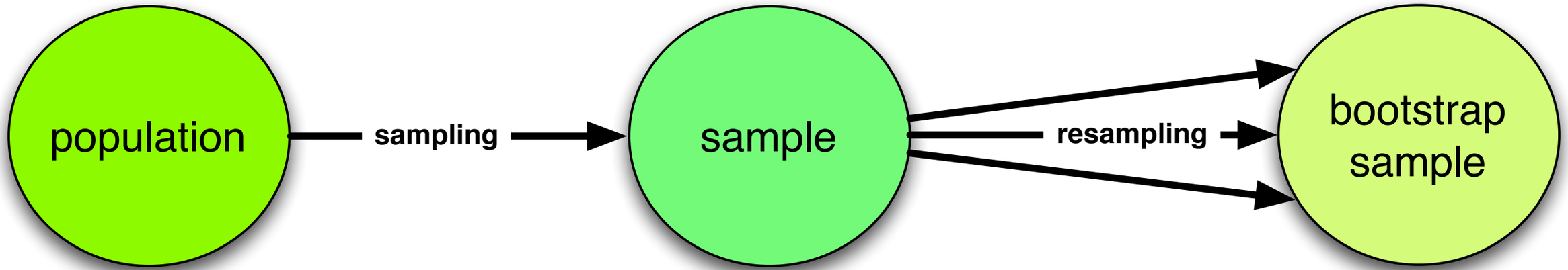
Bootstrapping

for fun and profit

Bios301 5 December 2011


population
sampling
sample
resampling
bootstrap
sample

jackknife
cross-validation
randomization test
permutation test

The Annals of Statistics
1979, Vol. 7, No. 1, 1–26

THE 1977 RIETZ LECTURE

BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

We discuss the following problem: given a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from an unknown probability distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of the observed data $\mathbf{x}$. (Standard jackknife theory gives an approximate mean and variance in the case $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, θ some parameter of interest.) A general method, called the "bootstrap," is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear discriminant analysis, ratio estimation, estimating regression parameters, etc.

1. Introduction. The Quenouille–Tukey jackknife is an intriguing nonparametric method for estimating the bias and variance of a statistic of interest, and also for testing the null hypothesis that the distribution of a statistic is centered at some prespecified point. Miller [14] gives an excellent review of the subject.

This article attempts to explain the jackknife in terms of a more primitive method, named the "bootstrap" for reasons which will become obvious. In principle, bootstrap methods are more widely applicable than the jackknife, and also more dependable. In Section 3, for example, the bootstrap is shown to (asymptotically) correctly estimate the variance of the sample median, a case where the jackknife is known to fail. Section 4 shows the bootstrap doing well at estimating the error rates in a linear discrimination problem, outperforming "cross-validation," another nonparametric estimation method.

We will show that the jackknife can be thought of as a linear expansion method (i.e., a "delta method") for approximating the bootstrap. This helps clarify the theoretical basis of the jackknife, and suggests improvements and variations likely

