ISyE 6404 – Nonparametric Statistics

Lecture #1 (08/21/18) – Introduction/Motivation and Background Review

1. Introduction – Syllabus-Items Explanations

1. Instructor: Dr. Jye-Chyi (JC) Lu

2. Class time and Location: 9:30 – 10:45 am, TR, ISyE Annex #228

3. Reference/Textbook: P. H. Kvam and B. Vidakovic, *Nonparametric Statistics with Applications to Science and Engineering* (ISBN: 978-0-470-08147-1)

4. Office Hours and Location: 13:25 – 14:50 pm TR at Groseclose #312

5. Email: jclu@isye.gatech.edu (I will be the main contact for this course)

6. Class web page: We will use GT's Canvas (ISyE 6404) for class materials

7. Teaching Assistant: TBA

8. Course prerequisites: Master-Level statistics and probability.

9. Course Topics: Rank-based Methods, ANOVA, Nonparametric Inference, Smoothing, and Nonparametric (and Semi-Parametric) Regression

10. Software Application Projects: Students will use software such as Matlab, R, Excel, Minitab, SAS, SAS -JMP, and other publically available statistical packages to conduct a few hands-on data analysis projects.

11. Detailed Topics: This is a Master-Level *methodology and application* course. The course will cover most nonparametric statistics popular in applications such as bio-medical studies and reliability engineering.

Part I - Review: Probability and Distribution Theory, Common Distributions, Large-Sample Theory;

Part II - Rank-based Methods: Ordered Statistics, Goodness-of-Fit, Rank Test;

Part III – ANOVA, The CDF, Bootstrap and Jackknife;

Part IV – Empirical Likelihood (NPMLE), Density Estimation (DE), Proportional Hazards Regression;

Part V – Nonparametric Regression: Kernel Regression, Local Polynomials, Penalized Regression, Regularization and Splines, Smoothing Using Orthogonal Functions, Wavelet-based Smoothing.

11. Grade Distribution:

a) Part I & II - Review, Rank-based Methods **(Exam #1 – 25%)**

b) Part III & IV – ANOVA, CDF, NPMLE, DE, PH-Regression **(Exam #2 – 25%)**

c) Part V - Nonparametric Regression and Applications **(Take-home Exam #3 – 25%).**

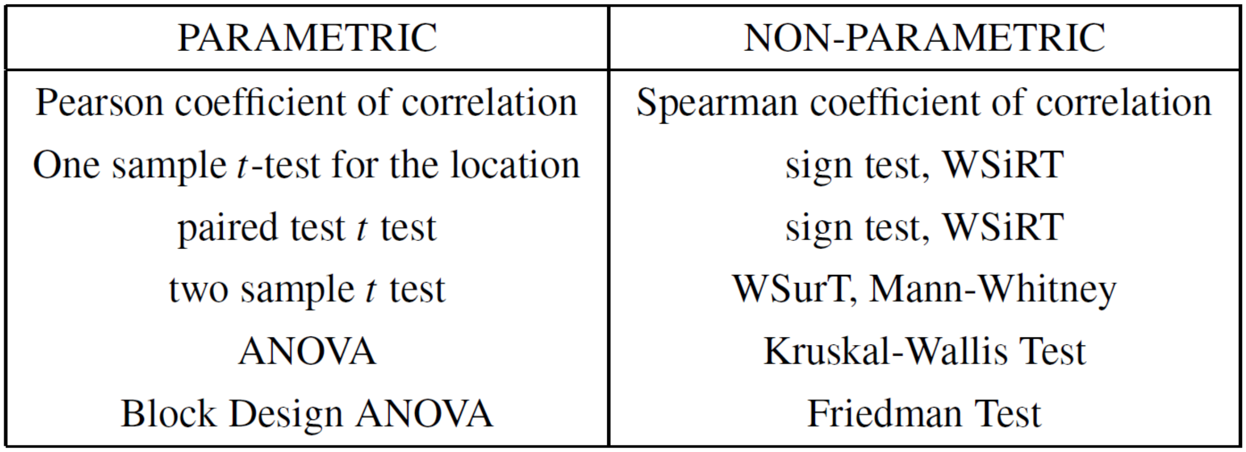
d) Enrichment Project **(*two* Team Projects – 15% (7.5% each))**

e) Computing Project **(individual project – 7%)**

f) Attendance and Survey: 3% (6 attendance will be checked *randomly*; students are allowed up to 2 missing attendance). 1% will be allocated to instructional survey.

