

# Bootstrap

G. Jogesh Babu

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

Director of Center for Astrostatistics

<http://astrostatistics.psu.edu>

- 1 Motivation
- 2 Resampling
  - Jackknife
  - Bootstrap
- 3 Bootstrap Variance
- 4 Bootstrap distribution
- 5 Bootstrap at its best
- 6 Confidence Interval
- 7 When does bootstrap fail
- 8 Linear Regression

# Motivation

- It is often relatively easy to devise an estimator  $\hat{\theta}$  of a parameter  $\theta$  of interest, but it is difficult or impossible to determine the distribution or variance (sampling variability) of that estimator. Variance helps in assessing the accuracy of the estimators.
- 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.
- Classical statistics focused 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  $\mu$  and variance  $\sigma^2$ .

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

Data vs. Sampling distribution of  $\bar{X}$

Sampling (unknown) distribution  $G_n$  of  $\bar{X} - \mu$  is given by

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

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

- 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 methods construct hypothetical 'populations' derived from the observed data, each of which can be analyzed in the same way to see how the estimates depend on plausible random variations in the data.
- Resampling methods help evaluate statistical properties using data rather than an assumed Gaussian or power law or other distributions.

# Monte Carlo simulation from data

- 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 observed/given data, without any assumption on the underlying population.
- Resampling procedures are supported by solid theoretical foundations.

Bias reduction

Estimation of Variance

$\hat{\theta}$  estimator of  $\theta$

*Jackknife* estimation of variance of  $\hat{\theta}$ :

Estimate  $\hat{\theta}_{-i}$  from  $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$

$$\text{Var}_J(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{-i} - \hat{\theta})^2$$

In general  $\text{Var}_J(\hat{\theta}) \approx \text{Var}(\hat{\theta})$ ; but not always

Example:  $\hat{\theta} = \text{Sample median}$

# What is Bootstrap

Bootstrap is a resampling procedure.

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

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

$\hat{\theta}$  is an estimator of  $\theta$

$\theta^*$  is based on  $X_i^*$

Examples:

$$\hat{\theta} = \bar{X}_n, \quad \theta^* = \bar{X}_n^*$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad \theta^* = \frac{1}{n} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$$

$$\theta^* - \hat{\theta} \quad \text{behaves like} \quad \hat{\theta} - \theta$$



# Nonparametric and Parametric Bootstrap

Simple random sampling from data is equivalent to drawing a set of i.i.d. random variables from the empirical distribution. This is **Nonparametric Bootstrap**.

**Parametric Bootstrap** if  $X_1^*, \dots, X_n^*$  are i.i.d. r.v. from  $\hat{H}_n$ , an estimator of  $F$  based on data  $(X_1, \dots, X_n)$ .

Example of Parametric Bootstrap:

$$X_1, \dots, X_n \text{ i.i.d. } \sim N(\mu, \sigma^2)$$

$$X_1^*, \dots, X_n^* \text{ i.i.d. } \sim N(\bar{X}_n, s_n^2); \quad s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

$N(\bar{X}_n, s_n^2)$  is a good estimator of the distribution  $N(\mu, \sigma^2)$

# Bootstrap Variance

$\hat{\theta}$  is an estimator of  $\theta$  based on  $X_1, \dots, X_n$ .

$\theta^*$  denotes the bootstrap estimator based on  $X_1^*, \dots, X_n^*$ .

$$\text{Var}^*(\hat{\theta}) = E^*(\theta^* - E(\theta^*))^2$$

In practice, generate  $N$  bootstrap samples of size  $n$ .

Compute  $\theta_1^*, \dots, \theta_N^*$  for each of the  $N$  samples.

$$\bar{\theta}^* = \frac{1}{N} \sum_{i=1}^N \theta_i^*$$

$$\text{Var}(\hat{\theta}) \approx \frac{1}{N} \sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2$$

# Bootstrap Distribution

Statistical inference requires sampling distribution  $G_n$ , given by  $G_n(x) = P(\sqrt{n}(\bar{X} - \mu)/\sigma \leq x)$

statistic	bootstrap version
$\sqrt{n}(\bar{X} - \mu)/\sigma$	$\sqrt{n}(\bar{X}^* - \bar{X})/s_n$
$\sqrt{n}(\bar{X} - \mu)/s_n$	$\sqrt{n}(\bar{X}^* - \bar{X})/s_n^*$

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

For a given data, the bootstrap distribution  $G_B$  is given by

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

$G_B$  is completely known and  $G_n \approx G_B$ .

## Example

If  $G_n$  denotes the sampling distribution of  $\sqrt{n}(\bar{X} - \mu)/\sigma$  then the corresponding *bootstrap distribution*  $G_B$  is given by

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

### Construction of Bootstrap Histogram

$M = n^n$  bootstrap samples possible

$X_1^{*(1)}, \dots, X_n^{*(1)}$	$r_1 = \sqrt{n}(\bar{X}^{*(1)} - \bar{X})/s_n$
$X_1^{*(2)}, \dots, X_n^{*(2)}$	$r_2 = \sqrt{n}(\bar{X}^{*(2)} - \bar{X})/s_n$
$\ddots \quad \ddots$	$\ddots \quad \ddots$
$X_1^{*(M)}, \dots, X_n^{*(M)}$	$r_M = \sqrt{n}(\bar{X}^{*(M)} - \bar{X})/s_n$

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

# Confidence Interval for the mean

For  $n = 10$  data points,  $M = \text{ten billion}$

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

– Babu and Singh (1983) Ann. Stat.

