

From 1809 and to 1925—
two essays on the history of the theory of errors

John Aldrich
Economics Division
School of Social Sciences
University of Southampton
Southampton
SO17 1BJ
UK
e-mail: john.aldrich@soton.ac.uk

Abstract: These essays extract two threads from the complicated history of the theory of errors in the 19th and early 20th centuries. The first extends from the *Theoria motus corporum coelestium* (1809) of Gauss through a piece by his student Richard Dedekind (1860) to a paper by Carl Lüroth (1876). The second extends back from Fisher's *Statistical Methods for Research Workers* (1925) to Student (1908) and Fisher (1912) to 19th century textbooks. The Gauss-Dedekind-Lüroth line produced Bayesian analysis for the normal regression model while the Fisher line developed sampling theory analysis. The first line worked on problems from astronomy and geodesy, the second on problems from biology and agriculture. The second line saw the merging of the theory of errors and biometry into a new discipline, mathematical statistics.

July 2011

General introduction

The work of Gauss and other error theorists was recalled in two landmark writings on the history of statistics: Plackett's (1949) "Historical note on the method of least squares" showed that Gauss—and not Markov—was "the first who justified least squares as giving those linear estimates which are unbiased of minimum variance" while Seal's (1967) "Historical development of the Gauss linear model" showed how "many of the mathematical results of least squares (or error) theory" were rediscovered by Fisher and his associates at Rothamsted Agricultural Station. These recollections were necessary because the subject was in new hands: Plackett and Seal published in *Biometrika*, a general statistical theory journal originally founded for the study of biological problems.

The error theory literature was substantial: already in 1877 Merriman recorded 408 contributions, including 153 in German, 110 in French, 90 in English and 16 in Latin. Stigler (1986), Sheynin (1996), Hald (1998), Farebrother (1999) and others have retrieved much of this material as well as recording some of the later growth. This international subject had distinct national traditions and there was diversity both in set-up and in inference principles. The two essays below treat Bayesian and sampling theory inference for the normal regression model and for random sampling from a normal population; other approaches to fitting relations to data are considered by Farebrother (1999). The essays engage primarily with the German tradition—Essay I literally and Essay II at one remove through an English offshoot. The French and English contributions listed by Merriman did not much impinge on the developments traced below; one of the French lines is followed by Heyde and Seneta

(1977).

Essay I begins with Gauss's *Theoria Motus Corporum Coelestium* (1809), which appeared 60 years after the first of Merriman's contributions, and traces the elaborations of its Bayesian analysis by Dedekind and Lüroth, apparently to go no further. Essay II works back from the post-astronomy, post-geodesy world of Fisher's *Statistical Methods for Research Workers* (1925a). This did not begin as a purple line—it came out of textbook teaching derived from German models—but it produced some of the most influential ideas in 20th century mathematical statistics. These were sampling theory ideas and when Plackett wrote only one person was in sight working in the Bayesian line from Gauss, Harold Jeffreys whose *Theory of Probability* had been received coolly in 1939; neither Plackett nor Seal could have foreseen how this approach would rival Fisher's sampling theory approach.

The essays below treat separate eras and different inference principles. They also exhibit contrasting forms of development: in one, there appears to have been an inexorable process in which the potential of the original scheme is realised—it was work within a paradigm; in the other a paradigm was created out of confusion. The two lines do not connect but they are related for Fisher's sampling theory development had its origins in Student's attempt to solve a problem that had already been solved in the Dedekind-Lüroth Bayesian development. The Student-Fisher collaboration was an unusual one for there was no agreement on fundamentals. A clean sampling theory path back from Fisher through Pizzetti (1891), to Helmert (1876) to the earliest sampling theory work can be plotted—see Hald (2000)—but as virtual, not as real, history.

Essay I: Lüroth, Dedekind and the Bayesian line in the theory of errors

1 Introduction

One of the achievements of statistics in the early 20th century was to develop a comprehensive inference theory for the normal regression model and for random sampling from a normal population. In fact two theories were developed, one on sampling theory principles and one on Bayesian. Their architects, Ronald Fisher and Harold Jeffreys, were not students of history and it has since been shown that many of the new results could be found in the old literature on the theory of errors—see for instance Seal (1967) and Hald (1998).

Sections 2-4 examine four Bayesian contributions published between 1809 and 1876. Gauss's contributions of 1809 and -16 never disappeared from view but Lüroth's of 1876 was rescued from obscurity by Pfanzagl and Sheynin (1996) who hailed it as a forerunner of the t -distribution. The fourth contribution is another rescue, an 1860 paper by Richard Dedekind, a student of Gauss, and the link between Gauss and Lüroth. Finally Section 5 considers related contributions by other authors and asks why these contributions were not consolidated, why there was no broad, long Bayesian line before Jeffreys.

2 An incomplete system: Gauss 1809 and -16

19th century astronomers devised two schemes for the reduction of observations based on the law of error, one for *direct* observations and one for *indirect* observations. A modern textbook, like Gelman et al. (2003), formulates them so

$$y_i | \mu, \sigma^2 \sim IN(\mu, \sigma^2) \quad (2.1)$$

$$y | \beta, \sigma^2, X \sim N(X\beta, \sigma^2 I). \quad (2.2)$$

In the 19th century they were not written like this. The symbols and language of “normality” and “regression” came from the English biometric school and the matrix formalism from A. C. Aitken; for these transformations see Essay II below and Aldrich (1998, 2003, 2005a) and Farebrother (1997). I will use the modern formulations without further comment.

Scheme (2.2) came first, in the *Theoria motus corporum coelestium* (1809) of Carl Friedrich Gauss (1777-1855). Gauss treats the precision of the observations (measured by $h = 1/\sigma\sqrt{2}$) as an unknown constant and makes inferences *only* about the regression coefficients. Gauss (1809, §§179-80: 255-61) assumes a uniform prior for the regression coefficients and shows that the maximum posterior values are given by the method of least squares. He (§182: 264-66) further showed that the marginal distribution of an individual coefficient is normal with a precision that can be expressed in terms of the unknown precision of the observations. This ‘precision formula’ is an important part of the story. For the *Theoria motus* and Gauss’s work in the theory of errors generally see Seal (1967), Stigler (1986), Hald (1998, 2007) and Sheynin (1996, 2009); the standard life is Dunnington (1954).

A second astronomer, Friedrich Wilhelm Bessel (1784-1846), gave a method for estimating the precision of an observation and a method for stating the precision of “das Endresultat”. Bessel’s articles (1815 and -16) contain no theoretical development comparable to that in the *Theoria motus* and concepts and methods simply appear with the empirical results based on them; Hald’s (1998: 360-1) commentary fills some of the gaps. Bessel (1815: 234) introduced the probable error (“der wahrscheinliche Fehler”) as a measure of precision: for the error curve or normal distribution the probable error gives the limits either side of zero between which 50% of the errors fall. Bessel (1816: 141-2) discusses other measures of dispersion, including the mean absolute deviation (ε) and the root mean square deviation (ε') and indicates their relationship to each other and to the probable error (ε'') and Gauss’s h :

$$\varepsilon'' = 0.6745\varepsilon' = 0.8453\varepsilon = 0.4769\frac{1}{h} \quad (2.3)$$

for the case of the error curve. These numbers were as familiar in the 19th century as 1.96 in the 20th.

To estimate the probable error of an observation Bessel used the formula

$$0.8453\frac{\sum |y_i - \bar{y}|}{m} \quad (2.4)$$

presumably by analogy with the relationship between ε'' and ε in (2.3); Bessel does not say so or indicate why he chose this particular relation. Of course the squared deviation formula involves heavier calculations.

Bessel used the estimated probable error of an observation and the precision formula to obtain the probable error of an estimate: Bessel (1815) treated the case of direct observations,

i.e. scheme (2.1), and the “Endresultat” is the mean. For a sample of size m the precision formula takes the form, the probable error of the mean is $1/\sqrt{m} \times$ the probable error of an observation. Bessel does not derive the formula but it follows easily from Gauss’s (1809, §179-80: 255-61) analysis of (2.2) for the case of a single coefficient. To calculate the probable error of the mean Bessel multiplies the estimated probable error of an observation (from (2.4)) by $1/\sqrt{m}$; as we see in Section 4 below, the procedure of plugging into the precision formula and the analogous one for scheme (2.2) would be widely adopted.

Gauss (1816) responded to Bessel with both a Bayesian analysis and a new non-Bayesian argument. His analysis focused on precision—“knowledge of h is interesting and instructive” he wrote—and involved no mean or regression coefficients. The treatment (1816, §§1-2: 41-2) begins with the error curve and the probable error, r , defined in terms of its quantiles. The Bayesian argument (§3: 42-3) follows the earlier pattern: the posterior distribution and the most probable value of h and of the probable error (r) are found assuming a uniform prior for the parameter h and the law of error for the observations. The most probable values depend on the sum of squared errors, not on Bessel’s absolute errors. Gauss (§4: 43-4) finishes the Bayesian analysis by developing a large sample normal approximation to the posterior and constructing 50% limits for h and r : it is an “even bet” that the true values lie between these limits. Gauss then changed direction and—under the influence of Laplace—presented a new inference theory based on sampling theory comparisons of alternative estimators of precision (§§5-8: 45-9). These large sample results also favoured the squared error estimate; see Hald (1998: 455-9) and Sheynin (1979) for details.

These contributions of Bessel and Gauss had the effect of making the estimation of the accuracy of observations an important issue in the theory of errors—after the proof of the principle of least squares for the estimation of the coefficients, the *most* important. Although Gauss gave two justifications for the squared error estimate, he appears not to have adopted, or even commented on, the practice of plugging in the estimated probable error of an observation into the precision formula to obtain the probable error of the final result. This did not change when he went on to produce a system of sampling theory.

Gauss's *Theoria combinationis observationum* (1823) brought a new sampling theory justification for the method of least squares free from a hypothetical form of error distribution and valid for all sample sizes; the first consideration made it an improvement on the *Theoria motus*, the second an improvement on Laplace's work of which Gauss was aware. The new work contained sampling theory inferences for all the parameters, Chebyshev inequality-like results (§§ 9-10: 13-19) for the magnitude of errors, the "Gauss-Markov" proof of least squares (§§ 19-21: 37-45) and a new estimate of precision based on dividing the least squares residuals by the sample size less the number of regression coefficients (§§ 37-8: 83-9). The precision formula is still valid (§ 21: 45) although it now relates a parameter of the sampling distribution of an estimator to a parameter of the error distribution; see Hald (1998: 459-484).

Coming to Gauss from Fisher or from Jeffreys, it is striking how little he made of the change from the earlier Bayesian approach to the later sampling theory approach. Gauss (1816: 45) introduced the sampling theory argument with the remark, "the matter may also

be viewed from a different angle” and the *Theoria combinationis* (§ 17: 31-33) emphasises the generality of the new approach. In 1838 Bessel published a new exposition of the law of error based on the central limit theorem and Gauss responded with some reflections on the theory of least squares. Writing to Bessel, Gauss (1839: 146) used the term “metaphysics” in connection with the Bayesian method but whether this was a criticism as it would have been for Fisher or a neutral observation as it would have been for Jeffreys is unclear.

Apart from the new estimate of precision the *Theoria combinationis* did not aim to change practice but rather to provide better—or, at least, additional—justifications for the old practices. While Gauss’s only subsequent publication on the theory of errors was the 1828 supplement on geodesy, he went on discussing and teaching the subject for the rest of his life: there are extracts from his correspondence in the *Werke* and there is a list of the courses he taught in Dunnington (1954: 405-10). In the class of 1850-1 was a first year undergraduate, Richard Dedekind, who recalled the experience fifty years later in a well-known memoir (1901); Dunnington (359-361) translates an extended excerpt. However, in an earlier and less noticed article Dedekind (1860b) had extended Gauss’s Bayesian theory.

3 Completing the system: Dedekind 1860

Dedekind would become “one of the greatest mathematicians of the nineteenth-century, as well as one of the most important contributors to number theory and algebra of all time”—Reck (2008: 1); for biographical information see Landau (1917), Biermann (1971) and Knus (1982). Dedekind’s doctoral dissertation on Eulerian integrals was written under Gauss;

though Gauss mostly lectured on astronomy, he supervised some mathematical students. The new doctor stayed on in Göttingen, teaching and attending the lectures of Gauss's successor: G. L. Dirichlet, another great number theorist. Probability was one of the subjects Dedekind taught and he (1855) published a paper on a controversy between Cayley (1853) and Boole (1854). The paper was noticed in Czuber's (1899) history of probability and more recently by Dale (1999: 384-6).

In 1858 Dedekind was called to the Zürich Polytechnikum (now the ETH) as professor of mathematics and among five rather miscellaneous communications to the *Züricher Vierteljahrsschrift* were one on probability (1860a) and one on least squares (1860b). Unlike Gauss and Bessel who were astronomers and geodesists, Dedekind had no practical interest in least squares. His interest was in mathematical organisation and he found stimulation both in what he was taught and what he had to teach. The "Dedekind cut," his best known creation, originated in his efforts to teach calculus and the editing of other mathematicians, particularly of Dirichlet, proved most productive. As Dedekind (1901: 295) recalled, it was Gauss's custom to teach both theories of least squares and Dedekind's least squares paper (1860b) tidied up the first theory. Dedekind has only one non-Gauss reference, to an exposition of least squares by Wittstein (1849) in an appendix to a calculus textbook.

The main aim of Dedekind (1860b) was to extend Gauss's Bayesian (1816) reasoning on the accuracy of observations to the case of indirect observations. In Dedekind's (1860b: 96) formulation of (2.2) there are m observations, k_1, \dots, k_m (the y of (2.2)), on linear functions, v_1, \dots, v_m , of the n unknowns x, y, z, \dots (the β vector) with known coefficients (the X matrix);

the errors are normally distributed and the prior distribution for x, y, z, \dots and h is uniform.

For the sum of squared errors Dedekind (96) writes

$$\Omega = (k_1 - v_1)^2 + (k_2 - v_2)^2 + \dots + (k_m - v_m)^2.$$

The minimum value of Ω , i.e. the sum of squared residuals, is denoted by Ω_0 .