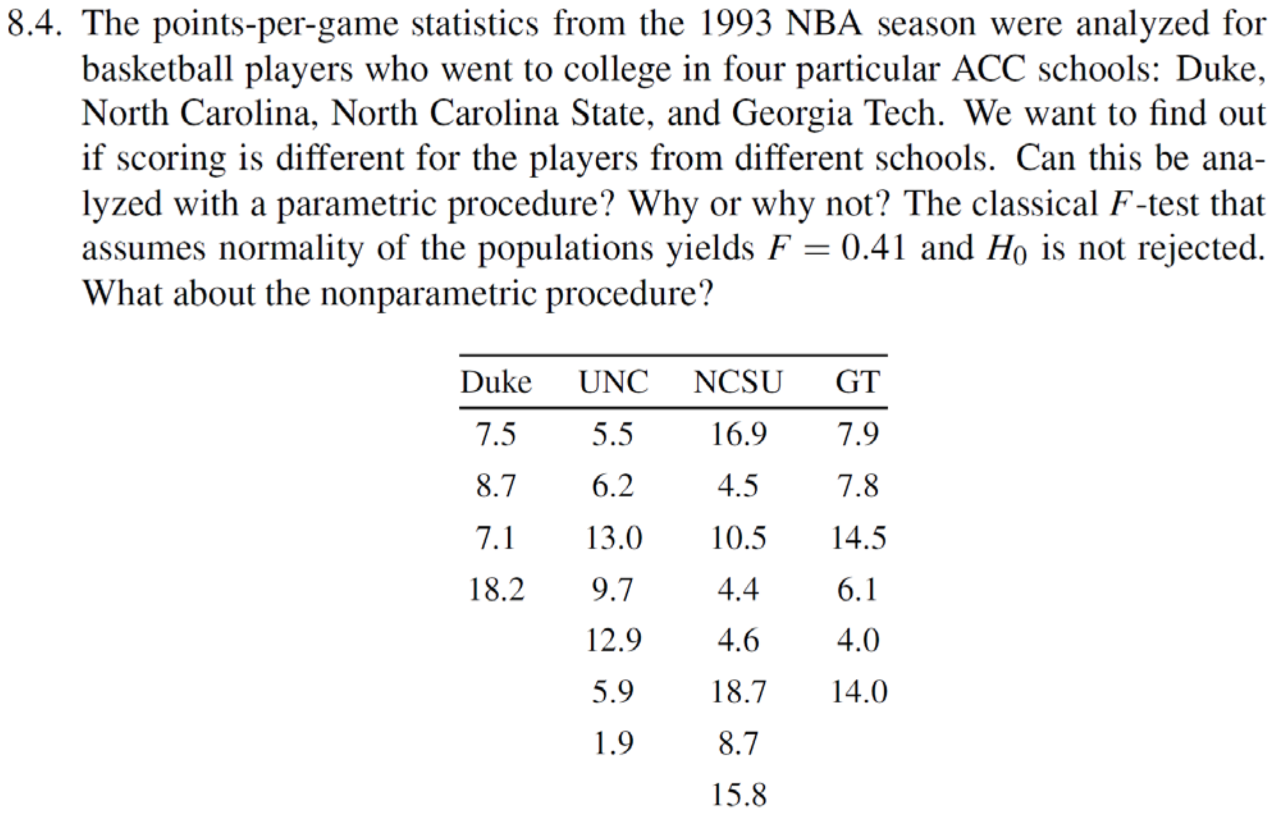
2. Motivation - Examples

a) Rank-based Methods:



