

MATH 5060: Exploratory Data Analysis (EDA)
Fall 2006 (Kafadar)

Course Motivation:

This course will introduce students to philosophy and methods comprising exploratory data analysis (“data mining”). We will relate these methods to analogous concepts in theoretical statistics, emphasizing “exploratory” versus “confirmatory”, and students will learn and apply analysis tools on real data and actual case studies. At the end of the course, you should be able to recognize appropriate tools to use on various sets of data, construct meaningful graphical displays, propose and fit suitable “models” for data, prepare reports, use R (statistics software package), and prepare oral and written reports based on the data analysis.

Course Objectives:

- Understand philosophies of “exploratory” versus “confirmatory” data analysis
- Learn the tools of EDA and how to apply them to real data
- Develop expertise in analyzing data: how to answer questions posed by the data; how to identify further questions raised by the data
- Learn to use a statistical software package (e.g., R)
- Improve problem-solving and presentation skills
- Collaborate on analyses in teams
- Know basic commands and the grammar of R

The primary objective of this course is to impart the philosophy of exploratory data analysis, to show its connections with conventional statistical analysis, and to illustrate some of the most-useful EDA methods. A consequence of this objective is the need to become familiar with a statistical software package that can facilitate the computations, especially with large data sets (such as data from microarray experiments and internet traffic data that we will be analyzing). Because “R” is free (www.cran.r-project.org), contains most of the tools that we need to achieve these objectives, and has gained rather wide acceptance, I will illustrate these methods in class using “R” commands — *but you will not be required to use R if you wish to use another software system*. So long as you can carry out the assignments, you are free to use whatever software you prefer. You may find, however, that many conventional statistical software systems do not include several methods commonly associated with EDA. The analyses in my examples will be conducted using R.

Course Information:

Time: Tuesday and Thursday, 4:00-5:15, CU-Denver 626

Prerequisites:

- Calculus
- Linear algebra
- Previous courses in statistical methods

Textbooks (first is required; others are recommended):

- Hoaglin, D.C.; Mosteller, F.; Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis (UREDA)*, Wiley.
- Hoaglin, D.C.; Mosteller, F.; Tukey, J.W. (1985), *Exploring Data Tables, Trends, Shapes (EDTTS)*, Wiley.
- Hoaglin, D.C.; Mosteller, F.; Tukey, J.W. (1991), *Fundamentals of Exploratory Analysis of Variance (FEAV)*, Wiley.
- Tukey, John W. (1977), *Exploratory Data Analysis (EDA)*, Addison-Wesley.

Handouts will come from *EDA*, *EDTTS*, and *FEAV*, when needed, for those students who do not have access to them. All four books can be found in the Auraria Library.

Course Expectations:

Most material will be presented via lectures (be prepared for sudden questions!). Three important components in a statistician's work are collaboration, problem solving, and communication; this course will provide you opportunities in all three areas.

1. *Homework assignments* (30%): Problem sets will be assigned weekly and due on Tuesdays. Collaboration is encouraged. *To ensure grading consistency, late homeworks cannot be accepted.* Please do **not** ask for exceptions to this policy (or expect such requests to be categorically denied).
2. *Midterm exam* (25%): The midterm will take place on Thursday, October 20. take-home.
3. *Final exam* (30%): date to be announced.
4. *Final project* (15%) The final project will be a **10-minute** presentation to the class based on a report of an analysis of data from your own choosing (subject to professor's approval). *You should solicit approval for your selected data set no later than October 13.* The accompanying report should describe the data, the questions that were posed, the analysis designed to answer them, and a summary of the findings (instructions will be given).
5. *Seminars*: Though not required, students should make every effort to attend statistics seminars on Friday afternoons.

Incompletes: Consistent with the general policy of the department on this issue, a grade of "incomplete" will **not** be an option, except under the most extreme (and completely unavoidable and unforeseen) circumstances. *I do not expect to assign any "incomplete" to any students.* Be prepared to have any such requests denied.

