Karl Pearson and “Applied” Statistics[*]

Radhendushka Srivastava

Karl Pearson’s statistical innovations have led to the development of several mathematical statistics techniques. His mathematical contributions to the theory of evolution include a family of univariate probability distributions (referred to as Pearson’s family of distributions), which is found to be useful even today for fitting on data arising from various scientific disciplines (physics, biology, anthropology, economics, etc.). Frequency curve (histogram) for continuous data, a commonly used non-parametric density estimator, owes its origin to him. Pearson’s correlation coefficient is a popular descriptive measure of the linear relationship between two continuous random variables. This article aims to highlight a few of his fundamental ideas.

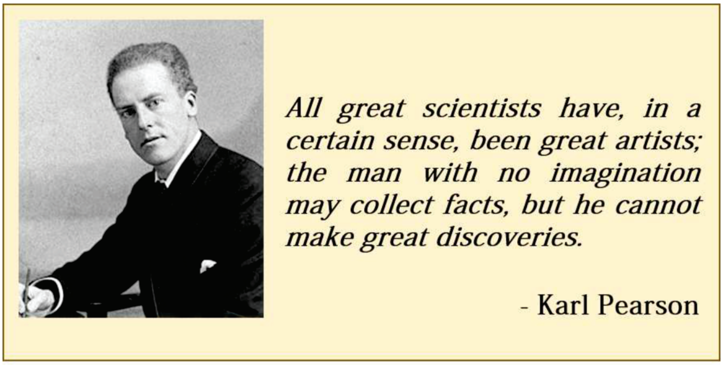
Radhendushka Srivastava is a postgraduate (2005) in statistics from the University of Lucknow. He holds a PhD (2006–2011) in statistics from Indian Statistical Institute, Kolkata, and was a post-doctoral associate at Cornell University (2011–2013). He subsequently joined the Department of Mathematics at IIT Bombay (2013–till date).

Introduction

Karl Pearson (1857–1936) was an English mathematician and philosopher. He was a professor of applied mathematics and mechanics at University College, London (UCL). His collaboration with W. F. R. Weldon led him to statistics, and he worked on the theory of evolution. He was immensely influenced by Sir Francis Galton’s ideas on eugenics. Many of Pearson’s contributions connected to the theory of evolution came out as articles during 1893 to 1912 (Some of his works [1, 2, 3] are listed in the suggested reading section). These articles (related to correlation, regression, chi-squared test, etc.) contain probably his greatest statistical ideas. In 1896, he was elected a Fellow of the Royal Society (FRS). He was awarded several prestigious awards for his innovative statistical ideas. Weldon, Galton and Pearson founded the

Keywords
Statistics, testing of hypotheses, histogram, correlation, regression, Pearson’s family of distributions, chi-squared test, correlation coefficient.

[*] Vol.28, No.2, DOI: https://doi.org/10.1007/s12045-023-1542-3



All great scientists have, in a certain sense, been great artists; the man with no imagination may collect facts, but he cannot make great discoveries.

- Karl Pearson

Figure 1. Karl Pearson

Weldon, Galton and Pearson founded the journal *Biometrika* in 1901 to publish developments in statistical theory, quantitative biology and eugenics. Today, *Biometrika* is known as one of the best statistics journals for publishing innovative statistical methodology and applications.

journal *Biometrika* in 1901 to publish developments in statistical theory, quantitative biology and eugenics. Today, *Biometrika* is known as one of the best statistics journals for publishing innovative statistical methodology and applications. The well-known statistician D. R. Cox [4] wrote an article in 2001 that discusses the development of the first 100 years of *Biometrika*. Karl Pearson also established the first (in the world) Department of Statistics (now renamed the Department of Statistical Science) at UCL in 1911. While Karl Pearson made important contributions to hypothesis testing, it is the work of his son, Egon Pearson, along with Jerzy Neyman, that is now studied as the first theorem on that subject (Neyman–Pearson lemma on most powerful tests). There are several articles written about Karl Pearson and his statistical innovations (a few are listed in suggested reading [5, 6, 7, 8]).

We illustrate a few contributions of Karl Pearson, especially on a specific family of distributions arising from biological data, frequency curve (histogram), and statistical testing of hypotheses. As mentioned above, his contribution to mathematical statistics is immense and can not be described in a short article. Readers are requested to go through the suggested list of readings and references therein to learn more about the contributions of Karl Pearson.

