Psychology 340/341
Advanced Statistical Methods
David C. Howell
2001--2002

Bootstrapping

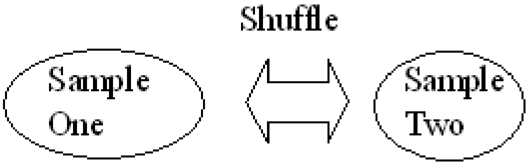
4/30/2002

The Context

I started this year with a talk on randomization tests, which are part of the more general area of Resampling Statistics. I want to end it with a talk on bootstrapping, which is the other leg of the resampling stool.

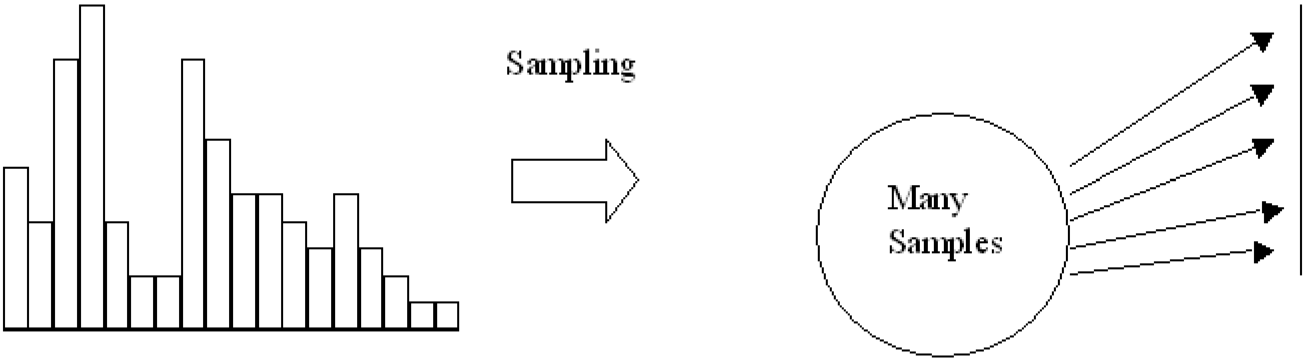
- Both randomization tests and bootstrapping represent the results of many random samples from some population.
- They differ in how that random sampling is done, which has the effect of producing a major difference in the kinds of populations to which they refer, and the purpose of each procedure.
- **Randomization tests** are only concerned with the **original data,** and simply ask about the distribution of some statistic if we assume that each observation is as likely to end up in one group as in another. (Notice that I didn't use the phrase "null hypothesis" in that sentence. If we have a null hypothesis, it is of the form "treatments do not affect outcome.")
    - Here we take the original data and sample **without replacement** to create two (or more) groups.
    - By sampling without replacement, we are simply rearranging (shuffling) the data across groups.
    - We aren't saying anything about the population from which the data were drawn, but just about how such data could have been distributed if the treatment had no effect..

- **Randomization Test Schemata**



- **Bootstrapping** actually pays attention to the **population** from which the data were drawn (sort of), but it does not assume that the population is normal, or any other particular shape. In fact, we generally assume that the population is of exactly the same shape as our sample.
    - Here we take the original data and sample **with replacement**.
    - By sampling with replacement we are allowing a particular observation to appear multiple times (or no times) in our sample.
    - We are basically acting as if the sample data are a perfect reflection of the population.
        - We could either think of each observation being replicated millions of times, to build up a huge population, or we could sample the one set of scores with replacement after each observation is drawn. The result would be exactly the same.

    - **Bootstrapping Schemata**



For a list of references on bootstrapping, go to my web pages (http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html) and click on the link to resampling procedures. The most important person in the history of bootstrapping is Bradley Efron, and his book with Tisbshirani is a classic. Lunneborg's book is also excellent.

Purpose

Although randomization tests are generally thought of as a way of testing a null hypothesis, bootstrapping is generally thought of as a way of creating confidence limits on some statistic. It doesn't have to be this way, but it usually is.

Bootstrapping has two overlapping purposes.

1. Deal with nasty distributions that are not normal.
2. Estimate parameters that we don't know how to estimate analytically.

The second of these is by far the more important in terms of where we will see bootstrapping's use in the future. This is the whole reason why I am talking about this topic today.

I will illustrate this below.

When we have normal distributions and common (well behaved) statistics such as the mean and variance, our normal parametric procedures can do a good job of setting confidence limits, and we don't need bootstrapping or any other resampling procedure.

1. Here we have normality.
2. Here we have an analytical solution.
    1. By "analytical" I mean a solution that comes about by use of a formula.
        1. For example, we know the standard error of the mean because it can be estimated by $s_{\bar{X}} = s/\sqrt{n}$.
        2. Also, students already know that
            - $CI_{.95} = \bar{X} \pm t_{.025}(df) * s_{\bar{X}}$
            - gives us the 95% CI on the mean of a population.

- But, to obtain confidence intervals this way we have to assume that our distribution is reasonably normal, and that we are looking for a CI on the mean, as opposed to the median or mode.
- If we were looking for a CI on the median, we would have an analytical solution **if** the distribution were normal. Otherwise we would not have an analytical solution.
    - ○
- If we were looking for a CI on the *difference* between two medians, there is not a neat analytical solution that would give it to us, regardless of the shape of the population.

This is where bootstrapping really comes in--i.e. where there is no analytical solution, or the one there is is too dependent on untenable assumptions.

