Introduction to Bootstrap Resampling

Bob Stine

University of Pennsylvania
ICPSR, June 29, 1998

1

Preliminaries: Keeping in touch

- Mail
  Bob Stine
  3014 Steinberg-Dietrich Hall
  Department of Statistics, Wharton School
  University of Pennsylvania (Wharton)
  Philadelphia, PA 19104-6302
- Electronically
  stine@stat.wharton.upenn.edu
  http://www-stat.wharton.upenn.edu/~bob

2

Research interests

- Model selection
  - Picking the right variables for a large regressions
  - Large means 10,000 or more
- Time series and prediction
  - Resampling for dependent data
  - Judging the uncertainty in forecasts
- Statistical software design and graphics
  - Perception in 3-D graphics

3

Special things for the summer

- Office in 214 HN, but more often getting coffee
- Software
  - Lisp-Stat, AXIS (See Stine & Fox 1997)
  - JMP-IN (statistical spreadsheet)
- T-shirts
  - I have lots more, so we'll see a few each day.
  - You have to guess how many I have!

4

Illustrative examples

- Is a special summer reading program effective?
  - Children's test scores (n=30)
  - Data are pre/post intervention
- Is justice becoming more swift, or slower?
  - Time from referral to completion of prosecution for FBI districts (n=90)
  - Data are for two years, 1995 and 1996

5

Evaluating the education program

- Evaluating the education program 30 students in data file EducScores.dat
- Initial comparison statistics

	Average	Change
pre	102	
post	112.7	10.7

- Typical hypotheses

  H_0: $\mu_{pre}=\mu_{post}$ –or– $\mu_{pre}-\mu_{post} = 0$
- Two sample t = 0.88. What's wrong here?

6

Evaluating education program, cntd

- Work with differences since matched pairs.
- Standard error of average = 5.5
  paired t p-value is 0.06.
- Rough 95% CI includes zero:
  $10.7 \pm 2\ (5.5) = [-0.3, 21.7]$
- Zero inside: no significant impact.
- What assumptions are required for the validity of our conclusions from these statistics???

7

Length of time for prosecution

- 90 FBI districts (prosecute.dat)
- Number of cases unchanged, with
- 31 day increase in average time for prosecution.
- Two-sample analysis (t = 1.4) inappropriate
  - because of the pairing.
- Within-district change: t = 2.31, significant.
- t-interval (SE = 13.6) [4.4 – 58]
  ("t-interval diff 0.95" typed in evaluation box)

8

Assumptions, assumptions

- Interval does not include zero, so significant
- Validity?
- Assumptions of t-test
  - Independence
  - Constant variance (equal precision)
  - Normality (central limit theorem)
  - Sampling done properly

9

Running some checks

- Symmetry plots
- Normality
  - QQ plots
  - Kernel density estimates
- What about constant variance?
- What about independence?

10

Research method: hard part

1. Question about some group or phenomenon
   "Did instruction work?"
   "Is justice becoming slower?"
2. Convert question
   "Are test scores higher?"
   "Are times for prosecution lengthening?"
- "Threats to validity" (Campbell & Stanley)
3. Gather data (sample)

11

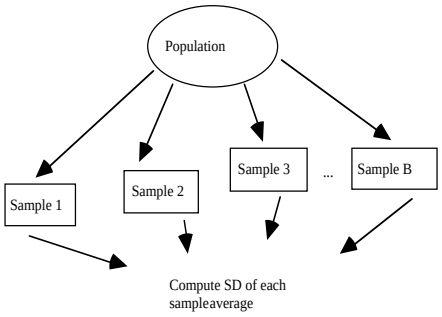
Research method: easy part

4. Compute statistics
   - Get data into your favorite package
5. Make an inference, description…
   - Obtain standard error
   - Get a confidence interval
   - Get a p-value
   - Decide how many stars to add for the journal

- What are the origins of methods for inference?

12

Idealized sampling model



13

Mathematical models

- Existential experiment + math implies
  $X_i \sim N(\mu,\sigma^2) \Rightarrow$ X-bar $\sim N(\mu,s^2/n)$
- Mathematical model of sampling variation
  - Describes sample-to-sample variation
  - Standard error SE = SD/√n
  - Confidence interval average ± 2 SE
- Assumptions are there for the convenience of the mathematics
  - With computers, we can reproduce the mathematics

14

Monte Carlo simulation alternative

- Replace the existential mathematics with real samples generated on a computer.
- Computer generated samples
  - Possible to draw "samples" from *specified* population
  - Interesting way to train yourself:
    What do normal samples look like?
- Set up utopian data $X_i \sim N(0,1)$
  - Create new data set, add a new icon
  - Name icon "norm" and define a formula
- Explore variation among normal samples.

15

Checking the simulation...

- See how well the simulation reproduces something we know
  - Averages of normal samples are normally distributed around μ with dispersion given by usual standard error
- Use the "bootstrap" command in AXIS with
  - estimator = mean
  - sampling rule = normal-rand n
  - number of trials = as many as you want

16

Simulation results

- Generate a new data set with simulation results
  - Variable contains simulated mean values
  - One mean for each simulated sample
- Results
  - Standard deviation of simulated sample *is* the simulated estimate of the standard error.
  - SE agrees with usual formula (σ^2/n)
  - Sampling distribution of estimator matches that predicted by theory.

