10 Galton and Regression

10.0 Texts

Stigler, Stephen(1986) *The History of Statistics : The Measurement of Uncertainty Before 1900*
The Empire of Chance : How Probability Changed Science and Everyday Life (Ideas in Context); Gerd Gigerenzer, et al
Hacking, Ian (1990) *The Taming of Chance* Cambridge University Press, Cambridge.
Pearson, Karl(1930) *The Life, Letters, and Labours of Francis Galton*, 3 vol.
Galton, Francis(1869) *Hereditary Genius* London, MacMillan
Galton, Francis(1875) *Statistics by intercomparison, with remarks on the law of frequency of error* Philosophical Magazine, 4th series 49:33-46
Galton, Francis(1877) *Typical laws of heredity*. Nature 15:492-495, 512-514, 532-533
Galton, Francis(1886) *Regression towards mediocrity in heriditary stature.* Journal of the Anthropological Institute 15:246-263

10.1 Galton the man

From the Britannica:

Sir Francis Galton (1822-1911)
English explorer, anthropologist, and eugenicist, known for his pioneering studies of human intelligence. He was knighted in 1909.

Early life. Galton's family life was happy, and he gratefully acknowledged that he owed much to his father and mother. But he had little use for the conventional classical and religious teaching he received in school and church. Indeed, he later confessed in a letter to Charles Darwin that the traditional biblical arguments had made him "wretched."

His parents had planned that he should study medicine, and a tour of medical institutions on the Continent in his teens--an unusual experience for a student of his age--was followed by training in hospitals in Birmingham and London. But at this time, in Galton's words, "a passion for travel seized me as if I had been a migratory bird." A visit to the University of Giessen, Germany, to attend lectures on chemistry was broken off in favour of travel in southeastern Europe. From Vienna he made his way through Constanza, Constantinople, Smyrna, and Athens, and he brought back from the caves of Adelsberg (present-day Postojina, Slovenia) specimens of a blind amphibian named Proteus--the first to reach England. On his return Galton went to Trinity College, Cambridge, where, as a result of overwork, he broke down in his third year. But he recovered quickly on changing his mode of life, as he did from similar attacks later.

Travels and exploration. After leaving Cambridge without taking a degree, Galton continued his medical studies in London. But before they were completed, his father died, leaving him "a sufficient fortune to make me independent of the medical profession." Galton was then free to indulge his craving for travel. Leisurely expeditions in 1845-46 up the Nile with friends and into the Holy Land alone were preliminaries to a carefully organized penetration into unexplored parts of southwestern Africa. After consulting the Royal Geographical Society, Galton decided to investigate a possible opening from the south and west to Lake Ngami, which lies north of the Kalahari desert some 550 miles east of Walvis Bay. The expedition, which included two journeys, one northward, the other eastward, from the same base, proved to be difficult and not without danger. Though the explorers did not reach Lake Ngami, they gained valuable information. As a result, at the age of only 31, Galton was in 1853 elected a fellow of the Royal Geographical Society and, three years later, of the Royal Society. In 1853, too, Galton married. There were no
children of the marriage. Galton wrote 9 books and some 200 papers. They deal with many diverse subjects, including the use of fingerprints for personal identification, the correlational calculus (a branch of applied statistics)--in both of which Galton was a pioneer--twins, blood transfusions, criminality, the art of travel in undeveloped

countries, and meteorology. Most of Galton's publications disclose his predilection for quantifying; an early paper, for example, dealt with a statistical test of the efficacy of prayer. Moreover, over a period of 34 years, he concerned himself with improving standards of measurement.

Advocacy of eugenics. Although he made contributions to many fields of knowledge, eugenics remained Galton's fundamental interest, and he devoted the latter part of his life chiefly to propagating the idea of improving the physical and mental makeup of the human species by selective parenthood. Galton, a cousin of Charles Darwin, was among the first to recognize the implications for mankind of Darwin's theory of evolution. He saw that it invalidated much of contemporary theology and that it also opened possibilities for planned human betterment. Galton coined the word eugenics to denote scientific endeavours to increase the proportion of persons with better than average genetic endowment through selective mating of marriage partners. In his Hereditary Genius (1869), in which he used the word genius to denote "an ability that was exceptionally high and at the same time inborn," his main argument was that mental and physical features are equally inherited--a proposition that was not accepted at the time. It is surprising that when Darwin first read this book, he wrote to the author: "You have made a convert of an opponent in one sense for I have always maintained that, excepting fools, men did not differ much in intellect, only in zeal and hard work." This book doubtless helped Darwin to extend his evolution theory to man. Galton, unmentioned in Origin of Species (1859), is several times quoted in Darwin's Descent of Man (1871). Galton's conviction that mental traits are no less inherited than are physical characteristics was strong enough to shape his personal religious philosophy. "We cannot doubt," he wrote, "the existence of a great power ready to hand and capable of being directed with vast benefit as soon as we have learned to understand and apply it."