Why would we really care, anyway?

In the good old days (35 years ago) most psychologists ran straightforward experiments with a couple of groups, and all we wanted to know was whether means were different. Or, we ran simple multiple regression studies in which we wanted to know the optimal equation for predicting one variable from several others. Everything was nice and simple.

But things are really not that simple any more. We now calculate all sorts of statistics that are far more complex (and less well understood) than those of the past. This whole class illustrates some of the ways that statistics has changed over the course of my career--and people keep asking me if statistics ever changes, or do I write new editions just to pass the time.

> For example, we now do Structural Equation Modeling, Hierarchical Linear Modeling, Nonlinear Regression, Mediated relationships, etc., for which our test statistics are not well developed, at best. You can see this by the fact that many tests put out competing test statistics--think of Manova with Pillai's trace, Hotelling's trace, Wilk's lambda, and Roy's Largest characteristic root. If you ran a nonlinear regression, we don't have any decent test to tell whether your coefficients are significant. Similarly, you already know that Wald's test in logistic regression is far from optimal.
>
> But if we don't have a good test, or if we don't have *any* test, what are we to do? Bootstrap!!!
>
> A few weeks ago we looked at mediating effects in regression. We were interested to know if the path from maternal care to maternal self-efficacy was mediated by self-esteem. So we took the product $b_1 * b_2$ as the measure of that indirect path. I gave a formula for testing the significance of this path, but it is a formula that depends upon parametric assumptions. There is even debate over the correct formula for the standard error of that path. Perhaps we can improve on that test.
>
> A bootstrapping approach to investigating that path would be to draw a very large number of samples from the original data (with replacement), compute $b_1 * b_2$ for each sample, and then compute confidence limits from the distribution of $b_1 * b_2$.
>
> The mechanics of doing that calculation could take a lecture in themselves, and there are a number of competing ways to estimate the confidence limits that vary in bias and precision. But

that isn't really the issue. I don't want people to walk away knowing exactly how to calculate a bootstrapped confidence interval. I want them to understand the logic that lies behind those calculations, and to not be surprised when they come across such an approach in the future. (SPSS already uses bootstrapping in some of its more complex analyses, as do Amos and Lisrel.)

For the past year I have constantly said things like "If the value of *t* exceeds the critical value at $\alpha = .05$, we can say that if we collected data over and over again with the null hypothesis true, only 5% of the time would we have a result more extreme than the critical value." When I come to hypothesis testing below, I am going to say exactly the same thing, only this time I really do collect a great many samples--though I do it with computer software rather than by finding huge numbers of participants. The major difference is that with a *t* test I imagine myself drawing these samples from a normally distributed population with certain characteristics, whereas with bootstrapping I actually do draw the samples from a population that looks exactly like the sample distribution.

Examples

One sample

It's best to start simple, so we will start with bootstrapping the median of one sample--e.g. the birthweight of infants of mom's who smoke. To do this we will use the data from Hosmer and Lemeshow, which we have seen earlier.

Emphasize the idea that with something like a **mean** with a **normal** population, we are asking *analytically* what the distribution of sample means would look like if we sampled from a population whose parameters match our sample, and whose distribution is normal. We just whip out our formula and get to work.

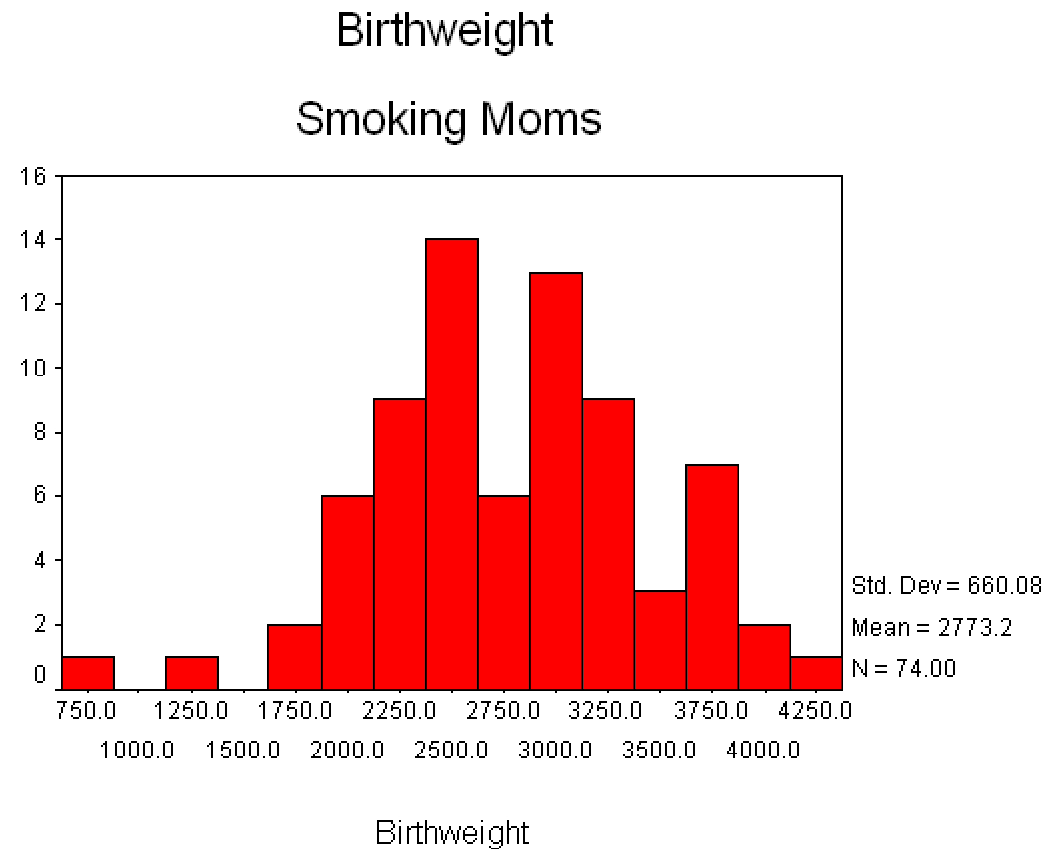
With other statistics, such as a median, we can't do that analytically, because we don't have a formula.

