Exploring the Bootstrap

Questions from Lecture 1

Review of ideas, notes from Lecture 1

- sample-to-sample variation
- resampling *with* replacement
- key bootstrap analogy

Topics for today

More examples of "basic" bootstrapping

- averages (proportion is an average)
- How many bootstrap samples?

Two-sample tests and the bootstrap

- two ways to resample
- one is "random" and one is "fixed"?

Calibrating BS intervals

- use the bootstrap to check BS intervals
- lots of computing

Efficient methods

If I summarize with something other than the average, might the CI get shorter?

More Summer Program t-shirts

Comments on Software

R

Free, fast, familiar (at least here)

Capable of doing bootstrap resampling

Also has "prepackaged" resampling tools

Key things needed for resampling

Essential

- sample with replacement.
- iterate the calculation of a statistic
- accumulate the results

Very helpful

- add your own simple functions

Testing the Bootstrap

Pick a problem where we know the answer.

Confidence interval for the mean of a normal population based on a sample average.

$$\left[\bar{Y} \pm 2\frac{s}{\sqrt{n}}\right]$$

Simulate data from normal

R has random number generator.

- `rnorm` generates normals

Fun to see at the variety of normal samples.

Normal confidence interval

```
-0.39095771  0.05018254
```

Bootstrap interval

```
-0.38447221  0.03901192
```

Comments

Did't use the SE in finding the BS interval

SE "settles down" quickly, CI takes longer.

Analysis of Osteoporosis Data

Analysis of mean value

Classical and bootstrap results agree (Lec 1)

	SE	CI
Classical	0.16	[–1.7, –1.1]
Bootstrap	0.16	[–1.7, –1.1]

Conclude:
Population of postmenopausal women *on average* have relatively low bone mass.

How many have severe osteoporosis?

Define severe as t-score < –2.5.

Prevalence in sample
13/64 = 0.203, about 20%.

What about an inference to the population?

Analysis of sample proportion

Average of 0/1 indicator is proportion.

Standard analysis
SE = 0.05
95% CI = [0.10, 0.30] = 0.20±0.1.

Bootstrap Analysis

How to do the resampling?

As before, except keep track of the percentage of observations in each bootstrap sample whose t-score is less than –2.5.

Results are similar to classical

Bootstrap interval agrees with prior result, namely [0.10, 0.29]. (really close)

Plots of the bootstrap distribution of the estimator show the values that determine the bootstrap interval.

- A *kernel density* shows a smooth estimate of the shape of the population.
- The kernel density is not "tricked" into giving artificially large bins like histogram. (e.g. Some bins have several grid points.)

Bootstrap distribution agrees with the normal, and how closely the bootstrap confidence interval agrees with the classical result.

- Quantile plot shows similarity

What happens in a more extreme case?

Define severe as t-score < -3.

Now only 3 of the 64 are "severe"

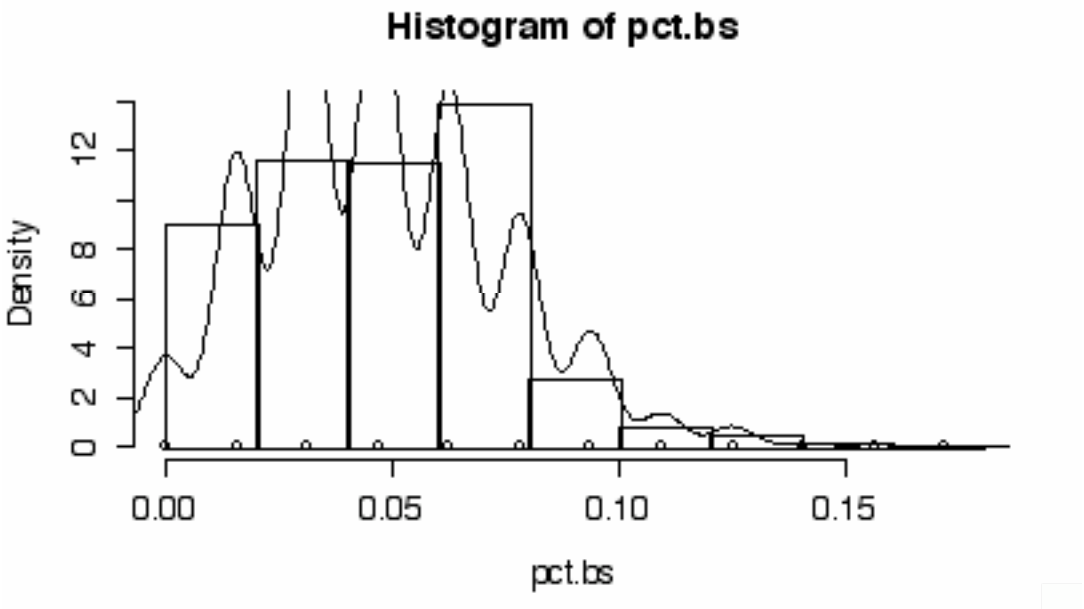
Classical interval is [-0.006, 0.100]

It has a *negative* limit!

