

# Scientific Computing for Biologists

## Resampling Methods

Paul M. Magwene

April 4, 2018

# Goals of Resampling Methods

## Goal

We have estimated some statistic of interest,  $S$ , for a set of observed data. We want to know whether the value of  $S$  we estimate,  $\hat{S}$ , is likely to have been generated by chance or under a model captured by an appropriate null hypothesis.

## Approach

Randomization tests and related methods allow one to address these questions for many types of statistics that are not amenable to classical analysis.

# Advantages of Resampling Methods

In general, resampling methods are:

- Computer intensive
- Require few assumptions about the distributional properties
- Robust

# Overview of Randomization Tests

## Basic Idea

Compare  $\hat{s}$  to the distribution of estimates of  $S$  obtained by randomly reordering the data.

## Randomization tests: Example

Consider the following measurements of mandible length (in mm) from skeletons of male and female golden jackals (*Canis aureus*).

- Male: 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
- Female: 110, 111, 107, 108, 110, 105, 107, 106, 111, 111

### Question

Are the mean values for the two sexes different?

## Randomization tests: Example cont

The means for the two sexes are:

- $\bar{x}_{male} = 113.4$
- $\bar{x}_{female} = 108.6$

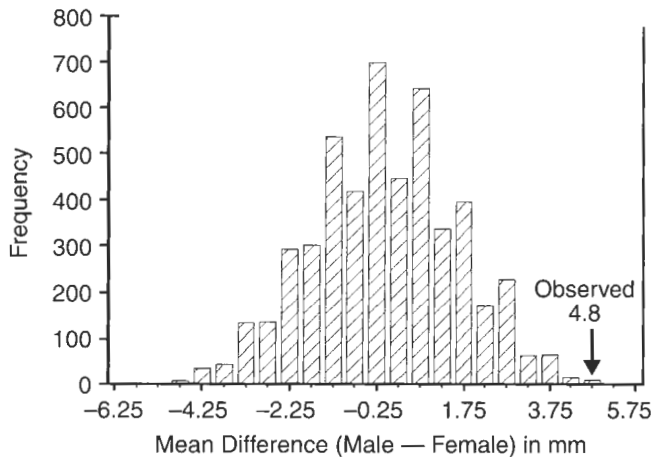
Standard deviations:

- $s_{male} = 3.72$
- $s_{female} = 2.27$ .

To construct the randomization test:

- generate a large number of samples where we randomly reallocate 10 of the specimens to the male group and the remaining 10 we designate as female.
- for each randomized sample calculate the difference in means between the male and female groups
- Examine the distribution of the randomization distribution and ask whether the observed difference in means is atypical.

## Randomization tests: Example cont



# Advantages and Limitations of Randomization Tests

## Advantages:

- Valid even without a random sample
- Can be designed to take particulars of a particular statistic into account
- When there is a classical, parametric equivalent there is often good agreement

## Limitations:

- Not possible to generalize to a population of interest



# Overview of Jackknife Methodology

## Basic Idea

Turn the problem of estimating any parameter of interest for  $n$  observations into a problem of estimating a sample mean.

## Basic Mechanics

Calculate pseudovalue of  $S$ ,  $s^*$ , leaving out a single observation at a time. Calculate mean and standard errors of pseudovalue to estimate confidence intervals for  $\hat{s}$ .

# Jackknifing I

- Assume we're interested in some arbitrarily complex statistic that is a function of the  $n$  data points:

$$\hat{\Theta} = \phi(x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)$$

- We define the  $i$ th **partial estimate** of  $\Theta$  as

$$\hat{\Theta}_i = \phi(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

- We define the  $i$ th **pseudovalue** as:

$$\widehat{\Theta}^*_i = n\hat{\Theta} - (n-1)\hat{\Theta}_i$$

We will treat the pseudovalues as if they were independent random variables with mean  $\hat{\Theta}$ .

