Bootstrap Resampling

SPIDA
Toronto June, 2005

Bob Stine
Department of Statistics
The Wharton School of the University of Pennsylvania

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

Plan for Talk

- Ideas
  - Bootstrap view of sampling variation
  - Basic confidence intervals and tests
- Applications
  - More ambitious estimators
  - Survey methods
  - Regression
  - Longitudinal data
- Moving on
  - Better confidence intervals

Truth in Advertising

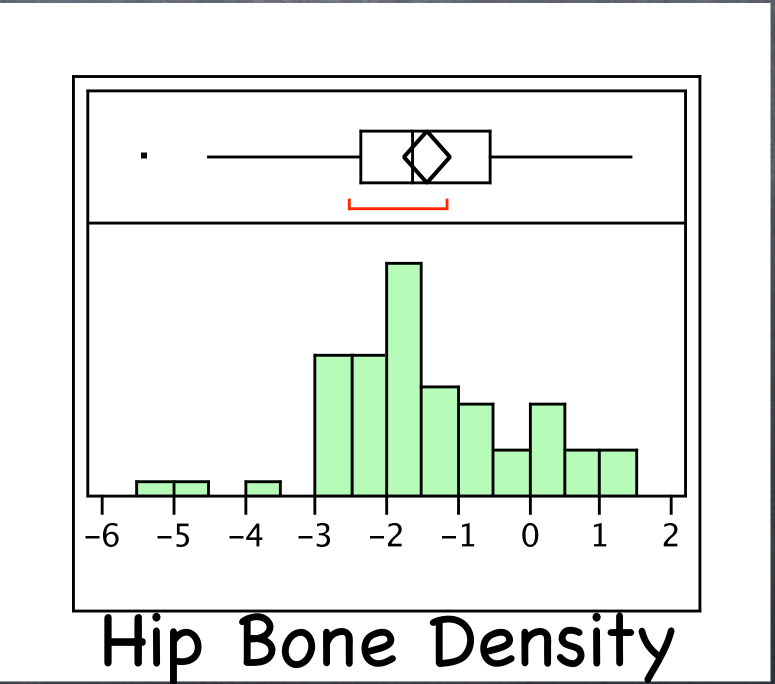
- Emphasis
  - Wide scope
  - Pique your interest
- Background
  - Time series modeling
  - Developed bootstrap-based method to assess the accuracy of predictions
- I've become a data miner
  - Build predictive models from large databases
  - Objective is prediction, not explanation

Research Question

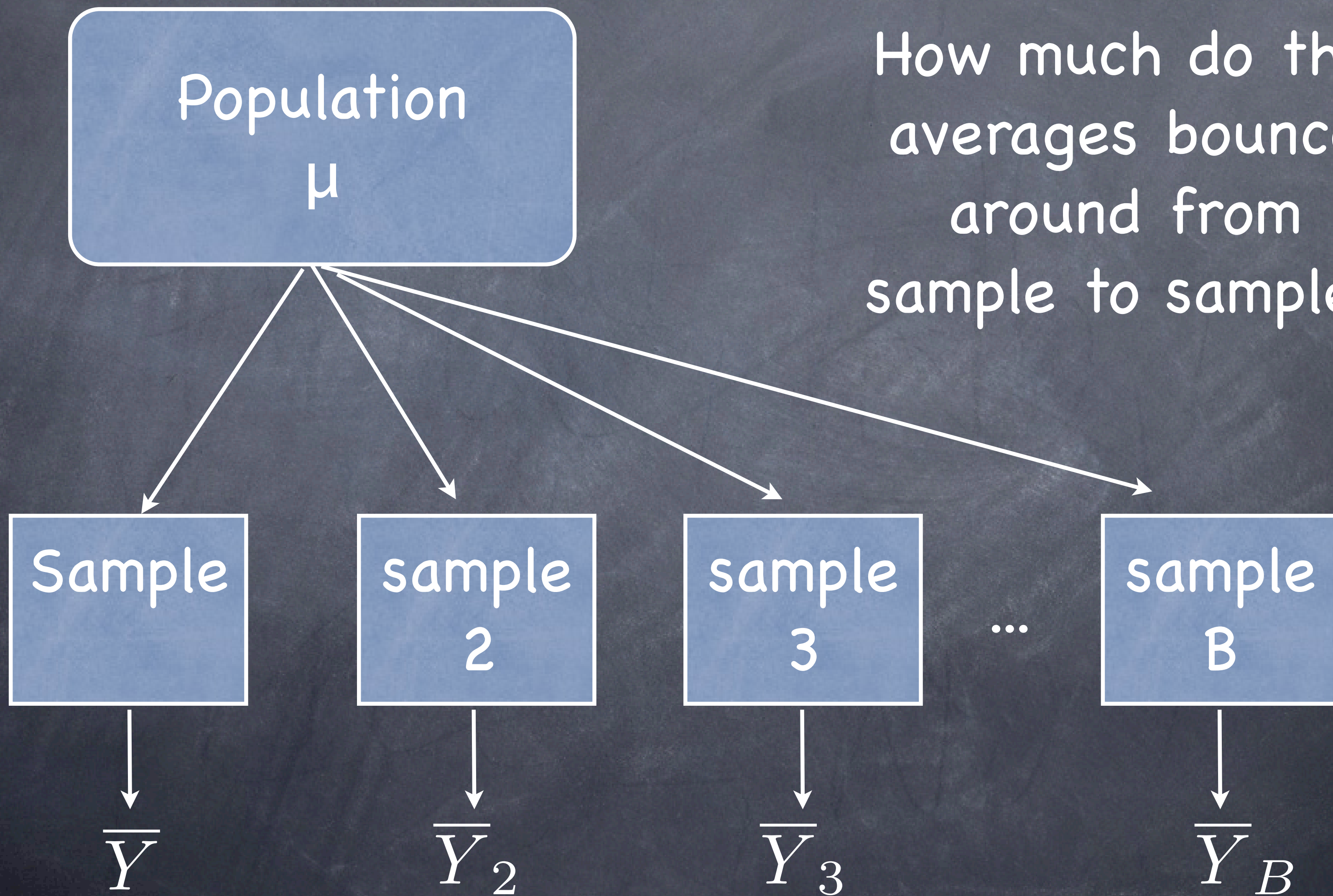
- Osteoporosis in older women
  - Measure using X-ray of hip, converted to a standardized score with ideal mean 0, sd 1
- Sample of 64 postmenopausal women
- What can we infer about other women?

Y-bar = -1.45

s = 1.3

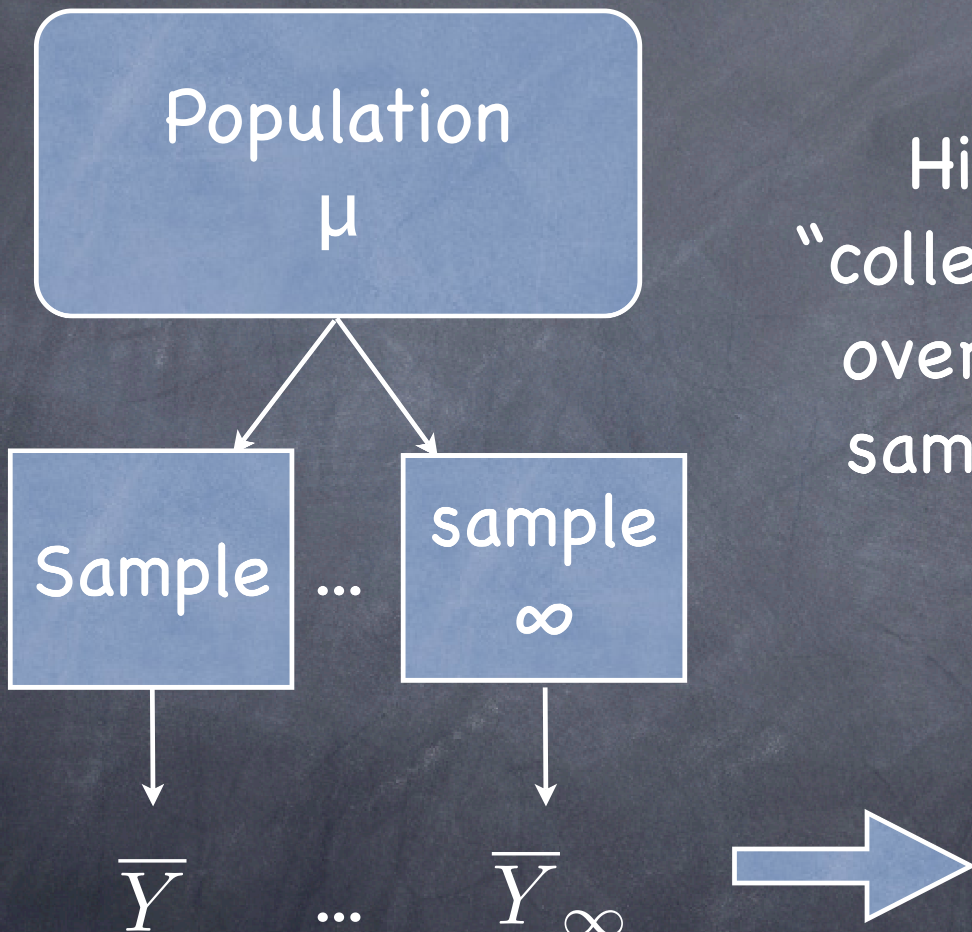


Statistical Paradigm

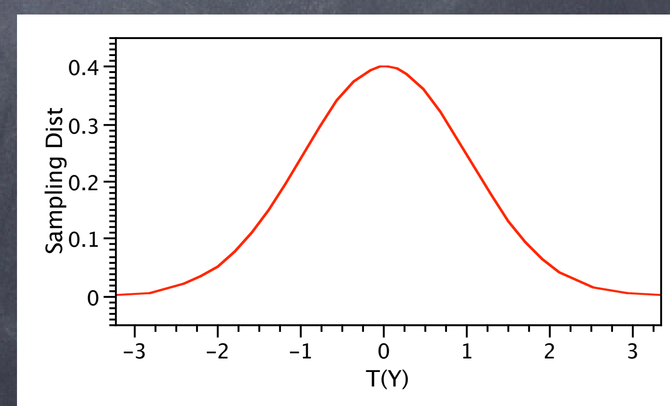


How much do the averages bounce around from sample to sample?