Galton's Inquiries into Human Faculty (1883) consists of some 40 articles varying in length from 2 to 30 pages, which are mostly based on scientific papers written between 1869 and 1883. The book can in a sense be regarded as a summary of the author's views on the faculties of man. On all his topics, Galton has something original and interesting to say, and he says it with clarity, brevity, distinction, and modesty. Under the terms of his will, a eugenics chair was established at the University of London.

Reputation. In the 20th century Galton's name has been mainly associated with eugenics. In so far as eugenics takes primary account of inborn differences between human beings, it has come under the suspicion of those who hold that cultural (social and educational) factors heavily outweigh inborn, or biological, factors in their contribution to human differences. Eugenics is accordingly often treated as an expression of class prejudice, and Galton as a reactionary. Yet to some extent this view misrepresents his thought, for his aim was not the creation of an aristocratic elite but of a population consisting entirely of superior men and women. His ideas, like those of Darwin, were limited by a lack of an adequate theory of inheritance; the rediscovery of the work of Mendel came too late to affect Galton's contribution in any significant way.

10.2 Galton on the normal distribution

(From Stigler):

Following Quetelet, Galton proposed that the appearance of a normal distribution in data indicated the appropriateness of classifying the data together as a group. In Hereditary Genius(1869), which consists mainly of long lists of famous people and their relatives, he illustrates the normal law for height with a dot diagram, and writes:(p29)

It clearly would not be proper to combine the heights of dissimilar races, in the expectation that the compound results would be governed by the same constants. A union of two dissimilar systems of dots would produce the same kind of confusion as if half the bullets fired at a target had been directed to one mark, and the other half to another mark. Nay, an examination of the dots would show to a person, ignorant of what had occurred, that such had been the case, and it would be possible, by aid of the law, to disentangle two or any moderate number of superimposed series of marks. The law may, therefore, be used as a most trustworthy criterion, whether or no the events of which an average has been taken, are due to the same or to dissimilar classes of conditions.

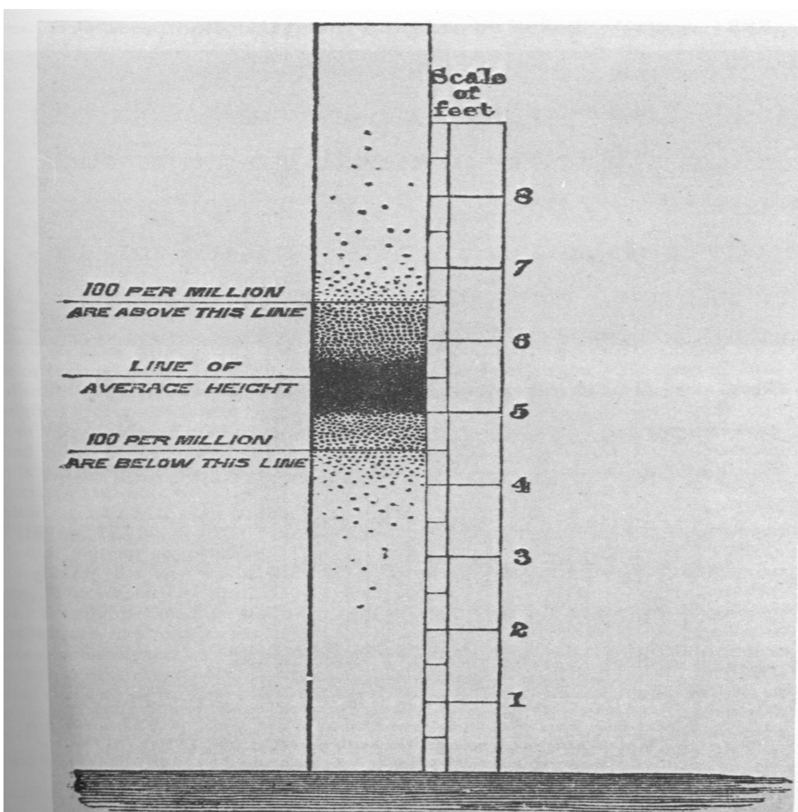

Figure 8.1. Galton's illustration of the "law of deviation from an average," showing the heights of a million hypothetical men. (From Galton, 1869, p. 28.)