## Jackknifing II

- From the pseudovalues we can calculate a **jackknife estimate** of  $\Theta$  as follows

$$\widehat{\Theta}^* = \frac{1}{n} \sum_{i=1}^n \widehat{\Theta}^*_i$$

- The variance of the pseudovalues,  $\widehat{\Theta}^*_i$ , is given by:

$$Var(\widehat{\Theta}^*) = \frac{(\sum \widehat{\Theta}^*_i - \widehat{\Theta}^*)^2}{n - 1}$$

- We can approximate the standard error of  $\widehat{\Theta}^*$  by calculating the standard error

$$SE_{jack} = \sqrt{\frac{Var(\widehat{\Theta}^*)}{n}}$$

- An approximate  $(1 - \alpha)\%$  confidence interval is given by

$$\widehat{\Theta}^* \pm t_{\alpha/2, n-1} SE_{jack}$$

where  $t_{\alpha/2, n-1}$  is the valued that is exceeded with probability  $\alpha/2$  for the t-distribution with  $n - 1$  degrees of freedom.

## Jackknife Estimates: Example

Suppose we have a random sample of size  $n = 20$  that consists of the following observations: 3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27, 0.50, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20, 1.21, 0.72

Let's use the jackknife to estimate confidence intervals for the standard deviation,  $\sigma$  of this sample.

- We calculate the jackknife pseudovalues,  $\widehat{\sigma}_1^*, \dots, \widehat{\sigma}_{20}^*$
- The mean of the jackknife pseudovalues,  $\widehat{\sigma}^* = 1.096$ .
- The variance of the pseudovalues,  $Var(\widehat{\Theta}^*_i) = 1.488$
- The standard error of the pseudovalues,  $\sqrt{\frac{Var(\widehat{\Theta}^*_i)}{n}} = 0.273$ .
- The approximate 95% confidence interval is:  
 $1.096 \pm 2.09 \times 0.273 = (0.53, 1.67)$

# Advantages and Limitations of Jackknife Estimates

## Advantages:

- Simple to calculate and not particularly computer intensive
- Jackknife estimates reduce bias

## Limitations:

- Works best when observed sample is moderately large
- Jackknife confidence intervals can sometimes over or underestimate the true confidence interval so simulation studies to test the robustness of the jackknife for a statistic of interest are often warranted

# Bootstrap: Overview

## Basic Idea

- In the absence of a priori knowledge about a population, the distribution of values in random samples is the best guide to the distribution of values in the population as a whole.
- If we are unable to resample the population to learn about the distribution of a statistic,  $S$ , the best proxy is the resample our original sample. Resampling is done *with replacement* in contrast to randomization tests.

## Basic Mechanics

Generate a large number of **bootstrap samples** by repeatedly resampling the original set of observations. Approximate the standard error of  $S$  based on the set of bootstrap samples. Estimate bias and/or confidence intervals based on bootstrap samples.

# Bootstrap: Sampling with replacement

- Original sample,  $\vec{x} = \{x_1, x_2, x_3, \dots, x_n\}$
- Create a large number of bootstrap samples
  - Bootstrap sample  $\vec{x}_{b_1} = \{x_8, x_1, x_4, x_1, \dots, x_2\}$
  - Bootstrap sample  $\vec{x}_{b_2} = \{x_2, x_3, x_5, x_9, \dots, x_{11}\}$
  - Bootstrap sample  $\vec{x}_{b_3} = \{x_{10}, x_5, x_1, x_{12}, \dots, x_{15}\}$
  - etc...

Each sample is of length  $n$  (size of original sample)

## Bootstrap: Calculate statistic of interest on each bootstrap sample

- $\hat{\Theta}_1 = \Theta(\vec{\mathbf{x}}_{b_1})$
- $\hat{\Theta}_2 = \Theta(\vec{\mathbf{x}}_{b_2})$
- $\hat{\Theta}_3 = \Theta(\vec{\mathbf{x}}_{b_3})$
- ... etc ...



# Bootstrap: Bootstrap distribution

- The bootstrap distribution,  $\hat{\Theta}$ , is an estimator of the sampling distribution of the statistic of interest.
- The standard deviation of the bootstrap distribution is thus an estimator of the standard error of the statistic of interest.

