Statistics 104 - Term Project - Fall 2004

- *Proposal: due Monday December 13, 2004 in class. (See instructions below.)*
- *Completed project: due Friday, January 14, 2005 by 4:00 PM.*

General description

The term project should allow you to see statistics applied to your field of interest. There are many possible projects; some are described here. You should work individually on this project, which should result in a 5 page paper (not counting tables and figures). Length is not an asset if it is not associated with increased content! The objective is for you to learn something and tell us about it.

You are encouraged to talk to the instructor or your section leader about ideas for a project. You may get a referral to another TF or faculty member who is particularly expert in the field you are interested in. Also, your instructor in another course you are taking in a field you are interested in may be able to help you develop some ideas.

You should hand in a **project proposal** by the due date given above. It is a paragraph describing what you plan to do. Be as specific as possible. The proposals will be returned within a week. The proposal is regarded as a homework assignment, i.e. it is required and should be done on time.

You will hand in your proposal, and perhaps do some of the work on the project, before the end of the lectures. Keep in mind that by the time you finish your project, you will have seen a lot more than we have done so far on hypothesis testing and estimation, including regression, analysis of contingency tables, and comparisons of means of several groups.

Some projects will involve considerably more effort than others. Generally speaking, an acceptable project on a safe but limited topic (for example, if you just report on an article without getting into new statistical ideas or going beyond the analysis presented in the article) will earn an average grade. More intellectually challenging projects have greater potential to earn an outstanding grade. While the project counts for 15% of your final grade, the biggest payoff of a more challenging project is in the opportunity it provides you to solidify and extend your understanding of the material in the course and to obtain practical experience in applying it to your own research concerns. In doing this project you may well have better access to statistical consulting than you will when doing your thesis!

You may want to do a project using data you have from another course (whether from an experiment or through access to a data set somebody else has collected). This may be a good way to apply statistics to something you have thought about. If you do a project of this sort, you must make very clear which part of the project is done specifically for your statistics course and which part is just a review or copy of work you have done for another course. Normally you would *not* just hand in the same paper to two courses.

If you are doing a collaborative project or if you are doing a project on a subject or data set that ties in with another class, you should read the relevant rules on pages 61–62 of this year's *Handbook for Students* concerning collaboration and submission of work to more than one course.

If you plan to do a project involving a data set or article from a book that is listed here, please copy the pages you need and return it to the library, so that others can use it!

Project format

Put a cover sheet on the front with title, name of project, date, and name and section of author or authors. If this project is being used in part for another class, make sure you note this on the title page. Also the project should be word processed, not handwritten.

Computer output should be incorporated in the usual way, i.e. put tables in the text or at the end but do not hand in a pile of unedited computer output. Tables and figures should be numbered and captioned. Bulky binders are not appreciated, nor are large numbers of superfluous tables.

Some ideas for projects

1. **Analysis of a data set that is available to you:** Perform a statistical analysis of some data set from an experiment, survey, or secondary data source. You should pay critical attention to issues concerning how the data were collected as well as to the statistical analysis. (Depending on the nature of the data and your own relationship to it, you may want to give more or less emphasis to explanation of the data collection.) You should make sure that your data set has enough complexity (more than just a couple of variables, and a decent number of observations) to support an interesting analysis.

   With this type of project you must be careful to explain who collected the data and to give proper credit for the part of the work that was already done elsewhere, even if you did it yourself. See the Harvard rules referenced in the introduction.

   A word of warning: some of you may be interested in working with time series data (such as stock market prices, exchange rates, cost of living index varying over time in a single country). This course does *not* teach the appropriate methods for formal statistical analysis of data of this type. Be very careful about using these data because you could easily slip into a nonsensical analysis. Cross-sectional data (e.g. from many different countries, areas, or other units, at the same time) can generally be analyzed better within the framework of methods we have touched on. (You can repeat the cross-sectional analysis for a couple of times if you just want to make a comparison.)

   Data in publications and books: There are many economic data sets, usually consisting of some kind of aggregates, in various statistical abstracts and data books, such as yearbooks of data that are published for various industries, handbooks of country data, population handbooks, etc. These can be found in libraries such as Littauer. A few of the most important sources of this kind are the following:

   - *World Development Report*, published annually by the World Bank. A good source of cross-country data on income, employment, health and other development indicators.
   - *Human Development Report*, published annually for the United Nations Development Program. There are a number of other statistical reports from the UN and other international agencies like the International Labor Organization.
   - *International Financial Statistics*, published monthly, with annual yearbooks, by the International Monetary Fund. A good source of cross-country and time-series data on financial indicators: exchange rates, money, interest rates, balance of payments, etc.
   - *Government Finance Statistics*, annual from the International Monetary Fund. A good source (the only source?) of cross-country and time-series data on public finance: government expenditure and revenue by category and level of government, debt, deficits, etc.
   - *Economic Report of the President*, annual. Tables at the end of the report provide data on macroeconomic indicators for the United States only.
   - *Survey of Current Business*
   - *Statistical Abstract of the United States*. Full of all sorts of statistical tables.