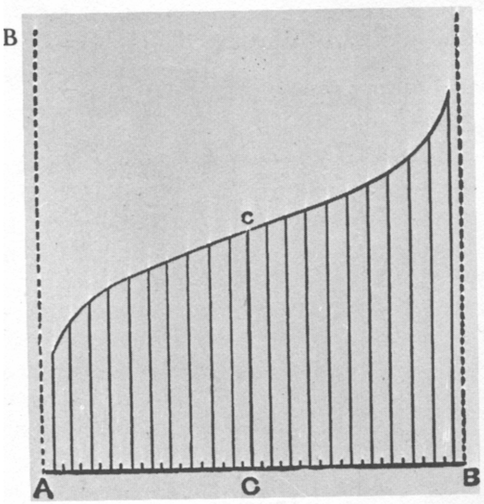

Figure 8.2. Two of Galton's renditions of the ogive. (A) The earlier, from 1875, shows the median and the quartiles (Galton, 1875). (B) The later drawing shows spikes with heights representing 21 equally spaced ideal data values (Galton, 1883, p. 51).

From Stigler, p 271

Even in 1869, though, Galton knew that recognizing a set of data as arising from the same species was the beginning, not the end, of an analysis, At one extreme he argued that "there is sufficient uniformity in the inhabitants of the British Isles to bring them fairly within the grasp of this law" (Galton, 1869, p. 29). Yet it was his basic aim to distinguish between Englishmen, to investigate the differences among them and the inheritance of these differences, not to lump them all together as a homogeneous whole, To this end, and in the first of his departures from Quetelet, Galton turned Quetelet's phenomenon to a novel use. If data from the same species arrayed themselves according to this curve and if the unity of the species could be demonstrated by showing that measurable quantities such as stature or examination scores followed the curve, then, once such a species was identified on the basis of measurable quantities, the process could be inverted with respect to qualities that eluded direct measurement! Qualities such as talent or "genius" that were at most susceptible to a simple ordering could, by Galton's method, be assigned a value on a "statistical scale." If a hundred individuals' talents were ordered, each could be assigned the numerical value corresponding to its percentile in the curve of "deviations from an average": The middlemost (or median) talent had value 0 (representing mediocrity), an individual at the upper quartile was assigned the value I (representing one probable error above mediocrity), and so on. Galton later called this method of analysis "statistics by intercomparison" (1874a, 1875), and it was to become the most used (and abused) method of scaling psychological tests (see, for example, Boring, 1920).

From Galton(1875,p 38),

Considering the importance of the results which admit of being derived whenever the law of frequency of error can be shown to apply, I will give some reasons why its applicability is more general than might have been expected from the highly artificial hypotheses upon which the law is based. It will be remembered that these are to the effect that individual errors of observation, or individual differences in objects belonging to the same generic group, are entirely due to the aggregate action of variable influences in different combinations, and that these influences must be (1) all independent in their effects, (2) aJI equal, (3) all admitting of being treated as simple alternatives 'above average' or 'below average;' and (4) the usual Tables are calculated on the further supposition that the variable influences are infinitely numerous.

