Annotated Bibliography of Sources of Real-World Datasets Useful for Teaching Applied Statistics

Judith D. Singer and John B. Willett
Harvard University


Ninth grade student achievement (40 subjects) on tests of seventh grade material, with and without idioms and figures of speech. Useful for bivariate analysis.


Several extensive data sets describing the blood chemistry (cholesterol, blood pressure, etc.), cardiovascular state, socioeconomic status, and year of death. Some censored cases, could be used in the teaching of survival analysis. Other datasets include body flexibility, diet, testosterone levels in right and left testes of mice (!), weaning of rats, and infant cognitive development.


Depression scores and selected covariates for 294 participants in the Los Angeles Depression Study. Data set includes individual item responses for a 20 question depression scale, person background characteristics and selected health variables.


A large variety of categorical data sets including: tenure in American universities, dolphin sightings, transitions between Piagetian stages, college expectations and participation in high school athletics, political preferences, religion and marijuana, sudden infant death.

1An earlier version of this bibliography was published in The American Statistician, 1990, 44(3), 223-230. The order of the authors has been determined by randomization.

Data concerning the effect of the "Personalized System of Instruction" on course grades in an intermediate macroeconomics course, useful for logit analysis and log-linear modeling.


Average brain weights and body weights for 62 species of mammals. Both variables are very skewed, but logarithmic transformations alleviate the skewness and improve the linearity of the scatterplot.


71 sets of raw data. Many substantive areas are included, but the emphasis is generally on the physical and natural sciences. Several interesting social science examples are given, including: unemployment statistics, insurance rate information, literary data sets (the Federalist papers data set and another on Platonic prose rhythm) and the birthday/deathday problem.


Measures of the moral integration, ethnic heterogeneity, crime, welfare effort, integration and mobility of residents in 43 American cities.


Salary data by rank, sex, and tenure status for faculty at 1,901 colleges and universities. Institutions are categorized according to Carnegie classifications.


Contingency table of the relationship between gestational age and neurological status for 505 babies. Also see detailed log-linear analysis of these data in: Green, J. A. (1988), Loglinear analysis of cross-classified ordinal data: Applications in developmental research, *Child Development*, 59, 1-25.


Multiple multiway contingency datasets on the use of proverbs and mechanical cliches used by 133 college freshmen in 1937. Data are broken down by urban/rural, gender, etc.

One of many sources describing the more than 1,500 four-year colleges in this country. Relevant data include: number of applicants, number of students accepted, number of students enrolling, mean SAT scores of incoming freshman, mean class rank of incoming freshmen, faculty/student ratios, financial aid available, number of part-time students and faculty, percent of faculty with doctorates, sex composition of student body. Can be supplemented with information from American Association of University Professors salary survey and endowment data given in the Digest of Education Statistics.


Individual data for 25 college sophomores at the University of Texas on the speed and accuracy with which they solved four types of arithmetic problems (addition, subtraction, multiplication and division).


Scores on nine tests for 37 of the "best" students and 37 of the "worst" students, with notations of class rank, designed "to be of assistance to college authorities in aiding freshmen to adjust themselves to their environment" (p. 381).


A large variety of non-social science data sets (lawn service, real estate market, professional sports, foreign food), with some social science data sets scattered here and there: categorical data on health issues in children by graduating class of pediatrician, starting salaries of MBA graduates, etc.


Dozens of data sets relevant for economics on topics such as prices of color television sets, common stocks, gold, and other commodities. An added benefit is that all the data sets are entered onto a floppy disk included with every book.


Multidimensional data on leisure activities by occupation. Useful for multidimensional scaling, cluster analysis, factor analysis, principal components analyasis, discriminant analysis, etc.

A variety of educational growth data sets suitable for repeated measures/ MANOVA analysis, including data on responses to inkblot plates by grade and IQ over time, longitudinal (4 grades) data on scaled vocabulary scores for boys and girls, and so on. (Data repeated in Finn, J. D., & Mattsson, I. (1978). *Multivariate Analysis in Educational Research*. Chicago, IL: National Educational Resources.


Multi-dimensional categorical data describing perceptions of course performance among men and women in physics and chemistry courses at Stanford University.


Raw data for 46 children divided into four groups -- stutterers, well-adjusted, medium adjusted and poorly adjusted. Measures include age, achievement, receptivity to education, physical condition, social-personality traits, insightfulness, family background, somotype and anthropometrics.


33 raw datasets. Many substantive areas are included, and many of these data sets are just plain interesting, such as the ages of signers of the Declaration of Independence, murder/suicides by crashing private airplanes, heights of singers in the New York Choral Society.


