

Department of Accounting & Law State University of New York at Albany

Acc 522. Statistical Methods for Business Decisions
Fall, 2001
J Gangolly / S Dutta

"There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot it's a sign our knowledge of them is very small and confus'd; and when a Mathematical Reasoning can be had it's as great a folly to make use of any other, as to grope for a thing in the dark, when you have a Candle standing by you."

-- John Arbuthnot, *Of the Laws of Chance*, 1692.

Welcome

Welcome to the exciting world of exploratory data analysis, traditional as well as Bayesian statistics, and datamining. The emphasis in the course will be on the use of statistics and the powerful graphics provided by the object-oriented language S-Plus for the analysis and visualisation of data of special interest to auditors, information system auditors, computer security professionals, and other professionals involved in data warehousing and webmining.

The course will involve considerable amount of programming, and it is important that you get familiar working in unix as well as windows environments. In the class we will be working mostly in the unix/CDE environment, but you are responsible for familiarising yourself with the user interfaces provided by the windows version of S-Plus, which will **NOT** be covered in the class.

This course is co-ordinated with Acc 681 to the extent knowledge of unix operating system and unix shell scripting is covered there. It is therefore imperative that you take Acc 681 very seriously to benefit fully from this course. The course is very fast paced, and rather formal in terms of statistical as well as programming constructs used.. It is therefore important that you keep up with the class at all times and not be left behind. Should you need help, seek it immediately. Both of us instructors are here to help you learn.

Use the wonderful facilities in the Arthur Andersen Laboratory for Accounting Information Systems. Enjoy!

Administrivia

Semester: Fall, 2001

Time: T 4:15 — 7:05 PM

Room: BA 223 (PriceWaterhouseCoopers Classroom)

Instructors: Jagdish S. Gangolly & Saurav Dutta

Graduate assistants: TBD

Office: BA 365C & BA xxx

Phone: (518) 442-4949 / (518) 442-xxxx

Fax: (707) 897-0601 / (518) 442-3944

Office Hours: T 3 – 4:15 PM. or by appointment

Instructor Homepage: <http://www.albany.edu/acc/gangolly>

Course Homepage: <http://www.albany.edu/acc/inf703fall2001.doc>

Newsgroup: sunya.class.acc522

Class Conduct:

The course consists of lectures, solution of problems, discussion of homework and book assignments. You are expected to do the readings well ahead of the class. Class time is to be used for the clarification of any doubts that you may have. Do not expect to merely listen to the instructor and gain knowledge. Applied statistics is a practical field backed by robust theory. A good understanding of the theory and its use in practice is essential to excel in the field. This is a hands-on course, and you are required to demonstrate competence in the topics covered in order to receive an acceptable grade. *We shall be giving occasional homework assignments. We also shall be calling upon some of you to come to the board and discuss problems either in the textbooks, other sources, or homework assigned.*

Software:

We shall be using the S-Plus system running under solaris in our unix cluster, and expect you to use *cayley.bus.albany.edu* to write and test most programs. We shall also be providing hints on doing the same things under Microsoft windows 2000, but it is your own responsibility to learn the user interface in windows. When using windows version in the class, we shall mostly be using command line mode and not the windows user interface, except for some graphics. You may use the unix version of S-Plus on the designated PCs in the Arthur Andersen Lab via HummingBird Exceed.

Newsgroup/e-mail:

We shall be using the class newsgroup (sunya.class.acc683) extensively for making announcements regarding tests, homework, quizzes, added links to this course homepage, etc. In fact, the newsgroup will be the primary means of communication between us outside of the class. You should post to the newsgroup all your questions and doubts for clarification. You are strongly encouraged to answer queries posted by others, and such responses will count towards class participation points for grading. You should communicate with me via e-mail only for individual problems and questions.

The Arthur Andersen Laboratory for Accounting Information Systems Access:

As a graduate student in the Department, you have access to the Arthur Andersen Laboratory. You will need to get from Ms. Lisa Scholz the password to enter the lab. Contact her in BA 365 as soon as possible. Should you have special requirements for software (DBMS servers) or hardware (Windows 2000 Servers) for your projects, let me know, and arrangements will be made for your access.

You also will need logins to the University unix cluster and the Department's Windows 2000 server. You will need to apply on-line for an account on the unix cluster, and contact J Gangolly regarding login for the Windows 2000 server. You can not use any machine in the lab without these logins.

Course Objectives:

- Understanding of *exploratory data analysis*
- Understanding of the *language S-Plus*
- Understanding of *unix shell scripting for preprocessing of data*
- Understanding of basic *traditional statistics & testing of hypothesis*
- Understanding of *Bayesian analysis and networks*
- Understanding of the basics of *multivariate methods*

Catalog Description:

Extensive coverage of sampling techniques for decision making. Includes simple random sampling, stratified sampling, cluster sampling, treating unequal clusters, area sampling, imperfect frames, questionnaire design, and field operations.

