

STATISTICS 210 -- Final ReportThe Data Set

Each student is required to obtain a data set to be used as the basis for a practicum. This data set must meet the following minimal specifications:

1. It must contain at least 100 independent observations.
2. The sample should meet minimal standards for scientific validity.
3. It must contain at least one quantitative dependent variable and one binary qualitative dependent variable. In the case that a qualitative dependent variable is not available, then one can be constructed from the quantitative dependent variable by defining two categories, such as those which are normal vs. abnormal, high vs. low, etc. by relevant criteria. For example, hypertension is classified as a systolic blood pressure over 140; poverty as an annual income less than some \$ amount, etc. So for these examples you could do an analysis of factors associated with SBP or \$ income as a quantitative dependent variable, and then hypertension (yes or no) or poverty (yes or no) as a qualitative dependent variable.
4. It must contain at least one quantitative independent variable and one qualitative independent variable of primary interest.
5. It must contain at least one other covariate in addition to the independent variables in (4). Preferably it should obtain both a quantitative and a qualitative (or binary) covariate.
6. One or more relevant scientific hypotheses must be stated for each of the qualitative and quantitative dependent variables.

Some of you may have difficulty getting access to a fresh data set. Many students get a data set from their work, but others from a friend. Don't be bashful, ask around. Students who find a fresh data set do a much better job (and get a better grade) because they are able to become more familiar with the questions and the data. Another advantage is that you would have a substantive "scientist" to interact with, they collected the data for some reason.

However, if you wind up with a data set from the www or a book, then that's OK. Be sure to specify the precise source.

Be careful when selecting the data set. Data sets with "survival" data are not appropriate; neither are data with repeated measures. Some of you may consider economic data from some source. Often such data will not meet the requirement that the observations be statistically independent. For example, daily stock prices for stocks are not independent and have autoregressive errors. Such data are not acceptable.

By lecture 7 a report is due which describes the study objectives for which the data was collected, the nature of the variables and the sample of observations and the analyses to be performed. Separate analyses should be conducted for the quantitative dependent variable using appropriate methods which may include 2 sample tests, multiple regression, ANOVA for unbalanced data or ANCOVA/GLM models. For the qualitative dependent variable, simple 2-way contingency tables, stratified analyses or logistic regression models should be employed as appropriate.

The Final Report

The final report should consist of a scientific report and an appendix with SAS programs and the *principal* output (not everything), just like the exercises. The scientific report should be written as intended for a more general scientific or administrative audience, the same as for the exercises. The appendix should document your analyses and any hand calculations. Please label where every Table or Figure is generated either in the SASLOG or in the listing file.

In addition to the written technical report, you should prepare a 15 min. presentation of the scientific report for the entire class on the evening of the "final." Use handouts for the class or transparencies for an overhead projector.

The following is an outline of the types of topics to be addressed in each report.

Scientific Report

This report should consist of four sections, using the standard format for scientific reports and publications. It should be typed, not to exceed 20 double spaced pages (10 single spaced), excluding tables and figures.

1. Introduction. This presents the background and rationale for the study. It states the objectives of the study, and their etiology. Hypotheses are stated and it is specified which are exploratory and which are confirmatory.

2. Methods. This describes all aspects of study design and execution. Issues concerning the nature of the samples and the measurements are addressed. The independent variables and dependent variables are defined. Some important issues are the representativeness of the sample, the precision (reliability) of the measurements, and the adequacy of the sample size in terms of the precision of confidence interval estimates

or the power of the most important tests of significance. The principal statistical methods of estimation and testing are described. Significance levels to be employed for significance are defined for use in cases where multiple tests of hypotheses arise (e.g. Bonferroni). If exploratory analyses are to be performed, procedures for model validation are described.

The assumptions of all statistical methods should be clearly stated, as well as the precise nature of all statistical models employed. Where appropriate, tests of the assumptions of each analysis should be described. Where appropriate, the statistical methods applied should incorporate all of the activities in the course exercises, and your analyses should cover all of the topics addressed in the exercises.

All of this should be stated concisely in a way that can be understood by a general audience.

3. Results. This section describes the results obtained in the study, i.e. the results of the statistical analyses. Remember that this is written for a non-statistical audience. Basic descriptive statistics (e.g. the means or proportions, etc.) are presented and significant differences or associations are cited. Tables and figures (i.e. graphs or charts) should be used to present and summarize the results. Refrain from interpretation in this section. Just present the analysis results and stick to the facts. Address the nature, strength and significance of all findings.

4. Discussion. Now interpret and discuss the results in light of the originally stated objectives and hypotheses. Consistencies (or inconsistencies) within this data, and the strengths and weaknesses of the study and its results are described. General recommendations are given. All of this should consider the more general literature about the subject of investigation.

Appendix

The principal SAS programs and output from the analyses should be included with a cross-reference to the results presented in the report. For example, point out which output is used for each Table, Figure and major statement in the text. I don't want a box of paper; but I do want to make sure that the analyses were properly implemented. This is to be limited to less than 1 inch thick; there is a limit to the amount of paper I can lug around.

All hand calculations used to support statements in the introduction and methods sections should be presented (e.g. confidence intervals).

If you want me to return your graded reports, include a self addressed envelope with proper postage for the report to be mailed to you; indicate whether you want the computer output to be returned or for me to discard it. Otherwise you may pick it up at

the GW Stat. Dept. office or at the Biostatistics Center (specify which). After 2 weeks, unclaimed reports will be destroyed .