Karl Pearson and the Scandinavian School of Statistics

Peter Guttorp and Georg Lindgren
University of Washington and Lund University

Abstract
The relationship between Karl Pearson and the Scandinavian statisticians was more of a competitive than a collaborative nature. We describe the leading statisticians and stochasticists of the Scandinavian school, and relate some of their work to the work of Pearson.

KEY WORDS:
Optimal design, genetics, survey sampling, Gram-Charlier series, jump process, telephone theory.

1. Karl Pearson and Scandinavia

Karl Pearson was infatuated with the Norwegian landscape (Porter (2004)). He went there on his honeymoon, and he learned enough Norwegian (from a Swedish family friend) to be able to read Ibsen in original. However, we have found no connections with Norwegian statisticians. On the other hand, there are several connections to Danish workers. For example, *Harald Ludvig Westergaard* (1853-1936), a professor of political science at the University of Copenhagen, visited University College in 1925 and gave two lectures on vital statistics. At the time, Pearson gave a formal dinner for the visitor, and commented on how well organized Westergaard's statistical laboratory was (Porter (2004) p. 290). Westergaard responded by being one of the speakers at Pearson's retirement dinner (Westergaard (1934)). In terms of their main areas of research, it seems to mainly overlap in terms of the history of statistics (e.g. Westergaard (1932), Pearson (1978)). Westergaard argued in his 1890 text that one should be able to divide any statistical material up into subgroups that are normally distributed, while Pearson of course used his system of curves to graduate data sets.

Among Pearson's students was one *Kirstine Smith (*1878-1939), whom he (in a letter to Fisher, Pearson (1968)) describes as "a student of Thiele's, one of the most brilliant of the younger Danish statisticians". Smith became Thiele's secretary after she finished her mathematics degree in 1903 until his death in 1910. She came to London in 1915 to work at the Biometric laboratory, where she produced a paper on minimum chi-squared estimation of the correlation coefficient (Smith, 1916). The same estimation idea had been put forward by Engledow and Yule (1914) for a different parameter, namely the recombination fraction in genetics, and by Slutsky (1913) for regression coefficients (Slutsky's paper had been rejected by Pearson; see section 5 of the paper by Seneta in this volume). Fisher did not like Smith's paper, and tried to publish a rejoinder, pointing out that the procedure would depend on the grouping used, but Pearson refused to accept it. This was one of two refusals that created the rift between Fisher and Pearson (Pearson 1968). Later Smith wrote a dissertation (published as Smith (1918)) inventing optimal design (Kiefer, 1959), where she computed G-optimal designs for polynomial regression of order up to 6, and explicitly calculated some of these designs. She also published work on fraternal and paternal correlation coefficients (important from Pearson's point of view to study natural selection; cf. section 3 below). After finishing her degree work she moved back to Denmark, and worked at the Carlsberg Institute for several years, before obtainingher teaching credentials and leaving research to become a high school teacher (Pearson (1990) p.124).[1]

It is interesting to note that when Swedish statisticians and economists in 1921 discuss contemporary British mathematical statisticians they mention Edgeworth (Cambridge) and Bowley (London), but not Pearson (Nordisk Statistisk Tidskrift[2] (1922)).

[1] Some additional information about Smith can be found at http://www.webdoe.cc/publications/kirstine.php (last accessed Dec. 4, 2008)

[2] There is no author for this description of a formal academic discussion, but presumably it was written by the editor *Thor Andersson* (1869-1935), a political economist and industrialist (Sjöström, 2008).

2. The Scandinavian School of Statistics

There are five names that we would like to put forward as representing the Scandinavian school of statistics around the turn of the century 1900: T.N. Thiele, J.P. Gram, A.N. Kiær, J. Hjort and C.V. Charlier. Two other names, F. Lundberg, and A.K. Erlang, fall in another category, representing the emerging stochastic process theory. Their work will be described later in this article. Schweder (1980) contains brief biographies of most of these workers, as well as a few more. Hald (2005) describes the leading Danish statisticians. The history of official statistics in Norway can be found in Lie and Roll-Hansen (2001), while Swedish statisticians are discussed in Sjöström (2002).