Prerequisite: *Msi 220 or Mat 108 or equivalent.*

An Honest Description:

Gangolly:

Data acquisition and preprocessing for statistical analysis. Exploratory descriptive data analysis using the language S-Plus. Basic graphics commands in S-Plus including trellis graphics. Descriptive data exploration and statistical modeling. Data preprocessing for Datamining & Concept Description. Data Cleaning. Data Integration & Transformation; Data reduction & Data Compression; Concept hierarchy generation. Association Rules in Large Databases. Classification & Prediction. Multivariate Methods: Clustering & other multivariate statistical methods.

Dutta:

Fundamentals of Probability & Introduction to Bayesian Decision Theory: Probabilities: joint, conditional & marginal; Bayes' Theorem and Likelihood ratio. Nomenclature of decision trees (or Bayesian Networks or Influence Diagrams). Bayesian Decision Theory & the Audit Risk Model: the mechanics of Decision Trees (or an equivalent method). The construction of trees, method of Folding back, Conversion of given Probs. to usable Probability metrics. Audit Risk Model as an exercise in understanding decision trees. Traditional Distribution Theory & Testing of Hypothesis: Distributions, Variances & confidence Intervals and how they relate to audit decisions.

Textbooks and Readings:

- **Required:**
 - **The Basics of S and S-Plus (Statistics and Computing)**
Andreas Krause, Melvin Olson
2nd edition (May 15, 2000)
Springer Verlag
ISBN: 0387989617
 - **Data Mining: Concepts and Techniques**
Jiawei Han, Micheline Kamber
1st edition (August 2000)
Morgan Kaufmann Publishers
ISBN: 1558604898
- **Recommended:**
 - **S-Plus 4.5 Student Edition**
Duxbury Press

We shall also be placing materials on reserve in the library and/or provide links on this coursepage as the semester progresses

Requirements

The classes will consist of lectures, solution of problems, discussion of papers and programming exercises. I shall be dividing the class into groups of 3 each, balanced in terms of skills in accounting, programming, facility with computers, mathematical maturity, needs of the projects selected, and other such attributes. The groups will work through out the semester on three substantial homework projects, each group member taking turns to be the lead on the assignment

Grading

The final course grade is dependent on the following factors:

- 200 points: Test (In class open book/notes. Details will be announced in the class and updated here)
- 150 points: Group Homework Assignments
- 100 Points: Group Project & written report
- 0 - 50 points: Pop-quizzes, when given
- 25 points: Class participation
- 475 - 525 points: Total points (max)

The final course grade is strictly relative, based on the total points scored.

The grades, once assigned can not be changed except in case of errors in grading. **Under no circumstances is it possible to do extra credit work to improve the grade.**

About the Instructors:

Jagdish S. Gangolly is currently an *Associate Professor of Accounting and of Management Science & Information Systems* in the School of Business, and a Senior Program Faculty member of the Ph. D Program in Information Science. He holds a Bachelor's degree with a major in Mathematical Statistics, a master's degree with a major in Operations Research, and a Ph. D degree in Accounting. He is also a Certified Internal Auditor. He has previously taught at the University of Pittsburgh, University of Kansas, Claremont McKenna College & the Claremont Graduate School, and California State University at Fullerton. He has worked in senior executive positions in management services in the pulp & paper industry as well as in soft-drink franchising. His articles have appeared in *Journal of Accounting Research*, *Auditing: Journal of Practice & Theory*, *Journal of the Operational Research Society*, *Critical Perspectives on Accounting Expert Systems with Applications: An International Journal*, *Artificial Intelligence in Accounting & Auditing*, and *the New Review of Applied Expert Systems & Emerging Technologies*. In 1989, he was the guest editor of *Advances in Accounting* currently he serves on the editorial board of the American Accounting Association journal *Issues in Accounting Education*, the *International Journal of Digital Accounting Research*, and is an Associate editor of the *e-Services Journal*. He also serves on the *E-Commerce Curriculum Committee* of the **International Federation for Information Processing (IFIP)**. His current research activities are primarily in the areas of *conceptual information organisation, markup languages supporting electronic commerce, and the formal specification of control in accounting information systems*. He also has collateral research interest in the *relationships between Accounting and Legal Philosophy*.