Dedekind (1860b: 97) argues that it cannot be correct to take Gauss's formula for the most probable value of h for the case of known errors, viz. $\sqrt{m/2\Omega}$, and replace Ω by Ω_0 because when $m = n$, the magnitude Ω_0 is 0 and the most probable value of h is infinitely great. So Dedekind goes back to first principles. The joint density of the m observations is proportional to

$$h^m e^{-h^2\Omega} \tag{3.1}$$

which is also the posterior density for all the parameters given the uniform prior Dedekind explicitly assumes. He then integrates out x, y, z, \dots to give the posterior density of h , a function proportional to

$$h^{m-n} e^{-h^2\Omega_0}. \tag{3.2}$$

From which the maximum probability value of h is seen to be

$$\sqrt{\frac{m-n}{2\Omega_0}}. \tag{3.3}$$

Having reached formula (3.3), Dedekind (1860b: 83) could have finished but he went on to derive the marginal posterior for a regression coefficient to show that the least squares value is the maximum probability value for each coefficient separately. For this limited objective

a very schematic treatment sufficed. Dedekind writes the sum of squared residuals as

$$\Omega_0 = Y^2 + Z^2 + \dots + X^2 + \Omega$$

where Y^2 is a function of all n unknowns, Z^2 a function of $n - 1$ unknowns ... and X^2 a function of one unknown, x . This is the triangular form corresponding to Gaussian elimination (from Gauss (1809)) and the least squares values are found by solving the equations $Y^2 = Z^2 = \dots = X^2 = 0$. Dedekind integrated out the variables h and $y, z, ..$ from the joint posterior (3.1) and obtained the marginal density for x as a constant times

$$\frac{1}{(X^2 + \Omega_0)^{\frac{m-n+2}{2}}}. \tag{3.4}$$

This corresponds to Student's distribution with $m - n + 1$ degrees of freedom. Dedekind's only interest in the function was the location of the maximum which is clearly at $X^2 = 0$, so that the posterior marginal has its maximum at the least squares value.

Dedekind's 1860 paper extended the Bayesian analysis of Gauss (1816) by introducing regression coefficients and presenting a new estimate of precision based on maximising the marginal distribution of precision. In addition he found the marginal posterior of a regression coefficient and confirmed that the maximum value is given by least squares—thus giving solutions to the two biggest problems in the theory of errors. After 1860 Dedekind published no more research on probability or least squares. His article gets a sentence from Merriman (1877: 206), “The most probable value of the measure of precision h is found to be $\sqrt{\frac{n-1}{2\sum x^2}}$ and not $\sqrt{\frac{n}{2\sum x^2}}$ ” and so he must have perused it.

4 Probable errors: Lüroth 1876

Gauss's least squares teaching had fallen into a pattern in the 1820s and generations of students had experienced it before Dedekind and perhaps the surprising thing about his contribution was that it was so long in coming. While Dedekind's paper seems to have come out of a Göttingen time capsule, its sequel, Lüroth (1876), reflected changes in the wider world, among them the appearance of textbooks and the development of a journal literature where disputed points could be considered.

In 1832-4 the astronomer Johann Franz Encke (1791-1865) published a series of papers forming, in Merriman's (1877: 180) words, "a treatise on the Method of Least Squares, from which many text-books have been compiled." The textbooks Student and Fisher used decades later—see Essay II—descended from Encke. Encke had studied with Gauss in 1811-13 and, though he acknowledged the work of Laplace and Poisson, he (1832: 317) held that the writings of Gauss and Bessel "combine the strictest theory with the happiest practical applications of theoretical truths."

Encke picked and mixed from Gauss and Bessel. To derive least squares he (1832: 339-41) revived Gauss's original combination of normal errors and inverse probability but, for estimating the precision of observations, he (345) followed the sampling theory of Gauss's *Theoria combinationis*. Encke took from Bessel the 'sample mean \pm probable error' scheme, based on the precision formula but with Gauss's estimate of precision as plug-in. Encke's (1832: 347-50) worked example considered the value of the easterly deviation of the path of falling bodies due to the rotation of the earth. The mean of the 29 observations, and thus the

most probable value of the deviation, is $5'''.086$. The probable error of the mean is $0'''.950$ and thus it is “an even chance that the true deviation lies between $4'''.136$ and $6'''.036$.” Encke (349-50) made some interesting additional points: the experiments agree with theory for the interval contains $4'''.6$, the value given by theory; the existence of an easterly deviation borders closely on certainty for the mean is more than five times the probable error; to obtain narrower 50% limits more observations would be needed—to obtain a probable error of $'''.01$ about 2600 observations would be needed. Encke went on to reproduce the sampling theory results from Gauss (1816).

The main journal for German error theory was the *Astronomische Nachrichten*, founded in 1821 by Gauss’s student, H. C. Schumacher (1780–1850), and edited from 1854 by Bessel’s student, C. A. F. Peters (1806-1880). Dedekind never belonged to this community but Jacob Lüroth (1844-1910) had been an apprentice astronomer—he contributed an article in 1862 before turning to pure mathematics. In 1869 and -76 Lüroth contributed papers on probable errors. The first was one of a series taking a sampling theory approach to the estimation of precision; see David (1999) for a review. Peters (1856) kicked off by considering the expected value of Bessel’s mean absolute deviation measure; his modification of the Bessel formula for the probable error of the mean, viz.

$$0.845 \frac{\sum |x_i - \bar{x}|}{\sqrt{m(m-1)}}$$

became known as Peters’ formula. The series, which also included contributions from W. Jordan and von Andrae, finished with a paper (1876) by the geodesist F. R. Helmert (1843-1917) which contained, inter alia, the exact distribution of Gauss’s 1822 estimator for the

case of normal errors; see Sheynin (1995: 93-100) for a detailed account.

Lüroth's first paper (1869) extended Peters' formula to the case of indirect observations and it secured a place in the literature; see Czuber (1891). Lüroth (1876) took a completely different line and compared Encke's formula for the probable error with one based on Dedekind's distribution; it did not fit into the series and dropped from sight. The paper gives nothing away about its origins or motivation: Lüroth does not refer to his earlier paper and, though he refers to the Dedekind density at the appropriate point in the derivation (1876: 211n), there is nothing to indicate when, or how, he discovered Dedekind's work. I have found no links between the men though two years later they were discussing Cantor's work on dimension but not with each other—see Dauben (1979).

Pfanzagl and Sheynin (1996) give a through review of the 1876 paper and so I can be brief. Lüroth compares two formulae for the probable error of a regression coefficient, the usual one based on the most probable estimate of the precision and another based on averaging across values of the precision and of the other coefficients. Lüroth presents the formulae without advocating either or quoting any authorities. The usual (“gewöhnlich”) formula is that given by Encke while the second was new and based on the density obtained by Dedekind. Although the kernel of the “precursor of the t -distribution” was already in Dedekind, Lüroth put formula (3.4) into usable form. Lüroth also made a conceptual advance by introducing exact inference for intervals; its only forerunner, the interval estimate of precision of Gauss (1816), used a large sample normal approximation to the posterior.

Having developed the new interval formula Lüroth devoted the rest of the paper to

comparing the width of the two 50% intervals. To this end he (1876: 220) obtained the inequalities

$$r_1 > R_1 > r_1 \sqrt{\frac{p}{p+1}}.$$

where $p = m - n$ and r_1 and R_1 are the probable error of the first regression coefficient based on the most probable value of the precision, the usual method, and that based on averaging, the new method. The conclusion (p. 220) is:

so dass hier durchgeführte Methode der Berechnung das wahrscheinlichen Fehlers ihm stets kleiner liefert als die gewöhnliche, aber höchstens so, als ob eine Beobachtung mehr vorhanden wäre und die letztere Methode der Berechnung angewandt würde.

The force of these inequalities seems to be do that it is not worth the effort to compute R_1 (in astronomical practice values of p usually exceeded 20) and Lüroth provided no tables for computing it. Unlike Student (1908a) the interval was narrower than the usual one; see Essay II.

Lüroth's choice of place and time was perfect—the *Astronomische Nachrichten* was the main forum for the theory of errors and the probable errors was a live subject—but the paper provoked no discussion and the only contemporary comment I have found is Merriman's (1877a: 227) unilluminating, "The usual formula is compared with a new formula and shown to give larger values." Dedekind and Lüroth figure in Czuber's history of probability theory (1899), the first for his 1855 article and the second for his 1869 article; the 1876 article appears

in the bibliography but is not mentioned in the text. The *Jahrbuch über die Fortschritte der Mathematik* (the *Mathematical Reviews* of the day) did not notice the 1876 article.

What is to be made of the line from Gauss to Lüroth? Gauss presented partial analyses of scheme (2.2) in 1809 and -16; half a century later Dedekind produced a complete analysis and presented answers to the two leading questions—how to estimate β and how to estimate h ? Dedekind followed Gauss and did not consider interval statements about the mean. Lüroth extended the analysis to treat interval inference for the mean and compared the resulting interval with the one in general use: he advocated neither and showed that there was little practical difference between them.

5 Contributions without consolidation

The Dedekind-Lüroth results on the regression posterior marginals were obtained—with variations—several times in the period between Gauss and Jeffreys: Laplace (1820), Pizzetti (1889) and Bennett (1908) obtained the marginal for precision, Edgeworth (1883) and Burnside (1923) the marginal for the mean and Lhoste (1923) the marginals for both. Why were there so many initiatives leaving so little trace?

The 1820 argument is described in Hald's (1998) survey of the writings of Pierre-Simon Laplace (1749-1827). The later German literature paid little attention to Laplace but in the early days he and Gauss fed off each other: Gauss (1809) drew on Laplace, Laplace was an influence on the sampling theory of Gauss (1816) and the analysis in the "Mémoire sur le flux et le reflux de la mer" (1820) reads like a response to Gauss and possibly to

Bessel (1815). Laplace (1820: 486-) considers both mean and precision, k ($= h^2$) and obtains the joint posterior. Laplace integrates out the mean—so far so like Dedekind—however, protesting against “several geometers” who chose the most probable value, Laplace (487) declared for the posterior mean and derived it. Laplace did not integrate out the precision to obtain the marginal posterior for the mean but plugged the precision estimate into the normal distribution. Hald (1998: 422-3) comments that Laplace “fell for the easy normal approximation.” Laplace does *not* mention using an approximation and he may have thought he was improving on Bessel by providing a better insert, as Encke would do in 1832.

Laplace and Gauss produced new arguments without repudiating their old ones. Arguments in the theory of errors so proliferated that Merriman (1877: 159-60) identified thirteen different proofs of the principle of least squares; Knochloch (1992) suggests reasons—philosophical, mathematical and practical—why so many approaches coexisted. The Bayesian approach was transmitted by a variety of means. Some authors, like Encke, gave a Bayesian treatment of one topic and a sampling theory treatment of another, while others presented alternative treatments of the same topic; thus Bertrand (1889) discusses a number of ways of estimating the precision of errors, among them (196) the mean of the posterior of k following Laplace (1820) but for the known mean case of Gauss (1816). The textbooks went on presenting examples of Bayesian analysis with enough gaps in the presentation to encourage the reader to fill them—and they did.

In 1883 F. Y. Edgeworth (1845-1926) was just starting in probability and statistics and his “Method of least squares” belonged to his first reconnaissance of the field; for Edgeworth

see Stigler (1978, -86). Edgeworth knew the English literature on least squares, some of the writings of Laplace and Gauss, but not the latest German work. He published in the *Philosophical Magazine*, an applied mathematics journal, although “philosophical” suits the cast of Edgeworth’s interest—see Baccini (2009). Edgeworth had a unique background in mathematical ethics and economics and wrote like a learned commentator with an acute sense of system: the “primary object” of the least squares paper (1883: 361) is to “illustrate this application of mathematics to psychical quantity” with a “secondary purpose” to “classify the problems falling under our title ... and to offer some contributions to their solution.” Like Lüroth, Edgeworth published in a very visible place but wrote in a way to minimise his impact.

The problem of inference to the mean and modulus $c (= 1/h)$ from direct observations falls in sub-class I. A (3) of the taxonomy of Edgeworth (1883). Edgeworth was not explicit about using a uniform prior but later in the paper he (374) remarks that “constants do in general *in rerum naturá* as often present one value as another.” Like Dedekind, Edgeworth was most exercised by the estimation of precision but with the “greatest diffidence” he (367) insisted against the “most distinguished” authorities that the maximum probability estimate involves the sample size and *not* the sample size less one. Edgeworth did not consider the marginal density for the modulus but he obtained the marginal density for the mean with a view to obtaining its probable error. Having obtained the t -like curve, he (368) summarised:

it appears that what may be called the measure of the probable error is the sum of the squares of the apparent errors divided by n , not, as some might seem to

imply by $(n-1)$; although this *modulus*, as I think it may be called with propriety is not to be multiplied by the usual factor .476, but by the length of the abscissa which halves the half-area of the curve just indicated.

So, unlike Lüroth, Edgeworth insisted on a correction but, the correction made, he moved on to other sub-classes and, in later papers, to other topics. Apparently his contribution was first noticed by Welch (1958) on the jubilee of Student (1908); it has since been discussed by Pfanzagl and Sheynin (1996: 892-3) and Hald (2007: 69-71).

Seal (1967: 11) and Farebrother (1998: 118) have noted the contribution of the Italian geodesist Paolo Pizzetti (1860-1918). Bertrand's textbook *Probabilités* was Pizzetti's (1889) starting-point although he knew the work of Gauss and Laplace. His goal was the same as Dedekind's but Pizzetti stopped when he had found the maximum of the posterior of the precision in the regression case. Pizzetti did not carry the analysis into his treatise, *Fondamenti* (1892), which presented a sampling theory for precision, a theory Hald (2000: 213) interprets as "bridging the gap between Helmert and Fisher."

In early 20th century England Gosset and Fisher found new applications for the theory of errors, developed new theory and merged the subject into mathematical statistics which, in Fisher's hands, was aggressively anti-Bayesian; this radical mutation is described in Essay II below. The writings of Bennett (1908) and Burnside (1923) were entirely traditional and are known today chiefly because Fisher wrote about them. Bennett's treatment of precision was taught in Cambridge and presented in Brunt's (1917: 32) textbook. Like Dedekind, Bennett found the most probable value of Gauss's h (for the simpler case of direct observations); from

this and the precision formula, the value of the probable error of the mean “immediately follows” writes Brunt. Burnside’s (1923) paper on the theory of errors resembled Lüroth (1876) but was not so technically accomplished; see Aldrich (2009). Burnside objected to the assumption underlying the “standard formula” and he (1923: 486-7) replaced it by the assumption that a priori “all values of the precision-constant are equally likely.” Like Lüroth, Burnside compared the width of the 50% intervals but he computed the exact values for sample sizes up to 10. Burnside judged these intervals and the normal one “materially” different yet he (1923: 487) left it as “a matter of individual judgement” which assumption was the more reasonable.

Harold Jeffreys (1892-1989) resembled Laplace in being a physical scientist and mathematician and Edgeworth in attending to the writings of philosophers. However Jeffreys was not a learned and even-handed commentator: his aim was to re-found the theory of errors on correct lines, extending the argument of Gauss (1809) as presented by Whittaker and Robinson (1924); see Aldrich (2005b). Having recast the theory for the mean in 1931 and for regression in 1932, Jeffreys encountered Fisher’s work—see Essay II below—and embarked on the project of re-establishing mathematical statistics on Bayesian foundations.

