# Primer on Analysis of Experimental Data and Design of Experiments

## *Lecture 7. Bootstrap, Cross-Validation, and Goodness of Fit*

Muhammad A. Alam
alam@purdue.edu

# Course Outline

$$\bar{y} = f(\bar{x}) \qquad \bar{x} = x_1, x_2, \dots x_n \qquad \bar{y} = y_1, y_2, \dots y_m$$

Lecture 1: Introduction

Lecture 2: Collecting and plotting $x_1, x_2, \dots x_n$

Lecture 3: Physical and empirical $f, F, df/dx, \dots$

Lecture 4: Model selection between $f_1, f_2, \dots$

Lecture 7: Model Selection: Cross-validation and Bootstrapping method

Lecture 5: Scaling theory with known $f$, $f(\bar{x}) = f(\bar{X})$

Lecture 6: Scaling theory with unknown $f$, $\bar{x} \rightarrow X$

Lecture 8: Design of experiments to determine $\bar{y}_{\max} = f(\bar{x})$

Lecture 9: Machine learning … Statistical approach to learn f

Lecture 10: Physics-based machine learning $f = f_{\text{physics}} + \Delta f$

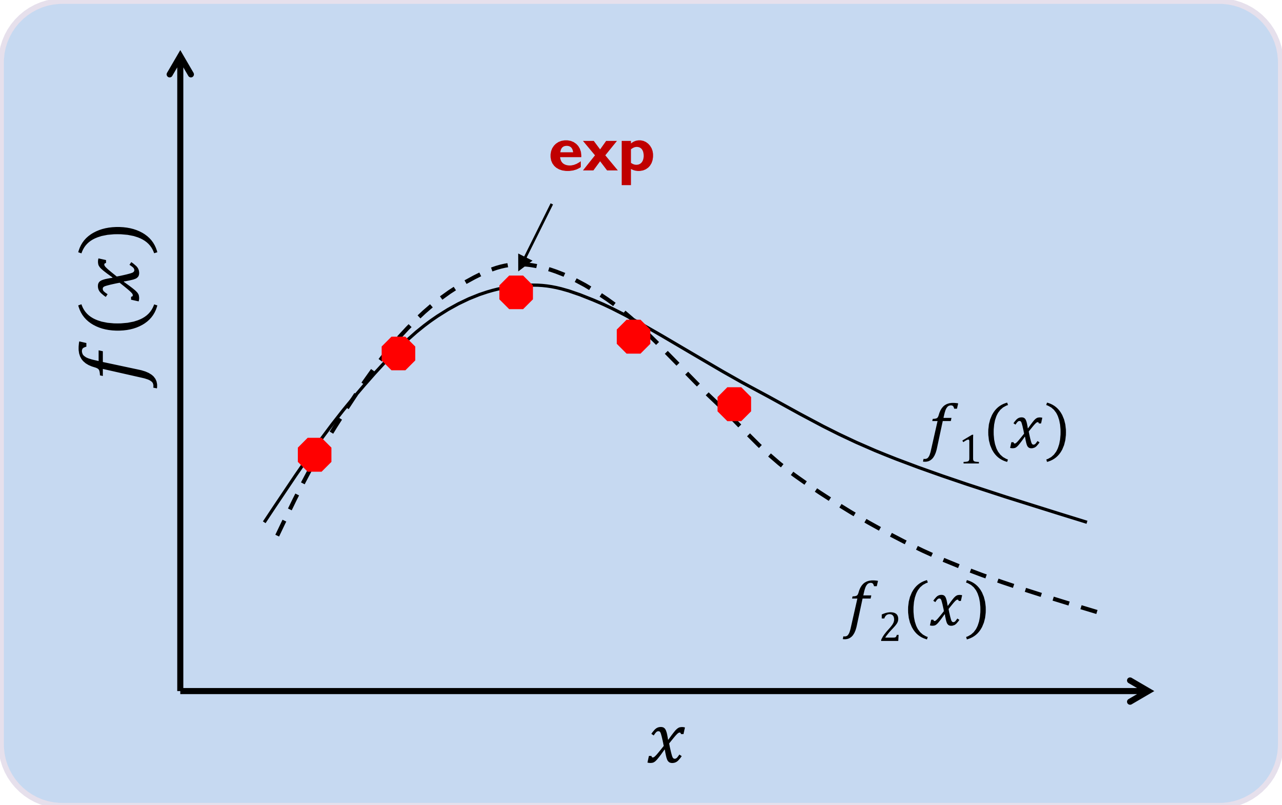
Lecture 11: Principle component analysis for classifying $\{y\}$.

Lecture 12: Conclusions

# Outline

1. Introduction
2. Goodness of Fit: Adjusted R-square, AIC methods, etc.
3. Cross-validation: Another way to compare models
4. Bootstrap method to generation population properties based on sample characteristics
5. Parametric vs. non-parametric distribution
6. Conclusions

# Recall: MLE can be used to fit any model to the data



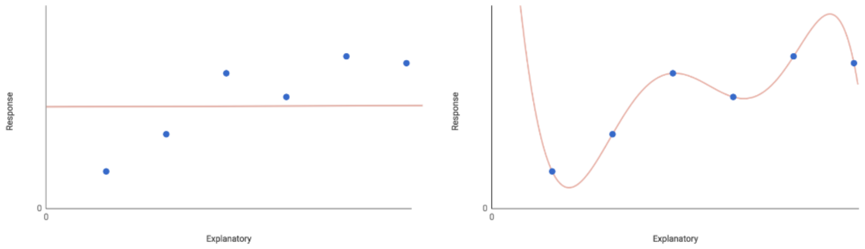

Each model can be checked for $\chi^2$, $KS$, or $QQ$ tests.
What if two or more models passes the test. Which one is better?