10.3 Galton’s quincunx

The quincunx

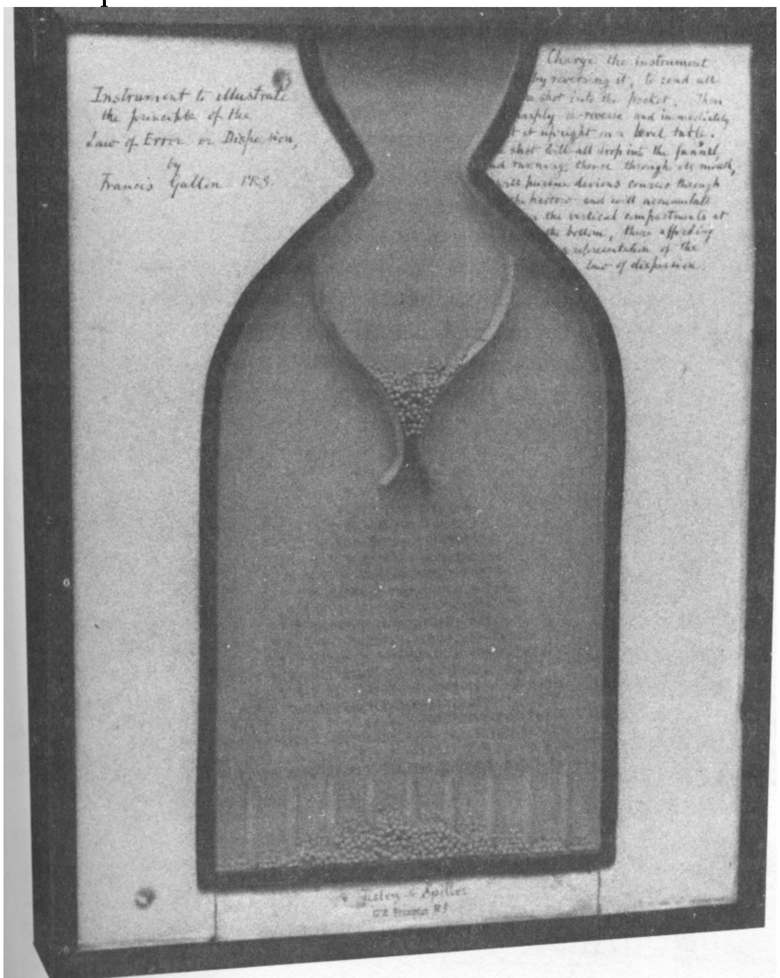

Figure 8.4. The original quincunx, apparently made for Galton in 1873 by Tisley & Spiller. Although it once had an opening at the top through which the shot could be poured, the top is now sealed with the shot inside. The glass has become cloudy with lead dust over the years. The caption, in Galton's handwriting, reads:

Instrument to illustrate
the principle of the
Law of Error or Dispersion
by
Francis Galton F.R.S.

Charge the instrument by reversing it, to send all the shot into the pocket. Then sharply re-reverse and immediately set it upright on a level table. The shot will all drop into the funnel, and running thence through its mouth, will pursue devious courses through the harrow and will accumulate in the vertical compartments at the bottom, there affording a representation of the law of dispersion.

Tisley & Spiller
172 Frampton Rd.

