The Origins of Statistical Computing

1 September 2006 622 views No Comment

David Alan Grier

Interest in statistical computing began not with the invention of the personal computer in the 1980s or even with the rise of the large mainframe computer during the 1960s. Statistical computing became a popular field for study during the 1920s and 1930s, as universities and research labs began to acquire the early IBM mechanical punched card tabulators. They used these machines not only for tabulating and computing summary statistics but also for fitting more complicated statistical models such as analyses of variance and linear regressions.

These labs proved to be important places for advancing statistical methodology. They helped make Galton's and Pearson's ideas on correlation practical tools that could be used for scientific research. They encouraged researchers to think in terms of large problems with extensive datasets. Without them, modern statistical methodology could easily have languished as an interesting theory, useful for small problems but otherwise impracticable.

In addition to advancing statistical methodologies, these labs helped to advance scientific computing in general. Many of these labs offered their services to physicists and astronomers as well as to biologists and social scientists. Some created tables of higher mathematical functions. Others solved complicated differential equations. A few of these labs, most notably those at Iowa State University and Columbia University, became test beds for early computer scientists, who experimented with new ideas for computing machines and for numerical algorithms.

Most of these labs were small ad hoc organizations. Many were nothing more than a creative professor who had arranged to use the tabulating machines of the university business office during a second or third shift. The largest of these labs were substantial institutions funded by donations from private individuals or small foundations. In the 1920s, such gifts were almost the only funds a researcher could expect to find. There was no National Science Foundation, no National Institutes of Health. There were no instrumentation grants for the mathematical sciences. The scientific infrastructure developed by Vanevar Bush during and after World War II simply did not exist. The only source of government money for scientific research was the Department of Agriculture, an organization that proved to be very supportive of empirical research and that helped to establish the largest and most sophisticated of the statistical laboratories, the Statistics Lab at Iowa State University.

The names associated with these early labs are familiar to very few—James Glover, H. T. Davis, A. E. Brandt, Howard Tolley. They published little and made only marginal contributions to the theory of statistics or the development of computers. Yet these researchers held a deep faith that the combination of computing technology and mathematical statistics would radically change science. These tools remain linked today, but during the 1920s and 1930s the combination not only helped establish the field of statistics on the American continent, it also promoted computing as an important tool for scientific research.

The First Statistical Labs: Before the Tabulator

Most of the earliest statistical laboratories were founded to study economic phenomena, though they quickly began to apply their tools and techniques to problems in the social, biological, and behavioral sciences. Economic applications paid the bills, leased the equipment, and provided the salaries for laboratory workers. One of the first of these laboratories was founded at the University of Michigan by James Glover, a professor of mathematics. Glover was a pioneer actuary and student of financial risk. He began teaching advanced statistics courses in 1904, though they included little of the mathematical statistics that was being developed in England by Karl Pearson and R. A. Fisher. The University of Michigan was one of the founts of -statistical knowledge in the United States. The University trained many of the early pioneers of American statistics, including George Snedecor, founder of the Iowa State University laboratory.

Glover had a small lab operating by about 1910. It was staffed by his students and followed the model of the computing labs found in observatories and astronomy departments. At these labs, junior personnel acted as computing assistants to the astronomers. These computing assistants would transform observations into stellar coordinates and would use the least squares methods of Gauss and Laplace to estimate the orbits of comets and planets. In Glover's lab, they reduced data to summary statistics, created actuarial tables, cross-tabulated data, and made projections from simple statistical models.

Glover rarely had more than two computing assistants. Most were his students and many were women. Even in the early 1900s, the University of Michigan was a coeducational institution, and Glover felt that statistical study was especially appropriate for young women. At times, nearly half of his students were women. A few of these women had surprisingly long and prosperous careers in industry, in the U.S. Census, or in actuarial firms. For the most part, the largest cohort of these women became human computers, the clerical workers who did mathematical calculations before the advent of electronic computers.

Though Glover ultimately had a distinguished career as an actuary and as the head of Teachers Insurance and Annuity Association, he is a marginal figure in the history of statistics. Glover's lab, however, seems to have inspired one of the founders of the field, Henry Reitz, to organize a computing lab at the University of Illinois, where he worked until 1918. Like Glover, Reitz was interested in actuarial work and did projects for insurance companies in Chicago and Springfield. Glover also taught the statistical courses at the University of Michigan when George Snedecor attended the university between 1910 and 1913.

The First Card Tabulators

The earliest statistical laboratories used mechanical adding machines or calculators, such as those made by Monroe, Marchant, or Sunstrand. They began using punched card tabulators in the early 1920s. These devices had been invented by Herman Hollerith for the 1890 U.S. Census. Hollerith formed the Hollerith Tabulating Machine Company to manufacture and market these devices. This company was merged with two other firms in 1911 to form the Computing Tabulating and Recording Company or C.T.R. In 1924, C.T.R. was renamed International Business Machines.