So what we will do is to draw many samples, with replacement, from some population, calculate the median for each sample, and then use our sampling distribution of the median to obtain confidence limits.

But, what population do we draw from?

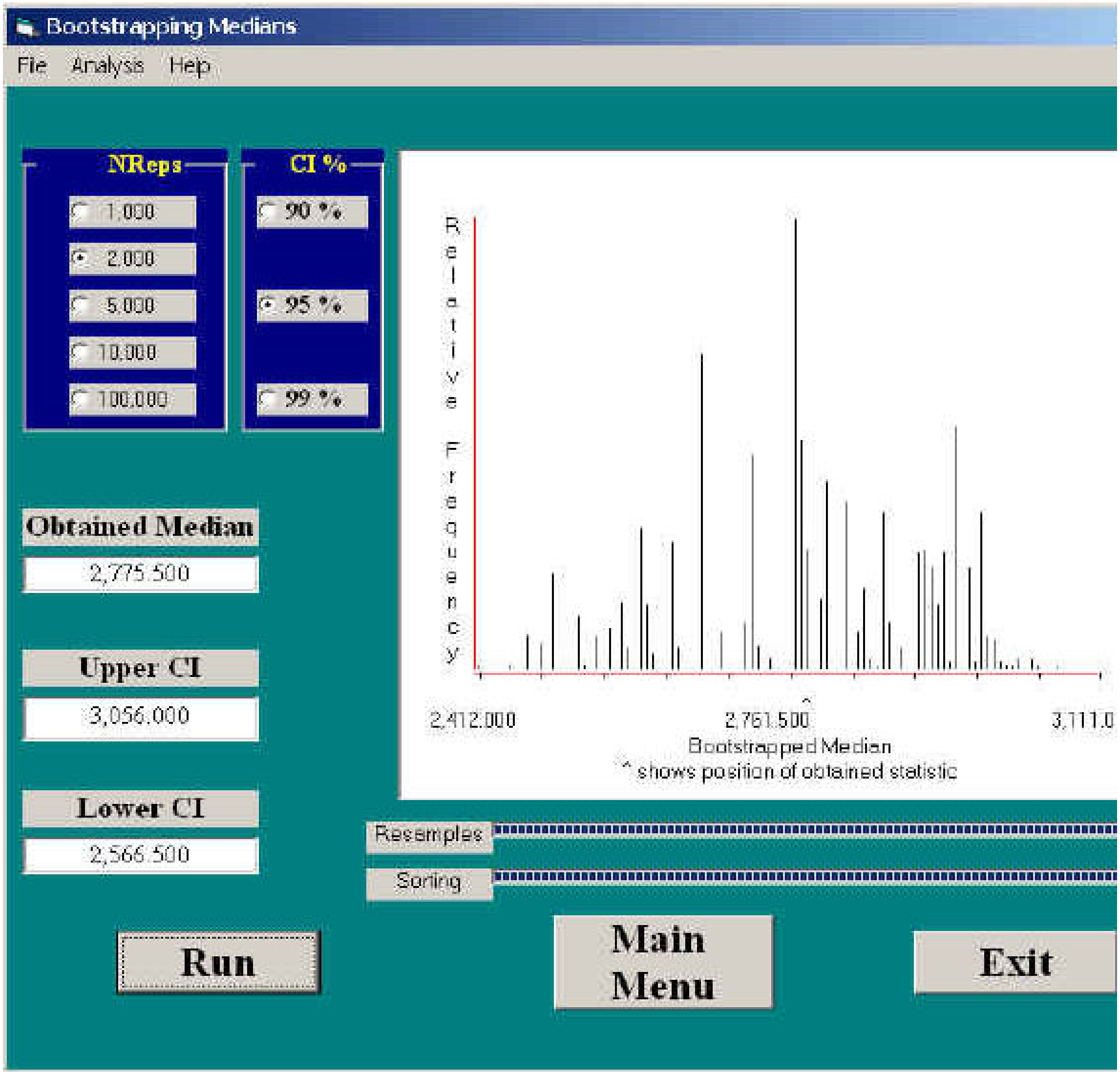
The *best* (maximum likelihood) guess is that the actual population looks exactly like our sample. So why not draw resamples from a "pseudopopulation" that matches our sample perfectly, and see what kinds of medians we get?

The following is the distribution of birth weights from data on smoking mothers provided by Hosmer and Lemeshow..



These data may be roughly normal, but they certainly have some outliers that would drastically influence the mean. So I bootstrapped the median.