Course Instructor: Professor K. Kafadar, 617 CU-Denver Bldg, 303-556-2547, kk@math.cudenver.edu
Office Hours: Tuesday and Thursday 2-3:45 (or by appt)

Topics:

- Single batches: stem-and-leaf display, letter value display, boxplots, qq plots, rootgrams
- Re-expressions: diagnostics and fitting; outlier detection
- Multiple batches: comparative boxplots, draughtsman displays
- Robust smoothing; smoothing scatterplots/boxplots: residuals
- Two-way and three-way analyses: median polish, diagnostic plots, interaction terms
- Exploratory analysis of variance
- Count data; fitting counts of counts
- Regression: fitting; adjustment; carriers, proxies; diagnostic plots, residuals; jackknife
- Standardization (indirect and direct)
- Exploratory analysis of variance

Fall 2006 Registration and Academic Deadlines, and Important Policies:

- CLAS students must always have an accurate mailing and e-mail address: Go to www.cudenver.edu/registrar to update and/or change.
- Students must complete and submit a drop/add form to make any schedule changes. Students are not automatically dropped from a class if they stop attending or do not make tuition payments. The student is ultimately responsible for verifying their schedule prior to officially published drop dates.
- Late adds will be approved only when circumstances surrounding the late add are beyond the student's control and can be documented. This will require a petition and documentation from the student.
- Late drops will be approved only when circumstances surrounding the late drop are beyond the student's control and can be documented. This will require a petition and documentation from the student.
- Students who wish to graduate in December of 2006 **MUST** meet with their academic advisor to obtain a graduation application. The application must be completed and submitted by September 6, 2006.
- Students are responsible for completing financial arrangements with financial aid, family, scholarships, etc. to pay their tuition. Students will be responsible for all tuition and fees for courses they do not officially drop using proper drop/add procedures and forms.

- August 24, 2006 (midnight) Last day to be added to the wait-list for a closed course.
- August 24 - September 6, 2006 Students are responsible for verifying an accurate fall 2006 registration via SMART. Students are NOT notified of their wait-list status by the University. All students must check their schedules prior to September 6, 2006 for accuracy.
- August 31, 2006 (midnight) Last day to add courses via the web SMART system.
- September 6, 2006 (5:00 pm) Last day to add structured courses without a written petition for a late add. This is an absolute deadline. This deadline does not apply to independent study, internships, and late-starting modular courses.
- September 6, 2006 (5:00 pm) Last day for undergraduates and graduates to apply for December 2006 graduation. This is an absolute deadline.
- September 6, 2006 (5:00 pm) Last day to request pass/fail or no credit option. This is an absolute deadline.
- October 30, 2006 (5:00 pm) Last day for NON-CLAS students to drop a fall 2006 course without a petition to their home college and receiving their Dean's approval.
- November 10, 2006 (5:00 pm) Last day for CLAS students to drop a fall 2006 course. Treated as an absolute deadline. Dean's approval required.
- November 10, 2006 (5:00 pm) Last day to withdraw (drop all courses) without a written petition.

See Academic Calendar for details on registration and payment deadlines:
www.cudenver.edu/registrar

Tentative schedule (subject to change, as needed) follows.

Lecture	Date	Topic	Reading (UREDA)
1	08/22	Philosophy: Confirmatory vs Exploratory	Intro
2	08/24	Stem-and-Leaf; 5-number summary, boxplot	Ch.1,2A,3A
3	08/29	Comparing batches: Letter values, boxplots	Ch.2A-C,3B
4	08/31	QQ-plots, spread-vs-level plots	Ch.3,C-E
5	09/05	Re-expressions: Ladder of transformations	Ch.4
6	09/07	Plots for diagnosing transformation	Ch.4
7	09/12	Mathematics of letter values (orstats)	Ch.2E-G
8	09/14	Mathematics of transformation	Ch.8A-E
9	09/19	Fitting lines: Robust-resistant line	Ch.5
10	09/21	Multiple carriers: sweeping out	EDTTS Ch.7
11	09/26	Diagnostics for regression	
12	09/28	Patterns in residuals	Ch.7
13	10/03	Median polish	Ch.6
14	10/05	Diagnostic plots; plotting fits	Ch.6H,8F
15	10/10	L1 and other two-way fitting	Ch.6D-G
16	10/12	Fitting 3-way,higher-order tables	EDTTS Ch.4
17	10/17	Median-based smoothers	EDA Ch 7
18	10/19	Linear vs nonlinear smoothing	handout
19	10/24	Characterizing uncertainty: Jackknife	handout
20	10/26	Midterm	Ch.1-6
21	10/31	Fitting discrete distributions	EDTTS Ch.9
22	11/02	(Poissonness plot; examples)	EDTTS Ch.9
23	11/07	Families of non-Gaussian distributions	EDTTS Ch.11A-C
24	11/09	Fitting g-h distributions	EDTTS Ch.11D-G
25	11/13	Standardization: motivation	handout
26	11/14	Direct/indirect standardization	handout
		Fall break Nov 20-24 (Thanksgiving)	
29	11/28	Robust estimation	Ch.9AB,10
30	10/30	L-, M-, R-estimates of location	Ch.11
31	12/05	Presentations	
32	12/07	Presentations	