STAT 400 Data Analysis Nov 7, 2003

The Bootstrap:

Suppose you have sample data and a population parameter for which the distribution theory of the desired parameter estimator is unknown. (It may be unknown because the population class of distributions is unknown, or because the theory of this estimator is Intractable, or because you just don't know the theory!) The choice of estimator is assumed to be known – the issue is that you want to know the variability of your estimator, since that tells you about the precision of your estimate. The bootstrap gives a way to find this precision, based only on the sample data you have used to come up with the estimate.

The operation of the basic bootstrap (the only version described here) is simple: You just compute the estimate for a large number of resamples of your original sample, and observe the variability (by computing the SD of the resample estimates). A "resample" here is a sample of the same size as the original sample, selected from the original sample at random and with replacement.

The justification for this procedure is this:

1. The ECDF is an estimate of the CDF.
2. Resampling is the same as using the inverse-CDF-transform method of simulating a distribution for a given CDF. The "given CDF" in this case is the ECDF.
3. The resample estimates thus reflect approx. the same variability as if the CDF itself had been sampled.

The inverse transform method of generating a random sample from a population whose CDF is F( ) is justified by the Probabilty Integral Transform Theorem. It says that if X is a RV having CDF F( ), then F(X) has a U(0,1) Distribution. The proof is done by computing P(F(X)<t) = P(X<inverse fcn of F at t) = F(inverse function of F at t) =t which is the CDF of the U(0,1).

A demo of this can be had using the following program, since it shows that the bootstrap can estimate that the sample mean of 20 values from a distribution has a SD that turns out to be very close to s/sqrt(n), and yet it does not use this formula.

```
Gmacro
boot2.mac     #estimates SD of 90th percentile
brief 0
n c1 k6
let k9=round(0.9*k6)
sort c1 c2
let k10=c2(k9)
do k5=1:1      # do it 3 five times but with the same one sample
do k1=1:1000
sample k6 c1 c2;
repl.
#let k2=mean(c2)
sort c2 c6
let k2=c6(k9)    # est 90th percentile in n=k6
let c3(k1)=k2
enddo
brief 2
let k4=stdev(c3)
let c5(k5)=k4
enddo
dotplot c1
print k10
print c5
#erase c1-c6
endmacro
```

In a situation where the theory is unknown, such as the variability of the sample 90th percentile (from an unknown population, perhaps), the following bootstrap procedure will provide an estimate of the SD of the sample 90th percentile.

```
Gmacro
boot2.mac     #estimates SD of 90th percentile
brief 0
n c1 k6
let k9=round(0.9*k6)
sort c1 c2
let k10=c2(k9)
do k5=1:1      # do it 3 five times but with the same one sample
do k1=1:1000
sample k6 c1 c2;
repl.
#let k2=mean(c2)
sort c2 c6
let k2=c6(k9)    # est 90th percentile in n=k6
let c3(k1)=k2
enddo
brief 2
```

```
let k4=stdev(c3)
let c5(k5)=k4
enddo
dotplot c1
print k10
print c5
#erase c1-c6
endmacro
```

Try them out!

You need to specify your own sample (of any size) in column 1.