> I drew 2000 samples, with replacement, from this pool of cases, calculated their medians, and plotted the medians. I also found those sample medians that comprised the middle 95% of the distribution, and these represent the confidence limits. (There are slightly better ways of doing that, but it would not make a noticeable difference here.)



NReps
○ 1,000
● 2,000
○ 5,000
○ 10,000
○ 100,000

CI %
○ 90 %
● 95 %
○ 99 %

Obtained Median
2,775.500

Upper CI
3,056.000

Lower CI
2,566.500

The 95% CI on the **median** is 2566.50 < median* < 3056.000.

The 95% CI on the **mean** would be 2773.243 ± 76.7322 = 2620.316 < μ < 2826.17.

Notice that the CI on the means is narrower, but that it irrelevant if what you want is a CI on the median. (One of the advantages of means is that their CI's are narrower in general.)

So now I can say with confidence that when I draw many samples from this population (a population that exactly resembles the sample), 95% of the time I would draw medians that lie between 2,566.5 and 3,065.0.

I said that there are better ways of getting the CI. I have no intention of discussing what those are here. They fall in the category of things that keep my brain alert, but that students don't really have to know here. But those alternative approaches still go through exactly the same process of pulling multiple samples and producing a sampling distribution. They just focus on ways of producing somewhat better estimates. I like that stuff, but it is not central to understanding what the procedure is all about.

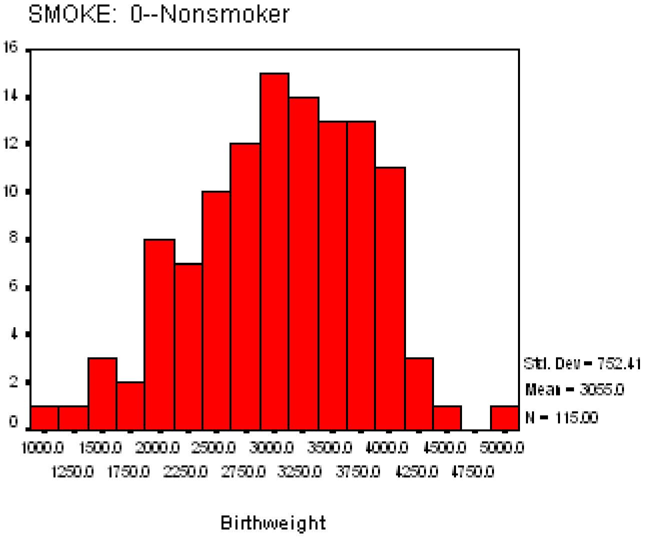
Median Differences from Two samples

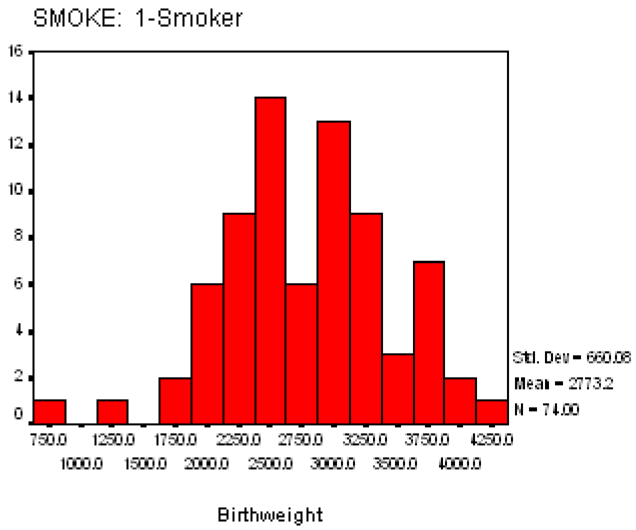
It is a relatively simple step to bootstrapping the difference between two medians. It is straightforward, and an extension on what we have already done.

Use Hosmer & Lemeshow's data on birthweight as a function of whether or not mom smoked. It is important to know whether maternal smoking has an effect on the fetus, and one of the major effects would be low birthweight--which is a known risk factor for a host of problems.

There are 115 cases in Hosmer and Lemeshow's data where mom did not smoke, and 74 cases where she did. We might wish to compare the *medians*, instead of means, because of some belief that there are serious outliers in birthweight due to medical complications, and we don't want these to influence the measure of central tendency. Or we might think that the two populations have quite different shapes. Or we might think that the really tiny infants are probably the result of factors more serious than smoking, and the really heavy babies are not of interest. In other words, we might wish to focus in on the center, leaving aside outliers. (An alternative dependent variable of interest might be the percent below 2500 grams, but that's a different story.)

