The Enigma of Karl Pearson and Bayesian Inference

John Aldrich

1 Introduction

"An enigmatic position in the history of the theory of probability is occupied by Karl Pearson" wrote Harold Jeffreys (1939, p. 313). The enigma was that Pearson based his philosophy of science on Bayesian principles but violated them when he used the method of moments and probability-values in statistics. It is not uncommon to see a divorce of practice from principles but Pearson also used Bayesian methods in some of his statistical work. The more one looks at his writings the more puzzling they seem.

In 1939 Jeffreys was probably alone in regretting that Pearson had not been a bottom to top Bayesian. Bayesian methods had been in retreat in English statistics since Fisher began attacking them in the 1920s and only Jeffreys ever looked for a synthesis of logic, statistical methods and everything in between. In the 1890s, when Pearson started, everything was looser: statisticians used inverse arguments based on Bayesian principles and direct arguments based on sampling theory and a philosopher-physicist-statistician did not have to tie everything together.

Pearson did less than his early guide Edgeworth or his students Yule and Gosset (Student) to tie up inverse and direct arguments and I will be looking to their work for clues to his position and for points of comparison. Of Pearson's first course on the theory of statistics Yule (1938, p. 199) wrote, "A straightforward, organized, logically developed course could hardly then [in 1894] exist when the very elements of the subject were being developed." But Pearson never produced such a course or published an exposition as integrated as Yule's *Introduction*, not to mention the nonpareil *Probability* of Jeffreys. His most encompassing work, the *Tables for Statisticians and Biometricians* (1914 and 1931), assembled methods without synthesising ideas.

Students of Pearson's very full career have concentrated on particular periods and topics. The period 1893-1900 has attracted most attention for it was then that Pearson produced the ideas that transformed statistics; for this period Stigler's *History* (1986) is particularly useful. The older Pearson has often been read more with an eye to his impact on Ronald Fisher, the next great transformer; see e.g. Aldrich (1997) and Stigler (2005). Pearson's career does not divide into a Bayesian period and a non-Bayesian period for, as Yule (1936, p. 85) noted, Pearson "showed repeated interest in problems of inverse probability." Pearson is a large detail in Fienberg's (2006) tapestry of Bayesian history and his Bayesian writings get a close examination in Dale's *History from Thomas*

Bayes to Karl Pearson (1999). However these accounts do not consider how Pearson managed his Bayesian and non-Bayesian sides. Pearson is at large in Hald's *History* (1998), either looking back to Bayes and to Laplace or forward to Fisher, but he is never the centre of attention. Yule's (1936) obituary, Egon Pearson's (1936/8) biography and Eisenhart's (1974) *DSB* article remain the only attempts to grasp the whole of Pearson the statistician. Eisenhart does not register Pearson's Bayesian activities, Yule records them but Egon does more, he conveys a sense that he felt them. As a beginner, he entered into his father's Bayesian projects and later he became a leader in the movement away from Bayesian ideas.

Sections 2 and 3 below consider Pearson before he became a statistician. Pearson learnt about probability as an undergraduate and it was probability with a Bayesian spin. He first wrote about probability and Bayesian arguments when he was reflecting on physics–this was the work that impressed Jeffreys. When Pearson started in statistics he used Bayesian methods but did not rely on them and was open to inspiration from any source. His basic techniques, significance testing, the method of moments and the Gaussian rule, are discussed in Sections 4-6. How these techniques worked with or against his earlier ideas on probability provides the material for Sections 7-11. Along the way notice is taken of how Pearson's contemporaries combined direct and inverse analysis in their own work. This study of the Bayesian element in Pearson's work offers a perspective on his approach to statistical inference and on the condition of Bayesian statistics in his time. The chapter concludes with some observations along these lines.

2 Probability with a Bayesian spin

In the Pearson Papers there are notes on the theory of probability belonging to "Carl Pearson of King's College." The 80 pages are not notes a student makes on his reading but lecture notes. Probability and the theory of errors were on the syllabus of the Mathematics Tripos and there was a system of intercollegiate lectures: in Lent 1878 lectures were offered on the "Theory of probabilities" and then in Easter 1878 and -79 on the "Theory of chances." It seems likely that Pearson's notes came from these lectures. The lecturer was not one of the Cambridge probability names–Todhunter, Venn or Glaisher–but Henry Martyn Taylor (1842-1927) a fellow of Trinity and third wrangler in 1865. Taylor was a geometer and appears never to have published on probability; his life is recalled by Lamb (1928). Pearson does not mention Taylor in any of his writings and there are no letters or other matter connected with him in the Pearson Papers.

The notes are in two parts: I-VII on probability and VIII on the theory of errors. I-VII contain no references although the treatment resembles that in Todhunter's *Algebra* (1858) and *History* (1865). Part VIII is described as "from" Thomson & Tait (1869) and Chauvenet (1863). VIII is in a separate booklet and were it not for the numbering it might seem to be another course. After stating the basic propositions of probability the notes proceed to "mathematical

& moral expectation" with applications to gambling and insurance. A section on conditional probability leads into the "probability of future events based on experience" and "applications to witnesses and juries." After a discussion of the solution of "questions involving large numbers" and a section on problems Part VIII finishes with "errors of observation & method of least squares."

The notes begin with two definitions:

> Defns. (1) "*chance*" If we can see no reason why an event is more likely to happen than not to happen, we say it is a "chance" whether the event will happen or not.
>
> (2) "*Probability*" The mathematical probability of any event is the ratio of the no. of ways in which the event may happen to the whole no. of ways in which it may either happen or not.

The second classical definition is the one that matters as there is no formal development of "chance." Later in the notes the terms seem to be used interchangeably.

Inference, as we would call it, follows Bayesian principles, as we would call them. It is convenient to use the term "Bayesian" although it has been in use only since the 1950s; see Fienberg (2006). Carl was introduced to the two main examples of Bayesian reasoning: the "probability of future events based on experience" which went back to Bayes (1763) and Laplace (1774) and the treatment of "errors of observation" which went back to Gauss (1809). The problems would now be described as inference from past Bernoulli trials to future ones and inference to the value of the coefficients in normal regression; the analysis in both cases is based on a uniform prior.

In the first class of problem the events are draws from an urn; this was Laplace's scheme, Bayes's scheme in which balls are rolled on a table did not survive Bayes and we will not need to notice it until much later when Pearson revived it–see Section 10 below. The urn contains an infinite number of white balls and black balls in an unknown ratio and the prior for the ratio is assumed to be uniform on the unit interval. The first problem considered is, given that a white ball has been obtained m times and a black ball n times, to "find the probable composition of the urn and thence the probability of drawing a white ball at a future trial." The solution, the mean of the posterior distribution as it would now be called, is given as

$$\begin{aligned} \pi &= \frac{(n+m+1)!}{n!m!}\int_0^1 x^{m+1}(1-x)^n dx \\ &= \frac{n+1}{n+m+2}. \end{aligned}$$

The extension to the case of obtaining m' white balls and n' black balls in $m' + n'$ future trials is also given. These results were obtained by Laplace (1776) and are reported in Todhunter (1865, pp. 466-7). Venn (1866) called the specialisation to the case $m = 0$ the "rule of succession" and he ridiculed its use. Carl's lecture notes contain no hint of controversy but when Karl wrote about probability–in 1888–he knew of the controversy and he was still referring to it in the 1920's.

Continuing with the same assumptions, the "probability after the event that the chance lies between x & dx" (sic) is:

$$\pi = \frac{x^m(1-x)^n dx}{\displaystyle\int_0^1 x^m(1-x)^n dx}.$$

No name is given to this result but, as it was the main result in Bayes (1763), it was generally known as "Bayes' theorem." In his own later writings Pearson referred to it as "Bayes' Theorem" and I will adopt his capital letters to remind us that the expression is not being used in the modern sense for one of the elementary theorems in conditional probabilities. That sense seems to have been introduced into the English literature by Burnside (1924) although it had long been current in the French literature; see Section 10 below.

Finding the chance that the probability lies in an interval involves integrating this density. The lectures give an approximation for the probability that the probability is between $\frac{m}{m+n} \pm \alpha$ for the case when "α is small so that α^2 can be neglected." The integrals involved are incomplete B-functions and the evaluation of such integrals was Pearson's longest running interest in probability; 55 years later the publication of the *Tables of the Incomplete Beta-Function* (1934) closed the book on the subject–see Section 10 below.

Another lasting presence would be the Gaussian method used in the theory of least squares. The methods of Part VIII of the notes came ultimately from Gauss (1809). Gauss had assumed a uniform prior but although the lecturer describes the problem as one in "inverse chances" he does not introduce a prior. Textbooks, such as Chauvenet, often adopted this prior-less Bayes approach; see Aldrich (1997). Carl's notes set up the simplest problem as follows:

> Let $\psi(x)dx$ be the chance that an error lies between x & $x+dx$ and let $a_1, a_2, ..., a_n$ be the observations made & let T = true result. To find the most probable form of the density function of the errors (ψ) and the most probable value of the true measurement (T) is a problem in inverse chances.
>
> ...
>
> Observed event is that errors $a_1 - T$, $a_2 - T$ etc. have been obtain, chance of this
>
> $$= \frac{\psi(a_1-T)\psi(a_2-T)\psi(a_3-T)...dT^n}{\displaystyle\iiint ...\psi(a_1-T)\psi(a_2-T)...dT^n}$$
>
> & this is to be made a maximum.

At this stage the notes treat both T and ψ, the distribution of the errors, as unknowns. But the notes continues by observing that while "no one has been able to find ψ" T can be found by assuming the normal form for ψ. The prior-less Bayes approach was not the only version of least squares available in English. Airy (1862) gives Gauss's second proof of least squares, what is now called the Gauss-Markov theorem, but that did not become part of the regular instruction.

In old age Pearson (1936a, p. 29) recalled that when he was an undergraduate he often went to the original sources and Porter (2003, p. 212) suggests that he did so in probability. The notes reveal nothing either way. The fact that Pearson never wrote about his undergraduate study of probability suggests that it made little impression on him at the time. On the other hand, he kept returning to the Bayes-Laplace arguments to examine their basis and to improve the approximations involved in getting results. The least squares/inverse chances analysis was not lodged so deeply in his consciousness.

In 1883 after much wandering geographical and intellectual–see E. S. Pearson (1936) and Porter (2003)–Pearson settled at University College London as the professor of applied mathematics. An important part of the job was to teach mechanics to engineering students and for this Pearson found the new "graphical statics" very suitable. His enthusiasm for graphical statics went beyond its value for teaching and its value for mechanics: as Porter (p. 235) observes, "For the first few years Pearson's statistical ambitions were encompassed within his program for graphical statics." Pearson's earliest work on probability was not linked to his graphical projects; it was essentially philosophical but it was linked to mechanics because his philosophical projects came out of mechanics. He first cultivated his interest in philosophical mechanics when he worked on Clifford's manuscript *Common Sense of the Exact Sciences*–see Pearson (ed.) (1888).

