through function boot( ) in order to run it with some R replications.
3. Use the function boot.ci( ) for obtaining confidence intervals

Sane's STATS Academy of Statistics

Function boot( )

Syntax :

boot(data, statistic, R, …)

Where

data : The data as a vector, matrix or data frame

statistic : A function which when applied to data returns a vector containing the **k** statistic(s) of interest

R : No. of replications

Object returned by function boot( ) contains:

t0 : The observed value of statistic applied to data

t : A matrix of dimension R x k, with k statistics and R replications

Sane's STATS Academy of Statistics

Function boot.ci( )

Syntax :

boot.ci(boot.out, conf = 0.95, type = "all“, …)

Where

boot.out : Object returned by function boot( )

conf : Confidence Coefficient

type : type of confidence intervals. Other possible values are "norm","basic", "stud", "perc", "bca"

Sane's STATS Academy of Statistics

Illustration with Single Statistic

```
> results <- boot(vect, statistic = Average , R = 1000)
> print(results)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = vect, statistic = Average, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*     47.9 -0.0869    8.466425
> ## Components t0 and t
> results$t0
[1] 47.9
> str(results$t)
 num [1:1000, 1] 33.5 42.2 43.7 46.3 34.6 38.4 43.1 55.1 48.1 47 ...
```

```
> boot.ci(results, type = "bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "bca")

Intervals :
Level       BCa
95%   (31.2, 64.6 )
Calculations and Intervals on Original Scale
```


Sane's STATS Academy of Statistics