The actual distributions:

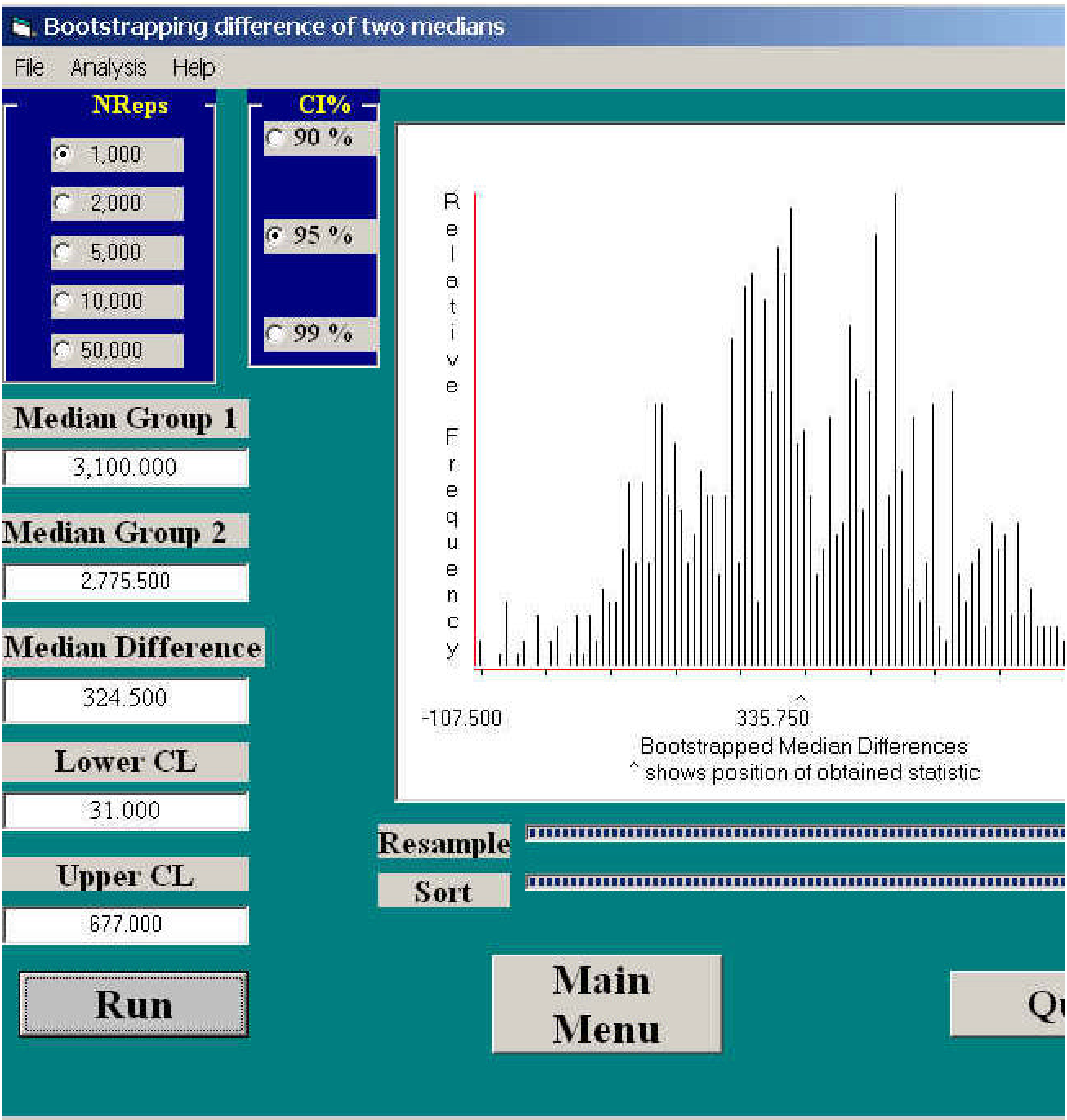




The process:

1. We act *as if* the data on mothers who did not smoke were a perfect image of the population of nonsmoking mothers. Similarly we act *as if* the data for smoking mothers were a perfect image of the population of smoking mothers.
2. We draw a sample, *with replacement*, from the nonsmoking mothers and calculate the median. We also draw a sample, *with replacement,* from smoking mothers and calculate the median.
3. We calculate the difference in the two medians and store that away.
4. We repeat the above process 1000 times, which gives up 1000 median differences for samples drawn from populations with those particular shapes and other characteristics.
5. We now know the sampling distribution of differences in medians for samples from these populations.
   1. We could report the confidence limits
   2. We *could* retain or reject the null based on whether the confidence limits included 0.
      1. Elaborate.

I carried out this process exactly as I have stated it here.



Bootstrapping difference of two medians

NReps: ● 1,000 ○ 2,000 ○ 5,000 ○ 10,000 ○ 50,000

CI%: ○ 90 % ● 95 % ○ 99 %

Median Group 1: 3,100.000

Median Group 2: 2,775.500

Median Difference: 324.500

Lower CL: 31.000

Upper CL: 677.000

Notice the shape of the sampling distribution of differences between medians.

The confidence limits on the *median* difference are 31 < median* < 677.

The confidence limits on the *mean* differences are 70.6927 < μ < 492.7338. Much smaller. Again, that is the nature of the mean, as opposed to the median. But if we don't care about the mean, that isn't important.

Notice how wide the confidence limits are (in either case). We can reject the null, but we don't know a lot about the difference in median birthweight between smoking and nonsmoking moms, except that it is positive.

Another Two Group Example

A nice example can be based on an example of waiting times for parking spaces, which I have used elsewhere in this course. In that example I had data that replicated a study by Ruback and Juieng (1997) on the amount of time it took drivers to leave their parking space when there was, or was not, someone waiting to take it when they left. Suppose that we have a particularly rude or thoughtless driver who decides that he will let you sit there while he places a call on his cellular phone. That call could take a couple of minutes, and would provide an outlier in the data. I can very easily imagine such an observation. To simulate this, I replaced the last observation in the Waiting group (42.06 seconds) with an extreme score (242.06 seconds).

