

Bootstrap

G. Jogesh Babu

Penn State University
<http://www.stat.psu.edu/~babu>

Director of Center for Astrostatistics

<http://astrostatistics.psu.edu>

Outline

- 1 Motivation
- 2 Simple statistical problem
- 3 Resampling
- 4 What is bootstrap
- 5 Regression
- 6 Fortran code
- 7 References

Motivation

- It is often relatively easy to devise a statistic (estimator of a parameter) that measures the property of interest, but is difficult or impossible to determine the distribution or variance (sampling variability) of that statistic.
- One might fit a parametric model to the dataset, yet not be able to assign confidence intervals to see how accurately the parameters are determined.
- In the past statisticians concentrated on estimators which have a simple closed form and which could be analyzed mathematically. Except for a few important but simple nonparametric statistics, these methods involve often unrealistic assumptions about the data; e.g. that it is generated from a Gaussian or exponential population.

Simple Statistical Problem

X_1, \dots, X_n are random variables from a distribution F with mean μ and variance σ^2 .

Sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ estimates the population mean μ

Data vs. Sampling distribution of \bar{X}

Sampling (unknown) distribution of $\bar{X} - \mu$

$$G_n(x) = P(\bar{X} - \mu \leq x)$$

If F is normal, then G_n is normal. Otherwise, for large n

$$G_n(x\sigma/\sqrt{n}) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}y^2} dy$$

G_n may not be symmetric in the non-normal case.

How to improve the approximation?

Resampling

- Resampling methods help evaluate statistical properties using data rather than an assumed Gaussian or power law or other distributions.
- Resampling methods construct hypothetical ‘populations’ derived from the observed data, each of which can be analyzed in the same way to see how the statistics depend on plausible random variations in the data.
- Astronomers have often used *Monte Carlo methods* to simulate datasets from uniform or Gaussian populations. While helpful in some cases, this does not avoid the assumption of a simple underlying distribution.

Resampling

- Resampling the original data preserves (adaptively) whatever distributions are truly present, including selection effects such as truncation (flux limits or saturation).
- Resampling procedure is a Monte Carlo method of simulating datasets from an existing dataset, without any assumption on the underlying population.
- Resampling procedures are supported by solid theoretical foundation.

What is Bootstrap

$\mathbf{X} = (X_1, \dots, X_n)$ - a sample from F

$\mathbf{X}^* = (X_1^*, \dots, X_n^*)$ - a simple random sample from the data.

$\hat{\theta}$ is an estimator of θ

θ^* is based on X_i^*

Examples:

$$\begin{aligned} \hat{\theta} &= \bar{X}, & \theta^* &= \bar{X}^* \\ \hat{\theta} &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, & \theta^* &= \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2 \\ \theta^* - \hat{\theta} & \text{ behaves like } & \hat{\theta} - \theta \end{aligned}$$

Correlation Coefficient

Sample correlation coefficient $\hat{\rho}$

$$\hat{\rho} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i Y_i - \bar{X}_n \bar{Y}_n)}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right) \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2\right)}}$$

Its bootstrap version

$$\rho^* = \frac{\frac{1}{n} \sum_{i=1}^n (X_i^* Y_i^* - \bar{X}_n^* \bar{Y}_n^*)}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2\right) \left(\frac{1}{n} \sum_{i=1}^n (Y_i^* - \bar{Y}_n^*)^2\right)}}$$

Statistical inference requires sampling distribution, G_n given by $G_n(x) = P(T_n \leq x)$

$$\begin{array}{cc} T_n & T_n^* \\ \sqrt{n}(\bar{X} - \mu)/\sigma & \sqrt{n}(\bar{X}^* - \bar{X})/s_n \\ \sqrt{n}(\bar{X} - \mu)/s_n & \sqrt{n}(\bar{X}^* - \bar{X})/s_n^* \end{array}$$

where $s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ and $s_n^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}^*)^2$

Bootstrap distribution (Histogram) G_B given the data

$$G_B(x) = P(T_n^* \leq x | \mathbf{X})$$

$G_n(x) \approx G_B(x)$, G_B is completely known

Example

G_n denotes the sampling distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$

$$G_n(x) = P(\sqrt{n}(\bar{X} - \mu)/\sigma \leq x)$$

G_B is the corresponding *bootstrap distribution* (Histogram) given the data

$$G_B(x) = P(\sqrt{n}(\bar{X}^* - \bar{X})/s_n \leq x | \mathbf{X})$$

$$G_n(x) \approx G_B(x),$$

G_B is completely known

Bootstrap Distribution

$M = n^n$ bootstrap samples possible

$$X_1^{*(1)}, \dots, X_n^{*(1)} \quad r_1 = \sqrt{n}(\bar{X}^{*(1)} - \bar{X})/s_n$$

$$X_1^{*(2)}, \dots, X_n^{*(2)} \quad r_2 = \sqrt{n}(\bar{X}^{*(2)} - \bar{X})/s_n$$

