Statistical Inference Using Bootstrapping

BOOTSTRAPPING

- Bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample.

- The method is especially useful when sampling distribution of estimator is not standard distribution.

- The method can be used in any statistical inference problem.

- The use of the term 'bootstrap' comes from the phrase “To pull oneself up by one's bootstraps ” - generally interpreted as succeeding inspite of limited resources.

BOOTSTRAPPING

- Many conventional statistical methods of analysis make assumptions about normality, including correlation, regression, *t* tests, and analysis of variance. When these assumptions are violated, such methods may fail.

- Bootstrapping method is used when assumptions of parametric tests are not satisfied and/or the parameter of interest does not have standard sampling distribution. (example- median or difference between 2 medians)

BOOTSTRAPPING

- Bootstrapping, a data-based simulation method, is steadily becoming more popular as a statistical methodology. It is intended to simplify the calculation of statistical inferences, sometimes in situations where no analytical answer can be obtained.

- As computer processors become faster and more powerful, the time and effort required for bootstrapping decreases to levels where it becomes a viable alternative to standard parametric techniques.

BOOTSTRAPPING

In following scenarios bootstrapping can be used:

- Small sample sizes with unknown population distribution
- When assumption of normality does not hold
- Skewed data
- A non-linear combination of variables (e.g. a ratio)
- A location statistic other than the mean(median,difference in median)

BOOTSTRAPPING

Original Sample
12,23,11,29,34, 38,41,45,6

Median=29.00

	Sample 1	Sample 2	Sample 3			Sample B
	23	11	6			41
	23	29	45			45
	29	11	11			11
	29	29	29			34
	34	34	11			34
	38	34	38			38
	41	41	41			41
	45	45	41			45
	41	11	6			6
Median	34.00	29.00	29.00			38.00

Here sampling distribution of sample median is Generated.
Assuming B=1000,
25th value and 975th value
Will provide 95% confidence Interval for median.

Sample size of original and each bootstrap sample is 9

DATA SCIENCE INSTITUTE

BOOTSTRAPPING IN R

```
#Import csv data set "bootdata"


bootdata<-read.csv(file.choose(),header=T)


#Install and call package "boot"


 install.packages("boot")
 library(boot)
```

BOOTSTRAPPING IN R

```
#Create function to find median. The function boot calls this function
 R=1000 times.

f<-function(data,i){
d<-data[i,]
med<-median(d)
return(med)
}
#The function boot sends data and random number vector to function f
1000 times.Each time random number vector is different.

bootobject <- boot(data=bootdata,statistic=f,R=1000)
```

BOOTSTRAPPING IN R 95% CI

```
#calculate 95% confidence interval
boot.ci(bootobject,type="perc")
```

```
Intervals :
Level     Percentile
95%   ( 0.90,  1.85 )
Calculations and Intervals on Original Scale
```

THANK YOU!!

DATA SCIENCE
INSTITUTE