Introduction

Contacting me

Mailing address

Bob Stine
444 Huntsman Hall
Department of Statistics, Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340

Electronically

stine@wharton.upenn.edu
http://www-stat.wharton.upenn.edu/~stine

Some background

Research interests

- model selection (a.k.a., data mining)
- time series, accuracy of prediction
- resampling methods

T-shirts

Software

- Lisp-Stat (Stine & Fox 1997)
- R (the *free* version of S-plus)

Overview of Lectures

Syllabus for a short week

- Optimistic
- Try to look at the review questions.

Bibliography

- Even more optimistic!
- Resampling has proliferated since these.

Other topics for out-of-class conversation

- Bootstrapping with missing data
- Handling measurement error in regression
- Structural equation models
- Logistic regression and other GLIMs

Lecture Notes

- On my web page at Penn.

  www-stat.wharton.upenn.edu/~stine/mich

Inference

Confidence intervals

- Statements about population features
- Main ingredients: std. error and t-value.

What is standard error all about?

- Sample-to-sample variation…
  How much would my statistic change if I got a second sample?
- Hard part about SE:
  You'll only ever have one sample.

Why is the t-value in the confidence interval?

- Determines how big is big enough…
  Values within about 2 SEs of the observed statistic are "close"

Today, you get to see standard error and let the computer figure out how many std errors are needed for a deviation to be large…

Illustrative Research Question

How can one answer these questions?

- Is osteoporosis present in older women?

- Does taking estrogen affect osteoporosis?

Population and sample

- Gather a sample of *n* = 64 post-menopausal women from patients at various clinics.

- Measure osteoporosis by hip x-ray, converted to a "t-score":
  "young normal" has mean 0 and SD 1

- Darn, this is an *observational study*

Data analysis

- Descriptive statistics via standard software

- Initial summary statistics

Mean t-score	–1.41
Standard deviation	1.2

Classical Approach to First Question

Tests and confidence intervals

Test the null hypothesis that the population mean level of osteoporosis is less than 0.

H_0 : □ □ 0

Form a confidence interval for □

See if zero lies inside the interval

How are these things done?

Standard error and normality

Form a confidence interval as
(sample avg) ± 2 (standard error of avg)

Standard error measures precision
How far do you expect the sample average to be from the population value?

Confidence interval requires
mean is approximately normally distributed

Standard package output gives

SE = 0.15

95% CI is [–1.7, –1.1]

Role of Statistics in Research

Research paradigm

1. Question about something (population)
   "Do older women have lower bone mass?"

2. Convert question (operationalize)
   "Is the mean t-score less than zero?"

3. Gather data (sample)

4. Compute statistics

5. Draw an inference, reach a conclusion...
   "Conclude population mean < 0."

Meta-questions

- Distinction between concept and data.

- Validity of data for the question at hand

- "Threats to validity" (Campbell & Stanley)

Where do you get the standard error and CI?

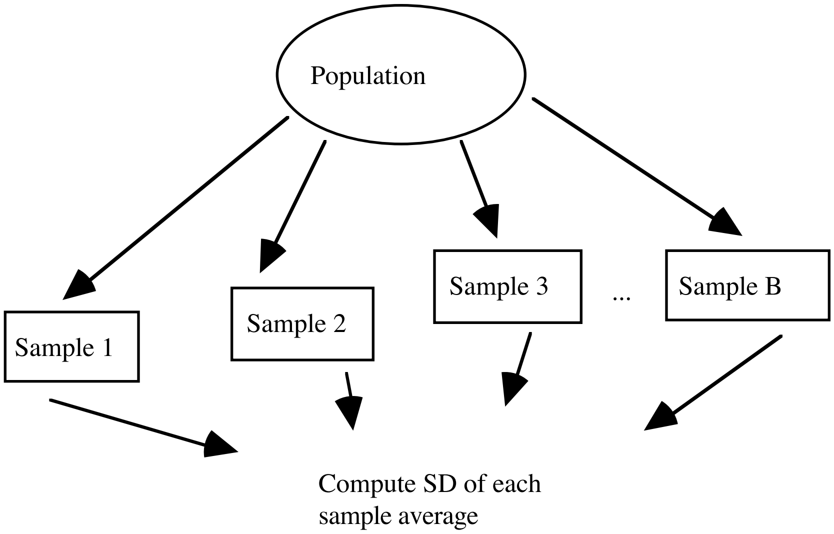
What's the role for bootstrap resampling?

Statistical Inference

Models and assumptions

Standard methodology (t-test) assumes
- Independent observations
- Constant precision (equal variance)
- Normal population

Sample taken correctly, right population



Existential experiment + math implies

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Mathematical model of sampling variation
- Describes sample-to-sample variation
- Standard error SE = SD/$\sqrt{n}$
- Confidence interval avg ± 2 SE

Alternatives to Classical Methods

Simulation

Use computer to sample some population.

- Sample from a normal population
- Calculate average from each

Advantage: See how your statistic behaves.

Problem: What to use for the population?

