

Graduate School of Library and Information Science
Spring 2000

Data Analysis for LIS Research

(LIS 450)

Section DA Thursday, 12–2:50 PM Room 111, Speech and Hearing Building

David Dubin

Office: LIS 222

Office hours: Tuesdays, 1–3 PM

Phone: 217–244–3275 (217–BIG–EARL)

E-mail: dubin@alexia.lis.uiuc.edu

Web: <http://www.lis.uiuc.edu/~dubin>

This document is Copyright © 2000 by David Dubin and the Trustees of the University of Illinois. In addition to this syllabus, this course is governed by the rules and guidelines set forth in the document A Handbook for Graduate Students and Advisers which students receive upon admission to the program. Students should also consult, and take to heart, the Professional Guidelines and Codes of Ethics for Library and Information Science Professionals available from the GSLIS main office.

This syllabus is provided to UIUC students as part of the materials for a particular class. However, it may be copied, redistributed, and modified under the terms of the [OpenContent License](http://www.opencontent.org) (Version 1.0). The text of that license is available on the Worldwide Web at www.opencontent.org. Resources that are linked to or referenced from within this syllabus (e.g., readings, outlines, discussions) are not covered by the OpenContent License, unless specifically labeled as such.

REQUIRED TEXTS

Hartwig, F. and Dearing, B. E. *Exploratory Data Analysis* (Sage Publications, 1979).

Jacoby, W. G. *Data Theory and Dimensional Analysis* (Sage Publications, 1991).

Knoke, D. and Kuklinski, J. H. *Network Analysis* (Sage Publications, 1982).

Dubin, D. (instructor). *Reading Packet for LIS450-DA* (Campus Publishing Services, 2000).

SCOPE AND OBJECTIVES

This class is a survey of data analysis issues, tools, and techniques for research in Library and Information Science. Students will locate and work with a data set of their choice, review the literature of recommended analysis methods, and prepare an analysis appropriate to the data set they have chosen.

Objectives

- Survey techniques for data collection, elicitation, analysis, and visualization.
- Review assumptions underlying inferential analysis methods.
- Develop research strategies for honest and skeptical data analysis.

THIS SYLLABUS

The official syllabus for this course is the SGML version that is linked off the class web page. Expressions of the syllabus in other formats are derived from the SGML version. The current SGML version should be consulted to resolve any inconsistencies among other renditions.

BASIS FOR EVALUATION

Students are responsible for their performance in meeting their own educational goals and those of the course; instructors are responsible for providing guidance, expertise, and support to help students reach those goals. Students are expected to participate in class exercises and discussions. Satisfactory work will receive a grade in the C range, good work will receive a grade in the B range, and superior work will receive a grade in the A range.

Final grades will be calculated as follows:

- Annotated Bibliography 30%
- Class Presentation 20%
- Term Project: 40%
- Class Participation: 10%

Annotated Bibliography

Write a one-page overview of the data set you have chosen for the focus of assignments in this class. Include a description of the data's origin, and the type of data it represents (e.g., survey data, word frequency data, etc.). Outline the types of inferences a researcher might wish to draw from the data, and the inferential tools or methods that one would apply.

Review the literature of recommendations for analysis of the type of data you have selected. Include tests on the assumptions underlying the inferential methods described in the overview. Include whatever recommendations are relevant to approaching your data with honesty and skepticism.

The overview and annotated bibliography are to be prepared in a structured, plain text format that will be specified by the instructor. This assignment will be submitted as a machine-readable file.

Class Presentation

Schedule a class presentation relating to the data set you have chosen and its analysis. Presentations will take place during one of the final three class meetings prior to the course wrap-up and evaluation. Assign one or two readings from your annotated bibliography to the class no later than three weeks before your presentation. Place a copy of the readings on reserve no later than two weeks before your presentation.

Term Project

Analyze your data set according to the recommendations of the literature that you have reviewed. Prepare a research paper (approximately 20 pages in paper form) that reports the results of your analysis. Include

whatever graphs or other visualizations of the data are illustrative of your findings. Document the paper with appropriate references.

Renditions of the term project can take whatever form (paper or electronic) are most suitable for conveying the results of the analysis. However, the format in which the project is authored must be expressive enough for academic writing.

Class Participation

The class participation grade is based on consistent attendance, contribution to in-class and/or online discussions, and providing assistance to classmates outside of class. Please alert the instructor if a classmate has been of help to you outside of class.

SEMESTER OUTLINE

Part I: **Introduction**

January 20

Overview of the class

Readings: Syllabus

Part II: **Exploratory Data Analysis 1**

January 27

Displays of distributions, skewness, outliers.