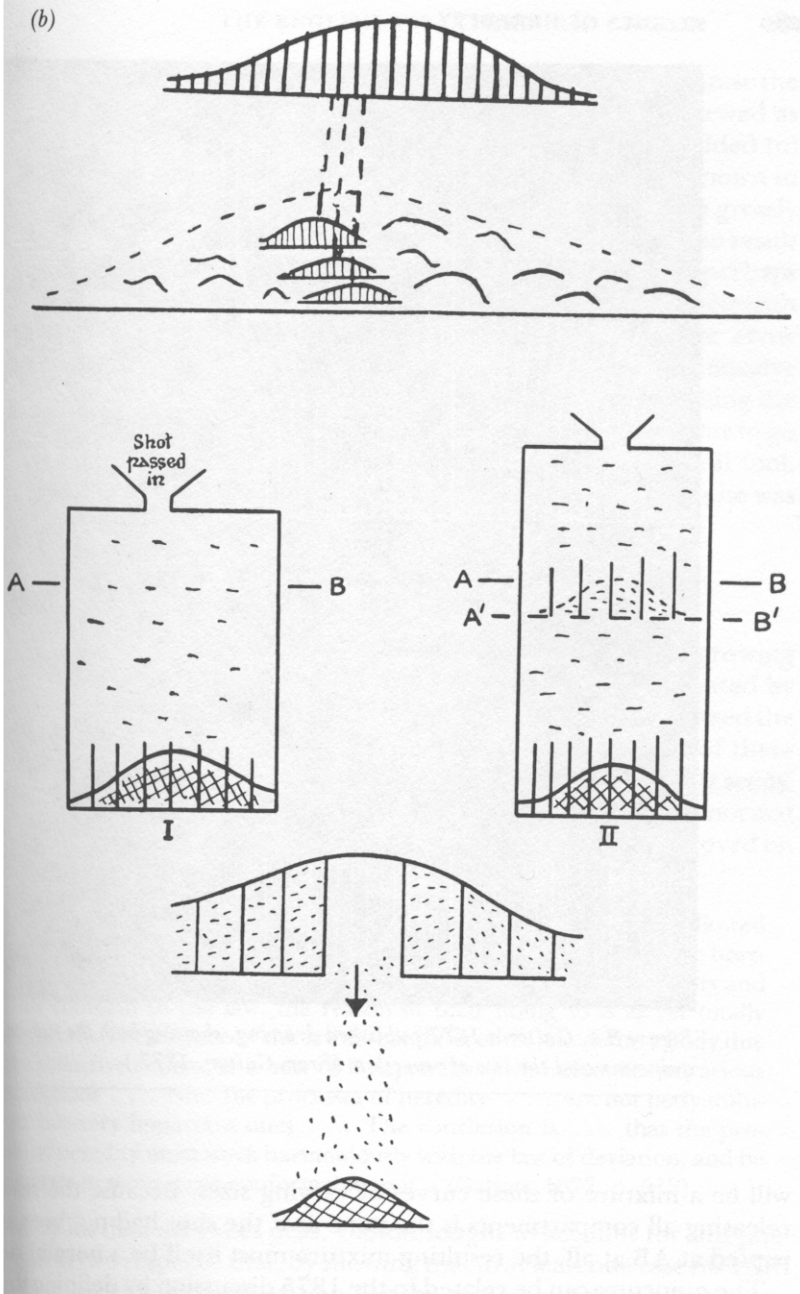

Figure 8.5. (a) Galton's explanation of the working of the two-stage version of the quincunx in a 12 January 1877 letter to his cousin, George Darwin. (From Galton Archives.) (b) Drawings by Karl Pearson, based on Galton's original letter. (From Pearson, 1914–1930, vol. 3B, pp. 465–466.)

10.4 Galton invents regression

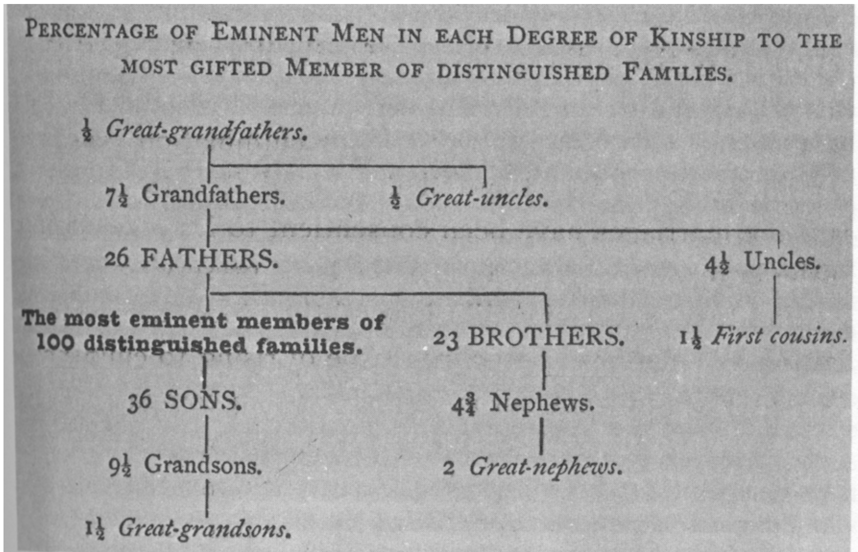

Figure 8.3. Galton's illustration of how the percentage of eminent men declines as the distance from the most eminent man in the family increases. Thus in 36 percent of the cases the son of the most eminent man was also eminent, whereas only 9.5 percent of the grandsons achieved eminence. (From Galton, 1869, p. 83.)