Sampling Distribution



Histogram of the "collection" of averages over samples reveals sampling distribution



Notation

- Data
  - Observe sample $Y = Y_1,\ldots,Y_n$
  - Y_i iid sample from population F_θ
  - θ = population parameter
- Statistic
  - $T(Y)$ = statistic computed from data Y
  - Estimates θ
- Sampling distribution
  - G_θ is sampling distribution of $T(Y)$

Using Sampling Distribution

- Hypothesis test
  - Sampling distribution G_θ implies a rejection region under a null hypothesis
  - Under H_0: $\theta = 0$ then

    $Pr(G_0^{-1}(0.025) \leq T(Y) \leq G_0^{-1}(0.975)) = 0.95$
  - Reject H_0 at the usual $\alpha=0.05$ level if

    $T(Y) < G_0^{-1}(0.025)$ or $T(Y) > G_0^{-1}(0.975)$
- Confidence interval
  - Invert test: CI are those θ_0 not rejected

What Sampling Distribution?

- Classical theory
  - Based on idea that averaging produces normal distribution, and most statistics are averages of one sort or another
  - "Asymptotically normal"
- Monte Carlo simulation
  - Pretend we know F_θ, and simulate samples from F_θ under a given value for θ
  - Repeat over and over to construct sampling distribution for estimator

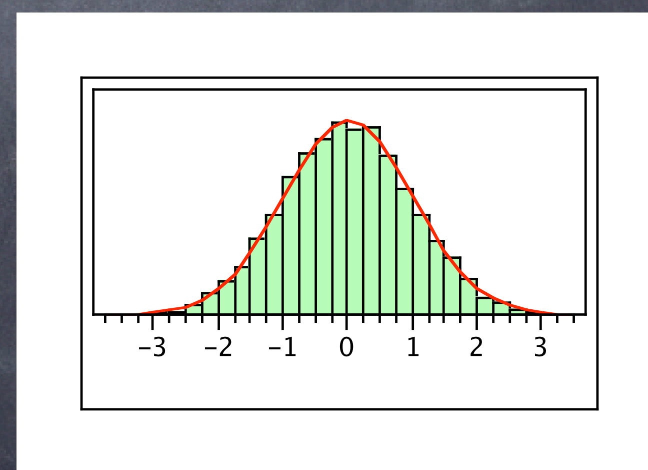
Simulation

Chosen population θ_0

Sample ... sample B

$T(Y)$ ... $T(Y)_B$

Histogram of averages over samples simulates sampling distribution under H_0



Limitations

- Classical theory
  - Works very nicely for averages, but...
  - Easy to find estimators for which it is quite hard to find sampling properties
  - Example: trimmed mean
- Simulation
  - How will you know the shape of the population when you don't even know certain summary values like its mean?
  - What is the distribution for hip X-ray?

Bootstrap Approach

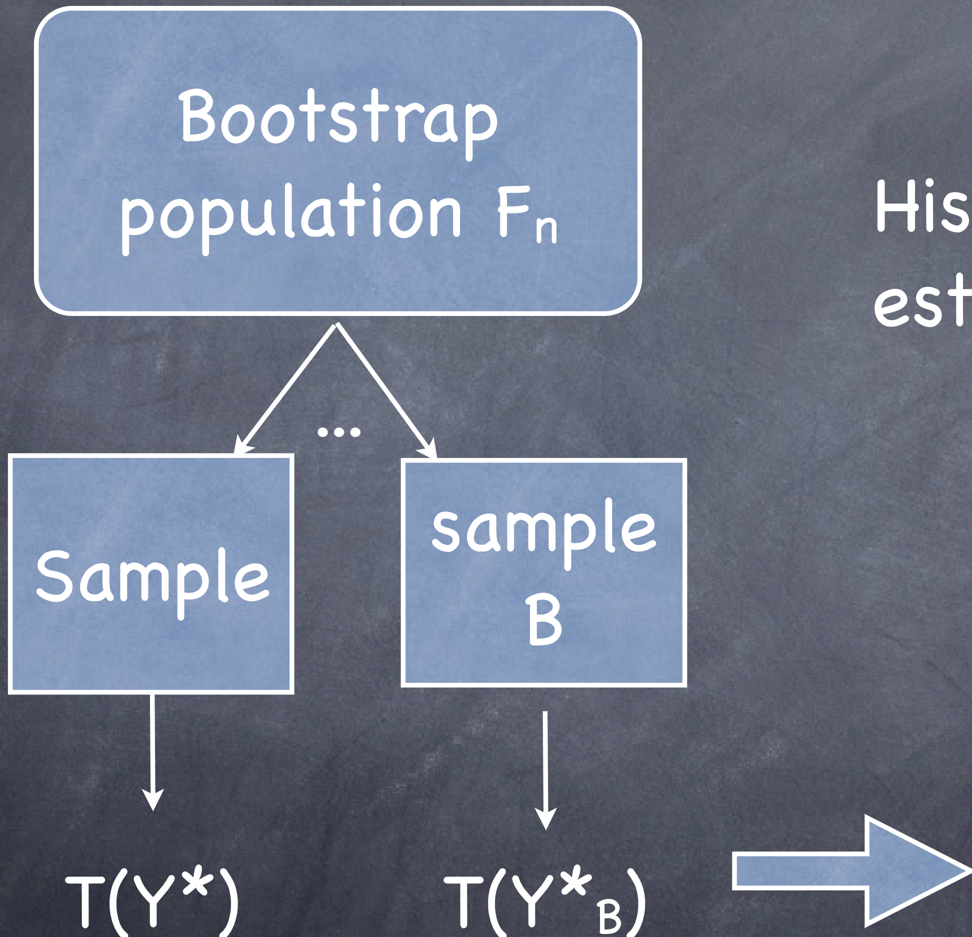
- Let the observed data define the population
  - Rather than think of $Y_1,\ldots,Y_n$ as n values, let these define the population of possible values
  - Assume population is infinitely large, with equal proportion of each Y_i
- Data define an empirical distribution function
  - F_n is the empirical distribution of $Y_1,\ldots,Y_n$

$$F_n(y) = \#\{Y_i \leq y\}/n$$

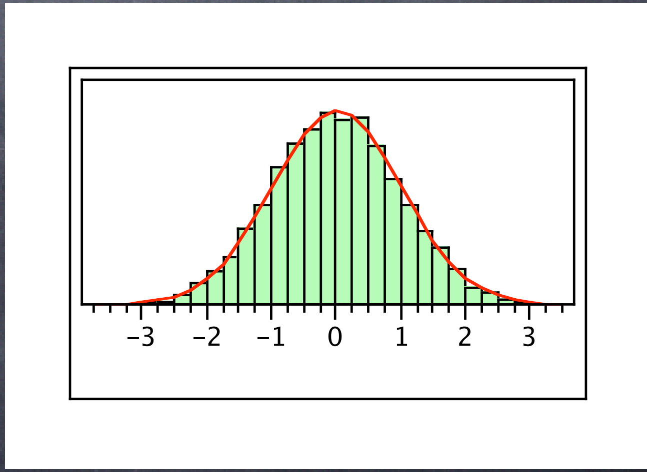
  - If Y^* is a random draw from F_n, then

$$P(Y^* = Y_i) = 1/n$$

Bootstrap Sampling Distribution



Histogram of $T(Y^*)$ estimates sampling distribution



Comments

- Bootstrap does not have to mean computing
  - All we've done is replace F_θ by F_n
  - No more necessary to compute the sampling distribution in the bootstrap domain than in the usual situation
  - But its a lot easier since F_n observed!
- There's no hypothesis nor parametric assumptions to constrain F_n in what we have at this point
  - Not hard to add that feature as well

Bootstrap is Max Likelihood

- Without assumptions on continuity or parametric families, the bootstrap estimates the population using F_n
- Empirical distribution function F_n is the nonparametric MLE for the population CDF
- Connection to MLE shows up in various ways, such as in variances which have the form

$$\Sigma x_i^2/n$$

rather than

$$\Sigma(x_i^2)/(n-1)$$

Osteoporosis Example

- Average hip score -1.45 with SD 1.3, n=64
  - Standard error of average = $s/\sqrt{n}$ = 0.16
  - Classical t-interval assuming normality
    -1.45 ± 0.32 = [-1.77, -1.13]
- Bootstrap approach
  - Bootstrap standard error is "usual formula"
    $$\begin{aligned} Var^*(\bar{Y}^*) &= Var^*(Y^*_1 + \ldots + Y^*_n)/n^2 \\ &= Var^*(Y^*_1)/n \\ &= n/(n-1)\, s^2/n = 0.162^2 \end{aligned}$$
  - Confidence interval?
  - Shape of sampling distribution?

Bootstrap Sampling Distribution

- Draw a sample $Y^*_1, \ldots, Y^*_n$ from F_n
  - Easiest way to sample from F_n is to sample with replacement from the data
  - Bootstrap samples will have ties present, so your estimator better not be sensitive to ties
- Compute the statistic of interest for each bootstrap sample, say $T(Y^*)$
- Repeat, accumulating the simulated statistics in the bootstrap sampling distribution.

