MATH 3870. Inventing Statistics

Proposed Syllabus

Instructor. Joseph Kung, GAB 471c. Three credit hours. Office hours: four during the week, and other times by appointment.

Course objective.

This course will critically examine the central concepts of statistics using original books, papers, and letters, the historical situation, and the biographies of some of the creators of statistics. Statistics is a human construction, contingent on a consensus which is determined by historical, social, and personal factors. We will show this by retracing the origins and rise of statistics through original documents, the historical and social context (usually of Victorian England), and the biographies of some of the creators of statistics. Although this is a mathematics course in spirit, it is about how human beings do statistics and are affected by statistics.

Grading and assessment.

There will be three components. The first consists of reading a part of a book, or a paper, taking a short quiz each week about the assigned reading, and participation in class discussions (42%, 3% per week). The second consists of a mid-term test and a final (15% each), and the last is a substantial essay and a presentation (to the general student body or internet public) based on the essay (28%).

Code of conduct. Students are expected to behave in accordance with the UNT student code of conduct. There will be zero-tolerance of violations, especially on tests and finals.

The estimate cost of taking this cost. There is no required textbook. There will be a package of readings, costing about $10 to $20.

A tentative syllabus.

The course consists of four parts, each taking about three or four weeks, with a week for the midterm and in-class presentations.

Weeks 1 to 4. Setting the stage. This has two aims, to set the stage historically and to establish a common mathematical background.

Readings: Extracts from Francis Galton, ``The Art of Travel; Or, Shifts and Contrivances Available in Wild Countries,'' Charles Dickens, ``Hard Times,'' a modern essay on John Snow, the Broad Street pump and cholera (from Thomas Korner, ``The pleasures of counting''), Karl Pearson's letter about the lack of statistical data about the oldest profession, Playfair's graph of the how British national debt incurred during the Napoleonic wars was paid-off.

Lectures on (a) Galton's study of heredity and their relation to the debate on evolution, the debate between phenomena (or phenotype) and genetics, how Galton opposed Mendelian theory as a pernicious doctrine; (b) the great geographical expeditions (to determine the major axis of earth, to find the source of the Nile, and so on), Humboldt's network to measure the temperature of the world's oceans, (c) the rise of a bureaucracy to collect data about the state (or statistics), and their claim to a social calculus which would measure and improve the common welfare scientifically.

Leveling lectures on probability and statistics. These will review concepts the students should already know, like conditional probability, expectation, inverse probability. We will use examples like the Monty Hall car-or-goat problem, Simpson's paradox, Bayes' formula for inverse probability and the law of insufficient reason, Ulam's calculation of the statistically correct fine or punishment for littering (and its accidental implementation in Singapore).

Discussion. Does the use of statistics help or hinder science? Statistical methods bring quick results, but do they discourage analyses of the underlying theory? Does having more data necessarily improve the common good? Is data necessarily objective? Does quantification necessarily turn the common good into a set of numbers? Were the scientific expeditions of the Victoria era thinly disguised exercises in imperialism? To what extent is the concept of an ``average man'' dehumanizing? What does the dialectic between statistical uncertainly and statistical laws say about free will?

Weeks 5 to 7. From mathematical abstraction to statistical law; or how the normal distribution become normal.

Readings. Extracts from papers of de Moivre, Gauss, Laplace and Le Gendre. Francis Galton's ``The measure of fidget'' (Nature, 32,174-175), extract from the work of Francis Galton and Karl Pearson about the normal distribution.

Lectures on (a) de Moivre's theorem, that the binomial distribution is approximated by the normal distribution, de Moivre's formula for the error; (b) further work by Laplace, Le Gendre, and Gauss on the theory of errors. Rediscovery of this theory by Galton, Pearson, Quetelet, and others; (c) controversies about whether other distributions fit reality better and how hypotheses should be tested; (d) the mathematical justification by the central limit theorem giving a precise mathematical description of the technical conditions under which the normal distribution is applicable.

Practical Experiments. What does 5% or 1% probability really mean? Try it, by tossing coins and observing runs. Observing a ``quincunx'', a physical -- nowadays computer -- simulation of the normal

distribution on an internet site like Mathematica. [How does one involve an audience in this effectively?]

