How John Venn's attempt to refute historical determinism with probability theory fostered the development of applied statistics

In the mid-nineteenth century, Henry Buckle caused a stir in Britain with his *Introduction to the History of Civilization in England,* which argued for a historical determinism applicable to all of human civilization. Buckle's work was largely based upon conclusions drawn by the Belgian astronomer, turned sociologist, Aldophe Quetelet in his *Sur L'Homme et le Dévelopment de ses Facultés*, published in 1835. This work prompted John Venn to write his treatise, *The Logic of Chance*, in 1866, setting forth the frequentist position on probability and his interpretation of what conclusions can be legitimately drawn from such statistics as those put forth by Quetelet. Venn was at pains to demonstrate that statistical regularity in nature and in matters of human behaviour did not imply determinism. He argued that even if statistics show a predictability in human behaviour in the aggregate, that individual events remain unpredictable and irregular. Venn was defending a belief in free will, but ironically, his work also showed that correlations and predictive models could be developed empirically and independently and put to use without regard to possible implications for human freedom or fatalism, thus clearing a philosophical path for statistical studies of society and nature. This paper will trace this path from error theory in astronomical observations, to proto-sociological studies of data on human behaviour, to theories of history, through the empiricist inductivism of the mathematically trained logican John Venn, to the flourishing ubiquitous rise of statistics in the late nineteenth century.

* * *

This paper begins with statistics and ends with statistics and therefore shows the connection of statistics with statistics, which does not suggest anything remarkable at all. However the statistics at the beginning is really part of a mathematical smoothing technique used on astronomical observations of star positions. The statistics at the end is that which uses correlations between human behaviour in one realm with behaviour in another as a tool to evaluate abilities, predict longevity, and direct mass marketing of consumer goods. The route from one to another goes through the beginnings of sociology, theories of history, philosophical questions of free will, and arguments over the meaning of probability, before once again being applied to the manipulation of data and drawing inferences therefrom.

I begin with the phenomena of astronomical observations. It comes as no surprise to us in the 21st century to hear that astronomers recording the positions of stars in the heavens are not always in perfect agreement. We take it as natural that different observers will record slightly different results and that, with the best of intentions and efforts, such measurements will vary slightly from one another. That realization itself took some time for astronomers to accept, and even longer to make provisions for in their records. But by the 18th century such provisions were made, and a mathematical theory had been devised to cope with a spread of measurements, all attempting to record the same position. This was known, in the time of Pierre Simon de Laplace, as the Theory of Error. The name is significant, because it underscores the working assumption that there is a single, correct position of a star, and the raw data that differ from that correct position are indeed errors.

It is perhaps all the more understandable, then, that when the Belgian Adolphe Quetelet, who later became the Royal Astronomer of Belgium, began to study data collected on human beings, that he would regard differences in the individual data collected as being errors from some "correct" value that inhered in Nature somewhere, somehow. That individual people were different was something that Quetelet could not deny, but the fact that measurements of individual differences tended to cluster around central values in just the same way that astronomical measurements of star positions did, and fit the same distributions as were provided for in Laplace's Error Theory, suggested to Quetelet that there was some model of humanity, like a Platonic form of the ideal human being, that was being expressed by individuals, who deviated from the "correct" measure for unknown reasons. Quetelet

expressed this notion as *l'homme moyen*, the Average Man, in his *Sur L'Homme et le Dévelopment de ses Facultés*, published in 1835. Just as astronomers sought to record the true position of stars in the heavens by taking multiple measurements and using error theory to hone in on the correct values, the "social physicist" as Quetelet called his researchers, would be able to deduce Nature's intended model for humanity by collecting and analyzing data.

Among the data that Quetelet collected were some surprising regularities in the human condition. Quetelet expected to find normal distributions in vital statistics such as height, weight, and girth, as well as mortality statistics. He was rather taken aback at statistical regularities in matters of human behaviour. Among the surprises that he reported were amazing constancies in the number of suicides year after year in major cities, such as Paris and London, where such statistics were kept. Another was the peak age at which people seem to commit murder—25 years of age is what he found. These Quetelet was prepared to ascribe to some inherent organic propensity for destruction that might be further elucidated by phrenology. Quetelet's work did much to launch a "science of society," or, to use the name we have settled upon, sociology.