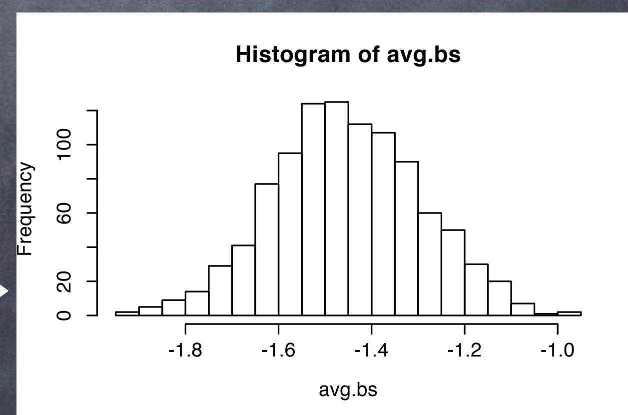
Bootstrap Sampling Distribution

Osteoporosis sample F_n

Sample … sample B

Avg(Y*(1)) … Avg(Y*(B))

Histogram of Avg(Y*) estimates sampling distribution



Computing

- Generally not too hard to do it yourself as long as the software allows you to
  - Draw random samples
  - Extract results, such as regression slopes
  - Iterative calculation
  - Accumulate the results
- Specialized packages

Sample Code in R

- Load data

```
osteo <- read.table("osteo.txt", header=T)
attach(osteo)
```

- Bootstrap loop to accumulate results

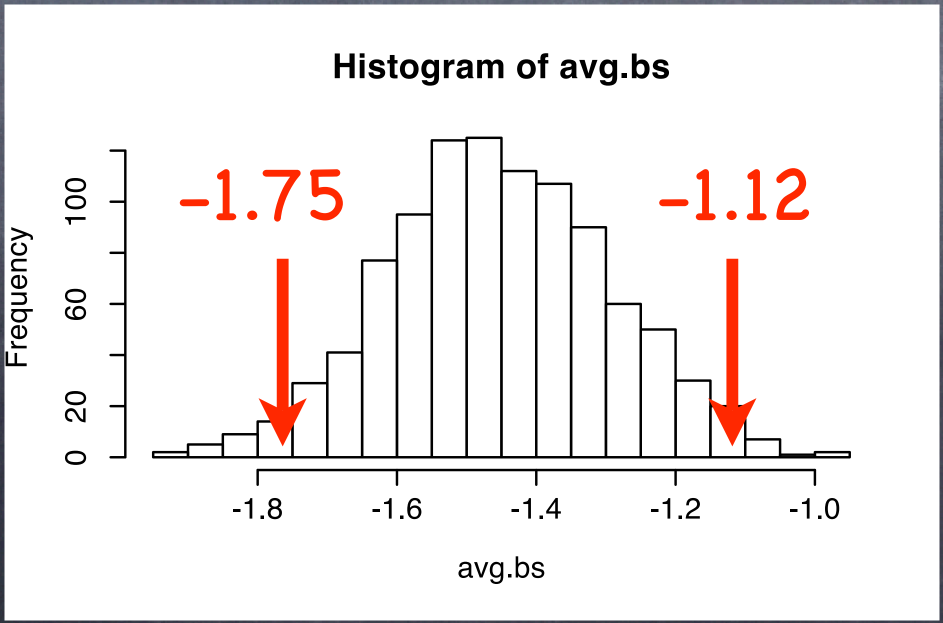
```
avg.bs <- c()
for(b in 1:1000) {
  yStar <- sample(hip, 64, replace=T)
  avg.bs <- c(avg.bs, mean(yStar)) }
```

- Compute summary statistics, generate plots

```
sd(avg.bs)          gives simulated SE = 0.159
hist(avg.bs)        draws histogram on prior page
```

What about a CI?

- Hope for normality, with BS filling in SE
  - $-1.45 \pm 2 \cdot 0.159 = [-1.77, -1.13]$ = t-interval
- Invert hypothesis tests... humm.
- Build bootstrap version of t-distribution...
- Use the sampling distribution directly



Bootstrap Percentile Intervals

- Computed directly from the bootstrap sampling distribution of the statistic
- Order the bootstrap replications
  $T_{(1)}(Y^*) < T_{(2)}(Y^*) < \cdots < T_{(B)}(Y^*)$
- To find the 95% confidence interval, say, use the lower 2.5% point and the upper 97.5% point.
- Need "a lot of replications" to get a reliable interval because you're reaching out into the tails of the distribution

How many replications?

- Enough!
- Don't want the bootstrap results to be sensitive to simulation variation

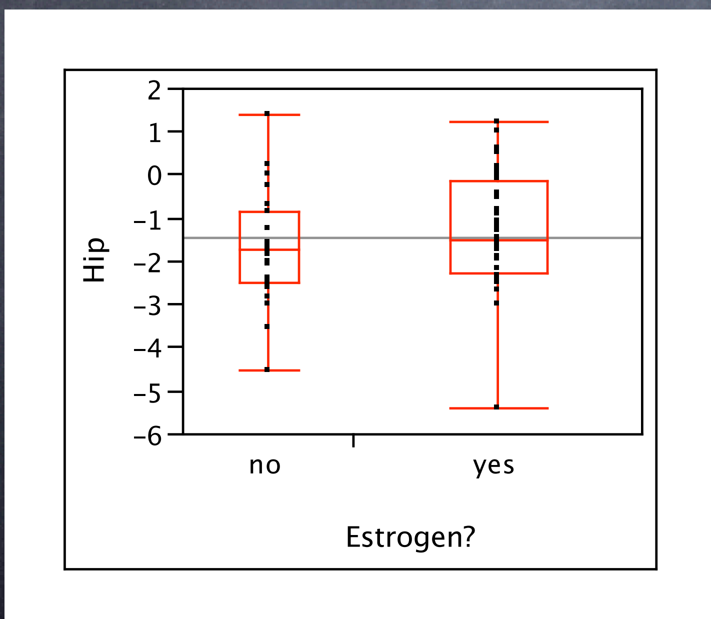
	B=100 SE	B=2000 SE	B=100 CI	B=2000 CI
Trial 1	0.176	0.160	-1.79,-1.08	-1.76,-1.12
Trial 2	0.145	0.164	-1.71, -1.17	-1.76,-1.12
Trial 3	0.169	0.162	-1.74,-1.10	-1.78,-1.14

Testing Hypotheses

- Key notion
  Need to be able to do the resampling in a way that makes the null hypothesis of interest true in the sampled distribution
- Example
  - Do women who have taken estrogen have higher bone mass than those who have not?
  - Standard approach would set
    $$H_0: \mu_1 = \mu_2$$
    and use a two-sample t-test

Two-sample t-test

- Two-sample test does not reject H_0
  - Difference in means is only about 1 SE away from zero
  - p-value (two-sided) is about 0.3

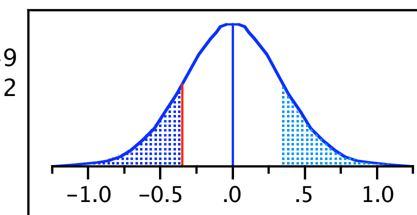


t Test

no-yes

Assuming unequal variances

Difference	-0.352	t Ratio	-1.049
Std Err Dif	0.335	DF	49.732
Upper CL Dif	0.322	Prob > |t|	0.299
Lower CL Dif	-1.026	Prob > t	0.85
Confidence	0.95	Prob < t	0.15



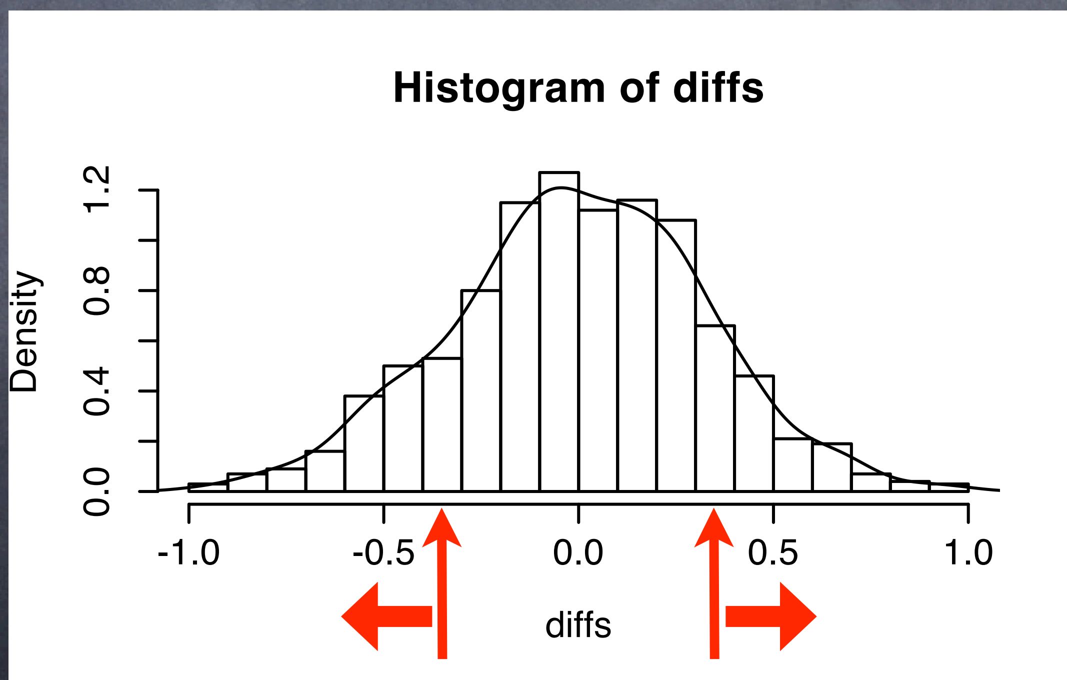
Bootstrap Comparison

- Need to do the resampling in such a way that the null is true
  - Mix the two samples, assuming that the variances are comparable
  - Force the two populations to have a common mean value (eg, grand mean)
- Draw bootstrap sample from each group
- Compute difference in means
- Repeat

Distribution of Differences

- Bootstrap probability of mean difference larger than the observed difference

$$P_0^*\left(|\overline{Y}^*_{no} - \overline{Y}^*_{yes}| > 0.35\right) = 0.28$$



Caution

- Hypothesis testing requires that you impose the null prior to doing the resampling
  - Not always easy to do
  - Example: How would you impose the null of no effect in a multiple regression with collinear predictors?
- Confidence intervals are direct and do not require "enforcing" a hypothesis

Big Picture

- Bootstrap resampling is a methodology for finding a sampling distribution
- Sampling distribution derived by using F* to estimate the distribution of population
  - Treat sample as best estimate of population
- Computing is attractive
  - Draw samples with replacement from data and accumulate statistic of interest
  - SD of simulated copies estimates SE
  - Histogram estimates the sampling distribution, providing percentile intervals

Does this really work?