Many Scandinavian contributors to statistical thinking and stochastic modeling (to use an anachronistic terminology) in the late 19th and early 20th century worked for insurance companies. *Thorvald Nicolai Thiele* (1838-1910) combined his appointment as Director of the Copenhagen Astronomical Observatory with duties as chief actuary and mathematical director of the life insurance company Hafnia, of which he was a founder. He introduced cumulants (which he called "half-invariants") about 30 years before their rediscovery and exploitation by R.A. Fisher. Thiele published in 1889 what may be the first mathematical statistics textbook (Thiele(1889)). Also *Jörgen Pedersen Gram* (1850-1916) worked for the Hafnia Company, in addition to founding an insurance company of his own.

Gram developed the Gram-Schmidt orthogonalization (although discovered by Chebyshev, see the article by Seneta in this issue) and applied it to the derivatives of the normal density, leading to the series expansion, now named after him and the Swedish astronomer/statistician C V L Charlier (although developed by Laplace, Poisson and Bienaymé; cf. Hald, 2002). Also Thiele worked on this series. Pearson, in reading Westergaard's (1890) statistics text book heard about Thiele's work on these series expansions, promptly bought Thiele's (1889) book on the theory of observations, read it

(being able to read Norwegian also enabled him to read Danish to some extent) and used a data set from the book to compare his distribution theory to Thiele's expansion (Pearson, 1895). Some of Thiele's work is available in English, in particular three papers (and two historical comments) in Lauritzen's 2003 book.

The Norwegian *Anders Nicolai Kiaer* (1838-1919) was a statistician and the first director of Statistics Norway. In several sessions of the International Statistical Institute, from 1895, 1899, and 1903, he explained and advocated the early ideas of survey sampling, or "the representative method". A representative investigation should be regarded as "a photograph which reproduces the detail of the original in its true relative proportions". Kiaer's ideas included "stratified sampling" – information of the population should be used in the sampling design, the selection of units should be made objectively and according to a well defined protocol, and the reliability of the study should be reported. At the ISI session in Berlin 1903, the representative method was recommended – survey sampling became an officially accepted method! Interestingly, Kiaer stopped using his method after it had been criticized by a Norwegian mathematical statistician (Lie and Roll-Hansen (2001), Schwede (2003), both in Norwegian; for an English account, see Desrosière (2002)). This example of a gap between mathematical/probabilistic arguments and "statistical thinking" may well have delayed the development of a theory of survey sampling. For a discussion of Kiaer's (apparently limited) influence on sampling in Russia, see Seneta (1985).

As the fourth representative of the Scandinavian School we would like to mention the Norwegian *Johan Hjort* (1869-1948), actually a very prominent marine biologist in Oslo. His theories on the fluctuations in fishery are still the basis for fish resource management and they contain all the important elements in building and testing stochastic models. The following quotation from Schweder (1999) gives a lively account of the contribution:

> "...To fishermen, the great fluctuations in their catches is a problem, while to marine biologists the cause of variation was one of the great challenges early in the century. According to the migration hypothesis, the abundance of fish was practically unlimited, but due to variation in the migration pattern, catches would fluctuate. Another hypothesis was that fertile females were fished and

> consequently the production of eggs was hampered. Fish hatching was proposed as a solution to the problems, both with respect to harvest quantity and variability. The proponents of cod hatching were pressed to conduct experiments to prove their case. These proponents understood testing in this way, and concluded that hatching indeed improved matters. Hjort and his colleagues insisted on the experiment being controlled, and took a more sceptical approach in the interpretation. They actually argued convincingly that the proponents had capitalised on natural variability and over-interpreted the data in the favour of the hatching hypothesis. Hjort knew that enormous numbers of eggs were produced by each female, and that only a very small fraction of the eggs would develop to a cacheable fish. He also knew the variability from year to year of the environment for these eggs, larvae and juveniles, and developed the variable year class hypothesis. The idea of using demographic concepts like cohort, mortality etc was new in fisheries. Hjort had, however, developed a life insurance program for fishermen and took the demographic and statistical way of thinking to fisheries. To test his hypothesis, he needed methods to age the individual fish. Methods to count the year-rings on the scales on herring were developed and validated for ageing. Samples of size 25 were collected from as many schools of herring as possible over the years 1907 to 1913, and the yearly age distribution of mature herring was estimated. When putting his yearly age distributions on the same graph, a clear picture emerged. The herring stock was mainly made up of two strong year classes, one from 1899 and one from 1904. The distribution was bi-modal for the years 1907 to 1910, and then uni-modal from 1911 to 1913. The modes moved beautifully one year up for each year, with the class of 1899 basically disappearing at the age of 11. Hjort had a clear concept of hypothesis testing: "Could these results be due to randomness …, or are they due to a general law?" His variable cohort hypothesis was tested by excellent and convincing descriptive graphics. He did not use any probability. Hjort grouped his data by area, and found the same picture (a strong 1904 year class) in all areas. He also gathered data for 1914, and the age-distribution for that year was as predicted from his hypothesis."