Another study of human behaviour that was ripe for being conquered by the march of science was history. Quetelet's findings could also be used to support a theory of determinism in human affairs. This is the sweeping formulation that was proposed by the young, self-educated whiz-kid of mid-nineteenth Britain, Henry Thomas Buckle. Buckle, born in 1821 to a wealthy London merchant and being of very delicate health, was largely educated at home. He became fluent in many

different languages and made a name for himself as one of the world's best chess players of his day. His father's wealth enabled Buckle to pursue the career of his choice. Buckle decided that he would commit himself to writing a definitive work of history, to which he devoted several years of intense labour. In 1857, at the age of thirty-five, Buckle published the first volume of his *History of Civilization in England.* The second volume appeared four years later. These were to be but the beginning of a mammoth undertaking, but Buckle died soon after the appearance of the second volume, so the work remained incomplete.

What was published, however, contained Buckle's philosophical views on the nature of history, which were very much informed by Quetelet's social statistics. Buckle argued very strenuously that there was a science of human history, on the grounds that human beings, as a society, behaved according to entirely predictable and fixed rules. There was, in short, a determinism in human affairs that individuals were powerless to stop or to alter. In support of this thesis, Buckle cited many of the statistics published by Quetelet, giving especial attention to those which indicated that matters on which people commonly thought that they had total freedom of choice were dictated by events totally out of their control. For example, after citing Quetelet's data on the age at which criminals are most active, Buckle continued:

> Nor it is merely the crimes of men which are marked by this uniformity of sequence. Even the number of marriages annually contracted is determined, not by the temper and wishes of individuals, but by large general facts, over which

individuals can exercise no authority. It is now known that marriages bear a fixed and definite relation to the price of corn; and in England the experience of a century has proved that, instead of having any connexion with personal feelings, they are simply regulated by the average earnings of the great mass of the people: so that this immense social and religious institution is not only swayed, but is completely controlled, by the price of food and by the rate of wages. In other cases, uniformity has been detected, though the causes of the uniformity are still unknown. Thus, to give a curious instance, we are now able to prove that even the aberrations of memory are marked by this general character of necessary and invariable order. The post-offices of London and of Paris have lately published returns of the number of letters which the writers, through forgetfulness, omitted to direct; and making allowance for the difference of circumstances, the returns are year after year copies of each other. Year after year the same proportion of letter-writers forget this simple act; so that for each successive period we can actually foretell the number of persons whose memory will fail them in regard to this trifling and, as it might appear, accidental occurrence. (*Buckle 1904*, p. 18)

Buckle's work had an extraordinary impact on British intelligentsia when it first appeared. It quickly became an item on everyone's must-read list. Among those who were persuaded to read it was a young man who had just started a career as an evangelical cleric in the Church of England, the same year in which volume one of Buckle's *History* was published. This was John Venn, now known to most of us as the inventor of Venn diagrams, and little else.

According to his own account, and buttressed by comments from several people who knew him well, Venn had had a very limited education when he went up to Cambridge as a young matriculant in 1853. His father was a prominent evangelical cleric, the leader of the Clapham Sect and the Secretary of the Church Missionary Society. Venn showed a talent for mathematics, so that is what he read at Cambridge, taking a first class degree as Sixth Wrangler in 1857. But soon after finishing his degree, Venn forsook mathematics and sold most of his mathematics books. Instead he took up his father's profession, which, he later said, he had never considered not doing, as it was traditional in his family. But he did not last long as a Churchman. After five years as a curate, Venn returned to Cambridge and took a position at his college, Gonville & Caius, as a lecturer in the moral sciences, where he remained for the rest of his life.

It was either during that brief time as a curate, or during his first years as a tutor at Caius College that Venn read Buckle's *History* and it had a great impact on him. Buckle was arguing for historical determinism and against the notion of personal

free will. At about the same time that Venn read Buckle, he was also introduced to John Stuart Mill's *System of Logic*.