- Yes!
- Key to success is to make sure that the bootstrap resampling correctly mimics the original sampling
- Bootstrap analogy

$$\theta(F):\theta(F_n) \quad :: \quad \theta(F_n):\theta(F^*)$$

- Key assumption is independence

Variations on a Theme

- I emphasize the “nonparametric” type of bootstrap which resamples from the data, mimicking the original sampling process
- Alternatives include
  - Parametric bootstrap, which mixes resampling ideas with Monte Carlo simulation
  - Computational tricks to get more efficient calculations (balanced resampling)
  - Subsampling, varying the size of the sample drawn from the data

Some Origins

- Several early key papers are worth a look back at to see how the ideas began
  - Efron (1979), "Computers and the theory of statistics: thinking the unthinkable", Siam Review
  - Efron (1979), "Bootstrap methods: another look at the jackknife", Annals of Statistics
  - Diaconis & Efron (1983), "Computer intensive methods in statistics", Scientific American

Bootstrap Always Works?

- No
  - It just works much more often than any of the common alternatives
- Cases when it fails
  - Resampling done incorrectly, failing to preserve the original sampling structure
  - Data are dependent, but resampling done as though they were independent
  - Some really weird statistics, like the maximum, that depend on very small features of the data

Reasons to Bootstrap

- Using non-standard estimator
- Diagnostic check on traditional standard error
  - Compute SE, CI by traditional approach
  - Compute by bootstrap resampling
  - Compare
- Provides way to justify new computer on research grant

Bigger Picture

- Once you're willing to "let go" of traditional need for formulas, you can exploit more interesting estimators
- Example... trimmed mean
  - Robust estimator
  - Trim off the lowest 10% and largest 10%
  - Take the average of the rest
  - Median trims 50% from each tail
- Standard error?
  - Formula exists, but its a real mess

Same Paradigm

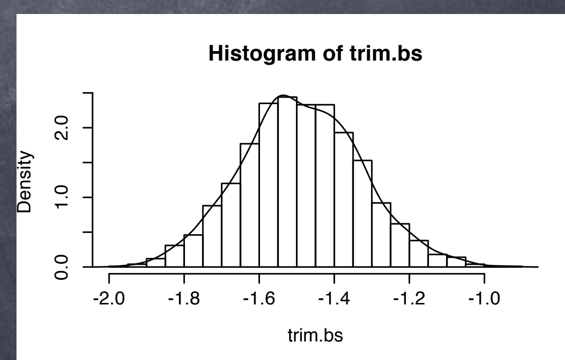
Osteoporosis sample F_n

Sample ... sample B

Trim(Y*(1)) ··· Trim(Y*(B))

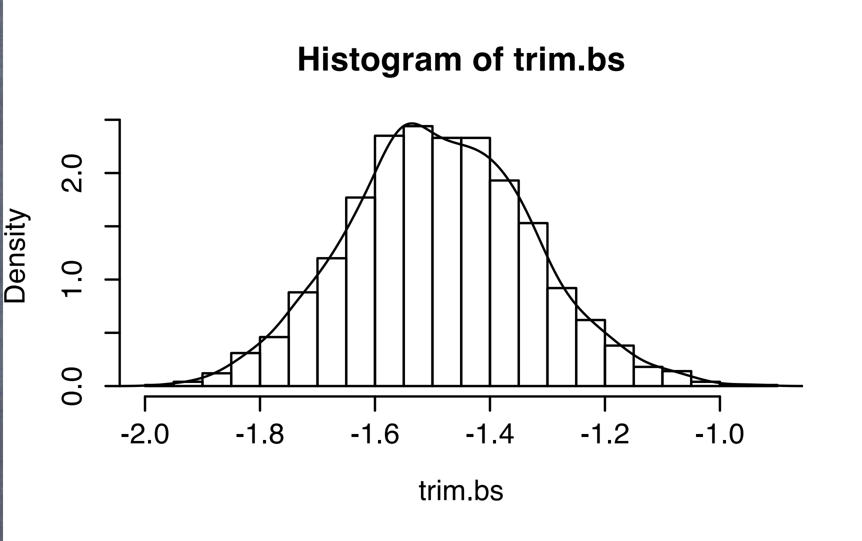
Just replace average by trimmed mean

Histogram of Trim(Y*) estimates sampling distribution, SE



Results for Trimmed Mean

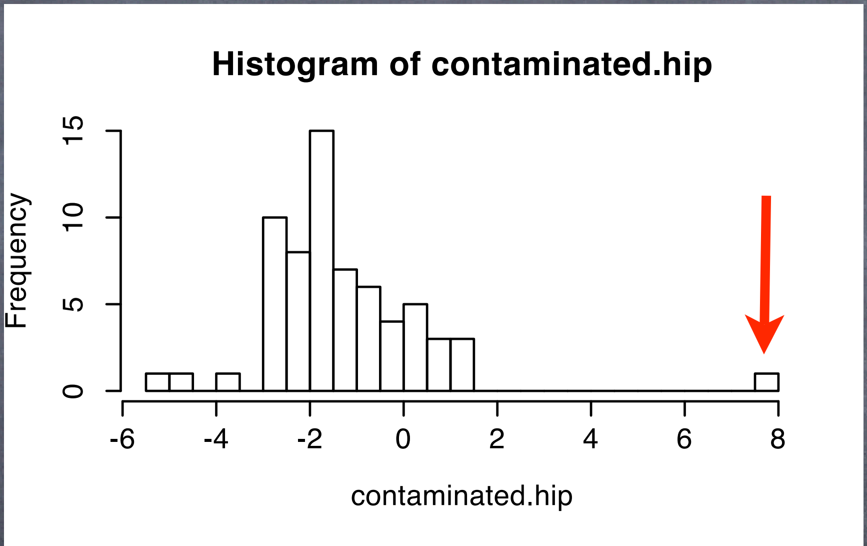
- Bootstrap B=2000 replications



- Results similar to using an average
  - Bootstrap SE = 0.16
  - Percentile interval = -1.79 to -1.17

But what about an outlier?

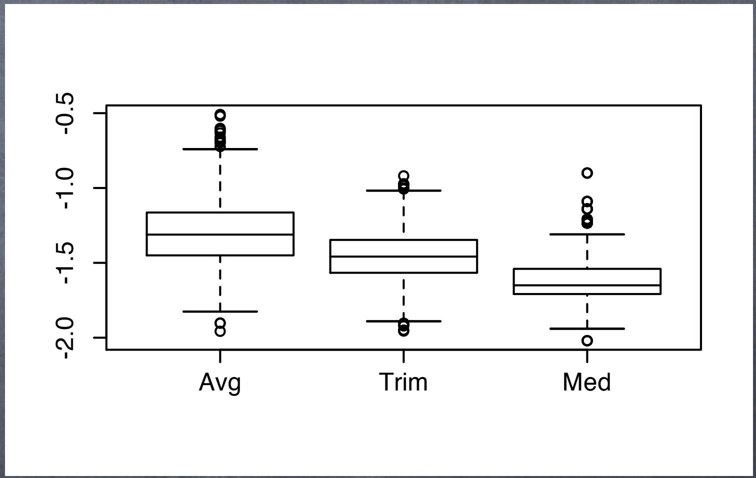
- Add one point that's a large outlier far from the rest of the data.



- Let's see how several estimates of location compare in this situation

Bootstrap Comparison

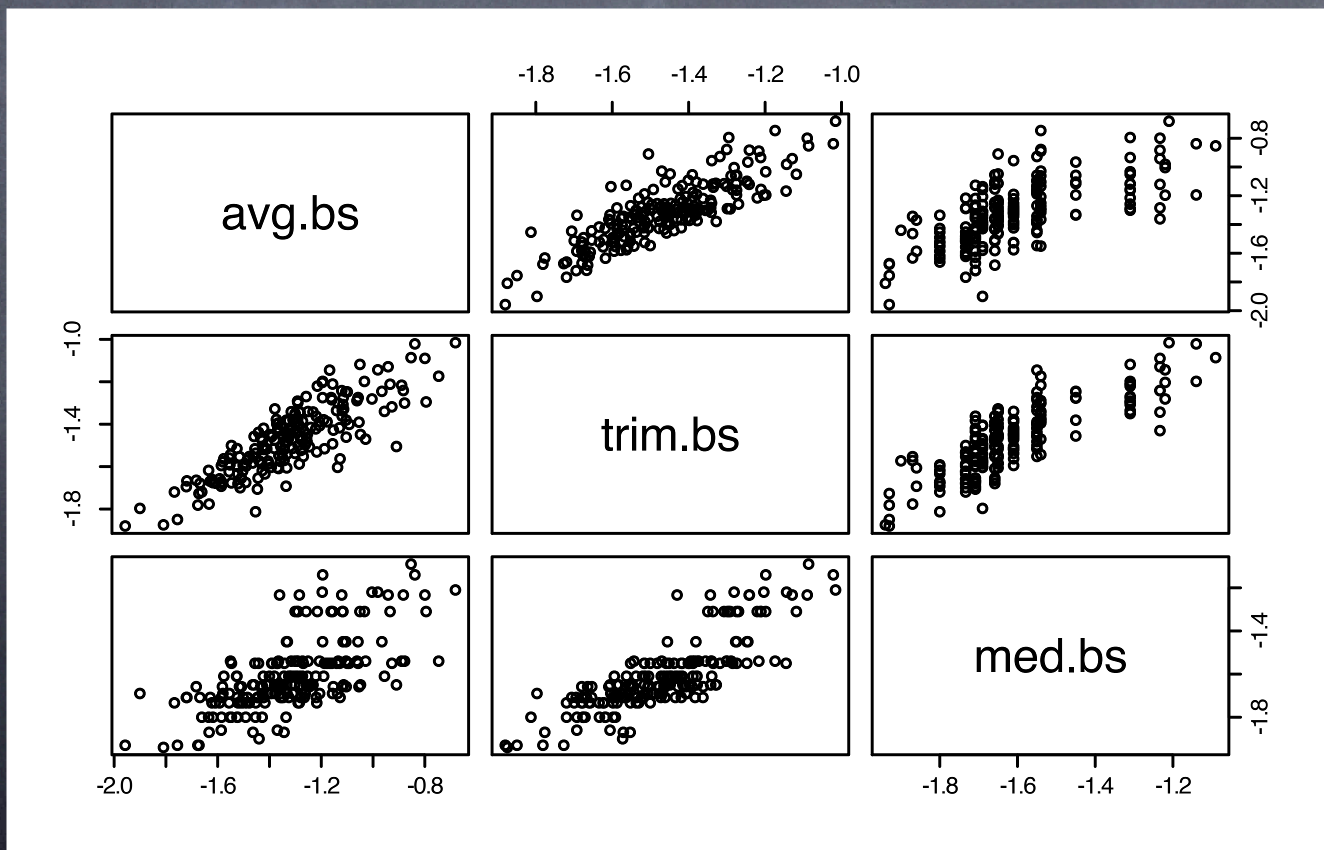
- Bootstrap 3 estimators, 2000 samples
  - Mean, trimmed mean, median
  - Compute all three for each bootstrap sample
- Trimmed mean has the smallest SE
  - SE*(Mean) = 0.21
  - SE*(Trim) = 0.16
  - SE*(Median) = 0.18



- Percentile interval for trimmed mean almost same as before, -1.76 to -1.15

Interesting Looks at Stats

- Bootstrap resampling makes it simple to explore the relationships among various statistics as well



Managing Expectations

- Bootstrapping provides a reliable SE and confidence interval for an estimator
  - Explore properties of estimators
  - Focus on problem, not formulas
- Bootstrapping does not routinely
  - By itself produce a better estimator
  - Generate more information about population
  - Cure problems in sampling design
  - Convert inaccurate data into good data

Questions?

