

Anthropological Sciences 192
Data Analysis in the Anthropological Sciences

Winter 2006

Draft: 11 January 2006
(v. 3.1)

Instructors:

James Holland Jones
Department of Anthropological Sciences

Office: Building 360, Room 361I
Phone: 3-4824
email: jhj1@stanford.edu
Office Hours: Friday 1:15-3:15, or by appointment

Ian G. Robertson
Department of Anthropological Sciences

Office: Building 360, Room 361E
Phone 4-4994
email: igr@stanford.edu
Office Hours: Thursday 1:00-3:00, or by appointment

Course Description:

In this course, we will develop a statistical toolkit appropriate for anthropologists. The emphasis will be on practical data analysis and the development of a problem-solving approach to inference. Students taking the course will gain skills that significantly improve their ability to:

- read quantitative arguments in an informed, critical way
- use computers to manage and analyze anthropological data
- convey quantitative arguments in scholarly publications
- design research projects to generate data that can be quantitatively analyzed

No extensive prior knowledge of statistics or computers is assumed, but a reasonable background in anthropology and/or archaeology is. The course will require a substantial commitment of time, including extensive use of computers for analysis and presentation work.

Readings

There is one text for this class. The text (Dalgaard 2002, below) will be supplemented with occasional readings from the primary scientific literature.

Dalgaard, P. 2002. *Introductory Statistics in R*. New York: Springer.

Software

We will use a freely-available software package called R. R has a number of advantages: (1) it is free, (2) it is essentially platform independent, (3) R represents the cutting-edge in statistical computation, (4) there is extensive (free) documentation and a large online community of users to provide support, (5) it is used overwhelmingly by professional statisticians and therefore learning R facilitates conversations with statisticians for consulting and collaborative purposes. *Did we mention it's free?*

R may be downloaded from one of the CRAN mirror website. The closest mirror lives across the Bay.

<http://cran.cnr.berkeley.edu/>

If you find going to a Berkeley website for the software too distasteful, perhaps you would prefer the UCLA mirror:

<http://cran.stat.ucla.edu/>

Grading

50% Weekly Problem Sets
25% Midterm (Take Home)
25% Final (Take Home)

Problem Sets

Much of the practical benefit from this course will come from working through problem sets; these will be assigned almost every week and will be due in class on the assigned date. Since we will be discussing the problems in class, late assignments will only be accepted under unusual circumstances.

Problem sets will be of two general types: (1) practice problems and (2) vignettes. For the latter, we will expect a more involved write-up. The goal of these vignette problem sets is to get students to think about integrating quantitative and inferential thinking into their scholarly argumentation. Marks for vignette problem sets will reflect the quality of the analysis and the quality of the arguments made concerning relevant anthropological questions. Be thorough (but to the point) in answering all questions posed in the problem set handout. Results of analyses should be presented in a brief, typed essay that references supporting tables and figures and does not exceed three double-spaced pages of 11-point or larger font, and 1-inch margins (tables, figures, citations are not included in page limit). Especially in the case of non-computer based calculations, show clearly the steps that were followed in arriving at the answer. Figures and tables do not have to be of publication quality, but they should also be clear, neat and appropriately labeled. Practice problem sets do not need to be so formally presented, but should similarly fulfill the spirit of clarity, brevity, and explication of the steps leading to the final answer (i.e., show your work!).

In your problem set write-ups, briefly state the problem at hand. You may assume a reader's knowledge of methods that have been discussed or presented in class, but beyond that, you need to summarize the results and general argument as you would in an article for publication. Be exact about procedures that have been used—don't say "correlation" if you mean "product-moment correlation coefficient;" don't say "factor analysis" if you have just done a principal components analysis. Be clear about the set of data you have analyzed, any transformations

you have used, what has been shown quantitatively, and any substantive conclusions you arrive at. With respect to conclusions, be careful not to overstate things or to claim too much—learning when to rein yourself in is part of what we want to teach here. Finally, you should adopt a style of presentation that is relatively serious and formal—again, the model you should have in mind is an article for a professional journal.

You may find it is helpful to coordinate your computing time with other students so that you can assist each other in solving technical problems associated with the use of R. Discussing the methods and issues with each other is fine, but you need to execute procedures and write up problem sets independently—much of the pay-off in this course lies in eventually struggling through difficult concepts on your own.

Course Website

The course website will be a repository for R code, data sets, problem set answers, links to useful online resources, etc. The url for the course website is:

<http://anthsci192.stanford.edu>

Syllabus

Lectures are on Monday and Wednesday from 1:15-3:05 in Building 360, room 361J.

Week 1. Introduction to R

01.11 Workshop on R installation

Dalgaard, chapter 1.

Week 2. Working with Data: Descriptive Statistics. Frequency Distributions, Histograms, Transformations, Correlation

01.16 NO CLASS: Martin Luther King Jr. Holiday

Dalgaard, chapter 2, 5.

01.18 R Basics: Input and Manipulation & Elementary Data Analysis

Whallon, Robert. 1987. Simple Statistics. In *Quantitative Research in Archaeology: Progress and Prospects*, edited by M. S. Aldenderfer, pp. 135-150. Sage Publications, Beverly Hills, California.