From Stigler,

In Heriditary Genius, Galton produced a table expressing a notion of regression. He found that selecting eminent men for study identified a peak in their families. " *The statistics show that there is a regular average increase in ability in the generations that precede its culmination, andas regular a decrease in those that succeed it*". He describes the phenomenon numerically in a table, and in " *Speaking roughly, the percentages are quartered at each successive remove, whether by descent of collaterally. Thus in the first degree of kinship the percentage is about 28; in the second around 7; and in the third, 1/12".*

10.5 The Galton Sweet Pea Experiment

In 1877, Galton divided sweet pea seeds into 7 packets by weight, and gave 10 seeds from each packet to 7 friends. How many experiments did Galton do on the way to St Ives? The answer is none, his friends had to do them all, Galton merely analysed them.

First let me point out a fact which Quetelet and all writers who have followed in his paths have unaccountably overlooked, and which has an intimate bearing on our work to-night. It is that, although characteristics of plants and animals conform to the law, the reason of their doing so is as yet totally unexplained. The essence of the law is that differences should be wholly due to the collective actions of a host of independent petty influences in various combinations. ..Now the processes of heredity ...are not petty influences, but very important ones. ..The conclusion is. ..that the processes of heredity must work harmoniously with the law of deviation, and be themselves in some sense conformable to it.

The seven groups of progeny were normally distributed, but with means different from their parents' weight. The average deviation of the progeny from the population average weight was 1/3 the deviation of the parents from the population average. Thus Galton discovered reversion, later called regression, toward the mean. This reversion is necessary if the overall population is not to become more variable over time. (There must be a certain relation between the overall population variance, the progeny variance, and the amount of reversion if the overall variance is to be preserved.)

Table 2. Raw Data on Diameters of Parent and Daughter Seeds Generated from Galton (1894) Table 2 in *Natural Inheritance*.

Diameter of Parent Seed (0.01 inch)	Diameter of Daughter Seed (0.01 inch)	Frequency
21.00	14.67	22
21.00	15.67	8
21.00	16.67	10
21.00	17.67	18
21.00	18.67	21
21.00	19.67	13
21.00	20.67	6
21.00	22.67	2
20.00	14.66	23
20.00	15.66	10
20.00	16.66	12
20.00	17.66	17
20.00	18.66	20
20.00	19.66	13
20.00	20.66	3
20.00	22.66	2
19.00	14.07	35
19.00	15.07	16
19.00	16.07	12
19.00	17.07	13
19.00	18.07	11
19.00	19.07	10
19.00	20.07	2
19.00	22.07	1

18.00	14.35	34
18.00	15.35	12
18.00	16.35	13
18.00	17.35	17
18.00	18.35	16
18.00	19.35	6
18.00	20.35	2
17.00	13.92	37
17.00	14.92	16
17.00	15.92	13
17.00	16.92	16
17.00	17.92	13
17.00	18.92	4
17.00	19.92	1
16.00	14.28	34
16.00	15.28	15
16.00	16.28	18
16.00	17.28	16
16.00	18.28	13
16.00	19.28	3
16.00	20.28	1
15.00	13.77	46
15.00	14.77	14
15.00	15.77	9
15.00	16.77	11
15.00	17.77	14
15.00	18.77	4
15.00	19.77	2

10.6 Galton discovers the bivariate normal distribution

In 1886, Galton cross tabulated mid-parent's height with the heights of 928 adult children, to produce a bivariate normal distribution and the regression lines we are now accustomed to, for the first time:(p 254-255)

I found it hard at first to catch the full significance of the entries in the table, which had curious relations that were very interesting to investigate. They came out distinctly when I "smoothed" the entries by writing at each intersection or a horizontal column with a vertical one, the sum of the entries in the four adjacent squares, and using these to work upon. I then noticed . ..[Figure 8.7]) that lines drawn through entries of the same value formed a series of concentric and similar ellipses. Their common centre lay at the intersection of the vertical and horizontal lines, that corresponded to 68 inches. Their axes were similarly inclined. The points where each ellipse in succession was touched by a horizontal tangent, lay in a straight line inclined to the vertical in the ratio of 2/3; those where they were touched by a vertical tangent lay in a straight line inclined to the horizontal in the ratio of 1/3. These ratios confirm the values of average regression already obtained by a different method, of 2/3 from mid-parent to offspring, and of 1/3 from offspring to mid-parent, because it will be obvious on studying. ..[Figure 8.7] that the point where each horizontal line in succession is touched by an ellipse, the greatest value in that line must occur at the point of contact. The same is true in respect to the vertical lines. These and other relations were evidently a subject for mathematical analysis and verification. They were all clearly dependent on three elementary data, supposing the law of frequency of error to be applicable throughout; these data being (1) the measure of racial variability, whence that of the mid-parentages may be inferred as has already been explained, (2) that of co-family variability (counting the offspring of like mid-parentages as members of the same co-family), and (3) the average ratio of regression. I noted these values, and phrased the problem in abstract terms such as a competent mathematician could deal with, disentangled from all reference to heredity, and in that shape submitted it to Mr. J. Hamilton Dickson, of St. Peter's College, Cambridge. I asked him kindly to investigate for me the surface of frequency of error that would result from these three data, and the various particulars of its sections, one of which would form the ellipses to which I have alluded.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	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	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

