Scientific Computing for Biologists

Lecture 11: Resampling methods and other non-parametric appraoches

Instructor: Paul M. Magwene

13 November 2012

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 $\widehat{s}$ 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, 11, 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:

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

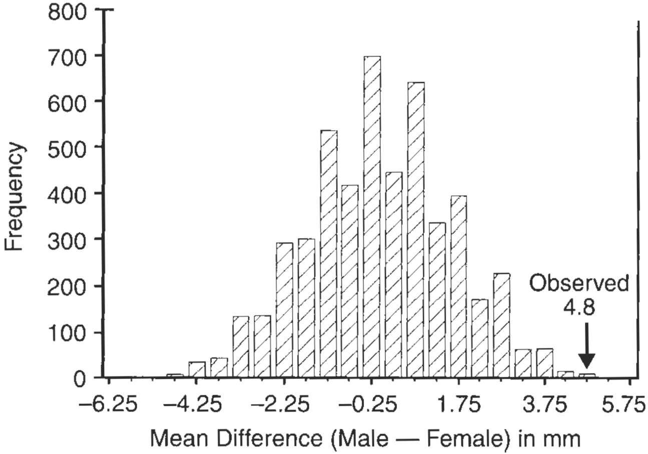
Standard deviations:

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

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

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

Motivation for Jackknifing

- Let $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ denote the mean for a sample of size n.
- Let
$$\bar{x}_{-j} = \frac{1}{n-1} \sum_{i \neq 1}^{n} x_i$$
denote the sample mean with the jth observation removed.
- If we know both $\bar{x}$ and $\bar{x}_{-j}$ we can compute the value of the jth point as
$$x_j = n\bar{x} - (n-1)\bar{x}_{-j}$$

Motivation for Jackknifing II

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

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

- We define the jth **partial estimate** of Θ as

$$\widehat{\Theta}_j = \phi(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots x_n)$$

- By analogy with the previous formula we define the jth **pseudovalue** as:

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

Motivation for Jackknifing III

- From the pseudovalues we can calculate a **jackknife estimate** of Θ as follows
$$\widehat{\Theta*} = \frac{1}{n}\sum_{i=1} n\widehat{\Theta*}_i$$

- We can approximate the standard error of $\widehat{\Theta*}$ by calculating the standard error
$$SE_{jack} = \sqrt{\frac{Var(\widehat{\Theta*}_j)}{n}}$$

- 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, σ of this sample.

- We calculate the jackknife pseudovalues, $\widehat{\sigma_1^*}, \ldots, \widehat{\sigma_{20}^*}$
- The mean of the jackknife pseudovalues, $\widehat{\sigma^*} = 1.096$.
- The variance of the pseudovalues, $Var(\widehat{\Theta^*}_j) = 1.488$
- The standard error of the pseudovalues, $\sqrt{\frac{Var(\widehat{\Theta^*}_j)}{n}} = 0.273$.
- 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: Standard confidence limits

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

Standard Bootstrap Confidence Limits:

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

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

Calculate confidence intervals directly from the bootstrap samples themselves.

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

Efron's percentile confidence limits:

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

Bootstrap: Examples

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 the confidence interval 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
- Standard 95% confidence limits:
  $1.06 \pm 1.96(0.25) = (0.57, 1.55)$

Bootstrap: Advantages and Limitations

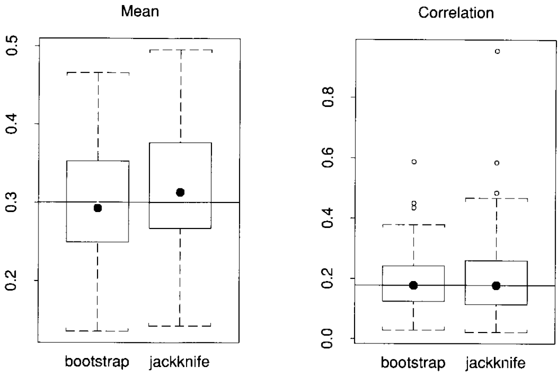
Advantages:

