

Stat 471: Lecture 15

Bootstrap I.

Moo K. Chung
mchung@stat.wisc.edu

October 10, 2003

1. In previous lecture, we generated pseudo populations based on parametric model $N(\mu, \sigma^2)$. Now we show a technique called *bootstrap* where no parametric assumptions are made. See Efron and Tibshirani (1993) An introduction to the bootstrap for detail. Given a random sample $\mathbf{x} = (x_1, \dots, x_n)$, The CDF of true population, F was estimated as the proportion of the total sample points that is equal or less than x , i.e.

$$\hat{F}(x) = \sum_{i=1}^n \mathbb{I}_{(-\infty, x_i]}(x).$$

Note that $\hat{F}(x_{(i)}) = i/n$ and $\hat{F}(x_{(i+1)}) = (i+1)/n$. So we are treating each sample point to occur with equal probability $1/n$. Based on this idea, we resample *with replacement* from $\mathbf{x}$ and we will denote the new sample as $\mathbf{x}^*$. If we resample m times, we denote them by $\mathbf{x}^{*1}, \dots, \mathbf{x}^{*m}$.

```
boot=inline('x(unidrnd( length(x),
m,length(x)))','x','m');
x=[5 8 3 2];
>>bs=boot(x,3)
bs = 2      3      8      8
     3      8      3      2
     2      8      3      3
```

2. Suppose $\mathbf{X} = (X_1, \dots, X_n) \sim F$. Suppose we are estimating θ by $\hat{\theta}(\mathbf{X})$. The distribution of $\hat{\theta}(\mathbf{X})$ would be an interest for statistical inference but it might be difficult to compute analytically. In this situation, we generate m bootstrap data $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*m}$ and obtain the bootstrap replications of $\hat{\theta}$,

$$\hat{\theta}^{*i} = \hat{\theta}(\mathbf{X}^{*i}).$$

These bootstrap resamples provide us with an estimate of the distribution of $\hat{\theta}$. In particular, we can estimate $\mathbb{E} \hat{\theta}(\mathbf{X})$ and $\text{Var} \hat{\theta}(\mathbf{X})$ by

$$\mathbb{E} \hat{\theta}(\mathbf{X}) \approx \bar{\theta}^* = \sum_{i=1}^m \hat{\theta}(\mathbf{X}^{*i})/m$$

$$\text{Var} \hat{\theta}(\mathbf{X}) \approx \sum_{i=1}^m \left(\hat{\theta}(\mathbf{X}^{*i}) - \bar{\theta}^* \right)^2 / (m-1)$$

From strength.data, let us estimate the population mean and variance based on 200 bootstrap replications. For population mean, we use $\hat{\theta}(\mathbf{X}) = \bar{X}$. For population variance, we use $\hat{\theta}(\mathbf{X}) = \sqrt{n}\bar{X}$.

```
bs=boot(arm,200); %200x147 matrix
bsmean=mean(bs,2)
>>[mean(arm) mean(bsmean)]
ans =
    78.7517    78.9025

%MATLAB built-in function
>>mean(bootstrp(200,'mean',arm))
ans =
    78.7799
%For matrices, var(X) is a row
%vector containing the variance
%of each column.
>>[var(arm) 147*var(bsmean)]
ans =
   445.6040   423.0257
```

3. Read Bootstrap bias estimation and confidence interval from your text book p.219-227.