STAT 703/J703
April 7th, 2005
-Lecture 24-

Instructor: Brian Habing
Department of Statistics
LeConte 203
Telephone: 803-777-3578
E-mail: habing@stat.sc.edu

Today

Methods Based on the CDF cont…
- The Nonparametric Bootstrap
- Relation to Survival Functions

Remembering Bayes Theorem

The empirical distribution function (or empirical cumulative distribution function) is defined as:

$$F_n(x) = \frac{1}{n}\{\# x_i \le x\}$$

Notes:

- $\sup_x |F_n(x) \rightarrow F(x)| \rightarrow 0$ as $n \rightarrow \infty$
- For each x, $F_n(x)$ is binomial with mean $F(x)$ and variance $(F(x)(1- F(x))/n$

<u>Nonparametric Bootstrap</u>

We previously examined the parametric bootstrap for the case when we assumed the data came from some distribution $F(\theta)$ with unknown parameter θ.

Estimating θ, we then generated "bootstrap samples" from the distribution $F(\hat{\theta})$. The statistic $\hat{\theta}^*$ is then calculated for each sample.

We then use the analogy that the sampling distribution of $\hat{\theta}$ is to θ sampling distribution of $\hat{\theta}^*$ is to $\hat{\theta}$.

The nonparametric bootstrap uses the same basic analogy… except that we don't have a specific distribution in mind for *F*.

Because of this we the parameter θ that we are focusing on is usually something like the mean, variance, or median that is "universally defined."

Example: Estimate the variance and bias of the sample standard deviation *s* for the sample:

```
3.95  3.79  3.75  2.71  5.52
6.12  1.74  6.05  3.92  5.69
```

*Generated using x<-10*rbeta(10,3,4) so the population has mean 30/7≈4.29 and variance 150/49 ≈3.06 (sd ≈ 1.75).*


```
sdboot<-function(x,nboots=10000){
sampsize<-length(x)
bootsamps<-
  matrix(sample(x,sampsize*nboots,
  replace=T),ncol=sampsize)
bootstats<-apply(bootsamps,1,sd)
est.bias<-mean(bootstats)-sd(x)
est.se<-sd(bootstats)
c(est.bias,est.se)
}
```

How well does it work?

When can it have trouble?

- Small sample sizes (but doesn't everything?)
- Statistic is not smooth

Section 10.2,2: Survival Functions

Let T=the survival time

$S(t)=P(T>t)=1-F(t)$

Hazard Function: The probability that an individual alive at time t will die in the time interval (t, $t + \varepsilon$)

For next time… Recall the

Law of Total Probability:

Let $B_1, B_2, \ldots, B_n$ be disjoint and exhaustive so that $\cup_{i=1 \text{ to } n} B_i = \Omega$, $B_i \cap B_j = \phi$ for $i \neq j$.

Then for any A,

$P(A) = P(A|B_1)\, P(B_1) + P(A|B_2)\, P(B_2) + \cdots + P(A|B_n)\, P(B_n)$.

Bayes' Rule: Let $B_1, \ldots, B_n$ be disjoint and exhaustive ($\cup B_i = \Omega$). Let A be any event. For any $j=1, \ldots, n$

$$P(B_j|A) = \frac{P(A|B_j)\, P(B_j)}{P(A|B_1)\, P(B_1) + \ldots + P(A|B_n)\, P(B_n)}$$