The Language of the English Biometric School

John Aldrich

Department of Economics, University of Southampton, Southampton, SO17 1BJ, UK. Email: john.aldrich@soton.ac.uk

Summary

This paper considers the language devised by Karl Pearson and his associates for discussing distributions, populations and samples, the basic language for frequentist inference. The original language—some of which is still in use—is described and also the changes it underwent under the influence of R.A. Fisher and of Russian and American mathematicians. The period covered is roughly 1890–1950.

Key words: Statistical notation; Karl Pearson; R.A. Fisher.

1 Introduction

This paper considers the role of language—the system of technical terms, symbols and idioms—in the development of statistical theory by examining a language from the past, that of the “English biometric school”, as Greenwood (1924, 1926) called the tradition of statistical research associated with Karl Pearson and Biometrika. The school went back to the 1890s but, even as Greenwood was writing, it was losing its identity to merge into the “English statistical school”, into the “Anglo-American school” and then ... such national tags fell out of use. I examine the segment of the language concerned with distributions, populations and samples—the basic language of frequentist inference—and also describe what became of it. This is only a small segment; there was much more—see chapter 7 of Walker’s History (1929).

The investigation serves two purposes. The translator’s notes provided here may help the reader through the school’s literature, separating different ways of seeing things from different ways of expressing them. More speculatively a study of the development of a particular language may cast light on the role of language in the development of statistical theory. Statistical language has not received much attention from historians. The most ambitious history of statistical ideas—Hald (1998)—abstracts from the history of language by stating propositions in modern terminology and writing mathematical expressions in modern notation. David’s “First (?) Occurrences” (2001) and Miller’s “Earliest Uses” provide information on individual terms and symbols but do not treat systems of such. The account below draws on all these sources as well as on Stigler (1986) and the large secondary literature on the work of the school; for further references, see Aldrich (2001).

Weldon’s (1893, p. 329) remark, “It cannot be too strongly urged that the problem of animal evolution is essentially a statistical problem”, expressed the original biometric impulse but by Greenwood’s time the school had long been a school of statisticians not of biologists. Weldon—and Galton too—were great influences on the original biometric project but Pearson had most impact on the language treated here. He kept something of the existing languages—he had been taught probability and the theory of errors and he read widely when he began work on biometry—but he made many changes, usually reflecting his background in applied mathematics (mathematics applied to physics). His language was still evolving when he died in 1936 but by then the language of his
community was out of his hands. It had been transformed by R.A. Fisher, classifiable perhaps as a rebellious insider, and numerous outsiders. From the 1910s mathematicians from the Russian Empire, formal or informal, and, from the 1920s, mathematicians from the U.S.A. worked to turn the ideas of the English school into respectable mathematics. Insiders considered them *pure* mathematicians. The Fisher effect and the pure mathematics effect were distinct. Indeed Neyman (1967) a leading ‘purist’ wondered that Fisher could be so hostile to the new mathematical spirit. See Seneta (1994) for a sketch of the Russian background and references.

I show the original language at work, indeed under pressure, in a middle-period piece, Student’s “The Probable Error of a Mean” (1908). This is perhaps the most remembered publication of the biometric school, discussed by Eisenhart (1979), Hald (pp. 664–8) and Lehmann (1999), amongst others, and excerpted in Kotz & Johnson’s *Breakthroughs* (1992). “Student’s distribution” has had a central place in statistics since 1925 when Fisher put it there but this was after a radical reconstruction and the original is remote in both technique and language. Student’s paper made a breakthrough in language—by carefully marking the distinction between population and sample quantities—but otherwise its voice is that of Pearson fitting curves and calculating moments. In Fisher’s voice it became an investigation of sampling distributions. The original incorporated translations from the language of the theory of errors for that is where the topic—inferece about the mean of the normal distribution—came from. This density of history makes the paper an ideal text on which to base a discussion of language change.

§§2–8 below treat the Pearsonian language for sampling and for distributions and §§9–14 consider how Fisher transformed it. The ‘pure mathematicians’ are in evidence throughout but most prominent in §6 on sample and population means.

2 Student’s “The Probable Error of a Mean”

“The Probable Error of a Mean” does not mean exactly what it says. The paper is not about the *probable error* of the mean—or of anything—least not directly. In the large sample normal theory of Pearson all that was required for inference was the estimate and its *probable error* (see §4 for the term) but the probable error had no place in Student’s derivations or in the new model practice which required special tables. Nor did it figure in Student (1908a) “Probable Error of a Correlation Coefficient” or in Soper (1913) “On the Probable Error of the Correlation Coefficient . . . “. The inertia was overcome when, beginning with Fisher’s “frequency distribution of . . . ” (1915), more descriptively apt titles appeared. Yet Student still used the old form in 1921; apparently it had come to mean what he required for inference—even if that was a sampling distribution not a probable error.

“The Probable Error of a Mean” investigates the distribution of the quantity \( z = \frac{x}{s} \) where \( x \) is “the distance of the mean of a sample from the mean of the population” and \( s \) is “the standard deviation found from a sample”. Student obtained the distribution of \( z \) from those of \( x \) and \( s \). The distribution of \( x \) was known and the distribution of \( s^2 \) was found by curve fitting using the method of moments (see §5 below)—that is, Student found a curve from the Pearson family, whose first moments matched those he calculated for \( s^2 \). This preliminary work on \( x \) and \( s \) provides our sample texts: in Extract I (§3 below) the distribution of \( x \)—known from the theory of errors—is expressed in the language of biometry; in Extract II (§5) a pattern is sought in the moments of \( s^2 \) which will identify its distribution; in Extract III (§6) the first of the moment determinations yields the expected value of \( s^2 \); in Extract IV (§8) the conclusion of the calculations is presented and the pattern for a Type III curve recognised. Student’s moment-matching procedure was made obsolete by the geometric change of variable procedure used by Fisher (1915) and years later the biometricians realised that the distribution of \( s^2 \) had been found by Helmert (1876)—see Pearson (1931), Hald (1998, p. 668) and David & Edwards (2001) for a translation of the original.
3 Populations, samples and $N$

Extract I. Student (1908, p. 7) expresses the distribution of $x$ in the language of biometry:

Now the means of these samples of $n$ ["drawn from a normal population with standard deviation $\sigma"$] are distributed according to the equation

$$y = \frac{\sqrt{n}}{\sqrt{2\pi} \sigma} e^{-\frac{n y^2}{2\sigma^2}}.$$

I consider first the role of $N$ and the population/sample terminology, leaving the biometricians’ language of normality to §4.

$N$ was not used in the theory of errors or in Edgeworth’s work in mathematical statistics; it came from physics. Maxwell, one of Pearson’s teachers, wrote $N$ for “the whole number of particles” when he (1860) discussed the distribution of particle velocities in gas theory. (For Maxwell and his relation to the theory of errors, see Porter (1986, chapter 5.).) In gas theory there were no observations on individuals (particles) nor was there a notion of sampling. In his articles on curve fitting Pearson (1894, 1895) considered a smooth curve as an approximation to an empirical frequency polygon and in that setting $N$ is most naturally interpreted as the sample size with the $N$ construction useful for showing the expected numbers in each interval. But Pearson wrote $N$ whether there was curve fitting or not. When he (1896, p. 264) formulated the bivariate normal distribution he wrote $N$ for “the total number of pairs” and $z \times \delta x \delta y$ for the “frequency of a pair falling between $x$ and $x + \delta x$, $y$ and $y + \delta y$” with

$$z = \frac{N}{2\pi \sigma_1 \sigma_2 \sqrt{1 - r^2}} e^{-\frac{1}{2} \left[ \frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - 2r \frac{x y}{\sigma_1 \sigma_2} \right]}.$$

Pearson followed the usual equation convention and wrote $z$ as depending on $x$ and $y$ (in the univariate case $y$ on $x$). Today some variation of “$f_x(x) =$” is often used but the notion of a “random variable” involved here only became domesticated in statistics in the 1940s. The movement towards a more specialised and stereotyped language is paralleled in other branches of mathematics.