Ten longitudinal data sets (10 waves) on measures of computation, color-naming and opposites-naming for 22 college-age males in New York at the turn of the century. Suitable for growth curve analysis.


Data for on the mother, father and children in eight families, with the age of each family member and their scores on five tests (addition, subtraction, multiplication, division and copying figures). The author concludes that "it is difficult to avoid the conclusion that ... likeness is due to heredity" (p.16).

Two large data sets: (1) a 20 variable subset of the PROJECT TALENT data (234 males, 271 females); and (2) the RECTANGLES data set on the dimensions of 100 rectangles, useful for factor analysis and principal components analysis.


39 raw datasets. Relevant examples include: educational plans of Wisconsin school boys, statistical aspects of literary style, satisfaction with housing conditions.


Educational and demographic descriptors for 32 large urban school districts, including data on how these characteristics have changed over time.


Categorical and continuous data on the number of nominals produced by native and non-native speakers. Data for experimental and control groups on pre- and post-test measures.


The data sets tend to be small, and many are from the sciences, but there are dozens of them. One interesting example is the movie production and promotion costs for "Dumb Movies," such as Revenge of the Nerds and Police Academy.


Some data sets with broad interest are submerged among many others, including: sex differentials in teacher pay, aptitude and age of first word, nutrition of preschoolers, ailments of university alumni, and the relationship between length of lifeline and length of life.


Data on the eating habits of 37 children, including age, sex, means (and standard deviation) for total amount of calories eaten and minutes spent at table.


Data for 300 participants in the National Longitudinal Study on reading, math, gender, race, college status, SES, High school program, High school grades, creativity, stress avoidance, etc.


Raw data for 1,031 school districts on enrollment, per pupil expenditures, and salaries for superintendents, central office administrators, principals, teachers, staff and support personnel.


An introductory textbook that melds together Tukey's exploratory data analysis and the more traditional confirmatory approaches. Many interesting data sets, including frequency of teacher criticism by student IQ, sex differences in reactions to hostile treatment by an experimenter, experimenter artifacts in social psychology research, characteristics of social networks.


Age, IQ scores and activity ratings for 16 boys and 16 girls. Two activity ratings are available for each child.


Many multiway categorical datasets on the crimes and police disposition of inner-city youth, by race, gender, type of crime, age, etc.


Raw data for four studies of relevance to the social sciences: Creativity and achievement, memory for words, essay grading practices and effects of programmed instruction.


Individual data on 60 subjects. Data include background characteristics (gender, age, IQ, etc.) and achievement (comprehension, reading, word knowledge, spelling, language,
study skills) for school-age children. Suitable for multiple regression, multivariate analysis, etc.


Several interesting data sets including: relationship between status, authoritarianism and conformity, methods to enhance recall of words, causes of the 1907 Romanian Peasant rebellion.


Aggregate cigarette smoking and cancer death rates, by type of cancer and state.


Sociodemographic characteristics for 353 Massachusetts towns and cities, including data on age, race, sex, income, labor force participation, voter registration, police, fire, crime, taxation, libraries and schools. The company publishes similar books for other states; write to them at Box 356, Wellesley Hills, MA 02181.


Several data sets used in employee discrimination cases against organizations including the University of Texas at Dallas, the city of Wichita Falls, and the Federal Reserve Bank.


State-level data on percentage of children in each category of special education and sociodemographic composition of the states. Washington, DC is a high-leverage outlier for the relationship between percent of students classified as educably mentally retarded and percent of population that is black.


Raw data for 61 maladjusted children on two IQ tests, as well as information on their sex, age, parentage (both Foreign, both American, Mixed) and maladjustment type (aggressive or asocial). Authors explore relationship between IQ and all these predictors.


Several multivariate data sets from a variety of disciplines including engineering, manufacturing, biology and mining. The volume includes several well-known data sets such as Fisher’s iris data (1936) and Rothkopf’s Morse-code confusion data (1957) which
have utility for the teaching of principal components analysis, factor analysis, multidimensional scaling and cluster analysis.


Citation frequencies and dates of birth for 187 prominent educational researchers.


Longitudinal data on rats -- some brighter than others -- trained to discriminate 18 cm and 12 cm circles.


Several categorical data sets of wide interest: Suicides by day of the week, homicides by month, stressful events, etc.


Individual data from an experiment on the effects of time lapsed between pre-tests and post-tests for 167 students in grades 4, 5, 6, and 7.