Discussion. To what extent is the protocol of hypothesis testing [the null hypothesis standing for innocence] a cultural artifact of the Anglo-Saxon legal system? If statistics were invented in France, would it have been data mining because of the Napoleonic code? Is the ``wisdom of crowds'' real? Propose other ways of making decision using statistical data.

Week 8. Pause for reflection and discussion; midterm test.

Weeks 9 to 11.

Readings. Extract from Gregory Chaitin, ``Exploring Randomness''; newspaper or magazine articles [about census undercount, the breast cancer screening controversy, assessment criteria for race-to-the top, the claims of SETE at UNT, or a current issue], a description of a parking meter fraud case in New York, Extract from Francis Galton, ``Regression towards mediocrity in hereditary stature,'' Journal of the Anthropological Institute, 15 (1895) 246-63.

Lectures on (a) regression and confounding; deconstructing ``correlation is not causation''; (b) whether Mendel (or his assistants) manipulated his data; (c) how to detect fraud using statistical tests or Benford's law; (c) randomness and complexity, are they related? (d) Order in randomness: Polya's theorem about random walks; (e) objective and subjective randomness: the fatal flaw in the German Enigma code during the Second World War, a statistical comparison of randomness in a Renoir landscape and a Jackson Pollock painting.

Discussion. Is randomness an objective or subjective attribute? What is probability, anyway? Does it matter whether Mendel threw away a peapod or two? Does science progress by induction or insight? Is there a danger that massive data analyses impede creative thought? If regression was invented to justify an oppressive class system, does it change the way we should use it? When a situation is non-linear, is there a justification for using regression just because it is the simplest and most accepted method?

Weeks 12 to 14. The early twentieth century.

Readings. Extracts from David Grier, ``When computers were human'', Hugo Steinhaus' ``Mathematical Snapshots'' , Constance Reid ``Neyman, from life''.

Lectures on (a) indirect inference in the Second World War: how Steinhaus, in occupied Poland, predicted the date the Second World War would end by reading only German newspapers, why manufacturer plates were almost always missing from shot-down planes, probabilistic arguments in operations research; (b) Confidence intervals versus the mean; (c) the rise of computers and computer methods like resampling; (d) ``quants'' in finance; do they cause and profit from stock market crashes?.

Discussion. Is science morally neutral, just a technique one can apply for good or ill? If science progresses most rapidly in wartime, why not give war a chance? What (if anything) was actually at stake

in the confidence interval controversy? Make your own distribution by resampling. In an interdependent connected world, is the normal distribution still valid? Is traditional risk analysis in finance valid anymore?

Week 15. Review and reflection. Presentations.

Final. There will be a short final.

Textbook? This course is a mixture of history, mathematics, and philosophy of mathematics. No one book will cover everything. There will be a package of readings, and many of the older books are freely available on the internet. The following is a list of relevant books:

Mathematics and history and philosophy of mathematics:

Michael Bulmer, Francis Galton. Pioneer of heredity and biometry, Johns Hopkins University Press, Baltimore,2003.

Gregory Chaitin, Exploring randomness, Springer, New York, 2001

David A. Grier, When computers were human, Princeton University Press, Princeton, 2005.

Thomas W. Korner, The pleasures of counting, Cambridge University Press, Cambridge, 1996.

William Playfair, The commercial and political atlas and statistical breviary, Cambridge University Press, Cambridge, 2005 (reprint of the 1801 edition).

Theodore M. Porter, The rise of statistical thinking 1820-1900, Princeton University Press, Princeton, 1988.

Theodore M. Porter, Karl Pearson. The scientific life in a statistical age, Princeton University Press, Princeton, 2005

Theodore M. Porter, Trust in numbers, Princeton University Press, Princeton, 1996.

Constance Reid, Neyman, from life, Springer, New York, 1982.

Hugo Steinhaus, Mathematical Snapshots, out of print [extracts will be provided].

Literature. These are books which you will enjoy reading. They will give you a sense of the Victorian age. Most are digitalized in some form and available freely on the internet.

Charles Dickens, Hard times.

George Eliot, Daniel Deronda.

Francis Galton, The art of travel; Or, shifts and contrivances available in wild countries.

Mrs. Gaskell, North and South.