1. Testing of Hypotheses

An important statistical query for a given data set is to decide on a parametric probability model that fits the data well. Karl Pearson [3] investigated a criterion for any distributional model of an observed data set and applied it to the determination of goodness of fit with the frequency curve. His method of the goodness of fit is now referred to as Pearson's chi-squared test. It consists of comparing the observed counts of data values in different intervals of the range with expected counts implied by the presumed distribution. Let O and E denote the observed as well as expected counts of the data values, then Pearson's χ^2 goodness of fit is computed as

$$\sum \frac{(O-E)^2}{E}.$$

A large value of the above statistics indicates that the data distribution may not be the same as the specified probability distribution.

Karl Pearson investigated a criterion for any distributional model of an observed data set and applied it to the determination of goodness of fit with the frequency curve.

One would like to point out that the method was developed for data arising from biological phenomena. Readers can go through [9] where Henry Inman compiled the letters exchanged between Karl Pearson and R. A. Fisher on the challenges in statistical testing of hypotheses. Jerzy Neyman and Egon Pearson later contributed to the fundamentals of statistical testing of hypotheses.

Pearson's χ^2 goodness of fit test is used even today for confirmatory analysis. Statistical testing of hypotheses is a well-known research area of modern mathematical statistics. With the advent of modern computing facilities and the availability of a huge volume of data, statistical testing of hypotheses remains a vibrant area of modern statistical research.

A histogram is a graphical view of the frequency distribution of grouped continuous data.

2. Histogram

A histogram is a graphical view of the frequency distribution of grouped continuous data. It is also a commonly used nonparametric density estimator. Histograms often help statisticians visualize how the data are distributed, the description being more informative when the sample size is large. The term histogram was introduced by Karl Pearson in (either 1891 or 1895; see [10, 7]). Funkhouser [10] in his article quotes the following quote by Karl Pearson, "[It was] introduced by the writer in his lectures on statistics as a term for a common form of graphical representation, i.e., by columns marking as areas the frequency corresponding to the range of their base." Histograms are also known as block diagrams and stair-case charts. Magnello [7] cites that Karl Pearson introduced the 'histogram', a term he coined to designate a 'time-diagram' in his lecture on 'Maps and Chartograms'. One may confuse between the histogram and bar chart. A histogram is constructed for continuous data, and there is no gap between the vertical bars. In contrast, the bar chart is typically used for categorical (or grouped) data and is drawn with a separation between the bars.

A histogram is constructed for continuous data, and there is no gap between the vertical bars. In contrast, the bar chart is typically used for categorical (or grouped) data and is drawn with a separation between the bars.

In modern statistical data analysis, histogram plays a major role in understanding the distribution of data and then choosing an appropriate probability model that suits the data best.

3. Pearson's Family of Distribution

In his early work on mathematical contribution to the theory of evolution [1], Pearson pointed out the geometrical relationship between the normal curve of frequency

$$z = z_0 e^{-x^2/2\sigma^2}.$$

and symmetrical point of binomial $\left(\frac{1}{2} + \frac{1}{2}\right)^n$ under the following set of assumptions.

1. The chances of any 'contributory cause' giving its unit of deviation in excess or in defect are presumed to be equal.

2. The number of 'contributory causes' are supposed to be indefinitely great.

3. The 'contributory causes' are all supposed to be independent.

In the absence of assumption 1, a curve, which is related to the skew binomial $(p+q)^n$ in precisely the same manner as the normal curve, is related to the symmetrical point of binomial $\left(\frac{1}{2}+\frac{1}{2}\right)^n$, is given by

$$z = z_0\left(1+\frac{x}{\alpha}\right)^p e^{-\gamma x}.$$

If α is the total frequency, and μ_r denotes the r^{th} moment about its centroid vertical of the frequency curve, then for this curve

$$2\mu_2(3\mu_2^2-\mu_4)+3\mu_4=0.$$

This relation must be satisfied or nearly satisfied if a series of observations or measurements are to be fitted with the skew curve, which is related to the skew point of the binomial, as the normal curve is to the symmetrical point of the binomial. For fitting a skew point-binomial, we must have

$$\mu_4 < 3\mu_2^2 + 3\mu_3^2/(2\mu_2).$$

For the normal curve $\mu_4 = 3\mu_2^2$. But a great number of statistical returns, especially in anthropometry and zoometry, give

$$\mu_4 > 3\mu_2^2 + 3\mu_3^2/(2\mu_2).$$

Hence they differ from the normal curve in the opposite direction to the skew point of the binomial and its corresponding frequency curve.