Part of Jeffreys’s probability project was prefigured a decade earlier in the writings of the French army officer Ernest Lhoste (1880-1948) as Broemeling and Broemeling (2003) have recently shown. Today the *Revue d’artillerie* seems a remote and unlikely publication but it was closer to the error theory mainstream than *Biometrika*: up to 1914 the *Revue* was covered by the *Jahrbuch*. In the new Fisher and Jeffreys mainstreams only one modern non-

English language work achieved a place—Behrens (1929)—and that barely had an independent existence for it was usually subsumed into the Behrens-Fisher composite. The 70 year gap between Lhoste’s papers of 1923 and their sighting in the English statistical literature by Villegas (1990) somehow symbolises the transformations of the 20th century.

Broad, long lines came with Jeffreys as they had in sampling theory with Fisher. What had changed? Unlike most of their predecessors, Jeffreys and Fisher were system builders and intolerant of the wrong system; they were not providing “another angle” (Gauss) or leaving the choice as a “matter of individual judgement” (Burnside) but getting things right. They wrote system books consolidating their discoveries, *Statistical Methods for Research Workers* and *Theory of Probability*.

Essay II The theory of errors as mathematical statistics—Fisher and Student 1908-25

1 Introduction

The theory of errors was reborn in R. A. Fisher's *Statistical Methods for Research Workers* of 1925. This was one of a series of biological monographs and manuals but it contained so many advances that much more was involved than a relocation from astronomy and geodesy. In developing and departing from traditional error theory Fisher had a forerunner and an associate—W. S. Gosset, who published as “Student.” Fisher (1939: 1) traced to Student's paper of 1908 “a fundamentally new approach to the classical problem of the theory of errors the consequences of which are still only gradually coming to be appreciated in the many fields of work to which it is applicable.” In the early 20s the two worked together to produce “Student's t .”

Merriman (1877a) recorded some significant English contributions to the literature on least squares but they were barely a memory when Gosset and Fisher began. Though there was little research activity in the theory of errors, there was plenty in the related fields of statistics and biometry. The statistician F. Y. Edgeworth had cut his teeth on the subject in the 80s—see Essay I—and the biometrician Karl Pearson had learnt the theory of errors as an undergraduate in the late 1870s but both had moved on. Gosset and Fisher would emerge in the orbit of Pearson and be viewed as Pearsonians who specialised in small-sample

theory. The fundamental paper of 1908 appeared in *Biometrika* away from the mainstream of error theory and, though Fisher placed some papers in more conventional journals, he was not much interested in the traditional applications and from the 20s he saw himself as a mathematical statistician.

What follows is a familiar story told from a different angle: the extensive secondary literature on the Student-Fisher development—surveyed in Aldrich (2003/10)—ranges from the classic biographies by E. S. Pearson (1939, -90) and Box (1978) to works marking the centenary of Student (1908), such as Zabell (2008) and Hanley, Julien and Moodie (2008). There was a third party in the story, Karl Pearson; for a guide to the Pearson literature see Aldrich (2001/10). Most of the technical developments are treated in Hald (1998).

2 Textbook error theory

In 1899 William Sealy Gosset (1876-1937) joined Guinness, the Dublin brewer. Gosset had a chemistry degree from Oxford but he had spent his first two years studying mathematics, gaining a first in Mathematical Moderations. The subjects for “Mods” were algebra, trigonometry, plane geometry, elementary differential and integral calculus and mechanics. The theory of chances and the theory of errors came at the next level in “Greats” and Gosset later taught himself these subjects from textbooks.

In 1904 Gosset produced a report for the Guinness Board on “The application of the ‘law of error’ to the work of the brewery.” This has references to Merriman’s *Text-Book on the Method of Least Squares* (1882), Airy’s *Theory of Errors of Observations* (1879) and

Lupton's *Notes on Observations* (1898); see E. S. Pearson (1990: 10-14). These were very different works—Merriman a lucid all-round introduction, Airy a digest of theoretical results and Lupton a rapid tour taking the physical scientist from philosophy of science to least squares calculations by way of references to Pearson, Edgeworth and Sheppard and sceptical comments (1898: 16, 30 & 81) on the practical use of the rule of succession and of Bayes' theorem. Neither Merriman nor Airy mentioned inverse probability and Lupton did not link Bayes's theorem to least squares; the error distribution of Gauss (1809) survived but the inference principles—see Essay I—had melted away.

The first version of the textbook by Mansfield Merriman (1848-1925) appeared in 1877 at the same time as the bibliography with a second version, which went through 13 editions, appearing in 1882; for a brief biography see Stigler (1985). The original version did not reflect the latest developments recorded in the bibliography, such as Helmert (1876) or Lüroth (1876), and later versions/editions did not try to keep up with the literature. Hald (2007: 105-9) places Merriman in a textbook tradition from Encke and Chauvenet (1863) in which the normal specification was retained but the Bayesian element attenuated, producing what Aldrich (1997: 163) calls “bowdlerised Gauss” and Hald (2007: 108) “likelihood.” Airy's book does not have even the ghost of Bayesian reasoning: a useful test is his treatment of the probable error, the central but slippery concept introduced by Bessel; see Essay I. Airy (1861: 21-23) introduces the probable error as a parameter of the error curve: 50% of the errors lie inside the interval, mean $\pm p.e.$; the concept is extended (41) to the sampling distribution of the mean and it is shown that the *p.e.* of the mean is \sqrt{n} times the *p.e.* of an observation.

Airy's (47) estimate of the probable error of the mean is $0.6745 \times \sqrt{\frac{\text{sum of squares of apparent errors}}{n(n-1)}}$

where the “apparent errors” are the deviations from the sample mean. In modern terms, this is a point estimate of a parameter of the sampling distribution of the mean.

Airy does not actually consider the main use for the estimated *p.e.* i.e. in constructing 50% intervals for the mean and for regression coefficients; Airy gives no examples involving data. Merriman (1882: 89-91) constructs an interval for an angle based on 24 observations with a mean of $116^\circ 43' 49''.64$ and probable error $0''.28$. Merriman interprets these values as follows:

The precision of the mean of these twenty-four observations is such that $0''.28$ is to be regarded as the error to which it is liable; that is, it is an even wager that the mean differs from the true value of the angle by less than $0''.28$.

The wager does not take into account that the probable error is estimated—of which more later—but the wager on the error leaves it unclear whether the probability is associated with the sample mean or the true value. The former seems the more likely but Merriman (170) has one example where it seems clear that the wager is on the true value. The example is unusual for at issue is the value of a “constant error” in a measuring instrument. The true value of the angle is known to be 90° but observations on an angle give $89^\circ 59' 57'' \pm 0''.8$ making it “extremely probable that a constant error of about $-3''$ exists in the instrument.” After some calculation Merriman concludes that “it is a wager of ... almost 10 to 1, that the mean is between the limits $89^\circ 59' 55''$ and $89^\circ 59' 59''$ ” and further “it may be shown that it is a wager of 39 to 1 that there is a constant error between $0''$ and $-6''$.” These look like

Bayesian wagers on a parameter value.

To return to Gosset's report, E. S. Pearson (1939: 215) comments, "All this is simply Airy or Merriman put by Gosset into the form most useful for his fellow brewers." "Imaginatively" might be more apt than "simply" for, though Gosset presents nothing absolutely new, it was more than a rehash of his reading. An instance is what Gosset (1904: 7-8) wrote about the sample size required to ensure a specified level of accuracy for the mean:

For this purpose we must first decide—

1. Within what limits of accuracy we desire to know the results.
2. What certainty we require that it will fall within those limits.

To illustrate the reasoning Gosset suggests, "it might be maintained that Laboratory produce should be within .5 of the true result with a probability of 10 to 1." Assuming that the "modulus" (i.e. Gauss's measure of precision $1/\sqrt{2}\sigma$) is known to be .8, Gosset gives the odds for a number of different sample sizes, showing that less than 4 observations would be insufficient but that with 4 the odds in favour of a smaller error than .5 are 12 : 1. The subject had been considered by Encke (see Essay I) but it was not in Gosset's reading.

Gosset (1904: 11) also discusses limits for the modulus of error using brewery data—samples of 82 and 159—to illustrate the formulae from Gauss (1816); see Essay I. At this stage there is no suspicion that using an estimate of the modulus would have implications for inferences on the mean or that the Gauss formulae were inappropriate for small samples.

3 Error theory and biometry I

The Guinness Board was impressed with the possibilities Gosset revealed and he was sent to consult an authority on outstanding issues. The person chosen, Karl Pearson (1857-1936), had mixed credentials: he had been a conventional applied mathematician (i.e. applying mathematics to physics) and had studied the theory of errors as an undergraduate—see Aldrich (2007)—but he had taken up biology and come to the opinion that error theory was finished and that the future was with skew curves and correlation. The Student-Fisher revival of the theory of errors is a surprising story and the first surprise is that a critic of the theory should have contributed to it.

For his meeting with Pearson—in May 1905—Gosset prepared four questions; see E. S. Pearson (1939: 215-6). One concerned the formula for the probable error of the probable error and here Gosset’s doubts about the effect of sample size on inferences surfaced:

E.g. if n were infinite, I could say “it is 10 : 1 that the truth lies within 2.6 of the result of the analysis. As however n is finite and in some cases not very large, it is clear that I must enlarge my limits, but I do not know by how much.

The point that the accepted formulae depends on infinite n was not in the textbooks.

In reply to another query Pearson introduced Gosset to correlation analysis, a subject related to bivariate error theory but which had developed independently. Gosset soon produced a report for the Board on “The Pearson co-efficient of correlation.” This was another expository piece but it (1905: 28) contained an acute observation on the applicability of

Pearson and Filon's (1898: 242) probable error formula:

When derived from large number of cases, if r is $2\frac{1}{2}$ to 3 times its probable error, the connection may be said to be very probable, the odds that it exists being of the order 20 : 1 to 50 : 1. But with only a few cases *I expect[†] a larger ratio is required.*

†For consider as an extreme case a correlation table with only four instances. If r happened to come high, say .9 the P.E. would be .064, $\frac{1}{14}$ of r . Yet no one would claim any certainty from four experiments unless the process had been already proved free from error.

Some years later Gosset explained to Fisher his interest in such small samples; McMullen (1970, letter 1) and E. S. Pearson (1968: 447; 1990: 25)):

[The work] concerns such things as the connection between analysis of malt or hops, and the behaviour of the beer, and which takes a day to each unit of the experiment, thus limiting the numbers, demanded an answer to such questions as "If with a small number of cases I get a value r what is the probability that there is really a positive correlation of greater value than (say) .25?"

It was settled that Gosset would spend 1906-7 with Pearson in London where he found—and embraced—new ideas and methods, including tests of significance, inverse probability, curve-fitting by the method of moments and χ^2 goodness of fit tests. These processes did not combine readily with those Gosset brought from the theory of errors, or even together

and Student's papers have a copy and paste quality that contrasts with the tight Bayesian productions described in Essay I and with Fisher's sampling theory productions of the 1920s; Pearson's ideas on inference are discussed by Aldrich (2007).

4 Probable errors of mean and correlation: 1908

Two papers on small sample inference came from Gosset's time with Pearson, "The probable error of a mean" (1908a) and "Probable error of a correlation coefficient" (1908b)—their titles echoing Pearson and Filon's (1898) "Probable errors of frequency constants." The two Student papers were conceived together to be executed in parallel. They would use Pearson's methods to find the exact sampling distribution of a statistic, something Pearson had never sought. Both papers are unsatisfactory, though in different ways: the correlation paper maps out a grand project but cannot realise it; the mean paper aims lower and achieves more, yet how, or whether, it fits into the grand scheme of the companion paper is unclear.

In his correlation report Gosset had pointed to a small-sample problem and now Student (1908a: 1-2) indicated one for the mean:

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to s/\sqrt{n} , where s is the standard deviation of the sample, and to use the tables of the probability integral.

But, as we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to an increasing error, until judgments reached in this way may become altogether misleading.

The “usual method” (essentially Encke’s) evidently rests on the presumption that the posterior mean of the population is $N(\bar{x}, \frac{s^2}{n})$. After Airy and Merriman, Student’s writing appears distinctly Bayesian and his objection to the method is that it cannot be accurate when the number of experiments is small, not that it is Bayesian.

The correlation paper (1908b: 302) also states its object in Bayesian terms, “We require the probability that R for the population from which the sample is drawn shall lie between any given limits.” For Student this was a simpler problem than for modern Bayesians who—following Jeffreys (1939)—embed it in a five parameter bivariate normal distribution and simpler too than the problem for the mean when the precision is unknown. Student started from the large sample theory of Pearson and Filon in which only the correlation and the sample size figure and he appears to have assumed that correlation is a one-parameter problem. Pearson’s lectures—see also Pearson (1907)—gave Gosset a model for one-parameter Bayesian analysis: when inferring to the probability of a success in Bernoulli trials combine the prior with the sampling distribution of the number of successes. Adapting this scheme to correlation required (1) the sampling distribution of the sample correlation, r , and (2) a prior distribution for the population correlation R . Student (1908b: 303) had some suggestions about (2) but postponed the issue to concentrate on (1). He had limited success for, while he (1908b: 304-8) could conjecture—correctly—the sampling distribution for the case of $R = 0$,

he got nowhere with the non-null case. However, with the null distribution Student (1908b: 302 and -8) could answer one question, whether it is “safe” to take .50 as the “limit of significance” for a correlation coefficient when the sample size is 21: “we may expect to find one case in 50 occurring outside the limits $\pm .50$ when there is no correlation and the sample numbers 21.” Student in 1908, unlike Jeffreys in 1939, saw no incompatibility between the Bayesian approach and significance testing; indeed both 1908 papers report P values for χ^2 goodness of fit tests.

While the correlation paper did new things, the mean paper re-did old things and none too expeditiously. In his jubilee piece Welch (1958) noted that Edgeworth had obtained the form of the posterior of μ in 1883 and earlier work by Dedekind and Lüroth has been uncovered; see Essay I for details. Student did not reproduce this analysis but produced a new sampling theory, part of which reproduced the work of Helmert (1876); Pearson (1931) pointed this out. In the correlation paper Student treated the sampling distribution of r as part of the solution to the Bayesian problem and his intuition may have told him that the sampling distribution of z (see below) would contribute to the posterior for μ . There is a remark (1908a: 23) suggesting that the sampling distribution *is* the posterior when there is no prior information, “These odds are those which would be laid, and laid rightly, by a man whose only knowledge of the matter was contained in the two experiments.” Student had only seen examples of Bayesian reasoning involving one parameter and would not have known how to treat a prior for two parameters.