Applications in Surveys

- Ratio estimator
  - Estimator is a ratio of averages obtained from two different surveys
- Sampling design
  - Adjust for the effects of sample weights on statistical summaries
  - Clustered sampling
  - Rao and Wu (1988, JASA) summarize the more technical details and results

Ratio Estimation

- Common to take ratio of summary statistics from different samples
- Example
  - Ratio of incomes in two regions of US
  - Weekly income reported in US Current Population Survey, April 2005
  - Homogeneity reduces sample size
  - NE/Midwest = 721.4/673.5 = 1.071
  - Weekly earnings in NE 7% larger

Level	Number	Mean	Std Dev	Std Err Mean
Midwest	164	673.5	490	38.3
NE	167	721.4	539	41.7

Classical Approach

- Some type of series approximation
- For ratio of averages of two independent samples, leads to the normal approximation

$$\sqrt{n}\left(\frac{\overline{Y}_1}{\overline{Y}_2} - \frac{\mu_1}{\mu_2}\right) \sim N\left(0,\ \frac{\sigma_1^2}{\mu_2^2} + \frac{\sigma_2^2\,\mu_1^2}{\mu_2^4}\right)$$

Details for the curious

$$g(\overline{Y}_1, \overline{Y}_2) \approx g(\mu_1, \mu_2) + \nabla g(\mu)\ \cdot\ (\overline{Y}_1 - \mu_1, \overline{Y}_2 - \mu_2)$$

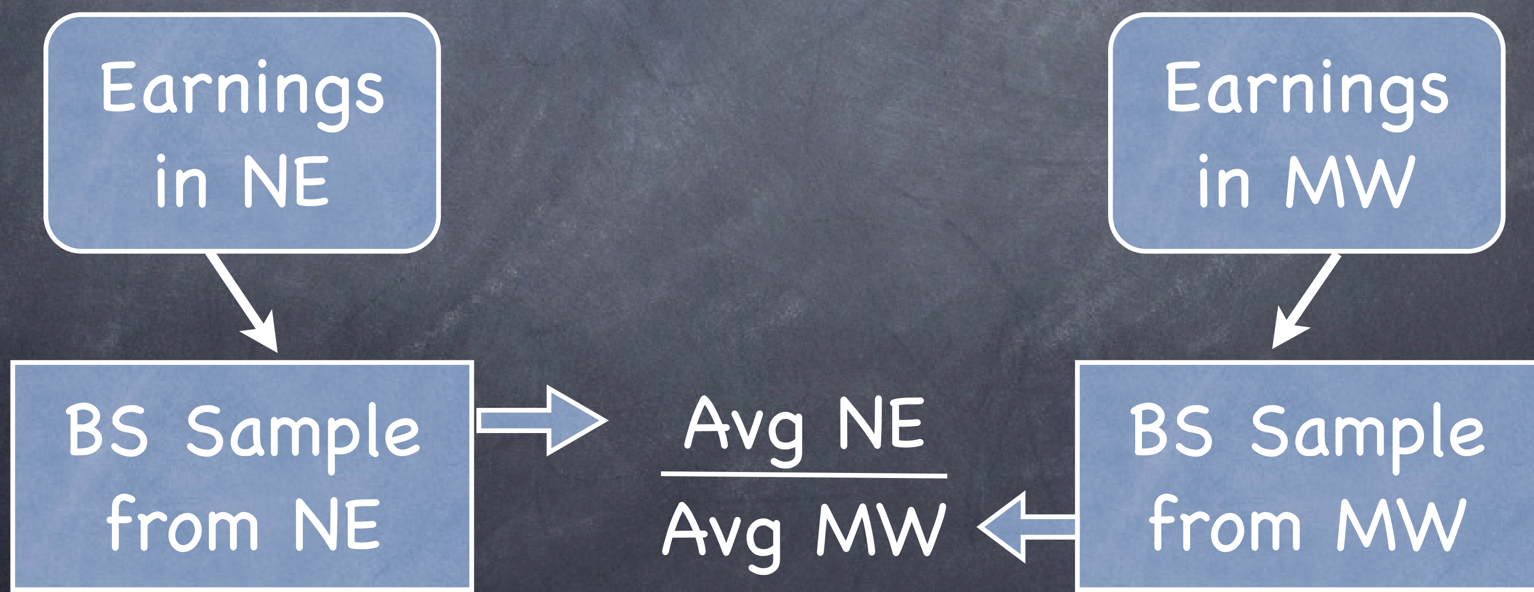
$$g(a,b) = \frac{a}{b}, \quad \nabla g(a,b) = \left(\frac{1}{b}, -\frac{a}{b^2}\right)$$

Classical Results

- Unbiased
  Estimate the ratio μ_{ne}/μ_{mw} by ratio of averages, 1.071
- Standard error
  Estimate SE of ratio of averages by plugging in sample values (eg s^2 for σ^2) to obtain SE $\approx$ 0.083
- Confidence interval
  Confidence interval requires that we really believe the normal approximation

Bootstrap Approach

- Two independent samples
- Resample each separately
- Compute ratio of means
- Repeat

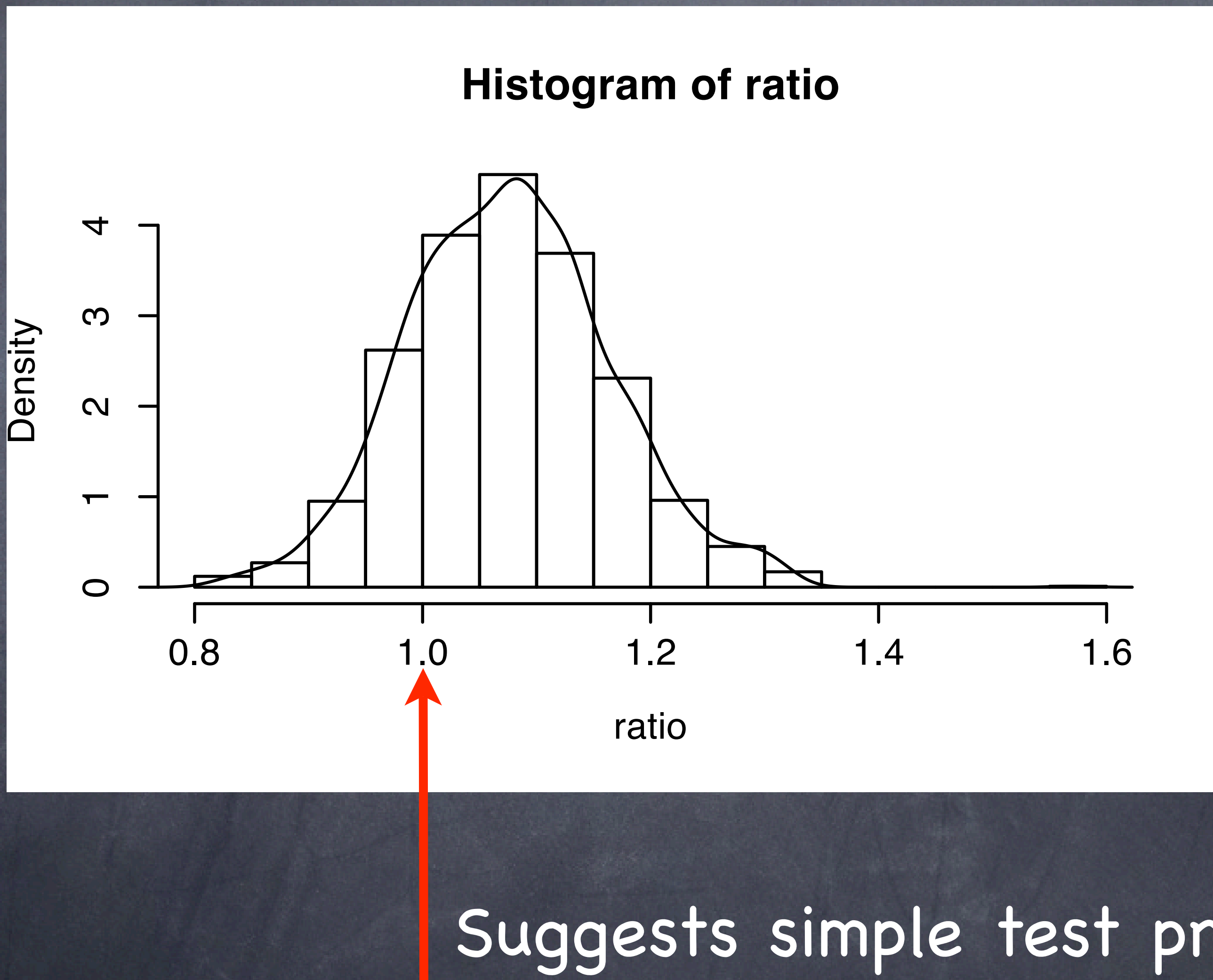


Bootstrap Results

- Repeat with 2000 ratios, with numerator from NE and denominator from MW
- Bias?
  Evidently not much, as the average bootstrap ratio is 1.076
- SE
  Similar to delta method, SE*(ratio) = 0.089
- Percentile interval is slightly skew