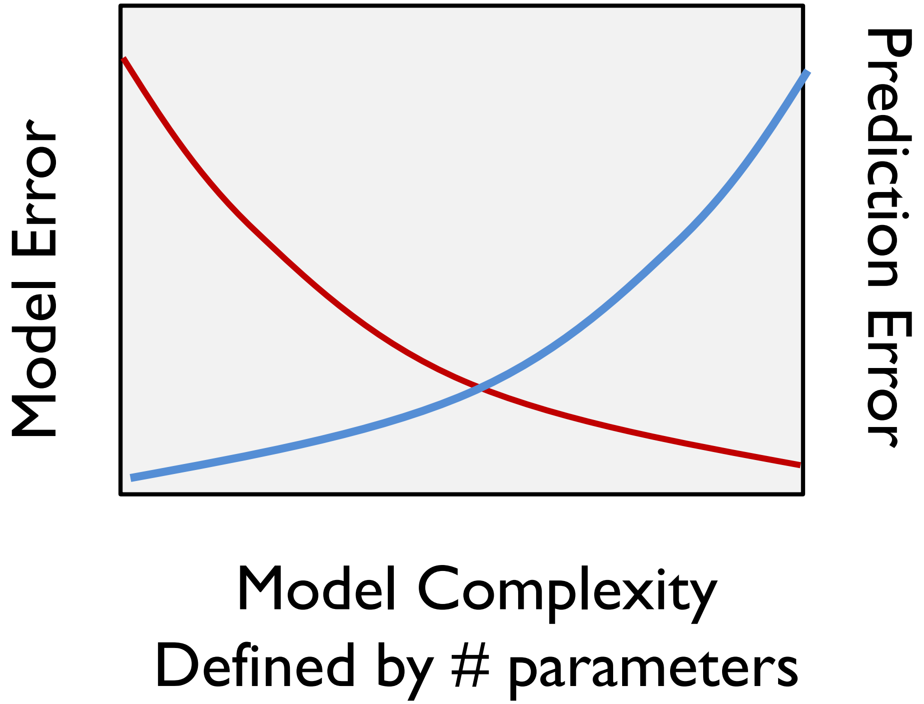
# Principle of Parsimony

Aristotle: Nature operates in the shortest way possible.

George Box: All models are wrong, but some are useful.

Occam's Razor: "given two or more equally acceptable explanations for a phenomenon, work with the one which introduces the fewest assumptions."





