**Bootstrapping**

Jack-knife

• The jackknife is a tool for estimating standard errors and the bias of estimators

• As its name suggests, the jackknife is a small, handy tool; in contrast to the

bootstrap, which is then the moral equivalent of a giant workshop full of tools

• Both the jackknife and the bootstrap involve resampling data; that is, repeatedly creating new data sets from the original data

The jackknife deletes each observation and calculates an estimate based on the remaining n − 1 of them

• It uses this collection of estimates to do things like estimate the bias and the standard error

• Note that estimating the bias and having a standard error are not needed for things like sample means, which we know are unbiased estimates of population means and what their standard errors are

It has been shown that the jackknife is a linear approximation to the bootstrap

• Generally do not use the jackknife for sample quantiles like the median; as it has been shown to have some poor properties

**The bootstrap**

• The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics

• For example, how would one derive a confidence interval for the median? • The bootstrap procedure follows from the so called bootstrap principle

Suppose that I have a statistic that estimates some population parameter, but I don’t know its sampling distribution

• The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution

• In practice, the bootstrap principle is always carried out using simulation

• The general procedure follows by first simulating complete data sets from the observed data with replacement

• This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution

• Calculate the statistic for each simulated data set

• Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error

Example

• Consider again, the data set of 630 measurements of gray matter volume for workers from a lead manufacturing plant

• The median gray matter volume is around 589 cubic centimeters

• We want a confidence interval for the median of these measurements

• Bootstrap procedure for calculating for the median from a data set of n observations

i. Sample n observations with replacement from the observed data resulting in one simulated complete data set

ii. Take the median of the simulated data set

iii. Repeat these two steps B times, resulting in B simulated medians

iv. These medians are approximately draws from the sampling distribution of the

median of n observations; therefore we can

• Draw a histogram of them

• Calculate their standard deviation to estimate the standard error of the median • Take the 2.5th and 97.5th percentiles as a confidence interval for the median

**Summary**

• The bootstrap is non-parametric

• However, the theoretical arguments proving the validity of the bootstrap rely

on large samples

• Better percentile bootstrap confidence intervals correct for bias

• There are lots of variations on bootstrap procedures; the book “An Introduction to the Bootstrap” by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information