3 Probability and belief in invariable sequences

Physicists used probability in the theory of gases and in the reduction of observations. In his scientific research Pearson did not use probability but when he thought *about* science he thought probability. His thoughts on probability and science are set out in the *Grammar of Science* (1892) which would later make such an impression on Jeffreys. But they are already visible in "The prostitution of science" (1888), a polemic against the "physico-theological proof of the existence of a deity" advanced by the Cambridge physicist G. G. Stokes (1887). Stokes's principal fault was to misunderstand the nature of "scientific law" and a principal objective of the *Grammar* was to give an accurate account of this notion. A law, Pearson (1892, p. 135) explained, is "a brief description in mental shorthand of as wide a range as possible of the sequences of our sense-impressions." Right at the end of his life Pearson linked this conception of law to his statistical practice–see Section 12 below–but the teaching of the *Grammar* that was most evident in his statistical writing and lectures was the teaching on probability.

Pearson (1888, p. 35) held that there is "no absolute knowledge of natural phenomena" only "knowledge of our sensations." Probability entered in the following way:

> our knowledge of the 'invariability' [of our sensations] is only the result of experience and is based, therefore, on probability. The probability deduced from the sameness experienced in the sequence of one repeated group of sensations is not the only factor, however, of this invariability. There is an enormous probability in favour of a general sameness in the sequences of all repeated groups of sensations.

The essentials of the probability argument of the *Grammar* are already here. Pearson refers to Boole's (1854) criticisms of the "probability basis of our knowledge of sequence in natural phenomena" and registers his belief that they have been "sufficiently met" by Edgeworth (1884). Edgeworth was more exercised by Venn's criticisms and he found arguments that he (pp. 230-4) claimed "suggest that the particular species of inverse probability called the 'Rule of Succession' may not be so inane as Mr. Venn would have us believe."

In 1890 Pearson had an opportunity to develop the ideas from the *Common Sense* and from his occasional writings on science. In his application for the Gresham Lectureship in Geometry he proposed lectures "on the elements of the exact sciences, on the geometry of motion, on graphical statistics, on the theory of probability and insurance." (see E. S. Pearson (1936, pp. 212-3.)) The first two topics were treated in the lectures given in November 1891 which formed the basis of the book of 1892. Pearson was hugely enthusiastic about the "graphical representation of statistics" and made this the theme of his probationary lecture on "applications of geometry to practical life" (1891). The final topic, the theory of probability and insurance, seems more like padding. Pearson had something to say about probability but there is no sign that he had anything to say about insurance. However by the time the probability lectures came round Pearson had taken off in an unforeseen direction and was collaborating with the zoologist W. R. Weldon.

The *Grammar of Science* (1892) was not a treatise comparable to Jevons's *Principles of Science* or Mill's *System of Logic* but was more of a philosophical supplement to the chapters on mechanics in the Clifford volume. The *Principles* and the *Logic* treat probability with the techniques of science. The *Grammar* did not discuss techniques but it made probability central to science by giving it a fundamental role in connection with causation and belief in invariable sequences.

The chapter on "Cause and Effect–Probability" grew out of the remarks in the 1888 essay. Pearson (1892, p. 136) lays out the ground:

> That a certain sequence has occurred and recurred in the past is a matter of experience to which we give expression in the concept *causation*; that it will continue to recur in the future is a matter of belief to which we give expression in the concept *probability.* Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated.

Probability is about belief; it is cognate with certainty which he holds is attainable in the world of conceptions, not in the world of perceptions. This probability is the quantity of belief interpretation of Jevons and De Morgan, rather than the classical notion of Pearson's undergraduate notes or the "frequency interpretation" generally associated with his statistical work and found by some, including Dale (1999, p. 504), in the *Grammar*. Pearson's probability has no connection with indeterminism.

Pearson did not linger over the concept of probability but went straight into a discussion of how belief in regularity can be justified. The starting point (1892, p. 169) was the rule of succession, the result he had taken down as an undergraduate: "Laplace has asserted that the probability that an event which has occurred p times and has not hitherto failed is represented by the fraction $\frac{p+1}{p+2}$." After considering some examples Pearson concludes that this value does not "in the least" represent the degree of belief of the scientist which is actually much stronger. His point is that the scientist does not only have specific experience of just those trials but the general experience that such regularities exist. Laplace's rule had to be applied in a different way (p. 177):

> Suppose we have experienced m sequences of perceptions which have repeated themselves n times without any anomy. Suppose, further, a new sequence to have repeated itself r times also without anomy. Then in all we have had $m(n-1)+r-1$ repetitions, or cases of routine, and no failures; hence the probability that the new sequence will repeat itself on the $(r+1)$th occasion is obtained by putting $p = m(n-1)+r-1$ and $q = 0$ ... Therefore if m and n be very great, there will be overwhelming odds in favour of the new sequence following routine, although r, or the number of times it has been tested, be very small.

This extension/application of Laplace can be interpreted as the use of an informative prior based on experience with the other m sequences but that is not how Pearson presents it.

In his discussion Pearson examines the basis of Laplace's argument "a little more closely." He (pp. 174-5) accepts Boole's objection to the procedure by which we "distribute our ignorance equally" but asks whether Laplace's principle is not based on the *knowledge* that "in cases where we are ignorant, there in the long run all constitutions will be found to be equally probable." He concludes that this is so, appealing to a passage in Edgeworth (1884, p. 231):

> We take our stand upon the fact that probability-constants occurring in nature present every variety of fractional value; and that natural constants in general are found to show no preference for one number rather than another.

The notion that the uniform prior is justified only when it represents experience was one that Pearson held throughout; the uniform prior was an informed prior, not an ignorance prior. Pearson accepted the conclusion that when experience does not have this character the uniform prior should not be used but he used a non-uniform prior only once, in 1917; see Section 9 below.

The *Grammar* has plenty of references, both current and historical. Todhunter (1865) is recommended as a guide to Laplace's *Théorie Analytique* and Pearson even mentions Todhunter's sections on Prevost and Lhulier. Venn's *Logic of Chance* is among the references but Pearson does not square up to its very critical treatment of the rule of succession. This reading list lasted Pearson for the rest of his life: he went on writing as though Edgeworth's answer to Boole and Venn was the latest word. The reading list even had an extended life for Fisher took it over, changing the signs so that Boole and Venn went from bad to good; their arguments are reviewed by Zabell (1989).

The *Grammar* does not discuss statistics and its probability matter made no contact with the "geometry of statistics" the Gresham topic for the academic year 1891-2. Probability arrived in the following session's lectures on the "laws of chance." In the first lecture, "The Laws of chance, in relation to thought and conduct," Pearson (1892/1941, p. 90) indicates the relation between the two previous courses, "We shall find that statistics are the practical basis of much, if not all scientific knowledge, while the theory of chance is not only based on past statistics but enables us to calculate the future from the past, the very essence of scientific knowledge." Pearson (1892/1941, p. 94) contrasts two views, "According to Dr Venn probability deals with the laws of things, while according to De Morgan probability has to do with the laws of our thought about things." The *Grammar* had not mentioned probability in "things" but now Pearson (p. 90) advocated Edgeworth's position as a "middle road" between those–De Morgan, Mill and Jevons–who are "pushing the possibilities of the theory of probability in too wide and unguarded a manner" and Boole and Venn who are "taking a severely critical and in some respects too narrow view of them." Pearson argued that probability in thoughts–"subjective chance"–approximates probability in things–"objective chance"–when there are enough observations. The second lecture on "Chance and Ignorance" discussed the disputed principle of the "equal distribution of ignorance." The approach is that of the *Grammar* although Pearson no longer considers only the case of a run of successes. Bayes's Theorem duly appears.

4 Significance tests–more Edgeworth

At the beginning of 1893, after a year of graphical statistics and a term of probability, the Gresham lecturer turned to statistical inference. Pearson began by applying a test of significance to "the frequency of the improbable" with particular reference to games of chance. The results on roulette were published in his "Science and Monte Carlo" (1894): the verdict of "Science" was that Monte Carlo roulette is "a standing miracle, not a game of chance." (p. 43)

The inference scheme, expounded by Pearson (1894, p. 49), was based on the probability-value Jeffreys so objected to:

> Let the results of a large but definite number n of trials be known

> and be represented by s successes; let the result of an indefinitely great number of trials be also known and let it be represented on the average by S successes for every n trials. How great must the deviation $S - s$ be in order to lead us to assert that we are not dealing with a game of chance? What are the odds against such a deviation?

The observed value of $S - s$ is converted into a number of standard deviations and the probability of a deviation "as great as or greater than" (p. 51) is calculated using the large sample normal approximation to the binomial.

Pearson learnt about tests from F. Y. Edgeworth (1845-1925). Stigler (1986, pp. 316 and 327) describes their early encounters and some of their correspondence is printed in E. S. Pearson (1965). Edgeworth, who had been working in statistics for a decade, guided Pearson's reading but Pearson never became a close student of Edgeworth's own work and only a few of Edgeworth's writings ever touched him. Their minds were very different. Edgeworth was much more interested in foundations and perhaps more excited by arguments than by results. He was very self-conscious and ready to share his thoughts with the reader; he paid very close attention to the literature and it was often difficult to know where he stood himself. There are overviews of Edgeworth's statistical work in Stigler (1978, and 1986 ch. 9) and of his Bayesian productions in Dale (1999, pp. 439-447). From the beginning, Edgeworth used both Bayesian reasoning and sampling theory reasoning: Stigler (1978, p. 296) comments, "Like Laplace, Edgeworth was inconsistent in his application of inverse probability, reverting to sampling distributions when comparing means by significance tests."

One of Edgeworth's concerns in testing was whether the difference between two samples was "accidental" or "comes by cause." In the second case he (1885, pp. 187-8) described the difference as "significant." Pearson adopted the term and followed Edgeworth's large-sample normal procedure involving probable errors; see Sections 6-7 below. Pearson's own unique contribution to testing, the χ^2 test of 1900, did not involve probable errors but it involved probability-values. A characteristic statement–from Pearson (1900, p. 170)–was, "In 56 cases out of a hundred such trials we should on a random selection get more improbable results than we have done." The issue here was the fit of a curve and Pearson concluded, "Thus we may consider the fit remarkably good." Other illustrations involve the fairness of chance set-ups–like the early Monte Carlo investigations, though without their glamour.

Returning to the session of 1892/3, it was then that Pearson began collaborating with the biologist Raphael Weldon on statistical studies of evolution. The method of moments was the first result of their collaboration.

5 The method of moments and "good results"

The Bayes-Laplace link between past and future and the test of significance were great constants in Pearson's thinking. The method of moments was a third and perhaps more special for it was wholly Pearson's own. He devised it in the course of 1893 and applied it in the first two of his "Contributions to the mathematical theory of evolution." The inspiration came from mechanics: the concept of a moment was fundamental to the mechanics of beams and the calculation of moments was one of the display pieces of graphical statics. In the first contribution Pearson used the method of moments to dissect an abnormal frequency-curve into two normal curves and in the second to fit skew curves.