Pearson introduced the generalized frequency curve that may be more suitable than normal for frequency curves arising from different disciplines in the following type of differential equation

$$\frac{1}{z}\frac{dz}{dx} = -\frac{x}{\beta_1+\beta_2x+\beta_3x^2},$$

where β_1,β_2,β_3 are real numbers. The solutions to the differential equation are sometimes referred to as Pearson's family of probability density (continuous) curves. Now there are several types of Pearson family of distributions that include well-known probability density functions like beta, gamma, exponential, etc., along with sampling distributions t, χ^2, and F.

Pearson introduced the generalized frequency curve that may be more suitable than normal for frequency curves arising from different disciplines.

4. Correlation Coefficient

The correlation coefficient quantifying the relationship between two numeric variables based on paired measurements/observations is perhaps the most well-known statistical concept commonly associated with Karl Pearson.

The correlation coefficient quantifying the relationship between two numeric variables based on paired measurements/observations is perhaps the most well-known statistical concept commonly associated with Karl Pearson. Ironically, he had only chronicled a history of correlation in 1920 [11]. The origin of the correlation coefficient is attributed to Sir Francis Galton, though Auguste Bravais had published strikingly similar ideas several decades before Galton.

Pearson also made his own contributions to the development of the correlation coefficient. If the correlation coefficient computed from a bivariate data set is regarded as an estimator of the unknown population correlation coefficient, there is an associated estimation error. Pearson was the first to provide a reliable measure of the precision of this estimation error (see [12] for details).

Acknowledgement

The author would like to thank Prof. K.S. Mallikarjuna Rao of IIT Bombay and Prof. Debasis Sengupta of ISI Kolkata and the editor for their valuable comments that improved this write-up significantly.

Suggested Reading

[1] **Karl Pearson, II. Mathematical contributions to the theory of evolution. II. Skew variation in homogeneous material, *Proceedings of the Royal Society of London*, Vol.57, pp.257–260, 1885.**

[2] **Karl Pearson, III. Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Science*, Vol.185, pp.71–110, 1894.**

[3] **Karl Pearson, X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Vol.50, No.302, pp.157–175, 1900.**

[4] **D. R. Cox, *Biometrika: The First 100 Years*, *Biometrika*, Vol.88, No.1, pp.3–11, 2001.**

[5] E. S. Pearson, Karl Pearson: An appreciation of some aspects of his life and work, *Biometrika*, Vol.28, No.3/4, pp.193–257, 1936.

[6] A. W. F. Edwards Galton, Karl Pearson and Modern Statistical Theory, Keynes, M. (eds) Sir Francis Galton, FRS, *Studies in Biology, Economy and Society*, Palgrave Macmillan, London, 1993.

[7] M. Eileen Magnello, Karl Pearson and the origins of modern statistics: An elastician becomes a statistician, *The New Zealand Journal for the History and Philosophy of Science and Technology*, Vol.1, 2006.

[8] M. Eileen Magnello, Karl Pearson and the establishment of mathematical statistics, *International Statistical Review*, Vol.77, No.1, pp.3–29, 2009.

[9] Henry F. Inman, *Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from nature*, *The American Statistician*, Vol.48, No.1, pp.2–11, 1994,

[10] H. Gray Funkhouser, Historical development of the graphical representation of statistical data, *Osiris*, The University of Chicago Press on behalf of The History of Science Society, Vol.3, No.3, pp.269–404, 1937.

[11] Karl Pearson, Notes on the history of correlation, *Biometrika*, Vol.13, No.1, pp.25–45 1920.

[12] Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press: Cambridge, 1986.

[13] Karl Pearson, Mathematical contributions to the theory of evolution. XIX. Second supplement to a memoir on skew variation, *Philosophical Transactions of the Royal Society of London, Series A*, Containing Papers of a Mathematical or Physical Character, Vol.216, pp.429–457, 1916.

Address for Correspondence
Radhendushka Srivastava
Email:
rsrivastava@iitb.ac.in