STAT 400 Data Analysis November 10, 2003

Today: More about the bootstrap

The bootstrap demos from last day were a bit rushed, and I have more to say about them, and more examples to display.

1. Does the Bootstrap produce correct results for the SD of the sample mean?

We computed the SD of the sample mean, based only on one particular random sample of sinze n=20. We showed that this method gave a value similar to the theoretical value and the usual estimator of it. ($\sigma/\sqrt{n}$ and $s/\sqrt{n}$). These latter formulas did not need to be known to do this.

For example, for a particular sample of N(0,1) data, we found:

Standard deviation of sample, s = 1.14

$s/\sqrt{20} =$ 0.26

bootSD of sample mean = 0.24

Since in this case we know Population SD = 1,
Population SD of mean = $1/\sqrt{20}$ = 0.22

So the bootstrap has performed well here.

2. Does the Bootstrap produce correct results of the SD of the 90th percentile?

We compute the SD of the sample 90th percentile, based only on one random sample.

The theoretical formula is not widely known and so this was an example of how the bootstrap could estimate this SD in such a situation. Today we will check this bootstrap approach by simulating the correct SD since we will start with a known distribution.

For example:

Here is a sample n=20 from N(0,1):

```
     .              :    :::       .:      . . .:   :       .
   -+---------+---------+---------+---------+---------+-----C1
 -2.40     -1.60     -0.80      0.00      0.80      1.60
```

The 90th percentile could be estimated by the 18th order statistic, in this case 1.22. But how precise is this estimate? Bootstrap it!

BootSD = 0.27 so we might say our estimate of the 90th percentile is 1.22±.27 (mean ± SD)

But is this right? Lets simulate some samples from N(0,1), estimate the 90th percentile from each sample (18th order statistic) and compute the SD when we have enough of them.

simSD90 0.35

So our estimate of .27 was a bit low. But of course, the .27 was based on a sample of n=20, and the .35 is a population-based SD, so we don't expect perfection. The question is, is this bootstrap method seriously biased in this case?

We can investigate this with simulation. I did the whole experiment (generate 20 N(0,1), bootstrap the 90th percentile estimate and compute its SD) 15 more times. Here are the 15 values of the bootstrap estimates of SD of the 90th percentile:
.36, .35, .34, .27. .24, .14, .40, .36, .49, .46, .16, .40, .36, .38, .34

and the average of these is .34 – pretty close to .35. If the method is biased, the bias would be modest in this case. But there is no reason to expect unbiasedness – it would depend on the situation.

Does the good performance of the bootstrap depend on the N(0,1) sample we used?

These demos used "data" that was generated from a N(0,1) distribution, but this information was not used. Someone asked if the performance would be as good for some less symmetrical distribution, so I will show how it works with Gamma (3,1)

Here is a Gamma (3,1), n=20, which is strongly right skewed:

```
        :  : .    ..:. . .    ...     .  .         .      ..
-------+---------+---------+---------+---------+---------C1
     1.2       2.4       3.6       4.8       6.0       7.2
```

In this case the bootstrap SD of the 90th percentile estimate is

bootSD 1.07

The true SD (using the n=known population Gamma(3,1) of the 90th percentile is simSD90   0.82

Our one sample produced a pretty good estimate of the SD of the 90th percentile. To see if it is usually this good, you can do the simulation using the “boot2” program 15 times, starting with a new Gamma sample each time in column 1.

(The program to produce the theoretical SD (0.82 above) is

```
Gmacro
bootcheck.mac
rand 20 c1-c500;
gamm 3 1.
Do k1=1:500
sort ck1 c501
let k2=c501(18)
let c502(k1)=k2
enddo
stdev c502 k3
name k3 'simSD90'
print k3
endmacro
```

Is the variability of the bootstrap estimate of SD of a statistic too variable to be useful?

