22S:166
Introduction to the Bootstrap

Lecture 8
September 18, 2009

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Resources

- Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Number 38 in CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Davison, A.c. and Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, New York: Cambridge University Press.
- materials listed under Web Resources

Review concepts

- suppose we have one sample of n data values: $y_1, \cdots, y_n$
- sample values considered outcomes of i.i.d. random variables $Y_1, \ldots, Y_n$
- probability density function (pdf) or probability mass function (pmf) f
- cumulative distribution function (cdf) F
- sample will be used to make inference
  - about population characteristic θ
  - using statistic T whose value in sample is t
- questions of interest regarding T
  - bias?
  - standard error?
  - quantiles?
  - how to compute confidence limits for θ?

  - likely values under a null hypothesis of interest?

Two classes of statistical methods

- parametric
  - particular mathematical model for behavior of random variables Y_j
  - pdf or pmf f is completely determined by values of unknown parameters ψ
  - quantity of interest in statistical analysis θ is a component or function of ψ
- nonparametric
  - uses only the fact the Y_js are i.i.d.
  - no mathematical model for their distribution
  - (may be useful to do a nonparameteric analysis even if a reasonable parametric model exists)
    - to assess sensitivity of conclusions to assumptions of parametric model

The empirical distribution

- puts probability mass $\frac{1}{n}$ at each sample value y_j
- empirical distribution function (edf) or $\hat{F}$
  - nonparametric mle of F
  - sample proportion $\hat{F}(y) = \frac{\#\{y_j \leq y\}}{n}$
    - where $\#$ denotes the number of items in a set
- edf plays role of fitted model when no mathematical form is assumed for F

Example of edf

```
> library(QRMlib)
> help(edf)
> data <- sort(rnorm(100) )
> plot( data, edf(data), type = "s" )
> qs <- seq(-2.5,2.5,by=0.005)
> lines( qs, pnorm(qs), lty = 2 )
```


Example for the nonparametric bootstrap: City population data

- for each of $n = 49$ U.S. cities, two data values
  - u_j = population in 1920 (in 1000s)
  - x_j = population in 1930 (in 1000s)
- population of interest is all U.S. cities
- the 49 cities are assumed to be a simple random sample from this population
- define (U,X) as pair of population values for a randomly selected city
- then if we knew $\theta = \frac{E(X)}{E(U)}$ and the total 1920 population for the U.S., we could estimate the total 1930 population of U.S.
- want to estimate θ without assuming any parametric model for X and U
- sample-based statistic is $T = \frac{\bar{X}}{\bar{U}}$

- observations 1 to 10 of this dataset are included with the **boot** package for R


```
> library(boot)
> data(city)
> city
     u   x
1  138 143
2   93 104
3   61  69
4  179 260
5   48  75
6   37  63
7   29  50
8   23  48
9   30 111
10   2  50
```


The non-parametric bootstrap

- goal: to get an idea of the sampling distribution of the statistic T under repeated sampling from the population of interest
- basic idea: our sample data gives us all the information we have about the whole population
- steps:
  1. calculate statistic of interest (call it $\hat{\theta}$) from dataset as a whole
  2. fit edf $\hat{F}$
  3. Draw a "bootstrap sample" from $\hat{F}$ and calculate statistic of interest on bootstrap sample
     - i.e., draw a sample of size n from original dataset **with replacement**
     - $Y_1^*, Y_2^*, \ldots, Y_n^* \sim \hat{F}$
     - $\hat{\theta}^* = \hat{\theta}(Y_1^*, Y_2^*, \ldots, Y_n^*)$

4. repeat step 2 independently a large number B of times obtaining bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$
5. Use bootstrap replications to:
   - estimate standard error of $\hat{\theta}$
   - estimate bias
   - obtain confidence interval