PARAMETRIC	NON-PARAMETRIC
Pearson coefficient of correlation	Spearman coefficient of correlation
One sample t-test for the location	sign test, WSiRT
paired test t test	sign test, WSiRT
two sample t test	WSurT, Mann-Whitney
ANOVA	Kruskal-Wallis Test
Block Design ANOVA	Friedman Test

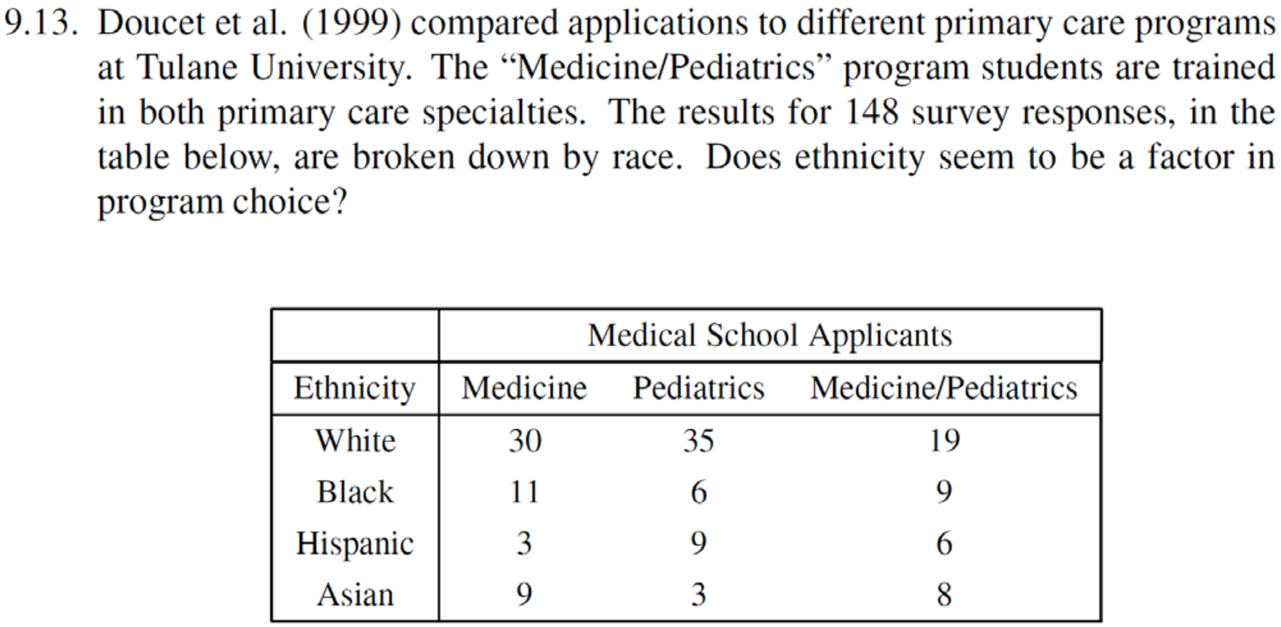
i) One-Way ANOVA Example (Textbook page 152):



8.4. The points-per-game statistics from the 1993 NBA season were analyzed for basketball players who went to college in four particular ACC schools: Duke, North Carolina, North Carolina State, and Georgia Tech. We want to find out if scoring is different for the players from different schools. Can this be analyzed with a parametric procedure? Why or why not? The classical F-test that assumes normality of the populations yields $F = 0.41$ and H_0 is not rejected. What about the nonparametric procedure?