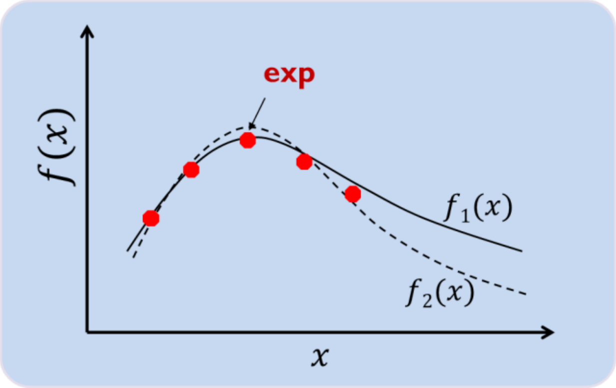
# Parameter number vs. goodness of fit

n = number of samples, M=number of parameters

$$R \equiv \sum_{n=1}^{n} \left(t_i - t_{i,fit}\right)^2$$

1) Method of adjusted residual …

$$R_{adj}^2 = \frac{(n-1)R^2 - (M-1)}{n-M}$$




2) Akaike Information Criterion

$$M \to p + 1$$

$$AIC = n \times \ln\left(R^2 / n\right) + 2M$$

Error penalty

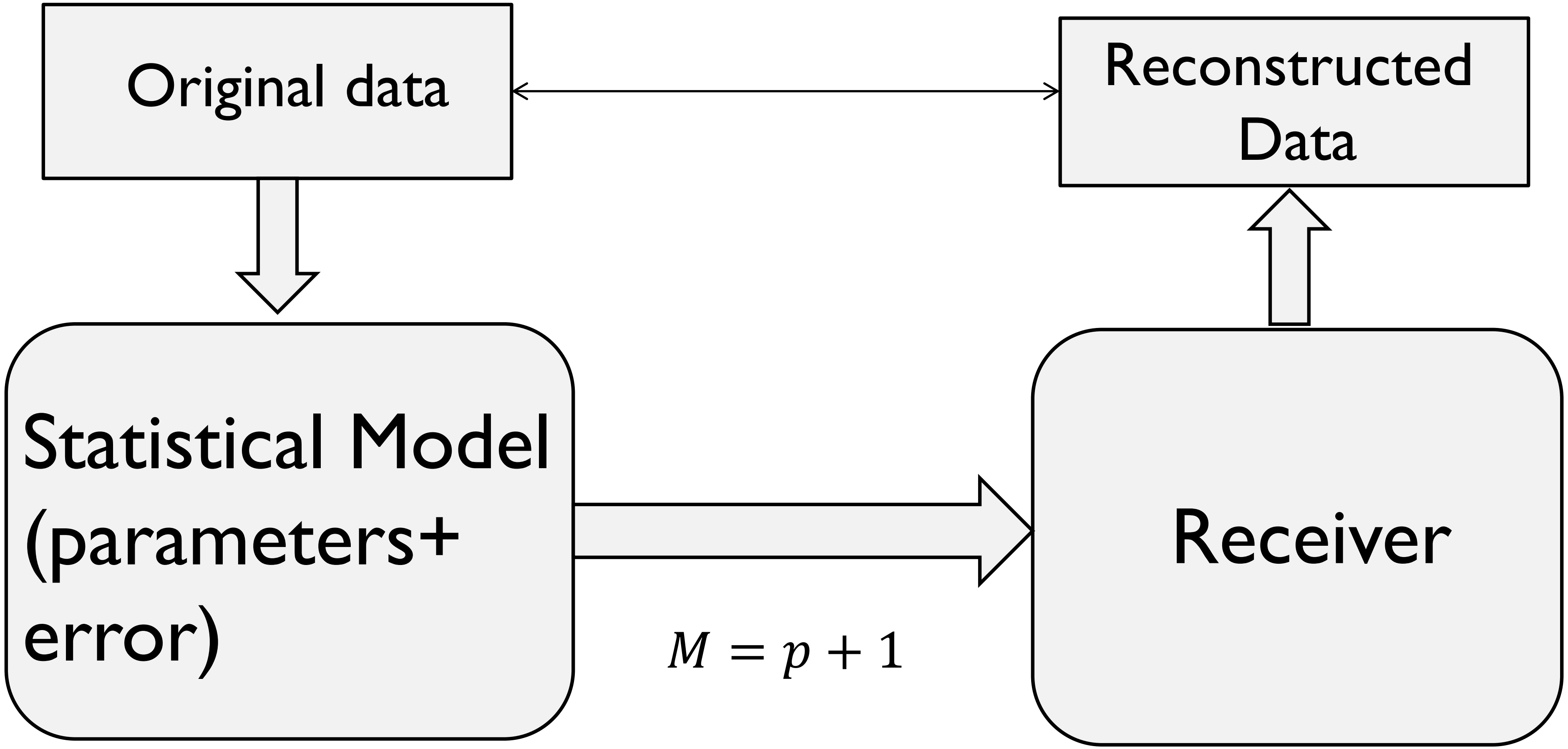
Parameter Penalty

2) Schwarz Information Criterion

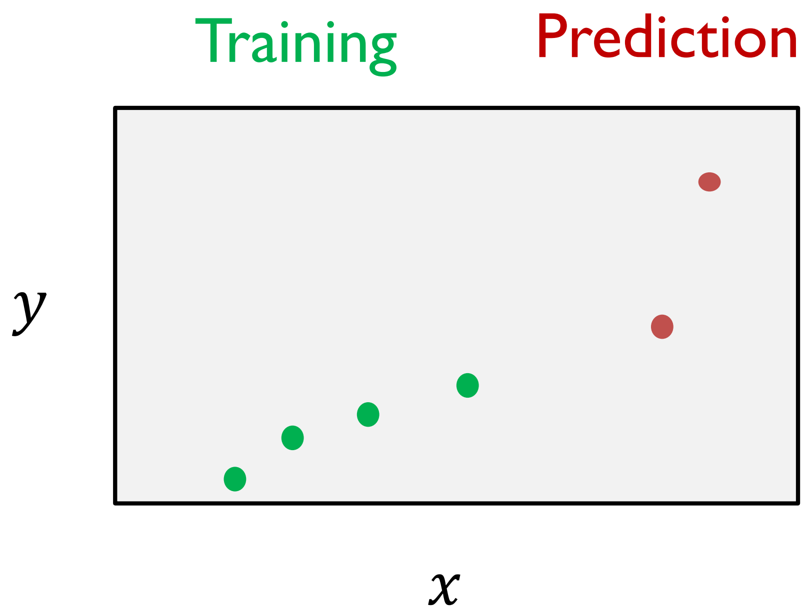
$$BIC = n \times \ln\left(R^2 / n\right) + M \times \ln n$$

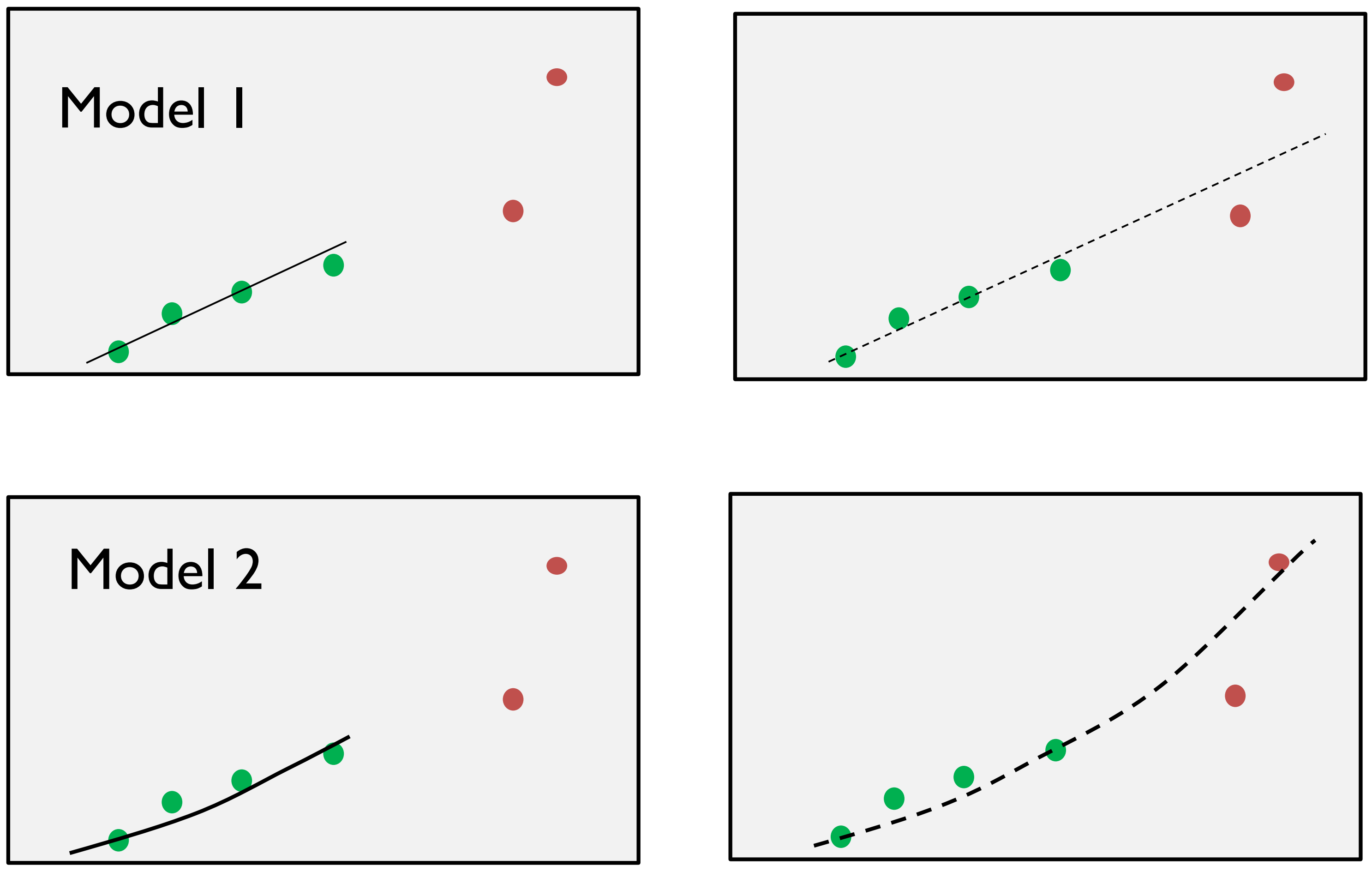
*Ref. Les Kirkup, Data Analysis with Excel, Cambridge Univ. Press. P. 304*