Bootstrap in extreme case

Never gives a negative limit, here it gives the interval [0,0.11].

Plot of bootstrap distribution shows why standard interval fails: *skewness* in the distribution of the estimator!



Other procedures

- Standard interval for the proportion is

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Fails when the sample proportion is near zero since it can be negative.

- Other methods exist that guarantee the right coverage without needing the bootstrap.

Powerful notion

You learn a lot about your estimator by looking at its bootstrap distribution.

If the distribution is close to normal

- Use your favorite tool to check for normality
- If the bootstrap distribution is close to normal, then
- Bootstrap results usually resemble the classical results.

What About Estrogen?

Standard analysis

Two groups
44 who used estrogen, 20 did not

Users of estrogen have higher t-scores (less osteoporosis) than those who did not.

	Avg	*SE*	*CI*
use estrogen	-1.3	0.18	[–1.8,–1.1]
did not use	-1.5	0.27	[–1.9,–0.7]

What's the right test/CI to compare?

Some side issues to consider

- What about outliers?
- With outliers in the data, should we compare sample means?

Two-Sample Comparison

Two-sample t-test

Welsch interval

```
-0.4858549   0.8253224
```

Standard two-sample t-test

```
-0.4665447   0.8060122
```

The confidence interval for the difference in population means includes zero

- or the p-value is larger than 0.05

Conclude that the difference is *not* significant.

Can you do this with the bootstrap?

How to do the resampling?

Resampling is more complex

Rather than a single collection of numbers we have two sets of values: the t-scores *and* the responses to the question of estrogen use.

Key issue

Do we preserve the group sizes?
Are they fixed, or are they random?

Random resampling

Resample cases, not just the hip t-scores.

Bundle each subject's data together

Parallel the observational nature of data.

The data was gathered by sampling 64 women and learning that 20 had not used estrogen. Not experimental (i.e., chose to sample 44 who used estrogen.)

Alternatively

If we were to go back and get another sample, would we have to get 20 no's and 44 yes's? No.

Fixed (experimental resampling)

Suppose instead that the data had been gathered as part of an experiment that fixed the number of women in each group.

Then we should preserve this feature of the data in the bootstrap samples (i.e., sample from the two populations separately).

This issue returns in regression.

Bootstrap Results for Difference

Bootstrap differences: fixed sampling

```
     2.5%      97.5%
-0.447453  0.8232793
```

Bootstrap differences: random sampling

```
     2.5%      97.5%
-0.3974236  0.2971010
```

- Should the results be identical?

How many bootstrap samples?

How many seem needed for stability?
Would the results differ (in an important way) if someone tried to reproduce your analysis?

Crude rule of thumb:
200 for SE, 2000 for CI

Fun with Sample Sizes

How large to make the bootstrap samples?

The usual situation.

Use the original sample size if you want to understand the properties of a statistic computed on samples of this size.

Use other sizes to play "what if" games.

- e.g. power calculations

What might happen with a much larger sample?

Would the comparison be significant if I had more data, say 1000 observations rather than just 64?

Investigate with larger bootstrap samples.

Beware of assumptions in case your small sample is not representative.

Osteoporosis comparison

Generate BS samples of 1000 observations.

How many do you need in order to find a significant different… a "bootstrap power analysis".

Robust Estimators

Outliers and estimators

What should we do about an outlier?

- Ignore it
- Remove it
- Compromise

Why use the sample average if it is so sensitive to the presence of outliers?

- It is the standard in all the books
- Optimality properties
- We know how to get a SE and CI.

Robust statistics offer insurance

- Compromise choice
- Results in an automatic downweighting of unusual observations.

Choice of an estimator

Which is best: mean, median, or mode?

Optimal depends on shape of population.

Mean optimized for normal.

Median tolerates outliers, but inefficient.

Trimmed Means

Definition

Trimmed mean is mean of the "middle"

- drop lowest 5%
- average the middle 90%
- drop the highest 5%

Special cases

- median = 50% trimmed mean
- average = 0% trimmed mean

Compromise as insurance against outliers.

Influence functions –you can read more

Influence functions graph how the value of a statistic depends upon the location of the data values.

- Mean's influence function is linear
- Median is step function.

Trimmed mean is compromise

- Center of data is normal, so linear there
- Extremes have outliers, so count

Other choices are obvious (biweight).

Picking a Robust Estimator

If robust estimators are so good, why don't more people use them?

Have not been taught about them.

They believe in the normality of data.

If the mean gives you an answer you like, why use something else?

The mean is easy to work with since you have simple formulas for both SE and CI.

Which estimator should I use?

You'd like to use the one that gives the shortest CI (or has smallest SE).

But, this sort of optimality depends on the population, which you don't know.

Bootstrap to the rescue…

Use the bootstrap to get a confidence interval for whatever estimator you want to use.