Weldon posed the dissection problem–the estimation of a mixture of normal distributions–at the end of 1892 and by August 1893 he (1893) was announcing results. A full account of the method was ready by October and appeared the next year as the first of Pearson's Contributions. E. S. Pearson (1965, pp. 8ff) and Magnello (1996, pp. 50-2) describe the biological background but the process by which Pearson arrived at the method is more obscure; the Gresham public saw nothing until it was all over.

A frequency-curve registers the number of specimens falling in a small range and Pearson (1894, p. 72) writes the equation for the compound curve as

$$y = \frac{c_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x_1-b_1)^2}{2\sigma_1^2}} + \frac{c_2}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x_2-b_2)^2}{2\sigma_2^2}}$$

where c_i is the (unknown) number of specimens belonging to the i-th component normal curve. Pearson's notion of a frequency-curve is discussed in Aldrich (2003).

To estimate the unknowns Pearson devised the method of moments–"chosen after many trials and errors." Pearson (1894, p. 75) does not indicate what was tried beyond saying

> Other methods produce exponential equations the solution of which seems more beyond the wit of man than that of a numerical equation even of the ninth order.

The method of inverse chances was an obvious choice. Although individual problems had been treated by other methods, this was still the only general method available. The mixture problem is easy to set up as one in inverse chances but not easy to solve. It produces an exponential equation and my conjecture is that Pearson tried this approach and then gave up.

In the abstract/preview Pearson (1893, pp. 331-2) describes the method of moments and says something by way of justification:

> The method adopted for the dissection is based on equality of the first five moments and of the areas of the abnormal curve and of its two components. This method is justified in the same manner as the determination of the normal curve by fitting any series of observations by aid of the area and the first two moments (*i.e.*, the first moment gives the mean, and the second the error of mean square) is justified.

Behind the method is the elastician's intuition: the area enclosed by a frequency-curve and the horizontal axis can be thought of as a material lamina with mechanical characteristics expressed by the value of the moments; now imagine a visible lamina determined by the frequency-curve of the observations formed from two invisible component laminae; the moments of the visible lamina are determined by the moments of the invisibles and vice versa. So calculate enough moments of the observed frequency-curve to allow the moments of the components to be determined. The determination involved an equation of the ninth order. The moments fed into this equation can be generated graphically and so graphical statistics was not just descriptive statistics.

On the matter of justification the main paper (1894, p. 75) adds little beyond some rather confusing reflections:

> while the mathematical solution should be unique, yet from the utilitarian standpoint we have to be content with a compound curve which fits the observations closely, and more than one such compound curve may arise. All we can do is to adopt a method which minimizes the divergences of the actual statistics from a mathematically true compound. The utilitarian problem is to find the *most likely* components of a curve which is not the true curve, and would only be the true curve had we an infinite number of absolutely accurate measurements.

The "utilitarian problem" does not seem at all like an expression of the "utilitarian standpoint." The first suggests maximising a posterior distribution and second least squares or some other way of maximising goodness of fit. Of course both roads lead to the sample mean in the canonical problem of estimating the mean of the normal distribution. The main recommendation for the new method is founded on analogy for the passage continues:

> As there are different methods of fitting a normal curve to a series of observations, depending on whether we start from the mean or the median, and proceed by "quartiles," mean error or error of mean square, and as these methods lead in some cases to slightly different normal-curves, so various methods for breaking up an abnormal frequency-curve may lead to different results. As from the utilitarian standpoint good results for a simple normal curve are obtained by finding the mean from the first moment, and the error of mean square from the second moment, so it seems likely that the present investigation, based on the first five or six moments of the frequency-curve, may also lead to good results.

Over the years Pearson's belief in the method hardened. He never investigated whether the method was "good" although he welcomed any news that it was; see Section 9 below. His final paper (1936b) bears the defiant epigraph, "*Wasting your time fitting curves by moments, eh?*"

Pearson submitted the main paper and the abstract together in October 1893. The abstract (1893) also reported research on a second problem, the formulation and fitting of asymmetrical frequency distributions. Shortly before Edgeworth had approached Pearson with "some skew price curves" asking "if I could discover any way of dealing with skewness" (Pearson's recollection is quoted by E. S. Pearson (1965, p. 10). This was a long-standing interest of Edgeworth's as Stigler (1986, pp. 330-331) notes. Pearson's suggestion appeared in a letter published in *Nature* (1893); the topic also figured in the new round of Gresham lectures (E. S. Pearson (1938a, p. 150). The communications of October 1893 were followed by a long memoir on "Skew variation in homogeneous material" (1895), the second of the Contributions. The "Pearson curves" had come to stay and more memoirs followed. The curve work was moments all the way: the curves were fitted by the method of moments and the different types were distinguished by moment conditions. Pearson may have contemplated alternative methods of estimation but he never wrote about them. In most situations the method of inverse chances would require the maximum to be found iteratively and although Gauss had developed iterative methods for least squares Pearson never seemed to consider using similar methods.

6 Correlation, "inverse chances" and "best results"

The third of the Contributions, "Regression, heredity and panmixia" (1896), read in November 1895, used correlation to investigate some of the phenomena of heredity. After two expeditions with the method of moments this was a return to the beaten track and the method of "inverse chances"–see Section 2 above. Correlation was itself on the beaten track for Galton and Edgeworth had proposed methods for estimating correlation and Weldon(1893) was using Galton's method in his empirical work; Stigler (1986, chapters 9 and 10) reviews all this activity.

Pearson (1896, p. 265) chooses the value of the correlation for which "the observed result is the most probable" and calls it the "best" value of r. Yule (1938, p. 198) calls the method "maximum likelihood" but he was writing after Fisher had introduced the name and popularised the method. Edgeworth (1908, p. 395) understood it as Bayes with an implicit uniform prior: "The application of the genuine inverse method to determine the coefficient of correlation was introduced by Professor Karl Pearson." Gauss (1816) used this method in an investigation which was a model for Pearson's reasoning. When Pearson introduced the method of moments he pointed to the analogy of the mean but now his (1896, p. 264) analogy was the mean square error:

> The problem is similar to that of determining σ for a variation-curve, it may be found from the mean error or the median, but, as we know the error of mean square gives the theoretically best results.

The report in Pearson's (1894-6, vol. 1, p. 71) lectures was that "Gauss has shewn that the second mean error gives better results than the first." Gauss (1816) had obtained the "second mean error" by the method of inverse chances and then compared its performance under repeated sampling with that of other estimates of precision.

Like Gauss, Pearson (1896, p. 266) worked with a single parameter, in his case the correlation. For the other four parameters he implicitly assumed that they were known, identifying the sample moments with the population moments. Applying a quadratic Taylor approximation to the logarithm of this concentrated density function for the correlation yields a normal density centred on its maximum value (given by the product-moment formula). This normal approximation to the posterior density of the correlation generates a probable error as well as an estimate. The argument followed the pattern of Gauss (1816) which Pearson (1894-6, vol. 1, p. 89) translated as "the standard deviation of the standard deviation". Pearson used probable errors to register the accuracy of estimation and to support significance tests. Thus he (1896, p. 267) concluded from the small probable errors computed from Weldon's shrimp data that, "The samples taken were sufficiently large to determine r with close practical accuracy." He also compared pairs of samples using the probable error of the difference, "With these probable errors the identity of the first pair of r's is unlikely; the identity of the second excessively improbable." Elsewhere in the paper Pearson reports the probable errors of the various means and standard deviations and comments on the "significance" or otherwise of the differences between samples.

The correlation paper was an outstanding contribution to Bayesian statistics. But it was not the work of a Bayesian statistician in the sense that Jeffreys was a Bayesian statistician. The method of inverse chances was in Pearson's repertoire but so was the method of moments and he used that more often. Any thoughts on justification seemed to run in terms of sampling properties. Not that Pearson did any justifying, he mentioned parallels but did not investigate them and there would never be a Pearsonian theory of best methods comparable to Edgeworth (1908/9) or Fisher (1922) or, in a different direction, to Jeffreys (1939).

In 1816 Gauss moved untroubled between posterior distributions and sampling distributions but he knew the difference. Whether Pearson did is unclear: one thing is clear, that by the time he produced a careful explanation of what a "probable error" is–in 1903–he was describing a feature of a sampling distribution.

7 Probable errors, Sheppard and Edgeworth

Unlike the method of inverse chances, the method of moments could not support significance tests because it generated no probable errors. To remedy this, Pearson & Filon (1898) applied the Gaussian method for calculating probable errors to the values found by the method of moments. Pearson seems to have understood the probable errors as sampling theory probable errors.

The "general theorem" of Pearson & Filon (1898, pp. 231–6) extends the results from the specific one-parameter situations treated by Pearson (and Gauss) to the general multi-parameter situation: the extension allowed Pearson's (1896) formula for the probable error of the correlation which was based on one-parameter analysis to be corrected. The big conceptual change was that in the new paper the Taylor expansion is not around the maximum probability value but around the true value and the discrepancy between estimate and true value is referred to as an "error." The probable error is now interpreted as a property of a sampling distribution. In the correlation case there was the ghost of a Gaussian argument and a link to a particular form of estimation but in the case of the skew curves there had never been a Gaussian argument and there was no link between the method of moments and the formulae for the probable errors. The general theorem is discussed in more detail by Stigler (2008) and by Aldrich (1997) who relate it to the argument of Fisher (1922) on the asymptotic distribution of maximum likelihood. A large part of Fisher (1922) is a reconstruction of Pearson & Filon for Fisher also demonstrates that the method of moments applied to the Pearson curves is inefficient and includes the correct large sample probable errors for the method.

Pearson produced three further papers on calculating the "probable errors of frequency constants"–for univariate distributions (1903), for bivariate distributions (1913) and for truncated distributions (1920). The series begins (1903, p. 273) with a clear statement of what a probable error is and it is clearly a sampling theory concept:

> The simple idea involved in the probable error of a statistical constant is of the following kind: If the whole of the population were taken, we should have certain values for its statistical constants, but in actual practice we are only able to take a sample, which should if possible be a "random sample." If a number of random samples be taken any statistical constant will vary from sample to sample, and its variation is distributed according to some law round the actual value of the constant for the total population. This variation would be very properly measured by the standard deviation of the constant for an indefinitely great series of random samples.

Previously Pearson had assumed the concept was too familiar to need explaining.

The 1903 paper by-passes the general theorem of Pearson & Filon and "endeavours to give simple proofs of the main propositions of the subject." It shows how to calculate the standard deviations of sample moments and the correlations between errors in moments. The derivations are based on sampling theory. Pearson (1903, p. 280) concludes by describing the "general problem" which is to find the probable error of any constant c_i of a frequency distribution and the correlation of any two constants c_i and $c_{i'}$. "Any constant will be a function of the mean h and the moments μ_2, μ_3, μ_4,.., μ_q ... about the mean" and so it is possible to use the δ method to obtain the needed expressions. Thus we have "the means of determining the probable errors and the correlations of the constants of any frequency distribution whatever." Pearson indicates how the probable errors of the constants of the Pearson curves fitted by the method of moments can be worked out but he does not go through the calculations; it is implicit that the formulae in Pearson & Filon are correct.