Duke	UNC	NCSU	GT
7.5	5.5	16.9	7.9
8.7	6.2	4.5	7.8
7.1	13.0	10.5	14.5
18.2	9.7	4.4	6.1
	12.9	4.6	4.0
	5.9	18.7	14.0
	1.9	8.7	
		15.8	

ii) Contingency Tables (Textbook page 178):



9.13. Doucet et al. (1999) compared applications to different primary care programs at Tulane University. The “Medicine/Pediatrics” program students are trained in both primary care specialties. The results for 148 survey responses, in the table below, are broken down by race. Does ethnicity seem to be a factor in program choice?

	Medical School Applicants		
Ethnicity	Medicine	Pediatrics	Medicine/Pediatrics
White	30	35	19
Black	11	6	9
Hispanic	3	9	6
Asian	9	3	8

b) Smoothing and Regression:

i) Curve-Fitting (Textbook page 247):

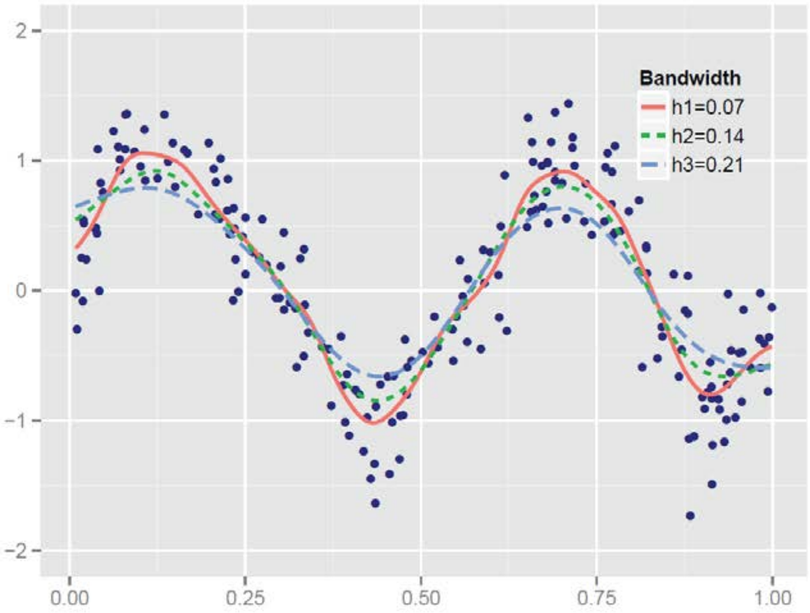


Figure 13.3 Nadaraya-Watson Estimators for different values of bandwidth.

ii) Nonparametric Regression (J.-L. Wang (2003) in Nonparametric Regression Analysis for Longitudinal Data):