$\dots \quad \dots \quad \dots \quad \dots$

$$X_1^{*(M)}, \dots, X_n^{*(M)} \quad r_M = \sqrt{n}(\bar{X}^{*(M)} - \bar{X})/s_n$$

Frequency table or histogram based on r_1, \dots, r_M gives G_B

For $n = 10$ data points, $M =$ ten billion

$N \sim n(\log n)^2$ bootstrap replications suffice

– Babu and Singh (1983) Ann Stat

Confidence Intervals

Compute

$$\sqrt{n}(\bar{X}^{*(j)} - \bar{X})/s_n$$

for N bootstrap samples

Arrange them in increasing order

$$r_1 < r_2 < \cdots < r_N \quad k = [0.05N], \quad m = [0.95N]$$

90% Confidence Interval for μ is

$$\bar{X} - r_m \frac{s_n}{\sqrt{n}} \leq \mu < \bar{X} - r_k \frac{s_n}{\sqrt{n}}$$

Bootstrap at its best

Pearson's correlation coefficient

$$\hat{\rho} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i Y_i - \bar{X} \bar{Y})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2\right)}}$$

Smooth function model

$\hat{\rho} = H(\bar{\mathbf{Z}})$, where $\mathbf{Z}_i = (X_i Y_i, X_i^2, Y_i^2, X_i, Y_i)$

$$H(a_1, a_2, a_3, a_4, a_5) = \frac{(a_1 - a_4 a_5)}{\sqrt{((a_2 - a_4^2)(a_3 - a_5^2))}}$$
$$\mathbf{Z}_i^* = (X_i^* Y_i^*, X_i^{*2}, Y_i^{*2}, X_i^*, Y_i^*)$$
$$\rho^* = H(\bar{\mathbf{Z}}^*)$$

Studentization

Functions of random variables or statistics are often normalized (divided) by the standard deviation to make these units free. When standard deviations are estimated, the normalization is known as studentization.

$$t_n = \sqrt{n}(H(\bar{\mathbf{Z}}) - H(\mathbb{E}(\mathbf{Z}_1)))/\hat{\sigma}_n$$

$$t_n^* = \sqrt{n}(H(\bar{\mathbf{Z}}^*) - H(\bar{\mathbf{Z}}))/\sigma_n^*$$

$\hat{\sigma}_n^2 = \ell'(\bar{\mathbf{Z}})\Sigma_n\ell(\bar{\mathbf{Z}})$ and $\sigma_n^{*2} = \ell'(\bar{\mathbf{Z}}^*)\Sigma_n^*\ell(\bar{\mathbf{Z}}^*)$ are the variances of the numerators.

$\ell = \partial H$ vector of first partial derivatives of H

Σ_n sample dispersion of \mathbf{Z}

Σ_n^* dispersion of bootstrap sample \mathbf{Z}^*

$\hat{\theta} = H(\bar{\mathbf{Z}})$ is an estimator of the parameter $\theta = H(\mathbb{E}(\mathbf{Z}_1))$

Randomly choose $N \sim n(\log n)^2$ bootstrap samples

Compute $t_n^{*(j)}$ for each

Arrange them in increasing order

$u_1 < u_2 < \dots < u_N$ $k = [0.05N]$, $m = [0.95N]$

90% Confidence Interval for the parameter θ is

$$\hat{\theta} - u_m \frac{\hat{\sigma}_n}{\sqrt{n}} \leq \theta < \hat{\theta} - r_k \frac{\hat{\sigma}_n}{\sqrt{n}}$$

This is called bootstrap PERCENTILE - t confidence interval

Theory

Under $\ell(\bar{\mathbf{Z}}) \neq 0$

$$P(t_n \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}p(x)\phi(x) + \text{error}$$

$$P^*(t_n^* \leq x) = \Phi(x) + \frac{1}{\sqrt{n}}p_n(x)\phi(x) + \text{error}$$

$$\sqrt{n}|P(t_n \leq x) - P^*(t_n^* \leq x)| \rightarrow 0$$

\hat{F}_n an estimator of F we could Bootstrap from \hat{F}_n

If $F \sim N(\mu, \sigma^2)$, then $\hat{F}_n \sim N(\hat{\mu}, \hat{\sigma}^2)$

Same theory works.

- Babu and Singh (1983) Ann Stat
- Babu and Singh (1984) Sankhyā
- Babu and Singh (1990) Scand J. Stat

When does bootstrap work well

- Sample Means
- Sample Variances
- Central and Non-central t-statistics
(with possibly non-normal populations)
- Sample Coefficient of Variation
- Maximum Likelihood Estimators
- Least Squares Estimators
- Correlation Coefficients
- Regression Coefficients
- Smooth transforms of these statistics

When does Bootstrap fail

- $\hat{\theta} = \max_{1 \leq i \leq n} X_i$ non-smooth statistic
 - Bickel and Freedman (1981) Ann. Stat.