- There are many references (starting with the business pages of today's newspaper) that contain information on stock prices, mutual fund indices, exchange rates, etc.

Of course, you may do an analysis combining information on the same countries from more than one of these sources. For Economics concentrators, data of this type may be applicable to fitting or testing some of the relationships you learned about in Social Analysis 10 or Economics 1010.

Some books in other areas that include data sets are the following:

- *Data: a Collection of Problems from Many Fields for the Student and Research Worker*, by D.F. Andrews and A.M. Herzberg
- *Case Studies in Biometry*, edited by Lange, Ryan, Billard, Brillinger, Conquest and Greenhouse

If you are interested in sports statistics, you probably know where to find them; there are a plethora of almanacs containing this kind of information.

Also, articles in books and journals sometimes contain the original data set and you may have an idea for a different analysis than the one which the author did. See below for articles and journals, and remember that you should distinguish carefully between what you did and what was in the original article.

Data from other Harvard sources: You may have a data set from experimental or survey research you worked on in another course or from your job or volunteer work, or your adviser or instructor in another class may have something you can work with.

There are survey research data sets at the Government Data Center (in Littauer) and at Government Documents (Lamont Library, first floor). These are mostly large data sets from government surveys and privately conducted opinion polls. Some of these may be in an inconvenient form for use (9-track tapes and CD-ROM), but the staff at these two centers should be able to help you. The Murray Research Center (in the Radcliffe Yard) has a lot of data sets as well, mostly from research studies on personality development, family history, and similar topics; in some case they can make these data available on diskette. Many of the data sets that are available at these locations contain individual-level data, and each data set should be accompanied by some description of how it was collected. These data sets are already entered in computer-readable format, so you don't have to spend your effort entering them into the computer; instead, you may need to cut down the data set to a manageable size.

Data sources on the Internet: There is an increasing amount of data available on the Internet. As with other Internet materials, there is some gold out there and a lot of pure junk. If you would like to browse around for data on a topic you are interested in, you can start from the Statistics Department home page, "`http://www.stat.harvard.edu/`", and look under "Non-academic Resources." As with the large data sets available at Harvard, these have the advantage of being already entered into computer-readable format, but the disadvantage that they may be very large (more than you need).

2. **Review of an article:** Read an article or several articles in which statistics or probability is applied to a problem that interests you. You may want to focus on describing what was done in the application. Alternatively, you may take a more critical approach, discussing fallacies in the use of statistical

reasoning in a paper and other analyses and hypothesis that you think should have been tested. Of course, the more of your original ideas you can incorporate the better the project (if the ideas make sense!).

When you do this project, you want to show how the methods of Stat 104 can be applied to the problem, *not* your mastery of some nonstatistical subject matter. Don't choose a topic which has little statistical content, and don't spend all your time arguing about nonstatistical aspects of a book or paper.

The following are some sources of readings for this type of project. You may also use any article from your field of study in which statistical arguments are used. Your professors or TF's in other fields may be able to suggest good articles that involve statistics at the appropriate level. Make sure you include enough description to make it clear to the person grading your project what the article was about; you can attach a copy of the article if you like.

You may find these books and journals in various libraries – try looking them up with HOLLIS. As noted above, you should copy what you need and leave the book in the library, since many students will be trying to get access to these same books.

Books:

- *Econometric Models and Economic Forecasts*, Pindyck and Rubinfeld
- *Statistics: A Guide to the Unknown*, Tanur et al. — this book contains short articles; you should read 2–3 related articles.
- *Statistics and Public Policy*, Fairley and Mosteller
- *Statistics and the Law*, DeGroot, Fienberg and Kadane — a series of readings on various legal topics.
- *Mathematics in the Archaelogical and Historical Sciences*, Hodson, Kendall and Tautu — a survey of quantitative methods used in these fields.
- *The Mismeasure of Man*, Gould — a historical review that contains chapters at a variety of technical levels.
- *Quantitative Analysis of Social Problems*, Tufte
- *Quantification in American History*, Swierenga
- *Statistical Problems of the Kinsey Report*, Cochran, Mosteller, Tukey
- *The Theory of Committees and Elections*, Black
- *Artificial Intelligence and Statistics*, Gale
- *Pygmalion in the Classroom*, Rosenthal and Jacobsen — a study of how teacher expectations affect student performance (discussed briefly by Professor Rosenthal in his guest lecture).
- *Social Statistics in Use*, Hauser
- *Statistical and Mathematical Aspects of Pollution Problems*, Pratt
- *Farmwork & Fieldwork: American Agriculture in Anthropological Perspective*, Chibnik
- *Evolutionary Operation*, Box and Draper
- *Environmental Health: Quantitative Methods*, Whittemore
- *Optimal Strategies in Sports*, Ladany and Machol
- *The Data Game: Controversies in Social Science Statistics*, Maier – chapters focusing on the definition, collection, and interpretation of various kinds of policy-relevant statistics.
- *The Fascination of Statistics*, Brook, Arnold, Hassard, and Pringle