Two "fundamental memoirs" are named by Pearson (1903, p. 273)–Pearson & Filon and a paper by W. F. Sheppard. Sheppard was the only English mathematical statistician of the time to base inference exclusively on sampling theory. Sheppard (1899, p. 102) noted that some of his formulae have already been obtained by Pearson "but by a different method." Sheppard did not comment on that method and generally he did not argue with his Bayesian colleagues. For Pearson (1903) the difference between his new approach and that of both fundamental memoirs is that the memoirs used information on the form of the distribution of the observations and not just on their moments. Thus when Pearson (1913) re-derived the probable errors associated with correlation he reported that "they have been reached independently of any system of Gaussian distribution."

Edgeworth also had views on probable errors and in 1908-9 he published a multi-part paper on the "probable-errors of frequency constants". This made no impression at the time and has only been read as people have asked how much of Fisher's (1922) account of the efficiency of maximum likelihood it contains. It is curious that the paper made no impression on Pearson for it could have been written for him: his work is referred to and the paper treats the question he had raised–and left–in the mid-90s, what is the "best" value of a frequency constant? Although the paper did not affect Pearson, it is worth pausing over because it summarises many of the thoughts on inference Edgeworth had been accumulating over the previous 25 years.

Edgeworth (1908, p. 387) still defended the uniform prior:

> I submit that very generally we are justified in assuming an equal distribution of apriori probabilities over that tract of the measurable with which we are concerned. And even when a correction is required by what is known about the a priori probability, this correction is in general, as I have elsewhere shown, of an order which becomes negligible as the number of the observation is increased.

Edgeworth had already made the point about the negligibility of the correction in the large sample case in the 1884 paper. He was also aware of the difficulty that a uniform prior in a different parametrisation will generally produce different inferences; this was a difficulty that Fisher (1922) made much of. For Edgeworth it was another small-sample problem.

The question of whether Edgeworth anticipated Fisher arises because Edgeworth did not restrict himself to inverse analysis; see Pratt (1976) for details. Edgeworth (1909, p. 82) makes this remark on the validity of direct and inverse methods :

> The general theorem which comprehends the preceding proposition as a particular species may itself likewise be proved by a direct

> method free from the speculative character which attaches to inverse probability.

Edgeworth was not frightened by the speculative element but he liked to have things clear.

Edgeworth thought long and systematically about inverse probability as the basis for statistical inference. Pearson appeared to have no interest in this particular project although that did not stop him from thinking unsystematically; see below Section 9. However phrases like "the ultimate basis of statistical theory" did appear in Pearson's writing and they appeared in his writing about Bayes Theorem–how that could be so we consider next.

8 "The stability of statistical ratios"

Pearson may have once put himself on the "middle road" between the objectivists in probability and the subjectivists but this was not evident to a critic like J. M. Keynes–at the time a kind of proto-Jeffreys–who (1908, p. ii) put together and attacked "the theories founded by Dr Venn and Professor Karl Pearson on conceptions of statistical frequency." Moreover there did not seem to be much of the Bayesian statistician about Pearson: the method of moments owed nothing to Bayes, the probabilities he computed were sampling probabilities and he had retired, even if he had not repudiated, the Gaussian method for producing estimates and probable errors.

And yet Pearson went on teaching Bayes' Theorem: Gosset wrote it down in 1906/7, as Yule had in 1894. This was not inertia for Pearson had an active Bayes' Theorem project. In a manuscript of 1907 he described the "fundamental problem in statistics" as the "permanence of statistical ratios without change of conditions" just as the "fundamental problem in physics" is the "permanence of repeated sequences without change of conditions." In his paper "On the influence of past experience on future expectation" (1907) he wrote about the stability of statistical ratios. Pearson (1907, p. 365) explained why the matter was important, "One and all, we act on the principle that the statistical ratio determined from our past experience will hold, at any rate approximately, for the near future." The permanence of statistical ratios was "all important" in statistical theory and part of the "ultimate basis of statistical theory" (pp. 365 & 366.)

His paper (1907, p. 365) set out to

> put into a new form the mathematical process of applying the principle of the stability of statistical ratios, and to determine, on the basis of the generally accepted hypothesis, what is the extent of the influence which may be reasonably drawn from past experience.

The "generally accepted hypothesis" was the "equal distribution of ignorance." Although Pearson (p. 366) recognised that "It is perfectly easy to form new statistical algebras with other clusterings of chances" he was still satisfied with the old argument from Edgeworth

The paper took a theme from Pearson's pre-statistical days and extended it to statistics but the real business of the paper was with improving the mathematical process. To this end it applied Pearson's (1899) ideas on the hypergeometrical series to the approximation of the predictive density; there was a parallel paper, Pearson (1906), which did the same for direct distributions. (Dale (1999, pp. 507-8) analyses the mathematics.) Pearson (1907, p. 365) wanted to expose the inaccuracy of the normal approximation and to improve on that approximation. "Very crude approximations of the principle are occasionally made in vital statistics, or even suggested in medico-statistical textbooks." Thus when past experience consists of np occurrences and nq non-occurrences of an event, the "result anticipated" in m further trials is "mp individuals of the given character with a probable error $.67449\sqrt{mpq}$." The result is often given "without any regard to the relative magnitudes of m and n, or again of p and q."

In the paper there is a hint of a new statistical method based on Bayes' Theorem. Illustration II (pp. 374-6) involves a sample with $n = 1000$ and $np = 20$ and a "second sample" of $m = 100$ trials in which there are 6 occurrences:

> [A] percentage as large as this actually occurs with a frequency of 24.8 in the thousand... On the other hand, if we adopt a Gaussian distribution we should only expect a deviation as large as 4 from the mean to occur in excess 3.5 times in the 1000 trials; or the odds are 285 to 1 against its occurrence. Such long odds would reasonably lead us to suppose some modification in the population since the first sample was drawn.

The use of Bayes' Theorem for testing the equality of two proportions is one of the applications for the tables printed in Pearson's *Tables for Statisticians* (1914). The person who developed the hint and computed the tables was Pearson's medical statistician follower Major Greenwood (1880-1949).

Greenwood (1913) looked at situations in which normal approximations work badly, among them Bayes' Theorem. Greenwood produced tables for small values of m and n and he (1913, pp. 80-1) also described the significance test for investigating whether the difference in outcomes between a treated and a control group is an "effect." When Pearson (1914) reproduced Greenwood's tables he (p. lxxii) added his own variant of the significance test:

> Of 10 patients subjected by a first surgeon to a given operation only one dies. A second surgeon in performing the same operation on 7 patients, presumably equally affected, loses 4 cases. Would it be reasonable to assume the second surgeon had inferior operative skill?

Pearson calculates the probability that 4 or more of the patients will die and gives the odds against the occurrence as 20 to 1.

Pearson and Greenwood did nothing further with this argument and Greenwood moved out of Pearson's orbit. Greenwood had first mentioned this use of Bayes Theorem when he (1912, p. 648) was discussing a paper by Yule on measures of association. On the same occasion he criticised Yule for his unfair treatment of Pearson's assumption that behind any dichotomous response is a continuous latent variable. Greenwood seems to have had a conversion of some kind for when he next wrote about association it was with Yule and their paper–Greenwood and Yule (1915)–has neither Bayes' Theorem nor reverence for Pearson. Greenwood and Yule collaborated on three further papers one of which–Greenwood and Yule (1917)–had a Bayesian treatment of the Poisson distribution based on a uniform prior, "by analogy with Bayes' postulate" they (p. 41) said.

Yule was taught Bayes' Theorem by Pearson and his attitude to inverse probability is worth notice. As a working statistician he was eclectic: amongst his efforts were a Bayesian paper with Greenwood, one on minimum χ^2 (for measuring linkage) with Engledow and measures of association that were intuitively appealing rather than derived from any principle. Yule took more of a position in the *Introduction to the Theory of Statistics* (1911) and it is a Bayesian position. Sample to population inference is treated in Part III of the book on the "Theory of sampling"; here the ultimate concern is with the "standard deviation of the constant around the observed value."

Instead of the usual paraphernalia of prior, likelihood and approximations to the posterior Yule has only standard errors and an argument that the inverse standard error, the real object of interest, reflects the sample standard error and the prior standard error with the first dominating in large samples. The relevant paragraph is headed, "The inverse standard error, or standard error of the true proportion for a given observed proportion: equivalence of the direct and inverse standard error when n is large" (p. xi; p. 273). This part of the book (p. 349) ends with a warning:

> Finally it must be remembered that unless the number of observations is large, we cannot interpret the standard error of any constant in the inverse sense, i.e. the standard error ceases to measure with reasonable accuracy the standard deviation of the constant around the observed value...If the sample is large the direct and inverse standard errors are approximately the same.

With its emphasis on standard errors this conception seems to belong to the large sample normal world of Pearson (1896) and Pearson & Filon (1898). Yule was 'there' when Pearson wrote those papers and it is possible that his notion of the two kinds of standard error was his interpretation of Pearson's cross-over. Pearson, of course, did not register any change.

It may be making things too sharp to say that Yule the reflective statistician was a Bayesian and Yule the working statistician an opportunist ready to try anything that might work, including Bayesian methods. However it is worth defining such a position to help clarify Pearson's. Pearson the working statistician was also an opportunist prepared to use Bayesian methods, e.g. Bayes' Theorem for testing. Bayes' Theorem was important to Pearson the reflective statistician not so much because it provide a paradigm for statistical inference–it is not clear that anything did for Pearson–but because it was the basis for any belief in the permanence of statistical ratios. Permanence, not Bayes' Theorem per se, was "all important" in statistical theory. This line of thinking, which is evident in 1907, grew out of the *Grammar*, would reappear again in 1920–see below Section 10–and would finally appear in the preface to Part II of the *Tables* (1931, p. vi) with some small adjustment for the modern situation in which physics is becoming a "branch of mathematical statistics."

Another of Pearson's students–'Student'–was Bayesian both as a working statistician and as a reflective statistician. He and Pearson disagreed about Bayes, as we see in the next section.

9 "Bayes' Theorem ought only to be used ..."

Pearson's next publication on Bayes' Theorem–in 1917–seemed to break the mould. For thirty years Pearson had been defending Bayes' Theorem on the ground that the uniform prior reflects experience but now he described a case where experience called for a non-uniform prior. The setting was new too, inference to the value of a correlation coefficient instead of prediction from past Bernoulli trials to future ones. In the -10s the estimation question seemed to be more open than at any time since Pearson took his first steps twenty years before. R. A. Fisher was involved in these events and they have usually been examined for their impact on him; see e.g. E. S. Pearson (1968) and Aldrich (1997). At the time, though, Pearson was more concerned with Student who was insisting that he follow the Bayesian line.

