Version of March 4, 2006

Speaking of Graphics

Chapter 7

Quetelet, Galton, Pearson and Correlation

In order to sketch the history of multivariate data analysis, we must return to the last decades of the nineteenth century, when the concept of correlation was first developed. Correlation is a measure of association between two observed variables, such as between body stature and body weight, between level of education and general mental ability, between level of income and political preference, etc. It expresses how much of the variation in two observations (or measurements) is common. Correlation does not determine how much of the variation in one observation is caused by the variation in another, or vice versa. Correlation does not specify the direction from cause to effect. There even may be no direct cause-effect relationship between two correlated variables. It may well be, for example, that body stature and body weight both depend to some extent on the level of functioning of endocrinal glands (e.g., the thyroid), that level of education and mental ability are both to some degree conditioned by one's family, which in turn is determined in part by the level of education and by the abilities of its members.

7.1 (Lambert) Adolphe Quetelet (1796-1874) and the error curve

From the time of Auguste Comte [1], nineteenth century sociologists envied the exact scientists for their use of deterministic cause-effect relationships. The latter provide a mathematical description of the observed phenomena which allows one to make predictions and emit hypotheses which subsequently can be tested in further experiments. For example, when a body of mass m is submitted to a force F it is accelerated by an amount a according to Newton's law $F = ma$. This law can be tested repetitively in the laboratories, for example, by spinning a mass around and by measurement of the resulting centrifugal force under carefully controlled conditions. These would account for all possible sources of interference, arising from unequal distribution of the mass, from nonlinearity of the velocity and force sensors, etc.

In sociology and psychology no such deterministic laws had been found. Furthermore, their study often addresses the functioning and behavior of individuals or groups in their natural environment of which they are an interactive part. Around the time of the foundation of positivist philosophy, Adolphe Quetelet introduced his idea of a 'social physics' based on probability laws. This Belgian astronomer acquainted himself during a brief stay in Paris with the statistical work of Pierre Simon de Laplace. In particular, he tried to apply Laplace's ratio method for the purpose of estimating the number of inhabitants in Belgium. (The ratio method required the determination of the birth or death rate in a few selected areas and proposed to estimate the size of the whole population from the observed rate and from the total number of registered births or deaths in the country.)

After severe criticism of the ratio method by de Keverberg, who pointed to the many sources of inhomogeneity in various regions, Quetelet abandoned the idea. He remained fascinated, however, by the remarkable regularities that emerged from large collections of demographic data. Quetelet is now remembered mostly for his practice of fitting curves to data presented in the form of histograms, such as the heights and chest widths of army conscripts. He remarked that these curves strongly resembled to what was then known as the error curve (and which is now referred to as the normal probability density curve or Gaussian curve). These curves are symmetric with respect to the central mode which represents the average or mean value of the tabulated data. Quetelet concluded that, although there exists an infinite number of influences which are the cause of an individual observation, the average of a large number of them can be regarded as a constant. In his view, individual observations were scattered around the mean in the same way as repeated astronomical observations tend to err in a predictable way around its true value.

At last, sociologists got their hands upon observable quantities and a law that governed their manifestation, although the latter possessed a probabilistic, rather than a deterministic nature. In Quetelet's view, the manifestation of the error curve in a property of a group of individuals bore evidence that similar causes where operational on all the individuals of the group. Whenever there appeared a shift in the average value or in the shape of the curve this was then attributed to the influence of new or extraneous causes. However crude, Quetelet's approach had a great influence on the development of statistics, especially in England after the translation of his work on social physics in 1842, which originally had appeared in French [2]. Florence Nightingale was one of his great admirers and exerted her influence in academic circles to promote Quetelet's statistical method in England[3].

Notes on Quetelet

[1] Auguste Comte, Cours de Philosophie positive. Baillière, Paris, 1839.
Comte is the founder of positivist philosophy.

[2] (Lambert) Adolphe Quetelet, Sur l'Homme et le Développement de ses Facultés, ou Essai de Physique sociale. Bachelier, Paris, 1835.
An English translation appeared in 1842, which promoted Quetelet's statistical ideas in England during the second half of the nineteenth century : Adolphe Quetelet, A Treatise on Man and the Development of his Faculties. Chambers, Edinburgh, 1842.
See also the chapter on F. Nightingale in this book.

[3] A detailed historical account of the statistical work of Quetelet is found in :
Stephen M. Stigler, The History of Statistics. The Measurement of Uncertainty before 1900. The Belknap Press of Harvard Univ., Cambridge, Mass., 1986.

Biographical Notes on Quetelet (1796-1874)

----	Studies of mathematics at the University of Gent.
1823	Three-months visit to Paris, studies with Fourier and Laplace, as part of a project for a new observatory.
1824	Application of Laplace's ratio method for the indirect estimation of the size of a population from the birth and death rates in selected regions.
1835	Social physics. Elaboration of the concept of an Average Man and the practice of fitting normal curves to categorized data.
1842	Translation into English of his work on social physics ('A Treatise on Man and the Development of his Faculties.')

7. 2 Francis Galton (1822-1911) and correlation

7.2.1 Life and work of Galton

One of the scholars who were receptive to the ideas of Quetelet in England, was Francis Galton. His concept of regression and correlation in the study of inheritable properties was a great contribution to statistical thinking. Galton was an intellectually precocious child with an estimated IQ of the order of 200 . He studied mathematics and medicine at Cambridge. After his (unfinished) studies he earned a distinction from the National Geographical Society as an explorer of South-West Africa. The fortune he inherited allowed Galton to pursue his intellectual interests as a 'gentleman scientist'. According to his biographer [1], he combined an endless curiosity about the phenomena of nature with mechanical ingenuity and inventiveness. He designed a meteorological chart and investigated the effectiveness of prayer. In the last part of his active life he also developed a system of classification of individuals on the basis of fingerprints.

His main contribution, however, resulted from an extensive study of heredity, and for this, he is still best remembered today. Galton possessed a broad outlook and at the same time he was endowed with a profound sense for subtle detail, which seems to be the right stuff of which great scientists are made [1].