17

What about in the "real world"

- What to use for the population?
  - Can't use a normal population since don't know the population.
- Alternative methods for getting standard errors
  - Nonparametric methods
    - based on ranks, not applicable to many problems
  - Jackknifing
    - Tukey's 1958 abstract
    - Re-compute statistic, leaving out one each time
    - Does not generalize well (Fails for the median)

18

Bootstrap approach

- Resample the observed data with replacement to compute the sampling distribution.
- Make the existential experiment real
  - Avoid pretending you know the population
- Use your best estimate of the population.
  - Sample data is all you know about population.
  - "Sample the sample"

19

Key analogy

- Behavior of statistic when sampling from true population (do not observe) is similar to the behavior of the statistic applied to samples from the data (do observe).
- Sampling with replacement (resample function)
  - (1,2,3,4,5,6) --> (2,5,3,3,1,5)
  - Many, many samples possible

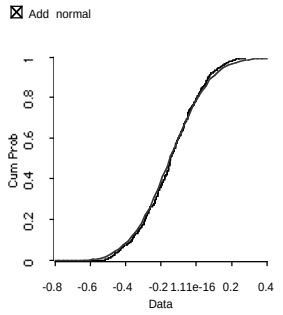
20

Bootstrap in a normal problem

- Use normal sample "norm"
  - Be sure to "convert to values"
    (double click icon & "convert to values")
- Normal assumptions + mathematics
  - SE(sample mean of "norm") $\approx s/\sqrt{n} = 0.17$
  - 95% confidence interval "t-interval norm 0.95"
    [avg + t × se] = [-.151 , .540]
- Settings in BS dialog:
  - estimator $\Rightarrow$ mean, sampling rule $\Rightarrow$ resample norm

21

Bootstrap reproduces normal theory



22

Bootstrap the mean for normal

- Each BS sample has 30 obs, just like original.
- SE^*(sample mean) = 0.17
- Bootstrap confidence interval from bootstrap replicates
  - Message ":percentile-interval 0.95"
  - 95% percentile interval = [-0.159, 0.513]
  - interval formed by extreme 2.5%, 97.5%.
- Virtually identical to classical SE/CI!

23

Application to education data

- Pairing reduces to a one-sample problem
- Compute bootstrap standard error and CI
- Use lots of trials for the CI (B ≈ 1000)
  - Classical 5.5 [–.5, 22]
  - Bootstrap [,]

24

Bootstrap = perspective

- Key analogy is fundamental, not the computing
- Sampling from the sample resembles the process that generated the original data.
- Independent observations, with equal variance
  - Truth of these in our two examples?

25

Role for mathematics remains

- Mathematics shows what the bootstrap does for the sample mean.
  - Can derive bootstrap SE for x-bar in same way that its done in the text for usual sampling problems.
- Averages are easily handled with algebra
  - Regression estimates are classically treated as averages
- Other, often better estimators, are not so simple.
  - Many interesting, useful estimators are not averages

26

Handouts

- Optimistic syllabus
  - Try to look at the questions.
- Even more optimistic bibliography
  - Many publications in literature last year.
- Notes
  - Distribute PDF files on server: f:\bootstrap
  - Last year's version are on my Web page
  - This year's will replace them.

27

Summary for today

- The bootstrap produces reliable standard errors, CI's for virtually any estimator.
- Relies on repeated resampling of the data.
  - sampling with replacement
- Resampling parallels the original DGP.
- Frees time to think about problem rather than worry over math.
- Works reliably if resampling done properly. (e.g. paired vs two sample method)

28

Next time

- Reaching out for new methods that offer better solutions for old problems.
- Since bootstrap takes care of finding the SE and offers a CI, we can use other estimators even if we lack the standard formula.

29

Review Questions

1. What crucial assumptions underlie bootstrap resampling? Analogies?
   Resampling the data parallels sampling the population. Key analogy is ... $m:\mu$ as $m^*:m$
2. Does the bootstrap provide a new method for estimation? Testing?
   No. It is a method for assessing a statistic, allowing more flexibility in your choice.

30

Questions, continued

3. How does the bootstrap differ from simulation?
   BS resamples the data rather than a hypothetical population. Parametric BS.
4. Do you need a computer to bootstrap?
   No. It's better thought of as a point of view rather than a computing method.
5. How do you get many samples from one?
   Sample WITH replacement.

31

Questions, continued

6. How many samples does the bootstrap require?
   As a rule of thumb, about
   200 for SE and
   2000 for intervals
   depending on how far into the tails you want to look. The main idea is to be confident that the results would not be different in a meaningful way if you repeated the BS simulation.
7. Where do these names come from?
   Jackknife is Tukey's name for a rough-and-ready tool; bootstrap from " to pull oneself up by the bootstraps".

32