There were counts of individuals, members of families, in Galton’s Natural Inheritance (1889), an early model of biometric research, but Galton did not articulate the concept of sample nor discuss the relationship between sample and population. He (p. 35) elided the distinction between them when he emphasised that the science of heredity is concerned with “large populations”. Pearson began conjoining “population” and “sample” around 1900. The object of his $\chi^2$ test was to determine “whether the sample may be reasonably considered to represent a random system of deviations from the theoretical frequency distribution of the general population” (1900, p. 164). “Population” entered his vocabulary somewhat earlier, the “normal population” of Pearson & Filon (1898, p. 277) replacing the “normal group” of Pearson (1894). Weldon (1890 and 1893) had referred to collections of observation units (e.g. 1000 adult female crabs) as “samples” and wondered if his were large enough. Pearson (1896, p. 264) replied but did not add “sample” to his own vocabulary. Pearson & Filon’s (1898) “On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation” developed a large sample normal approximation to the joint distribution of some statistics, based—naturally and without remark—on independent observations. The “random selection” discussion treats the distribution of random selections of one statistic given fixed values of the others. “Selection” was a term from the theory of evolution and in the “Mathematical Contributions to the Theory of Evolution” the modelling of the biological process was at least as important as the statistical inference. The phrase “random sampling” appears in the statistical paper of 1900 which, like the probable error paper of 1903, is free of “organs” and other biological paraphernalia.

By 1900 the abstract idea of a population of anything was established, yet an appeal to population
size still seemed necessary. Thus Pearson (1903, p. 273) wrote “If the whole of a population were taken we should have certain values for its statistical constants, but in actual practice we are only able to take a sample . . . .” His paper’s results were for "random samples" from a finite population of size $N$ but they were applied to the normal case where there is no finite population. When Student (1908, p. 6) mentions that "$N$ as usual represents the total frequency" he is referring to a total frequency of sample means or of sample standard deviations. Soper (1913, p. 91) refers to the "mean value of correlation coefficient for $N$ samples of size $n$".

In the theory of errors it was not usual to refer to the number of errors or the population of errors or use $N$. Fisher brought this practice of distributing relative frequencies into biometry (see §9 below) but he fell in with the local language of populations and samples, writing about an "indefinitely large population" (1915) and "hypothetical infinite populations" (1922, p. 311). In the 20s $N$ seemed to be dying but the revivals of Elderton (1906) and Yule (1911) revived it too: $N$ is in Elderton & Johnson (1969, p. 70) and Yule & Kendall (1950, p. 251). I have not found any criticism of the use of $N$. Perhaps it was seen as an empty formality—look past $N$ to the working part of the expression—or as another manifestation of the English unsoundness on probability; cf. Chuprov’s views discussed in §6.

Associated with Pearson’s populations were frequency curves, or surfaces, with “frequency constants”. Pearson & Filon (1898, p. 231) consider an $m$-dimensional multivariate data distribution. “Let this frequency surface be given by $z = f(x_1, x_2, x_3, \ldots, x_m; c_1, c_2, c_3, \ldots, c_p)$ where $c_1, c_2, c_3, \ldots, c_p$ are $p$ frequency constants.” As the $x$’s are measured from their means $h_1, h_2, h_3, \ldots, h_m$, there are $m + p$ constants in all. The $c$ for constants notation seems natural although some error theory works, including Chauvenet (1863) and Brunt (1917) (see §§4 & 9 below), viewing the constants as unknowns to be found denoted them by $x, y, z$. Jeffreys’s Theory of Probability (1939, §§3.4/5) mixes conventions: $x$ for the mean and $x_1, \ldots, x_m$ for the regression coefficients appear with $\sigma$ for the standard deviation.

4 Normality, Standard Deviations, Probable Errors

The word “normal” (Extract I) had been used by Galton but Pearson made it a standard term; see Kruskal & Stigler’s “Normative Terminology” (1997) for more on its origins. The books Pearson was recommended as a student in 1874–7, Chauvenet (1863) and Thomson & Tait (1867), referred to the “probability curve” or “law of error”. “Normal” may be a sound abbreviation for according to the law (of error) and Pearson (1893, p. 329) may have begun by thinking that as a rule the normal held but he kept the name even after he had concluded that it did not; by ’95 he (p. 360) was saying, “to deal effectively with statistics [data] we require generalised probability curves which include the factors of skewness and range.” In the 1920s Fisher would reverse Pearson’s allocation of tasks—non-normal for data description and normal for inference—and the normal distribution would become the norm for data description; see §12 below. The discovery that $\chi^2$ derived from normally distributed data followed the Pearson Type III curve set the pattern for this development.

According to Stigler (1986, p. 328), Pearson settled on the term “standard deviation” after trying “standard divergences”. Pearson (1894, p. 88n) said he had “always found it more convenient to work with the standard-deviation than with the probable error [= $(2\pi)^{-1/2}$] or the modulus [h $= (2\pi)^{-1/2}$], in terms of which the error-function is usually tabulated.” His measure appears in Airy (1861) as the “error of mean square”. In Brunt’s Combination of Observations (1917, p. 30) it is the “mean square error” represented by $\mu$. Pearson’s choice of $\sigma$ followed the same first letter rule: use the first letter of the name of the symbolised, possibly via transliteration. Fisher’s (1918, p. 399) “variance” reinforced rather than displaced the standard deviation, for the new measure was defined as the square of the standard deviation and written $\sigma^2$. Walker (1929, pp. 49–57) describes the various measures.
Moments (discussed in §5 below) arrived at the same time as the standard deviation and so a moment equivalent in the form of $\sqrt{\mu_2}$ was always available. However it was convenient to have a name and simple symbol for such a useful concept. The situation reflected that in mechanics where there was an analogous quantity the “swing radius”, as Yule (1896, p. 328) explained in an exposition for statisticians, based on Pearson’s lectures (1894–6).

The new symbol changed the look of the law of error. Chauvenet had used Gauss’s original notation and, slightly modernised, this appears in Brunt (1917, p. 13): “the error curve” is written

$$\phi(x) = \frac{h}{\sqrt{\pi}} e^{-hx^2}$$

so that $\phi(x)dx$ is “the proportion of errors between the limits $x - \frac{1}{2}dx$ and $x + \frac{1}{2}dx$”.

The biometricians had no mean parameter and so the “normal curve” remained an error curve—notoriously errors are centred on 0. Fisher’s Statistical Methods for Research Workers (1925) brought the non-central normal into the canon. The quantity $x$ of Extract I, distributed as $N(0, \frac{\mu_2}{n})$, is a “mean” because the observations are measured from the mean of the population—see Extract III in §6 below. Today $x$ might be written as $\frac{1}{n} \sum (X_i - \mu)$ or $\bar{X} - \mu$ and “the equation” (probability density function) corresponding to $(\bar{X} - \mu) \sim N(0, \frac{\mu_2}{n})$ derived from that for $X \sim N(\mu, \frac{\mu_2}{n})$. However Student and his contemporaries were used to working with deviations and for them the equation for $\bar{X}$ was the exotic. Equations for non-central normal distributions of any kind were uncommon in the biometric literature though they did appear, as when Pearson (1894, p. 72) wrote a mixture of non-centrals.

Pearson’s large-sample normal inference theory was about deriving and manipulating (newly minted) standard deviations. Yet Pearson retained the “probable error” as a measure of precision: half the area of the normal curve is within one probable error of the mean. Galton (1889, p. 58) lamented that “this cumbrous, slip-shod and misleading phrase” (dating from 1815) was “too firmly established”. Pearson (1903, p. 273) also bowed to the weight of “custom” and in his Tables (1914, p. xxii) used the stupendous symbols $\chi_1 \sigma$ and $\chi_2 \sigma$ for the probable errors of the mean and standard deviation ($\chi_1 = .6744898/\sqrt{n}$ and $\chi_2 = .6744898/\sqrt{2n}$). Out of the probable error and standard deviation Yule created the standard error, first (1897) for a regression residual and then (1911) for anything that could be estimated. Fisher followed, saying (1925, p. 48) curtly “The common use of the probable error is its only recommendation.”

Pearson & Filon (1898) represent the standard deviations of the frequency constants by symbols on the pattern $\sigma_{e_1}$ and $\sigma_{e_1}$ etc. In Student’s (1908, p. 7) notation these are registered as standard deviations of sample quantities—thus $\sigma_{^2\bar{x}}$. When Pearson and Filon consider $\sigma_1$ and $\sigma_2$, the standard deviations of the bivariate normal, they write $\Sigma_{\sigma_1}$ and $R_{\sigma_1\sigma_2}$ for the “standard deviation of errors in $\sigma_1$” and the “coefficient of correlation between errors in $\sigma_1$ and $\sigma_2$”. The symbols reflect two features of the Pearson style: the resort to upper case when a second symbol is needed and the use of $\Sigma$ in a non-summation role—see §5 and §7 below.