# Essence of the information theoretic approach



# Cross validation method




Vopnik-Chervonenkis (VC) dimension

# Outline

1. Introduction
2. Goodness of Fit: Adjusted R-square, AIC methods, etc.
3. Cross-validation: Another way to compare models
4. Bootstrap method to generation population properties based on sample characteristics
5. Parametric vs. non-parametric distribution
6. Conclusions

# Statistics of Sample vs. Population

Distribution-free statistical measure of data ….

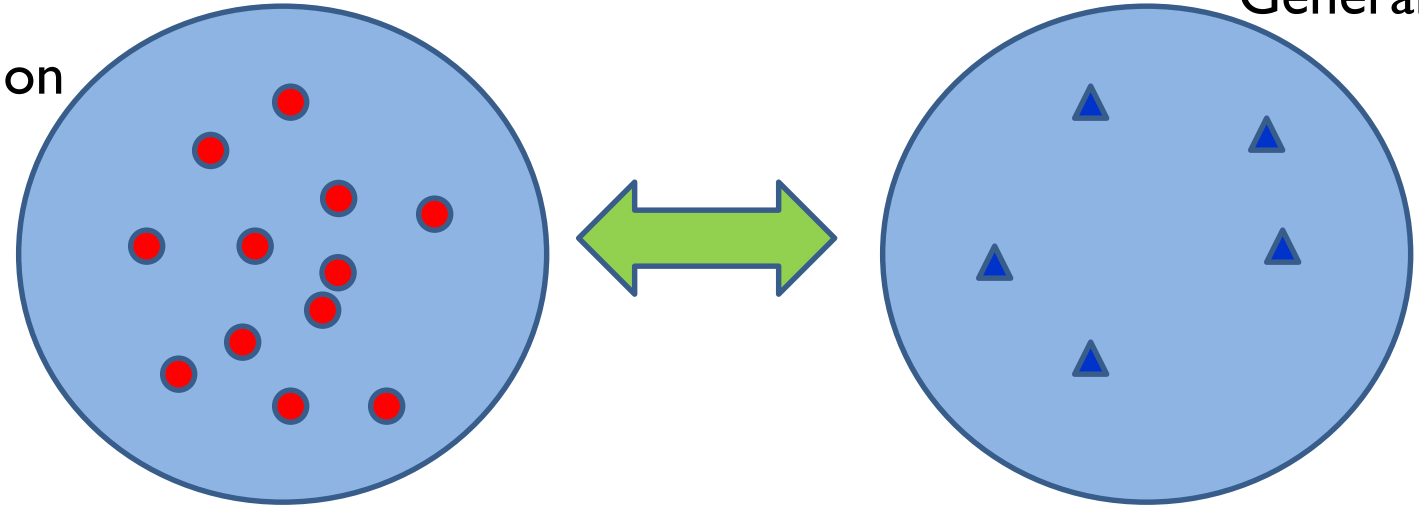
Parameter-space

$$\langle \boldsymbol{t} \rangle = \frac{\sum_{j=1,N} \boldsymbol{t}_j}{N} \qquad \boldsymbol{s}^2 = \frac{\sum_{j=1,N} (\boldsymbol{t}_j - \langle \boldsymbol{t} \rangle)^2}{\boxed{N-1}}$$

$$\delta_{T_K} = \sqrt[k]{\frac{\sum_{j=1}^{N} (t_i - \langle t \rangle)^k}{\boxed{N-k+1}}}$$