The sampling theory occupying the space between the introduction and the illustrations

is for the distribution of a quantity z “obtained by dividing the distance between the mean of a sample and the mean of the population by the standard deviation of the sample”, i.e. $z = x/s$, where $x = \bar{x} - \mu$ and $s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$. The terminology of “sample,” “standard deviation,” etc. is Pearson’s; the language is discussed in Aldrich (2003). Student did not explain, or even mention, that his z was *not* the basis for the “usual method” which involved dividing x by s/\sqrt{n} , with s calculated with divisor $(n - 1)$ and not n . With reference to the divisor, Fisher (1939: 2) later remarked that Student had followed the “foolish convention” of the biometric school. In 1908 the differences probably seemed immaterial and it was only in the 1920s and at Fisher’s urging that Student reverted to the usual form, known since those days as the t form; see Section 9 below.

Gosset was an astute critic and a brilliant applied statistician but not one of those “very brilliant mathematicians who have studied the Theory of Errors” as Fisher (1939: 6) called them. However Gosset began with an ingenious plan for obtaining the distributions: Pearson (1895) had shown how to fit curves to data using the method of moments and Student proposed using this method to fit curves to artificial samples of z - and r -distributed variates. Section VI of the mean paper (1908a: 12-19) describes the procedure for constructing two sets of 750 samples of size 4 and gives some empirical results. But this was only the ghost of the plan for in the event Gosset found the densities of z and s analytically and so the section, “Practical test of the foregoing equations” is headed. The correlation paper followed the plan and combined curve-fitting to the Monte Carlo results and theoretical analysis of some very special cases.

Regarding the z ratio, it was known—Student (1908a: 7) cites Airy—that the numerator is normal with mean 0 and standard deviation σ/\sqrt{n} . Student did not know that the distribution of s^2 had been found by Helmert (1876) and applied by Pizzetti (1891) and he adapted Pearson’s empirical curve-fitting procedure to solve this theoretical problem. Student (1908a: 3-4) found the first four moments of s^2 analytically and, as these matched those of a Pearson Type III curve with density

$$y = Cx^{\frac{n-3}{2}}e^{-\frac{nx}{2\sigma^2}},$$

he concluded that he had found the distribution of s^2 ; from this he (7-8) derived the density of s analytically. By computing more moments Student (1908a: 6-7) showed that the numerator and denominator of the z ratio are uncorrelated and, then using the formula for a ratio of independent random variables, he took a final analytical step (8) to obtain the density of z ,

$$y = \frac{C}{(1+z^2)^{\frac{n}{2}}}.$$

This was also a Pearson curve—a Type VII—and perhaps Student would have found it had he followed his original curve-fitting plan; Hanley, Julien and Moodie (2008) describe his sampling experiments.

The entire derivation rests on a very creative use of Pearson’s ideas but there was a more specific debt, or so Gosset said later; E. S. Pearson (1990: 17) quotes a letter from him:

I gained a lot from [Karl Pearson’s] ‘rounds’: I remember in particular his supplying the missing link in the error of the mean paper—a paper for which he disclaimed any responsibility.

Gosset did not identify the “missing link.”

Student produced tables giving the probability that z lies between $-\infty$ and values from .1 to 3.0 for sample size, n , from 4 to 10. The “illustrations of method” show how to use the tables. and they entail evaluating the probability that an effect is positive. In the first on the soporific effects of a treatment applied to 10 cases there is a sample mean of $+.75$ and a standard deviation of 1.70 (1908a: 20-1):

let us see what is the probability that [treatment] 1 will on the average give increase of sleep; i.e. what is the chance that the mean of the population of which these experiments are a sample is positive. $\frac{+.75}{1.70} = .44$ and looking out $z = .44$ in the table for ten experiments we find by interpolating between .8697 and .9161 that .44 corresponds to .8873, or the odds are .887 to .113 that the mean is positive.

The probability evaluation resembles that in his correlation report—“the odds that it exists ...”—but Student appears not to have used it again. In Fisher’s *Statistical Methods* (1925a: 107-8) the example is reworked as a significance test and that was a possibility both conceptually and practically in 1908 and it is not clear why Student did not opt for it.

In going from a z -probability to a μ -probability, Student presumably used manipulations like the following:

- i) Interpolating in the table yields $P(z < .44) = .8873$.
- ii) Putting $z = \frac{\bar{x} - \mu}{s}$ and substituting gives $P(\mu > \bar{x} - .44s) = .8873$.
- iii) In the present case $\bar{x} = .44s$, and so $P(\mu > 0) = .8873$.

Student did not mark any conceptual gap between the sampling distribution of $Z = (\bar{X} - \mu)/S$ (as it might be written today) and the probability distribution of μ . Merriman had made the same transition in his treatment of the constant error and it may be that both were confused. That is possible but I am inclined to think that Student was gesturing towards an unformulated Bayesian argument. The transition also looks forward to the fiducial argument, a form of reasoning introduced in Fisher (1930), and Fisher (1939: 4) and Jeffreys (1939: 310) found fiducial elements in Student’s paper.

Whatever Student thought he was doing, his sampling theory for z was new and historians have not yet found ‘forerunners.’ Given that there had been related sampling theory since Gauss and Laplace, waiting for Gosset was a very long wait. Pfanzagl & Sheynin (1996: 892) identified one obstacle:

it was not an obvious idea to consider the errors of \bar{x}_m and $\hat{\sigma}_m$ simultaneously, so as to obtain a confidence interval with exact coverage probability $\frac{1}{2}$ from the distribution of $(\bar{x}_m - \mu)/\hat{\sigma}_m$.

One great difference between Student (1908a) and Lüroth (1876) and the main tradition—including Dedekind (1860b), Helmert (1876), Pizzetti (1891) and even Gosset in 1905—was that dispersion was no longer a primary focus. In Student (1908a) the distribution of s^2 is only an intermediate result and there is no pause to consider how the distribution might be used to correct the usual large sample formula. Pearson (1915) reverted to type, judging the distribution of s to be the main result of Student (1908a).

Fisher’s (1939: 5) comment that Student’s z was received with “weighty apathy” is just.

The table was reprinted in Pearson's *Tables for Statisticians and Biometricians* (1914) but this was no accolade for all tables published in *Biometrika* were reprinted in that volume. The fortunes of z changed when Fisher began promoting it in the 20s; he was, however, first involved long before.

5 Error theory and biometry II

Gosset was in contact with a practitioner of traditional error theory, the Cambridge astronomer F. J. M. Stratton. The link was through agriculture: Guinness had an interest in the cultivation of its inputs and Gosset worked with agricultural scientists—see the appendix by Student (1911) to Mercer and Hall's "The experimental error of field trials"—while Stratton had branched out to apply the theory of errors in collaboration with the agriculturalist T. B. Wood—see Wood and Stratton (1910). Stratton was probably unaware of Student's z . Stratton's lectures to the mathematics undergraduates on the combination of observations were a main source for David Brunt's (1917) textbook and, while this has some biometric excursions—Pearson curves and correlation—it does not mention z . Stratton, however, made a connection of another kind—he brought together Gosset and Ronald Fisher a third-year undergraduate; see Box (1987) and E. S. Pearson (1990).

In 1912 Ronald Aylmer Fisher (1890-1962) had already published an article on estimation in the Cambridge journal *Messenger of Mathematics* and written an insightful essay on biometry and Mendelism (unpublished until 1967). The article "On an absolute criterion for fitting frequency curves" proposed the prior-less Bayes method from textbook error theory

(see Section 2 above) as a *general* method and explained why its invariance (‘absoluteness’) made it superior to methods like Pearson’s method of moments; the paper is described in Aldrich (1997: 162-6). Fisher produced only one worked example, the mean and dispersion (Gauss’s h) of the ”normal curve of frequency of errors” where the absolute criterion (= maximum likelihood) gives the n form for the estimate of dispersion; his references are to Chauvenet and to Bennett (1908)—for the latter see Essay I. Fisher (1912: 160) disallowed Bennett’s argument for $n - 1$, based on maximising the marginal posterior for precision, because it is not possible “to obtain an expression for the probability that the true values of the elements should lie within any given range”—a declaration that appears to be disallow all Bayesian arguments.

Error theory and Pearsonian biometry contributed to the intellectual formation of Fisher and Gosset in different ways: one had lectures on the theory of errors and read up Pearson, the other did the opposite; one used methods from the theory of errors to solve Pearson’s problems—or at least to further the biometric project—and the other used Pearson’s techniques to solve a problem in error theory. Fisher and Gosset might easily have had nothing to say to each other for one was interested in point estimation and one in interval estimation and they had a fundamental disagreement over the Bayesian argument: Student had gone from Merriman and Chauvenet to a more Bayesian position, Fisher had gone in the opposite direction. However, there was something in Fisher’s work that gave him an interest in Student’s results, or at least in one of them. Fisher (1912: 160) entertained a variant of the absolute criterion according to which the density of a statistic is maximised, not the

density of the observations. He wanted to apply this criterion to the estimation of h using the density of s^2 and it appears that he obtained the density for just this purpose; this was another unnecessary task for Brunt (1917: 59) refers to Helmert (1876) though without indicating that Helmert had derived the density of s^2 . The z -density had no estimation interest but, as Fisher was by now aware of Gosset's work, its derivation may have seemed a nice extension.

6 Fisher's proofs

In September 1912 Gosset forwarded Fisher's derivations of the distribution of s^2 and z to Pearson saying, "It seemed to me that if it's all right perhaps you might like to put the proof in a note. It's so nice and mathematical that it might appeal to some people." (Quoted in E. S. Pearson (1968: 446).) Pearson was unobliging and the full proof did not appear until 1923—see section 12 below—although parts appeared in Fisher's 1915 paper on correlation and in a 1920 paper on estimating σ^2 . Like Student, Fisher devised a single method of attack for both mean and correlation and neither method came from the error theory textbooks. Fisher would use his method on a string of problems but Student's method was never used again.

To judge from Fisher (1915, -20, 23), the "so nice and mathematical" argument of 1912 began with the chance of the sample falling in a region of volume dx_1, dx_2, \dots, dx_n , as given

by

$$df = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\sum(x-m)^2}{2\sigma^2}} dx_1 dx_2 \dots dx_n.$$

Fisher uses a geometric argument to show that the frequency with which \bar{x} and s (still as defined by Student with $\frac{1}{n}$ as divisor) fall into the ranges $d\bar{x}$ and ds is proportional to

$$e^{-\frac{n}{2\sigma^2}(\bar{x}-m)^2} d\bar{x} \cdot s^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds.$$

From this expression it is evident that \bar{x} and s are independent (something Student assumed from absence of correlation) and the density for s pops out. By a change of variable to s and $z = (\bar{x} - m)/s$, Fisher obtains a joint density proportional to

$$\frac{s^{n-1}}{\sigma^n} e^{-\frac{ns^2}{2\sigma^2}(1+z^2)} ds dz.$$

Integrating with respect to s from 0 to ∞ yields the density of z as found by Student, i.e.

$$\frac{\frac{n-2}{2}!}{\frac{n-3}{2}! \sqrt{\pi}} \cdot \frac{dz}{(1+z^2)^{\frac{n}{2}}}.$$

Fisher's appeal to geometry became the trade-mark of his distribution theory.

In publishing the very tentative correlation paper Student (1908b: 302) hoped to “interest mathematicians who have both time and ability to solve [the problem].” After some work by Soper (1913), Fisher (1915) produced a very elaborate geometrico-analytic derivation and this time Pearson published the analysis. Student explained how he intended to use the distribution but Fisher's purposes are not so clear for he (1915: 520-1) only shows how the newly-derived distribution can be used for estimating the population correlation, ρ , via the second version of the absolute criterion; for this see Aldrich (1997: 166-7).

Gosset told Fisher how he saw the matter (September 1915: Letter 1 of McMullen (1970) and E. S. Pearson (1968: 447; 1990: 25)):

I am very glad that my problem is a step nearer solution ... but there still remains the determination of the probability curve giving the probability of the real value (for infinity population) when a sample of x has given r . Of course this would have to be worked out for two or three á priori probabilities and if otherwise convenient I would try $y = y_o(1 - x)^{\frac{m-4}{2}}$ (giving m the values 3, 4 and 6 in succession) as the á priori distribution of the probability of x being the real value of r .

But of course anything almost would do if it gave an integrable expression.

As E. S. Pearson (1990: 25) points out, the term $(1 - x)$ was surely a slip for $(1 - x^2)$. Unfortunately only one side of the correspondence survives and we are left to speculate how, if at all, Fisher replied.

There was an appearance of z in the correlation paper but Student did not comment on it. When considering re-expressing the frequency curve for r using transformations of r and ρ , Fisher (1915: 518) remarks of one of these transformed cases: “It is interesting that in the important case, $\rho = 0$, the frequency reduces to $\frac{dt}{(1+t^2)^{\frac{n-1}{2}}}$ and the curves are identical with those found by “Student” for z ...” (In the original $r = 0$ is written.) Fisher did not develop the idea of using Student’s z for testing for absence of correlation but it foreshadowed the further uses of z that Fisher would find in the 1920s.

7 Extending the theory of errors

In 1919 Fisher took a job at Rothamsted Experimental Station where he would associate with biological researchers and devise new methods for agricultural science; his new interests brought him closer to Gosset. He continued to pursue his old interests and, before considering the new work, we should note a paper he addressed to astronomers. Fisher and Gosset began with a legacy of intellectual capital from the traditional theory of errors and they did not go back for more. Gosset paid no further attention to the subject but Fisher descended upon it twice; the second occasion is described in Section 12 below. His “Mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error” (1920) responded to a claim from Arthur Eddington, error theorist and astronomer, that the mean error estimate was superior to the mean square estimate. The estimation of accuracy had always attracted attention—see Essay I—and in refuting the claim Fisher unknowingly re-derived results of Gauss (1816) and Helmert (1876). Fisher’s paper introduced an important new concept, sufficiency, which became a cornerstone of his theory of estimation.

In the years 1921-5 Fisher created a system of mathematical statistics and gave it practical form in *Statistical Methods for Research Workers*. The Pearson revolution of the 1890s had dethroned the normal distribution and by restoring the theory of errors Fisher was re-enthroning it. The restored theory was an enhanced theory for it incorporated the biometricians’ regression and included a new technique, the analysis of variance. The reader of *Statistical Methods* was not given the distribution theory that underpinned the development

but only the tables required for the new statistical practice. Fisher’s “Mathematical foundations” (1922a) presented a new theory of estimation but I won’t discuss it here—it went beyond the theory of errors and the only return was a new gloss on the old method of least squares; see Aldrich (1997) and Stigler (2005 and -7). The theory gets some slight attention in *Statistical Methods* when Fisher (1925a: 24-6) describes the application of maximum likelihood to a genetic example involving a curved multinomial distribution.

8 The analysis of variance and modifying χ^2

Fisher’s analysis of variance enriched the theory of errors by adding new interpretations of the basic scheme of indirect observations ((2.2) of Essay I) and by devising inferences for sets of coefficients. These extensions had precedents: Edgeworth (see Stigler (1978: 299ff)) and Thiele (see Lauritzen (2002: 222)) had both developed analysis for categorical explanatory variables and Bienaymé (see Heyde and Seneta (1977: 66ff)) had considered inference for sets of coefficients. Fisher made them core repertory.