Prof. Saurav Dutta is currently an *Associate Professor* in the Department of Accounting, Business Law and Taxation at the State University of New York at Albany. He holds a Bachelor of Technology Degree in Aerospace and Mechanical Engineering from the Indian Institute of Technology (Bombay, India) and a Ph. D. in Accounting from the University of Kansas. He is also a Certified Management Accountant and received the Robert Beyer Silver Medal in 1989 for securing the second highest total score in the CMA examination held in June 1989. He was also the recipient of the National Talent Search Scholarship awarded by the National Council of Education, Research and Training (NCERT, India) in 1979. Prior to joining the faculty at Albany, he was teaching at Rutgers University. His research and teaching interests are primarily in the areas of Auditing, Information Systems, Behavioral Finance and Management Accounting. His hobbies include mountaineering, rowing, racquetball and bridge.

Department of Accounting & Law
State University of New York at Albany

Acc 522. Statistical Methods for Business Decisions
Fall, 2001
J Gangolly / S Dutta

Tentative Schedule

August 28, 2001 (Gangolly)

Theme: *Data preprocessing for Statistical Analysis & Introduction to S-Plus*

Topics: *Unix shell scripting, tr, join; ftp, character coding, text processing*

Readings: *Unix man pages for tr, join, ftp; basic use of emacs or vi editors, etc.*

To Do: *Homework 1.* (Due September 15, 2001)

September 4, 2001 (Gangolly)

Theme: *Data & Graphics in S-Plus*

Topics: *Matrices, Frames, Arrays, and their manipulation; Basic graphics commands in S-Plus; Trellis graphics*

Readings: KO: Ch 3 – 6.

September 11, 2001 (Dutta)

Theme: *Fundamentals of Probability & Introduction to Bayesian Decision Theory*

Topics: *Probabilities: joint, conditional & marginal; Bayes' Theorem and Likelihood ratio. Nomenclature of decision trees.*

Readings: Readings on Probability Concepts and Decision Trees to be distributed.

September 18, 2001

Holiday, No Class

September 25, 2001 (Dutta)

Theme: *Bayesian Decision Theory & the Audit Risk Model*

Topics: *Continuation of previous class; the mechanics of Decision Trees. The construction of trees, method of Folding back, Audit Risk. Evidence aggregation in auditing.*

Readings:

"Achieved Audit Risk and Audit Outcome Space" William R. Kinney.

"Toward a more Consistent Model for Audit Risk" by John T. Sennetti.

October 2, 2001 (Gangolly)

Theme: *Descriptive Data Exploration & Statistical Modeling*

Topics: *Use of S-Plus Graphics; Regression.*

Readings: KO: Ch 7 – 8

To Do: *Homework 2* (Due October 31, 2001)

October 9, 2001 (Gangolly)

Theme: *Preprocessing for Datamining & Concept Description*

Topics: *Data Cleaning, Data Integration & Transformation; Data reduction & Data Compression; Concept hierarchy generation.*

Readings: HK: Ch. 3 – 5.

October 16, 2001 (Gangolly)

Theme: Association Rules in Large Databases

Topics: Association Rule Mining

Readings: HK: Ch. 6.

October 23, 2001 (Gangolly)

Theme: Classification & Prediction

Topics: Classification by Decision Tree Induction, Tree pruning, Extracting decision rules from decision trees, Basics of Bayesian Classification.

Readings: HK: Ch. 7.

October 30, 2001 (Gangolly)

Theme: Multivariate Methods I: Clustering

Topics: Hierarchical & Partitioning methods for Clustering, and their use in S-Plus.

Readings: HK: Ch. 8. and Clustering in S-Plus online manuals.

To Do: Homework 3 (Due November 15, 2001).

November 6, 2001 (Gangolly)

Theme: Multivariate Methods II: Other Methods

Topics: Multi-Dimensional Scaling, etc.

Readings: HK: Ch. 9.

November 13, 2001 (Dutta)

Theme: Continuous Probability Distributions: An Introduction.

Topics: Finding Probabilities from a Standard normal distribution, computation of z-values. Normal approximations of Binomial Distribution.

Readings: Fundamentals of Normal Distribution, To be distributed

To Do: TBD

November 20, 2001 (Dutta)

Theme: Confidence Interval and Test of Hypotheses.

Topics: Confidence interval estimation, sample size determination, Type I error and Type II error (alpha and beta errors).

Readings: Confidence Intervals and Hypotheses Testing, To be distributed.

November 27, 2001 (Dutta)

Theme: Application of Continuous Probability Distributions to Audit Decisions.

Topics: Concepts of Audit Risk and Materiality.

Readings: "Considering Multiple Materialities for Account Combinations in Audit Planning and Evaluation" Saurav Dutta and Lynford Graham

December 4, 2001

Test

December 11, 2001

Theme: Project Presentations

Updated on August 23, 2001 by Jagdish S. Gangolly (j.gangolly@albany.edu)