Although least squares was an important application for the early statistical labs, Tolley and the others at the Bureau of Agriculture were initially more interested in the statistical methods of Frederick Winslow Taylor than they were in the methods of Galton, Pearson, and Fisher. Taylor was an engineer from Philadelphia, whose writings on scientific management were highly influential in the first decades of this century. He proposed a means of studying the methods of workers and developed some crude statistical techniques for gathering and analyzing data. These techniques were filled with heuristic and ad hoc methods and were often criticized by Taylor's detractors. They were relatively effective at the time, however, and were studied by many managers who wished to improve production at their plant or in their office.

Tolley was interested in applying Taylor's ideas to fruit markets in New York and cold-storage warehouses along rail lines. Yet, he clearly understood the limits of Taylor's methods and knew that these statistical methods were unable to help him in situations with large amounts of variability, such as estimating crop production and weather damage. His training at the Coastal Survey helped him to understand the relationship between correlation analysis and least squares. During his early years in the Department of Agriculture, he worked to promote least squares analysis. Although researchers were generally interested in this method,

they occasionally found it difficult to apply because the punched card tabulators of the early 1920s were unable to multiply. Tolley apparently found a practical method to compute correlations that required both a punched-card tabulator and a desk-top calculator.

The lab in the Department of Agriculture inspired two Iowans, George Snedecor and Henry A. Wallace, to experiment with punched-card statistical computations. Henry Wallace eventually rose to prominence as the Vice President of the United States, but during the 1920s, he was the publisher of his family's farm journal, Wallaces' Practical Farmer. He was also a self-taught statistician and was interested in the interplay of biology and economics in farm management. During the 1910s, he learned the methods of correlation studies and least squares regression by reading Yule's book, An Introduction to the Theory of Statistics (London: Griffin, 1911). Finding in that book no easy method for solving the normal equations for regression, Wallace devised his own, using an idea that Gauss had applied to an astronomical problem.

In 1923, Henry A. Wallace learned of the new statistics lab at the Department of Agriculture while he was visiting his father, Harry Wallace, who was then the Secretary of Agriculture. Intrigued with the machines, he borrowed a tabulator at a Des Moines insurance firm and taught himself how to use the device to calculate correlations.. He would punch data cards and would then take them to the offices of the insurance company for tabulating. During the first years of the 1920s, he published ever more sophisticated statistical studies in the pages of Wallaces' Farmer, studies that must have baffled many of his loyal readers, who tended to be modestly educated farmers. The last, published in January 1923, was a detailed study of land values in the state.

The study of Iowa land values marked the maturity of Wallace's statistical ability. By the time he published it, Wallace had become a friend of George Snedecor, who taught the statistics courses at Wallace's alma mater, then named Iowa State College. Impressed with Wallace's knowledge of least squares, Snedecor invited him to teach an advanced course on those methods to college faculty. This class, which met for 10 consecutive Saturdays over the fall and winter of 1924, ended with a demonstration of punched-card calculation. After the class, Snedecor helped Wallace prepare a manuscript on his algorithm for solving normal equations. They jointly published the manuscript in 1925 with the title Correlation and Machine Calculation.

The title of Wallace's and Snedecor's pamphlet tends to mislead modern readers. For the most part, the machines to which it refers are desk calculators, not tabulating machinery. Part of Wallace's methods were easily adapted to tabulating machines. By computing sums of squares and sums of cross-products, a mechanical tabulator could quickly produce a set of normal equations. The same tabulator, however, could not be easily used to solve these equations. It was extremely awkward, if not impossible, to use a 1920s vintage tabulator to solve matrix arithmetic problems. Such problems were solved by human computers who used desk calculators.

Fig. 2

Inspired by Wallace, Snedecor devoted much effort to acquiring tabulating machines for his university. He was able to secure them in the Fall of 1927 and established a statistical computing lab within the Department of Mathematics. (See Fig. 2) This first lab seems to have been a cooperative effort by several college

departments and may have been partly supported by local IBM officials, who were interested in placing their equipment at universities. IBM helped many schools establish computing labs at that time. The first was at Cornell, which leased tabulating machines to form a lab in 1926. Next came Iowa State College, Columbia University, and the University of Michigan, who acquired these machines in 1927. Shortly thereafter came the University of Texas, Harvard University, Stanford University, and the University of Tennessee.

The Statistics Lab At Iowa State College

During 1927, Snedecor exhibited the same kind of exuberance that we now attribute to someone who has just acquired a fancy new personal computer. He used the tabulating equipment for every possible application that he could find and proudly presented a detailed report to his chair. He tabulated basic agricultural statistics, tracked the results of agricultural county fairs, and started a punched-card livestock breed book. A colleague used the tabulator to evaluate higher mathematical functions. Another interpolated a function with polynomials.