5 Curves, Moment Coefficients, Moments

Extract II Student (1908, p. 4) conjectured the distribution of $s^2$ by establishing that the pattern in its moments is that required for a Pearson Type III curve, viz. $2\beta_2 - 3\beta_1 - 6 = 0$:

$$\beta_1 = \frac{M_3^2}{M_2^2} = \frac{8}{n - 1}, \quad \beta_2 = \frac{M_4}{M_2^2} = \frac{3(n + 1)}{n - 1}$$

Thus

$$2\beta_2 - 3\beta_1 - 6 = \frac{1}{n - 1}[6(n + 3) - 24 - 6(n - 1)] = 0$$
Consequently a curve of Prof. Pearson’s Type III may be expected to fit the distribution of $s^2$.

Moments (represented by $M$’s) and these particular functions of them ($\beta$’s) were standard accessories in Pearsonian curve-fitting, to which we now turn.

Here Pearson was not over-writing old terminology—as with normality—but introducing new terms for a new project. The Type III curve was one of a family of curves, characterised by their first four moments, introduced by Pearson (1895). The object, he (1916, p. 529) recalled, was to obtain a series of curves such that one or other of them would agree with any observational or theoretical frequency curve of positive ordinates to the following extent:—(i) The areas should be equal; (ii) the mean abscissa or centroid vertical should be the same for the two curves; (iii) the standard deviation … about this centroid vertical should be the same for the two curves, and (iv) to (v) the third and fourth moment coefficients should be the same.

Although the curves had their genesis in a probability argument—see Stigler (1986, pp. 333–341)—the curve-fitting imperative dominated. The frequency curves $y = f(x)$ are integrals of the differential equation (1895, p. 381):

$$\frac{1}{y} \frac{dy}{dx} = \frac{-x}{c_1 + c_2x + c_3x^2}.$$

The type depends on the nature of the roots of the quadratic in the denominator, which can be expressed in terms of moments. Pearson (1895) started with five types of curve but the number grew to seven (1901) to twelve (1916). In the 1916 paper the classification into types is represented on a diagram subdividing the $\beta_1\beta_2$ plane. The numbering was a system of naming only; there was no numerical structure to the system.

“Moments” came from mechanics and Pearson had originally used methods from graphical statics for calculating them. Mechanics, central to the mathematics of the time, had already provided the inspiration for Chebyshev’s “method of moments”—see Seneta (1994)—but Pearson seems not to have known of this. Pearson (1895) treated a “frequency-curve” as an object with mechanical properties, essentially as the sheet enclosed by the curve and the horizontal axis, the area given by the number of observations. Pearson used the symbols $\mu_i$ etc. for the moment-coefficients around an arbitrary vertical line $y’$ and $\mu_i$ etc. around one going through the “centroid” or centre of gravity of the sheet: “if $y$ be the length of a strip parallel to $y’y’$ and $x$ its distance from $y’y’$ then the $n$th moment $= \int x^ny’dx’$. The system of priming indicates that the moments around the centroid had primacy. The term “moment coefficient” (Extract III below) and $N$ go together. Moments, as moments of ‘absolute’ frequency curves, involved a factor $N$ and the coefficient of $N$ is the “moment coefficient”. When $N$ died the moment coefficient became the moment—see §14 below.

The $M$’s of Extract II represent moment coefficients with the subscript representing the order of the moment. There was a system according to which subscripts, primes and capitals could modify the root symbol $\mu$. The unprimed moment coefficient was taken about the centroid (mean) and the primed around the origin. The $M$’s (capital $\mu$’s) are used in accordance with Pearson’s practice of using capitals when he needed a second symbol after the primary $\mu$ had been used. The $M$’s were not related to the $\mu$’s in any fixed way—they just came second. Student’s $M$’s happen to be moment-coefficients of the distribution of $s^2$ and are secondary to the moment coefficients of the observations, represented by $\mu$’s (see Extract III below).

The investment in moment notation did not stop with the moments. Functions of the moments were used to determine the type of curve to be fitted. The $\beta$’s were introduced in the 1895 paper while the quantity Student evaluates, $2\beta_2 - 3\beta_1 - 6$, was designated $\kappa_1$ in the 1901 supplement. The symbol $\kappa$ for criterion followed the first letter rule but the skewness and kurtosis coefficients, $\beta_1$ and
\( \beta_2 \) were not similarly grounded. At first \( \beta_1 \) appeared to signify only the first of the \( \beta \)'s though later a formula for the numerical subscripts was given. While the moments and these functions of moments were denoted by Greek letters, the constants of the frequency curves were a mixture of Greeks and Latins. Thus the Type III curve was usually written \( y = C x^p e^{-\lambda x} \) with constants \( p \) and \( \gamma \), as in Student (1908, p. 4).

The moment symbols were more salient than the curve symbols and much more care went into their design, thus \( \beta_2 \) was more familiar than \( p \) and more lovingly crafted. Indeed the Type III—and the others—were simply what the application of the moment criteria led to. The curve derived its identity from the criteria it satisfied; the “equation” was accidental. Student did not consider how his \( z \) distribution, which was not obtained by curve-fitting, was placed in the Pearson family, nor had Pearson (1900) placed \( \chi^2 \). In the early 1920s the curves seemed set to rule the world of distributions but the opportunity passed; see §12 below.

Moments are averages and averaging—in its population and sample forms—will be discussed next. Here the biometricians followed contemporary applied mathematics usage.

6 Getting the Mean Value

Extract III The first of Student’s (1908, p. 3) moment determinations for \( s^2 \) is just finding its expected value:

If \( s \) be the standard deviation found from a sample \( x_1, x_2 \ldots x_n \) (all these being measured from the mean of the population) then

\[
s^2 = \frac{S(x_1^2)}{n} - \left( \frac{S(x_1)}{n} \right)^2 = \frac{S(x_1^2)}{n} - \frac{S(x_1^2)}{n^2} - 2 \frac{S(x_1 x_2)}{n^2}.
\]

Summing for all samples and dividing by the number of samples we get the mean value of \( s^2 \) which we will write \( \bar{s}^2 \),

\[
\bar{s}^2 = \frac{n \mu_2}{n} - \frac{n \mu_2}{n^2} = \frac{\mu_2(n-1)}{n},
\]

where \( \mu_2 \) is the second moment coefficient in the original normal distribution of \( x \); since \( x_1, x_2, \ldots \) are not correlated and the distribution is normal, products involving odd powers of \( x_1 \) vanish on summing, so that \( 2 \frac{\Sigma x_i^2}{n} \) is equal to zero.

This result was well known in the theory of errors and is familiar today in the form \( E S^2 = \sigma^2 \); there is an \( \frac{(n-1)}{n} \) wrinkle because the biometricians formed the sample standard deviation with \( n \) as the divisor. The argument is familiar but the notation is not and the commentary is quite weird. The bar and \( S \), the population and sample operators, are easily replaced with an expectation sign and \( \Sigma \) respectively but Student and his contemporaries talked about “getting the mean value” in ways now alien. As late as 1922, Pearson (p. vii) was defining the \( s \)-th moment coefficient for a continuous random variable by \( \mu_s = S(x - \bar{x})^s / N \), where \( N \) is the total population or frequency” and \( \bar{x} \) the population mean.

In applied mathematics/mathematical physics the bar was used in a variety of contexts with an underlying similarity of mathematical form. The mean values \( \bar{v} \) and \( \bar{v}^2 \eta^2 / \bar{v} \) (expectations) appear in Maxwell’s (1867) gas theory while Thomson & Tait (1867, pp. 170–1) define the coordinates of “the centre of inertia” of masses \( v_1, v_2, \ldots \) with coordinates \( (x_1, y_1, z_1), (x_2, y_2, z_2), \ldots \) as \( \bar{x} = \Sigma x \), \( \bar{y} = \Sigma y \) etc.; the same bar symbol is used when there is “a continuous distribution of matter”, with the sums replaced by integrals. Student did not use the bar for the sample mean although this was common at the time. The bar could appear in two capacities in the same piece, although the barrings were generally separated: Brunt (1917) has \( \bar{s}^2 = \frac{n^2}{n} \) \( \left( \text{read as } E(\bar{x} - \mu)^2 = \frac{\sigma^2}{n} \right) \) on p. 10 and \( \bar{x} \) for
the mean of the measurements on p. 21. I have not spotted \( \bar{x} = \bar{X} \) for \( E \bar{X} = E X \).