Finally, in Sweden we find *Carl Vilhelm Ludvig Charlier*, (1862-1934), astronomer in Lund and director of the observatory from 1897, who writes the following in the preface of his 1910 book (based on material from 1905):

> "Mathematical statistics is the tool whose help enables the statistician to draw conclusions from his statistical material."

He regards "mathematical statistics is just as necessary for the statistician as the knife is for the surgeon". His student, *Sven Dag Wicksell*, (1890-1939), became in 1915 the first statistician in Sweden to get the academic title "docent in Mathematical statistics." He later became Professor of Statistics in Lund, and formulated the Wicksell stereological theorem.

In the preface of his 1910 book, Charlier pays special tribute to Karl Pearson, whom he calls "an outstanding scholar". However, while assuring his great admiration for Pearson, he continues:

> "Without wishing to undertake a detailed critique of his investigations, which, moreover, I most highly admire, I nevertheless believe it necessary to remark that the methods of Pearson possess an essential error, which consists in lacking sufficient generality both in the choice of the starting point and in the practical application".

Charlier then continues to criticize the Pearson distribution families as "unquestionably admirable formulae of interpolation; but they are derived without reference to the genetic development of such laws of error". In his own work on the Gram-Charlier series, he got help with some of the mathematics from the mathematician Marcel Riesz, although Cramér (1972) points out a fatal error in the proof of his main result.

Charlier wrote in German, and his work was rather early translated into English, which made him one of the best known Scandinavian statisticians for an international audience. Charlier, however, refers neither to Gram nor to Thiele in his 1910 book.

3. Issues of Genetics

Wilhelm Ludwig Johannsen (1857-1927), the foremost Danish geneticist around 1900, was one of the leaders of a group of biologists opposing the views of Galton and Pearson on inheritance. Johannsen was a professor of plant physiology at the Royal Veterinary and Agricultural University in Frederiksberg, north of Copenhagen, and later at the University of Copenhagen. He specialized in studying pure lines of a self-fertilizing bean plant, and invented some of the main terms in modern genetics, such as gene, phenotype, and genotype.

The conflict with Pearson and the biometric school started when Johannsen (1903) found a normal distribution of seed size in a population without genetic variability. The biometric school held that normal distributions in populations demonstrated gradual genetic variation on which Darwinian selection could act. Pearson (1903) and Weldon

and Pearson (1903) criticised Johannsen's statistical knowledge, and claimed (incorrectly, as it happens) that his interpretation of inherited variability should imply perfect correlation between characteristics of parents and offspring (Roll-Hansen (1983) has a thorough discussion of the conflict). Johannsen, on the other hand, held that the biometricians failed to distinguish between hereditary variation (variation in 'biological type') and 'fluctuating' variation (due to differing environmental influence).

While Pearson criticized Johannsen's work, Yule (1904) came to the Dane's defense, calling his work

> one of the most important contributions to the theory of heredity of recent years, and his results should be studied and judged in the original by all who are interested in the subject. The mode of treatment is novel, and the study of 'pure lines' a thoroughly sound procedure well calculated to elucidate the nature of intraracial heredity.

This contributed to the conflict between these two biometricians (Porter (2004) p. 273; Roll-Hansen (1983)). Johannsen actually wrote to Pearson and asked if he could visit and learn some statistics from him, but Pearson responded that he

> "could not hope to teach him anything in the first four days of term, or in any four days at any time."