After a year of operation, after his ardor cooled somewhat, Snedecor turned the lab equipment over to the management of A. E. Brandt. Brandt was a student of Snedecor's and had been a professor of farm mechanics at Oregon State University. He enjoyed the subtleties of the tabulators and liked to find new ways of doing calculations. From the Economics Department, he recruited human computers to help operate the machines and to solve normal equations for regression problems. One of these clerks, Mary Clem, would remain with the Statistics Lab for the next 50 years and would ultimately be identified as the lead human computer of the group.

Fig. 3

The computing facility was an important part of a lab that was quickly building statistical expertise. Through the Department of Agriculture, it acquired funds to host summer institutes in statistical theory. The first of these was held in 1927 with British statistician R. A. Fisher. Fisher met with about 50 researchers who were eager to learn his methods. (See Fig. 3.) One of these researchers was Henry A. Wallace, who would shortly thereafter leave Iowa and become Secretary of Agriculture, following in his father's footsteps. By then, Wallace had become fascinated with the problems of weather prediction and had begun a very large study in which he attempted correlating heat, humidity, and wind direction with the position of the planets. He discussed the study at length with Fisher, who was little interested in such poorly grounded research. The work eventually became an embarrassment to Wallace when his political enemies branded it as "weather astrology."

Even though Wallace may not have made the best use of his contact with Fisher, he did become a champion of statistical studies as means of planning programs to address social and economic ills. As Secretary of Agriculture, he devised and prepared the Agricultural Adjustment Act, a radical proposal to support farm prices and to alleviate the effect of the Great Depression on American agriculture. This program was the first major piece of legislation in Franklin Roosevelt's New Deal, written and implemented within his first 100 days in office, and it became a model for subsequent programs. It required the Department of Agriculture to undertake large statistical studies of major farm products, including cotton, corn, tobacco, and pork.

The Agricultural Adjustment Act proved to be a boon for land grant colleges and state experimental farms because they were asked to do much of the local statistical work. It was especially helpful to the Iowa State College Statistical Laboratory. The lab, now independent of the Mathematics Department, acquired several

government contracts in the early days of the New Deal. As the demand for statistical work increased, the lab undertook increasingly larger and more important jobs. By 1936, it was negotiating with the Department of Agriculture to undertake major research projects, including a large master sample of the nation's farms. These projects increased the size of the lab. Over a short period of five years, its budget grew by a factor of 16.

John Atanasoff and an Early Electronic Computer

The rapid expansion of the lab allowed one Iowa State College faculty member to undertake some experiments in computation, experiments that would provide the basis for the modern electronic computer. That professor, John Atanasoff, held appointments in both the mathematics and physics departments. During the early 1930s, Atanasoff had been studying approximate solutions to differential equations. The last step of his approach required him to solve a large system of linear equations. Knowing that the Statistics Lab routinely solved such problems when it computed regression models, Atanasoff began to consider how such equations might be solved by using the lab's punched-card equipment.

Between 1934 and 1937, he and A.E. Brandt experimented with lab equipment. In the first experiment, Brandt and Atanasoff modified an IBM tabulator to analyze an atomic spectrum. To do this, they constructed a special circuit that allowed a tabulator to compute all possible differences from a list of numbers. Once they had completed this experiment, they began to work directly on the problem of solving linear systems. Atanasoff sketched a design for the necessary circuits but never completed the task. Before they made much progress on the project, Brandt left the lab to join the Bureau of Soil Conservation. With Brandt gone, Atanasoff probably lost his access to the punched card machines, for he undertook no further experiments. He ultimately decided that his design was too difficult to complete and abandoned it.

Even though Atanasoff lost interest in the punched-card machines, he did not forget about the problem of solving systems of linear equations. In what is a well-known story among computer historians, Atanasoff set off on a long drive across Iowa to think about this problem sometime during the winter of 1937–1938. Several hundred miles later, at a roadside bar in Illinois, he conceived the basic elements for a machine to solve systems of linear equations. The proposed machine had a lot of similarities with modern computers. It was electronic and had a memory unit, a central processor, and binary arithmetic. He built a small prototype of this machine in 1939 and prepared a proposal in 1940 for a full working model, a proposal he used to solicit funds for the machine.

The Iowa State Statistical Lab clearly influenced Atanasoff as he wrote this proposal. Atanasoff saw his machine within the context of a computing lab, much like the Statistics Lab, and that it would solve linear systems "at low cost and for technical and research purposes." Atanasoff then listed nine possible applications for his machine. Of these nine, the first three are statistical.