Fisher’s analysis of variance was an amphibian inhabiting both correlation theory and error theory: the chapter in *Statistical Methods*—“Intraclass correlations and the analysis of variance”—begins by considering whether a correlation is significant and ends by analysing a field experiment. The analysis idea first appeared in Fisher’s great essay on population genetics, “Correlation between relatives” (1918), while the error theory version came out of his work on agricultural and meteorological time series and then on agricultural experiments; see Box (1978: ch. 4).

The formulation in Fisher's first time series paper, "An Examination of the Yield of Dressed Grain from Broadbalk" (1921b: 110), covers both versions:

When the variation of any quantity (variate) is produced by the action of two or more independent causes, it is known that the variance produced by all the causes simultaneously in operation is the sum of the values of the variance produced by each cause separately. ... The ... property of the variance, by which each independent cause makes its own contribution to the total, enables us to *analyse* the total, and to assign, with more or less accuracy, the several portions to their appropriate causes, or groups of cause.

Later authors specialised Fisher's amphibian: the version from least squares theory became Eisenhart's (1947) Class I and Scheffé's (1956) fixed effects, while the one from correlation became Class II or random effects. For the random effects scheme Scheffé (1956) gave references to earlier work by Chauvenet and Airy and Hald (2000) to Pizzetti—further testimony to the richness of the error theory tradition.

The object of Fisher (1921b: 109) was to "establish the existence of large changes in the mean yield [of wheat], to show how they may be disentangled from the other types of change, and to suggest their possible cause." Fisher's time series studies were based a trend-polynomial or trigonometric-plus error scheme. Fisher (1916) had welcomed Student's (1914) variate difference correlation method which was based on a polynomial trend—see E. S. Pearson (1990: 29-34) and Aldrich (1995: 371-2)—but when he came to his own empirical work he fitted trends by least squares. The "theory of polynomial fitting" is expounded

in Part II of the 1921 paper and over the years Fisher devoted considerable attention to orthogonal polynomials; in the beginning he appeared unaware of the considerable literature—see Hald (1998, ch. 25)—on the subject. The Broadbalk data comprised 70 years of yields on 13 plots and Fisher fitted fifth order polynomials where the linear trend is interpreted as “deterioration”, the sum of the higher order terms as “slow changes” and the residual as “annual causes”. For each plot a P value for the different kind of change is presented: e.g. for plot 8 a P for slow changes of .0012 is recorded and a P for deterioration of .056; Fisher’s (1921b: 110) comment on the latter is that it “shows a deterioration which would not be expected more than once in eighteen random trials. It is, therefore, probably real [..]” These P values were based on a development in distribution theory: Fisher used the χ^2 distribution for the estimates of variance and the variance attributable to the slow changes reflects the sum of 4 independent terms.

Independently of this work Fisher was reconsidering Pearson’s χ^2 theory and his first paper on the subject, “On the interpretation of χ^2 from contingency tables, and the calculation of P ,” introduced the concept of the number of degrees of freedom (1922b: 88):

the value of n' with which the table should be entered is not now equal to the number of cells, but to one more than the number of degrees of freedom in the distribution. Thus for a contingency table of r rows and c columns we should take $n' = (c - 1)(r - 1) + 1$ instead of $n' = cr$. This modification often makes a very great difference to the probability (P) that a given value of χ^2 should have been obtained by chance.

Like the analysis of variance, the concept of degrees of freedom would jump from biometry into error theory; for more on the contingency table development see Lancaster (1969).

Pearson also used χ^2 in goodness of fit tests and the first part of Fisher's "The goodness of fit of regression formulae and the distribution of regression coefficients" (1922c) was a response to his work in the regression field; the background is further discussed in Aldrich (2005). Fisher (1922c: 611) now emphasised that, "In testing the fitness of regression lines account must be taken of the number of degrees of freedom which have been absorbed in the process of fitting." In this paper Fisher (1922c: 599-600) applied the second form of his absolute criterion to produce an estimate of σ^2 that would take into account the number of constants being fitted; this adjusted form had long been standard in the theory of errors—see Essay I. Although Fisher had developed this argument in 1912 (see Section 5 above), he used Student's estimate, and not $n - 1$, in Fisher (1915, -20). Fisher (1922c: 599-601) now considered how the distribution of the χ^2 goodness of fit statistic deviated from the χ^2 form when σ is estimated in this way. In the derivation of the variance ratio statistic Fisher used χ^2 to represent the statistic but having obtained the density he introduced no special symbol and just used the generic x . Fisher describes the nature of the approximation of the new distribution (a Type VI) to that of Pearson's χ^2 (a Type III) and considers how to calculate in the absence of tables for the new distribution. In his first time series paper Fisher (1921b) had used the χ^2 without taking into account that the true variance is unknown.

Fisher extended the basic model to categorical explanatory factors for applications to agricultural experiments: the extension first appeared in Fisher and MacKenzie's "The ma-

nurial response of different potato varieties” (1923). Commentators on Fisher’s analysis of variance, like Urquhart, Weeks and Henderson (1973), emphasise the role of later Rothamsted colleagues, Wishart and Irwin, in clarifying the basis of the procedure but right at the start Fisher had produced a clear statement of the assumptions underlying the analysis of variance in an account he wrote for Gosset and which is reproduced in Student (1923: 283n).

The passage begins:

The yield obtained in any experiment is the sum of three quantities, one depending only on the variety; a second depending only on the ‘trial’; and a third, which may be regarded as the ‘experimental error’ varying independently of variety and trial in a normal distribution with a standard deviation which it is desired to estimate.

The basic model of indirect observations ((2.2) of Essay I) applied to explaining the yields X_{pq} is thus

$$X_{pq} = A_p + B_q + \varepsilon_{pq}, \quad \varepsilon_{pq} \sim IN(0, \sigma^2), \quad p = 1, \dots, m; \quad q = 1, \dots, n$$

(Fisher has no symbol for ε_{pq} but otherwise the notation is his.)

Fisher estimates the quantities A_p, B_q by least squares, minimising

$$\sum (X_{pq} - A_p - B_q)^2 \quad (1)$$

He continues, “Evidently (1) will be a minimum if

$$A_p + B_q = \bar{X}_p + \bar{X}_q - \bar{\bar{X}}$$

where \bar{X}_p is the mean of the values obtained with variety p , \bar{X}_q the mean of the values obtained with trial q , and $\bar{\bar{X}}$ is the general mean.” Next he lays out the analysis of variance table:

The actual evaluation is most conveniently carried out in the following form of the analysis of variance:

Variance	Degrees of freedom	Sum of squares
(a) Due to variety	$m - 1$	$n \sum_1^m (\bar{X}_p - \bar{\bar{X}})^2$
(b) Due to trial	$n - 1$	$m \sum_1^n (\bar{X}_q - \bar{\bar{X}})^2$
(c) Random variation	$(m - 1)(n - 1)$	$\sum_1^m \sum_1^n (X_{pq} - \bar{X}_p - \bar{X}_q + \bar{\bar{X}})^2$
(d) Total	$mn - 1$	$\sum_1^m \sum_1^n (X - \bar{\bar{X}})^2$

The sum of squares in line (c) being calculated by subtracting the values of lines (a) and (b) from the total. If either variety or ‘trial’ were without significant effect on the yield, the corresponding mean square would not differ significantly from that of line (c). To test the significance of such a difference we may use the fact that the estimates of variance in (a), (b) and (c) are all independent, and when m and n are fairly large the natural logarithms of the mean square has standard deviation $\sqrt{2/n_1}$ where n_1 is the number of degrees of freedom.

The table, the in-built identities and the idea of testing whether distinct estimates of a variance are significantly different would all become very familiar. The account in *Statistical*

Methods (1925a: 203-4) made no reference to least squares although it made the advance of replacing approximate distributions with especially tabulated exact distributions.

9 Student's z extended and transformed

In April 1922 Student's z took on extra life when Gosset asked Fisher:

I want to know what is the frequency distribution of $r\sigma_x/\sigma_y$ for small samples, in my work I want that more than the r distribution now happily solved ...”

Instead of a follow-up to Fisher (1915) came a follow-up to Student (1908a). As error theory, Student's (1908a) analysis of direct observations begged to be extended to the case of indirect observations; see Essay I and cf. Jeffreys's progress in 1931-2. When Fisher took the step in the second part of “The goodness of fit of regression formulae and the distribution of regression coefficients” (1922c) he described the result as regression theory. Fisher's reinterpretation of the Pearsonian concept of regression is treated at length in Aldrich (2005).

The regression extension of z went quickly into print but the project of constructing and publishing improved tables took years to be realised; Box (1978: 119-22) and Eisenhart (1979: 8-9) follow its slow progress. The pattern of repeated delays was already evident in October 1922 (Letter 12) when Gosset was admitting, “I haven't yet had time to do anything with the type VII.” Both were computing, Gosset using his old method and Fisher a new method of applying “correction formulae” to the value of the corresponding normal deviate;

the values which Gosset compares in his letter of the November 7th (letter 13) are identical to those published in Student (1925) and Fisher (1925c). However they were no longer computing values for z but for the transformed statistic, $z \cdot \sqrt{n-1}$. In his letter Gosset used both t and x to refer to the new statistic and it seems, as Eisenhart suggests, that Gosset introduced the special letter t where Fisher had used the generic symbol x . Subsequently they both used t .

Student (1925: 105) stated later that the transformation was Fisher's idea but the surviving letters do not indicate why they adopted it. The coarseness of the z -scale probably weighed most with Gosset although the defect could have been repaired in other ways, most simply by taking smaller steps than .1. Fisher had an interest in a standardised statistic because his new computational method was based on deviations from a standard normal distribution. But limiting standard normality could have been achieved by multiplying z by \sqrt{n} which would have fitted in with Student's view from 1908. For Fisher the appeal of multiplying z by $\sqrt{n-1}$ was presumably the interpretation

$$z \cdot \sqrt{n-1} = \frac{\bar{x} - \mu}{\sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} / \sqrt{n}} = \frac{\bar{x} - \mu}{\text{est.s.e.}(\bar{x})}$$

reflecting his conviction that the best estimate of σ^2 is $\frac{1}{n-1} \sum (x_i - \bar{x})^2$ and conforming to error theory practice. Gosset eventually came round to this estimate for in a 1926 letter (letter 73) he refers to the "correct $n-1$ formula." However it is unclear when the change occurred.

In his mature t publications Fisher used n for the number of degrees of freedom—the fundamental parameter—and n' for the sample size. Student's letter of November 1922 still

used the original notation of n for the sample size. Apparently Fisher had not yet developed this notation and most probably had not yet seen the attraction of extending the concept of degrees of freedom from χ^2 to z . In the tables of Student (1925: 105) the argument was changed from z to t and n represented the number of degrees of freedom and n' the number of observations but otherwise the tables were on the same pattern as those of 1908 and -17. Fisher (1925c: 114ff) laid out his table in the same way.

10 A system of distributions: 1923-4

In May 1923, when he was concentrating his thoughts on the analysis of variance, Fisher wrote to Gosset (letter 42, reproduced in Box (1978: 118)) describing how the various distributions he had been working with were interconnected. He elaborated on the theme in his 1924 conference paper “On a distribution yielding the error functions of several well known statistics” (1924/8) which introduced the distribution of a transformed variance ratio, and described its relationship to established distributions. Student’s z was now dead: in its original role it had been replaced by t and Fisher now re-assigned the letter z to the new distribution.

In Fisher’s previous papers there were manipulations of densities—integration or limiting arguments—but here Fisher collected and arranged known results—known to him, that is. The pattern could have been called the normal distribution and related distributions for Fisher (1924/8: 809) writes (his italics):

one series of modifications of the normal distribution gives the χ^2 distributions,

a second series of modifications gives the curves found by “Student”, while if both modifications are applied simultaneously we have the double series of distributions appropriate to z .

Fisher’s produced z from χ^2 which appears in the distribution of the sample variance. Having recalled χ^2 in its familiar goodness of fit role Fisher (1924/8: 806) gives an elementary characterisation of the χ^2 distribution:

if we have a number of quantities x_1, \dots, x_n , distributed independently in the normal distribution with unit standard deviation, and if

$$\chi^2 = S(x^2),$$

then χ^2 so defined will be distributed as is the Pearsonian measure of goodness of fit; n is, in fact, the number of independent squares contributing to χ^2 .

This proposition could have been extracted from Pearson’s 1900 paper but nobody had found any point in doing so.

Having described the χ^2 distribution as the distribution of ns^2/σ^2 where n , the number of degrees of freedom is one less than the number in the sample and s^2 is the best estimate of the variance σ^2 , Fisher (p. 807) passes to the general z distribution by the “most direct way” which is “to consider two samples of normal distributions, and how the estimates of the variance may be compared. ” The estimates are s_1^2 and s_2^2 with $n_1s_1^2/\sigma_1^2$ denoted by χ_1^2 and similarly for the second sample. He envisages two independent sums of standardised

squares S_1 and S_2 based on n_1 and n_2 independent quantities respectively. Writing

$$\begin{aligned} n_1 s_1^2 &= \sigma_1^2 \chi_1^2 = \sigma_1^2 S_1(x^2) \\ n_2 s_2^2 &= \sigma_2^2 \chi_2^2 = \sigma_2^2 S_2(x^2) \\ e^{2z} &= \frac{s_1^2}{s_2^2} = \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{n_2 S_1(x^2)}{n_1 S_2(x^2)}. \end{aligned}$$

Fisher (pp. 808-8) proceeds to obtain (transformed versions of) the χ^2 and t as special cases.

Fisher (1925b) presented another characterisation of Student's distribution. Tables of the new z appear for the first time in *Statistical Methods*.

11 The new theory of errors: 1925

The 40 odd examples of *Statistical Methods for Research Workers* showed how the theory Fisher had developed at Rothamsted was to be applied; Edwards (2005) gives an overview of the book. The book combines a reworking of the Pearsonian statistics of contingency tables and correlation with a greatly extended theory of errors. The error theory component included a system of inference for the normal regression model and for random sampling from a normal population; the book's "further applications of the analysis of variance" included "modern methods of arranging field experiments."

The book's methods are predominantly methods of significance testing based on the four key distributions, the normal, χ^2 , t and z . The estimation basics were covered but the emphasis on significance testing gave the book an appearance unlike anything in the error theory literature, unlike Brunt's book or the chapter on the "method of least squares" in

Whittaker and Robinson (1924). The language and notation were thoroughly biometrised: even the probable error was retired in favour of Yule's standard error.