Student (William Sealy Gosset) had been a student of Pearson's in 1906/7. His now celebrated *Biometrika* papers, "The probable error of a mean" (1908a) and "The probable error of a correlation coefficient" (1908b), owe much to Pearson for technique and nomenclature–see Aldrich (2003 and -08c)–but the only piece by Pearson that approaches them for deliberately Bayesian reasoning is his (1907). Student's standpoint is clearest in his (1908b) where inference proceeds by combining a sampling distribution and a prior; this variation on the usual approach is discussed by Welch (1958), E. S. Pearson (1990) and Aldrich (2008c). Fienberg (2006, p. 7) writes of Pearson's "strong influence" on Student but this Bayesian package does not come from Pearson's writings or from the notes he took of Pearson's lectures. However Student was in London while Pearson was working on Bayes Theorem and he may have taken something from Pearson's enthusiasm even if he took the idea in a different direction; Pearson went along with Student's thinking–at least in the sense that he published it in *Biometrika*!

Of Student's small-sample papers the significant one for Pearson was the correlation paper for it related to his own work. The first to follow it up was not Pearson but Herbert Soper. However Soper (1913, p. 115) acknowledged, "I am indebted to Professor Pearson for drafting the lines of this investigation and for critical supervision." Soper's investigation, "On the probable error of the correlation coefficient to a second approximation" was designed to improve on the *first* approximation, viz. the normal distribution, as an approximation to the sampling distribution of r.

Among Soper's (1913, p. 108) results was a formula for the mode $\widetilde{r}$ of the distribution of the sample correlation, r:

$$\widetilde{r} = \rho\left(1 + \frac{3(1-\rho^2)}{2n'} + \frac{(41+23\rho^2)(1-\rho^2)}{8n'^2}\right)$$

where ρ is the population value, n is the sample size and n' is $n-1$. There is a hint of a better way of estimating ρ in one of Soper's remarks (p. 91) on the implication of the "markedly skew character" of the distribution of r:

> The value of r found from a single sample will most probably be neither the true r of the material nor the mean value of r as deduced from a large number of samples of the same size, but the modal value of r in the given frequency distribution of r for samples of this size.

This suggests using the relationship between $\widetilde{r}$ and ρ to find the most reasonable value of ρ. Soper does not propose this but something similar is suggested in Pearson's "Appendix I to papers by 'Student' and R. A. Fisher" (1915).

Appendix I which is printed immediately after Fisher's (1915) paper on the exact distribution of r is a detailed study of the small-sample distribution of the standard deviation, a distribution Fisher had mentioned in his paper. Pearson (1915, p. 525) proposed a new way of estimating σ based on the relationship between the mode, $\widetilde{\Sigma}$, of the density for Σ, the sample standard deviation $\left(= \sqrt{\frac{1}{n}\sum(x-\overline{x})^2}\right)$ and the true parameter value,

$$\widetilde{\Sigma} = \sqrt{\frac{n-2}{n}}\sigma.$$

Pearson argues that if we take $\widetilde{\Sigma}$ as "that which has been most likely to have been observed" the "most reasonable value" for σ is given by $\sqrt{\frac{n}{n-2}}\Sigma$. He includes a table to assist in the calculations: thus "if Σ be observed, and $n = 20$, then the most reasonable value to give σ is $\Sigma/.9487$."

In a letter, reproduced in E. S. Pearson (1990, p. 26), Gosset argued against this procedure and for maximising the posterior density obtained by multiplying the sampling density by a fairly flat prior; this was in line with his 1908b procedure. Pearson's reply came in a long footnote to Appendix II, "On the distribution of the correlation coefficient in small samples" by Soper, Young, Cave, Lee & Pearson (1917). This extensive "cooperative study" took a while to produce for it involved heavy calculations at a time when the laboratory was losing staff to the services and was itself doing war work. The study appeared in the May 1917 issue of *Biometrika* although a year before Pearson had mentioned a finished paper to Fisher; see E. S. Pearson (1968, pp. 451-2).

The cooperative study abandoned the modal method: "The Editor" (p. 353n) reports that Student had pointed out that the "best value" of σ is obtained by maximising the density of Σ with respect to σ and concedes that this is a "desirable criticism." However the solution proposed by Student was not acceptable! The uniform prior is not "in accordance with our experience" a point Pearson discusses at some length, concluding (Soper et al. (1917, p. 354n)):

> To justify the equal distribution of our ignorance, we should have to assume that we neither knew the exact character measured, nor the unit in which it was measured, and such ignorance can only be very exceptional in the present state of our knowledge.

Student disagreed and was not satisfied with the proposal that an informative prior be used; see E. S. Pearson (1990, p. 27).

For Pearson the standard deviation was really only a warm-up for the correlation coefficient. "Suppose we have found the value of the correlation in a small sample to be r, what is the most reasonable value $\widehat{\rho}$ to give to the correlation ρ of the sampled population?" Section (8) of the paper answers that it is found by choosing the value of ρ which makes the product of $\phi(\rho)$, "the law of distribution of ρ's" and the density of r a maximum. The cooperators insist that correlation analysis is rarely performed in circumstances of radical ignorance and they give some illustrative calculations using an informative prior for ρ. The corollary is that

> Bayes' Theorem ought only to be used where we have in past experience, as for example in the case of probabilities and other statistical ratios, met with every admissible value with roughly equal frequency. There is no such experience in this case.

Here "Bayes' Theorem" refers to any situation where a uniform prior is used, not just to Bernoulli trials. For Pearson the theorem had always been associated with predictive inference but now he makes contact with the more common view that the Theorem could serve as a paradigm for statistical inference in general.

From the beginning Pearson had emphasised the importance of respecting past experience. In 1907 he had found it natural to follow the mathematicians but now Soper et al. (1917, p. 359) re-distribute the emphasis:

> Statistical workers cannot be too often reminded that there is no validity in a mathematical theory pure and simple. Bayes theorem must be based on experience, the experience that where we are *à priori* in ignorance all values are equally likely to occur.

This sermon was perhaps meant less for Student than for the mathematician Fisher whose (1915, pp. 520-1) use of the "absolute criterion" Pearson understood to be based on a uniform prior. Pearson upset both Gosset and Fisher: the former did not like the way his actual views had been ridiculed–see E. S.

Pearson (1990, p. 27)–while Fisher was angry that his views had been misrepresented. This misunderstanding and its consequences for Fisher, both in his work and in his relations with Pearson, are traced in Aldrich (1997).

Pearson had not finished his re-examination of estimation for he also was supporting an attack on Gauss's method from another direction. Kirstine Smith's "On the 'best' values of the constants in frequency distributions" (1916) was a very Pearsonian production but it came from a different side of Pearson: its title seemed to pick up from where Pearson (1894 & -6) left off, its criterion of 'bestness', minimum χ^2, derived from Pearson (1900) and the examples were based on work from the laboratory. Smith's best value *is* the minimum χ^2 value but the examples all show that the method of moments gives a good approximation.

There was something in Pearson's past that came in for criticism. Referring to the case of the parameters of the normal distribution, Smith (1916, p. 262) stated, "if we deal with individual observations then the method of moments gives, with a somewhat arbitrary definition of what is to be a maximum, the 'best' values for σ and $\overline{x}$ [the population mean]." The arbitrary definition is the "Gaussian test" and the objection (1916, p. 263n) is:

> while the Gaussian test makes a *single ordinate* of a generalised frequency surface a maximum, the χ^2 test makes a real probability, namely the whole volume lying outside a certain contour surface defined by χ^2 a maximum. Logically this seems the more reasonable, for the above product used in the Gaussian proof is not a probability at all.

The phrase "somewhat arbitrary" stung Fisher and he sent Pearson a note on Smith's paper; the note is reprinted with other documents in E. S. Pearson (1968, p. 454-6). Pearson's reply went beyond defending a young co-worker: he said that, although he had followed the Gaussian rule himself, "I very much doubt its logic now." He repeated Smith's point about a "real" probability, although he added in a postscript:

> Of course the reason I published Frøken Smith's paper was to show that by another test than the Gaussian, the method of moments gave excellent results, i.e. her second conclusion.

Aldrich (1997, pp. 167-8) and Stigler (2005, pp. 39-40) discuss the Pearson-Fisher exchange as an important part of the Fisher story. The exchange had no direct consequences for Pearson. His letter to Fisher shows his openness to argument on fundamental matters and his lack of means for resolving such questions.

The response to Student/Fisher and the endorsement of Smith each its own local logic and connected with something from Pearson's past but they did not fit with each other. Informative priors and minimum χ^2 satisfied Pearson as counters to Student and Fisher but he did no further work with either; these were not to be new methods of moments. Pearson did not forget about them: he reproduced the cooperators' tables in the *Tables for Statisticians* (1931, pp.

cliv-clxxx: pp. 252-3) and he (1928, p. 165) endorsed the method in his last discussion of inverse probability; he (1936b, p. 47) brought up Smith's "able, but not sufficiently appreciated" paper when he re-engaged Fisher over the Gaussian rule twenty years later.

10 "What Bayes meant by his own Theorem"

After the dispute with Student and Fisher there was an extraordinary change in Pearson's thinking, a new conception of Bayes' Theorem and related results which seemed to contradict what he had been saying against them, indeed what he had been saying for thirty years. Pearson's (1920, p. 1) subject was the problem of 1907, now called the "fundamental problem of practical statistics":

> An "event" has occurred p times out of $p + q = n$ trials, where we have no *a priori* knowledge of the frequency of the event in the total number of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?

He (1920, pp. 5-6) had an extraordinary discovery to report about this probability, denoted by C_r:

> the fundamental formula
>
> $$C_r = \frac{B(p+r+1, q+s+1)}{B(p+1, q+1)B(r+1, s+1)},$$
>
> of Laplace in no way depends upon the equal distribution of ignorance.

Pearson announced the discovery as he was presenting his latest ideas on evaluating B-functions. The tables of the incomplete Γ function (Pearson (1922)) had just been finished and one of use for the tables was for approximating B-function. Pearson had last published on the problem in 1907 and, as then, the new paper has a preamble emphasising the importance of the project. The title, "The fundamental problem of practical statistics," emphasised that much was at stake and there was a new idea, or at least a new emphasis, when Pearson (1920, p. 3) implied that the concept of a probable error depended upon solving the problem:

> Notwithstanding the criticisms of Boole and Venn all branches of science have adopted the theory of "probable errors": they have applied past experience of limited samples to predict what deviations are likely to occur from this past experience in future experience, and mankind acts in accordance with a firm conviction in the relative stability of statistical ratios.

Pearson's probable errors–see above Section 7–were sampling theory probable errors both in design and execution and no Bayes' Theorem went into their calculation. Presumably Pearson believed that to license the application of these quantities to out-of-sample prediction Bayes' Theorem was required.

Previously–in 1917–Pearson had broadened his conception of Bayes' Theorem but now he (1920, p. 3) took it back to its original meaning, indeed to the very way Bayes presented it:

> But any numerical appreciation of the reasonableness of this conduct is apparently based on the "equal distribution of ignorance" or ultimately on such a quaint idea as that of Bayes that his balls might roll anywhere on the table with equal probability.