Galton made effective use of visual and graphical displays for the forging of his ideas and observations into a coherent system. The quincunx and the diagram of mid-parental and filial heights certainly meant more than a mere illustration of ideas already worked out. They developed along with the progression of thought, as instruments in the hands of a skilful and persistent investigator.

His studies of heredity led Galton to coin the term 'eugenics'. He was convinced that : 'A eugenic program to foster talent and healthiness and to suppress stupidity

and sickliness was a sine qua non in any society that wished to maintain, let alone promote, its quality and status.' In this he was opposed by those who believed that nurture-not-nature contributed to improvement of talent and health. He was also critisized by the Mendelians who maintained that inherited traits are transmitted by particles [2, 3]. When Galton died in 1911, he endowed by his will a chair of eugenics at London's University College, with the expressed wish that it should be offered to his friend Karl Pearson, who had championed and extended his ideas.

7.2.2 Regression to the mean

While many people stood in awe for the remarkable stability of the distributions of inheritable traits (such as body stature) from one generation to another, this was a great concern to Galton. In his view, if a physical characteristic or mental ability is transmitted across generations then one would expect the spread of the observations around their mean value to become broader and broader (while the mean value itself remains constant). The spreads around the mean were found to be remarkably constant, however, in the absence of external causes, as had already been demonstrated by Quetelet. This precisely was Galton's concern. To make his point, he designed a mechanical device which consisted of a board with about 20 rows of pins arranged in the form of a triangle (in the same way as numbers are arranged in Pascal's triangle). Small lead pellets (obtained from the shot in a hunter's cartridge) could be directed through a funnel to fall upon the apex of the triangle. The pellets ricocheted on the pins and through gravity collected at the base of the triangle where they piled neatly one upon another in the way of the error curve. This device, called 'quincunx' by Galton, simulates the transmission of an inheritable trait which in the first generation is highly concentrated (the loading of shot at the apex) and which becomes more and more diluted after each generation through the effect of random events (which force a pellet to bounce off from a pin either to the left or to the right). Theoretically, after an infinitely large number of generations (rows of

pins in the quincunx) the spread of the distribution also becomes infinitely large. Yet, in reality, the spread of the distribution of inheritable traits around their mean value appears to be remarkably constant. In terms of his statistical simulation, Galton postulated a mechanism which tended to counteract the progressive broadening of the flow of pellets, such as to revert them again within narrow bounds around the mean. He called this principle reversion or regression toward mediocrity. (Nowadays this effect is referred to as regression to the mean.)
We briefly describe three ways in which Galton investigated his problem, each more sophisticated than the other.

7.2.3 The regression of famousness in 100 families

In an early study of inheritance, he investigated 100 families, each of which had a very eminent or famous member. (The study only addressed the male kinship.) He counted the number of famous fathers, sons, grandfathers, grandsons, etc. in these 100 families. It was found that the number of famous persons decreased by about a factor four by each generation upward or downward from the most famous member [4]. In particular, 1/4 th is inherited from each parent, 1/16 th from each grandparent, etcetera, the sum over all ascendants being equal to one. Galton concluded that although talent seems to run in families, there also appeared a regression toward mediocrity.

7.2.4 The regression of size of seed in sweat peas

In order to obtain further evidence for this hypothesis, Galton designed an experiment with sweet peas [5,6]. He obtained sets of seven batches of increasing diameter, each composed of ten seeds with more or less identical diameters. He distributed the sets of parental seeds to his friends with detailed instructions for sowing and harvesting of the offspring. The results of his experiment are reproduced

in Table 7.2.1 in which the seven batches are labeled from K to Q in decreasing order of the parental diameter.

Label of batch	Diameter of parent seed	Mean diameter of filial seeds	
		Observed	Smoothed
K	21	17.5	17.3
L	20	17.3	17.0
M	19	16.0	16.6
N	18	16.3	16.3
O	17	15.6	16.0
P	16	16.0	15.7
Q	15	15.3	15.4

Table 7.2.1 Diameters (in .01 inches) of parental and filial (first generation) seeds in Galton's 1877 experiment with sweet peas. Each of the seven batches of parental seeds contained 10 seeds of almost identical size. According to Galton's analysis the average deviation of the filial diameters from the common mean of 15.5 is about one third of the average deviation of the parental diameter from the same mean. This phenomenon was interpreted as a regression towards the mean [7].

Galton found his law of regression confirmed in the seeds of the first generation progeny. After smoothing the data he determined a common mean at 15.5 (hundredths of an inch). Parental seeds of this size gave an offspring with the same mean diameter of 15.5 . Parental seeds that deviated positively from 15.5 produced progeny whose deviation from the common mean of 15.5 was reduced to a fixed proportion of the parental deviation, as can be seen from the following calculations :

$$\frac{17.3 - 15.5}{21.0 - 15.5} = 0.327$$

$$\frac{16.3 - 15.5}{18.0 - 15.5} = 0.320$$

Galton concluded from this result that the regression ratio between two generations

must be equal to 1/3 . The only disappointing fact was that the regression worked only well for the larger seeds but barely showed in the smaller ones. Here Galton remarked that 'the smaller parent seeds were such a miserable set that I could hardly deal with them. Moreover, they were very infertile' [7].

7.2.5 The regression of height in 200 families

The experiment with sweet peas described above has been designed as a stratified experiment. This means that from the outset each category of diameter contained an equal number of seeds. Therefore, it was not possible from this experiment to establish a law between the spread of the parental distribution of diameters and that of the first generation offspring. It merely confirmed (partially) the hypothesis of regression to mediocrity and it yielded an estimate of the rate of regression. For this reason, Galton undertook a new study in which he addressed a general population. He obtained data of stature (body height) on 205 parents and on 928 of their adult children. Galton choose body height as his object of inquiry because it is the sum of about 50 individual bones, numerous cartilages interposed between the bones and the fleshy parts of scalp and soles [7].

It was assumed that the transmission of stature from one generation to the next depended only on the height of the parents. The distribution of height was known to be stable and its measurement easily obtained. In a first step of preprocessing, heights of female parents and children were multiplied by a constant of 1.08 in order to make male and female heights comparable. Next, mid-parent heights were computed by averaging the heights of both parents. This allowed to crosstabulate the data for different categories of mid-parent height and adult children (or filial) height (Table 7.2.2).