P values were already widely used and assessments of significance common and there was no great novelty in what Fisher said about the purpose of testing or about the way of interpreting results. He (1925a: 8) explains that by means of tests of significance "we can examine whether or not the data are in harmony with any suggested hypothesis." P values first appear in the book in its account of the normal distribution (47): "The value for which $P = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant." In the discussion of χ^2 Fisher (79) proposed a scale:

in practice we do not want to know the exact value of P [...] but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy.

This scheme of graduating P values was not a break with earlier thinking, more a codification. Thus Ziliak and McCloskey (2007: 200) find a similar scheme involving probable errors in Pearson's unpublished lecture notes from around 1905. How Fisher thought a value exceeding .9 should be interpreted is discussed in the next section.

This P orientation inspired Fisher (1925a: 76, 98-9 and 137) to add instructions on the calculation of P to the normal table and to completely reorganise the χ^2 and t tables. Thus the t -table differs from Student's tables (1908a, -17 and -25) and from the one in Fisher (1925b): it extends from $P = 0.9$ to $P = 0.1$ at intervals of 0.01 together with the values for .05, .02 and .01 for degrees of freedom $n = 1$ to $n = 30$. For the more complicated double entry z distribution Fisher (1925a: 210) gave only the .05 values. Stigler (2008) has some remarks on Fisher's role in the popularisation of the 5% level.

From his first publication on z Fisher (1923: 658) had seen the goal as a "test of the significance of the departure of \bar{x} from its hypothetical value m ." By contrast, Student (1908a) had used z to construct probability intervals: the first of his illustrations compared the soporific properties of two drugs and he (21) concluded that "the odds are about 666 to 1 that 2 is the better soporific." In his reworking Fisher (1925a: 107-8) found a t value of 4.06 and concluded, "For $n = 9$ only one value in a hundred will exceed 3.250 by chance, so that the difference between the results is clearly significant." *Statistical Methods* has examples of intervals based on large sample normal approximation—they are described in Aldrich (2000)—but interval inference based on t appeared only in the 1930s. By the end of the decade there were three versions: the fiducial interval of Fisher (1935), the confidence interval of Neyman (1934) and the Bayesian interval of Jeffreys (1939).

By 1925 Fisher and Student had made their fundamental contribution to the Gauss linear model and the distribution theory and applications would be central to statistics for decades. In works like Scheffé's *Analysis of Variance* (1959) and Draper and Smith's *Applied*

Regression Analysis (1966) F replaced Fisher's z and use of matrices smoothed the exposition but the foundations were those laid in 1925.

12 An unusual collaboration?

Fisher and Gosset disagreed on fundamentals and so their collaboration may seem surprising. The 70 surviving letters from 1922-6 (all but one from Gosset to Fisher) testify to the reality of the collaboration and the two clearly found virtues in each other and they were committed to some of the same causes, including the promotion of Student's distribution. Their differences did not seem to have affected the course of the collaboration and so it is possible to collect their views on fundamentals having described their collaboration.

Fisher formed and broadcast general views on fundamental principles and on the nature of the discipline but, for Gosset, "my work" involved individual practical problems. The articles of 1908 express a position on fundamentals—see Section 4 above—but Gosset's later publications add, or subtract, little: to modern ears he may sometimes sound like Harold Jeffreys or Robert Schlaiffer but he created no systems like their's. Egon Pearson (1990: 95) who knew Gosset from 192? summed him up thus:

Gosset emerges as a practical man who combined probability measures derived from application of what were the theoretically correct methods for testing a statistical hypothesis or for estimation, with other considerations—vague prior knowledge, economic limitations, and approximateness of the mathematical model.

In Section 6 we left Gosset in 1915 awaiting the Bayesian conclusion to Fisher’s correlation work and they next discussed the Bayesian approach in 1922. In the meantime, however, they had been in contact with Pearson and exchanges with him led them to change, or to clarify, their positions on inference. At issue were methods for estimating the standard deviation and the correlation coefficient. In 1915 and -17 Pearson published “appendices” to Student (1908a) and to Student (1908a) and Fisher (1915). Gosset’s exchange with Pearson led him away from informative priors while Fisher’s led him to distinguish likelihood from probability and to emphasise his opposition to Bayesian thinking;. The Fisher development is treated in Aldrich (1997: 8ff) and here I focus on the Gosset-Pearson exchange.

In the first appendix Pearson (1915) proposed a new way of estimating σ : on the basis of the relationship between the mode of the density for the sample standard deviation and the true parameter value Pearson argued that the “most reasonable value” for σ was not Student’s s but s multiplied by $\sqrt{\frac{n}{n-2}}$. In a letter, reproduced in E. S. Pearson (1990: 26), Gosset demurred and argued for maximising the posterior density obtained by multiplying the sampling density by a fairly flat prior, leading in practice to his s .

Pearson, writing as “the Editor” (p. 353n) in the second appendix, the “cooperative study” of Soper, Young, Cave, Lee & Pearson (1917), conceded that Student had made a “desirable criticism.” However, Student’s counter-proposal was not acceptable because the uniform prior is not “in accordance with our experience.” Accordingly the cooperators (1917: 354n)) produced a “most likely” value of the correlation in the sampled population using an informative prior. Gosset told Pearson he was “not altogether sure that I quite agree with

all you say about Bayes.” Two of the three points in his letter (reproduced in E. S. Pearson (1990: 27)) concern the use of an informative prior:

(2) [...] but I am not sure that it is easy to make use of knowledge of similar correlations without destroying the independence of the result in question (vide infra) But would it not be possible to compare various a priori distributions of ignorance not equal. I take it we should all rule out U curves but possibly if we wrote for $\varphi(\rho)$ $1 - \rho^2$ or $(1 - \rho^2)^2$ or even $(1 - \rho^2)^3$ we should get distributions of ignorance more appropriate for correlations in general, i.e. of ignorance concerning the particular subject matter but not of correlation in general.

(3) The disadvantage of using actual knowledge concerning similar work is that you destroy the independence of the work before you. [...] Surely it is better to be able to say ‘From our general experience of the correlation coefficient the population of which this is a sample probably had a correlation coefficient of .58 but this is much higher than that found from similar populations which have a mean of about .40’ than to say ‘Combining our knowledge of similar populations with the actual result before us the population in question probably had a correlation coefficient of about .45.’

Point (2) echoes the letter to Fisher of 1915 and ultimately the correlation paper of 1908 but the idea of comparison is new, while point (3) introduced a consideration that seemed to undermine the use of informative priors.

Student's 1917 piece extending the 1908a tables also has an indication that he was exercised by the issue of "independence." The phrase "unique sample" appears in the title, "Tables for estimating the probability that the mean of a unique sample of observations lies between $-\infty$ and any given distance of the mean of the population from which the sample is drawn": Student (1917: 414) explained:

By unique I mean to say that all the information which we have (or at all events intend to use) about the distribution of the population is given by the sample in question.

This echoes the "only knowledge" qualification in the original article—see Section 4 above.

Fisher (1921a: 17) made a similar point when he insisted that the posterior value produced by the cooperators "depends almost wholly upon the preconceived opinions of the computer and scarcely at all upon the actual data supplied to him." However while Student was a Bayesian wary of *informative* priors, it is clear from Fisher (1921a) that he was not a Bayesian at all and his method, soon to be christened maximum likelihood, did not rest on Bayes' theorem.

Fisher sent Gosset his 1921a and Gosset replied on April 3 1922 (letter 5 of McMullen (1970)), recalling his work on the 1908 papers:

When I was in the lab. in 1907 I tried to work out variants of Bayes with a priori probabilities other than $G = C$ but I soon convinced myself that with ordinary sized samples one's á priori hypothesis made a fool of the actual sample (as the

cooperators found) and since then have refused to use any other hypothesis than the one which leads to your likelihood (where I could deal with the mathematics).

Then each piece of evidence can be considered on it's own merits.

Actually this had been his position only since 1917 when he saw the cooperators' carrying out his own suggestion! Fisher had already attacked the $G = C$ reasoning and would do so again in the "Mathematical Foundations" (1922a). Gosset did not respond to these criticisms or show any interest in Fisher's high theory. He appears to have thought that there was no practical difference between likelihood and Bayes with a uniform prior.

The Bayesian question returned in 1923 when William Burnside (1923) published a Bayesian treatment of Student's problem. Fisher (1923) produced a note to register Student's priority and to put his own derivation into print. On the argument he (1923: 658) made only the neutral observation:

The slight difference [...] is traceable to Dr Burnside's assumption of an a priori probability for the precision constant, whereas Student's formula gives the actual distribution of z in random samples.

The difference between his posterior distribution and Student's sampling distribution arose because Burnside's prior did not lead to an adjustment for the degree of freedom lost in estimating the mean. (Fisher and Burnside went on corresponding for several years—see Aldrich (2009)).

In 1908 Gosset might have welcomed the contribution but in 1923 he had only this comment, "It is interesting to see how à priori probability has got him just off the line." (July

3, letter 25). Burnside did not get exactly Student's density and his intervals, like Lüroth's, were narrower than the customary ones. Besides the contributions by Burnside and Fisher in the Cambridge mathematics journal there was a response in *Biometrika* from Ethel Newbold. Newbold (1923) generalised Burnside's argument to other parametrisations of precision and found a prior that would produce Student's density as the posterior. However she (p. 405) did not advocate the Bayesian scheme she had devised, observing that the Student case fits "without burdening itself with the *à priori* assumptions attached to the other cases." Gosset commented to Fisher (number 39 in McMullen (1970), undated and inserted between December 1923 and January 1924):

You have dealt much more kindly with Burnside than has Miss Newbold who is just a little acrimonious I don't quite know why.

But it is a good instance of the futility of *à priori* assumptions.

Whether the last sentence is a general renunciation of all "*à priori*" assumptions or only some is unclear. E. S. Pearson (1990: 72-3) recalls Gosset suggesting a uniform prior for σ to him in May 1926 and comments that Student's presentation of the subject was "at times a little contradictory." In the last year of his life Gosset had some correspondence with Jeffreys about the latter's "Relation between direct and inverse methods" which Jeffreys (1939: 311) interpreted as an endorsement of his fiducial reconstruction of Student's reasoning; the correspondence appears not to have survived

Turning to *Statistical Methods*, Gosset had opportunities to express his views for he read the proofs and reviewed (1926) the book. Gosset was not afraid of disagreeing but he made *no*

comments on Fisher's handling of Student's distribution. Student did not entirely agree with Fisher on significance tests. Egon Pearson's "practical man" objected to Fisher's reaction to small values of χ^2 . The curve fitter—and both had been curve-fitters—wanted a good fit, a small value of χ^2 and a correspondingly large value of P . Indeed in Chapter I of *Statistical Methods* Fisher (1925a: 9) wrote that "if [the available observations] are completely in accord with the deductions, the hypothesis may stand . . ." However in the chapter on χ^2 Fisher (p. 80) added the twist that it is a fallacy to believe that "the higher the value of P the more satisfactorily is the hypothesis verified"—indeed, with a value of .999 the hypothesis is "as definitely disproved as if P had been .001." Student (1926) could not "altogether endorse" this because

in the first place there is this fundamental difference between the values .001 and .999: —while it is generally easy to formulate any number of likely hypotheses to account for the low value, it is apart from blunders or cheating, exceptional to find an alternative explanation to account for .999 which is not itself extremely improbable: in the second place χ^2 , with small samples, can only take a limited number of values of which zero, corresponding to $P = 1$, is not unlikely to occur by chance.

Gosset's reaction to Fisher's proposal for treating this pathological case brought out concerns that Gosset would presumably have applied to all cases—the role of "alternative explanations" and the greater or less improbability attached to them. The same concerns inform a letter Gosset wrote to Egon Pearson on 11 May 1926. (E. S. Pearson (1938: 243)

recalled the letter as having an important influence on his (and Neyman's) thinking about tests.) Gosset had written:

In your large samples with a known normal distribution you are able to find the chance that the mean of a random sample will lie at any given distance from the mean of the population. ... That doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the chance is very small, say .00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say .05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true.

I can conceive of circumstances, such for example as dealing a hand of 13 trumps after careful shuffling by myself, in which almost any degree of improbability would fail to shake my belief in the hypothesis that the sample was in fact a reasonably random one from a given population.

In this case Student was dealing with impressively low values of P and why he would not always be impressed by them. This was as much an objection to Edgeworth and Karl Pearson as to Fisher.

Student's final publication on Student's distribution was a response to Karl Pearson's (1931) criticism of the z -test. Student (1931: 408) explained "what we actually ask ourselves" when we calculate z :

If the average difference between A and B in the population were zero, what would be the probability of obtaining a sample of differences giving a value of z as high as that observed? and if this probability is sufficiently small we say that the difference is significant.

This was how Fisher understood the test. However, it seems from Gosset's comments of 1926 that he would embed the results of a significance test in an informal Bayesian analysis, treating the hard P -value as evidence and combining it with informal assessments of one's "belief in the hypothesis" and in relevant alternatives and an assessment of the reasonableness of the explanation the alternatives offer. Gosset did not want to publish formal Bayesian analysis because he did not want the prior to make "a fool of the actual sample" but then he did not want the actual sample to make a fool out of him.

Statistical Methods contained no inference principle that Student rejected but Fisher completely extinguished the Bayesian line in Student's work. Looking back from after the inference wars between Fisher, Jeffreys, Neyman, etc., Student may seem puzzlingly incoherent. However, judged against his predecessors and older contemporaries in the theory of errors, biometry and statistics, Fisher's insistence on a single line made him the odd one out.

13 Placing Fisher

The theme of Fisher's *Statistical Methods* is announced in the Preface (1925a: vi):

The elaborate mechanism built on the theory of infinitely large samples is not

accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data. Such at least has been the aim of this book.

And the reviewers responded: “The book will undoubtedly prove of great value to research workers whose statistical series necessarily consist of small samples” wrote Isserlis (1926: 146). Only one reviewer was such a worker and Student (1926: 148) derived “particular pleasure” from the publication of the book. Student was also the only reviewer familiar with such novelties as the analysis of variance.

The book’s “practical data” came from biometry, agricultural meteorology, genetics, medicine, bacteriology, algology, agricultural experiments and social statistics, rather than from the traditional error theory sources. Indeed the book was not seen as belonging there: astronomers did not review the book and the reviewers in the general science periodicals—Anon. (1925) in *Nature* and E. S. Pearson (1926) in *Science Progress*—made no link. The reviewers appreciated that the book rested on Fisher’s own papers—ten are listed in the sources—but otherwise it bore the Pearson stamp. Two of the three modern distributions had Pearsonian origins, although Fisher (1925a: 16-7) could not resist pointing out that Pearson (1900) contained an error. One reviewer—Anon. (1926: 579)—so disapproved that he summoned Macaulay, “we have heard a baby, mounted on the shoulders of its father, cry out, ‘how much taller I am than Papa!’”