Student and Brunt’s descriptions of taking mean values emphasise the resemblance between the sample mean and the population mean. In conformity with his use of \( N \), Student writes of “summing for all samples and dividing by the number of samples”, as Maxwell (1860, p. 381) had written “add the velocities of all the particles together and divide by the number of particles”. Brunt’s gloss for the same step in his derivation of \( \bar{x}^2 = \frac{\sum x^2}{n} \) points to the law of large numbers: in a “large number of trials the mean value of the [cross-product terms] will be zero” (p. 9). These idioms are not algorithms, nor do they justify the move to the next line in the argument, they are more heuristic interpretations of what a mean value is or perhaps re-assuring rituals passed from generation to generation.

The bar was used for averaging single terms or simple expressions like \( \bar{x}^2 \bar{y}^2 \) but not for averaging expressions that require brackets. Thus the entire right-hand side of Student’s first equation in Extract II, which would now go into brackets preceded by \( E \), did not go under the bar because there was another symbol, the “vinculum”, in the way. The vinculum was used for bracketing terms so that the expression \( \frac{\bar{B} \bar{D} - \bar{D} \bar{l}}{2} \) in Fisher (1922, p. 329) is the product of a mean value and a squared difference. Statisticians could have developed symbols like \( (\cdot - \cdot)^2 \) for the mean value of a squared difference but it seems they did not. Although Fisher often used the bar for an expected value—see e.g. (1918, p. 420, 1932, p. 77, etc.)—he also used \( \bar{x} \) for the sample mean from the beginning (1912, p. 157) and his perseverance is probably responsible for \( \bar{x}' \)’s survival in this capacity. The use of the bar symbol is very restricted in modern (post-vinculum) statistics; there is no bar operator with the force of \( \frac{1}{n} \sum_i \).

English statistics took many years and perhaps two generations of “Russian” influence to change from the bar and mean value to \( E \) and the expectation. From 1914 Russians—in intellectual affiliation if not nationality—were writing in Biometrika about the efficacy of \( E \) for treating English problems. Anderson (1914) worked it into a treatment of the Pearson—Student variate difference correlation method and then Chuprov’s (1918/19) “On the Mathematical Expectation of the Moments of Frequency Distributions” massively extended Student’s moment results. American publications, like Rietz’s Mathematical Statistics (1927), helped establish \( E \) and the expressions “expected value” and “mathematical expectation” in statistical English. Less mathematical audiences were also exposed to \( E \)—e.g. \( E(x) \) for “the mean value of any variate \( x \)” appears in Neyman’s Statistical Society paper (1934, p. 594)—but \( E \) still seemed arcane in English statistics until after the Second World War.

Chuprov (1918, p. 140) explained the “small popularity” of the expectation method; “English scientific tradition rejects the concept of ‘mathematical probability’ . . . and the method of mathematical expectation has naturally shared the fate of the concept . . . on which it rests.” Chuprov considered the English work rooted in the philosophical ideas of Mill and Ellis in which probability is based on “empirical frequency”. I think this probably exaggerates the influence of philosophical ideas on practical statistics. Greenwood (1926, p. 322) compared the English and Russian styles with special reference to the expectation. He thought Chuprov’s accusation about the neglect of expectation in part a “verbalism”: “whether we choose to write this as \( \bar{x} \) or as \( E(x) \) is surely as much a matter of taste as whether one speaks of the ‘arithmetic mean’ or the ‘arithmetic average’, provided one distinguishes between an empirical and a universal [population] mean”—provided, that is, one satisfies Fisher (see §10 below).

There was something about the English organisation that might be called the rejection of probability. Even before the rise of biometry, probability was fragmented. Astronomers like Airy (1861, preface) aimed to take the theory of errors away from probability or at least from discrete probability. No English books had the range of Bertrand’s Calcul des Probabilités (1908) or Markov’s Wahrscheinlichkeitsrechnung (1912). The method of mathematical expectation actually had a presence in England. In the 1901 edition of his textbook Choice and Chance, Whitworth used a script \( E \) for the “expectation or mean value of a variable magnitude” and stated (p. 205) what are now called the linearity properties. However the application to random divisions did not interest the
biometricians and when the $E$ operator became established it was after *Erwartung* or *espérance*.

Expectations came into Fisher’s domain in 1925 when Romanovsky (1925) described how “l’espérance mathématique” and “la fonction génératrice des moments” might be used to find the distribution of the correlation coefficient. Fisher had already found it but he (1925c) praised the elegance of Romanovsky’s methods and by 1929 was using them—see §13 below. By contrast his first $E$ publication was in 1948—in French.

7 Summing: $S$ and $\Sigma$

The biometric school generally used $S$ (Extract III) for summation but the practice was older than the school and outlasted it. Edgeworth used $S$ (see e.g. his (1883)) before Pearson published anything in statistics and in 1928 Hotelling (p. 346) warned American readers of Fisher’s *Statistical Methods*, “The style includes such peculiarities as the use of $S$ for summation, which is now common amongst British statisticians although everybody else in the world, even in England, uses $\Sigma$.” Fisher (1932, p. 42) retorted that “[T]his symbol is the one regularly used in our subject” and went on using it. However $\Sigma$ appears in his some of his more mathematical publications and when the full detail of $\sum_{i=1}^{n}$ was wanted, as in Cornish & Fisher (1937).

The preference for $S$ was a matter of style only but $S$ belonged to a system of interrelated symbols and a choice in one part had consequences elsewhere. With $S$ covering summation, $\Sigma$ could be redeplored as a second standard deviation symbol (see §4 above) or $S$ and $\Sigma$ could stand for different kinds of summing. Student (1909) used $S$ for summing within the sample and $\frac{1}{N} \Sigma$ for “summing for all samples and dividing by the number of samples”—the barring of 1908. When Church (1925), replied to Chuprov on behalf of “the English statistical student”; he made $\Sigma$ operate with the force of Chuprov’s $E$: “summing for all possible samples . . . and dividing by the number of such samples”. Fisher (1935, pp. 42–43) does something similar.

The notational tensions of the inter-war years are reflected in the publications of Egon S. Pearson; his co-authors included Karl Pearson, the American S.S. Wilks and the Pole J. Neyman. The last partnership was built over an Anglo-Russian cultural fault; Neyman was used to $\Sigma$ (and $E$ but not $N$)—see e.g. his (1925). They set off in 1928 with $S$ for summation and $\Sigma$ for “the sample”. Gradually the $S$’s were eliminated and in Egon’s (1936 &1938) life of Karl, the father’s $S$’s gave way to $\Sigma$’s.

8 Sampling Problems: $\sigma$ and $s$, $R$ and $r$

Extract IV Student (1908, p. 7) writes the equation for the Type III distribution conjectured in Extract III:

$$y = \frac{C}{\sqrt{n}} n^{-\frac{3}{2}} e^{-\frac{x^2}{2n}}$$

as the equation representing the distribution of $s$, the standard deviation of a sample of $n$, when the samples are drawn from a normal population with standard deviation $\sigma$.

Student had the solution of what Fisher called a “sampling problem”, i.e. finding the sampling distribution of a statistic. Fisher saw the 1908 papers as forerunners for his own work on sampling distributions; the phrase “sampling distribution” appears in his (1922b, p. 598).

Student extended the biometric language with his careful distinction between “the mean of the sample” and the “mean of the population” and the use of distinct but related symbols to distinguish the standard deviations. Fisher (1939, p. 2) described the situation before 1908:

Sampling problems were . . . constantly and seriously obscured by the practice of using the same symbol for the quantity estimated and for the estimate of it derived from the
observations. . . . It is a tribute to “Student’s” penetration that, without this aid [separate symbols] in most of the literature he must have studied, he should yet have arrived at a clear insight into the nature of sampling problems.

The only literature Student cites is Airy’s Theory of Errors—for authority for the equation in Extract I. Airy (pp. 28–37) derived the probable error of the mean and argued that the mean is normally distributed but he did not write the equation or any other sampling problem equation. As noted in §3, Student did not actually write an equation for the sampling distribution of the sample mean, but of the deviation of the sample from the population mean.

In his first papers on probable errors—Pearson (1896) and Pearson & Filon (1898)—Pearson did not demonstrate a “clear insight” into sampling problems. The basic paper on normal correlation (1896) has Pearson’s first probable error derivation—for r—and the argument seems to lead to a normal approximation to a posterior distribution. The formula \( r = \frac{S(xy)}{(n\sigma_1 \sigma_2)} \) (p. 265) appears monstrous today (as usual the working is in deviations from the mean) but it was understood that \( r \) and the \( \sigma \)‘s are on sample duty. Pearson’s use of the same symbol did not seem to confuse the user of his results but they made the theoretical arguments hard to follow. Pearson & Filon (1898) extended the treatment to apply to fitting skew curves by the non-Bayesian method of moments. Here Pearson clearly thought he was dealing with a normal approximation to a sampling distribution but the argument is obscure and the conclusions wrong. Aldrich (1997, §17) discusses this phase of Pearson’s probable error work.

