A Brief History of Statistics (Selected Topics)

ALPHA Seminar

August 29, 2017

Origin of the word "Statistics"

- Derived from Latin *statisticum collegium* ("council of state")
- Italian word *statista* ("statesman" or "politician")
- German book *Statistik*, published in 1749, described the analysis of demographic and economic data about the state (*political arithmetic* in English)
- Was broadened in 1800s to include the collection, summary, and analysis of data of any type; also was conjoined with probability for the purpose of statistical inference.
- Recently, some believe this word is holding back our field

Early examples of statistical inference

- 5th century B.C. — Athenians estimated the height of ladders necessary to scale the walls of Platea by having multiple soldiers count the bricks, then multiplying the most frequent count (the mode) by the height of a brick.

- Al-Kindi (801-873 A.D.) wrote "Manuscript on Deciphering Cryptographic Messages" which showed how to use frequency analysis to decipher encrypted messages.

- John Graunt in *Natural and Political Observations Made Upon the Bills of Mortality* estimated the population of London in 1662 as $384,000$ using records on the number of funerals per year ($13,000$), the death rate (3 people per 11 families per year), and average family size (8).

"Correct values" from multiple measurements

- Mode (Athenians)
- Median
  - Originated in Edward Wright's 1599 book, *Certaine Errors in Navigation*, concerning the determination of a location with a compass.
  - Further advocated by Ruder Boskovic in his 1755 book on the shape of the earth, in which he showed that the median minimizes the sum of absolute deviations.
  - The term "median" was coined by Galton in 1881.

- Mean
  - The mean of two numbers was a well-known concept to the ancient Greeks
  - Extended to more than two numbers in the late 1500s in order to estimate the locations of celestial bodies.
  - Shown to minimize the sum of squared deviations by Thomas Simpson in 1755

Least squares

- Introduced by Adrien Legendre in 1805
- By 1825 it was a standard tool in astronomy and geodesy
- The dominant theme of mathematical statistics (called "the combination of observations") in the 1800s
- The topic of the first statistics course at U of Iowa (1895 or earlier)

Probability

- By 1700 the mathematics of many games of chance was well understood; major contributors were Fermat, Pascal, Huygens, Leibniz.

- But this early work did not consider the inferential problem: How, from one or several outcomes of the game, could one learn about the properties of the game and how could one quantify the uncertainty of the inferred knowledge of these properties?

- It was Jacob Bernoulli (1654-1705) who began down this road. His weak law of large numbers was published in 1713 (posthumously), and he put a lower bound on the probability that X/N (the proportion of successes in N repeated trials) was within a specified distance from the true probability of success.

- De Moivre (1667-1754) refined Bernoulli's bound and stumbled upon a discrete version of the normal curve as an approximation to binomial probabilities

Error distributions

- In 1755 Thomas Simpson introduced the notion of error curves, including what are now called probability density functions. His pdf of choice was triangular.

- Others proposed error curves of different shapes: semicircular (Lambert 1765), exponential (Laplace 1774), parabolic (Lagrange 1776), normal (Laplace 1778, although it was not called the normal distribution until 1873)

Synthesis of least squares and probability

- Probability needed to assess the accuracy of least squares
- In 1809, Gauss showed that the (sample) mean maximizes the likelihood of the errors only when the error curve is of the form

$$\phi(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2}$$

  for some positive constant h.
- In early 1810, Laplace came out with the Central Limit Theorem, which showed that the aggregate of a large number of errors will be approximately normally distributed.

- Later in 1810, Laplace read Gauss's work and made the revolutionary connection between the central limit theorem and least squares estimation: the error curve that had the most compelling rationale was also the one that led to the simplest statistical analysis!

- The Gauss-Laplace synthesis is regarded as one of the major milestones in the history of science, and it became a staple in astronomy and other physical sciences by the mid-1800s.

- It took until about 1900 before it had fully diffused into other scientific disciplines

Some important figures after Laplace and Gauss

- Francis Galton (1822-1911) — fitted normal curves to data, discovered reversion (later called regression) to the mean, and correlation
- Francis Edgeworth (1845-1926) — the first to compare means of two populations, using a precursor to the two-sample t test
- Karl Pearson (1857-1936) — Introduced moments, Pearson's correlation coefficient, P-value, Pearson's chi-square test, principal component analysis, among many other things
- Ronald A. Fisher (1890-1962) — Introduced randomization test, named and promoted the analysis of variance and the design of experiments

Earliest Statistics departments

- 1911: University College, London (Karl Pearson's department)
- 1918: Johns Hopkins Department of Biometry and Vital Statistics
- 1931: University of Pennsylvania Department of Economic and Social Statistics
- 1933: Iowa State Statistical Laboratory
- 1935: George Washington Statistics Department (first in a College of Liberal Arts and/or Sciences)

Largest departments in U.S.

- North Carolina State University (41 faculty, some joint with other departments)
- Iowa State University (40 faculty, some joint with other departments)
- Texas A&M University (27 faculty)

These departments generally have 4-8 “lecturers,” “instructors,” or “teaching professors” as well.

By comparison, U of Iowa has 15 faculty (3-4 are actuarial science rather than statistics faculty) and 6 lecturers.

Rankings of Statistics departments in U.S.

- U.S. News and World Report (Iowa is #34, tied with Yale and University of Illinois
  http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/statistics-rankings/page+2)

- National Research Council (Iowa is #31,
  http://www.stat.tamu.edu/∼jnewton/nrc_rankings/nrc41.html#area34)

- World Ranking Guide