A close study of Bayes's quaint idea led Pearson to conclude that the *apparently* of the second sentence could be extended to *but not actually.*

The claim that Laplace's solution to the fundamental problem did *not* depend on the "equal distribution of ignorance" was received with incredulity by Edgeworth (1921, pp. 82-3fn.) and Burnside (1924). Their point was that the posterior would only remain unchanged when the prior was changed if–through some subterfuge–there were a compensating change in the likelihood. Naturally Edgeworth acknowledged that the prior did not matter in the case of large samples for he had been saying as much for years. The reply from Pearson's (1921, p. 300) was that, "Some misunderstanding has arisen with regard to my paper ... I believe it is due to the critics not having read Bayes' original theorem." He (p. 301) declared further

> But if the critics say: Then this is not what we mean by Bayes' Theorem, I would reply: Quite so, but it is what Bayes meant by his own Theorem, and it probably fits much better the type of cases to which we are accustomed to apply it than what you mean by Bayes' Theorem.

Pearson's (1924) second response was to Burnside (1924); Burnside was an outsider, a distinguished group theorist who made a few forays into statistics–they are described by Aldrich (2006). Pearson (1924, p. 190) was perfectly correct when he said, "Dr. Burnside, I venture to think, does not realise either the method in which I approach Bayes' Theorem, or the method in which Bayes approached it himself." Indeed Burnside saw no reason why he should be interested in the historical Bayes; he took his statement of "Bayes' formula"–the one used today–from Poincaré.

By resurrecting Bayes's "quaint idea" Pearson became Bayes's first modern commentator. The table model had disappeared from the literature long before. Of course it described by Todhunter (1865, pp. 294-300) but he makes it clear that Bayes's own way of proving his theorem had been superseded. Pearson (1920) has been read by students of the historical Bayes–including Edwards (1978, p. 118), Stigler (1982, p. 255), Dale (1999, pp. 509-516) and Hald (1998, pp. 253-6)–and they have considered where he goes wrong and why Pearson found the argument attractive. Perhaps the root of the mistake was Pearson's identification of the "equal distribution of ignorance" with the notion that the balls "might roll anywhere on the table with equal probability." Bayes had assumed a uniform table but, as Edwards (1978, p. 118) and Stigler (1982, p. 255) point out, the argument goes through whether the table is undulating or flat provided all the balls are thrown on the same table. Pearson may have had this at the back of his mind and it may have been coming forward as he executed his all guns blazing retreat in 1921 and -24. Stigler (1982, p. 255) suggests that Pearson was drawn to the model both because he was deep in Bayes in connection with his work on the history of statistics–see Pearson (1921/33)–and because he used a similar model in his own research. One of the points at issue in the Pearson-Yule controversy over association–see Section 8 above–was the notion that an event is triggered when an underlying continuous variable passes a threshold. The threshold in Bayes is fixed by the stopping point of the first ball, while in the Pearsonian variant–see his (1921, p. 301)–"men [sicken] from a disease when their resistance falls below a certain level." There was *something* behind the reference to "the type of cases to which we are accustomed to apply" the Theorem but it was not what anybody else had in mind or even what Pearson had in mind in his earlier writings, especially in exchanges of 1915-17.

The 1920 paper is perplexing. There is nothing extraordinary in wishing that the Bayes-Laplace conclusion did not rest on such narrow foundations, or in wondering how far Bayes's latent variable mechanism could be generalised without affecting the result, or in going wrong in the intricate arguments involved in generalising it. What is extraordinary is Pearson's lack of self-control, that he was prepared to accept and proclaim a conclusion that his experience should have told him must be wrong. All Egon Pearson (1938, p. 210) could say was that, "Perhaps it was due to a temporary lack of clearness in thought, a fault to which, I suppose, all of us succumb at times!"

During the 1920s Bayes was a presence in University College. The B-function campaign which opened with Soper (1922) and ended with the publication of the tables in Pearson (1934) involved numerous "computers and collaborators." One of the collaborators, Wishart (1927), put together a Pearsonian dish of quadrature and history before he moved away into Fisher's orbit. Pearson put the substance of the 1920 paper into his lectures and Egon (1990, p. 75) recalled that after the lectures of 1922, "Oscar Irwin and possibly others had ... concluded that K.P. had slipped up ..." Egon, who had recently started at University College, was more than a spectator. His paper "Bayes theorem, examined in the light of experimental sampling" (1925) went back to the origins and considered the evidence for Edgeworth's informative uniform distribution. Egon (1990, p. 75) looked back on the experience:

> It must have been in 1922 or 1923 that I started on the long piece of experimental 'counting' based on the rather ingenuous idea of exploring Edgeworth's statement that his justification for assuming a uniform distribution of prior probabilities between 0 and 1 was his own personal experience. Of course as W. F. Sheppard pointed out

> ... my collection of several hundred results, following a U-shaped distribution in fact 'proved nothing', because a subjective element had inevitably entered into the characters which I chose to count.

Egon would change his tack and in the early days of the Neyman-Pearson partnership it was he who was the cooler towards Bayes; see Reid (1982, pp. 78-85).

The world beyond University College was changing. Fisher's "Mathematical Foundations of Theoretical Statistics" (1922) attacked Bayes Theorem and provided an alternative in the form of maximum likelihood. It (1922, p. 310) cites the disagreement between Pearson and Edgeworth as evidence of the confused status of inverse probability. The "Foundations" also criticised Pearson's work on probable errors and on the method of moments. Pearson did not engage with the "Foundations." There was no lack of energy; Pearson was 65 years old but he was still very active and wrote more than 100 further pieces. There was a lack of understanding. Pearson may have seen no need to reply to Fisher on Bayes because Fisher was not interested in the prediction problem which was Pearson's main concern. He did reply to Fisher on the method of moments–in 1936–and to prepare he asked his son what "principle" was behind Fisher's work; Stigler (2005, p. 46) describes the episode.

11 One more "very fundamental" problem

Aside from putting the cooperative material into the *Tables* Part II (1931) Pearson last Bayesian effort was the 1928 paper, "On a method of ascertaining limits to the actual number of marked members in a population of given size from a sample." After the will-o'-the-wisp of the real Bayes' Theorem this was a return to business as usual, the evaluation of posterior probabilities.

The problem was "a very fundamental one in statistical practice" (1928, p. 149):

> Suppose a population to number N individuals, of whom p are actually marked by a special characteristic and q not so; thus $N = p + q$. Now suppose a sample n is taken of this population, and that in this sample r are found marked and s not so.
>
> ...
>
> We ... ask on the basis of this experience what is the likelihood of various values of p and q in the actual population N. In other words, knowing r and s, we seek the distribution of p and q.

The new paper complemented Pearson (1907) which had asked, given $n = r + s$, what will be the distribution of r' and s' in further samples of size $n' = r' + s'$. Now (1928, p. 149) as then, there was a "current method" based on the normal distribution:

> the usual method is to consider the percentage of marked individuals to be
>
> $$100\frac{r}{n} \pm 67.449\sqrt{\frac{rs}{n^3}}$$
>
> and to suppose that the probable error thus determined will measure in a rough sort of way the possible deviation of the sample value $100r/n$ from the actual, and unknown $100p/N$.

Again Pearson was dissatisfied with a result based on the normal approximation and one that neglected the size of the population sampled.

The problem was receiving attention from sample survey workers–see Hald (1998, pp. 289-99)–but Pearson was not part of this discussion. At the time he was studying Laplace–see Pearson (1929 and 1921/33)–and for background he (1928, p. 163) merely notes, "While attention has been frequently been paid to the problem of future expectancy on the basis of past experience I do not know of anyone but Laplace who has attempted the present problem on the basis of inverse probabilities." In an appendix he explains why Laplace's attempt was unsatisfactory but he does not say why the problem requires the use of inverse probabilities. The mathematics of the paper is straightforward and complements the 1907 analysis; it is described by Hald (1998, pp. 294-7) and Dale (1991, pp. 389-91). As usual, there is an extra-mathematical discussion and this time it takes the form of an appendix, "A note on the theory of inverse probabilities." This is not a return to 1907 and is once again a somewhat perplexing discussion: Hald (p. 298) laments, "What Pearson has in mind is difficult to see." The obvious rationale for such a discussion is to justify the use of inverse probabilities in the main part of the paper and this seems to be Pearson's intention: after the usual sparring with Venn, he (p. 165) concludes, "where our knowledge is based upon a single sample only ... we must either decline like Dr Venn to make any use of this knowledge ... or we must appeal to the principle of inverse probabilities to give us some measure of the accuracy with which we are describing the sampled population." Pearson muses over the assumptions of his pre-1920 work–the 1920-4 adventure is not mentioned. He argues that, while analysis along the lines of the correlation work of 1917 can be useful, it is not relevant to the "single sample" case; here he is echoing Student's objection. The "Edgeworthian doctrine" that justifies a uniform prior on the ground that it reflects experience is not justified when our experience is not of that kind. Pearson considers the notion that populations are sampled from a super-population but finds that it contains arbitrary assumptions. He keeps returning to the thought (p. 164) that, "It is difficult to conceive any more reasonable weighting, *based on a single experimental result*, than that which weights them [possible populations] with the probability that if they had existed they would have given the observed result." This sounds like the prior-less Bayes position of his undergraduate notes. He writes as though he had provided grounds for such a position and that he had provided a new argument for inverse probability.

The 1928 paper was a disappointing end to forty years of writing about inverse probability. Pearson went on doing the mathematics but was not able to explain why.

12 The *Grammar* again

I began with Harold Jeffreys and the discord he found between Pearson's philosophy of science and his statistical practice. What Jeffreys admired in the *Grammar*–see Jeffreys (1963, p. 407)–was its attitude that "laws are not established with certainty but can have a high degree of probability on the data" and its outlining a "theory of how this can happen." What puzzled Jeffreys about Pearson's statistical practice was that it did not follow the same principles. For Jeffreys's own principles-led work in statistics see Aldrich (2005).

I think that Jeffreys misread Pearson and that there was more coherence between Pearson's statistics and his philosophy than Jeffreys perceived. The arguments that impressed Jeffreys were not actually designed to show how laws are established but to show how belief in permanence is justified. Pearson transferred those arguments from the *Grammar* and physics to statistics in his 1907 paper "On the influence of past experience on future expectation." In the last year of his life Pearson wrote about laws and how they are established, treating implicitly, at least, the link between philosophy of science and statistics–the Jeffreys issues. Egon Pearson (1938, p. 234) describes the circumstances

> Finally, we may note two letters to *Nature* [Pearson (1935a, b)] and a last contribution to *Biometrika* [Pearson (1936b)] on a problem which, if the word is understood in its widest sense, may be termed the problem of graduation. Here he sought again to emphasise the difference between the world of concepts and the world of perceptual experience. It is the teaching of *The Grammar of Science*, most clearly seen in the letters, but to be read, too, behind the thrusts of the *Biometrika* article.

None of these publications mention Bayes' Theorem for they draw on a different side of the *Grammar*. The letters to *Nature* treat the significance test as a way of assessing the process of graduation and the final contribution to *Biometrika* treats Pearson's favourite method of graduation, the method of moments.

