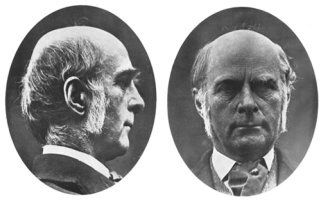
Hist. Stat. 6 The invention of correlation. Francis Galton (1822-1911), Karl Pearson (1857-1936)

Sir Francis Galton was a towering scientist of the Victorian age who did influential work in a wide variety of fields, including geography, meteorology, criminology, and psychology. He produced the first weather map. He invented the system for using fingerprints for criminal ID.



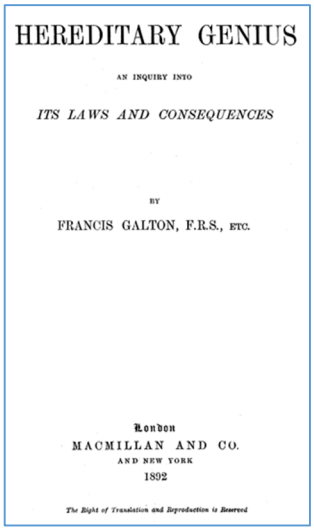
Photos of Galton from his anthropometric lab

He was first and foremost a student of heredity, developing statistical techniques to make sense of data in that field, especially methods of correlation and regression. "*The great stimulus for modern statistics came from Galton's invention of the method of correlation*, which, significantly, he first c onceived not as an abstract technique of numerical analysis, but as a statistical law of heredity." [Porter]

Galton firmly believed that all components of human nature and achievement were determined mainly by genetics, and he was a founder of the eugenics movement. (He, in fact, made up that word.)

Here is the introductory paragraph of his first important book, *Hereditary Genius* (1869).

> **I PROPOSE to show in this book that a man's natural abilities are derived by inheritance, under exactly the same limitations as are the form and physical features of the whole organic world. Consequently, as it is easy, notwithstanding those limitations, to obtain by careful selection a permanent breed of dogs or horses gifted with peculiar powers of running, or of doing anything else, so it would be quite practicable to produce a highly-gifted race of men by judicious marriages during several consecutive generations. I shall show that social agencies of an ordinary character, whose influences are little suspected, are at this moment working towards the degradation of human nature, and that others are working towards its improvement. I conclude that each generation has enormous power over the natural gifts of those that follow, and maintain that it is a duty we owe to humanity to investigate the range of that power, and to exercise it in a way that, without being unwise towards ourselves, shall be most advantageous to future inhabitants of the earth.**



HEREDITARY GENIUS

AN INQUIRY INTO

ITS LAWS AND CONSEQUENCES

BY

FRANCIS GALTON, F.R.S., ETC.

London
MACMILLAN AND CO.
AND NEW YORK
1892

The Right of Translation and Reproduction is Reserved

Galton thought that people like himself (for example, his older cousin Charles Darwin) were superior achievers largely because of heredity. His scholarly pursuits describing how humans could change over generations were influenced by Darwin's work on evolution.

> **I have no patience with the hypothesis occasionally expressed, and often implied, especially in tales written to teach children to be good, that babies are born pretty much alike, and that the sole agencies in creating differences between boy and boy, and man and man, are steady application and moral effort. It is in the most unqualified manner that I object to pretensions of natural equality. The experiences of the nursery, the school, the University, and of professional careers, are a chain of proofs to the contrary. *Hereditary Genius*.**

Galton was the first scientist to use the phrase "nature and nurture."

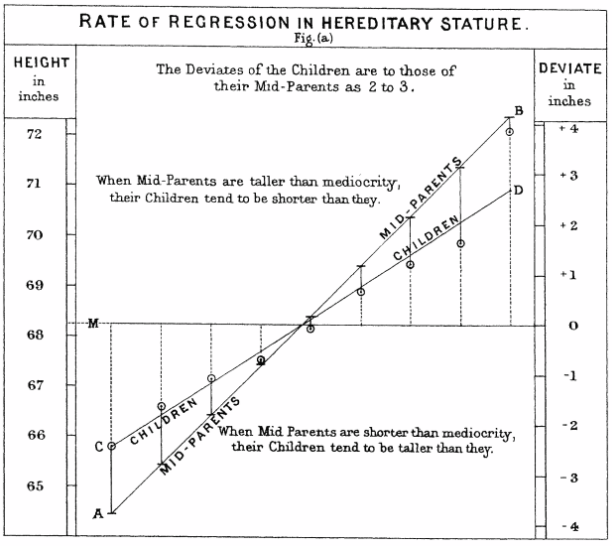
> **The phrase "nature and nurture" is a convenient jingle of words, for it separates under two distinct heads the innumerable elements of which personality is composed. Nature is all that a man brings with himself into the world; nurture is every influence from without that affects him after his birth. The distinction is clear: the one produces the infant such as it actually is, including its latent faculties of growth of body and mind; the other affords the environment amid which the growth takes place, by which natural tendencies may be strengthened or thwarted, or wholly new ones implanted. *English Men Of Science: Their Nature And Nurture* (1874)**