General formula

Population



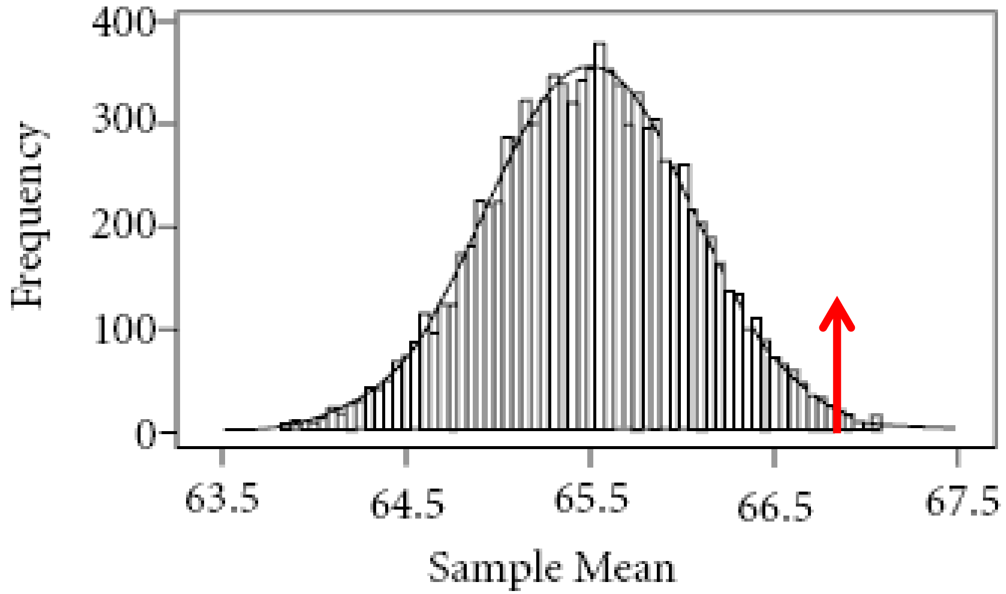
Parameter

Similar to Fourier Series, First used by Brahe for Alpha Aretis

Good for comparison, but not appropriate for projection

# Distribution of the Sample Statistic/Moment (e.g. Mean)

Sample Size =20
Number of samples=10k (from population)




Meaning of p-value

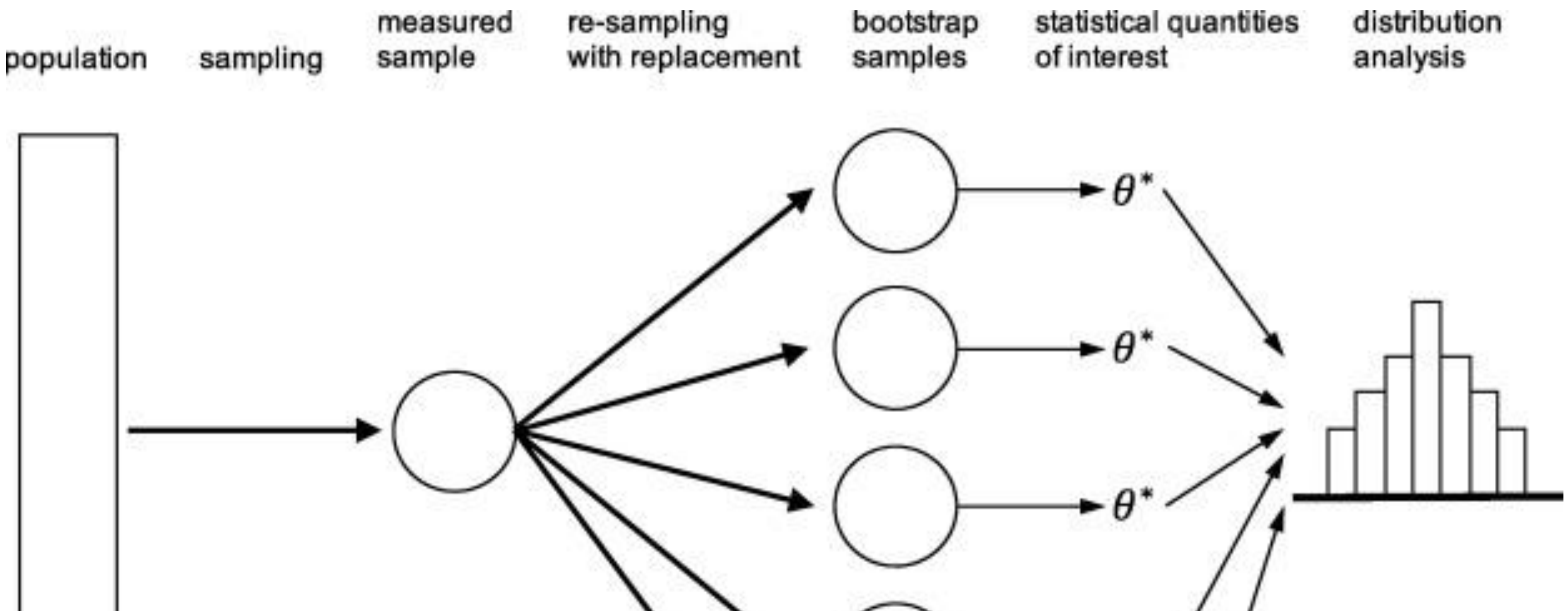
$\mu_x = \mu$

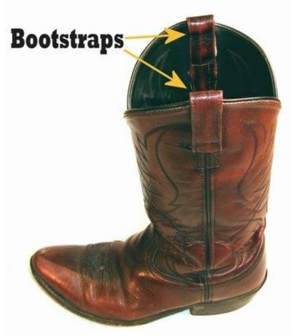
$\sigma_x = \sigma / \sqrt{N}$

Z=(X−$\mu$)/($\sigma / \sqrt{N}$) N > 30

Z=(X−$\mu$)/($s / \sqrt{N}$) N < 30

# Overall algorithm for bootstrapping



# Working with a single sample

0.2 -0.1 0.5 0.3 -0.6 All you have is a single sample ..