$$\mathbf{P} = \{x_1, x_2, \ldots, x_N\}$$

population

$$\mathbf{S} = \{x_1, x_2, \ldots, x_n\}$$

sample

$$n << N$$

$$\theta = h(\mathbf{P})$$

population parameter

$$T = h(\mathbf{S})$$

estimate

Problems


Non-robust

difficult

nonparametric

bootstrap

bootstrap sample

$$\mathbf{S}_1^* = \{x_{11}^*, x_{12}^*, \ldots, x_{1n}^*\}$$

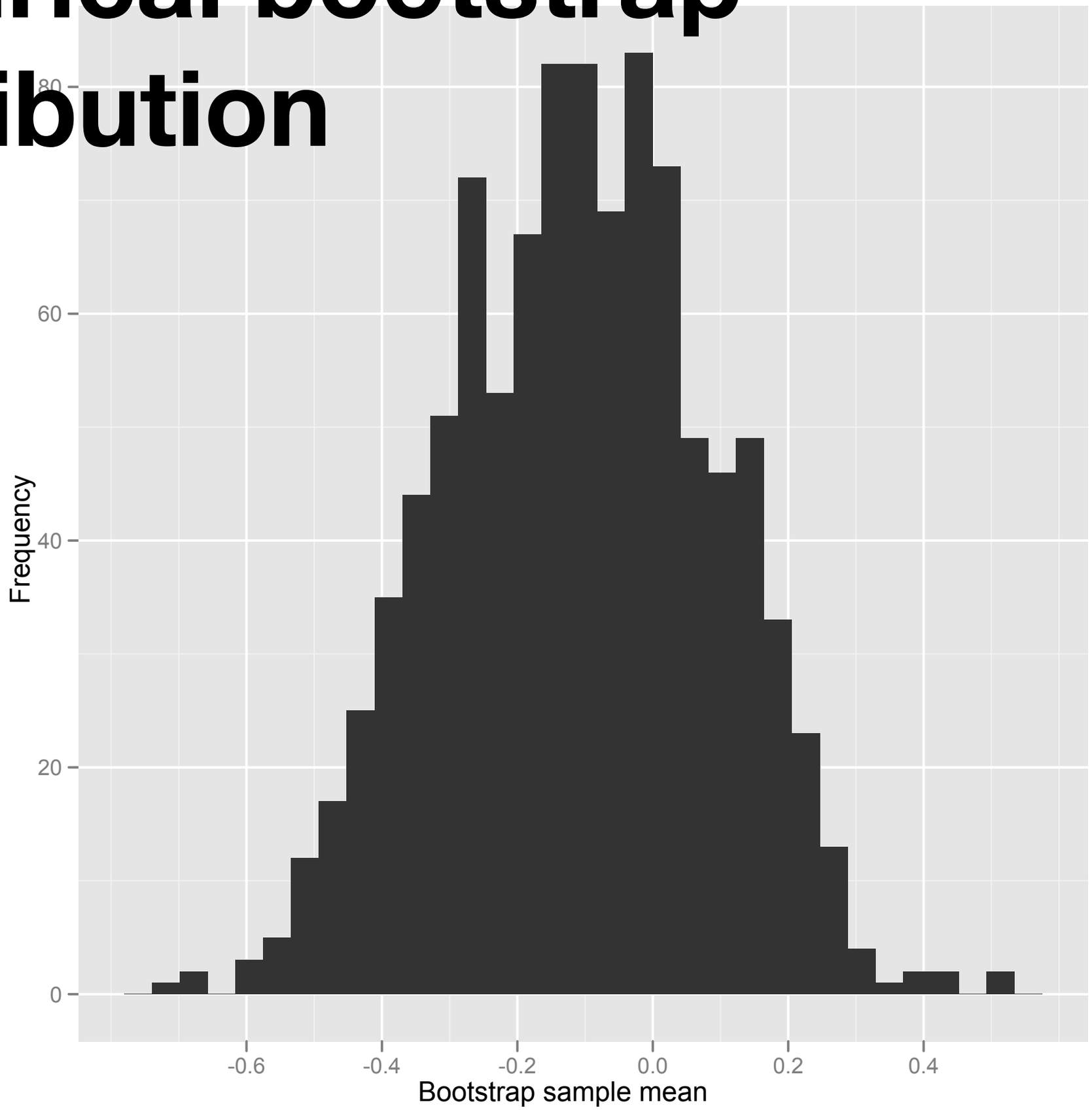
```
> x <- rnorm(10)

> x
 [1] -0.92808680  0.09901648  0.23525382
0.62914907  0.08515775 -0.42132747
 [7] -1.05194033 -0.71576518  0.21399354
0.82478246

> sample(x, 10, replace=TRUE)
 [1] -0.71576518  0.62914907 -1.05194033
-1.05194033  0.09901648 -0.92808680
 [7] -0.42132747 -0.92808680 -0.92808680
-1.05194033
```

population : sample :: sample : bootstrap sample

$$\mathbf{S}^* = \{\mathbf{S}_1^*, \mathbf{S}_2^*, \ldots, \mathbf{S}_m^*\}$$

$$T_i^* = t(\mathbf{S}_i^*)$$

empirical bootstrap distribution



$$\bar{T}^* = \hat{E}(T^*) = \frac{\sum_i T_i^*}{m}$$

$$\widehat{\mathrm{Var}}(T^*) = \frac{\sum_i (T_i^* - \bar{T}^*)^2}{m-1}$$

Bias

$$\hat{B}^* = \bar{T}^* - T$$

(estimate of $T - \theta$)

Error

(1) Sampling error

(2) Bootstrap error

bootstrap confidence intervals

$$(T - \hat{B}^*) \pm z_{1-\alpha/2} \widehat{\mathrm{SE}}^*(T^*)$$

$$\left(\widehat{\mathrm{SE}}^*(T^*) = \sqrt{\widehat{\mathrm{Var}}(T^*)} \right)$$

bootstrap percentile intervals

$$T^*_{(1)}, T^*_{(2)}, \ldots, T^*_{(m)}$$

bootstrap percentile intervals

$$T^*_{[(m+1)\alpha/2]} < \theta < T^*_{[(m+1)(1-\alpha/2)]}$$

bias corrected, accelerated percentile intervals

(1) Calculate:

$$z = \Phi^{-1}\left[\frac{\sum_{i=1}^{m} I(T_i^* \leq T)}{m+1}\right]$$

(correction factor)

bias corrected, accelerated percentile intervals

(2) Calculate:

$$a = \frac{\sum_{j=1}^{n} \left(T^*_{(-j)} - \bar{T} \right)^3}{6 \left[\sum_{j=i}^{n} \left(T^*_{(-j)} - \bar{T}^2 \right) \right]^{3/2}}$$

$$a_1 = \Phi\left[z + \frac{z - z_{1-\alpha/2}}{1 - a(z - z_{1-\alpha/2})} \right]$$

$$a_2 = \Phi\left[z + \frac{z + z_{1-\alpha/2}}{1 - a(z + z_{1-\alpha/2})} \right]$$

bias corrected, accelerated percentile intervals

(3) Calculate BC_a interval:

$$T^*_{[m \cdot a_1]} < \theta < T^*_{[m \cdot a_2]}$$

regression bootstrapping

$$y_i = \boldsymbol{X}_i \beta + \boldsymbol{\epsilon}_i$$

case resampling

$$\mathbf{X}_i^* = \{x_{i1}^*, x_{i2}^*, \ldots, x_{in}^*\}$$

$$\mathbf{y}_i^* = \{y_{i1}^*, y_{i2}^*, \ldots, y_{in}^*\}$$

$$\mathbf{y}_i^* = \mathbf{X}_i^* \beta + \epsilon_i$$

model-based resampling

$$\epsilon_i = \hat{y}_i - y_i$$

$$\epsilon_j^* = \{\epsilon_{j1}^*, \epsilon_{j2}^*, \ldots, \epsilon_{jn}^*\}$$

predictors **x** do not change in subsamples