The Victorian era was characterized by the desire to explain almost everything by numbers. Research was intoxicated with measurement. To make his ideas "scientific," Galton derived statistical methods to *quantify* relationships between traits of related people, such as fathers and sons. Here is Galton, in 1888,

explaining correlation pretty much as we still introduce it today, except that he only considered positive correlation. Here is the first appearance of the word correlation in the scientific literature.

> **"Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity … but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.**
>
> **Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. Thus the length of the arm is said to be co-related with that of the leg, because a person with a long arm has usually a long leg, and conversely. If the co-relation be close, then a person with a very long arm would usually have a very long leg; if it be moderately close, then the length of his leg would usually be only long, not very long; and if there were no co-relation at all then the length of his leg would on the average be mediocre. It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be nil. Between these two extremes are an endless number of intermediate cases, and it will be shown how the closeness of co-relation in any particular case admits of being expressed by a simple number. *Co-relations and their Measurement, chiefly from Anthropometric Data.* (1888)**

In 1884 Galton created the Anthropometric Laboratory in London, a center for collecting data on people who volunteered (and even paid – 3 pence) to participate. He managed to get data on more than 10,000 people. A clear pattern he observed in these data is that the children of parents who lie at the tails of a distribution will tend to lie closer to the middle of the distribution. He made up the term "*reversion towards mediocrity*" for this phenomenon. [He later replaced "reversion" by "regression."] Today we call it ***regression to the mean***. The figure shows an example from data he had on parent and child height, where he used the mean of the two parents' heights as "Mid-parent" height. You can see the adult children whose parents were 3 inches above the mean were on average about 2 inches above the mean. In general, the deviation from the population mean for the children was about 2/3 the deviation of the parents.



from *Regression towards mediocrity in hereditary stature* (1886)

In deriving the mathematics associated with the correlation between two variables, Galton standardized the two sets of measurements; that is, he re-expressed them in terms of a common unit of spread. Today we use *z*-scores to achieve a common unit of dispersion, $z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{\text{deviation from mean}}{\sigma}$. Galton did not use σ to describe the dispersion of a normal distribution, but a measure called *probable error*, which is equivalent to .674σ. Based on his notation, Galton invented what we now call the ***correlation coefficient***. He called it "index of correlation," and symbolized it by the letter ***r*** from *r*eversion. Here he sums up his findings.

> **To conclude, the prominent characteristics of any two co-related variables, so far at least as I have as yet tested them, are four in number. … (1) that y = rX for all values of y; (2) that r is the same, whichever of the two variables is taken for the subject; (3) that r is always less than 1; (4) that r measures the closeness of co-relation. *Co-relations and their Measurement, chiefly from Anthropometric Data*. (1888)**

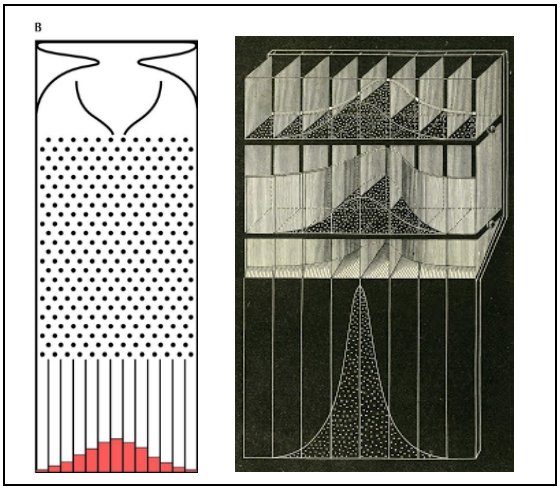
Galton's regression method is based on two essential factors. First, *human characteristics are assumed to be normally distributed,* and second, we compare individuals to a relevant *average*, not to some exterior standard. These were key ideas he got from earlier work by Quetelet.