As we did before, we assume that the data from the Waiting group accurately reflects the population of Waiting scores, and similarly with the NonWaiting group. We sample our data with replacement, and sometimes that observation will show up, and sometimes it won't. Because it is such a large value, it could cause the resulting median (or mean) difference to swing back and forth between positive and negative, creating an unusual sampling distribution. (Ask if it would have more effect on the mean or on the median difference.)

This can be demonstrated using the program that I have written. Do this in class.

The data file is named Waiting1extreme.dat

Question? Is this really a reasonable example?

> I don't know, but I guess that I don't have a good reason to believe that the truly rude person will be more likely to end up in the Waiting, as opposed to the NonWaiting, group. So perhaps I am unreasonable in my assumption that the sample data represent the true underlying populations.
>
> I deliberately used this as an example to illustrate that it is important to think about the underlying assumption before doing the analysis.

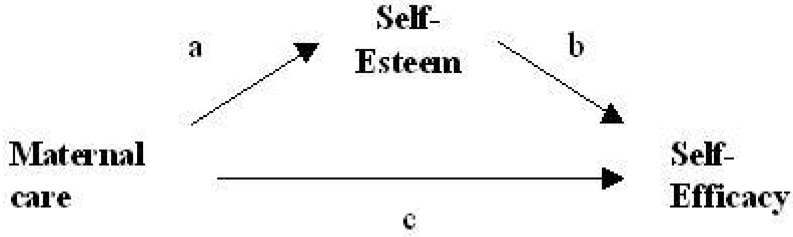
Esther Leerkes' Mediation Example.

This example contains an error. When I remove the error I will remove this note. The point is that the indirect path is the path from Maternal Care to Self-Esteem, time the semi-partial path from Self-Esteem to Self-Efficacy, partialling Maternal Care. The conclusions don't change.

This is the one that I am most excited about, because it represents an answer to a real, and important, problem. We had an analytical answer, from Baron and Kenny (1986), who developed it from Sobel, but we don't know how dependent that answer is on the shape of the distribution, nor how robust the test statistic is. So we are going to use a bootstrap approach, which does not depend on assumptions about the shape of the population. (A nice description and java program for the traditional Baron and Kenny approach can be found at http://quantrm2.psy.ohio-state.edu/kris/sobel/sobel.htm.)

I think that this is a particularly important and meaningful example, because it is close to what I expect you will find in computer software in the near future. It is an example of a situation where we don't have a very satisfactory analytical solution, so we use the brute force approach of our 2 gigabyte laptop.

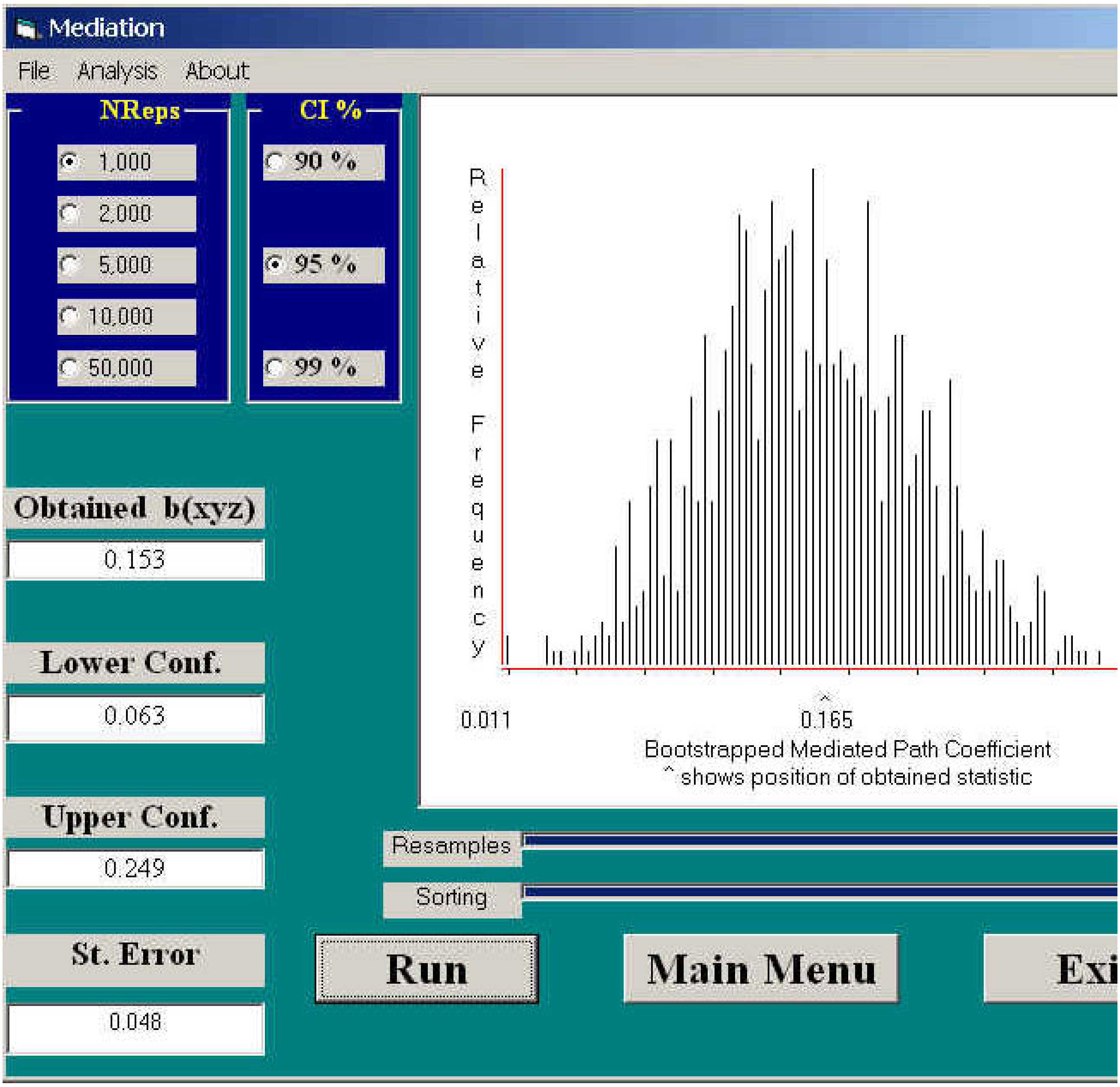
Two-three weeks ago we saw Esther Leerkes' problem in which she wanted to ask if there was a mediational path from maternal care to self-esteem to self-efficacy. She first calculated the path coefficient (beta) between maternal care and self-esteem. She then calculated the path from self-esteem to self-efficacy. Finally, she showed that when both maternal care and self-esteem were in the model, the *direct* path from maternal care to self-efficacy was reduced. This is shown in the following diagram.