Heights of the Mid-parents in inches	Heights of the Adult Children														Total Number of	
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children	Mid-parents
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205

Table 7.2.2 Crosstabulation of 928 adult children of various statures (heights) born of 205 parents of various statures. The mid-parent height is the average of the male and female heights. All female heights have been multiplied by 1.08 in order to make them commensurable with the male heights[7].

Finally, the data were smoothed and retabulated as shown in Fig.7.2.1 [8]. A highly remarkable feature of this figure is that it can be read at the same time as a table and as a bivariate diagram. When viewed as a diagram, the vertical and horizontal scales indicate the height of mid-parents and adult children, as well as their deviations from the median. The latter was determined to be equal to 68.25 inches in the two populations, as expected. The spread of the mid-parent heights is smaller by a factor of $\sqrt{2}$ than the spread of the adult children heights, again as expected [9]. These spreads amount to 1.22 and 1.70 inches respectively.

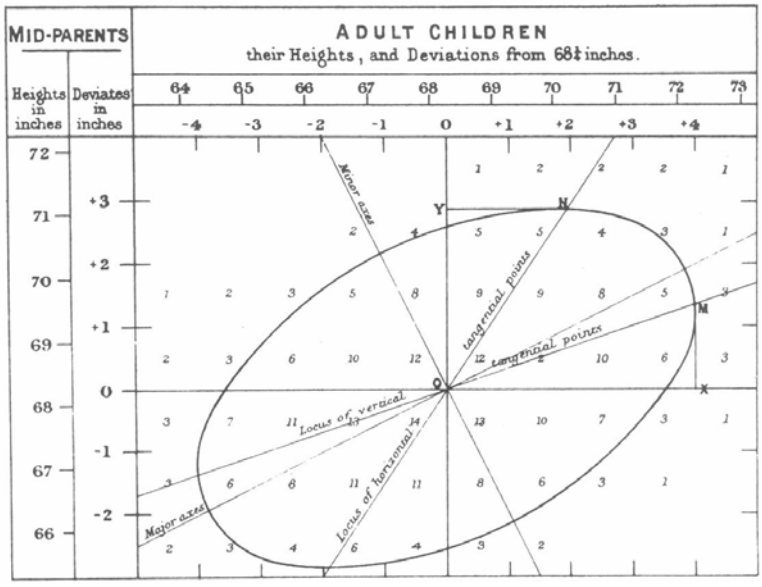

Figure 7.2.1 Smoothed, scaled and adjusted data obtained by Galton from the original Table 7.2.2. The data is crosstabulated by mid-parent height and by adult children height. At the same time they can be read as a bivariate diagram in which the coordinate axes represent deviates from the median height (which is equal to 68.25 inches for both mid-parents and adult children. The elliptic contour has been fitted to points of equal magnitude (close to a value of 4). The locus of horizontal tangential points ON estimates the probable filial height from a given mid-parental height. The tangent of this line with the vertical axis represents the rate of regression which was found to be equal to 2/3 [7].

On Figure 7.2.1 Galton fitted lines through points of equal magnitude much in the way as is done with isotherms and isobars on a geographic map. It was found that these equidensity contours formed concentric and symmetric ellipses. The center O of the ellipses coincided with the common mean of the mid-parent and filial distribution. Galton also indicated the two loci of vertical and horizontal tangential points. (Here, the locus is the collection of tangential points that can be drawn on all possible concentric and similar ellipses. In theory, only one ellipse needs to be considered, however, since the locus is a straight line through the origin.) The former is defined by the line OM passing through the origin O and through the vertical tangential point M. The latter is obtained by the line ON passing through O and the horizontal tangential point N. The line OM was fitted by Galton such as

to run through the maximal values of each column of the figure. The line ON was fitted such as to fit to the maximal values of the rows.

	Mid-parents	Adult children
Number	928	205
Mean	68.09	68.66
Standard deviation	2.54	1.94
Skewness	-0.065	0.127
Kurtosis	-0.218	0.124

Table 7.2.3 Descriptive statistics of Galton's data on parental and filial heights [7,9].

In a way, one could say that the locus of vertical tangential points defines the most probable midparent-height for a given adult children height. Similarly, the locus of horizontal tangential points prescribes the most probable adult children height for a given mid-parent height [10]. Galton found that the tangents of the angles subtended by the loci and the most proximate coordinate axes are 1/3 and 2/3 , respectively. The latter finding was of utmost importance, as it represented the rate of regression of height from one generation to another, so eagerly sought after by Galton. This indicated that the probable deviation from the mean (on either side) of the offspring was 2/3 of the deviation from the mean of the mid-parent height. Returning to the original problem, Galton devised a simple formula which related the rate of regression v , the constant spread of heights across all generations p , and the spread f produced by hereditary transmission at each generation (i.e., the widening of the pellet flow in the quincunx) :

$$v^2 \frac{p^2}{2} + f^2 = p^2$$

For every observed v and p, one could thus compute the unknown f. For example, in the case when $v = 2/3$ and $p = 1.70$ one finds that $f = 1.50$ [7].

7.2.6 A definition of correlation

A few years after the publication of the stature data, Galton realized that the two coefficients of regression became identical when both the mid-parental and children data were scaled to have the same spread. He then realized that the relationship between the two variables could be summarized by means of a single coefficient which he initially called co-relation, but later changed to the now familial term of correlation [11]. He proposed the symbol r, possibly referring to his concepts of reversion or regression toward mediocrity. Galton pointed toward the correlation as a consequence of the variation of two organisms which is partly due to common causes and he predicted its application to many diverse fields of investigation. The significance of this finding for the development of the human sciences is invaluable. It diverted the interest in cause-effect relationships towards the search for common causes of variation among correlated measurements. This heralded the advent of factor analysis, which Galton could not have foreseen. His interest shifted to other interests, such as the classification of fingerprints, while he left the statistical consequences of his discovery to his more mathematically oriented successors.