The approach of Pearson (1903) is definitely not Bayesian and the probable error determinations are exercises in calculating standard deviations. There is a clear sample/population distinction but moments of the sample and moments of the population are not distinguished in the notation. Pearson’s “simple proofs” operate in a finite universe and there is plenty of “dividing by the number of random samplings after summing for all such samples” even though the results are applied to normal curves.

Fisher made a great thing of the sample/population distinction and developed a system of linked symbols but Student did not have to work with the linked pair \( s \) and \( \sigma \), or any pair. He could have used only \( \sigma \) with some appropriate form of words for the sample quantity and used the utility symbol \( x \) instead of \( s \); he does use \( x \) for \( s^2 \) at one point (1908, p. 4). Helmert (1876, p. 122) wrote the probability for the sum of squares of the \( n \) deviations as

\[
\frac{h^{n-1}}{\Gamma \left( \frac{n-1}{2} \right)} \sigma^{n-1} e^{-h/2} d\sigma
\]

where \( h \) is the Gauss symbol for the modulus and \( \sigma \) is a utility symbol (replaced by \( u \) in the David & Edwards (2001) English translation to avoid confusion).

In a second paper Student (1908a, p. 302n) augmented Pearson’s notation by writing “\( r \) . . . for the correlation coefficient of a sample and \( R \) for the correlation coefficient of a population”. Student handled the established symbols differently in his two papers: \( r \) became the sample quantity and \( \sigma \) the population quantity. Perhaps Student thought that \( \sigma \) was fixed in its role in the normal curve and \( r \) in the formula \( S(xy)/(n\sigma_1 \sigma_2) \). In his pairing he may have considered \( \rho \) and \( \Sigma \) unavailable as both appeared in other roles in Pearson (1896), the correlation reference. Of course he may have just been satisfied that the notation worked then and there. \( \{R, r\} \) did not become the prototype for an upper-case lower-case system perhaps because \( R \), as a grander \( r \), had already been taken for the multiple correlation coefficient by Yule (1907). The next Biometrika contribution, Soper (1913), replaced \( R \) by \( \rho \) and this notation has stayed. \( R \) has stayed too, though it was an anomaly in the Graeco-Latin system for its Greek counterpart is \( P \). When Fisher (1928) gave the distribution of \( R \), he used \( \rho \) for the population quantity. The standard modern work, Anderson (1959/84), uses the historically insensitive \( \overline{R} \).

Student may have demonstrated the case for a dual notation—he did not put it into words—but the pairing \( \{R, r\} \) did not set and years passed before \( \{\sigma, s\} \) was adopted. The only significant early
reference on \( z \), Pearson’s *Tables* (1914, pp. xiii & 36) respected Student’s use of \( z \) but presented no subsidiary notation. Fisher (1915, p. 507) wrote \( \mu^2 \) not \( s^2 \) and Pearson (1915, pp. 522–3) paired \( \Sigma \) with \( \sigma \), capitalising for the second symbol to be introduced. In 1925 when Pearson wrote again on the bivariate normal he used \( \sigma_1, \sigma_2 \) and \( r \) for the sample quantities and \( \Sigma_1, \Sigma_2 \) and \( \rho \) for the population constants.

In neither the 1908 papers of Student nor in his “The Distribution of the Means of Samples which are not Drawn at Random” (1909) are there symbols corresponding to the modern \( \bar{X} \) and \( \mu \). The parsimony reflected perhaps two principles: write only central normals and introduce symbols only if they are to be used in serious mathematical expressions; see §4 above. The 1909 notation followed Pearson on capitalisation: lower-case Greek letters for the first lot of moments—of the parent distribution (\( \mu_2, \ldots, \beta_1 \), etc.)—and capitals (the Greeks and the Romans happen to be the same!) for the second lot—for the distribution of the mean (\( M_2, \ldots, B_1 \), etc.).

9 Fisher and the Theory of Errors

The effect the mature Fisher had on the biometric language can be seen by comparing Student (1908) with Fisher (1925a) or Pearson & Filon (1898) with Fisher (1922). However in Fisher’s first publications the language of the theory of errors was as prominent as that of biometry. This was a bilingualism he put to advantage.

Fisher’s tutor, the astronomer F.J.M. Stratton, lectured on the theory of errors. Judging by Brunt (1917), which was based on the lectures, the theory had changed little from Pearson’s Cambridge days. Brunt does describe some biometric developments but does not try to relate them to least squares theory or to coordinate the languages. There was no sense among the astronomer/error theorists that developments in Cambridge pure mathematics might be relevant. Yet probable errors turn up in a number theory paper by Hardy & Littlewood (1914, p. 187)—see von Plato (1994, p. 58) for its place in the history of the strong law of large numbers—and Hardy would later encourage Cramér to write a book on probability theory because there was no mathematically satisfactory book in English; see Cramér (1976, p. 516).

Fisher’s first paper (1912) was an error theorist’s incursion into biometry. Fisher (p. 157) chose the paradigm error theory problem to illustrate his solution to the Pearson problem of “fitting frequency curves”: find the “arbitrary elements” \( m \) and \( h \) of the “normal curve of frequency of errors”

\[
f = \frac{h}{\sqrt{\pi}} e^{-h^2(x-m)^2}.
\]

Here \( h \) was standard—see §4 above—but \( m \) wasn’t, though it was a natural first letter rule choice.

Fisher’s solution, the “absolute criterion”, involved maximising the likelihood (the term came later—for details of this and other terminology associated with maximum likelihood, see Aldrich (1997)). In the case of the example, maximising with respect to \( m \) and \( h \) gives

\[
m = \bar{x} \quad 2h^2 = \frac{n}{\sum v^2}
\]

where \( n\bar{x} = \sum x \) and \( v \) is written for \( x - \bar{x} \). The symbols \( m \) and \( h \) represent both possible values and maximum values. Later in the paper Fisher (p. 159) uses \( \mu^2 \) for \( \frac{\sum v^2}{n} \) (cf. Brunt’s usage in §4) and thus he had two symbols at his disposal for both mean and dispersion from the beginning. The \( h \) and \( \mu \) combination did not last long but \( m \) and \( \bar{x} \) went on until 1935. (see §14 below).

One piece of notation really lasted: \( \theta \) for the generic unknown parameter. It is not clear why Fisher chose \( \theta_1, \ldots, \theta_r \) for the arbitrary elements of the frequency function, \( f \), instead of, say, \( c \). The symbol \( \theta \) was generally used for an angle and there was no angle in the 1912 paper. The choice must have contributed to the adoption of the Greek for parameter convention. In 1912 Fisher had no occasion to introduce a symbol for an estimate of \( \theta \).
Correlation, the subject of Fisher’s first distribution theory paper (1915), came from deepest biometry. The bivariate normal density (cf. §3) is de-centralised, \( r \) has become \( \rho \), there is no \( N \) and the “frequency distribution of the population [is] specified by the form”

\[
df = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} e^{-\frac{1}{2}(\frac{(x - \mu_1)^2}{\sigma_1^2} + \frac{(y - \mu_2)^2}{\sigma_2^2} - \rho \frac{(x - \mu_1)(y - \mu_2)}{\sigma_1 \sigma_2})} \, dx \, dy
\]

where “\( df \) is the chance that any observation should fall in the range \( dx \, dy \)”. The “\( df = \ldots \, dx \)” construction became Fisher’s regular way of representing “frequency distributions” or “probability distributions” or even “error functions” (1924/8); the symbol \( f \) had no independent meaning. Kendall (1943) modernised the Fisher format, writing “\( dF = \ldots \, dx \)” and, following Cramér (1937), referred to the Stieljes integral and the distribution function \( F \).

Fisher (1915, p. 508) defined the “statistical derivatives” (statistics) thus

\[
\begin{align*}
n\bar{x} &= \sum_{i=1}^{n} (x_i), & n\bar{y} &= \sum_{i=1}^{n} (y_i), \\
n\mu_1^2 &= \sum_{i=1}^{n} (x_i - \bar{x})^2, & n\mu_2^2 &= \sum_{i=1}^{n} (y_i - \bar{y})^2, \\
nr \mu_1 \mu_2 &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),
\end{align*}
\]

using the error theory symbols \( \bar{x} \) and \( \mu^2 \) and taking \( r \) and \( \rho \) from Soper. (After this paper \( \Sigma \) gave way to \( S \) in most of Fisher’s writing.) For the mean value of \( r \) Fisher wrote \( \bar{r} \).