Baron and Kenny (1986) said that the critical test is whether or not the direct path from Maternal Care to Self-Efficacy (c) drops (or, better yet, drops out) when Self-Esteem is also used as a predictor. They don't give a direct way to test the drop. Instead they propose that we should test the significance of the indirect path from maternal care to self-esteem to self-efficacy (given by the product of the two beta weights). They give a formula for the standard error of this path, but there is disagreement about that formula and we don't know how robust it is to violations of any assumptions. (See the preceding link to find references to this test.)

So, I went ahead and created a bootstrap test.

> The path is $\beta_a*\beta_b$.
>
> I drew 1000 bootstrap samples from Esther's data, and calculated the product $\beta_a*\beta_b$ for each. I then calculated the standard deviation of those estimates and the 95% CI on the parametric equivalent of $\beta_a*\beta_b$. This is shown in the figure below.



NReps

- ● 1,000
- ○ 2,000
- ○ 5,000
- ○ 10,000
- ○ 50,000

CI %

- ○ 90 %
- ● 95 %
- ○ 99 %

Obtained b(xyz): 0.153

Lower Conf.: 0.063

Upper Conf.: 0.249

St. Error: 0.048

Notice that my estimate of the standard error (.048) is somewhat smaller than Baron and Kenny's (0.053). That is a reliable finding, not a fluke.

Notice the CI, and the fact that it does not include 0.

We can conclude that there is a significant path from maternal care through self-esteem to self-efficacy, and that conclusion does not depend on any dubious assumptions.

I don't want the brevity of this section to suggest that it isn't important. It is very important, because it is a hint of what is to come. I would anticipate that there will be many more refinements in the mathematics of bootstrapping, but that is really not an important issue to students. That is just a refinement of a technique. But what is important is that we have a way of testing hypotheses both when we know that whatever assumptions are lying around are violated, and when we don't have any nice neat formula to do what we want.

And Yet Another Approach

I want to talk about a different approach to this same problem. When I first outlined this page I had not done the programming that was required for its solution, but this just points to things that are coming in the future. I have since done that programming, but not within the same overarching program as the last few examples. I have created a program within a package called Resampling Stats, which is a very good package.

When I first talked about mediation several weeks ago, I said that the original test proposed by Baron and Kenny would be to ask if the path from maternal care to self-efficacy decreased significantly when we ran a multiple regression with both maternal care and self-esteem. I further said that Baron and Kenny apparently didn't have a test for that decrease, but proposed, instead, testing mediation by seeing if the indirect path was significant. That is what we have done above.

There is nothing to prevent a literal test of Baron and Kenny's statistic. We could

1. Run the simple regression of maternal care -> self-efficacy, and record the slope.
2. Run the multiple regression of maternal care and self-esteem -> self-efficacy, and record the maternal care -> self-esteem slope.
3. Calculate the difference between the slopes in these two models.
4. Repeat steps 1-3 1000 times (or more)
5. Examine the distribution of differences obtained in steps 3 and 4.
6. Calculate confidence limits on the difference.