We have shown that the bootstrap does provide variability estimates which are almost unbiased, but they do vary quite a bit in samples of size 20. However, any estimate of variability will vary quite a bit if based on such a small sample. For example, the sample SD as an estimator of the population SD has quite high variability, even in the case of N(0,1) data.

Here is the result of an experiment on this:

```
                        .
                        :
                        :.. .        .
                  ..  : ::: :.  : .  :: :   :
       .    :: :.:::..:.:::::::::.: :::::.:.:..::   ... .  .
    +---------+---------+---------+---------+---------+-------Sample  SD
  0.60      0.75      0.90      1.05      1.20      1.35
```

mean of sample SD = 1.00   SD of sample SD = 0.16 range in n=100 trials is (0.63, 1.41)
So the estimate of the SD which we know is 1 is only within ± .16
In our 15 repetitions of the bootstrap estimate of the 90th percentile, the

SD of the SD of the estimate was only 0.10. This is less than the 0.16 which is the variability of the sample SD that we habitually use as our estimate of the population SD.

Of course, larger samples provide much better estimates. The small sample used in these experiments (n=20) was used to demonstrate the surprising efficiency of the bootstrap – in a case where it might be expected to fail, it worked pretty well.

Can the bootstrap be used in more complex situations?

I mentioned last time that the bootstrap can be used in very complicated situations for which nobody knows the theory needed to compute the SD of the estimator, other than by the bootstrap. I will describe a case like this as well.

Stepwise regression is a popular technique with statisticians that do not like to make full use of the context of a data set (this reflects my bias against this method!). It is well-known that variables that are highly correlated (positively of negatively) will pre-empt each other in a stepwise regression. So the subset of variables you end up with will depend to some extent on chance – the particular errors observed in the sample. The subset of variables produced by a stepwise regression is often suggested as indicating the variables that are most predictive of the dependent variable, even though a different set might have been almost as good. How can one judge, in a particular instance, the seriousness of this problem? Bootstrap the whole stepwise procedure, and observe the variability of the outcome.

The demo of this is done with a data set on body measurements of 252 men. The aim is to find a model which estimates the body density (measured by immersing the men in water.)

There are 13 body measurements (hgt, wgt age, etc …) and one density variable. First we use a random subset of 25 of these men, and then do a stepwise regression. We find using stepwise regression that only one body measurement "abdomen" is significant. But can we judge how much this conclusion depends on the particular sample used, without referring to the population of 252 men? The bootstrap shows that different results might easily have been observed. This is the case even with using all 252 men in the sample.

The bootstrap has confirmed what is widely known to be true – stepwise regression does not produce good explanatory models with any consistency.

Continued next page ….

About the “men” data:

Here are the first few lines (I’ll send the unwrapped total data set by e-mail):

Density-Water	Eqn	Age	Wgt	Hgt	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59	37.3	21.9	32	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154	66.25	34	95.8	87.9	99.2	59.6	38.9	24	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
1.034	28.7	24	184.25	71.25	34.4	97.3	100	101.9	63.2	42.2	24	32.2	27.7	17.7
1.0502	20.9	24	210.25	74.75	39	104.5	94.4	107.8	66	42	25.6	35.7	30.6	18.8
1.0549	19.2	26	181	69.75	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
1.0704	12.4	25	176	72.5	37.8	99.6	88.5	97.1	60	39.4	23.2	30.5	29	18.8
1.09	4.1	25	191	74	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2

Ignore the “Eqn” variable (the second column):

There are a couple of gross errors – you can ignore them for now.

Length measurements are in cms. Except HGT which is inches.

The MINITAB command to select a random 25 rows (without replacement):

```
Sample 25 c1 c3-c15 c101 c103-c115
```

The Minitab Commands to bootstrap the stepwise regression:

```
Sample 25 c101 c103-c115 c201 c203-c215;
Repl.
Step c201 c203-c215
```

I think I want you to try this, so I’ll make it an exercise, but not due until Wed Nov 26. More details on Wednesday this week.