- *The Statistical Consultant in Action*, Hand and Everitt (some of the articles in this collection include data)
- *Data Analysis for Politics and Policy*, Tufte
- *Incomplete Data in Sample Surveys, Vol. 1: Report and Case Studies*, Madow, Nisselson and Olkin
- *Education and Class: the Irrelevance of IQ Genetic Studies*, Lewontin and Schiff
- *The Bell Curve*, Herrnstein and Murray
- *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*, Devlin, Fienberg, Resnick, and Roeder
- *Experiment in Plant Hybridization*, Gregor Mendel (the original research in genetics, republished 1965 by Harvard University Press)
- *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Arminger, Clogg and Sobel — statistical methodology for sociology
- *Quantitative Sociology: International Perspectives on mathematical and statistical modeling*, Blalock et al.
- *Model Building in Sociology*, Abell

Journals: There are many journals which include articles with statistical analyses at an accessible level; in some cases the original data sets are also included. For example, marketing is an area of economics / business management in which both survey and experimental research are prominent. Demography is another area in which there are many statistical articles, often with data sets attached. Psychology, biology and medicine are areas in which many articles will include at least some statistics. Talk to instructors in your field about what journals make use of statistical methods.

On the other hand, it probably doesn't make sense for most of you to use an article from a very sophisticated journal (such as an advanced econometric journal), which assumes more technical background than you have. If you decide to look for a journal article, you should be prepared to search through a number of articles to find one that you are interested in and understand that has substantial statistical content.

- *Journal of Consumer Research*
- *Journal of Marketing*
- *Journal of Marketing Research*
- *Population Studies*
- *Chance* (a popularly-oriented statistics magazine)
- *Ecology* (particularly Volume 74, No. 6, a special issue on statistical methods)
- *Journal of Experimental Zoology*
- *New England Journal of Medicine*
- *Public Opinion Quarterly*
- *Psychological Bulletin*
- *Journal of Applied Psychology*
- *Journal of Applied Behavior Analysis*
- *Political Behavior*
- *American Sociological Review*
- *Journal of Mathematical Sociology*

3. **Original application of probability and statistics:** Think about how statistics or probability might play a role in some topic of interest to you. For example: How can drug testing and AIDS testing be improved to avoid the "false positive" problem (confirmatory test, multiple tests)? Can and should we adjust the US Census for undercounted minorities? Are there good strategies for playing the lottery?

   While you couldn't do a complete statistical design of a research project in a short paper like this, you should lay out your thoughts on at least some aspects of the design and/or analysis of some sort of study on the topic you choose. At least be clear about the question you are trying to answer in the study you describe. It is recommended that you speak to a TF or the course head before doing this type of project.

4. **Report on a statistical method:** Read about and describe a technique that we don't discuss in class. Possible sources include other introductory statistics texts and some of the books listed above. If you use a section from a textbook, one way to show your understanding is by doing several problems from the end of the section. Of course it is up to you (with the help of your section leader) to make sure you are actually doing something that is not covered in Stat 104. Some suitable texts for this type of project that are more advanced than Moore and McCabe are the following:

   - *An Introduction to Mathematical Statistics and it Applications*, Larsen and Marx (a first statistics text at a more mathematical level)
   - *Applied Linear Regression Models*, Neter, Wasserman, and Kutner (a more advanced regression text)
   - *Applied Linear Regression*, Weisberg
   - *A First Course in Probability*, Ross (textbook on probability theory)
   - *Planning and Analysis of Observational Studies*, Cochran
   - *Making Decisions*, Lindley (introduction to statistical decision theory)
   - *Introductory Lectures on Choices under Uncertainty*, Raiffa (decision theory)
   - *Elementary Decision Theory*, Chernoff and Moses (similar topic but more difficult than the two just above)
   - *Random Walks in Biology*, Berg (somewhat mathematical discussion of motions of particles and molecules)

5. **Design/analysis of an experiment or survey:** An individual or group might design an experiment or survey to answer a particular question, carry it out and analyze the data. Some ideas: What is the fastest way to get from the river houses to the Science Center? Survey your classmates to determine if the amount of sleep one gets at night effects the frequency of sleeping in class. Organize an ice-cream taste test to determine the best ice cream in Harvard Square. Measure the flight time of different kinds of paper airplane. Survey the attitudes of Harvard students on some social or cultural issue or some aspect of Harvard life, and see how responses are related to some other variable.

   In this type of project, it is important to be thoughtful about (and provide an adequate description of) the methods and design of the study, and the possible biases associated with these methods, as well as the analysis of the data. You also need to be realistic in planning your research design: can you carry out what you have planned within a reasonable time period and investment of your own energy? The quality of the final product is what we are looking for, not just the amount of perspiration that went into it! Finally, you should make use of the concepts and methods learned in this course, and not just general knowledge, in planning and writing up this type of project.