In fact, in a letter to Weldon, Pearson (1905) exclaimed

> "I wish he would stick to Bateson and leave me alone!"

Pearson later (Pearson (1907)) made some rather disagreeable comments on some statistical comments by Johannsen on index numbers and cranial sizes.

Johannsen (1922) described Pearson's and his followers work:

> "...Pearson as leader of the 'biometric' school has continued [in the direction of Galton] using all the refined methods of higher mathematics; hereditary science and social statistics have flown together here, and correlation computations–using *Bravais*' formula as a starting point–are used extensively. Yes, heredity is defined as 'correlation between traits of parents and offspring'."

In Johannsen's own research, the idea was rather to use Mendelian tools to study the individual rather than the aggregate. He writes

> "Differences can drown in average relations, and these can create regularities which do not correspond to realities, when individuals or pure populations are considered. ... The biometric

> school has, almost defiantly, fought Mendelism and has thereby–quite indefensibly–dogmatically stuck to *Galton*'s word, that 'The science of heredity has more to do with combining sibling groups and larger populations than with studying the individual occasions.' This direction has thereby excluded itself form a deeper understanding of biological causes."

But Johannsen does by no means rule out statistical tools, which he used extensively in his own research. He writes

> "But we really do not want to avoid statistics! ... Here it is necessary that biologists never neglect the elements of statistical methodology; since only through them can one obtain clear expressions of the achieved results, and possibility of a closer criticism of the numerical relations in the materials studied."

Johannsen's 1913 book is described by Yule (1929, p. 361) as "Very largely concerned with an exposition of the statistical methods."

4. Other developments in the Scandinavian School

Lundberg, Erlang and Cramér are three Scandinavians who have had distinct but very different influences on probability and statistics. Their work on stochastic processes was closer to that of Yule than that of Pearson.

Enst Filip Oskar Lundberg, (1876-1965), was a Swedish forerunner in insurance mathematics and stochastic process theory. We quote from Cramér (1969):

> Filip Lundberg's work on risk theory were all written at a time when no general theory of stochastic processes existed, and when collective reinsurance methods, in the present day sense of the word, were entirely unknown to insurance companies. In both respects his ideas were far ahead of his time, and his works deserve to be generally recognized as pioneering works of fundamental importance.

Filip Lundberg's thesis from 1903 lies in time between Bachelier's (1900) and Einstein's (1905) works on continuous stochastic processes and Brownian motion, and treats processes with jumps caused by irregular claims. The Poisson process was given as a special case with jumps of equal size. He developed what was later to be known as the forward equations, and derived asymptotic distributions, with error bounds. In later works he introduced marked Poisson processes, studied extremes and tail behavior, and discussed barrier crossings.

Lundberg's work has had a long-standing reputation for being impossible to understand. It may perhaps be regarded as a lucky circumstance that *Harald Cramér*, (1893-1985), who started as a mathematician, working in number theory, got acquainted with Lundberg's work on risk theory after he took up a part time position at the Royal Insurance Board in 1919. One of Cramér's first duties on his new job was to explain some of Lundberg's work to the head of the board!

Cramér's professorship in Stockholm, inaugurated 1929, was named Insurance mathematics and mathematical statistics. Cramér did not interact with Karl Pearson, although he was well aware of his work (Cramér, 1981). In his unpublished memoirs[3] he writes:

> During the 20s and 30s so many new findings regarding statistical methodology had been published, particularly in England, where Karl and Egon Pearson (father and son), R. A. Fisher and Jerzy Neyman had an intensive production of novelties. I realized their great importance for applications, but felt very critical of their mathematics. Both Fisher and the two Pearsons seemed completely alien to the new probability theory which was founded upon the work of Russian and French mathematicians. I was tempted to try to produce a synthesis of the two lines of development.

Originally he was negotiating with Springer to publish the work in their yellow series, and mentions a table of content in German for the proposed book from 1937. However, the political developments in Germany made him reluctant to publish there, and the book (Cramér, 1946) was finally published after the war by Princeton University Press.