Buckle had argued that much of history had been written as mere reporting of events as though they occurred by happenstance, largely a matter of luck or Chance. Whereas, he, Buckle, had found that the march of history is inexorable and follows directly upon the path laid by the accumulation of knowledge. Just as the discoveries of science have brought us to see that Nature obeys laws, so will the regularities discovered by social science show us the inevitable course of history. To quote Buckle, "in the ordinary march of society, an increasing perception of the regularity of nature destroys the doctrine of Chance, and replaces it by that of Necessary Connexion' (*1904*, p. 5).

Venn found these arguments persuasive, and in Mill's *Logic,* he found what he considered a framework for handling the incomplete information represented by statistics and the transformation of that through the process of induction to knowledge. But he was troubled by the implications of both men's thought for free will. Both Buckle and Mill appeared to argue for a complete historical determinism. Chance represented at best incomplete information. Greater knowledge would lead to greater understanding, but never to individual freedom. And if there was not freedom of the will, then how could there be individual responsibility for actions? How could there be, for example, sin?

Mill's view of freedom of the will, was that it was a perception, but an erroneous one. In the *System of Logic*, Mill wrote "if we knew the person thoroughly, and knew all

the inducements which are acting upon him, we could foretell his conduct with as much certainty as we can predict any physical event" (*Mill 1987*, p. 13). Buckle took the view that individual actions might be individually chosen, but in the long run, actions that differed from central tendencies would cancel each other out, so that the overall effect, which determined the course of history, was predetermined.

According to Venn's son, J. A. Venn, it was this argument of historical determinism by Buckle that drove John Venn to re-awaken his mathematical interests and make a thorough study of chance, what it means, and the foundation that it provides for the theory of probability. It was during those first years back at Caius College, now in the role of Junior Fellow and Catechist in the Moral Sciences, in the early 1860s, that Venn undertook his analysis of chance. This was published in 1866 in what was Venn's first book, *The Logic of Chance.* The first edition bore the subtitle *An Essay on the Foundations and Province of the Theory of Probability with Especial Reference to Its Application to Moral and Social Science*.

What seemed to bother Venn greatly was Buckle's tendency to make sweeping generalizations about historical determinism on the basis of a few extraordinary examples. Buckle argued that human history is inevitable, and even predictable, once we know enough, because, as his examples show, people behave in predictable ways. They merely think they have freedom of choice. This Venn could not accept. In *The Logic of Chance*, he berates Buckle for arguing from bizarre examples to startle the reader into accepting his concept of determinism:

> Instances are selected which, though they possess no greater logical value, are, if one might so express it, *emotionally* more powerful. That the annual number of suicides should remain nearly the same is assumed to be strange enough, but what are we to say to the staggering fact that the number of misdirected letters annually sent to the post-office is about the same? … The aim of all such attempts is the same. It is by the help of statistical uniformity to establish the existence of causation (in the sense of invariable unconditional sequence) in individual cases (*1866*, pp. 328-329).

But in fact, Venn is caught in a web of contradictions of his own. Venn, the ordained priest of the Church of England, cannot accept a notion of human determinism that could be interpreted as removing personal responsibility for human actions, just because we are all driven by forces larger than ourselves. But at the same time, Venn liked to think of himself as a "man of science" and that committed him, in his own assessment, to the view that everything in Nature is subject to Nature's inexorable laws. From this view, there is no such thing as chance. In *The Logic of Chance*, Venn cited with approval David Hume's assessment of chance in the world as being an expression of our ignorance of the real causes.

Venn could not deny determinism in Nature without aligning himself with opponents of science, which he could not abide. Nor could he subscribe to determinism in the affairs of men without absolving sinners from responsibility for

their actions. His solution was to make the curious assertion that statistical regularities that apply to the aggregate have no bearing at all on the individual instances that make it up. Hence, there is, or at least there may be, freedom of choice at the individual level while aggregated results are predetermined.

This position, which basically I view as muddled doubletalk, had an extraordinary freeing effect on the development of statistics. Venn in effect said that one could collect statistics and discern correlations therefrom without committing oneself to all manner of implications about human nature.

When Quetelet collected statistics on human beings, he quickly took them as data distributed around some pre-ordained true value that represented Nature's plan. If a probabilistic measure emerged from his studies, it was a measure of inevitable likelihood, such as the likelihood to commit suicide, that was as objective as a measure of the composition of the gases in the atmosphere. Whereas when Venn saw a collection of data, the probabilistic measure that emerged from that was simply the frequency of a particular result within the entire sample. It might point to some underlying, but unknown cause, but then it might just have happened.