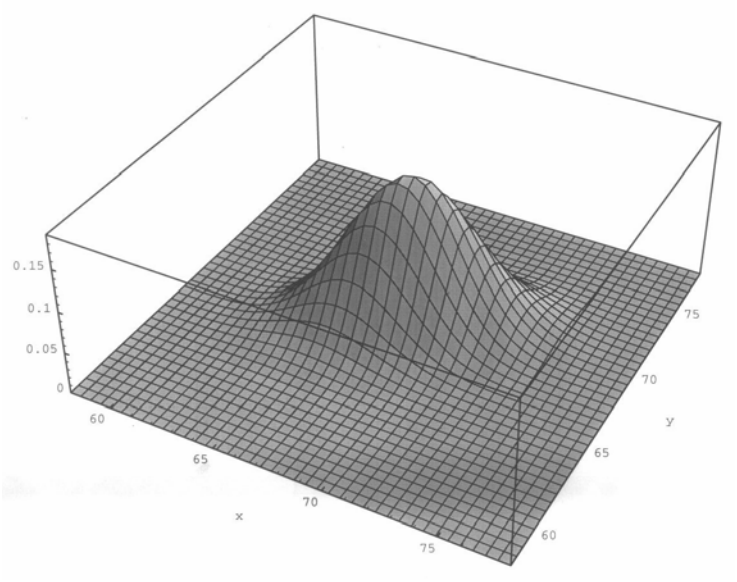

Figure 7.2.2 Bivariate normal distribution fitted to Galton's data on parental (x) and filial (y) heights, as reported in Table 7.2.2 [7].

Notes on Galton

[1] A detailed historical account of the statistical work of Galton is found in : Stephen M. Stigler, The History of Statistics. The Measurement of Uncertainty before 1900. The Belknap Press of Harvard Univ., Cambridge, Mass., 1986.

[2] Galton must have been aware of the work of his first cousin Charles Darwin, whose book 'On the origin of species by natural selection' appeared in 1859 (more than twenty years after the return from his voyage on the Beagle to the Galapagos). Darwin's book 'On the descent of man' appeared in 1871. Galton dismissed the idea of evolution, however, as being irrelevant to the study of heredity (in Regression towards mediocrity. Opus cit., 1885). He probably was also unaware of the results obtained by Gregor Mendel at Brünn in 1856 from genetic experiments with peas. Mendel's work was published in an obscure Austrian Journal and was only rediscovered by Hugo de Vries in 1900. Modern developments have made obsolete the practical results of Galton's research. Nevertheless, his statistical-graphical approach is still as valid today as it was in his own time.

[3] Norman T. Gridgeman, Francis Galton, Dictionary of scientific Biography. (Charles C. Gillispie, Ed.), Vol. 6, Ch. Scribner's Sons, New York, 1972, pp. 265-267. The particles that are inherited according to Mendel's particulate theory of heredity are now referred to as genes. It seems that Galton had played with the idea of latent and patent characteristics, which corresponds with our modern concepts of genotype and phenotype.

[4] Francis Galton, Hereditary Genius : An Inquiry into its Laws and Consequences. Macmillan, London, 1869.

[5] Gustav Hegi, Illustrierte Flora von Mitteleuropa. Band IV, 3. Teil, Dicotyledones. 2. Teil, Leguminosae - Tropaelauae. Paul Parey, Berlin, ..., 1975. Sweet peas (Lathyrus odoratus) reproduce by self-fertilization, i.e., by fusion of two haploid cells (which each contain a single genome).

[6] Francis Galton, Typical laws of heredity, Nature 15, 512-514, 1877.
The results of the experiments on sweet peas were also presented in the appendix of Galton's 1885 report of his analysis of the parental and filial heights of people.

[7] Francis Galton, Regression towards mediocrity in hereditary stature. J. Anthropological Institute, 15, 246-263, 1885. The experiment on sweet peas is described in the appendix. The main content of the paper deals with the analysis of the parental and filial body statures. The table of body statures contains a few suspicious entries. For example, the first row shows 4 adult children resulting from 5 couples within the smallest category of mid-parent height. (At least one parent couple did not raise an adult child.) The last row exhibits 14 adult children, apparently from the same couple with the largest mid-parent height. If one works out the ratio between the number of children and the number of couples, an inverse

trend is found between this ratio and the mid-parent height. The coefficient of correlation between the categorized mid-parental height and the number of adult children per parent pair equals -0.63 , which is significant at the 5 percent level of probability (p=.04). This suggests a bias in Galton's data such that taller parents tended to produce fewer off-spring. The bias may be artificial, however, as a result from the two discrepancies mentioned above.

[8] Francis Galton. Opus cit. , 1885.
The smoothed data was obtained from the original table. The smoothed value at the intersection of each row and column was first determined as the sum of the adjacent values. The result was then rescaled, rounded and adjusted such as to produce smooth progressions both in the horizontal and vertical directions of the table. Note that there is a likely error in the number at the intersection of the row at 1.5 deviations and the column at 0.5 deviations. The number should be between 10 and 12 . The error is perhaps made by the lithographer.
The graphical presentation of the data by Galton has been referred to as a data-based grid :
Edward R. Tufte, The visual Display of quantitative Information. Graphics Press, Cheshire, Conn., 1993, p.145.

[9] Galton used the deviations of the 25th percentile from the median (or the so-called 'probable error') as a measure of spread. This measure of spread is also called the semi-interquartile range. (The term 'standard deviation' still had to be defined by K. Pearson in 1893.) The descriptive statistics of parental and filial heights (Table 7.2.2) are presented in Table 7.2.3.
None of the measures for skewness and kurtosis reach statistical significance at the 5 percent level of probability. Hence, the stature data can be regarded as being normally distributed. The analysis was performed by means of :
SAS (Version 6.08) PROC UNIVARIATE, Statistical Analysis System. SAS Institute Inc., Cary, NC, 1985.