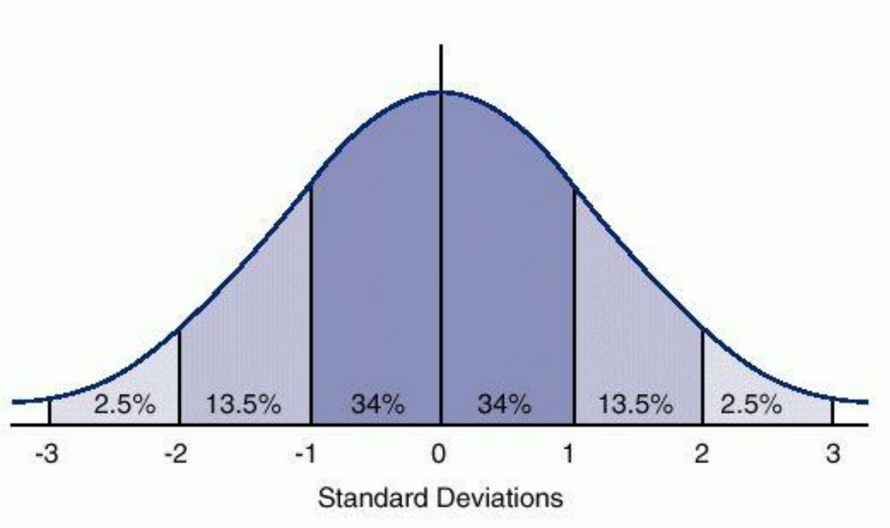
Generate synthetic samples from the original (with replacement)

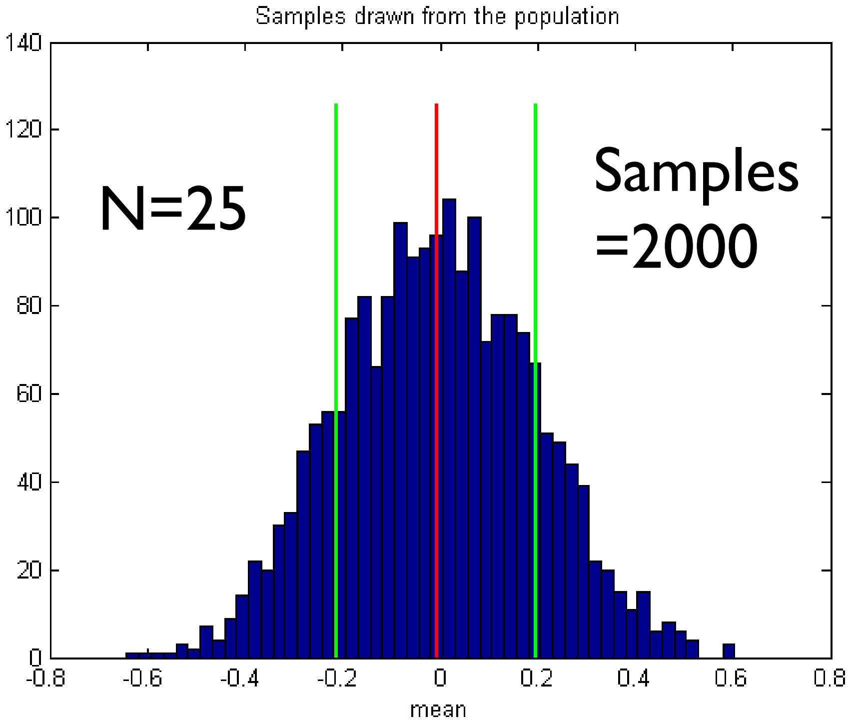
0.2 -0.1 -0.6 -0.1 0.5 Synthetic sample 1