Dozens of interesting data sets on topics such as the unemployment of graduates from 44 British universities, the drinking habits of 120 mothers and their adolescent children, average salaries and overall performance of 26 baseball teams, and marriage rates in the 50 states (Nevada is a very high outside value).


Raw data from 8 studies in psychology and psychiatry, on topics as diverse as headaches, smoking and Alzheimer’s disease.


Many unique and interesting data sets on the link between physical and psychological characteristics. Do blondes have more fun? Do lunatics eyebrows join together in the middle? Are manic depressives thin or fat?


Assorted education data by state and school district, including teacher salaries, high school graduation rates, functional illiteracy rates, and presence of computers in schools.

Aggregate data on school achievement and economic development for 18 nations. Includes measures of preschool influence, GNP, percent enrollment in primary and secondary school.


Ages and IQ scores of siblings in 44 families.


IQ scores over a six year period for 343 children.


Data on two parents and one child from each of 1296 families. Variables include information on height, weight, and psychological tests.


A host of data sets of different sizes on many different topics including: blood pressure and obesity of Mexican-Americans, baseball data, 1970/1 draft lottery, promotion rates among male and female pharmacists, leisure time companions of black women, ranking of rum brands by different nationalities, preference for Charlie’s Angel’s actors, longevity and environment, color of canned tuna, etc.


Productivity, reputation and size of psychology departments at 75 universities.


Body length of crickets as a function of geographical location and weather throughout the USA.


Original data from four studies of IQs of identical twins reared apart: Newman, Freeman & Holzinger (1937), Shields (1962), Juel-Neilsen (1965) and Burt (1955).


Twenty-five children were asked to build towers on each of two occasions. Each time they were given: (a) a set of cubes; and (b) a set of cylinders. Raw data are given on the number of blocks of each type used each time, and how many minutes it took to construct the tower.


Large four-way contingency table derived from a randomized experiment in which several brands of detergent were compared under a variety of user-related conditions.


"Quality" rankings and characteristics of university departments in the social sciences, by discipline. Data include number of faculty, number of students, productivity of faculty, number of grants awarded, follow-up placement of doctoral students.


Data for 38 infants on crying in early infancy and later measures of IQ.


Popularity scores for 17 children: percent each child was named first, percent each child name last, effects of ordering and the effects of sex.


Citation frequency data for 51 schools of education. Measures include mean citation rate, percentage of faculty with 10-100 citations and percentage of faculty with no citations. Kroc correlates these measures with five rankings of schools of education; these analyses could be recreated by abstracting data from the five sets of rankings cited in his bibliography.

Three measures of reading instruction for a sample of 53 learning disabled students, by curricular approach and school.


Life expectancy and per capita income for 105 nations divided into five wealth classifications (industrialized, petroleum exporting, higher, middle and lower).


Multiway contingency table data on the qualifications (Masters vs. doctorate) of art educators, by year and gender.


Individual data for 80 children on the sizes of 10 bones, measured by both x-rays and anthropometry at each of 3, 4, or 5 occasions, by sex. The authors construct lots of individual growth curves.


Lung cancer mortality by degree of urbanization and gender, in Louisiana.


Gesell adaptive scores and age at first word (in months) for 21 children with cyanotic heart disease. The data set contains some interesting outliers and high leverage cases.


"Dewey beats Truman" is one of the classic headlines of the twentieth century. The data contained in this book provide an opportunity to examine the pre-election polls to ask why this headline seemed plausible. Data are provided by the respondents' social class, as well as the month in which the pre-election poll was conducted. Also see analyses in Baker, S. G., & Laird, N. M. (1988), Regression analysis for categorical variables with outcome subject to nonignorable nonresponse, *Journal of the American Statistical Association, 83*, 62-69.

Raw data for 13 data sets across several disciplines. Relevant examples include a subset of 20 from the Coleman Report, educational expenditures for Massachusetts school districts, municipal bond data for 20 US cities.


Annual reports issued by the Department of Education providing descriptive information on education, often over time, sometimes by state, occasionally by school district. The data on university endowments can be used in conjunction with other university level data, such as that given in Barron’s (1987).


State-by-state listing of per-pupil expenditures (in 1955-56 and 1964-65) and percent of school revenues from state and local sources. The entire journal series contains a wealth of information on class size, salaries, public expenditures for education, etc.

National Education Association (1926). The ability of the states to support education. *Research Bulletin of the National Education Association, 4*.

Several chapters of state-level data documenting wealth, per capita income, educational expenditure, population size, number of adults per child, and so on.


Data for 61 school districts on educational policies and practices, as well as selected education and economic descriptors.