The third pioneer is *Agner Krarup Erlang,* (1878-1929), who worked as high school teacher in Copenhagen and other places, had 1904 taken up probability theory as a spare time exercise. Through the mathematician *Johan Ludwig Wilhelm Valdemar Jensen* (1859-1925), perhaps best known for Jensen's inequality, he was introduced to the managing director of the Copenhagen Telephone Company*,* and started to work for the company, applying probability theory to telephone traffic (Brockmayer et al. (1960)). His first paper on the subject came out 1909, and it dealt with the Poisson law for telephone

[3] H. Cramér (1978): *Korta minnen från ett långt liv*. Unpublished typescript.

calls, and waiting times in a telephone switch. Erlang, while first publishing in Danish, got his work translated into English, French, and German. Therefore his probabilistic approach to telephone traffic was soon recognized abroad. However, the arguments behind the results were not that easily understood, and often not even given in the paper. Cramér admits that he was not aware of the work by Erlang when he came across Lundberg's treatise, and started to work on the Poisson process.

References:

Bachelier (1900): Théorie de la spéculation. Ann. Sci. Ecole Norm. Sup. 17, 21-86.

Brockmeyer, E., Halstrom, H.L. and Jensen A. (1960): *The life and works of A.K. Erlang.* Acta Polytechnica Scandinavica, Mathematics and computing machinery series No. 6.

Charlier, C. V. L. (1910): *Grunddragen av den matematiska statistiken.* Lund: Statsvetenskaplig tidskrift. German editions 1920, 1931. English edition 1914.

Cramér, H. (1946): *Mathematical Methods of Statistics*. Princeton University Press.

Cramér, H. (1969): Historical review of Filip Lundberg's work on risk theory. Skand. Akt. Tidskr., 1969, suppl. 3-4, 6-12.

Cramér, H. (1972): On the history of certain expansions used in mathematical statistics. Biometrika, 59, 205-207.

Cramér, H. (1981) Mathematical probability and statistical inference. International Statistical Review, 49, 309–317.

Desrosière, A. (2002) Three studies on the history of sampling surveys: Norway, Russia-USSR, United States. Science in Context, 15, 377-382.

Einstein, A. (1905): On the Motion—Required by the Molecular Kinetic Theory of Heat—of Small Particles Suspended in a Stationary Liquid. Ann. Phys. 17, 549-560.

Engledow, F.L. and Yule, G.U. (1914): The determination of the best value of the coupling-ratio from a given set of data. Proc. Camb. Phil. Soc. 17, 436-440.

Erlang, A K. (1909): Sandsynlighetsregning og Telefonsamtaler. Nyt tidsskrift for Matematik B, 20,: 33-40. Later published in French: Calcul des probabilités et conversations téléphoniques. Revue Général de l'Electricité, 18, 1925.

Hald, A. (2002): *On the History of Series Expansions of Frequency Functions and Sampling Distributions, 1873-1944.* Matematisk-Fysiske Meddelser 49. Reitzel, Copenhagen.

Hald, A. (2005): *Nogle danske statistikeres liv og deres værker.* Matematisk-Fysiske Meddelelser 51. Reitzel, Copenhagen.

Johannsen, W. (1903): *Über Erblichkeit in Populationen und in Reinen Linien: ein Beitrag zur Beleuchtung schwebender Selektionsfragen*. Jena: Fischer. Summarized in English in Peters (1959), 20-26.

Johannsen, W. (1913): *Elemente der exakten Erblichkeitslehre.* 2te Ausgabe. Jena: Fischer.

Johannsen, W. (1922): Biologi og Statistik. Nordisk Statistisk Tidskrift, 1, 71-80.

Kiefer, J. (1959): Optimum experimental designs (with discussion). Journal of the Royal Statistical Society B, 21, 272-319.

Lauritzen, S. L. (2003): *Thiele: pioneer in statistics*. Oxford University Press.

Lie. E. and Roll-Hansen, H. (2001): *Faktisk talt. Statistikkens historie i Norge.* Oslo: Universitetsforlaget.

Lundberg, F. (1903): Approximerad framställning af sannolikhetsfunctionen. II. Återförsäkring af kollektivrisken. Uppsala: Almqvist & Wiksell.