Use the bootstrap to see if the choice matters much in your problem.

Bootstrap Comparisons

Key feature of comparisons

Apply each estimator to the *same* bootstrap samples.

Which is most variable? Least variable?

Which has the widest interval? Shortest?

Doing it

Just need three "containers" to hold the different bootstrap replicated estimators.

Design of simulation
Use the same bootstrap samples for each!

Once you've generated the bootstrap replicates of the estimators, compare them using standard tools.

Do enough replications so that you believe that the results are not going to materially change if you were to repeat the resampling simulation.

Comparison Results: Osteoporosis

Data (different data from in class)

Just work with the 40 in the "yes" group.

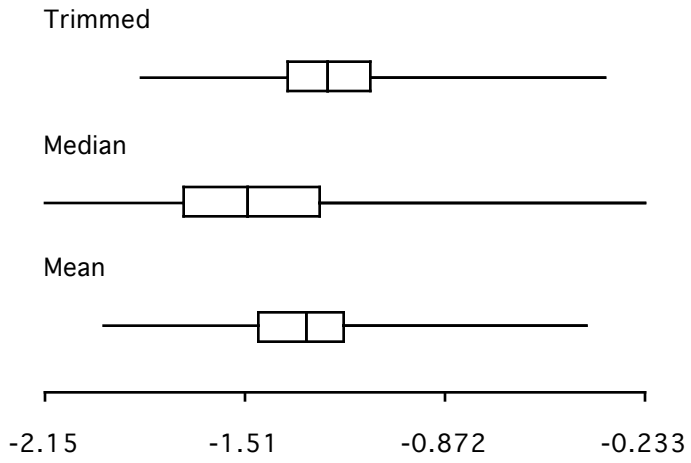
See if an outlier in that group has much effect.

Comparison boxplots in AXIS

Use the "compare" command

Use the "vertical" option.

Summary



Trimmed mean has the smallest SE.

Median has largest SE.

Take a more detailed look in class using more features of the compare command window.

Calibrating Bootstrap Intervals

How can you tell if the bootstrap intervals are working?

- Not so easy to make a test case as it was for the familiar problem of the average of normals.

- But hey, the bootstrap is a method for evaluating any statistical procedure, so…

The double bootstrap (a.k.a., calibration)

- Bootstrap of a bootstrap.

- Evaluate how well the bootstrap interval covers in the one place where the population is known: sampling from the sample.

- Idea: its just a simulation…
  - (a) Outer layer draws a sample from the sample (say, X*)
  - (b) Inner layer computes the boostrap interval from this sample, call it I(X*).
  - (c) Outer layer checks to see how many of the I(X*) intervals cover.

You only thought you had a fast computer!

Review Questions

How large should my sample be in order to use the bootstrap?

More is always better when it comes to sample size. If you have a small sample though, the theory shows that the bootstrap makes better use of the data than traditional methods (e.g., t-test/interval).

How do I decide how many bootstrap replications are needed?

You need enough so that the results are not materially changed when the bootstrap simulation is repeated.

Why bother to look at the bootstrap distribution? Isn't the confidence interval enough?

The bootstrap distribution will show you, for example, how well the simulated distribution of your statistic matches normality. In most cases, the closer the bootstrap distribution comes to normality, the closer the "standard" results match those from the bootstrap.

Does the size of each bootstrap sample have to match the size of the original sample?

No, and in fact using other sample sizes lets you explore alternative scenarios, such as what level of significance you might have seen with a larger sample (as an alternative to the standard power calculations).

What choice did we consider for doing the resampling needed for the two-sample test?

We considered two plans: one that fixed the number of women in the two groups, and another that allowed the group sizes to vary with the samples.

How can a trimmed mean be a better estimator of μ than the sample average?

The sample average is the "best" estimate of μ available, if the sample comes from a normal population. If the population's not normal, the sample average is really poor. The trimmed mean is a compromise, based on the premise that data is normal in the middle, and prone to rogue outliers at the extremes.

The trimmed mean limits the effects of outliers. If the data are not normal, the sample average

can be a very poor estimator with high sampling variation.

What's the problem with removing outliers when editing data prior to an analysis?

The confidence interval that you construct after editing the data does not know that the sample is not a real sample, but rather an edited sample. Perhaps your actions should lead to a longer interval, since the data have been manipulated. The usual interval assumes pristine data and does not account for the editing process.

How can you decide if an alternative estimator is going to produce a smaller standard error?

Bootstrap comparisons are one method: compare the bootstrap distributions to see which estimator has the smallest sample-to-sample variation (std. error).

What does the influence function tell you? What is the influence function for the average?

Influence functions describe how points affect an estimator. For example, the influence function for the mean is a line, indicating that

the most influential points are those farthest from the center of the data.

What's an outlier and where do they come from?

Outliers come from many sources, and it's hard without careful analysis and complete records to know. Some come from typing errors. Others appear because they are part of the process being measured: some observations are simply very different from the others.