## Bootstrap: Bootstrap standard confidence limits

Simple method for obtaining bootstrap confidence limits is to assume that  $\hat{\Theta}$  has an approximately normal distribution and the bootstrap sampling gives a good approximation to the sampling of the statistic of interest.

Standard Bootstrap Confidence Limits:

$$\text{Estimate} \pm t_{\alpha/2}(\text{bootstrap standard deviation})$$

where 'Estimate' is the estimate of the parameter of interest based on the initial sample and the 'bootstrap standard deviation' is based on the variance of the bootstrap samples.

# Bootstrap: Percentile confidence limits

Another approach is to calculate confidence intervals directly from the bootstrap samples themselves.

- Makes no assumptions about normality
- Requires fairly large bootstrap samples ( $\geq 1000$  for 95% CIs,  $\geq 5000$  for 99% CIs)

Efron's percentile confidence limits:

- $(\hat{\Theta}_{L,\alpha/2}, \hat{\Theta}_{H,\alpha/2})$  where  $\hat{\Theta}_{L,\alpha/2}$  is the estimate of  $\Theta$  from the bootstrap distribution such that a fraction  $\alpha/2$  of all bootstrap estimates are less than this value, likewise for the upper limit  $\hat{\Theta}_{H,\alpha/2}$

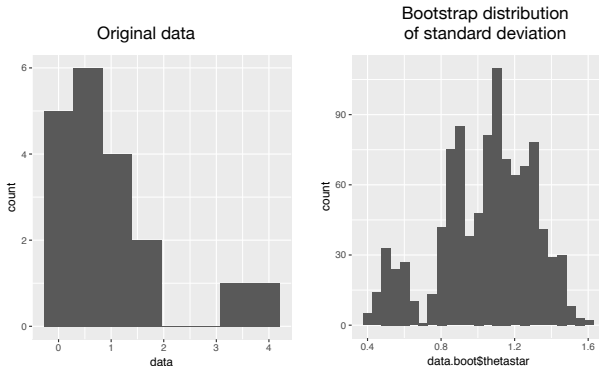
## Bootstrap: Example

Consider the same sample used to illustrate the jackknife with observations: 3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27, 0.50, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20, 1.21, 0.72

Let's use the bootstrap to estimate confidence intervals for the standard deviation of the distribution this sample is drawn from.

- the sample estimate of the standard deviation is,  $\hat{\sigma} = 1.06$
- generated 1000 bootstrap samples
  - mean of the bootstrap samples = 0.97
  - standard error of the bootstrap samples = 0.25

# Bootstrap: Example, cont.



- Standard 95% confidence limits for the standard deviation:  
 $1.06 \pm 2.10 \times (0.25) = (0.54, 1.59)$

# Bootstrap: Advantages and Limitations

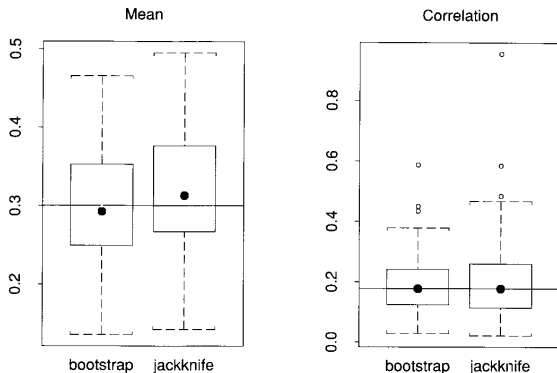
## Advantages:

- Robust
- Can be applied to arbitrary statistics of interest

## Limitations:

- Works best when observed sample is moderately large
- Requires a fair amount of computing power (not typically a problem these days)
- Variety of complex procedures for estimating confidence intervals, none clearly preferable in all situations

# Bootstrap vs. Jackknife



- Jackknife can be viewed as approximation to bootstrap
- Jackknife less computationally intensive to calculate
- Jackknife can fail if the statistic of interest is not 'smooth' (small changes in data cause only small changes in the statistic) - e.g. median