Nordisk Statistisk Tidskrift (1922): Statistiken vid Sveriges universitet. Nordisk Statistisk Tidskrift, 409-477.

Pearson, E. S. (1968): Studies in the History of Probability and Statistics XX: Some Early Correspondence Between W. S. Gosset, R. A. Fisher and Karl Pearson, with Notes and Comments. Biometrika, 55, 445-457.

Pearson, E. S. (1990): *'Student'. A Statistical Biography of William Sealy Gosset.* Edited by R. L. Plackett and G. A. Barnard. Oxford University Press.

Pearson, K. (1895): Contributions to the Mathematical Theory of Evolution.–II. Skew Variation in Homogeneous Material. Phil. Trans. Royal Society London, series A, 186, 343-414.

K. Pearson (1903): Professor Johannsen on Heredity. Nature, 69, 149-50.

Pearson, K. (1905): Letter from K. Pearson to W.F.R. Weldon, 10 April 1905, Watson Library, University College London, Pearson Papers (625).

Pearson, K. (1907): Review of "Über *Dolichocephalie* and *Brachycephalie*. Zur Kritik der Index-Angaben. By W. Johannsen." Biometrika, 5, 482.

Pearson, K. (1978): *The History of Statistics in the 17th and 18th Centuries, Against the Changing Background of Intellectual, Scientific, and Religious Thought* (ed. E. S. Pearson). London: Griffin.

Peters, J. A. (1959): *Classic Papers in Genetics.* Englewoods Cliffs: Prentice Hall.

Porter, T. M. (2004): *Karl Pearson. The scientific life in a statistical age.* Princeton University Press.

Roll-Hansen, N (1983): The Death of Spontaneous Generation and the Birth of the Gene: Two Case Studies of Relativism. Social Studies of Science, 13, 481-519.

Schweder, T. (1980): Scandinavian Statistics, Some Early Lines of Development. Scand. J. Statist., 7, 113-129.

Schweder, T. (1999) Early statistics in the Nordic countries–when did the Scandinavians slip behind the British? Bull. Int. Statist. Inst. Tome LVIII. 52nd Session, Helsinki, Finland (CD, isi99.pdf).

Schweder, T. (2003): Statistikkens historie i Norge – faktisk uten statistikere? Tidskrift for Samfunnsforskning, 44, 309-318.

Seneta, E. (1985): A sketch of the history of survey sampling in Russia. J. Roy. Statist. Soc., Series A, 148, 118-125.

Seneta, E. (2009): Karl Pearson in Russian contexts. This volume.

Sjöström, O (2002): *Svensk statistikhistoria – en undanskymd kritisk tradition.* Hedemora: Gidlunds.

Sjöström, O. (2008): History of statistics 1650-1930 - Europe and
Sweden: An obscure critical tradition of social science statistics. Paper available at http://www.gidlunds.se/HistOfStat.pdf. Last accessed Dec. 7, 2008.

Slutsky, E.E. (1913) On the criterion of goodness of fit of regression lines, and the best method of fitting them to the data. J. Roy. Statist. Soc. 77, 78-84.

Smith, K. (1916): On the 'best' values of constants in frequency distributions. Biometrika, 11, 262-276.

Smith, K. (1918): On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika, 12, 1-85.

Thiele, T. N. (1889): *Almindelig Iakttagelselære: Sandsynlighetsregning og mindste Kvadraters Methode.*. Reitzel, Copenhagen. Translated into English in Lauritzen (2003).

Weldon, W. F. R. and Pearson, K. (1903): Inheritance in Phaseolus Vulgaris. Biometrika, 2, 499-503.

Westergaard, H. (1890): *Die Grundzüge der Theorie der Statistik.* Fischer, Jena. Translation of *Statistikens Theorie i Grundrids*, same year.

Westergaard, H. (1932): *Contributions to the History of Statistics.* King & Son, London.

Westergaard, H. (1934): MS of speech made in Pearson's honor. Watson Library, University College London, Pearson Papers (39).

Yule, G. U. (1904): Professor Johannsen on Heredity. Nature, 69, 223-24.

Yule, G. U. (1929): *An Introduction to the Theory of Statistics*, 9th edition. London: Griffin.