Pearson (1935a) expounded the logic of significance tests and of the χ^2 test, in particular, in response to a challenge from the scientist H. J. Buchanan-Wollaston. The exchanges in *Nature* consist of a letter from Buchanan-Wollaston, Pearson's reply, a response from Fisher and Pearson's rejoinder to Fisher; all are reprinted with a commentary in Inman (1994). Fisher's *Statistical Methods for Research Workers* (1925) had given significance testing far greater prominence and Pearson wanted to distinguish his use of tests from Fisher's; there is a hint too of dissatisfaction with the new Neyman-Pearson theory of testing.

Pearson (1935a, p. 4) made a string of points about the χ^2 test ("the P, χ^2 test"), why he had devised it and how he thought it should be used:

> (i) I introduced the P, χ^2 test to enable a scientific worker to ascertain whether a curve by which he was graduating observations was a reasonable 'fit'....
>
> (ii) As a measure of 'goodness of fit' the P, χ^2 test does enable one to compare the relative advantages of any graduation curves. But personally I have never assumed that the better graduation curve was the one from which the material had actually been drawn.
>
> ...
>
> (vi) From my point of view tests are used to ascertain whether a reasonable graduation curve has been achieved, not to assert whether one or another hypothesis is true or false. ... The fact is that all these distributions by mathematical curves in no case represent 'natural laws'. They have nothing in this sense to do with 'hypothesis' or 'reverse of hypothesis'. They are merely *graduation curves*, mathematical constructs to describe more or less accurately what we have observed.
>
> (vii) The reader will ask: "But if they do not represent laws of Nature, what is the value of graduation curves?" He might as well ask what is the value of scientific investigation! A good graduation curve–that is, one with an acceptable probability–is the only form of 'natural law', which the scientific worker, be he astronomer, physicist or statistician, can construct. Nothing prevents its being replaced by a better graduation; and ever bettering graduation is the history of science.

The first point is a little odd for, while Pearson (1900) does contain assessments of goodness of fit, it also appears to consider the truth of hypotheses, such as that a particular set of dice are fair; see Section 4 above. However, following Fisher's (1936) intervention, Pearson (1935b, p. 6) made it clear that "graduation" had to be interpreted very broadly:

> I should [also] term graduation the fitting of a binomial to a series of observations, or the determining whether a system of correlation coefficients could be reasonably supposed to have arisen from samples of material drawn from a population with a given correlation coefficient.

Apparently graduation covers everything that is done with data, except for prediction.

The second letter also has a passage–Pearson (1935b, p. 6)–elaborating points (vi) and (vii) which is very reminiscent of the *Grammar*

> The 'laws of Nature' are only constructs of our minds; none of them can be asserted to be true or to be false, they are good in so far as they give good fits to our observations of Nature, and are liable

> at any time to be replaced by a better 'fit', that is, by a construct giving a better graduation.

Science has two faces: graduation based on past data and the calculation of the future from the past. The two letters to *Nature* covered the first which involved methods of fitting and testing. They did not consider the second but then nobody had asked him about that.

13 Perspectives

We have tried to see what was behind that "repeated interest in problems of inverse probability" that Yule noticed in his life of Pearson. Pearson did not publish much on these problems–less than 10 publications from over 400–and yet he believed that they dealt with important issues: one purported to solve the "fundamental problem of practical statistics." An interest in inverse probability was not unusual: most of Pearson's contemporaries, like their illustrious predecessors Laplace and Gauss, used inverse arguments as well as direct. Pearson seems to have been less of a Bayesian–in the modern sense–than Edgeworth, Yule or Gosset. They were more or less inclined to view the inverse argument as the fundamental principle in statistics, the basis for all inference, while Pearson was less sure of what the true foundations should be, perhaps even whether there could be true foundations. On the other hand, he was sure that permanence, or stability, underlay the application of statistics and that the argument for permanence rested on Bayes' Theorem. His contemporaries–and successors–attached no special importance to this argument.

It is ironical that Jeffreys, the most organised and principled of Bayesian thinkers, should have read Pearson and been inspired by him; Edgeworth with his greater subtlety would have been a more productive read, though not so simply inspiring. Of course Jeffreys was perplexed to find that the principle that supported all his work on inference did not support Pearson's. Pearson did not commit himself to any fundamental principle, or even principles; he could contemplate different principles, e.g. minimum χ^2 and the "Gaussian test," without feeling compelled to get to the bottom of things. The continuity in Pearson's work came from certain enduring convictions about high-level matters, that the Pearson curves are valuable for representing data, that the method of moments is a useful method of graduation and that any projection of permanence beyond the observational record rests on an appeal to Bayes' Theorem. But there was a mass of activity only loosely connected with these fundamentals. It did not seem a matter of principle that inference for a finite population should involve inverse probability but not for an infinite population or that prior knowledge should be incorporated in inferences concerning correlation but not elsewhere. In this activity there was opportunism in the sense of an intelligent response to opportunities. Thus the opportunity Pearson exploited in using inverse probability for the finite population problem was the availability of the mathematics he had developed in his earlier work on Bayes' Theorem. Pearson's interest in that body of mathematics seems to have been a force in its own right in sustaining his interest in Bayes.

Pearson had an enigmatic position in the history of inverse probability for, while he defended inverse probability against Venn, he did more to bring about its downfall–or temporary eclipse–than Venn ever did. He did this by promoting methods like the method of moments which had no basis in inverse theory. The view of Fienberg (2006, p. 6) that inverse probability was "the method of choice of the great English statisticians of the turn of the century, such as Edgeworth and Pearson" never really applied to Pearson, or at least the visible Pearson. He may have started out that way in 1893 before he discovered that he could not make the method work on the dissection problem but afterwards inverse chances was just a method among others. Looking at Pearson's positive contribution, the small body of Bayesian writing, it is hard to find any great or lasting influence. His enthusiasm for Bayes' Theorem may have had an effect on Yule and Gosset but they channeled the enthusiasm into more orthodox directions. Pearson clearly influenced others, including Greenwood, Egon Pearson and Wishart, to take on Bayesian projects but the experience left no permanent impression on their work. Pearson's quest for the real Bayes was a fiasco. His successors in English and then in Anglo-American statistics divided between those who rejected all Bayesian arguments, led by Ronald Fisher and later Jerzy Neyman, and those who rejected all but Bayesian arguments, the solitary Harold Jeffreys. Victory went to the non-Bayesian side, to Pearson's side in the sense that his practice was overwhelmingly non-Bayesian but it was not a victory for Pearson or his ideas. The first system of non-Bayesian statistics was the theory of estimation of Fisher's "Mathematical foundations" (1922). This attacked Pearson's practice at several points but it was not an instance of one system confronting another for there was no Pearsonian system. It was Fisher who found systematic things to say about the method of moments and inverse probability–invariably damning things. Spectators like Neyman (1938, p. 11) might deprecate Fisher's efforts to discredit Pearson's work but they did not find in it a system they were prepared to defend.

14 References

Aldrich, J. (1997) R. A. Fisher and the Making of Maximum Likelihood 1912-22, *Statistical Science*, **12**, 162-176.

_________ (2003) The Language of the English Biometric School, *International Statistical Review,* **71**, 109-130.

_________ (2005) The Statistical Education of Harold Jeffreys, *International Statistical Review*, **73**, 289-308.

_________ (2006) Burnside's Encounters with the "Modern Theory of Statistics", Southampton University Economics Discussion Paper.

_________ (2008a) R. A. Fisher on Bayes and Bayes' Theorem, *Bayesian Analysis*, **3**, 161-170.

_________ (2008b) Keynes among the Statisticians, To appear in *History of Political Economy.*

_________ (2008c) Student's "Fundamentally New Approach to the Classical Problem of the Theory of Errors", In preparation.

Bayes, T. (1763) An Essay towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society,* **53**, 370-418.

Boole, G. (1854) *An Investigation of the Laws of Thought*, London, Walton and Maberly. (Reprinted by Dover, New York, 1976.)

Buchanan-Wollaston, H. J. (1935) Statistical Tests, *Nature*, **136**, 182-183. Reprinted in Inman (1994).

Burnside, W. (1924) On Bayes' Formula, *Biometrika,* **16**, 189.

Chauvenet, W. (1863) *A Manual of Spherical and Practical Astronomy, vol. 2*, 5th. edition, Philadelphia: Lippincott.

Dale, A. I. (1999) *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, second edition. New York: Springer-Verlag.

Edgeworth, F. Y. (1884) Philosophy of Chance, *Mind*, **9**, 223-235.

_________ (1885) Methods of Statistics, *Jubilee Volume, Royal Statistical Society*, pp. 181-217.

_________ (1908-9) On the Probable Errors of Frequency-Constants, *Journal of the Royal Statistical Society,* **71**, 381-397, 499-512, 651-678; **72**, 81-90.

_________ (1921) Molecular Statistics, *Journal of the Royal Statistical Society*, **84**, 71-89.

Edwards, A W. F. (1974) A Problem in the Doctrine of Chances, reprinted in the expanded edition (1991) of *Likelihood*, Baltimore: Johns Hopkins Press.

_________ (1978) Commentary on the Arguments of Thomas Bayes, *Scandinavian Journal of Statistics*, **5**, 116-118.

Eisenhart, C. P. (1974) Karl Pearson, *Dictionary of Scientific Biography*, **10**, 447-73. New York: Scribner.

Fienberg, S. E. (2006) When did Bayesian Inference become "Bayesian"? *Bayesian Analysis*,.**1**, 1-40.

Fisher, R. A. (1915) Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, *Biometrika*, **10**, 507-521.

_________ (1922) On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309-368.

_________ (1925) *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

_________ (1935) Statistical Tests, *Nature*, **136**, 474. Reprinted in Inman (1994).

Gauss, C. F. (1809) *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Perthes et Besser, Hamburg. *Werke*, **7**, 1-280. Translated by C. H. Davis as *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections.* Little, Brown, Boston, 1857. Reprinted by Dover, New York, 1963.

_________ (1816) Bestimmung der Genauigkeit der Beobachtungen, *Zeitschrift für Astronomie und verwandte Wissenschaften*, **1**, 185-196. *Werke*, **4**, 109-117.

Translated as "The Determination of the Accuracy of Observations" in David & Edwards (ed) (2001). The page references in my text are to this translation.

Greenwood, M. (1912) Contribution to the Discussion of Yule (1912), *Journal of the Royal Statistical Society*, **75**, 647-649.

_________ (1913) On Errors of Random Sampling in Certain Cases not Suitable for the Application of a "Normal" Curve of Frequency, *Biometrika*, **9**, 69-90.

Greenwood, M., and Yule, G. U. (1915) The Statistics of Anti-typhoid and Anti-cholera Inoculations, and the Interpretation of such Statistics in General. *Proceedings of the Royal Society of Medicine* (Epidemiology), **8**, 113-190.

_________ (1917) On the Statistical Interpretation of Some Bacteriological Methods Employed in Water Analysis, *Journal of Hygiene*, **16**, 36-54.

Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*, New York: Wiley.

Inman, H. F. (1994) Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from *Nature*, *American Statistician*, **48**, 2-11.

Jeffreys, H. (1939) *Theory of Probability*, Oxford, University Press.