Fisher had derived the \( s \) distribution independently and when he (p. 507) described Student’s result (Extract IV) it was in his own notation: the frequency is proportional to

\[
\mu^{n-2} e^{-\frac{\mu^2}{2s^2}} d\mu.
\]

Although \( \mu^2 \) was eventually supplanted by \( s^2 \) Fisher was still using it when he (1921, p. 5) treated the intra-class correlation and even in (1924, p. 100) when recalling earlier results.

Fisher (1920) took the language of Biometrika to the Astronomical Society, the national home of the theory of errors, when he challenged Eddington’s claim that \( \sigma \) is better estimated from absolute than from squared deviations. Fisher (p. 189) did not refer to the error curve or gaussian distribution but to the normal distribution, writing “the chance of any observation falling in the range \( dx \)” as

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{(x - \bar{x})^2}{2\sigma^2}} \, dx.
\]

For the rival methods of “determining” \( \sigma \) from the observations he wrote

\[
n\sigma_1 = \sqrt{\frac{\pi}{2}} S(|x - \bar{x}|) : \quad n\sigma_2^2 = S(x - \bar{x})^2,
\]

using linked notation for the first time indicating what is being estimated and how. Eddington’s claim revived an old issue in error theory and part of Fisher’s reply was in Helmert (1876) and even in Gauss—see David & Edwards (2001). The earlier authors did not use linked notation when they discussed the superiority of \( \sigma_2 \).

By 1920 Fisher’s language was predominantly biometric, though the influence of the theory of errors can be seen in details such as the definition of “frequency distribution”. Soon he was rewriting Student (1908) and Pearson & Filon (1898) in a new language. He worked on two fronts, the theory of estimation (§§10 & 11) and the theory of sampling distributions (§§12 & 13). The vocabularies embodied some common principles but their notations were not unified until the late 1930s (§14).
From the beginning Fisher had serviceable notation and notational uniformity, like the use of Greek letters for parameters, seems to have had low priority. Uniformity is aesthetically appealing but does not eliminate dictionary time entirely for it is always necessary to check the reference of ‘pronouns’ like $x$.

10 “Verbal Confusion”: Parameter and Estimate

Fisher’s “On the Mathematical Foundations of Theoretical Statistics” (1922) reconstructed Pearson & Filon. It propounded new estimation criteria and showed how maximum likelihood satisfies them—beating the method of moments even on its home ground of the Pearson curves. The specifically estimation ideas and vocabulary are discussed in Aldrich (1997) but there was one general concern that extended to distribution theory.

In a technical footnote Fisher (p. 329n) criticised Pearson & Filon for drawing “no sufficient distinction . . . between the population and the sample” but he (p. 311) had a larger point about language:

it has happened that in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems; for it is customary to apply the same name, mean, standard deviation, correlation coefficient, etc. both to the true value which we should like to know but can only estimate, and to the particular value at which we arrive by our method of estimation; so also in applying the term probable error, writers sometimes would appear to suggest that the former quantity, and not merely the latter, is subject to error.

It is this last confusion . . . more than any other, which has led to the survival to the present day of the fundamental paradox of inverse probability . . .

There is a dig here at the “probable errors of frequency constants”. The hypothesis about the survival of inverse probability was perhaps natural for Pearson had been confused and so had Fisher; see Aldrich (1997). Yet Student’s unconfused correlation paper was Bayesian in design: $R$ was there because Student wanted to make probability statements about it.

Fisher might have said that the method of moments flourished in a world where this “verbal confusion” prevailed. More generously he (p. 321) said that the method involved “the most extended use of the criterion of consistency”. Fisher introduced the term “statistic” and popularised “parameter”; his earlier linguistic experiments are described in Aldrich (1997). Fisher’s use of parameter accorded with an established meaning which the Oxford English Dictionary traces to the mid-nineteenth century viz. “a quantity which is constant (as distinct from the ordinary variables) in a particular case considered, but which varies in different cases”; see Stigler (1976) for a discussion. Pearson (1894, p. 104) had once used “parameter” in the sense of estimate. However “statistic” was a new word and, to some, grotesque: “Where . . . did you get that atrocity, a statistic?” asked Arne Fisher (Bennett (1990, p. 312)). Pearson (1936) could not understand Fisher’s point in introducing the parameter/statistic pair. Fisher’s statistics are usually estimates but the relevance to testing of the point about distinguishing estimates from parameter values was clear and reflected in his practice. However his only general statement about the importance of the distinction for problems of distribution is the passage in his obituary of Student quoted in §8 above.

11 The Theory of Estimation: $\theta$, $\hat{\theta}$ and $T$

For all its emphasis on language, the “Foundations” did not rationalise the existing notation. Fisher had used $\theta$ for an arbitrary element (parameter) but it had been his choice too for an angle, thus $\cos \theta = -\rho r$ appears throughout his correlation work. It appears in both roles in 1922. Although
Fisher used $\theta$ for the parameter to be estimated from his first paper, the symbol had to struggle to win that role. Hotelling (1930) was the first to follow up Fisher’s work on estimation and he wrote $\rho$, for parameter presumably. However $\theta$ was adopted by such influential contributors as Neyman (e.g. 1934) and Wald (e.g. 1939) and it has stayed.

The notation of the “Foundations” changes by the paragraph but there was one notational constancy—the use of the caret or hat for the maximum likelihood estimate: $\hat{\theta}$, $\hat{\rho}$ and $\hat{m}$ duly appear; $m$ had been extended beyond the normal mean to the Poisson distribution and to the “location” parameter family. The caret originated in the co-operative study of Soper, Young, Cave, Lee & Pearson (1917) conceived as a sequel to Fisher (1915). The co-operators (pp. 352–360) used $\hat{\rho}$ for the value Fisher (1915, p. 520) found by applying his absolute criterion. They interpreted this value as the maximiser of the posterior distribution of $\rho$ assuming an “equal distribution of ignorance”. Fisher (1921, pp. 9 & 13) emphatically rejected this Bayesian interpretation but kept $\hat{\rho}$ for his own “optimum value”. In the “Foundations” Fisher used the hat for the maximum likelihood value but he altered the definition of likelihood; in 1915 and 1921 this had been based on the sampling distribution of $r$. The particular symbol $\hat{\rho}$ now had no role because on the new definition the maximum likelihood value of $\rho$ was the old $r$!

Fisher’s “Theory of Statistical Estimation” (1925b) also used hats. The (1925b) and (1922) were Fisher’s most influential contributions to estimation and so perhaps the future of $\hat{\theta}$ was assured. Fisher, however, gave up $\hat{\theta}$ for $T$. He (1925b) first introduced $T$ because he wanted to separate in notation the maximum likelihood value $\hat{\theta}$ and the sufficient statistic—in 1922 he thought they were the same. $T_1$ stood for a statistic “sufficient in estimating” $\theta$ (1925b, p. 713) and $T_2$ for any other statistic. In the chapter on estimation added to the 1928 edition of the Statistical Methods Fisher compared five estimation methods, $T_1$, .., $T_5$. Five affixes to $\theta$ could have achieved the same effect but Fisher was warming to a combination of the first letter rule and the Graeco-Latin convention: $T$ for theta. In Fisher’s last presentation of the theory of estimation (1956, pp. 163–9) $T$ represents the maximum likelihood estimate of $\theta$. The symbol $T$ is now associated with a post-Fisher convention: upper-case letter for a random variable, lower-case for the value it takes. Wilks (1944) used $X$ for the random variable and $x$ for the argument of the distribution function but the convention became popular only later—perhaps through the example of Feller’s Introduction to Probability Theory (1950). $T$ has lived on as a symbol for a statistic with $\hat{\theta}$ for an estimate. In Lehmann (1959, p. 36) a statistic $T$ is defined as a certain kind of transformation, suggesting a straight application of the first letter rule.

12 “Error Functions of Well Known Statistics”

In the early 20s Fisher found new applications for Student’s $z$, corrected and extended Pearson’s $\chi^2$ theory and found a new organising principle for distributions. With this went a search for a suitable language in which to express these discoveries. At first Fisher used the Pearson curves to catalogue the distributions. He seemed very taken with the system, perhaps because he had found new depths to it. The “Foundations” (1922, p. 314) praised “a body of work which has enormously extended the power of modern statistical practice” and over a third of the paper is given over to the curves. Yet by 1924 Fisher had abandoned typology. In the great summary paper Fisher’s own $z$ distribution was “yielding the error functions of several well known statistics”—not Pearson’s Type VI.