0.3 0.2 -0.6 0.2 0.5 Synthetic sample 2

0.5 -0.1 0.5 0.2 0.3 Synthetic sample 3

# Bootstrap method - Introduction



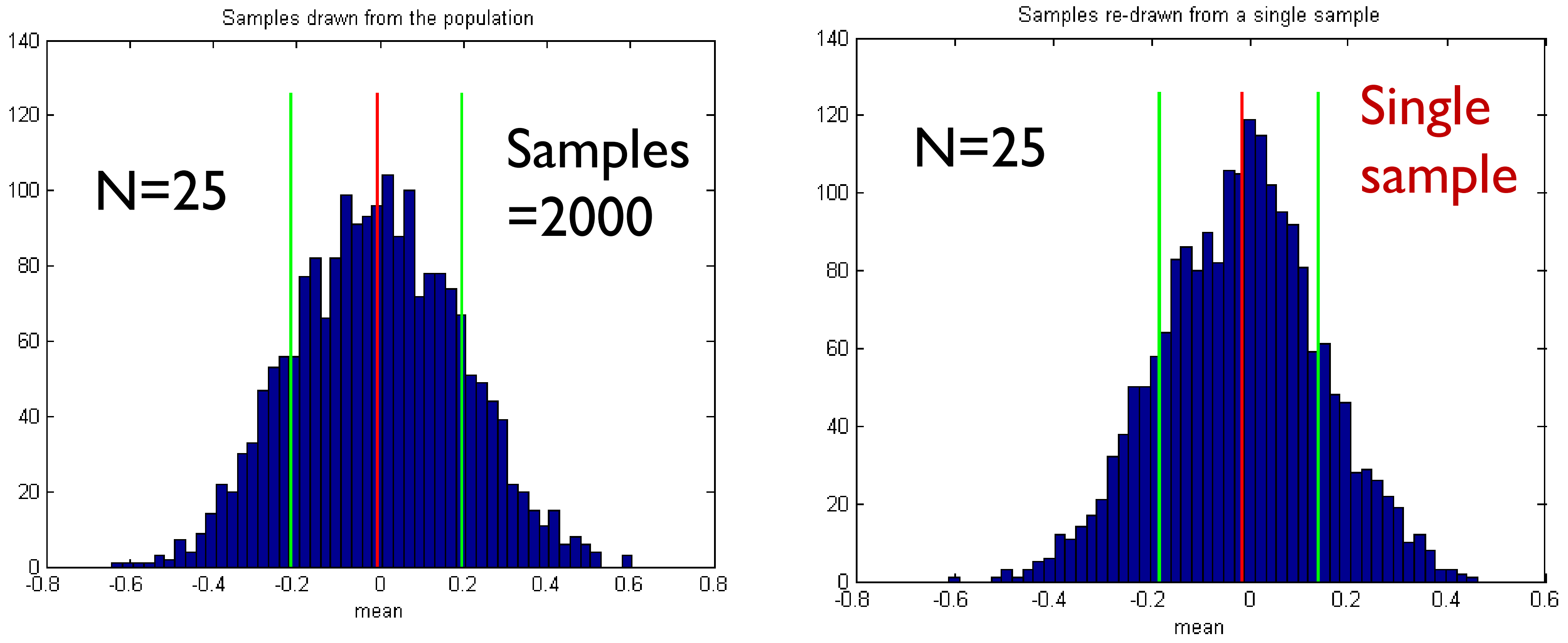




68% between +0.2 to -0.2
95% between +0.4 to -0.4

$$s = \sqrt{\frac{\sum_{j=1,N=25} (t_j - \langle t \rangle)^2}{N-1}} \sim \sqrt{\frac{1}{24}} \sim 0.2$$

*Ref. courses.washington.edu/matlab2/Lesson_6.html*

# Multiple sample vs. single sample




Bootstrap average is not zero!
And yet, the s~0.18, just from a single sample.
The success of the method relies on precision measurement

# Parametric vs. non-parametric Bootstrap

0.2 -0.1 0.5 0.3 -0.6

Fit the distribution of your choice by Maximum likelihood estimators (MLE) (obtain parameters, i.e. $\eta_0, \beta_0$)

Generate synthetic samples based on the parametric distribution

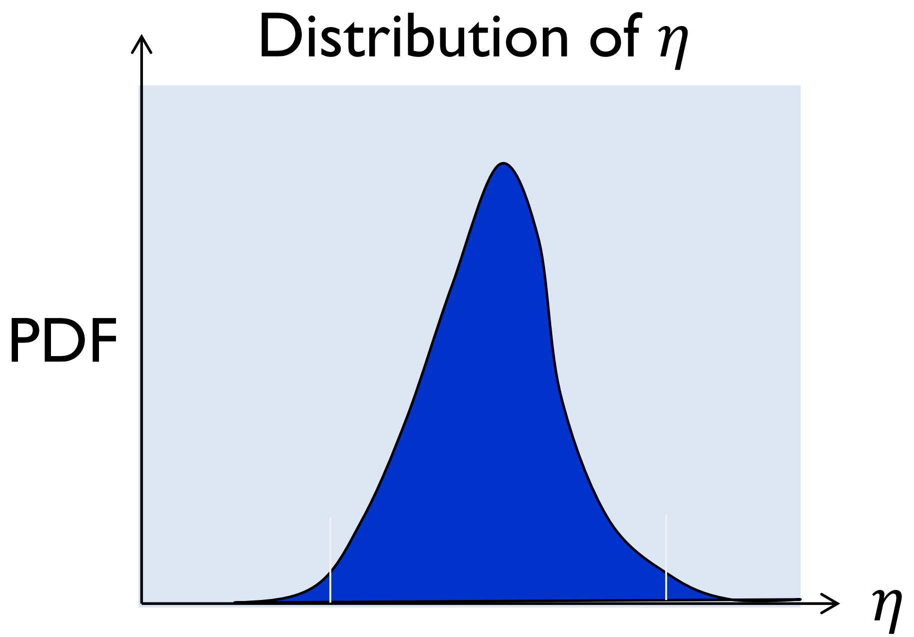
0.12 -0.17 -0.44 -0.71 0.52 Synthetic sample 1 (new $\eta_1, \beta_1$)

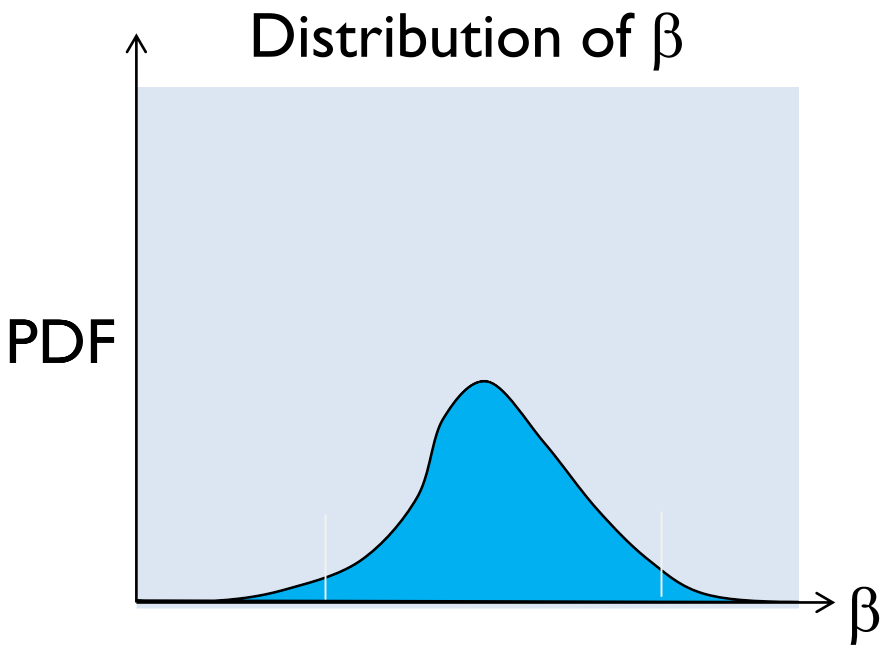
0.32 0.21 -0.69 0.23 0.58 Synthetic sample 2 (new $\eta_2, \beta_2$)