  0.91 to 1.27 = [1.07-0.16, 1.07+0.20]

Bootstrap Sampling Dist



Suggests simple test procedure

Bootstrap in Regression

- Familiar linear model with q predictors

  $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{iq} + \epsilon_i$

  In vector form

  $$Y = X\beta + \epsilon$$

- The OLS estimator is linear in Y, given X,

  $$b = (X'X)^{-1}(X'Y)$$
  $$= \text{weighted sum of } Y_i$$

- Residuals are

  $$e = Y - Xb$$

  with estimated variance $s^2 = \Sigma(e_i^2)/(n-q-1)$

Bootstrap Linear Estimator

- Bootstrap standard error can be gotten for any linear estimator without computing
- Assuming the model as specified,

$$Y = X\beta + \epsilon,$$

generate a bootstrap sample given X by resampling residuals

$$Y^* = X b + e^*$$

- Conditional on design of the model

$$b^* = (X'X)^{-1}X'Y^* = b+(X'X)^{-1}X'e^*$$

so that $SE^*(b^*) = (X'X)^{-1} \Sigma e_i^2/n$

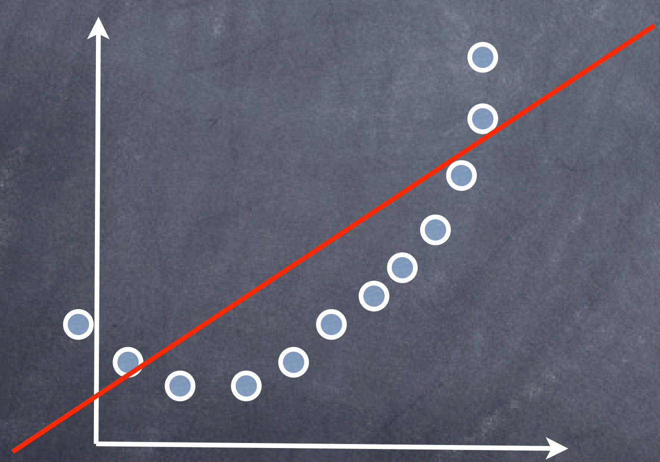
BS in Regression

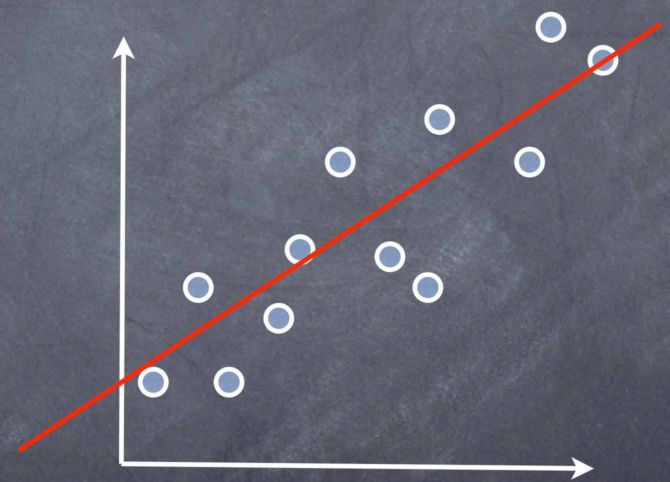
- Notice that this approach
  (1) Assumes the model is correctly specified, with the usual assumptions on the errors holding
  (2) Fixes the X design (conditional on X)
  (3) Produces a slightly biased SE, shrunken toward 0
- The first requirement is particularly bothersome
  - Believe have the right predictors?
  - Believe homoscedastic?

Wrong Model?

- Suppose that the data have this form:

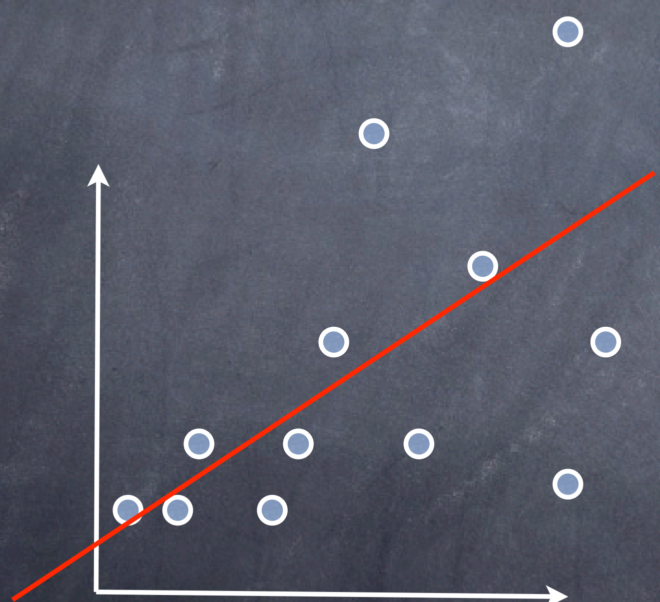
Then the resulting bootstrap sample will look like this



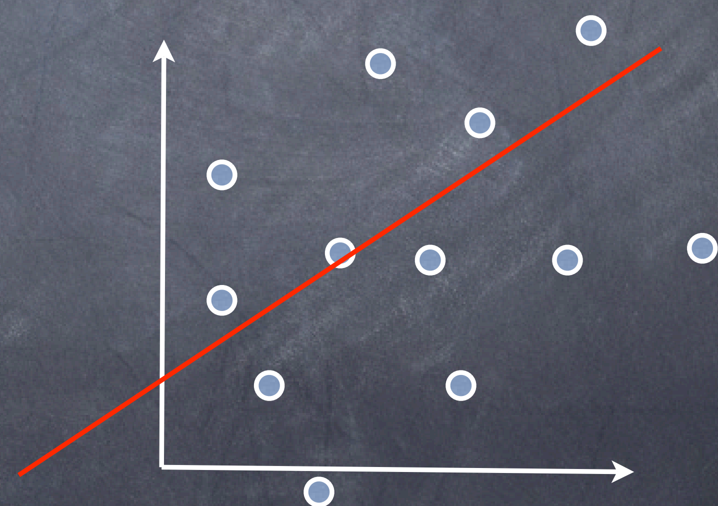


Wrong Error Structure?

- Suppose that the data do not have equal variance:



Then the resulting bootstrap sample will look like this

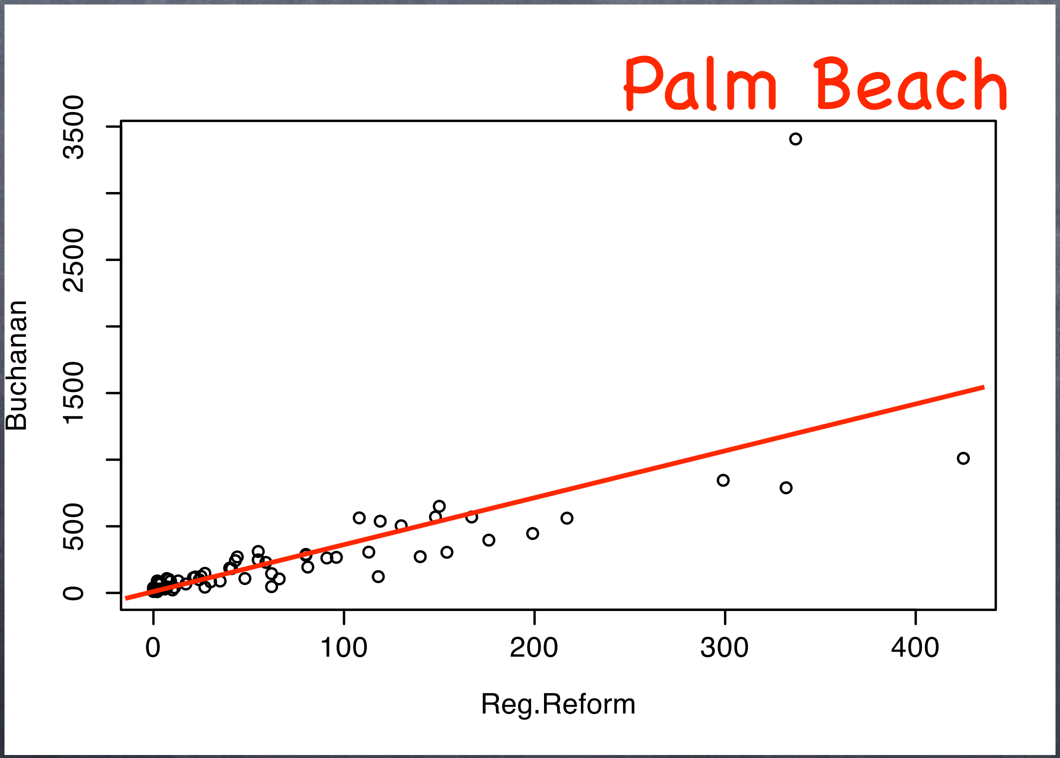


Model-Free Resampling

- Rather than resample residuals, resample observations
  - Resample from the n tuples (y_i, x_i)
  - Resulting data have different structure, one that keeps y_i bound to x_i
  - Random design
- Procedure now gets the right structure in the two previous illustrations
  - Model is not linear
  - Errors lack equal variance

Outlier Havoc

- Florida 2000 US presidential county-level vote totals for Buchanan vs number registered in Buchanan's Reform Party.