Fisher reworked the $\chi^2$ theory in a series of papers beginning with (1922a). Student used the $\chi^2$ criterion to assess his curve fitting but he did not notice any analytical connections with the $\chi^2$ distribution of 1900. The technique Pearson used was quite different from Student’s and more like Fisher’s—instead of curve-fitting there is a change of variable effected by a geometric construction. The distribution problem was conceptualised as finding a probability—“the probability of a complex system of $n$ errors occurring with a frequency as great or greater than that of the observed system”.

---

122  J. ALDRICH
The answer was given as

$$P = \frac{\int_x^\infty e^{-\frac{1}{2}x^2}x^{n-1}dx}{\int_0^\infty e^{-\frac{1}{2}x^2}x^{n-1}dx}.$$  

Pearson did not isolate the density, neither then nor later (1914, 1916a). Fisher (1924/8) did isolate the density and introduced the expression “Pearson’s $\chi^2$ distribution”.

The “errors” of Pearson (1900) are jointly normal and Pearson wrote their frequency surface as a constant times $e^{-\frac{1}{2}x^2}$ as he had in earlier papers. The multinormal came about as a large sample approximation to the multinomial and Pearson also used the symbol $\chi^2$ for the quantity $S \left( \frac{e}{m} \right)$ where the errors $e$ are the discrepancies between “observed frequencies” $m$ and “theoretical frequencies” $m'$. The fixing of the symbol $\chi$ in the $\chi^2$ complex came much later—for Pearson, $\chi$ was a utility symbol like $x$.

Pearson was concerned with evaluating $P$ for different values of $\chi$ and $n$. He did not place the new distribution in the curve system and nothing would have been gained if he had. The first of Fisher’s papers (1922a) examined contingency tables and here quite incidentally he mentioned that Pearson’s 1900 result was that the distribution of $\chi^2$ is “given by the Pearsonian curve of Type III”. Not that Fisher or Pearson reached it by the classical curve-fitting route. Fisher agreed with Pearson’s identification of type but corrected him on the number of degrees of freedom—a concept Fisher took from mechanics; see for example Jeans (1904, chapter V).

When Fisher examined the application of the $\chi^2$ distribution to regression goodness of fit he (1922b) found that the type was wrong—it should have been a Type VI, in effect a rescaled $F$. In the same paper Fisher indicated the link between Pearson’s $\chi^2$ and the $s$-distribution when he wrote $\chi^2 = (n - 2) \frac{s^2}{\sigma^2}$. He also mentioned that Student’s $z$ was a Type VII curve. These Type designations had the advantage that the constants bore no special interpretations associated with the number of observations or the number of cells. They brought little else beyond unhelpful associations—their moments were known but their destiny as curves-to-be-fitted did not entail that their distributions be tabulated.

Fisher (1924/8, 1925 and 1925b) gave a new importance to Student’s work but de-canonicalised his $z$. From 1919 Fisher had been using $z$ for a transformation of the correlation or other measure of resemblance, $z = tanh^{-1} r$. He was not then particularly interested in Student’s problem and in any case there seemed to be no danger of confusion. Fisher kept using $z$ and carried it into the analysis of variance where there was some danger of confusion. So he introduced a completely new symbol $t$ related to Student’s $z$ by $t = z \cdot \sqrt{n' - 1}$. Fisher (1925a, p. 91) wrote $t$ in terms of its components

$$t = (\bar{x} - m) \sqrt{\frac{n'}{s}}.$$  

Fisher took to referring to the distribution of $t$ as “Student’s distribution” and Student endorsed the symbol $t$ when he (1925) tabulated this quantity in preference to his own $z$.

Fisher used the root symbol $n$ for the number of degrees of freedom and wrote $n'$ for the number of observations—a practice he continued through all editions of the Statistical Methods. The number of degrees of freedom was the fundamental quantity because it was used to define the $\chi^2$ and $t$ distributions. His use of $n$ and $n'$ is reminiscent of Pearson (1900) where the distribution of the quadratic form in $n$ jointly normal variables appears before the demonstration that this normal could be obtained from a multinomial with $n + 1$ cells. Pearson’s tables for $\chi^2$ were in terms of the number of cells, $n' = n + 1$. Eisenhart (1979, p. 8n) describes how in the late 1930s E.S. Pearson and other “not Fisherian” statisticians rejected the use of $n$ for the number of degrees of freedom and introduced the Greek $\nu$.

Fisher’s architectonic “On a distribution yielding the error functions of several well known statis-
tics” (1924/8) presents the normal, the $\chi^2$ and Student’s distribution as special cases of the $z$-distribution. (Following Snedecor (1934) $z$ gradually gave way to $F = \sigma^2z$.) It could as well have been called “the normal distribution and related distributions” but on either view there was no need to mention the Pearson system. Nor does the system appear in Fisher’s applications book, the Statistical Methods. Of course Pearson-trained statisticians knew the language of the curves and they continued to refer to them but two post-war developments illustrate how the curves were marginalised. Elderton’s Frequency-Curves and Correlation (1906) had been an introduction to statistics from a Pearsonian viewpoint but its successor Elderton & Johnson’s Systems of Frequency Curves (1969) was a specialist book of curves. Weatherburn’s First Course in Mathematical Statistics (1946) treats the gamma and beta distributions as primary entities and gives their Type aliases in footnotes. The exponential family of distributions, the distributional complement to Fisher’s estimation theory, was introduced in 1935/6 but only became part of mainstream theory much later.

13 Graeco-Latin Stirrings

In his work on distributions Fisher used, not hats, but the Student–Soper pairings $\{\sigma^2, s^2\}$ and $\{\rho, r\}$ and new ones for regression, $\{\alpha, a\}$ and $\{\beta, b\}$. The regression symbols appear in the “Goodness of fit of regression formulae, and the distribution of regression coefficients” (1922b). The notation of the goodness of fit part—such as $x$ and $y$ for dependent and independent variable—derived ultimately from Pearson (1905). The specification and methods of the “distribution of regression coefficients” came from error theory but the language, including “regression”, was from biometry. Pearson (1896) had already used $\beta$ for a regression symbol but Yule (1907) reorganised the multiple regression notation around elaborately subscripted $b$’s. Fisher (1925, §29) preferred the simpler $\beta_1, \beta_2$, etc. As Pearson and Yule worked with deviations from the mean they had no symbol corresponding to $\alpha$. There is more on regression specifications and notations in Aldrich (1998).

Fisher (1925, p. 117) represented “the true regression formula which we should obtain from an infinity of observations” by $\alpha + \beta (x - \bar{x})$ and wrote the fitted value as $Y = a + b (x - \bar{x})$ where the “coefficients $a$ and $b$ are calculated by the equations”

$$a = \bar{y}, \quad b = \frac{S(y(x - \bar{x}))}{S(x - \bar{x})^2}.$$

Of course for these least squares/maximum likelihood values the symbols $\hat{\alpha}$ and $\hat{\beta}$ would have been equally legitimate. In a 1955 paper Fisher (p. 71) used expected values and—defining $\alpha$ and $Y$ differently—wrote $Y = E_i(y) = \alpha + \beta x$ and $E_i(y - Y)^2 = \sigma^2$ but this expectation notation was never incorporated into the Statistical Methods.

Fisher (1922b) extended Student’s $s^2$ and $z$ to regression and took the first step towards canonising $\{\sigma^2, s^2\}$. However it was a redefined $s^2$ for Fisher (1925a, p. 91) substituted

$$s^2 = \frac{S(x - \bar{x})^2}{n - 1},$$

for Student’s biometric formula with the number of observations as the divisor. For the regression case Fisher used a degrees of freedom based formula for $s^2$.

14 Moments, Cumulants & the Graeco-Latin Triumph

Curve-fitting was not important for Fisher. He disparaged the method of moments and did not use moments in his research, yet they stayed—like unwanted step-children. The moment definitions in the Statistical Methods (1925, p. 72)—“moment coefficient” was quietly dropped—do not distinguish around the mean and around the origin quantities; the same symbols are used for moment statistics and the “moments of theoretical distribution”. Nor did Fisher distinguish population and sample quantities when he discussed testing departures from normality. His symbols were variants of Pearson’s $\beta$’s:
\[ \gamma_1 = \mu_3 / \mu_2^{3/2} = \beta_1^1 \text{ and } \gamma_2 = \mu_4 / \mu_2^2 - 3 = \beta_2 - 3. \] The rationale seems to be, \( \gamma \)'s are not \( \beta \)'s—for which he had other work—but are like them.