Under the assumption that the probable error p of the distribution of heights is the same in both individual parents and in their offspring, Galton derived the probable error p_m for the mid-parent height as :

$$p_m^2 = \frac{1}{4}\left(p^2 + p^2\right) = \frac{p^2}{2} \qquad \text{and hence :} \qquad p_m = p/\sqrt{2}$$

[10] In modern terms, one would now state that Galton had fitted a bivariate normal probability surface (or normal correlation surface) to the smoothed data. The equidensity contours of the distribution are the ellipses in the diagram. Under these conditions one can regard the loci of the tangential points as the least squares regression lines. Hence the rates of regression are the slopes of the regression lines. Nowadays we use the term regression for the process of fitting a function to observed data in order to make predictions of a dependent variable from one (or more) independent variables. The original meaning of the term in the context of hereditary regression toward mediocrity has been lost.

[11] Francis Galton, Co-relations and their measurement, chiefly from anthropological data. Proceedings of the Royal Society of London, 45, 135-145, 1888.
Francis Galton, Kinship and correlation. North American Review, 150, 419-431, 1890.

The story is admirably told by : Stephen Stigler. Opus cit., pp. 297-299.
The coefficient of correlation was also referred to as Galton's function. It was shown afterwards that the correlation is equal to the square root of the product of the two regression coefficients. In the case of Galton's stature data this leads to a value of $(2/9)^{1/2}$ or 0.471 .

It is possible to determine the coefficient of correlation directly from the regression diagram even when the two scales do not have equal spread. For example, in Fig.7.2.1 the correlation coefficient is equal to the ratio of XM to OY or equivalently to the ratio of YN to OX . This property has been discovered by my colleague, Ludo Gypen, at the Janssen Research Foundation. There are 13 other ways of determining the correlation coefficient from experimental data, but this is by far the most simple and elegant way :
Joseph L. Rodgers and W. Alan Nicewander, Thirteen ways to look at the correlation coefficient. The Am. Statist., 42, 59-66, 1988.
Ludo Gypen, Comment on Rodgers and Nicewander. The Am. Statist., 42, 291, 1988.

Biographical Notes on Galton (1822-1911)

----	Studies of medicine. Design of a meteorological chart. Takes part in anthropological expeditions in Africa.
1869	Study of inheritance of eminence in families.
	Formulation of the concept of regression towards mediocrity.
1877	Study of regression (or reversion) of hereditary traits in seeds of sweet peas.
1885	Study of regression in anthropological measurements (stature). Publication of a normal correlation surface for the average heights of parents and that of their adult children.
****	
1888	Definition of the term co-relation, which he changed later into correlation.
1889	'Natural Inheritance', a synthesis of his ideas and observations on transmission of hereditary traits.
	Classification of fingerprints.

7.3 Karl Pearson (1857-1936) and the mathematical definition of correlation.

7.3.1 Life and work of Pearson

The long and tortuous road from causal relationships to correlations led from Quetelet, via Galton to Karl Pearson and thus prepared the way for factor analysis which is the search for common causes of correlated phenomena. Quetelet was the first to apply statistical methods to demographic and sociological data. He found that, although individual behavior is subjected to random events, populations or large samples as a whole show remarkable regularity. This led him to believe that the appearance of a normal distribution in classified data was proof of common and constant causes that acted upon each individual member of the group. Galton adhered to this concept in his study of inheritance, but disagreed on the matter of constancy of causes, which he thought to be subject to reversion, or regression toward mediocrity, when studied from one generation to another. He devised a coefficient of reversion in order to characterize the relationship between inheritable and observable traits (such as anthropological measurements of height, chest width, etcetera). The coefficient was later renamed into coefficient of correlation which is still in use today. The publication of Galton's 'Natural Inheritance' araised the interest of Karl Pearson in statistics and their application to biological data. His life and work has been described with affection and detail by his son and successor Egon [1].

Pearson studied mathematics at Cambridge under the distinguished A. Cayley, J.C. Maxwell and G.G. Stokes. He went to Heidelberg to acquaint himself with physics and metaphysics. He also studied law but never practised it. During his student years he wrote philosophical essays about Spinoza, Maimonides, Luther and Marx. After marriage, he took part in debates about the 'Women's Question'. Throughout his later life he advocated 'free thought' in the sense of a religious, but undogmatic, knowledge : 'It is human thought which dictates the laws of the

universe... We have to look upon the universe as one vast intellectual process.' Pearson had an ebullient character, with great capacity for hard work and with extraordinary productivity.

At the age of 27, in 1884, Karl Pearson was appointed professor at University College in London where he lectured on applied mathematics and mechanics. A turning point in his career, as we have already mentioned, was his acquaintance with the work of Francis Galton. This opened new and broad avenues for man's intellectual exploration of the living world [2] : 'This part of the inquiry may be said to run along a road on a high level, that affords wide views in unexpected directions, and from which easy descents may be made to totally different goals...'. Pearson reflected on this passage [1]: 'I interpreted that sentence of Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and this new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment'.

During the period of 1891-1892 Pearson had accepted to lecture 12 times a year at Gresham College in London (a sort of Open University). His lectures covered the sciences, the theory of probability and its application to insurance, and statistical graphics, which was then called the geometry of statistics. Pearson made reference to the work of William Playfair whom he described as the father of geometrical statistics and the founder of the English school of political arithmetic. In his lectures, he emphasized the role of geometry as a fundamental method for investigating and analyzing statistical material [3]. He discussed the graphical train schedules (then called Bradshaw) devised by the French engineer M. Ibry, and the flow diagrams of Charles Minard that have become classics of statistical graphics [4].

During his life time, Pearson contributed to the foundation of statistical theory, especially with respect to the biological sciences. He founded and headed the Biometric Laboratory from 1901 to 1906 and edited the prestigious Journal of Biometrics. In 1911, after Galton's death, Karl Pearson was nominated as Galton professor of Eugenics. He founded the department of 'applied statistics', a term proposed earlier by Florence Nightingale, and which encompassed eugenics and biometrics. In 1933, Pearson retired and the department was split into Statistics and Eugenics, headed respectively by his son Egon Pearson and Ronald Fisher.