Fisher (1925a: 3) described the theory of errors as “one of the oldest and most fruitful lines of statistical investigation” without saying how fruitful it had been for *Statistical Methods*.

Some readers saw the links. Isserlis (1926: 146) commented on the chapter on means and regression coefficients that it “is full of good matter, but the author, here as elsewhere, is very economical in his references to earlier work.” The American econometrician Henry Schultz (1929: 86) “regretted” that Fisher “did not see fit clearly to separate the propositions which are due to him from the general body of statistical theory.” Isserlis was exercised by Fisher’s failure to refer to Chebyshev for orthogonal polynomials and Schultz by his failure to refer more fully to Gauss; neither made any assessment of the advances Fisher had made on the traditional theory.

Elsewhere Fisher would write more on the error theory setting of his work. His memoir “Student” (1939) takes the long view and has reflections on Gauss and Helmert. For Fisher, writing up Student and the theory of errors went with writing down Pearson or at least separating himself from Pearson. When Maurice Fréchet was inquiring into correlation on behalf of the International Statistical Institute Fisher told him (letter of 6 November 1934):

it would be of service to many who imagine that the quadratic analysis of measurements or frequencies is bound up with the correlation technique of my distinguished predecessor to show that the methods developed by Gauss and the early exponents of the theory of errors are more widely applicable and more appropriate in the majority of cases.

When Karl Pearson died in April 1936 the *Annals of Eugenics* he had founded and which Fisher was now editing published no tribute but it gave W. F. Sheppard a fine send-off when he died six months later. Fisher’s (1937: 12) eulogy contains this remarkable diseulogy of

Pearson:

Throughout Sheppard's writings we find practically nothing that ought to be retracted, and very little that is now obsolete. This is the more remarkable as the period in which he wrote was one of rapid development, and one which threw up many passing fashions in statistics, which failed to sustain the enthusiasm with which they were promoted. The proliferation of correlation coefficients and correlation ratios—"equiprobable r ", "biserial r ", "biserial η ", etc.—evidently had no appeal for him. To a man of his knowledge we can understand that the so-called "calculus of correlations" appeared as no more than the older method of least squares, in a new and not always appropriate notation. His restraint is equally marked with respect to the once highly popular "Pearsonian frequency curves"; at the height of their popularity he published his table of the despised normal frequency distribution. Equally we find nothing in his writings on such temporary vogues as the variate difference correlation method; instead he developed methods for polynomial fitting.

Apart from the reference to the calculus of correlations for which Yule was responsible—see Aldrich (1998)—this is a long indictment of Pearson's work; in fact, Fisher himself fell for the variate difference correlation method and Fisher devoted around a third of his "Mathematical foundations" (1922a) to those Pearson frequency curves.

14 Postscripts

Fisher’s admiration for “Gauss and the early exponents of the theory of errors” did not lead him or his immediate followers to make a close study of their work and its rediscovery, outlined in the General Introduction, came only later. Fisher also contributed to the relocation of the subject. While he was primarily a biological scientist, he and many of his followers, like Pearson’s before them, cast themselves as statisticians and set up in the Royal Statistical Society; see Aldrich (2010) for the consequent reorientation of statistics.

Essay I considered the astronomers’ theory of errors, Essay II described where Fisher took that theory and an Essay III might examine the way astronomers and geodesists received the work of Fisher and the mathematical statisticians. Some notice was taken in the 30s, as by Birge and Deming (1934) and by Harold Jeffreys. Jeffreys’s *Theory of Probability* (1939) appeared in the International Series of Monographs on Physics but it was read more by statisticians than by physicists; see Aldrich (2005b) for the development of Jeffreys’s ideas and Aldrich (2003/-10) for the impact of his work. Sheynin (1996: 121) mused in his *History of the Theory of Errors*, “perhaps there still (?) exist two versions of the theory: an astronomic-geodetic, and a statistical one.” The history of English textbooks indicates a changing relationship. Four decades after publishing Brunt (1917)—see Section 5 above—Cambridge University Press published a new *Combination of Observations* by the astronomer W. H. Smart. This was untouched by developments in mathematical statistics—indeed it contained little that was not in Brunt. However the Press’s current offering, Wall and Jenkins’s *Practical Statistics for Astronomers* (2003) which first appeared six decades after

Brunt is, as the title suggests, greatly touched by those developments although its orientation owes more to Jeffreys than to Fisher. A noted geodesist has recently produced a *general* text on Bayesian analysis, Koch (2003).

Bibliography

Where a reprint or a translation is given the page references in the text above are to that later version. Many of the original publications are now available on-line.

Airy, G. B. (1861) *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*, third edition in 1879, London: Macmillan.

Aldrich, J. (1995) Correlations Genuine and Spurious in Pearson and Yule, *Statistical Science*, **10**, 364-376.

_____ (1997) R. A. Fisher and the Making of Maximum Likelihood 1912–1922, *Statistical Science*, **12**, 162-176.

_____ (1998) Doing Least Squares: Perspectives from Gauss and Yule, *International Statistical Review*, **66**, (1), 61-81.

_____ (2000) Fisher's 'Inverse Probability' of 1930, *International Statistical Review*, **68**, 155-172.

_____ (2001/10) *Karl Pearson: A Reader's Guide*, website

<http://www.economics.soton.ac.uk/staff/aldrich/kpreader.htm>

_____ (2003) The Language of the English Biometric School, *International Statistical Review*, **70**, 109-131.

_____ (2003/10) *A Guide to R. A. Fisher*, website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm>

_____ (2003/10) *Harold Jeffreys as a Statistician*, on the website

<http://www.economics.soton.ac.uk/staff/aldrich/jeffreysweb.htm>

_____ (2005a) Fisher and Regression, *Statistical Science*, **20**, (4), 401-417.

_____ (2005b) The Statistical Education of Harold Jeffreys, *International Statistical Review*, **73**, 289-308.

_____ (2007) The Enigma of Karl Pearson and Bayesian Inference, paper based on a talk given at the *Karl Pearson Sesquicentenary Conference*, the Royal Statistical Society March 2007.

_____ (2008) R. A. Fisher on Bayes and Bayes' Theorem, *Bayesian Analysis*, **3**, (1), 161-170.

_____ (2009) Burnside and his Encounters with "Modern Statistical Theory," *Archive for History of Exact Sciences*, **63**, (1), 51-79.

_____ (2010) Mathematics in the London/Royal Statistical Society 1834-1934, *Journal Electronique d'Histoire des Probabilités et de la Statistique*, **6**, (1): pp. 33.

Anon. (1925) Review of The Fundamentals of Statistics by Prof. L. L. Thurstone and Statistical Methods for Research Workers by Mr. R. A. Fisher, *Nature*, **116**, 815. Reproduced with notes by J. Aldrich on the website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/nature.htm>

Anon. (1926) Review of Statistical Methods for Research Workers (R. A. Fisher), *British Medical Journal*, **1**, 578-9. Reproduced with notes by J. Aldrich on the website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/bmj.htm>

Baccini, A. (2009) F. Y. Edgeworth's Treatise on Probabilities, *History of Political Economy*, **41**, 143-162.

Behrens, W. V. (1929) Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen, *Landwirtschaftliche Jahrbücher*, **68**, 807-837.

Bennett, T. L. (1908) Errors of Observation, Technical Lecture 4, Cairo, Ministry of Finance, Survey Department Egypt.

Bertrand, J. L. F. (1889) *Calcul des probabilités*, Paris: Gauthier-Villars.

Bessel, F. W. (1815) Ueber den Ort des Polarsterns. In *Astronomisches Jahrbuch oder Ephemeriden für das Jahr 1818* (J. E. Bode, ed.). pp. 233-241. Königliche Akademie der Wissenschaften, Berlin.

_____ (1816) Untersuchungen über die Bahn des Olbersschen Kometen, *Abhandlungen der mathematischen und physikalischen Klassen - Königliche Preussische Akademie der Wissenschaften zu Berlin*, 119-160.

_____ (1838) Untersuchung über die Wahrscheinlichkeit der Beobachtungsfehler, *Astronomische Nachrichten*, **15**, 369-404.

Biermann, K.-R. (1971) Richard Dedekind, *Dictionary of Scientific Biography*, **4**, 1-5 New York: Scribner.

Birge, R. T. & W. E. Deming (1934) On the Statistical Theory of Errors, *Review of Modern Physics*, **6**, 119-161.

Boole, G. (1854) Solution of a Question in the Theory of Probabilities, *London, Edinburgh*

and *Dublin Philosophical Magazine*, 4th series, **7**, 29-32.

Box, J. F. (1978) *R. A. Fisher: The Life of a Scientist*, New York: Wiley

_____ (1981) Gosset, Fisher and the t Distribution, *American Statistician*, **35**, 61-66.

_____ (1987) Guinness, Gosset, Fisher, and Small Samples, *Statistical Science*, **2**, 45-52.

Bravais, A. (1846) Analyse mathématique sur les probabilités des erreurs de situation d'un point, *Mémoires presentés par divers savants à l'Académie royale des sciences de l'Institut de France*, **9**, 255-332.

Broemeling, L. & A. Broemeling (2003) Studies in the History of Probability and Statistics XLVIII: The Bayesian Contribution of Ernest Lhoste, *Biometrika*, **90**, 720-723.

Brunt, D. (1917) *The Combination of Observations*, Cambridge: Cambridge University Press.

Burnside, W. (1923) On Errors of Observation, *Proceedings of the Cambridge Philosophical Society*, **21**, 482-487.

Cayley, A. (1853) Note on a Question in the Theory of Probabilities, *London, Edinburgh and Dublin Philosophical Magazine*, 4th series, **6**, 259.

Chauvenet, W. (1863) On the Method of Least squares. An Appendix to *A Manual of Spherical and Practical Astronomy*, Vol. 2: 469-566. Philadelphia: Lippincott.

Cochran, W. G. (1980) Fisher and the Analysis of Variance: 17-34 in S. E. Fienberg & D.

V. Hinkley (1980) (eds) *R. A. Fisher: An Appreciation*, New York: Springer.

Czuber, E. (1891) *Theorie der Beobachtungsfehler*, Leipzig: Teubner.

_____ (1899) Die Entwicklung der Wahrscheinlichkeitstheorie und ihrer Anwendungen,

Jahresbericht der Deutschen Mathematiker-Vereinigung, **7**, .

Dale, A. I. (1999) *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, expanded second edition, New York: Springer.

Dauben, J. W. (1979) *Georg Cantor: his Mathematics and Philosophy of the Infinite*, Cambridge, Mass.: Harvard University Press.

David, H. A. (1998) Early Sample Measures of Variability, *Statistical Science*, **13** (4), 368-377.

Dedekind, R. (1855) Bemerkungen zu einer Aufgabe der Wahrscheinlichkeitsrechnung, *Journal für reine und angewandte Mathematik*, **50**, 268-271. Gesammelte mathematische Werke, (page(s) 36-39.

_____ (1860a) Mathematische Mittheilungen III: Ueber die Elemente der Wahrscheinlichkeitsrechnung, *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, **5**, 66-75. *Gesammelte mathematische Werke*, **1**, 88-94.

_____ (1860b) Mathematische Mittheilungen: IV Ueber die Bestimmung der Präcision einer Beobachtungsmethode nach der Methode der kleinsten Quadrate, *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, **5**, 76-83. *Gesammelte mathematische Werke*, **1**, 95-100.

_____ (1901) Gauss in seiner Vorlesung über die Methode der kleinsten Quadrate, Festschrift zur Feier des hundertfünfzigjährigen Bestehens der Königlichen Gesellschaft der Wissenschaften zu Göttingen, 45-59. Reprinted in R. Fricke, E. Noether and Ø. Ore (eds) *Gesammelte mathematische Werke, Vol. II*: 293-306. Wieweg, Braunschweig, 1931.

Deming, W. E. and R. T. Birge (1934) On the Statistical Theory of Errors, *Review of Modern Physics*, **6**, 119-161.

Draper, N. R and H. Smith (1966) *Applied Regression Analysis*, New York: Wiley.

Dunnington, G. W. (1954) *Carl Friedrich Gauss: Titan of Science* (new edition 2004 with additional material by Jeremy Gray and Fritz-Egbert Dohse), New York: Mathematical Association of America.

Edgeworth, F. Y. (1883) The Method of Least Squares, *Philosophical Magazine*, **16**, 360-375.

Edwards, A. W. F. (2005) R. A. Fisher, Statistical Methods for Research Workers, 1925, ch. 67 and pp. 856-870 of I. Grattan-Guinness (ed.) *Landmark Writings in Western Mathematics : Case Studies, 1640-1940*, Amsterdam: Elsevier.

Eisenhart, C. (1947) The Assumptions Underlying the Analysis of Variance, *Biometrics*, **3**, 1-21.

_____ (1979) On the Transition from Student's z to Student's t , *American Statistician*, **33**, 6-10.

Encke, J. F. (1832-4) Über die Methode der kleinsten Quadrate, *Berliner Astronomisches Jahrbuch für 1834*, 249-312 für 1835, 253-320; für 1836, 253-308. There is a translation of the first part in pp. 317-369 of Richard Taylor (ed.) (1841) *Scientific Memoirs*, vol. 2, London:

Farebrother, R. W. (1997) A. C. Aitken and the Consolidation of Matrix Theory, *Linear Algebra and its Applications*, 264, 3-12.

_____ (1999) *Fitting Linear Relationships: A History of the Calculus of Observations 1750—1900*, New York: Springer.

Fisher, R. A. (1911) Paper on 'Heredity' (comparing methods of Biometry and Mendelism), unpublished paper reprinted as pp. 155-162 of Norton and Pearson (1976).

Fisher, R. A. (1912) On an Absolute Criterion for Fitting Frequency Curves, *Messenger of Mathematics*, **41**, 155-160.

_____ (1915) Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, *Biometrika*, **10**, 507-521.

_____ (1916) *Biometrika*, *Eugenics Review*, **8**, 62-64.

_____ (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance, *Transactions of the Royal Society of Edinburgh*, **52**, 399-433.

_____ (1920) A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error, *Monthly Notices of the Royal Astronomical Society*, **80**, 758-770.

_____ (1921a) On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample, *Metron*, **1**, 3-32.

_____ (1921b) Studies in Crop Variation. I. An Examination of the Yield of Dressed Grain from Broadbalk, *Journal of Agricultural Science*, **11**, 107-135.

_____ (1922a) On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309-368.

_____ (1922b) On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P , *Journal of the Royal Statistical Society*, **85**, 87-94.

_____ (1922c) The Goodness of Fit of Regression Formulae and the Distribution of

Regression Coefficients, *Journal of the Royal Statistical Society*, **85**, 597-612.

_____ (1923) Note on Dr. Burnside's Recent Paper on Errors of Observation, *Proceedings of the Cambridge Philosophical Society*, **21**, 655-658.