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  $\mu$  is

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

Pearson's correlation coefficient and its bootstrap version

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

## Example of Smooth Functional Model

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

$$H(a_1, a_2, a_3, a_4, a_5) = \frac{(a_1 - a_4 a_5)}{\sqrt{((a_2 - a_4^2)(a_3 - a_5^2))}}$$

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

# Smooth Functional Model: General case

$H$  is a smooth function and  $\mathbf{Z}_1$  is a random vector.

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

Division (normalization) of  $\sqrt{n}(H(\bar{\mathbf{Z}}) - H(\mathbb{E}(\mathbf{Z}_1)))$  by its standard deviation makes them units free.

Studentization, if estimates of standard deviations are used.

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

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

$$\hat{\sigma}_n^2 = \ell'(\bar{\mathbf{Z}})\Sigma_n\ell(\bar{\mathbf{Z}}) \text{ and } \sigma_n^{*2} = \ell'(\bar{\mathbf{Z}}^*)\Sigma_n^*\ell(\bar{\mathbf{Z}}^*)$$

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

$\Sigma_n$  sample covariance matrix of  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$

$\Sigma_n^*$  covariance matrix of bootstrap sample  $\mathbf{Z}_1^*, \dots, \mathbf{Z}_n^*$

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

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

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

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

Same theory works for *Parametric Bootstrap*.

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



# Bootstrap Percentile- $t$ Confidence Interval

In practice

- Randomly generate  $N \sim n(\log n)^2$  bootstrap samples
- Compute  $t_n^{*(j)}$  for each bootstrap sample
- Arrange them in increasing order  
 $u_1 < u_2 < \dots < u_N$ ,  $k = [0.05N]$ ,  $m = [0.95N]$
- 90% Confidence Interval for the parameter  $\theta$  is

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

This is called bootstrap PERCENTILE- $t$  confidence interval

# 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

$$\hat{\theta} = \max_{1 \leq i \leq n} X_i \quad \text{Non-smooth estimator}$$

– Bickel and Freedman (1981) Ann. Stat.

# When does Bootstrap fail

$$\hat{\theta} = \max_{1 \leq i \leq n} X_i \quad \text{Non-smooth estimator}$$

- Bickel and Freedman (1981) Ann. Stat.

$$\hat{\theta} = \bar{X} \text{ and } EX_1^2 = \infty \quad \text{Heavy tails}$$

- Babu (1984) Sankhyā
- Athreya (1987) Ann. Stat.

# When does Bootstrap fail

$$\hat{\theta} = \max_{1 \leq i \leq n} X_i \quad \text{Non-smooth estimator}$$

- Bickel and Freedman (1981) Ann. Stat.

$$\hat{\theta} = \bar{X} \text{ and } EX_1^2 = \infty \quad \text{Heavy tails}$$

- Babu (1984) Sankhyā
- Athreya (1987) Ann. Stat.

$$\hat{\theta} - \theta = H(\bar{\mathbf{Z}}) - H(\mathbb{E}(\mathbf{Z}_1)) \text{ and } \partial H(\mathbb{E}(\mathbf{Z}_1)) = 0$$

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  $\beta$  and  $\alpha$

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

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

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

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

# Classical Bootstrap

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

Draw  $e_1^*, \dots, e_n^*$  from  $\hat{e}_1, \dots, \hat{e}_n$ , where  $\hat{e}_i = e_i - \frac{1}{n} \sum_{j=1}^n e_j$ .

Bootstrap estimators

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

$V_B = E_B(\beta^* - \hat{\beta})^2 \approx \text{Var}(\hat{\beta})$  efficient if  $\sigma_i = \sigma$

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

## Paired Bootstrap

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

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

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  $\sigma_i$  are the same



- *The Classical Bootstrap*
  - Efficient when  $\sigma_i = \sigma$
  - But inconsistent when  $\sigma_i$ 's differ
  
- *The Paired Bootstrap*
  - Robust against heteroscedasticity
  - Works well even when  $\sigma_i$  are all different

```
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  $(x_i, y_i), i = 1, \dots, N$ .

## References

G. J. Babu and C. R. Rao (1993)

*Bootstrap Methodology*, Handbook of Statistics, Vol **9**, Ch. 19.

Michael R. Chernick (2007).

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

Michael R. Chernick and Robert A. LaBudde (2011)

*An Introduction to Bootstrap Methods with Applications to R*, Wiley.

Abdelhak M. Zoubir and D. Robert Iskander (2004)

*Bootstrap Techniques for Signal Processing*, Cambridge Univ Press.

A handbook on 'bootstrap' for engineers to analyze complicated data with little or no model assumptions. Includes applications to radar and sonar signal processing.