7.3.2 The definition of standard deviation

The contents of the Gresham lectures were published in extended form in 'The Grammar of Science' in which K. Pearson proposed a philosophical and ethical foundation of science. In this book he advocated freedom from dogmatic authority and direction from reason [5] : 'Our object in biology is the same as in physics, namely to describe the wide range of phenomena in the briefest possible formulae.' The Gresham lectures also stimulated Pearson to develop new statistical concepts, among which the definition of standard deviation as a measure of spread of measurements about their mean. Galton had defined spread as the probable error, by which he meant half of the interval between the 25th and 75th percentile of the distribution. In order to make his data independent of location and spread, Galton subtracted the median from each number and divided the result by this probable error. Pearson defined in 1893 the standard deviation s_x of measurements x as their root mean squared deviation from the mean $\bar{x}$, which reads in modern notation [6]:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - \bar{x}\right)^2 \quad \text{with} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where n is the number of measurements or observations. Using a rigorous

mathematical approach, Pearson re-examined Galton's anthropological data as well as other data that he collected himself. His first objective was to extract the correlation coefficient (or the Galton function, as it was then called) from bivariate data such as shown in Table 7.2.2 and Fig. 7.2.1 [7].

7.3.3 The mathematical definition of correlation

In 1896, Pearson provided a mathematical definition for his correlation coefficient which is also known as the product-moment correlation [8]:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x\, s_y}$$

where $\bar{x}$, $\bar{y}$ and s_x, s_y represent the means and standard deviations of the measurements x and y, respectively.
It can be verified that r_{xy} varies between -1 and +1. Perfect correlation results when x is proportional to y (plus or minus a constant). Perfect anticorrelation occurs when x is proportional to $-y$ (plus or minus a constant). Because the observations x and y are subject to random errors, the correlation coefficient r is itself a random variate. The exact distribution of r was derived by Ronald Fisher in 1915 [9]. This allowed to define upper and lower bounds for an estimated coefficient of correlation.

There are several variants of the correlation coefficient. One may replace the original variates x and y by their rank numbers. If one computes r by the formula above on the ranks of x and the ranks of y, this yields the rank order correlation, which is also called Spearman's r. It seems that Galton already had produced correlations of ranks before he had defined his correlation of variates [10]. So far

the coefficient of correlation (or Galton function) has been derived from fitting a bivariate normal distribution to the observed data (x and y) such as in Fig. 7.2.1 .

George Udny Yule (1871-1951), who was Pearson's student and assistant, pointed out in 1897 that the stringent assumption of bivariate normality of x and y can be replaced by the weaker assumption of linear relationship between x and y [11] . The laborious problem of fitting bivariate distributions is thus reduced to the much simpler one of fitting a straight line to x and y by means of the then known method of least squares. (The method of least squares had already been developed in the nineteenth century by Laplace, Legendre and Gauss [1].) Pearson initially objected to this idea as the thought that biological phenomena would not be amenable to simple linear relationships. Yule also considered the relationship of one variate with several others, and thus laid the foundation for multiple linear regression by means of the least squares method. This brought the biological sciences one more step nearer to the quantitative methods of the physical sciences.

7.3.4 Lines of closest fit to points in space

Another great contribution of Karl Pearson, which is relevant to the field of statistical graphics, is his fitting of lines and planes to a swarm of points in a Cartesian coordinate system [12]. Given a data table **X**, let's say with n rows and m columns, one can represent the row of the table as n points in an m-dimensional coordinate space. In the case when m equals 2, this results in a two-dimensional Cartesian diagram. We will restrict our discussion to this case, although the conclusions can be extended to any number of dimensions. Pearson considered the problem of drawing a line through the swarm of points such that the sum of squared distances from the line would be minimal. In Fig. 7.3.1, which is from Pearson's original publication, the points are labelled P_1 , P_2 , ... P_n and their distances to a

line AB are identified as p_1 , p_2 , ... p_n . The least squares criterion to be minimized here can be written as :

$$U = \sum_{i=1}^{n} p_i^2 \qquad \text{minimum}$$

From his familiarity with mechanics, a subject which he had taught before, Pearson knew that the expression U is the moment of inertia about the line AB . The line of minimal inertia always passes through the centroid of the swarm (i.e., the center of mass assuming that all points are endowed with unit mass). It is also the line for which the variance of the projections of the points upon it is maximal. (This can be proven by means of Pythagoras' theorem.)

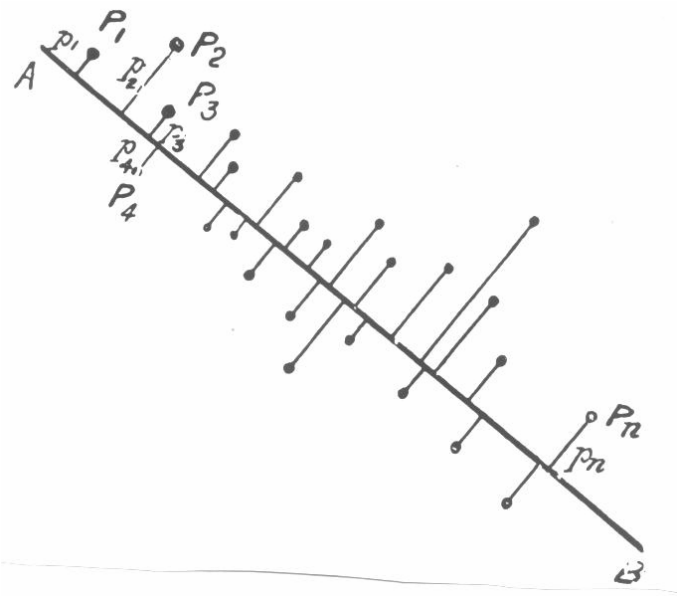

Figure 7.3.1 Line AB drawn through a swarm of points P_1 , P_2 ... P_n , with perpendicular distances p_1, p_2 ... p_n . The sum of squared distances represents the momentum of inertia of the points (assuming that these posses unit mass). The objective is to find a line such that the momentum is minimal [12].