The following output comes from my recent efforts to use Resampling Stats for this purpose.

```
' Program to test mediation hypothesis by comparing the regression
' coefficient for direct path alone and for direct path and mediator.
' Based on data from Esther Leerkes.
maxsize sest 500 mc 500 seff 500        ' Allow room for input
Read file "MediateData.dat" sest mc seff
Count sest >= -999 n                    ' Calculates sample size
print n

'Do the two regressions on the original data and print out the results.
regress noprint seff mc e1              ' Use one predictor
score e1 sob1 sob0
Print e1
regress noprint seff mc sest d1         ' Use two predictors
```

```
score d1 mob1 mob2 mob0
Print d1

'Calculate difference in coefficients
subtract sob1 mob1 diff                    ' Diff = difference in b(i)
print diff

'Now do the bootstrapping loop
Repeat 1000
Generate n 1,n row
take seff row seff$
take mc row mc$
take sest row sest$
regress noprint seff$ mc$ sest$ d
score d mb1 mb2 mb0
regress noprint seff$ mc$ e
score e sb1 sb0

subtract sb1 mb1 diff                      ' Diff = difference in b(i)


end

median diff meddiff                        ' Get median of difference in b(i)
percentile diff (5 95) CIdiff90            ' Get percentiles
percentile diff (2.5 97.5) CIdiff95
percentile diff (1 99) CIdiff99
print meddiff
print CIdiff90 CIdiff95 CIdiff99
histogram diff
```

```
Line 5: 92 records (0 missing values) read from MediateData.dat

N = 92

E1 = 0.111550     3.2604

D1 = 0.058167    0.14714      2.9278

DIFF = 0.05338              (This is the obtained difference.)

MEDDIFF = 0.052551          (This is the median of the differences.)
```

(Confidence Intervals)

CIDIFF90 = 0.020843   0.093739

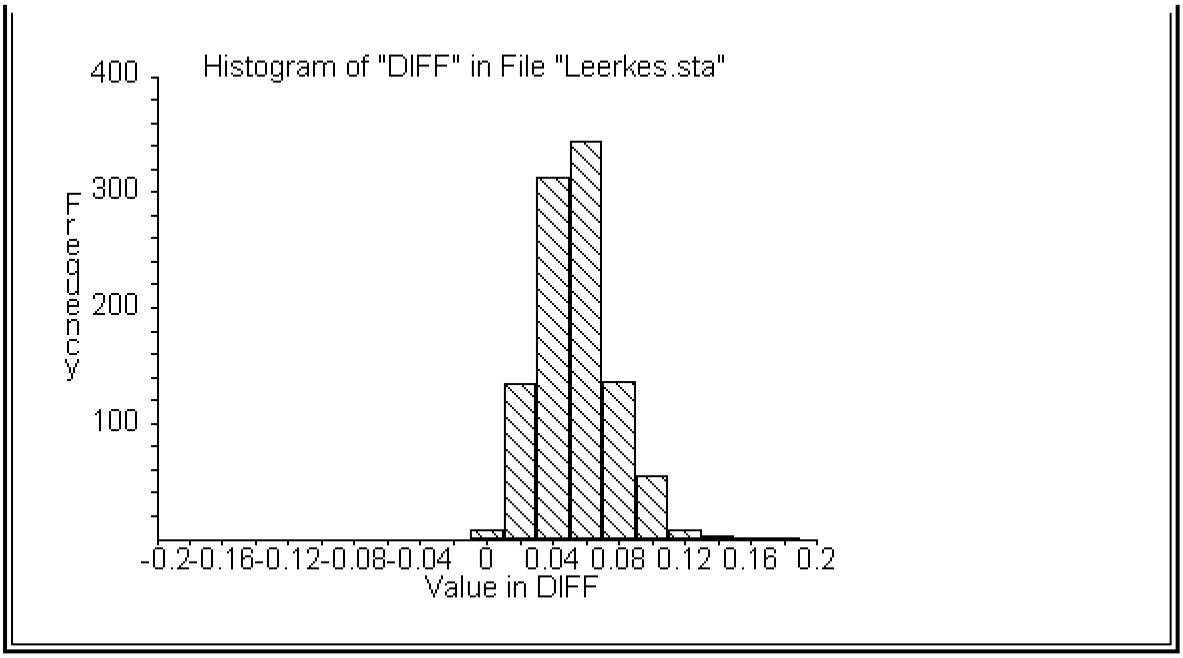
CIDIFF95 = 0.017117   0.10387

CIDIFF99 = 0.012092   0.11255

Vector no. 1: DIFF

```
        Bin                             Cum
     Center      Freq       Pct         Pct
                ---------------------------------------------
          0         8       0.8         0.8
       0.02       134      13.4        14.2
       0.04       312      31.2        45.4
       0.06       343      34.3        79.7
       0.08       136      13.6        93.3
        0.1        55       5.5        98.8
       0.12         8       0.8        99.6
       0.14         2       0.2        99.8
       0.16         1       0.1        99.9
       0.18         1       0.1       100.0
```

Note: Each bin covers all values within 0.01 of its center. Successful execution. (0.7 seconds)



This test leads to the same conclusion as the former test. The CI on the difference between the two coefficients does not include 0, and we can conclude that there is a reliable difference. This is actually a more direct test of what we were trying to do.

SAS apparently has procedures that will help you set this up as a macro, and, based on an e-mail exchange with SPSS, I suspect that SPSS is headed in that direction. At the moment it requires a bit of programming. I do not think that you could currently program this is SPSS.

Hosmer, D. W. & Lemeshow, S. (1989) Applied logistic analysis. New York: Wiley.

Return

Conclusion:

I read an article in the paper on Sunday about a company that makes small boats. The president of the company was saying that his predecessor had started the company and "thrown it as far as he could throw it." Then the current president had come along and thrown it as far as he could throw it. And now they were looking for a new CEO to pick it up and throw it the next step. In a sense I have done that with this particular class. I have deliberately chosen to end with a topic that is not covered in most similar courses, and I have "thrown" people ahead as far as I can. Now it's up to Heather Bouchey to toss it further. I can't think of a way that I would rather end.

And, Yes, this might be on the exam.

Last revised: 11/25/02