This relaxation of the relationship of the individual outcomes in any set of data to their aggregate frequencies permits a different view of individual differences to emerge. Statistical correlations may be useful information, even when the ultimate causes are unknown. Where Quetelet looked to the mean value as a best approximation of a Platonic ideal, others, notably Francis Galton, took the frequency interpretation to imply only a central point around which deviations extend for

unknown reasons. For Galton, those deviations were more interesting than the central point. In fact, with Galton, we get quite a different interpretation of the mean, one we are more familiar with today. Galton's mean is the mediocre, with the ideal of excellence being measured by the widest deviation from that mean in a particular direction, for example, the category of genius, which interested Galton greatly.

The Logic of Chance was published in 1866. For a time, the discussion that it engendered appeared confined to questions of probability theory, or, to be more precise, philosophical discussions of the foundations of probability theory. Its significance to statistics was not apparent until the work began to be discussed by a new generation of thinkers, including Francis Galton that I have already mentioned. Another significant figure in the late 19th century was a distant cousin of Galton's, the enigmatic Francis Ysidro Edgeworth, who is remembered today primarily for his contributions to economics. A third figure was Karl Pearson, who is often thought of as a philosopher of science, but who contributed considerably to the development of statistics as an applied science. Stephen Stigler, the eminent historian of statistics refers to a notable change in the intellectual climate of statistical methods and the construction of an empirical and conceptual methodology beginning in the 1880s, as principally the work of these three men: Francis Galton, Francis Ysidro Edgeworth, and Karl Pearson. (Stigler, 1986, pp. 265-66.)

Edgeworth and Galton corresponded with each other, and both corresponded with Venn and quoted Venn's *Logic of Chance* in their works in the 1880s as a benchmark on questions of probability as they applied to statistics. Venn quoted

Galton in the third edition of *The Logic of Chance*, published in 1888. In 1890, Karl Pearson, who was then professor of applied mathematics at University College, London, accepted the additional appointment of the lectureship in geometry at Gresham College. This entailed giving a series of public lectures annually on topics of his own choosing. In 1892, Pearson's topic was "Laws of Chance: Being the Elements of the Theory of Probability in its Relation to Thought and Conduct." (Stigler, 1986, p. 327.) These were elementary lectures to a general audience and contained nothing new. Significant, however, is that Pearson cited the work of Venn and Edgeworth in his talks. In 1893, Pearson took ill and was advised to forego his lectures for a time. Accordingly, he sought out people who could replace him effectively, talking on the topics he had announced, which were the applications of the laws of chance, that were discussed the previous year. In February of 1893, Pearson wrote to Venn, introducing himself as a "perfect stranger," and asking if Venn would be willing to undertake one of his lectures for him. Venn agreed and proposed a topic which he named "Curves of Error"—an interesting choice of title, as it preserved the historic sense of a cluster of data points representing errors from some true value, a philosophical position in a sense incompatible with Venn's own views, but one that was entrenched in the literature. Interestingly, Pearson heartily agreed to the topic and the abstract that Venn had sent him, but asked permission to change the title to "On Frequency Curves," which much better summed up Venn's outlook. Pearson argued that the term "curve of error" would confuse his popular audience. (Venn archives, letters dated Feb 28, March 2, and March 24, 1893.) Thus, on grounds of not confusing his audience,

Pearson signalled the change in thinking of what is implied by a set of data that cluster around a midpoint.

Venn's muddle, which effectively denied inevitable determinism in human affairs, opened the door to just looking at data and observing what appeared alongside what. From that, it was not an impossible leap to draw conclusions that had only a mathematical probability of being correct, but nevertheless worked out often enough that the aggregate results were reliable. Venn's interpretation led the way to using statistics to manage groups while leaving the status of the individual alone. It opened statistics to wide application. Venn himself did not develop those applications, but his work shifted the thinking of those who did.

Byron E. Wall, York University, July 2008

Quoted by A. J. Ayer in his Introduction to Mill's *Logic of the Moral Sciences.*