_________ (1963) Review of *The Foundations of Statistical Inference* by L. J. Savage and others, *Technometrics*, **5**, 407-410.

Keynes, J. M. (1908) *The Principles of Probability*, submitted as Fellowship dissertation to King's College Cambridge December 1908.

Lamb, H. (1928) Obituary Notices of Fellows Deceased: Henry Martyn Taylor, *Proceedings of the Royal Society*, **117**, xxix-xxxi.

Laplace, P.-S. (1774) Mémoire sur la Probabilité des Causes par les Évènemens, translated by S. M. Stigler as "Memoir on the Causes of Events", *Statistical Science*, **20**, 32-49.

Magnello, M. E. (1996) Karl Pearson's Gresham Lectures: W. F. R. Weldon, Speciation and the Origins of Pearsonian Statistics, *British Journal of the History of Science*, **29**, 43-64.

_________ (1998) Karl Pearson's Mathematisation of Inheritance: from Galton's Ancestral Heredity to Mendelian Genetics (1895-1909), *Annals of Science*, **55**, 35-94.

_________ (1999) The Non-correlation of Biometrics and Eugenics: Rival Forms of Laboratory Work in Karl Pearson's Career at University College London, (In two Parts), *History of Science*, **37**, 79-106, 123-150.

De Morgan, A. (1838) *An Essay on Probabilities : and their Application to Life Contingencies and Insurance Offices*, London: Longman.

Neyman, J. (1938) A Historical Note on Karl Pearson's Deduction of the Moments of the Binomial, *Biometrika*, **30**, 11-15

Pearson, E. S. (1925) Bayes Theorem, Examined in the Light of Experimental Sampling, *Biometrika*, **17**, 388-442.

_________ (1936/8) Karl Pearson: An Appreciation of Some Aspects of his Life and Work, In Two Parts, *Biometrika*, **28**, 193-257, **29**, 161-247.

_________ (1965) Some Incidents in the Early History of Biometry and Statistics 1890-94, *Biometrika*, **52**, 3-18.

_________ (1967) Some Reflections on Continuity in the Development of Mathematical Statistics 1885-1920, *Biometrika*, **54**, 341-355.

_________ (1968) Some Early Correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with Notes and Comments, *Biometrika*, **55**, 445-457.

_________ (1990) *'Student', A Statistical Biography of William Sealy Gosset*, Edited and Augmented by R. L. Plackett with the Assistance of G. A. Barnard, Oxford, University Press.

Pearson, K. (1874-7) Lecture Notes on the Theory of Probability, held in Manuscript Room University College London Library list number 46.

_________ (ed.) (1888) *The Common Sense of the Exact Sciences* by W. K. Clifford. London: Kegan Paul, Trench.

_________ (1888) *The Ethic of Freethought*, London: T. Fisher Unwin.

_________ (1888) "The Prostitution of Science" Chapter II and pp. 33-53 of *The Ethic of Freethought*, London: T. Fisher Unwin.

_________ (1891) The Application of Geometry to Practical Life, *Nature*, **43**, 273-276.

_________ (1892) *The Grammar of Science,* London, Walter Scott.

_________ (1892/1941) The Laws of Chance, in Relation to Thought and Conduct: Introductory, Definitions and Fundamental Conceptions Being: the First of a Series of Lectures Delivered by Karl Pearson at Gresham College in 1892, First printed in *Biometrika*, **32**, 89-100.

_________ (1893) Asymmetrical Frequency Curves, *Nature*, **48**, October 26th, 615-616 & **49**, 6.

_________ (1893) Contributions to the Mathematical Theory of Evolution (Abstract), *Proceedings of the Royal Society,* **54**, 329-333.

_________ (1894) Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society A*, **185**, 71-110.

_________ (1894) Science and Monte Carlo, *Fortnightly Review*, **55**, 183-193. (Reprinted in Pearson (1897) vol. 1. Page references are to this re-print.)

_________ (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A*, **186** 343–414.

_________ (1896) Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society A*, **187** 253-318.

_________ (1897) *Chances of Death and Other Studies of Evolution*, 2 vols. London: Edward Arnold.

_________ (1899) On Certain Properties of the Hypergeometrical Series, and on the Fitting of such Series to Observation Polygons in the Theory of Chance, *Philosophical Magazine*, **47**, 236-246.

_________ (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling, *Philosophical Magazine*, **50**, 157-175.

_________ (1903) (Unsigned editorial) On the Probable Errors of Frequency Constants, *Biometrika*, **2**, 273-281.

_________ (1904) On the Theory of Contingency and its Relation to Association and Normal Correlation, *Drapers' Company Research Memoirs, Biometric*

Series, I. Cambridge University Press, Cambridge.

_________ (1906) On the Curves which are Most Suitable for Describing the Frequency of Random Samples of a Population, *Biometrika*, **5**, 172-175.

_________ (1907) On the Influence of Past Experience on Future Expectation, *Philosophical Magazine*, **13**, 365-378.

_________ (1911) *The Grammar of Science: Part I–Physical,* third edition, London, A. & C. Black.

_________ (1913) (Unsigned editorial) On the Probable Errors of Frequency Constants: Part II, **9**, 1-10.

_________ (1914) *Tables for Statisticians and Biometricians*, Cambridge: Cambridge University Press.

_________ (1915) (Unsigned editorial) On the Distribution of the Standard Deviations of Small Samples: Appendix I to Papers by 'Student' and R. A. Fisher, *Biometrika,* **10**, 522-529.

_________ (1920a) The Fundamental Problem of Practical Statistics, *Biometrika,* **13**, 1-16.

_________ (1920b) (Unsigned editorial) On the Probable Errors of Frequency Constants: Part III, **13**, 113-132.

_________ (1921) Note on the "Fundamental Problem of Practical Statistics," *Biometrika,* **13**, 300-301.

_________ (1921/33) *The History of Statistics in the 17th and 18th Centuries against the Changing Background of Intellectual, Scientific and Religious Thought: Lectures by Karl Pearson given at University College, 1921-1933.* Edited by E. S. Pearson (1978). London: Griffin.

_________ (1922) (ed.) *Tables of the Incomplete* Γ*-Function. Cambridge University Press,* Cambridge.

_________ (1924) Note on Bayes' Theorem, *Biometrika,* **16**, 190-193.

_________ (1928) On a Method of Ascertaining Limits to the Actual Number of Marked Members in a Population of Given Size from a Sample, *Biometrika,* **20A**, 149-174.

_________ (1929) Laplace, *Biometrika*, **21**, 202-216.

_________ (1931) *Tables for Statisticians and Biometricians, Part II. Cambridge University Press,* Cambridge.

_________ (1934) *Tables of the Incomplete Beta-Function. Cambridge University Press,* Cambridge.

_________ (1935a) Statistical Tests, *Nature*, **136**, 296-297. Reprinted in Inman (1994). Page references are to this reprint.

_________ (1935b) Statistical Tests, *Nature*, **136**, 550. Reprinted in Inman (1994). Page references are to this reprint.

_________ (1936a) Old Tripos Days at Cambridge, as Seen from Another Viewpoint, *Mathematical Gazette*, **20**, 27-36.

_________ (1936b) Method of Moments and Method of Maximum Likelihood, *Biometrika*, **28**, 34-59.

Pearson, K. & Filon, L. N. G. (1898) Mathematical Contributions to the Theory of Evolution IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation, *Philosophical Transactions of the Royal Society A*, **191** 229-311.

Porter, T. M. (2004) *Karl Pearson: the Scientific Life in a Statistical Age*, Princeton: Princeton University Press.

Pratt, J. W. (1976) F. Y. Edgeworth and R. A. Fisher on the Efficiency of Maximum Likelihood Estimation, *Annals of Statistics*, **4**, 501-514.

Reid, C. (1982) *Neyman–from Life*, New York: Springer.

Sheppard, W. F. (1899) On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation, *Philosophical Transactions of the Royal Society A*, **192**, 101-167.

Smith, K. (1916) On the "Best" Values of the Constants in Frequency Distributions, *Biometrika*, **11**, 262-276.

Soper, H. E. (1913) On the Probable Error of the Correlation Coefficient to a Second Approximation, *Biometrika*, **9**, 91-115.

_________ (1921) *The Numerical Evaluation of the Incomplete B-function, Tracts for Computers No. VII*, Cambridge: Cambridge University Press.

Soper, H. E., A. W. Young, B. M. Cave, A. Lee & K. Pearson (1917) On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of "Student" and R. A. Fisher, A Cooperative Study, *Biometrika*, **10**, 328-413.

Stigler, S. M. (1978) Francis Ysidro Edgeworth, Statistician, *Journal of the Royal Statistical Society, A*, **141**, 287-322.

_________ (1982) Thomas Bayes's Bayesian Inference, *Journal of the Royal Statistical Society, A*, **145**, 250-258.

_________ (1986) *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge MA: Harvard University Press.

_________ (2005) Fisher in 1921, *Statistical Science*, **20**, 32-49.

_________ (2007) Karl Pearson's Theoretical Errors and the Advances They Inspired, forthcoming in *Statistical Science.*

'Student' (1908a) The Probable Error of a Mean, *Biometrika*, **6**, 1-25.

_________ (1908b) Probable Error of a Correlation Coefficient, *Biometrika*, **6**, 302-310.

Stokes, G. G. (1887) *On the Beneficial Effects of Light*, London: Macmillan.

Thomson, W. & P. G. Tait (1869) *Treatise on Natural Philosophy volume 1*, Oxford: Clarendon Press.

Todhunter, I. (1858) *Algebra for the Use of Colleges and Schools*, Cambridge: University Press.

_________ (1865) *A History of the Mathematical Theory of Probability : from the Time of Pascal to that of Laplace*, London: Macmillan.

Venn, J. (1888) *The Logic of Chance*, first and second editions in 1866 and -76, London: Macmillan.

Welch, B. L. (1958) "Student" and Small Sample Theory, *Journal of the American Statistical Association*, **53**, 777-788.

Weldon, W. F. R. (1890) The Variations Occurring in Certain Decapod Crustacea.– I. Crangon vulgaris, *Proceedings of the Royal Society*, **47**, 445-453.

_________ (1893) On Certain Correlated Variations in Carcinus maenas, *Proceedings of the Royal Society*, **54**, 318-329.

Wishart, J. (1927) On the Approximate Quadrature of Certain Skew Curves, with an Account of the Researches of Thomas Bayes, *Biometrika*, **19**, 1-38.

Yule, G. U. (1911) *Introduction to the Theory of Statistics*, London: Griffin.

_________ (1912) On the Methods of Measuring Association Between Two Attributes, (with discussion), *Journal of the Royal Statistical Society*, **75**, 579-652.

_________ (1936) Karl Pearson 1857-1936, *Obituary Notices of Fellows of the Royal Society of London*, **2**, 74-104.

_________ (1938) Notes of Karl Pearson's Lectures on the Theory of Statistics, 1884-96, by G. U. Yule, *Biometrika*, **30**, 198-203.

Zabell, S. (1989) R. A. Fisher on the History of Inverse Probability, *Statistical Science*, **4**, 247-256.