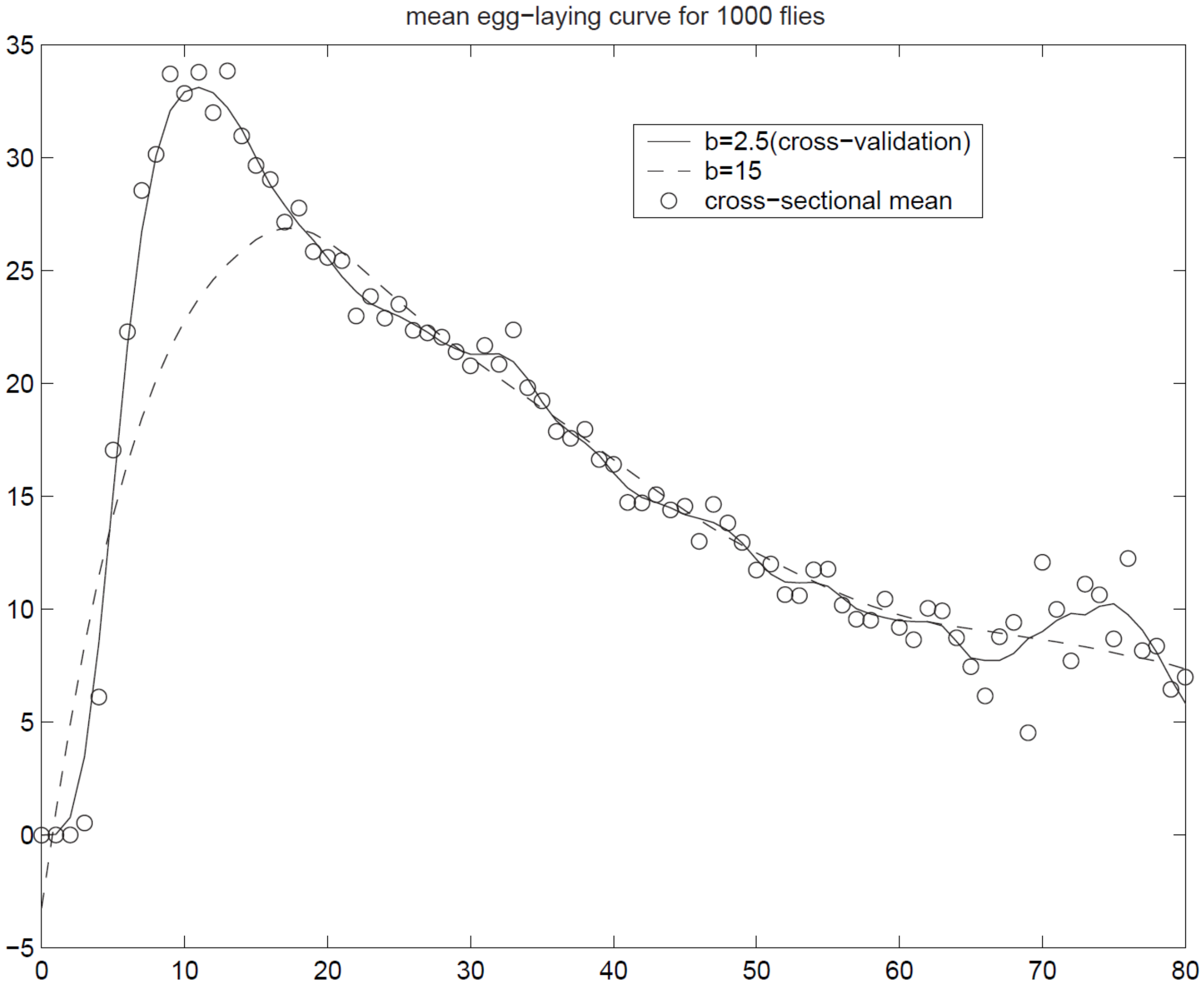


Figure 1: Mean egg-laying curves of 1,000 female Mediterranean fruit flies. (a) daily cross-sectional means of flies alive at the beginning of the day (circles) (b) smoothed mean curve with fixed bandwidth $b = 15$ (dash line) (c) smoothed mean curve with cross-validated bandwidth choice $b = 2.5$ (solid line).