↓ ↓

Plot distribution of statics $\eta_i, \beta_i$

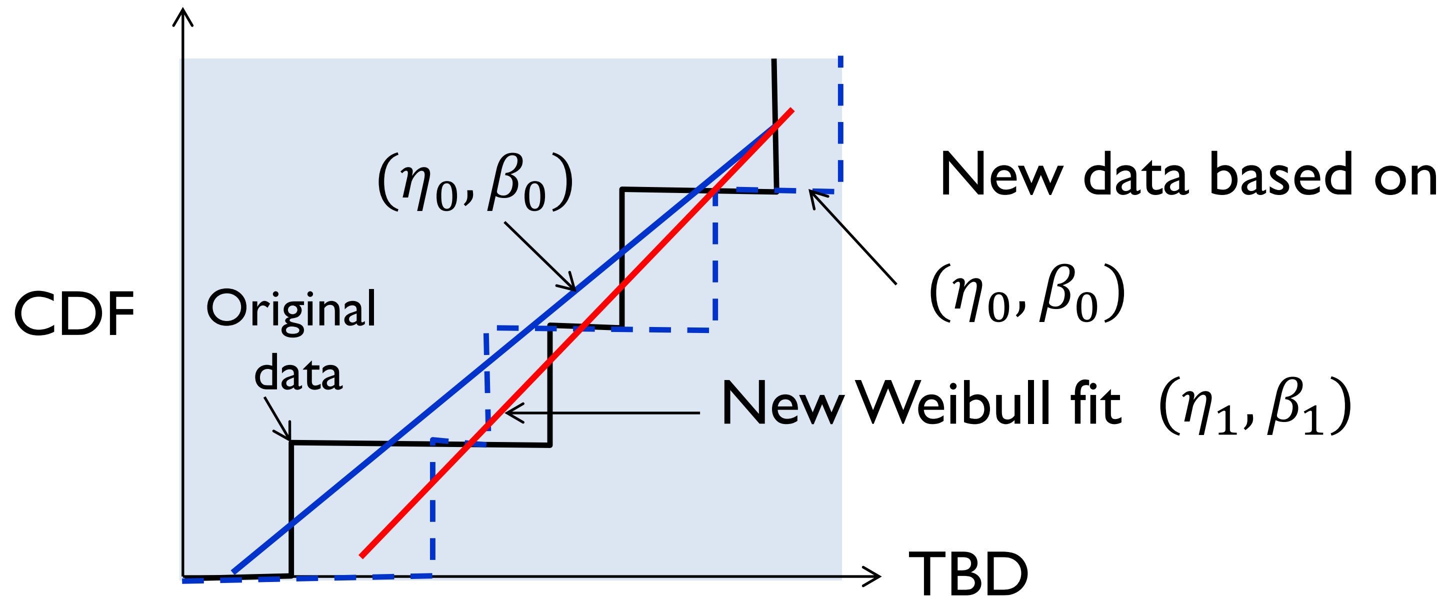
# Distributions of $\alpha$ and $\beta$




Same technique for polling and tenure rate of faculty!

# Why resampling from the same distribution generates new fit parameters




Samples taken from the same distribution $(\eta_0, \beta_0)$ generates datapoints that are fitted with new $(\eta_i, \beta_i)$

# Conclusions

1. If you have physics on your side, the game is over. If physics is unknown, then computer-based model comparison is helpful.

2. The approach selects the best among two or more models based on the principle of parsimony. It can be formulated in terms of information theoretic approach or cross validation approach.

3. The second approach relies on the Bootstrap or Monte Carlo approach. Here one generates the population statistics based on single sample. And then uses parametric or non-parametric approaches to define the goodness of fit.

# References


Goutte, Cyril. "Note on free lunches and cross-validation." *Neural Computation* 9.6 (1997): 1245-1249.

Wolpert, David H., and William G. Macready. *No free lunch theorems for search*. Vol. 10. Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

Efron, Bradley, and Gail Gong. "A leisurely look at the bootstrap, the jackknife, and cross-validation." *The American Statistician* 37.1 (1983): 36-48.

Efron, Bradley. "How biased is the apparent error rate of a prediction rule?." *Journal of the American statistical Association* 81.394 (1986): 461-470.

Diaconis, Persi, and Bradley Efron. "Computer-intensive methods in statistics." *Scientific American* 248.5 (1983): 116-131.

A simple and gentle tutorial on Vapnik-Chervonenkis (VC) dimension part 1 and part 2 of 2 on November 11, 2013 by panthimanshu17

Gershenfeld, Neil A., and Neil Gershenfeld. *The nature of mathematical modeling*. Cambridge university press, 1999.

Bollen, Kenneth A., and Robert A. Stine. "Bootstrapping goodness-of-fit measures in structural equation models." *Sociological Methods & Research* 21.2 (1992): 205-229.

Gershenfeld, Neil A. "Dimension measurement on high-dimensional systems." *Physica D: Nonlinear Phenomena* 55.1-2 (1992): 135-154.

# Review Questions

- What is the principle of parsimony?
- What does "B" stand for in the BIC metric?
- What theory underlies the AIC and BIC metrics?
- What is wrong with standard curve fitting? How does cross-validation address the problem?
- In what ways are cross-validation and bootstrap methods are similar?
- In what ways are the two methods different?
- What is the difference between parametric and non-parameteric bootstrapping method?