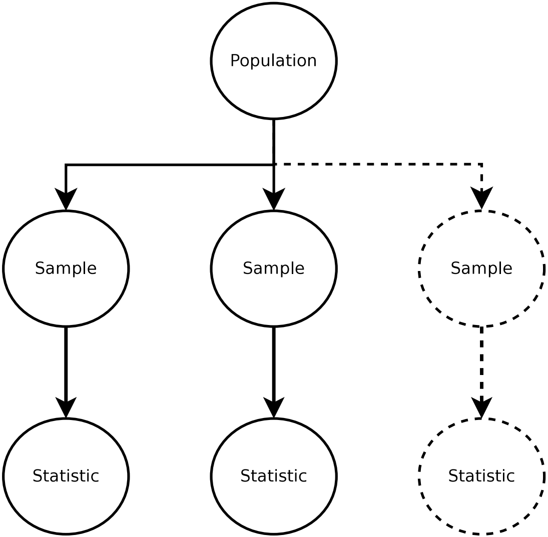
Empirical Justification of Statistical Inferences

Matthew S. Shotwell

Assistant Professor
Department of Biostatistics
Vanderbilt University School of Medicine

August 28, 2013

Classical Inference



Classical Inference

The algorithm:

1. make assumptions about data generating mechanism
2. deduce or approximate sampling distribution
   - exact
   - asymptotic
   - bootstrap (asymptotic *validity*)
3. check (*not verify*) assumptions in observed data
4. make inferences conditional on assumptions

Problems with Classical Inference

Non-technical problems

- restricted thinking
- unnecessary theoretical consideration

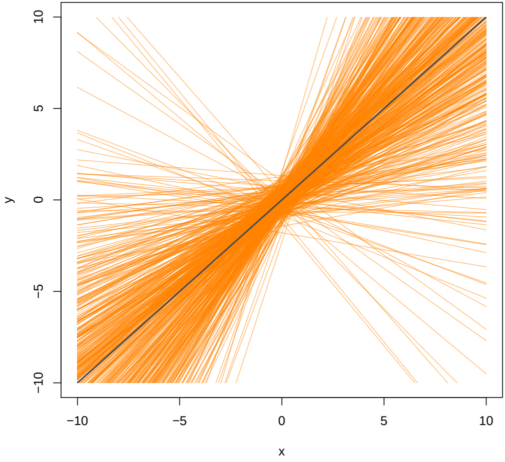
Problems with Classical Inference

Technical problems

- assumptions not met
- sample size insufficiency (asymptotic justification)

Sample Size Insufficiency Example

$$y = x + \epsilon \qquad \epsilon \sim N(0,1) \qquad x \sim N(0,1) \qquad n = 5$$



VANDERBILT UNIVERSITY

Sample Size Insufficiency Example

For intercept and slope ($\alpha = 0.05$):

estimate	$\hat{\beta}$, maximum likelihood estimate
interval	$\hat{\beta} \pm z_{\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}}$
test	H_0: $\beta = 0$, reject when $\lvert\hat{z}\rvert > z_{\frac{\alpha}{2}}$

VANDERBILT UNIVERSITY

Sample Size Insufficiency Example

In a simulation of $1e^6$ samples:

	Bias	Coverage	Type-I	Type-II
Intercept	$-3.7e^{-4}$	85%	16%	—
Slope	$-2.2e^{-3}$	86%	—	47%

VANDERBILT
UNIVERSITY

Sample Size Insufficiency Example

If we adjust (by guess-and-check) the value of α to 0.002:

	Bias	Coverage	Type-I	Type-II
Intercept	-1.8e^{-3}	95%	5%	—
Slope	-7.0e^{-4}	95%	—	71%

VANDERBILT UNIVERSITY

Sample Size Insufficiency Example

Confidence interval widths:

α	Intercept	Slope	Coverage
0.05	1.2 1.7 2.4	1.3 2.0 2.8	85%
0.002	1.9 2.7 3.6	1.9 2.9 4.5	95%