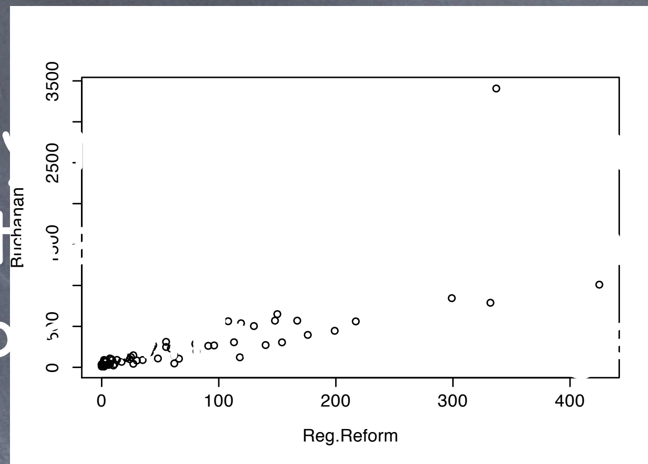


$b_1 = 3.7$

$SE = 0.41$

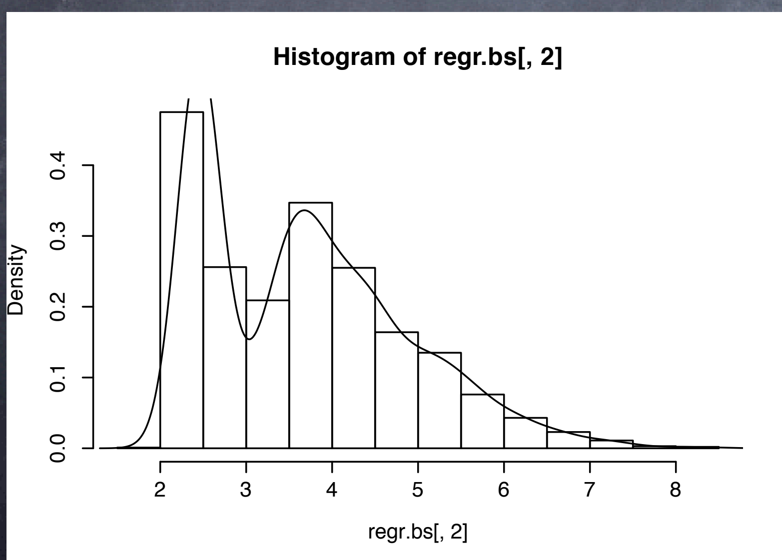
Which is which?

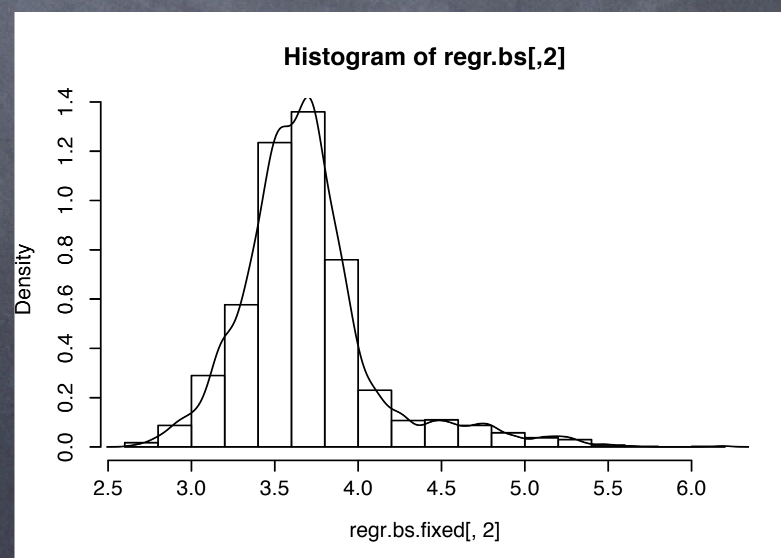
It certainl… …ethod is used
to assess t… …e slope in this
examp… …d outlier.



SE* = 1.17

SE* = 0.41





Comparison

- Two results "should" be close, but can be rather different in cases of outliers
- Resampling residuals
  - Fixes the design, as might be needed for certain problems (experimental design)
  - Closely mimics classical OLS results
  - But, requires model to hold
- Resampling cases (aka, correlation model)
  - Allows predictors to vary over samples
  - Robust to model specification

Longitudinal Data

- Repeated measurements
  - Growth curves
  - Panel survey
  - Multiple time series
- Data shape
  - n items (people, districts, ...)
  - T observations per item
- More general error structure
  - Items are independent, but anticipate dependence within an item

Longitudinal Modeling

- "Fixed effects" models
  - Econometrics

    $Output_{it} = \alpha_i + \beta_1 \text{ Trend} + \beta_2 \text{ Macro} + \ldots + \epsilon_{it}$
- "Random effects" models
  - Growth curves

    $Weight_{it} = a_i + \beta_1 \text{ Age} + \beta_2 \text{ Food} + \ldots + \epsilon_{it}$
- Hierarchical Bayesian models
- Functional data analysis
- Honest degree of freedom approach
  - Reduce to single value for each case

Bootstrap for Longitudinal

- Extend bootstrap to other types of error models
- Key element for successful resampling is independence
  - Conditional on data, resampled values are independent, so
  - Better make sure that the original sampling produced independent values
- Longitudinal models usually assume independent subjects

Longitudinal Example

- Stylized example tracking economic growth
  - 25 locations
  - Two years (8 quarters)
- Simple model for retail spending
  - $Spending_{it} = \alpha_i + \beta_1 Un_{it} + \beta_2 Yd_{it} + \epsilon_{it}$
- Simple model is probably misspecified
  - Suggests error terms may be highly correlated within a district

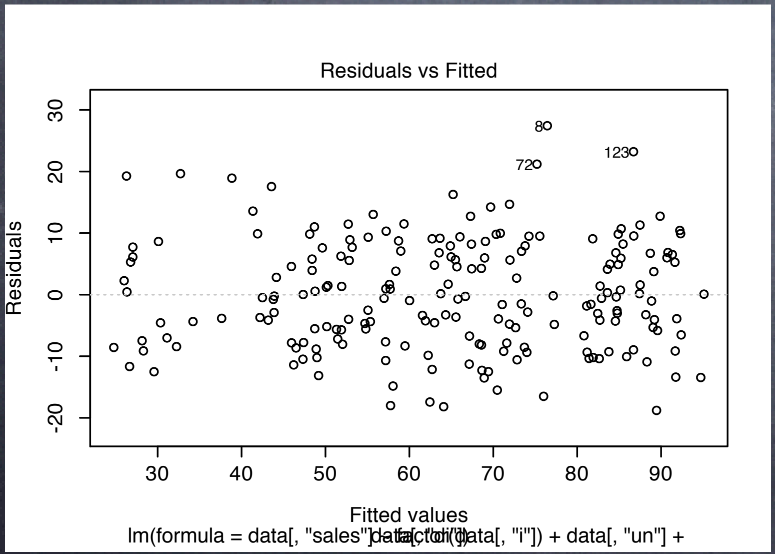
OLS Estimates

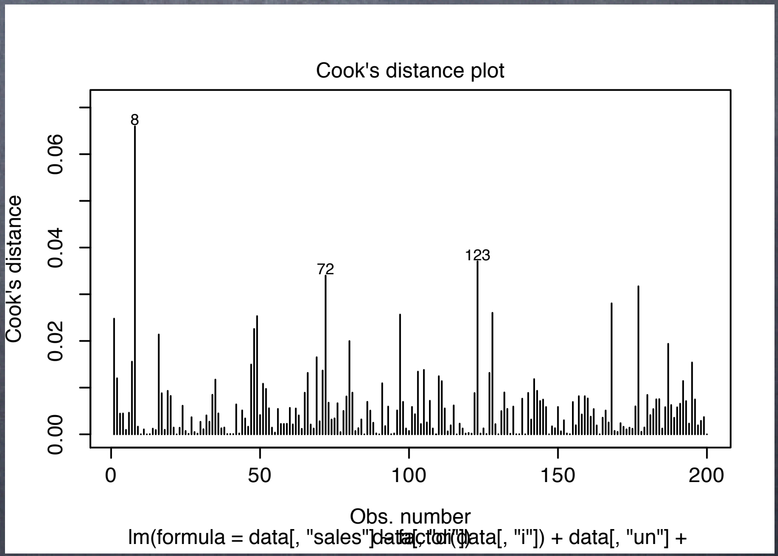
- Fit the usual OLS regression, with separate intercept within each district
- Find significant effects for employment and disposable income

Factor	Coef	SE	t
Avg Effect	43	11	4.0
Unemp	-97.7	29.1	-3.4
Disp Inc	0.29	0.087	3.3

Residual Issues

- Standard residual plots look fine,
- But "longitudinal" residual correlation is large at 0.5





Resampling Plan

- Exploit assumed independence between districts
  - Resample districts, recovering a "block" of data for each case
  - Assemble data matrix by glueing blocks together
- Bootstrap gives much larger SE for Disp Inc

Factor	Coef	SE	SE*	t*
Avg Effect	43	11	24	1.8
Unemp	-97.7	29.1	26.5	3.7
Disp Inc	0.29	0.087	0.144	2.0

What happened?

- Bootstrap gives a version of the "sandwich" estimator for the SE of the OLS coefficients
- Sandwich estimator

  $$Var(b) = (X'X)^{-1} X'(\text{diag } e_i e_i') X (X'X)^{-1}$$

- Note that both bootstrap and sandwich estimators presume districts are independent.