When does Bootstrap fail

- $\hat{\theta} = \max_{1 \leq i \leq n} X_i$ non-smooth statistic
 - Bickel and Freedman (1981) Ann. Stat.
- $\hat{\theta} = \bar{X}$ and $EX_1^2 = \infty$ heavy tails
 - Babu (1984) Sankhyā
 - Athreya (1987) Ann. Stat.

When does Bootstrap fail

- $\hat{\theta} = \max_{1 \leq i \leq n} X_i$ non-smooth statistic
 - Bickel and Freedman (1981) Ann. Stat.
- $\hat{\theta} = \bar{X}$ and $EX_1^2 = \infty$ heavy tails
 - Babu (1984) Sankhyā
 - Athreya (1987) Ann. Stat.
- $\hat{\theta} - \theta = H(\bar{\mathbf{Z}}) - H(E(\mathbf{Z}_1))$ and $\partial H(E(\mathbf{Z}_1)) = 0$

Limit distribution is like linear combinations of Chi-squares. But here a modified version works

 - Babu (1984) Sankhyā.

Linear Regression

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$E(\epsilon_i) = 0 \text{ and } \text{Var}(\epsilon_i) = \sigma_i^2$$

Least squares estimators of β and α

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 \sigma_i^2}{L_n^2}$$

$$L_n = \sum_{i=1}^n (X_i - \bar{X})^2$$

Classical Bootstrap

Estimate the residuals $e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$

$$\hat{e}_i = e_i - \frac{1}{n} \sum_{j=1}^n e_j$$

Draw e_1^*, \dots, e_n^* from $\hat{e}_1, \dots, \hat{e}_n$

Bootstrap estimators

$$\beta^* = \hat{\beta} + \frac{\sum_{i=1}^n (X_i - \bar{X})(e_i^* - \bar{e}^*)}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\alpha^* = \hat{\alpha} + (\hat{\beta} - \beta^*)\bar{X} + \bar{e}^*$$

$$E_B(\beta^* - \hat{\beta})^2 \approx \text{Var}(\hat{\beta})$$

Efficient if $\sigma_i = \sigma$

V_B does not approximate the variance of $\hat{\beta}$ under heteroscedasticity (*i.e.* unequal variances σ_i)

Paired Bootstrap

Resample the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$
 $(\tilde{X}_1, \tilde{Y}_1), \dots, (\tilde{X}_n, \tilde{Y}_n)$

$$\tilde{\beta} = \frac{\sum_{i=1}^n (\tilde{X}_i - \tilde{\bar{X}})(\tilde{Y}_i - \tilde{\bar{Y}})}{\sum_{i=1}^n (\tilde{X}_i - \tilde{\bar{X}})^2}, \quad \tilde{\alpha} = \tilde{\bar{Y}} - \tilde{\beta}\tilde{\bar{X}}$$

Repeat the resampling N times and get

$$\beta_{PB}^{(1)}, \dots, \beta_{PB}^{(N)}$$

$$\frac{1}{N} \sum_{i=1}^N (\beta_{PB}^{(i)} - \hat{\beta})^2 \approx \text{Var}(\hat{\beta})$$

even when not all σ_i are the same

FORTRAN code

```
PAIRED BOOTSTRAP RESAMPLING
NSIM = INT(N * ALOG(FLOAT(N))**2)
DO 20 ISIM = 1,NSIM
DO 10 I = 1,N
J = INT(RANDOM * N + 1.0)
XBOOT(I) = X(J)
10 YBOOT(I) = Y(J)
20 CONTINUE
```

FORTRAN code illustrating the paired bootstrap resampling for a two dimensional dataset $(x_i, y_i), i = 1, \dots, N$.

Comparison

- **The Classical Bootstrap**

- Efficient when $\sigma_i = \sigma$
- But inconsistent when σ_i 's differ

- **The Paired Bootstrap**

- Robust against heteroscedasticity
- Works well even when σ_i are all different

References

G. J. Babu and C. R. Rao (1993). *Bootstrap Methodology*, Handbook of Statistics, Vol 9, Chapter 19.

Michael R. Chernick (2007). *Bootstrap Methods - A guide for Practitioners and Researchers*, (2nd Ed.) Wiley Inter-Science.

Abdelhak M. Zoubir and D. Robert Iskander (2004). *Bootstrap Techniques for Signal Processing*, Cambridge University Press.

It is a handbook on 'bootstrap' for engineers, to analyze complicated data with little or no model assumptions. Bootstrap has found many applications including, artificial neural networks, biomedical engineering, environmental engineering, image processing, and Radar and sonar signal processing. Majority of the applications are taken from signal processing literature.