Fig. 4

Atanasoff built his machine between 1940 and 1942. (See Fig. 4.) With the start of World War II, he abandoned his creation when he left the college to join the Naval Ordnance Laboratory in Washington, D.C. Atanasoff's machine would have remained in obscurity were it not for John Mauchly, one of the inventors of

the ENIAC computer. Mauchly met Atanasoff at a conference in May 1941 and visited Atanasoff's lab in Ames, where he studied Atanasoff's machine. During this visit, Mauchly learned a great deal about electronics and about computing machines. Some of these ideas found their way into the ENIAC design and remained a point of contention between the two computer pioneers for the rest of their careers.

The Statistics Lab of the Cowles Commission

Although the statistical lab at Iowa State University became one of the major centers of statistical research, a similar lab at Indiana University grew to be a center of econometric research. This lab was founded by H. T. Davis and was championed by Colorado financier Alfred Cowles. Davis was a new member of the mathematics faculty when the dean asked him to form a statistical lab in 1927. A small local foundation had agreed to finance such a group so that they might better understand the fluctuations in the economy.

Like Snedecor, Davis was quick to use the lab to explore a diverse range of computational problems, most of which had little to do with the statistical study of the economy. One of the first was an optical interference computation for a physicist colleague. When he realized that his computers had used the wrong values in making the computation, Davis convinced his colleague to perform the experiment a second time, using the values that his computers had mistakenly employed. Although he did statistical calculations for business and economics professors, Davis took little interest in statistical research. Instead, he studied numerical analysis and used the lab to create tables of higher mathematical functions.

Alfred Cowles helped revive Davis's interest in statistics. Cowles, an amateur statistician much like Wallace, was interested in large regression studies of the economy. He envisioned studies that would collect thousands of observations and would fit regression models to 20 or 30 independent variables. In 1931, he approached Davis and asked him for help in undertaking such computations. Davis immediately realized that his small computing lab, staffed by human computers, would be unable to do the necessary computing. He urged Cowles to lease punched -card equipment and helped him to establish a statistical lab near Cowles's offices in Boulder, Colorado. Davis spent the next several summers with Cowles, helping him develop the necessary mathematical techniques. It seems likely that Davis used the methods of Snedecor and Wallace because their pamphlet was one of the few works on the computations needed for regression models.

In 1939, this organization, which was then known as the Cowles Commission for Economic Research, moved to the University of Chicago. It reestablished its computing lab at the university even though the University of Chicago supported one of the oldest organized computing groups in the nation. This group did calculations for the Physics and Astronomy Departments. It was run by Ardis Monk, the wife of a physics professor. She oversaw a dozen or so graduate students. These human computers worked with the traditional tools of the day, slide rules and logarithm tables. Their computations were analogue approximations, not the digital calculations of the Cowles Commission Laboratory. Over the next 14 years, during which time the Cowles Commission moved to Yale University, the lab did an increasingly important series of econometric computations, which supported the work of people like Kenneth Arrow and T. J. Koopmans.

The Last Days

World War II marked the glory days of the statistical lab and the start of its slow decline. The Office of Scientific Research and Development financed dozens of statistical projects and organized computing labs to find concrete numerical answers. At Columbia University, Abraham Wald operated a lab of 20 human computers to develop a theory of sequential testing. University of California statistician Jerzy Neyman worked with a New York computing group to help the Air Corps clear the Normandy beaches of mines in preparation for the D-Day landings. Neyman used a geometric model to estimate the number of mines that would survive being bombed by the Air Corps. The New York group, which had began operation as a W.P.A. project, calculated the actual estimates.

At the war's end, the statistical labs were deeply interested in the new electronic computers. At an International Statistical Institute Conference (I.S.I.) in the early summer of 1946, the Cowles Commission sponsored a session on electronic computers. At the session were representatives of many of the statistical labs, as well as some senior statisticians like Harold Hotelling. The featured speaker was John Mauchly, who had unveiled the ENIAC computer only a few months prior. Mauchly told the attentive group how the

ENIAC could summarize data, compute correlations, and solve linear regressions.

Yet, in some ways, the I.S.I. Conference in 1946 marked the start of the decline of the old statistical labs. As universities and corporations built centralized computing services, the statistical labs faded away. Their old calculators were shelved, and the punched-card equipment returned to IBM as statisticians purchased computer time from a computing center. This trend was reversed in the early 1970s, when inexpensive minicomputers first appeared on the market. The popularity of the personal computer and the widespread availability of statistical software ensured that not only every department would have a computing facility but that every statistician could do more computing in an hour than the old Department of Agriculture statistical lab could have done during the entire year of 1924.

Though modern computers are more powerful than the antiquated technology of punch-card tabulators, the old statistical labs were probably more important to science and technology. In the 1920s, the computing labs helped establish statistics on the American continent. Without them, even a modest study was beyond the ability of an individual statistician. At the same time, statistics labs often had the most powerful computing machines within their larger institution. They showed how organized computing could benefit science and provided a place for the earliest of computer scientists to test their ideas.