VANDERBILT UNIVERSITY

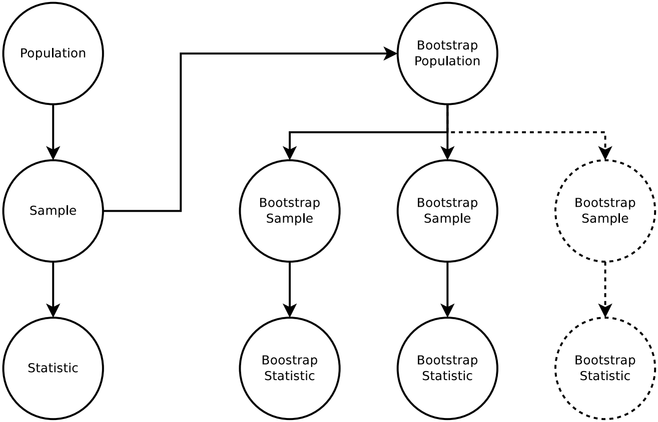
A Novel or Forgotten Idea

- checking assumptions of asymptotic/exact inferences does not directly assess the correctness of statistical inferences (i.e., coverage/type-I error)
- empirical checks of coverage/type-I error, followed by calibration of confidence limits/critical values may result in more accurate statistical inferences

VANDERBILT UNIVERSITY

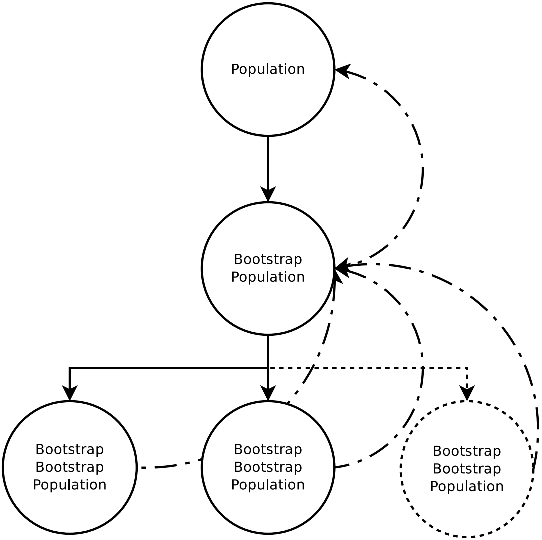
Origins of Empirical Calibration: The Double Bootstrap

The Bootstrap



Origins of Empirical Calibration: The Double Bootstrap

The Double Bootstrap



Empirical Inference: Building on the Double Bootstrap Idea

The algorithm

1. ~~make assumptions~~ approximate data generating mechanism
2. ~~deduce~~ approximate sampling distribution
   - ~~exact~~
   - ~~asymptotic~~
   - simulation/bootstrap (~~asymptotic~~ empirical *validity*)
3. ~~check (not verify) assumptions in observed data~~
4. make inferences conditional on ~~assumptions~~ approximation

VANDERBILT
UNIVERSITY

Sample Size Insufficiency Problem Revisited

Consider again the Sample Size Insufficiency Example, where we have a sample of 5 (x, y) pairs, and have estimated the intercept and slope that best characterizes the linear association between x and y, as well as the error variance. Hence, these estimates characterize our best guess regarding the data generating mechanism, and may be used to calibrate (via simulation) statistical inferences. But, does this really improve coverage/type-I error?

VANDERBILT UNIVERSITY

Sample Size Insufficiency Problem Revisited

For a basic implementation of the empirical inference algorithm (nominal $\alpha = 0.05$)

	Coverage	Type-I	Type-II
Intercept	95%	6%	—
Slope	98%	—	69%

VANDERBILT UNIVERSITY

Isn't Empirical Justification Enough?

If the desired inferences have the expected properties (i.e., coverage/type-I error) under a plausible data-generating mechanism or a set of plausible mechanisms, isn't this sufficient justification? What additional validity might be attained by considering a theoretical (asymptotic/exact) argument?

VANDERBILT UNIVERSITY

Empirical Inference Where Asymptotic Inference Fails

The canonical nonparametric bootstrap failure:
$U(0, \theta)$, where $\hat{\theta} =$ sample maximum

- bootstrap distribution does not converge to true sampling distribution
- coverage for the basic bootstrap interval is poor
- partially corrected by nonparametric double bootstrap
- fully corrected by parametric double bootstrap

Should we still avoid methods that are invalid from an asymptotic perspective?

Other Ideas

- robustness
  - integrating uncertainty
  - triple bootstrap ... and beyond
- classical inference perpetuating understated type I error?
- an insurance/trading/gambling perspective
- if only computers had been invented first ...
- the value of theoretical arguments

VANDERBILT UNIVERSITY