Since the distances of the points to the centroid are fixed, minimization of the sum of squared distances in the perpendicular direction to the line AB automatically maximizes the sum of squared distances in the direction of the line. Pearson showed how to compute the parameters of the line which minimizes the moment about the line and, hence, maximizes the variance of the projections upon the line. He referred to this line as the line of closest fit to the points in the swarm. Figure 7.3.2 reproduces Pearson's original illustration in his 1901 paper [12]. The centroid C is defined by the means $\bar{x}$ and $\bar{y}$ of the two variates x and y. The line of best (closest) fit is indicated as AA'. The line BB' perpendicular to it, is called the line of worst fit. The inertia about BB' is maximal and the variance projected upon it is minimal. The two regression lines are labelled EE' (y on x) and FF' (x on y) for comparison. The regression line of y on x is used for prediction of y when x is determined without error, while the regression line of x on y serves to predict x when y is given and free of error. The line of closest fit describes the regression between x and y when both are subject to error [13]. Note that the three regression lines on Fig. 7.3.2 also appear on the correlation surface of Galton's diagram of parental and filial heights (Fig. 7.2.1).

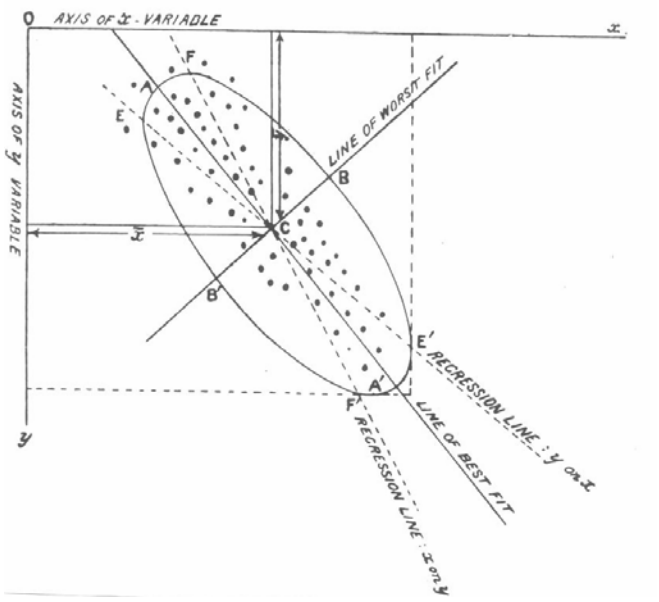

Figure 7.3.2 Line of closest fit AA' through a swarm of points in a two-dimensional coordinate space. The centroid C of the swarm has coordinates equal to the means $\overline{x}$ and $\overline{y}$. The line of closest fit possesses minimal momentum and accounts for a maximum of the variance of the data. The regression lines EE' and FF' are added for comparison. The line of worst fit is perpendicular to the line of best fit. It accounts for the residual variance in the data after accounting for the variance already explained by the line of best fit. The concept can be extended to multidimensional data [12].

Lines and planes of closest fit can be constructed in higher dimensional spaces. For example, in a three-dimensional space one can determine Pearson's line of best fit through the centroid of the points by means of the method of least squares. This is achieved by minimizing the sum of squared distances of the points perpendicularly to the line. Given this line, one can define a plane through the centroid and perpendicular to the line of best fit. The swarm of points can now be projected upon this plane. This way, one obtains a new swarm of points, this time in a two-dimensional space. Once again, one can determine the line of best fit in the residual space by the same method as outlined above. The two lines of best fit are by construction, perpendicular to each other. They define a plane of best fit. What remains is the line of worst fit, which is perpendicular to the plane of best fit. This

procedure can be extended to a multi-dimensional space, yielding hyperplanes of best fit. Each successive line of best fit accounts for a lesser amount of the variance in the data than its successors. The iterative process is continued until all the variance in the data is exhausted. This way, one obtains as many lines of best fit as there are dimensions in the original swarm of points. By construction, these lines are also the axes of inertia of the points (if one assumes that they are assigned equal masses).

In the case of elliptic or ellipsoidal swarms of points, the axes of inertia, which are also the lines of best fit, coincide with the axes of symmetry of the system of points. In the more general case the lines of best fit are referred to as the principal axes or principal factors of the swarm [14]. With this analysis, Pearson was well ahead of his time. It took several decades before it was realized that the lines and planes of closest fit of the 1901 paper correspond with the factors that account for the correlations between the variables. The path that led to this insight has been rather tortuous as will be explained in the next chapter on factor analysis.

Notes on Pearson

[1] Egon S. Pearson. Karl Pearson, An Appreciation of some Aspects of his Life and Work. Cambridge University Press, Cambridge, Engl., 1938.

Churchill Eisenhart. Karl Pearson, Dictionary of Scientific Biography. (Charles C. Gillispie, Ed.), Vol. 10, Charles Scribner's Sons, New York, 1974, pp. 447-473.

[2] Francis Galton, Natural Inheritance, MacMillan, London, 1889.

[3] Karl Pearson's emphasis on the role of geometry in statistical analysis of data is remarkable. The geometrical approach was discontinued, however, by Pearson's successors (Egon Pearson, Ronald Fisher) in favor of algebraic exposition. In more recent times there is a revival of the geometrical interpretation of statistical concepts, especially in the field of multivariate data analysis, which we address in subsequent chapters. It has also become fashionable nowadays to provide the algebraic formulation and the geometric representation of statistical concepts side-by-side. Pearson contrasted the English school of political arithmetic, with their emphasis in statistical charts, against the German school of state science (headed by Achenwall) and the French school of probability theory (founded by Laplace).

[4] A beautiful collection of the classic statistical diagrams is found in :
Edward R. Tufte, The visual Display of quantitative Information. Opus cit.

[5] Karl Pearson, The Grammar of Science. J.M. Dent, London, 1937. First published in 1882.

[6] The parameter s_x^2 is now called the variance of x . It is also the second moment of the distribution of x . In Pearson's original notation defined s_x^2 is defined in the form :

$$s_x^2 = \frac{1}{n} S\left(x^2\right) - \bar{x}^2 \quad \text{with} \quad \bar{x} = \frac{1}{n} S(x)$$

which is equivalent to:

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \quad \text{with} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where S represents a summation function over all elements of x . The formula is algebraically equivalent to the one shown in the main text, but is numerically more accurate when data is truncated or rounded to a fixed number of decimals.