Several data sets presenting time-to-failure of products, useful for examining industrial product reliability.


Number of first degree graduates with and without distinction, by school and faculty within school. Suitable for log-linear modeling.


Selected results from a study of opening lines used in singles bars in the St. Louis area. Two-way contingency table describing the relationship between type of opening line (compliments, propositions, etc.) and time of evening.


David Phillips has made a cottage industry of looking at what many might term coincidences -- birthdays and deathdays and copycat suicides after popularized accounts in the media. These are but a handful of articles, each listing the detailed raw data on deaths following these events that led him to his conclusions.


A large variety of categorical data sets including: fingerprints, family size, work conditions and work quality, behavioral problems and birth order, high school rank by gender and socioeconomic status.


Number of meals served, breads baked and ale brewed at the de Bryene household from October 1412-September 1413, by month. That's right, the fifteenth century.


Thirty data sets of small to moderate sizes, on topics ranging from education to cartoons. The educational data sets include information on school strikes and freshman SAT verbal and math scores.

Many multiway contingency tables dealing with students’ liking of school in Norway, broken down by type of school, SES, etc.


Categorical data on the language background and language proficiency of native and non-native speakers and how this influences their choice of writing device.


Rankings for 17 first ladies from Florence Harding through Nancy Reagan on 10 dimensions ranging from integrity, leadership and accomplishments.


Raw data for 100 children who were adopted at birth. Measures include: natural mother’s IQ and education level, foster mother’s IQ and education level, foster father’s occupation and child’s IQ on each of 5 occasions, from infancy through pre-adolescence.


Many artificial data sets as well as approximately 20 interesting small to moderately sized real data sets, including: pre/post data on the influence of Sesame Street, risk of reading problems among kindergartners, behavior reversal, programmed music instruction of elementary school children, IQ testing, etc.


Data for 28 boys given including: 2 IQ scores (Terman and Stanford-Binet), 2 socioeconomic status measures (parents education and father’s occupation), somatypes (endomorph, mesomorph, ectomorph), and drawing type.


Three-way contingency table on the relationship between race of victim, race of defendant and use of the death penalty, showing that the death penalty is not uniformly applied.


Aggregate data on the size of the national assemblies in more than 100 countries, with data on population, literacy, working age, etc.

Dozens of exercises for generating raw data from students in the class, on variables such as the distance between eyebrows, and the ability to judge the lengths of lines and distances between points.


Raw data for a handful of data sets gathered in educational settings, including: effects of delay in oral practice on second language learning (pp. 228-229), relationship between recall and sentence structure (p. 233), predictors of student performance on the Peabody Picture Vocabulary Test (p. 281).


Several real data sets on topics of political and social relevance, thoroughly analyzed using fairly elementary statistical methods. Topics include the relationship between bureaucracy size and population, the number of radio receiver licenses issued and the proportion of mental defectives in Britain (per 100,000 population), and the relationship between dietary fat consumption and death rates in selected countries around the world.


Sociodemographic, education, health and economic indicators for 130 countries.


The authors analyze data for the 50 states on the relationship between failure on the selective service exam administered during 1969-1970 and contextual and education descriptors of the states. The selection bias inherent in analyses of state level SAT scores is present here, but the data set is interesting.


Raw data for several interesting data sets including Cyril Burt’s IQ data, Allison and Cicchetti’s brain weight and body weight data, three time points for 26 boys and 32 girls who participated in the Berkeley Guidance study (anthropometric information only, however.)

Many characteristics in dozens of societies around the world, including age at weaning, toilet training, fear of ghosts, rituals, etc.


Age of person at death (in years) and the length of the person's lifeline (in centimeters) for 50 individuals. Not surprisingly, the test of $H_0: r=0$ cannot be rejected.


Chronological, mental age, social distance scores, and gender for twelve children tested three times during adolescence. Also indices for racial tolerance over three occasions of measurement for 39 racial groups.


Individual data for 100 students on these two IQ tests, with information on student sex, age and native language.

---

*Judith D. Singer and John B. Willett* are Associate Professors at the Harvard University Graduate School of Education specializing in quantitative methods. They have written and presented dozens of talks, workshops, and papers on the application of statistical methods in education and the social sciences. Together with other colleagues, they have written two books, *By Design? and Who Will Teach?*, both published by Harvard University Press. Their current research interests center on applications of survival analysis in the social sciences. They have recently received the Raymond B. Cattell Award and the Palmer O. Johnson Award from the American Educational Research Association and an NSF Visiting Fellowship from the ASA.