Comments

- Why the effect on the SE for the estimate of Disp Income but not for the slope of unemployment?
- Answer requires more information about the nature of these series
  - Within each district, unemployment rates vary little, with no trend
  - Within each district, disposable income trends during these two years
  - Trend gets confounded with positive dependence in the errors

Getting Greedy

- Generalized least squares
  - With the dependence that we have, suggests that one ought to use a generalized LS estimator
- Estimator requires covariance matrix for the model errors

$$b_{gls} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y)$$

$$Var(\epsilon) = \Omega$$

- But never see errors, and only get residuals after fit the slope…

Practical Approach

- Two-stage approach
  - Fit the OLS estimator (which is consistent)
  - Calculate the residuals
  - Estimate error variance from residuals, using whatever assumptions you can rationalize
- Estimate with V in place of Ω

$$b_{gls2} = (X'V^{-1}X)^{-1}(X'V^{-1}Y)$$

- But what is the SE for this thing?
  - $Var(b_{gls}) = (X'\Omega^{-1}X)^{-1}$
  - $Var(b_{gls2}) =?= (X'V^{-1}X)^{-1}$

Bootstrap for GLS

- Freedman and Peters (1982, JASA)
- Show that the plug-in GLS SE underestimates the sampling variation of the approximate GLS estimator
- Bootstrap fixes some of the problems, but not enough
- Bootstrap the bootstrap
  - Use a "double bootstrap" procedure to check the accuracy of the bootstrap itself
  - Find that SE* is not large enough

Dilemma

- OLS estimator
  - "Not efficient" but we can get a reliable SE by several methods
    - Bootstrap
    - Sandwich formula
- GLS estimator
  - "Efficient" but lack simple means to get a reliable SE for this estimator

Double Bootstrap Methods

- Return to the simple problem of confidence intervals
- Numerous methods use the bootstrap to get a confidence interval
  - Percentile interval
  - BS-t interval
  - Bias-corrected BS interval
  - Accelerated, bias-corrected BS interval
  - ...
- Use the idea of Freedman and Peters to improve the percentile interval

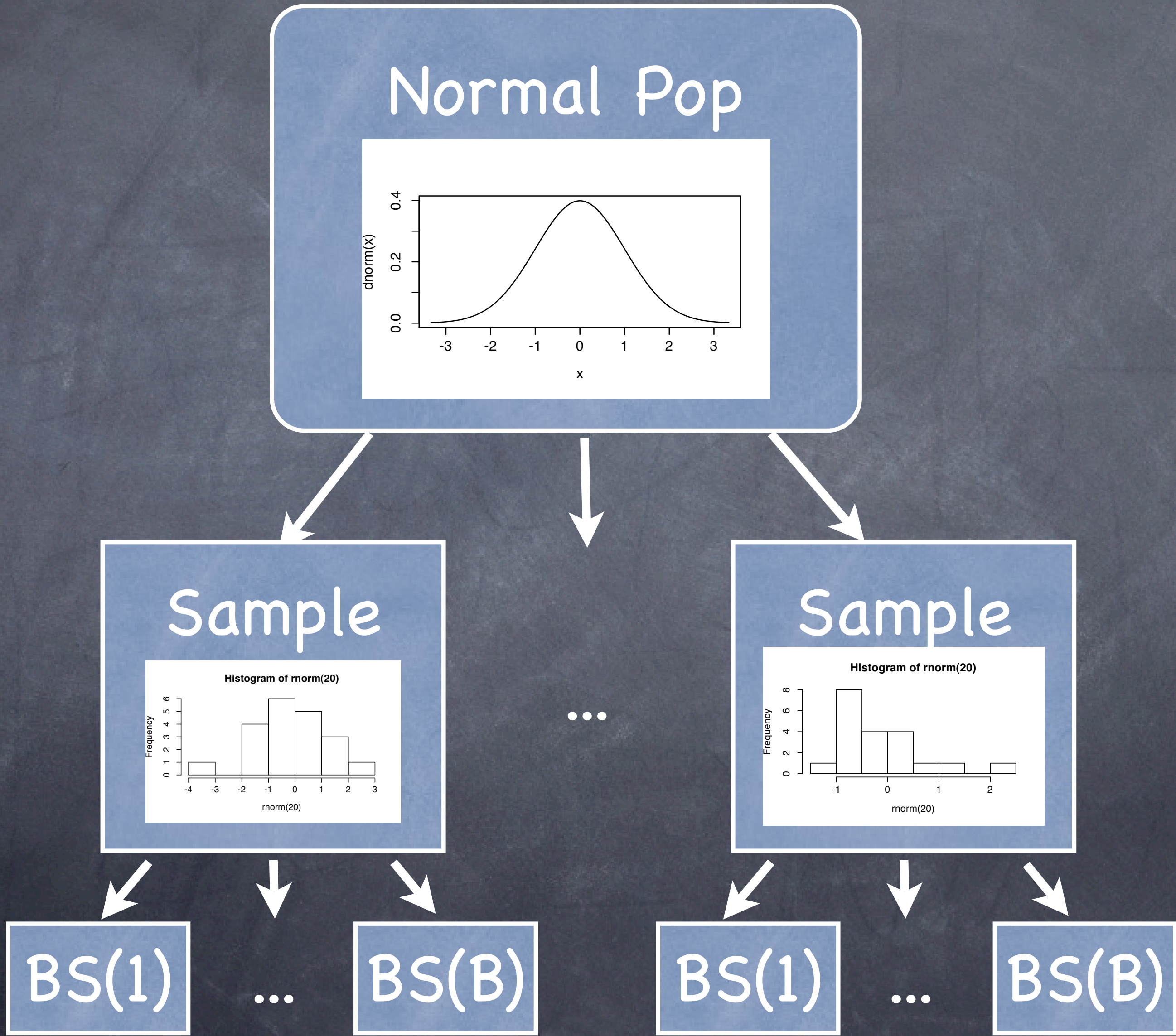
CI for a Variance

- Consider a problem with a known answer
  - $Y_1, \ldots, Y_{20}$ iid $N(\mu,\sigma^2)$
  - Get a 90% confidence interval for σ^2
- The standard interval uses percentiles from the chi-square distribution

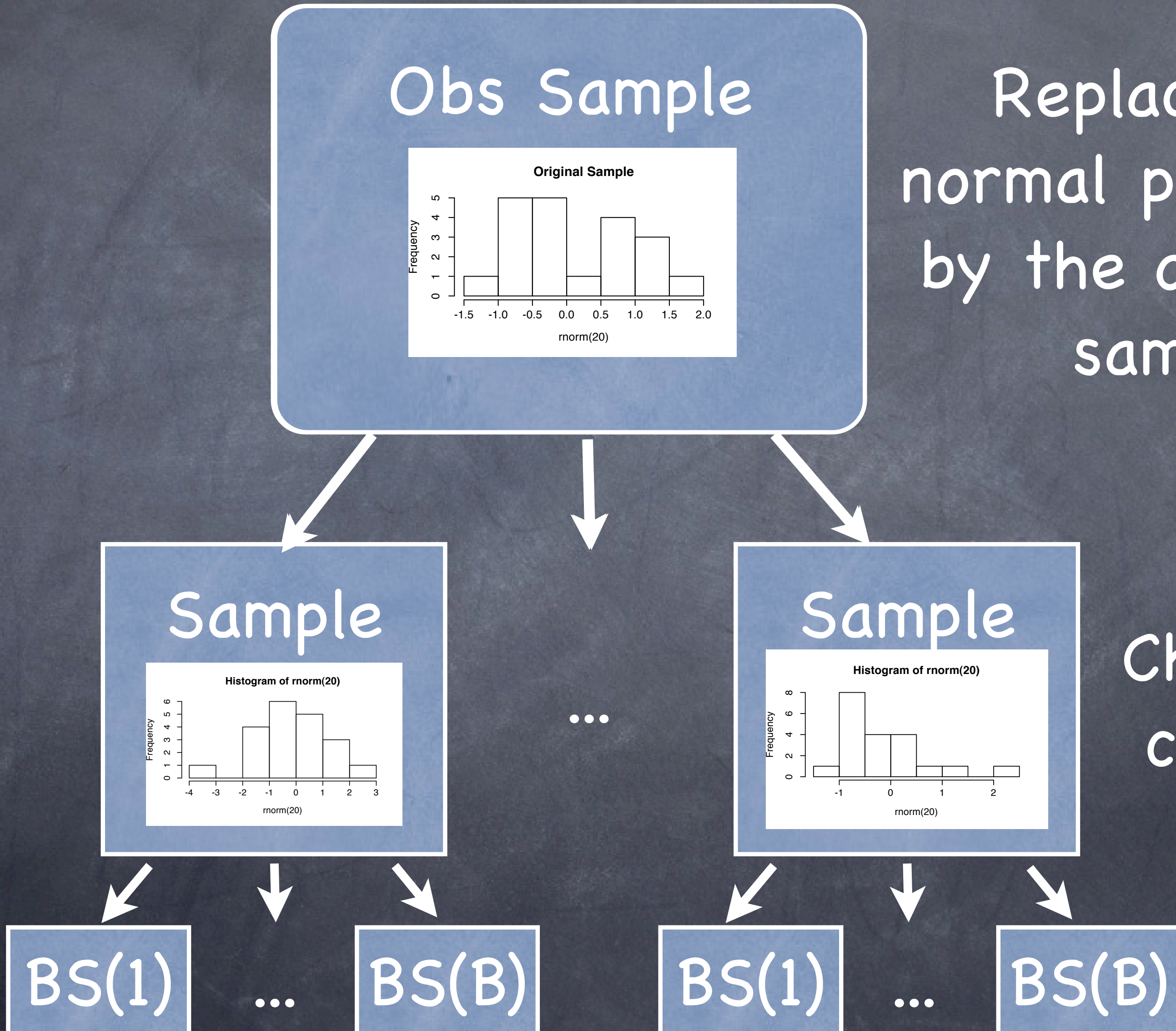
$$P\left(\frac{(n-1)\,s^2}{\chi^2_{0.95}} \le \sigma^2 \le \frac{(n-1)\,s^2}{\chi^2_{0.05}}\right) = 0.90$$

- The standard bootstrap percentile interval has much less coverage (Schenker, 1985)
  - Nominal 90% percentile interval covered σ^2 only 78% of the time

Simulation Experiment



Double Bootstrap



Replace the normal population by the observed sample

Check the coverage

Double Bootstrap Method

- Start with data, having variance s^2
  - Draw a bootstrap sample
  - Find the percentile interval for this sample
    - This is the second level of the resampling
  - Repeat
- Results for variance
  - Of 500 percentile intervals, only 81% cover bootstrap population value (which is s^2)
  - Need to calibrate the interval

Calibrated Percentile Interval

- If use the 0.05 and 0.95 percentiles of the values of s^{2*}, only covers 81% of the time
- So, adjust interval by using more extreme percentiles so that coverage is better

Lower	Upper	Coverage
0.05	0.95	0.81
0.04	0.96	0.83
0.02	0.98	0.88
0.01	0.99	0.895

Bootstrap Calibration

- Bootstrap is self-diagnosing
  - Use the bootstrap to check itself, verifying that the procedure is performing as advertised
- Now you really can justify that faster computer in the budget

Where to go from here?

- Bootstrap resampling has become a standard method within the Statistics community
- Focus on research problems, choosing the appropriate method to obtain a good SE and perform inference
- Books
  Efron & Tibshirani (1993) Intro to Bootstrap
  Davison & Hinkley (1997) Bootstrap Methods
- Software
  R has “boot” package

Questions?