- 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

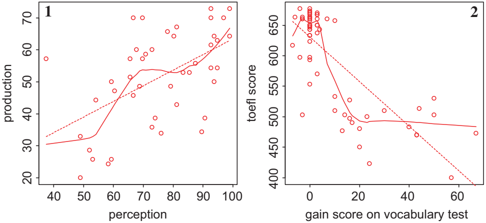
Bootstrap vs. Jackknife



- 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)

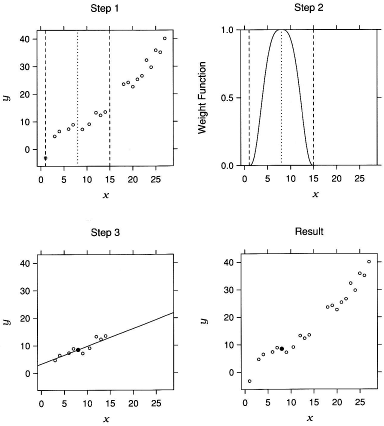
LOESS Regression

- A type of non-parametric regression
- Basic idea – fit a curve (or surface) to a set of data by fitting a large number of *local regressions*.
- Cleveland, W.S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots". Journal of the American Statistical Association 74 (368): 829-836. doi:10.2307/2286407.



Graphical overview of LOESS fitting, I

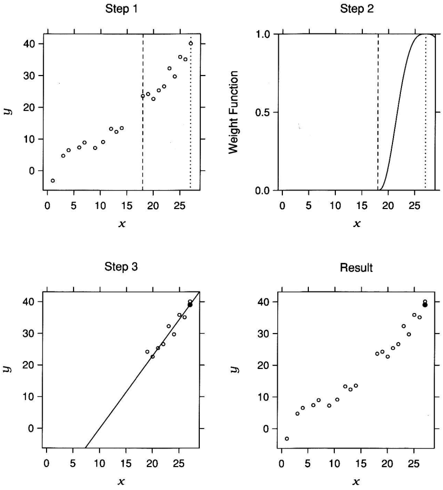
from Cleveland (1993)



3.49 HOW LOESS WORKS. The graphs show how the initial fit at x = 8 is computed. (Top left) α, which is 0.5, is multiplied by 20, the number of points, which gives 10. A vertical strip is defined around x = 8 so that one boundary is at the 10th nearest neighbor. (Top right) Weights are defined for the points using the weight function. (Bottom left) A line is fitted using weighted least-squares. The value of the line at x = 8 is the initial loess fit at x = 8. (Bottom right) The result is one point of the initial loess curve, shown by the filled circle.

Graphical overview of LOESS fitting, II

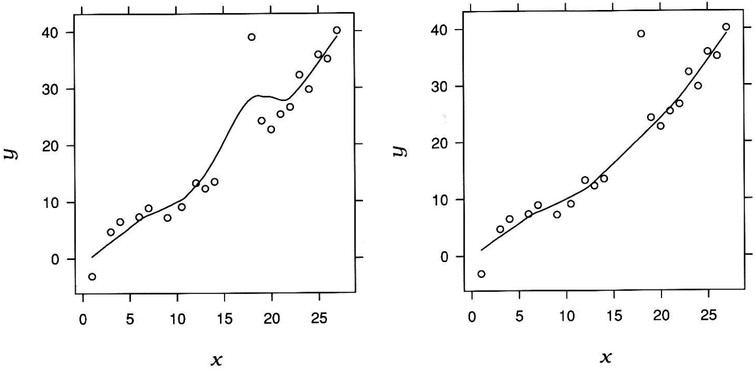
from Cleveland (1993)



3.50 HOW LOESS WORKS. The computation of the initial loess fit value at $x = 27$ is illustrated.

Graphical overview of LOESS fitting, III

from Cleveland (1993)



3.51 HOW LOESS WORKS. Loess employs robustness iterations that prevent outliers from distorting the fit. (Left panel) The open circles are the points of the graph; there is one outlier between x = 15 and x = 20. The initial loess curve has been distorted in the neighborhood of the outlier. (Right panel) The graphed curve is the fit after four robustness iterations. Now the fit follows the general pattern of the data.