Fisher eventually became interested in moments—in a negative way—and 5 editions later the notation was sorted out. His “Moments and product moments of sampling distributions” (1929) aimed to make moment calculations more tractable by introducing “cumulative moment functions” and a combinatorial technique to go with them.

Fisher (1929) wrote the “moment generating function” (after Romanovsky presumably):

\[
M = \int e^{t \phi(x)} dx = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \ldots
\]

with \( \mu_r = \int x^r \phi(x) dx. \)

The moment generating function and the characteristic function gave English writers an alternative to the geometric method Fisher had been using in distribution theory (1915, 1928 etc). The terminology took a while to stabilise, for the “moment generating function” was also used for the characteristic function, e.g. by Bartlett (1933). The history of these methods, which goes back at least to Laplace, is traced by Hald (1998).

Fisher expands the logarithm of \( M \) in powers of \( t \) to give

\[
K = \log M = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \ldots
\]

The “cumulative moment functions \( \kappa \)” are so named because \( \kappa_r \) involves only \( \mu_r \) and lower orders”. The Graeco-Latin convention is deployed for the statistics corresponding to \( \mu_r \) and \( \kappa_r \). Writing \( m_r = \frac{1}{n} \sum (x - \overline{x})^r \), he explains that the “mean value of \( m_r \) is not in finite samples equal to \( \mu_r \)”.

Fisher (p. 203) introduces the statistics \( k_1, k_2, k_3, \ldots \) “functions of the symmetric functions, of which the sampling means shall be \( \kappa_1, \kappa_2, \kappa_3, \ldots \)”. The pair \( \{ \mu_r, m_r \} \) is sanctioned by usage but \( \{ \kappa_r, k_r \} \) embodies an improved order.

Fisher introduced cumulant notation into the 1932 (fourth) edition of the Statistical Methods: “the inconveniences of the moment notation seem now definitely to outweigh the advantages formerly conferred by its familiarity.” (p. xi). In fact moments stayed (pp. 73-4) but the notation was reorganised in the ‘obvious’ way with \( \mu_1, m_1, \ldots, \mu_1', m_1', \ldots \) etc. In §14 he (p. 78) remarks that the Greek symbols are “best used not for statistics but for the parameters of which \( g_1 \) and \( g_2 \) are estimates.”

The \( \gamma \)'s and \( g \)'s are examples of Fisher’s idiosyncratic first letter rule; \( \gamma \) and \( g \) are also paired in (1934, p. 292) but in a different context.

In 1935 Fisher switched to \( \mu \) for the mean of the normal distribution—see e.g. his Design of Experiments. The symbol has become so familiar that it may seem odd that it was not adopted decades earlier. In the 1890s the contraction of \( \mu_1 \) to \( \mu \) may well have seemed a blundering break with the system and at the next promising moment (1912) \( \mu \) was already committed—to the standard deviation. Fisher’s new pairing was \( \{ \mu, \overline{x} \} \); the pairing \( \{ \mu, m \} \) would have been more logical but perhaps too confusing.

Disregarding the anomalous \( \overline{x} \) and granting correspondent status to \( g \)'s and \( T \)'s, Fisher’s notation was now thoroughly Graeco-Latin. The twin alphabets were deployed to greatest effect in Bartlett’s “Theory of Statistical Regression” (1933) where parallels were found for the \( r \)'s and \( \sigma \)'s of Yule’s “new system” of 1907. Kendall (1943, p. xii) went “where possible” for the Graeco-Latin system, though he admitted “Complete notational consistency can only be achieved at the expense of jetisoning a great deal of accepted statistical usage, and even then would probably result in some cumbersome symbols.” The Graeco-Latin convention gives a special status to a particular estimator. In the Fisher pairs \( \{ \sigma^2, s^2 \}, \{ p, r \}, \{ \alpha, a \} \) and \( \{ \beta, b \} \) there is no single principle behind the choice: three are maximum likelihood values while \( s^2 \) is preferred to the maximum likelihood value because it is
linked to degrees of freedom.

In 1940s Britain the Graeco-Latin system may have seemed the logical culmination of a long history but in 1965 it was not even mentioned in the *American Statistician*’s “Recommended standards for statistical symbols” (Halperin, Hartley & Hoel). The Greek for parameters rule was recognized, as were established pairs like $\sigma^2$ and $s^2$ but the recommended designator for an estimate was the caret, the device Fisher picked up then dropped.

15 Summary and Sequel

The language described above originated in the mid-1890s in an applied mathematician’s study of biological problems. I have focussed on the development of a language for sampling problems, those arising in frequentist inference, not in sample survey analysis in which the biometricians took little interest. The original activity was curve-fitting but Pearson had to extend the language to accommodate sampling. The language took a new turn with Student’s papers of 1908 and another when Fisher introduced some important generic terms in the 1920s. Overall it took about forty years for a systematic notation—the Graeco-Latin—to materialize. The details of the language explosions of the 1890s and 1920s are very different. Terms arrived in great packages but, while Pearson devised a notation which stayed essentially the same with ad hoc additions, Fisher started by improvising upon the existing notations and only gradually developed a systematic notation.

Yule (1907, p. 182) commended his “new system” of correlation notation because many new results were “suggested by the notation itself”. The present paper has not been concerned with notation that aims so high but with the more primitive activity of naming. However, names are usually chosen to be suggestive as well as to refer effectively. Some names have survived although the suggestion has ceased to be functional, as with “moment”, or become embarrassing, as with “normal”—see Pearson (1920, p. 25). Some expressions were full of dangerous suggestions, e.g. “summing for all samples and dividing by the number of samples”, though this one seems to have been as harmless as Student’s labelling of papers that were not about probable errors as “probable error of . . .”.

For Fisher language mattered and (see §8 and §10) he criticised the biometric language for not suggesting important distinctions or perhaps for suggesting they were not important. He thought that linguistic deficiencies had led to mistakes and that for effective work insight of the highest order was needed. These claims are hard to judge because it is difficult to know how much of the confusion was due to faulty language and how much to other causes; it is even harder to decide how much research was not done because the demands on insight were too great. The work of Helmert shows that the device Student introduced for dealing with the distribution of $s^2$ and which Fisher developed was not essential. Quite different languages can be used to express the same content.

Linguistic innovations usually appeared first in papers and were passed on through teaching and other personal contacts. Codification also went on through published works such as Pearson’s volumes of tables, his *Biometrika* editorials and at an introductory level, the textbooks Elderton (1906) and Yule (1911). There was no advanced textbook covering the work of the biometric school nor one covering Fisher’s work. However in the 1940s three advanced textbooks appeared—Kendall’s *Advanced Theory*, Cramér’s *Mathematical Methods* and Wilks’s *Mathematical Statistics*. They represent the intellectual traditions described above, different ways of negotiating the pure/applied mathematics divide. Kendall—like Pearson and Fisher, a Cambridge-trained applied mathematician—was most faithful to the English tradition; random variables arrive in chapter 7 and the expectation comes after moments and cumulants. Wilks was an outsider, a PhD student of Rietz who came into London and *Biometrika* when he collaborated with E.S. Pearson in the early 30s. Cramér represented the Continental mathematical tendency, coming after Chuprov, one of the generation that “created modern probability” (see von Plato (1994)); his English connection was with Hardy. By devoting the first half of the *Mathematical Methods* to probability he ended the separation that had prevailed in the English literature since Airy. By the end of the 40s the *Annals of Mathematical Statistics* had
replaced *Biometrika* as the leading theory journal for statistical theory.

**Acknowledgement**

I am grateful to the Editor and the referees for their comments.

**References**


Miller, J. (Ed.) (continuing) Earliest Known Uses of Some of the Words of Mathematics on the website:
http://members.aol.com/jeff570/mathword.html
Mille, L. (1923) (continuing) Earliest Uses of Various Mathematical Symbols on the website:
http://members.aol.com/jeff570/mathsym.html
Pearson, K. (1874–7). Lecture Notes on the Theory of Probability, held in Manuscript Room University College London Library list number 46.
Pearson, K. (1900). On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be reasonably Supposed to have Arisen from Random Sampling. Philosophical Magazine, 50, 157–175.


Student (1909). The Distribution of the Means of Samples which are not Drawn at Random. *Biometrika*, 7, 210–214.


**Résumé**

Cet article examine le langage inventé par Karl Pearson et ses associés pour discuter les distributions, les populations et les échantillons, le langage de base pour l’influence par les fréquences. Le langage original—une partie duquel est toujours en usage—est décrit ainsi que les changements qu’il a subis sous l’influence de R.A. Fisher et des mathématiciens russes et américains. La période couverte va environ de 1890 à 1950.

[Received January 2002, accepted September 2002]