> Our ordinary way of looking at individual differences is awry : thus we naturally, but wrongly, judge of differences in stature by differences in heights measured from the ground, whereas on changing our point of view to that whence the law of deviation regards them, by taking the average height of the race, and not the ground, as the point of reference, all confusion disappears, and uniformity prevails.
>
> It was to Quetelet that we were first indebted for a knowledge of the fact, that the amount and frequency of deviation from the average among members of the same race, in respect to each and every characteristic, tends to conform to the mathematical law of deviation.
>
> (*Typical Laws of Heredity,* 1877)

In his research to describe human and other natural phenomena, Galton decided to use the phrase "**normal distribution**" instead of the "error distribution," as the astronomers did. This reflects the attempt to describe what one "normally" observes. In this context he also replaced the word "error," as used in the physical sciences, with "deviation." In his influential book, ***Natural Inheritance***, Chapter 5 is called Normal Variability. Two quotes from that chapter:

> **I need hardly remind the reader that the Law of Error upon which these Normal Values are based, was excogitated for the use of astronomers and others who are concerned with extreme accuracy of measurement, and without the slightest idea until the time of Quetelet that they might be applicable to human measures.**
>
> **It has been shown that the distribution of very different human qualities and faculties is approximately Normal, and it is inferred that with reasonable precautions we may treat them as if they were wholly so, in order to obtain approximate results.**

Galton felt he needed to justify *why* it was sensible to assume normal distributions for genetic contributions, so he invented a clever device called the ***quincunx***, often now called Galton's bean machine. Pellets dropped down into it hit pins and bounced around randomly (mimicking random contributions to a given human characteristic), but always ended up making a normal distribution by the time the whole set piled up at the bottom. The figure shows a couple of schematics of his original apparatus. You can find working models in science museums today. "Quincunx" describes the arrangement of the pins – it's the name for the pattern of dots on the "5" face of a die.



The deeper, probabilistic justification for Galton's concept of correlation was provided by his younger friend and colleague, **Karl Pearson** (1857-1936), a significantly more adept mathematician and theorist. The *r* formula we use today is often called Pearson's *r*. Pearson made many other advances in statistical theory, as well. He was particularly successful in developing a categorization system for probability distributions, including non-normal distributions. He was first to derive the **chi-square distribution**, used to test goodness of fit for data that can be organized into categories – the so-called **chi-square test**. He derived the probability distributions for many hypothesis tests about standard deviations or variances. Pearson was the director of the Biometric Lab and the Applied Mathematics Departments at University College, London for many years. It was there that he helped W. S. Gosset develop the mathematics for the **t-test**. Karl Pearson was the preeminent British academic statistician for over 50 years. After his retirement, that role was next held by **Ronald Fisher** (1890-1962).

After the older Galton died in 1911, Pearson wrote his biography, published in three volumes (1914, 1924, 1930). Here is an excerpt from the preface to the last volume, written 20 years after Galton's death.

"It may be said that a shorter and less elaborate work would have supplied all that was needful. I do not think so ... I have written my account because I loved my friend and had sufficient knowledge to understand his aims and the meaning of his life for the science of the future. I have had to give up much of my time in the last twenty years to labour which lay outside my proper field, and that very fact induced me from the start to say, that if I spent my heritage in writing a biography it shall be done to satisfy myself and without regard to traditional standards, to the needs of publishers or to the tastes of the reading public." -- Karl Pearson *The Life, Letters and Labours of Francis Galton*

Exercises:

1. In fitting a normal curve to a data, Galton used the median of the data as the center, and the probable error as his descriptor of spread. The probable error is the value on the x-axis that cuts off the middle 50% of the distribution. Use a calculator or computer to find the z-score for the value of a normal curve that cuts off the middle 50% of the area. Choose any mean and any standard deviation you like. Explain why the answer does not depend on your choice.

Sources:

Porter, Theodore. *The Rise of statistical thinking: 1820-1900*. Princeton, 1986
Stigler, Stephen. *The Seven Pillars of Statistical Wisdom*. Harvard, 2016
Weisberg, Herbert. *Willful Ignorance*. Wiley, 2014
Galton, Francis. *Hereditary Genius*. Macmillan, London 1869
Galton, Francis. *Natural Inheritance*. Macmillan, London 1889
Galton, Francis. Co-relations and their Measurement, chiefly from Anthropometric Data. (1888) *Proceedings of the Royal Society of London*
Galton, Francis. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* (1886)
Plackett, R.L. Karl Pearson and the Chi-Squared Test. *International Statistical Review*, 51 (1983)
Images of Galton and Pearson. www.Galton.org
Image of Quincunx. http://evolution-textbook.org/content/free/figures/28_EVOW_Art/23_EVOW_CH28.jpg
Image of Quincunx. Stigler, S. .Galton Visualizing Bayesian Inference. *Chance* 2011