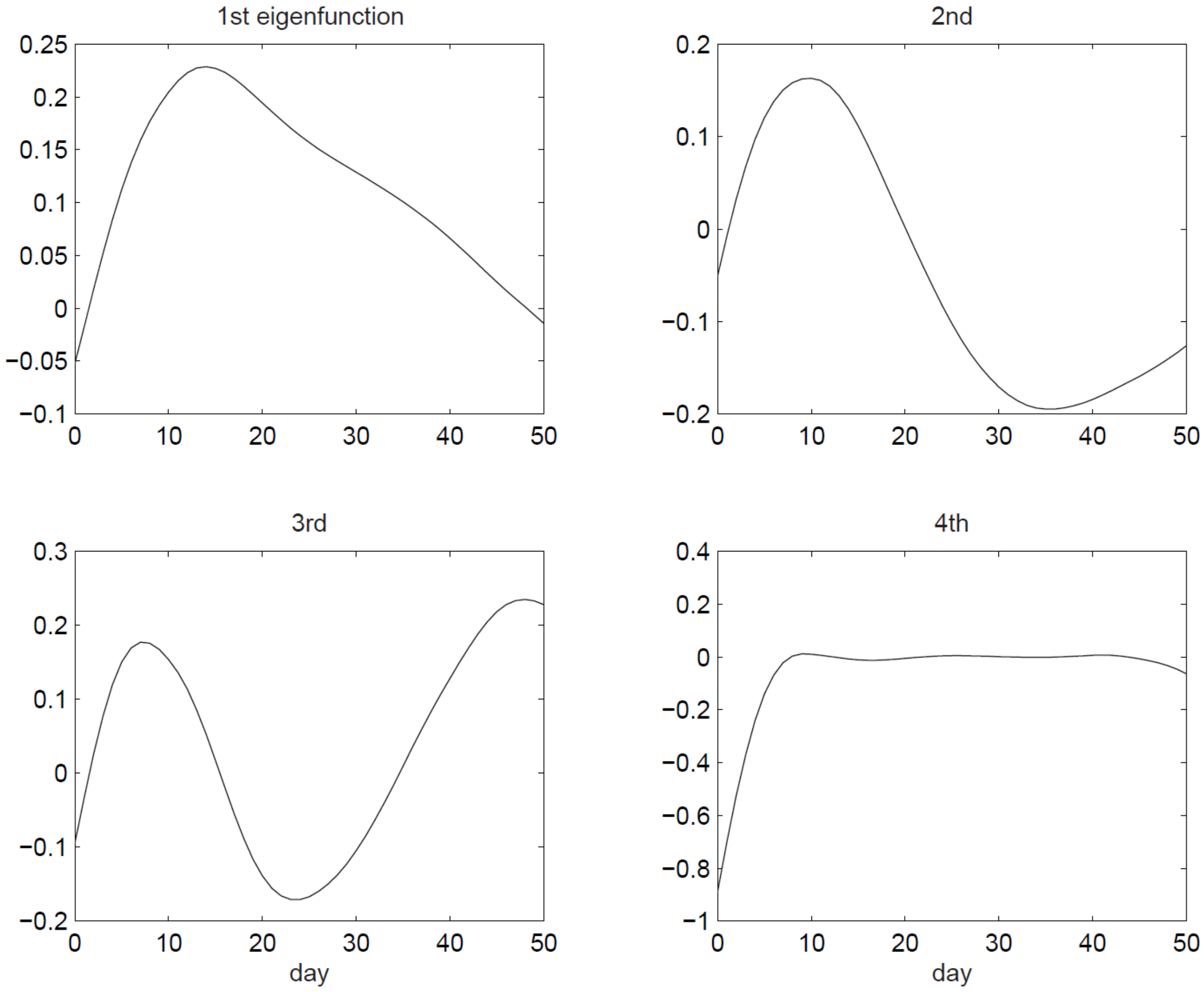


Figure 2: First four eigenfunctions for egg-laying data. The fraction of variation explained by each of these components are: 0.6183, 0.2090, 0.0779, 0.0536 respectively.

3. Background Review

a) **Probability:** Set Theory, Events, Probability and Random-Variables, Parametric Distributions (e.g., normal, exponential, Gamma (Chi-Square), F, Weibull; Binomial, Poisson, Negative Binomial)

b) **Statistics:** Summary statistics (e.g., mean, variance, skewness, correlation), parameter estimation, hypothesis-testing, confidence-interval, ANOVA, regression.

Note #1: The course will teach at the level that most students feel comfortable to understand the lecture materials. Since this is a Master-Level course, the course will emphasize on **methodology and application**, especially in using software to analyze data and write analysis reports.

Note #2: Students who are graduating with Master degrees should explore application examples, search for data, and apply software to analyze the data to gain practical experience. Students who are interested in the Ph.D. level contents will be encouraged in digging into more in-depth research studies in the ENRICHMENT PROJECTS. **For both groups of students writing reports in high quality is the key to do well in your job. This course will design enrichment computing/data-analysis projects for students to practice skills learned in the class.** Most projects will provide opportunities for students to select their own focused topics/application-areas.