Other methods for inference

Nonparametric

- based on ranks when you can do it.
- sacrifice power for *robustness*

Jackknife

- Tukey's 1958 abstract for estimating SE
- Re-compute statistic *n* times
  Leaving out one each time
- Does not generalize well
  Fails for the median
- An approximation to the bootstrap

Bootstrap

Resample the observed data repeatedly and compute the sampling distribution.

Bootstrap Resampling

Return to fundamental problem

Make the existential experiment real

- parallel the notion of repeated samples.
- observe sampling variation

Avoid pretending you know the population by directly using the data itself.

Key idea

Sample is your best estimate of population:
"Sample the sample"
Let the sample define a population with equal probability on each observed value.

Key bootstrapping analogy

The behavior of the statistic when applied to samples from the sample is analogous to its behavior when applied to samples from the original population.

Mathematics

Role is to describe when the analogy holds.

*So, how does one get **repeated** samples?*

Bootstrap Sampling and Notation

Constructing bootstrap samples

- Sample with replacement (`sample`)
- (1,2,3,4,5,6) □ (2,5,3,3,1,5)
- Many, many samples are thus possible

Original process

Population □ $(x_1, x_2, ..., x_n)$ □ $\bar{x}$

Resampling process

BS Sample 1: $(x_3, x_7, ..., x_2)$ □ $\bar{x}^*_1$

BS Sample 2: $(x_8, x_1, ..., x_1)$ □ $\bar{x}^*_2$

....

BS Sample B: $(x_4, x_9, ..., x_{11})$ □ $\bar{x}^*_B$

How many bootstrap samples are needed?

Bootstrap Results

Bootstrap a normal sample

Check to see that the bootstrap works when we know the right answer

Use a sample from a normal population.

Simulate normals (`rnorm`) in R

What are the right answers?

- Normal assumptions + mathematics
- SE(avg of n) ≈ SD/$\sqrt{n}$ = .12
- 95% confidence interval

[avg + t □ se] = [-.07, .42]

Bootstrap B=1000 times

Each BS sample has 64 observations, just like the original sample. We find from these that

- SE*(sample average) = 0.12
- 95% percentile interval = [-0.07, 0.42] (interval formed by 2.5% and 97.5% of BS replications)

Virtually identical to classical SE and CI!

Bootstrap Analysis of Osteo Data

Bootstrap analysis

- Read the data (read-table in R)
- Use lots of trials for the CI (B ≈ 1000)

Comparison to classical

	SE	CI
Classical	0.15	[–1.7, –1.1]
Bootstrap	0.15	[–1.7, –1.1]

R commands

Found in the file “Lecture1.R” on my web page. Hopefully, you can follow along later and reproduce the examples that we have done in class.

Discussion

Computer is not really needed

- Bootstrapping is a perspective, not computing
  Computing is getting faster and easier!
- Key analogy is fundamental
  The resampling must reproduce nuances of the original sampling process.

Assumptions remain

- Independent observations
- Equal variance (i.e., equally precise)

Bootstrap algebra

Role for mathematics remains

- Usual derivation of std. error expression
- Bootstrap derivation parallels usual.

Going further

Why use an average?

Averages offer easy algebra.

Other estimators are not so simple to analyze theoretically, but you can handle them easily using the bootstrap.

Things to Take Away

Bootstrap resampling does

Produces reliable standard errors and CI's for virtually any estimator.

Relies on repeated resampling of the data.

Presumes that the resampling parallels the original data generating process.

Frees time to think about the substantive problem rather than technical distractions.

Bootstrap resampling does not

Work if resampling done improperly.
(e.g. paired vs two sample test)

Make good things happen with bad data.

Fix flaws in your research paradigm.

NEXT TIME...

- More types of problems, questions.
- Questions of efficiency and diagnostics.
- Fancier bootstrap methods in R.

Some Review Questions

What are the crucial assumptions and analogies?

Resampling the sample data parallels the original sampling of the population.

Key analogy is ... $\bar{Y}^* : \bar{Y} :: \bar{Y} : \square$

Does BS provide new methods for estimation?

No. It is a method for assessing a statistic, which then often allows more flexibility in your choice of statistics to use.

How does BS differ from classical simulation?

Bootstrapping resamples the observed data rather than a hypothetical population.

Do you always need a computer for BS?

No. It's better thought of as a point of view rather than a computing method.

How do we get so many samples from one?

Sample WITH replacement.

How many bootstrap samples are needed?

As a rule of thumb, about 200 samples are needed for finding a standard error whereas on the order of, say, 2000 are needed for intervals (depending on how far into the tails you want to look). Some recent results suggest you can do well with many fewer (Ed Carlstein, UNC)

The crux is that you want to be confident that the results would not differ in a meaningful way if you were to repeat the BS simulation. (See Efron and Tibshirani's book as well as the Booth and Sarkar paper.)

How does the jackknife differ from the bootstrap? Where did these names originate?

The jackknife is a rough approximation to the bootstrap, obtained by leaving each observation out, one at a type (usually). The name is Tukey's idea for a rough-and-ready tool; bootstrap comes from the expression to pull oneself up by the bootstraps.