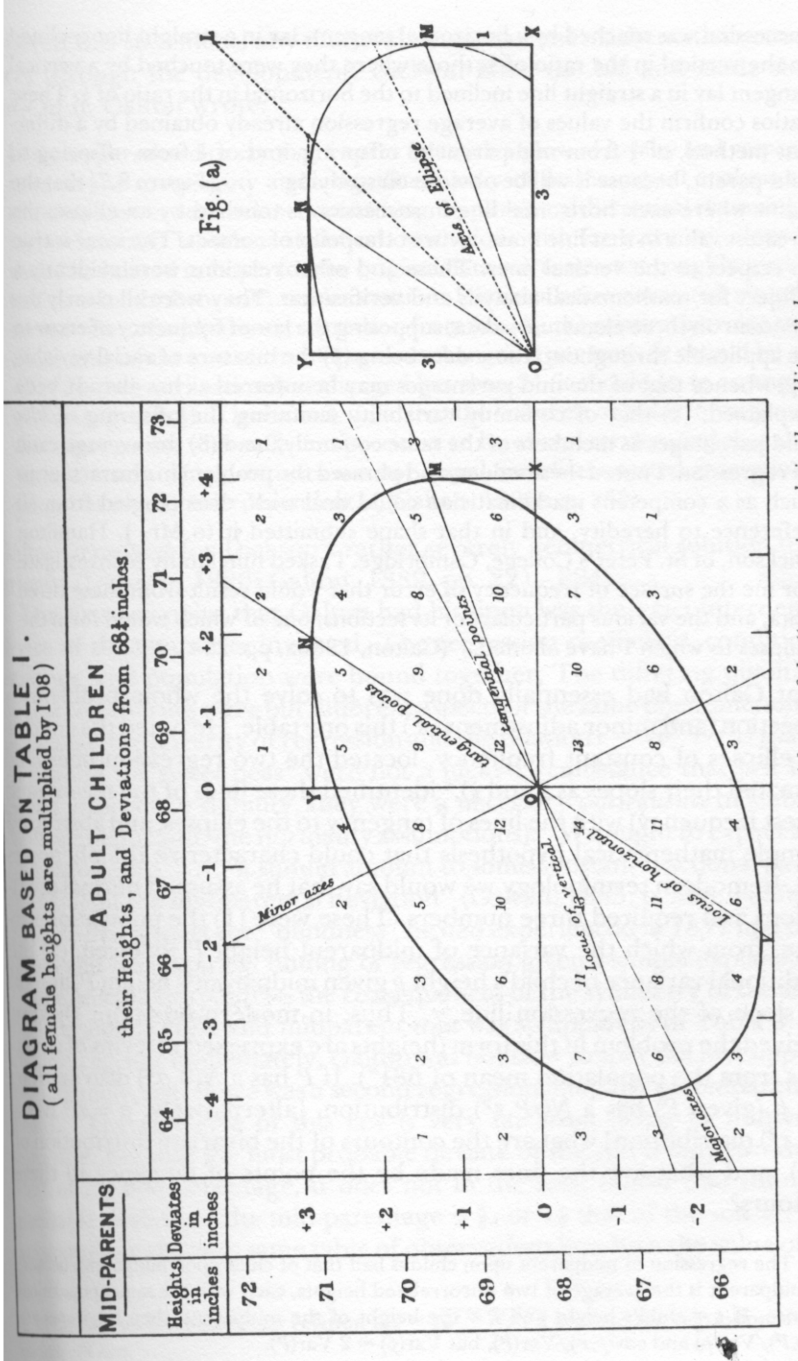

Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)