Week 3. Working with Pictures: Visual Presentation, Graphics, EDA

01.23 1000 Words

Dalgaard, chapter 3.

01.25 Exploring Data

Cleveland, William S. 1985. *The Elements of Graphing Data* (Chapters 1 and 2).
Wadsworth Advanced Books and Software, Monterey, California.

Week 4. Probability and Probability Distributions

01.30 Probability theorems, Bayes's Theorem, Continuous distributions (Normal, Gamma, Chi-Square, F)

Hilborn, R., and M. Mangel. 1997. *The ecological detective: Confronting models with data*. Vol. 28, *Monographs in Population Biology*. Princeton: Princeton University Press. Chapter XXX.

02.01 Likelihood, Discrete distributions (Binomial, Poisson, Negative Binomial)

Eliason, S. R. 1993. *Maximum Likelihood Estimation : Logic and Practice*. New York: SAGE.

Week 5. One and Two Sample Tests, ANOVA, Kruskal-Wallis

02.06 The Logic of (Statistical) Hypothesis Tests

Dalgaard, chapter 4-6.

02.08 Applying Hypothesis Tests

TAKE-HOME MIDTERM EXAM

Week 6. Regression

02.13 Multiple linear regression

Dalgaard, chapter 9.

Schreiber, Katharina J., and Keith W. Kintigh 1996. A Test of the Relationship Between Site Size and Population. *American Antiquity* 61(3):573-579.

Kvamme, Kenneth L. 1997. A Wider View of the Relationship between Settlement Size and Population in the Peruvian Andes. *American Antiquity* 62(4):719-722.

Harvey, P. H., and J. R. Krebs. 1990. Comparing brains. *Science*. 249:140-146.

02.15 More linear regression, logistic regression

Dalgaard, chapter 10 & 11.

Harpending, H. C., and R. Pennington. 1991. Age structure and sex-biased mortality among Herero pastoralists. *Human Biology* 63 (3):329-353.

Lydie, N., N. J. Robinson, B. Ferry, E. Akam, M. De Loenzien, and S. Abega. 2004. Mobility, sexual behavior, and HIV infection in an urban population in Cameroon. *JAIDS-Journal of Acquired Immune Deficiency Syndromes* 35 (1):67-74.

Week 7. Tabular Data, Loglinear Models

02.20 Data Tables and the Analysis of Proportions

Dalgaard, chapter 7.

Roberts, J. M., C. C. Moore, and A. K. Romney. 1995. Predicting Similarity in Material Culture among New-Guinea Villages from Propinquity and Language - a Log-Linear Approach. *Current Anthropology* 36 (5):769-788.

02.22 Parametric Models for Count Data: Poisson Regression

Morris, M. 1991. A Log-Linear Modeling Framework for Selective Mixing. *Mathematical Biosciences*. 107(2):349-377.

Ellis, M., and C. Muschkin. 1996. Migration of persons with AIDS - A search for support from elderly parents? *Social Science & Medicine* 43 (7):1109-1118.

Week 8. Multivariate Analysis

02.27 Exploring Data Structure

Madsen, Torsten 1988. Multivariate Statistics and Archaeology. In *Multivariate Archaeology: Numerical Approaches in Scandinavian Archaeology*, edited by T. Madsen, pp. 7-27. Jutland Archaeological Society Publications, vol. XXI. Jutland Archaeological Society, Moesgard, Denmark.

03.01 Reducing Dimensionality and Clustering

Patton, J. Q. 2004. Coalitional Effects on Reciprocal Fairness in the Ultimatum Game: A Case from the Ecuadorian Amazon. In *Foundations of Human Sociality Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, edited by Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. New York: Oxford University Press.

Week 9. Survival Analysis

03.06 The Messy Statistics of the Living: Censoring and Truncation

Dalgaard, chapter 12.

Wood, J. W., D. J. Holman, K. M. Weiss, A. V. Buchanan, and B. LeFor. 1992. Hazard models for human population biology. *Yearbook of Physical Anthropology* 35:43-87.

03.08 I'm Waiting for My Man: Waiting Time to Events and Their Statistical Analysis

Galdikas, B. M. F., and J. W. Wood. 1990. Birth spacing in humans and apes. *American Journal of Physical Anthropology* 83:185-191.

Jones, J.H., M.L. Wilson, and A.E. Pusey. 2006. The Demography of the Gombe Chimpanzees: An Analysis of 42 Years of Vital-Event Data. Submitted Manuscript.

Week 10. Distance Measures, Cluster Analysis, Trees, etc.

03.13 Distance Measures

Cowgill, George L. 1990. Why Pearson's r Is Not a Good Similarity Coefficient for Comparing Collections. *American Antiquity* 55(3):512-520.

03.15 Taxonomies and Phylogenies

Sneath, P.H. and R.R. Sokal 1973. *Numerical Taxonomy*. San Francisco: W.H. Freeman and Company. chapter?

Felsenstein, J. 2003. *Inferring phylogenies*. Sunderland, MA: Sinauer. Chapter 11.

Finals Week