[7] Karl Pearson, Mathematical contributions to the theory of evolution. III Regression, heredity and panmixia. Philos. Trans. R. Soc. London, 187A, 253-318, 1896.
In this paper, Pearson dealt with data that follow a bivariate normal distribution. He derived the algebraic form of the equidensity ellipse that appeared in Galton's paper of 1885 and which is reproduced in Fig. 7.2.1 :

$$\frac{x'^2}{s_x^2} + \frac{y'^2}{s_y^2} - 2r_{xy}\frac{x'\,y'}{s_x\,s_y} = c^2$$

where x' and y' are the deviations from the means of x and y, r_{xy} is the coefficient of correlation between x and y and where c is a constant. Figure 7.2.2 shows the bivariate normal distribution which is fitted to Galton's stature data, using the previously determined means and variances [4] and the coefficient of correlation derived by Galton [11]. The drawing has been produced by means of : Mathematica, A System for doing Mathematics by Computer. Wolfram Research Inc., Champaign, Ill., 1993.

Pearson noted that this result had already been derived nearly half a century ago by A. Bravais, although the latter did not define the coefficient of correlation, which can be credited entirely to Galton.

A. Bravais, Analyse mathématique sur les probabilités des erreurs de situation d'un point. Mémoires par divers Savans, Tome IX, Paris, 1846, pp. 255-332 .
The algebraic form of the elliptic contours was also confirmed in 1886 by J.D. Hamilton Dickson upon request by Galton. The results of A. Bravais were also reproduced independently by Francis Y. Edgeworth in 1892 in terms of Galton's coefficient of correlation. For more details on the history of the correlation coefficient one must read : Stephen Stigler, The History of Statistics. Opus cit.

[8] In Pearson's notation this would read as :

$$r_{xy} = \frac{\frac{1}{n}S(xy) - \bar{x}\bar{y}}{s_x\,s_y}$$

where S is a summation function which extends over all products between pairs of x and y, and where s_x and s_y are defined as above. The numerator in the expression is the product moment of the bivariate distribution of x and y.

[9] Rodriguez R.N., Correlation. Encyclopedia of Statistical Sciences, Vol. 2, J. Wiley, New York 1982, p. 196 .
Ronald A. Fisher derived in 1915 the so-called z-transform of the correlation coefficient r, which very rapidly approaches normality with increasing size of the sample n :

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

The mean of z is $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$, where r is the true (or population) mean. An approximate expression for the standard deviation of z is $1/(n-3)^{1/2}$.

[10] Rodriguez R.N., Correlation. Opus cit., p.197 and 199.
If x and y are normally distributed, then one may classify x or y or both into two disjoint (dichotomous) categories (e.g., larger than the median or not larger than the median). In this case, the Pearson correlation is referred to as the tetrachoric correlation. When only one of x or y is made dichotomous, then the Pearson correlation is called the biserial correlation.

[11] Some authors still believe that, in order for r to be valid, the underlying variables x and y should possess a bivariate normal distribution. Yule showed, however, that it is only required that x and y are paired, continuous and linearly related, e.g., in the form y = ax + b :

George Udny Yule, On the theory of correlation. J. Roy. Statist. Soc., 60, 812-854, 1897. The question is also discussed at length in :
Stephen Stigler, The History of Statistics. Opus cit., pp. 345-354 .

The assumptions of the correlation coefficient have also been discussed by :
M.D. Nefzger and J. Drasgow, The needless assumption of normality in Pearson's r . The Am. Psychol., 12, 623-625, 1957 .

[12]Karl Pearson, On lines and planes of closest fit to systems of points in space. Phil. Mag., Series 6, 2, 559-572, 1901.
The paper is also reprinted in facsimile in :
Bryant Edwin H. and Atchley William R. (Eds.), Multivariate statistical Methods : Within-Groups Covariation. Dowden, Hutchinson and Ross, Stroudsberg, Penn., 1975 (Distributed by Holsted Press - J. Wiley), pp. 17-30 .

[13] The regression line of y on x is obtained by ordinary least squares regression (OLS) which minimizes the distances of the points from the regression line in the direction of y . In a similar way, the regression line of x on y follows from minimization of the distances of the points from the regression line in the direction of x . In the construction of Pearson's line of best(closest) fit, the distances are perpendicular to the regression line, hence the name of orthogonal least squares regression. The principal axes can be obtained from a data table by means of an iterative procedure which will be described in greater mathematical detail in the chapter on factor analysis.

[14] Each iteration produces a principal axis and is followed by a projection of the points upon a (hyper)plane which is perpendicular to this axis. The projection in each iteration reduces the number of dimensions of the system of points by one. Iteration stops when no more dimensions are left. The result consists of as many principal axes as there are dimensions in the original swarm of points. Principal axes are mutually perpendicular and are produced in decreasing order of importance. The most important one accounts for the largest part in the variance of the data. The second principal axis explains the largest possible part of the residual

variation, in the (hyper)plane which is perpendicular to the first principal axis, etc. Principal axes have played an important role in factor analysis. They have been rediscovered thirty years after Karl Pearson's publication by the American factor analysts among which Louis Thurstone, who has been one of the prominent promoters of multiple factor analysis. The principal axes are also called principal factors or principal components, and form the basis of a popular method of exploratory data analysis, which is the subject of the chapters on multivariate data analysis.

Biographical Notes on Pearson (1857-1936)

----	Studies of mathematics at Cambridge under Cayley, Maxwell and Stokes.
1884	Appointed professor at University College in London, where he lectured on mathematics and mechanics.
1889	Acquaintance with Francis Galton following the publication of 'Natural Inheritance'.
1892	Presents the public Gresham Lectures in London on the Geometry of Statistics.
1892	'The Grammar of Science', a scientific and philosophical testimony.
1893	Definition of the term standard deviation.
1896	Definition of the product-moment correlation coefficient.
1900	Controversy between Mendelians and biometrists after the rediscovery of Gregor Mendel's work by H. de Vries.
1901	Founding of the Biometric Laboratory at University College and editorship of the Biometrics journal. 'Lines and planes of closest fit', first ideas on the principal axes or principal components of a system of points in space.
1911	Nominated Galton professor of Eugenics. Founding of the department of 'applied statistics', which included eugenics and biometrics.
1933	Pearson retires and the department of Applied Statistics is split into eugenics (Ronald Fisher) and statistics (Egan Pearson).