Readings: Hartwig and Dearing ch. 1–2; McNeil ch. 1–2

Part III: **Exploratory Data Analysis 2**

February 3

Relations and transformations.

Readings: Hartwig and Dearing, ch. 3–4, McNeil ch. 3

Part IV: **Exploratory Data Analysis 3**

February 10

Normality, transformations, multiple comparisons

Readings: Lunn and McNeil, ch. 2, 3, and 6; Hartwig and Dearing ch. 5–6

Part V: **Measurement Theory**

February 17

Scales, invariance, appropriate statistics, order, additive structure.

Readings: Stevens, 1959; Michell, 1986

Part VI: **Elicitation and Pre-Analysis**

February 24

Knowledge elicitation tools, pre-analysis, imputation

Readings: Dubin, Kwasnik, and Tangmanee, 1996; Banks and Parmigiani, 1992; Levy and Lemeshow ch. 13

Part VII: **Dimensional Analysis 1** *March 2*

data theory, measurement

Readings: Jacoby, ch. 1–3

Part VIII: **Dimensional Analysis 2** *March 9*

dimensionality, scaling methods

Readings: Jacoby ch. 4–7

Part IX: **Spring Break** *March 16*

Part X: **Network Analysis** *March 23*

Network models: data collection and analysis

Readings: Knoke and Kuklinski

Part XI: **Scaling and Clustering** *March 30*

scaling vs. clustering

Readings: Kruskal, 1977

Part XII: **Visualization** *April 6*

Projection pursuit, VIRIs, dotplot analysis

Readings: Swayne, Cook, and Buja, 1998; Church and Helfman, 1993; Korfhage, ch. 7

Part XIII: **Presentations** *April 13*

Readings: Student-assigned readings

Part XIV: **Presentations** *April 20*

Readings: Student-assigned readings

Part XV: **Presentations** *April 27*

Readings: Student-assigned readings

Part XVI: **Wrap-up and Evaluation** *May 4*

Final Projects: Due May 11 at 5 PM.

READING ASSIGNMENTS

- [Banks and Parmigiani, 1992] Banks, D. L. and Parmigiani, G. (1992). Pre-analysis of superlarge industrial data sets. *Journal of Quality Technology*, 24(3):115–129.
- [Church and Helfman, 1993] Church, K. W. and Helfman, J. I. (1993). Dotplot: A program for exploring self-similarity in millions of lines of text and code. *Journal of Computational and Graphical Statistics*, 2(2):153–174.
- [Dubin et al., 1996] Dubin, D., Kwaśnik, B. H., and Tangmanee, C. (1996). Elicitation techniques for classification research. In Fidel, R., Beghtol, C., Kwaśnik, B. H., and Smith, P. J., editors, *Advances in Classification Research*, volume 5 of *ASIS Monograph Series*, pages 33–68. Information Today, Inc., Medford, NJ.
- [Hartwig and Dearing, 1979] Hartwig, F. and Dearing, B. E. (1979). *Exploratory Data Analysis*, volume 16 of *Quantitative Applications in the Social Sciences*. Sage, Newbury Park, CA.
- [Jacoby, 1991] Jacoby, W. G. (1991). *Data Theory and Dimensional Analysis*, volume 78 of *Quantitative Applications in the Social Sciences*. Sage, Newbury Park, CA.
- [Knoke and Kuklinski, 1982] Knoke, D. and Kuklinski, J. H. (1982). *Network Analysis*, volume 28 of *Quantitative Applications in the Social Sciences*. Sage, Newbury Park, CA.
- [Korfhage, 1997] Korfhage, R. R. (1997). *Information Storage and Retrieval*. Wiley Computer Publishing, New York.
- [Kruskal, 1977] Kruskal, J. (1977). The relationship between multidimensional scaling and clustering. In Van Ryzin, J., editor, *Classification and Clustering*, pages 17–44. Academic Press, New York.
- [Levy and Lemeshow, 1999] Levy, P. S. and Lemeshow, S. (1999). *Sampling of populations: methods and applications*. Wiley, New York.
- [Lunn and McNeil, 1991] Lunn, A. D. and McNeil, D. R. (1991). *Computer Interactive Data Analysis*. Wiley, New York.
- [McNeil, 1977] McNeil, D. R. (1977). *Interactive Data Analysis: A Practical Primer*. Wiley, New York.
- [Michel, 1986] Michel, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100(3):398–407.
- [Stevens, 1959] Stevens, S. (1959). Measurement, psychophysics, and utility. In Churchman, W. C. and Ratoosh, P., editors, *Measurement Definitions and Theories*, chapter 2, pages 18–37. Wiley, New York.
- [Swayne et al., 1998] Swayne, D. F., Cook, D., and Buja, A. (1998). XGobi: interactive dynamic data visualization in the X Window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130.