_____ (1924/8) On a Distribution Yielding the Error Functions of Several Well Known Statistics, *Proceedings of the International Congress of Mathematics, Toronto*, **2**, 805-813.

_____ (1925a) *Statistical Methods for Research Workers*, (second edition in 1928, fourth in 1932) Edinburgh: Oliver & Boyd.

_____ (1925b) Applications of 'Student's' Distribution, *Metron*, **5**, 90-104.

_____ (1925c) Expansion of 'Student's' Integral in Powers of n^{-1} , *Metron*, **5**, 109-120.

_____ (1930) Inverse Probability, *Proceedings of the Cambridge Philosophical Society*, **26**, 528-535.

_____ (1935) *The Design of Experiments*, Edinburgh: Oliver and Boyd.

_____ (1937) The Character of W. F. Sheppard's Work, *Annals of Eugenics*, **8**, 11-12.

_____ (1939) "Student", *Annals of Eugenics*, **9**, 1-9.

_____ (1956) *Statistical Methods and Scientific Inference*, Edinburgh: Oliver & Boyd.

Fisher, R. A. and W. A. Mackenzie (1922) The Correlation of Weekly Rainfall (with discussion), *Quarterly Journal of the Royal Meteorological Society*, **48**, 234-245.

_____ (1923) Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties, *Journal of Agricultural Science*, **13**, 311-320.

Fricke, R., E. Noether and Ø. Ore (1930) Schriften von R. Dedekind in *Gesammelte mathematische Werke*, (pp. 505-507.

Gauss, C. F. (1809) *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Perthes et Besser, Hamburg. *Werke*, **7**, 1-280. Translated by C. H. Davis as *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*. Little, Brown, Boston, 1857. Reprinted by Dover, New York, 1963.

_____ (1816) Bestimmung der Genauigkeit der Beobachtungen, *Zeitschrift für Astronomie und verwandte Wissenschaften*, **1**, 185-196. *Werke*, **4**, 109-117. Translated as “The Determination of the Accuracy of Observations” in H. A. David & A. W. F. Edwards (ed.) (2001) *Annotated Readings in the History of Statistics*, Springer New York.

_____ (1823) *Theoria combinationis observationum erroribus minimum obnoxiae*, translated by G. W. Stewart as *Theory of the Combination of Observations Least Subject to Errors: Part One, Part Two, Supplement: Supplement Pt. 1 & Pt. 2*, New York: SIAM.

_____ (1839) Letter to Bessel, 28 February 1839. *Briefwechsel zwischen Gauss and Bessel*: 523-525. Engelmann, Leipzig, 1880; *Werke*, **8**, 146-147.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2003) *Bayesian Data Analysis*, 2nd edition, London: Chapman and Hall.

Gosset, W. S. (1904) *The Application of the “Law of Error” to the work of the Brewery*, privately printed.

_____ (1905) *The Pearson Co-efficient of Correlation*, privately printed.

Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

_____ (1999) On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares, *Statistical Science*, **14**, 214-222.

_____ (2000) Studies in the History of Probability and Statistics XLVII. Pizzetti's Contributions to the Statistical Analysis of Normally Distributed Observations, 1891, *Biometrika*, **87**, 213-217.

_____ (2007) *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713—1935*, New York: Springer.

Hanley, J. A., M. Julien and E. E. M. Moodie (2008) Student's z , t , and s : What if Gosset had R ? *American Statistician*, **62**, (1), 64-69

Helmert, F. R. (1876) Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Fehlers directer Beobachtungen gleicher Genauigkeit, *Astronomische Nachrichten*, **88**, 113-132.

Heyde, C. C. and E. Seneta (1977) *I. J. Bienaymé: Statistical Theory Anticipated*, New York: Springer

Hotelling, H. (1927) Review of Statistical Methods for Research Workers by R. A. Fisher, *Journal of the American Statistical Association*, **22**, 411-412. Reproduced with notes by J. Aldrich on the website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/hotelling.htm>

Isserlis, L. (1926) Review of Statistical Methods for Research Workers (R. A. Fisher), *Journal of the Royal Statistical Society*, **89**, 145-146. Reproduced with notes by J. Aldrich on the website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/isserlis.htm>

Jeffreys, H. (1931) *Scientific Inference*, Cambridge: Cambridge University Press.

_____ (1932) On the Theory of Errors and Least Squares, *Proceedings of the Royal Society, A*, **138**, 48-55.

_____ (1937) On the Relation between Direct and Inverse Methods in Statistics, *Proceedings of the Royal Society, A*, **160**, 325-348.

_____ (1939) *Theory of Probability*, Oxford: Oxford University Press.

Knoebloch, E. (1992) Historical Aspects of the Foundations of Error Theory: 253-279 of J. Echeverría, A. Ibarra, T. Mormann (ed.) *The Space of Mathematics: Philosophical, Epistemological, and Historical Explorations*, Berlin: de Gruyter.

Koch, K.-R. (2003) *Introduction to Bayesian Statistics*, 2nd edition, New York: Springer.

Knus, M.-A. (1982) Dedekind und das Polytechnikum in Zürich, *Abhandlungen der Braunschweiger Wissenschaftlichen Gesellschaft*, **33**, 43-60.

Lancaster, H. O. (1969) *The Chi-squared Distribution*, New York: Wiley.

Landau, E. (1917) Richard Dedekind, *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Geschäftliche Mitteilungen*, 50-70.

Laplace: S. (1812) *Théorie analytique des probabilités*, (2nd ed. 1814, 3rd ed. 1820, 4th supplement 1825), Paris: Courcier

_____ (1820) Mémoire sur le flux et le reflux de la mer. *Mém. Acad. R. Sci. Inst. Fr.* **3**, 1-90. *Oeuvres complètes*, **12**, 473-546, 14 vols. 1878-1912. Paris: Gauthier-Villars.

Lauritzen, S. L. (2002) *Thiele: Pioneer in Statistics*, Oxford: Oxford University Press

Lhoste, E. (1923) Le calcul des probabilités appliqué à l'artillerie, lois de probabilité a priori, *Revue d'artillerie*, Jul. 58-82.

Lupton, S. (1898) *Notes on Observations: Being an Outline of the Methods used for determining the Meaning and Value of Quantitative Observations and Experiments in Physics and Chemistry, and for Reducing the Results Obtained*, London: Macmillan.

Lüroth, J. (1862) Ephemeride der Calypso, *Astronomische Nachrichten*, **57**, 135.

_____ (1869) Bemerkung über die Bestimmung des wahrscheinlichen Fehlers, *Astronomische Nachrichten*, **73**, 187–190.

_____ (1876) Vergleichung von zwei Werten des wahrscheinlichen Fehlers, *Astronomische Nachrichten*, **87**, 209–220.

_____ (1880) Ein Problem der Fehlertheorie, *Zeitschrift für Vermessungswesen*, **9**, 432–438.

Mercer W. B. and A. D. Hall (1911) The Experimental Error of Field Trials, *Journal of Agricultural Science*, **4**, 107–132.

Merriman, M. (1877) A List of Writings Relating to the Method of Least Squares, with Historical and Critical Notes, *Transactions of the Connecticut Academy*, **4**, 151–232.

_____ (1884) *A Text-book on the Method of Least Squares*, (eighth edition 1903), New York: Wiley.

Miller, J. (Ed.) (continuing) *Earliest Known Uses of Some of the Words of Mathematics*, on the website

<http://jeff560.tripod.com/mathword.html>

Miller, J. (Ed.) (continuing) *Earliest Uses of Symbols in Probability and Statistics*, on the website

<http://jeff560.tripod.com/stat.html>

Newbold, E. (1923) Note on Dr Burnside's Paper on Errors of Observation, *Biometrika*, **15**, 401-406.

Neyman, J. (1934). On the two different aspects of the representative method, (with discussion). *Journal of the Royal Statistical Society* , **97** 558-625.

Norton, B. J. and E. S. Pearson (1976) A Note on the Background to, and Refereeing of, R. A. Fisher's 1918 Paper 'On the Correlation between Relatives on the Supposition of Mendelian Inheritance' *Notes and Records of the Royal Society of London*, **31**, 151-162.

Pearson, E. S. (1926) Statistical Methods for Research Workers (R. A. Fisher), *Science Progress*, **20**, 733-734. Reproduced with notes by J. Aldrich on the website

<http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/esp.htm>

_____ (1939) "Student" as Statistician, *Biometrika*, **30**, 210-250.

_____ (1967) Studies in the History of Probability and Statistics. XVII: Some Reflexions on Continuity in the Development of Mathematical Statistics, 1885-1920, *Biometrika*, **54**, 341-355.

_____ (1990) *Student, A Statistical Biography of William Sealy Gosset*, Edited and Augmented by R. L. Plackett with the Assistance of G. A. Barnard, Oxford: University Press.

Pearson, K. (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society A*, **186** 343-414.

_____ (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.

_____ (1907) On the Influence of Past Experience on Future Expectation, *Philosophical Magazine*, **13**, 365-378.

_____ (1914) *Tables for Statisticians and Biometricians*, Cambridge: Cambridge University Press.

_____ (1915) (Editorial) On the Distribution of the Standard Deviations of Small Samples: Appendix I to Papers by 'Student' and R. A. Fisher, *Biometrika*, **10**, 522-529.

_____ (1931) (Editorial) Historical Note on the Distribution of the Standard Deviations of Samples of any Size Drawn from an Infinitely Large Normal Parent Population, *Biometrika*, **23**, 416-418.

Pearson, K. & L. N. G. Filon (1898) Mathematical Contributions to the Theory of Evolution IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation, *Philosophical Transactions of the Royal Society A*, **191**, 229-311.

Peters, C. A. F. (1856) Über die Bestimmung des wahrscheinlichen Fehlers einer Beobachtung aus den Abweichungen der Beobachtungen von ihrem arithmetischen Mittel, *Astronomische Nachrichten*, **44**, 29-32.

Pfanzagl, J. & O. Sheynin (1996) Studies in the History of Probability and Statistics XLIV A Forerunner of the t -Distribution, *Biometrika*, **83**, 891-898.

Pizzetti: (1889) Sopra il calcolo dell'errore medio di un sistema di osservazioni, *Atti Reale Accademia dei Lincei, Series 4*: 5, 740-744.

_____ (1891) I fondamenti matematici per la critica dei risultati sperimentali. Atti della Università di Genova Genoa.-Università di Genova. Quarto Centenario Colombiano Biblioteca di 'Statistica'

Plackett, R. L. (1949) A Historical Note on the Method of Least Squares, *Biometrika*, **36**, No. 3/4, 458-460.

Rainsford, H. F. (1957) *Survey Adjustments and Least Squares*, London: Constable.

Reck, E. (2008) Dedekind's Contributions to the Foundations of Mathematics, *Stanford Encyclopedia of Philosophy*, E. Zalta, ed., 32 pp.

<http://plato.stanford.edu/entries/dedekind-foundations/>

Scheffé, H. (1956) Alternative Models for the Analysis of Variance, *Annals of Mathematical Statistics*, **27**, 251-271.

_____ (1959) *Analysis of Variance*, New York: Wiley.

Schultz, H. (1929) Applications of the Theory of Error to the Interpretation of Trends: Discussion, *Journal of the American Statistical Association*, **24**, Supplement. 86-89.

Seal, H. (1967) The Historical Development of the Gauss Linear Model, *Biometrika*, **54** 1-24.

Sheynin, O. (1979) Gauss and the Theory of Errors, *Archive for the History of Exact Sciences*, **20**, 21-72.

_____ (1995) Helmert's Work in the Theory of Errors, *Archive for History of Exact Sciences*, **49**, 73-104.

- _____ (1996) *The History of the Theory of Errors*, Egelsbach: Hänsel-Hohenhausen.
- Smart, W. H. (1958) *Combination of Observations*, Cambridge: Cambridge University Press.
- Soper, H. E. (1913) On the Probable Error of the Correlation Coefficient to a Second Approximation, *Biometrika*, **9**, 91-115.
- Soper, H. E., A. W. Young, B. M. Cave, A. Lee & K. Pearson (1917) On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of "Student" and R. A. Fisher, A Cooperative Study, *Biometrika*, **10**, 328-413.
- Stigler, S. M. (1978) Francis Ysidro Edgeworth, Statistician, (with discussion) *Journal of the Royal Statistical Society*, **141**, 287-322.
- _____ (1985) Mansfield Merriman in *Encyclopedia of Statistical Sciences* **5**, 437-8, New York: Wiley.
- _____ (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, MA: Belknap Press.
- _____ (2005) 1812, 1814 P. S. Laplace, *Théorie analytique des probabilités*, and *Essai philosophiques sur les probabilités*, ch. 24 and pp. 329-340 of I. Grattan-Guinness (ed.) *Landmark Writings in Western Mathematics: Case Studies, 1640-1940*, Amsterdam: Elsevier.
- _____ (2005) Fisher in 1921, *Statistical Science*, **20**, 32-49.
- _____ (2007) The Epic Story of Maximum Likelihood, *Statistical Science*, **22**, (4), 598-620.
- _____ (2008) Fisher and the 5% Level, *Chance*, **21**, (4), 12.

- Student (1908a) The Probable Error of a Mean, *Biometrika*, **6**, 1-25.
- _____ (1908b) Probable Error of a Correlation Coefficient, *Biometrika*, **6**, 302-310.
- _____ (1911) Note on A Method of Arranging Plots so as to Utilize a given Area of Land to the Best Advantage in Testing Two Varieties, Appendix to Mercer and Hall (1911) pp. 128-132.
- _____ (1914) The Elimination of Spurious Correlation due to Position in Time or Space, *Biometrika*, **10**, 179-180.
- _____ (1917) Tables for Estimating the Probability that the Mean of a Unique Sample of Observations Lies between $-\infty$ and any Given Distance of the Mean of the Population from which the Sample is Drawn, *Biometrika*, **17**, 414-417.
- _____ (1923) On Testing Varieties of Cereals, *Biometrika*, **15**, 271-293. Amendment and correction **16**, 1924: 411.
- _____ (1925) New Tables for Testing the Significance of Observations, *Metron*, **5**, 105-108.
- _____ (1926) Review of Statistical Methods for Research Workers (R. A. Fisher) *Eugenics Review*, **18**, 148-150. Reproduced with an Introduction by J. Aldrich on the website <http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/student.htm>
- _____ (1931) On the “z” Test, *Biometrika*, **23**, 407-408.
- Urquhart, N.S., D.L. Weeks and C.R. Henderson (1973) Estimation Associated with Linear Models: a Revisitation, *Communications in Statistics–Theory and Methods*, **1**, 303-330.
- Villegas, C. (1990) Bayesian Inference in Models with Euclidean Structures, *Journal of the*

American Statistical Association, **85**, 1159-1164.

Wall, J.V. and C. R. Jenkins (2003) *Practical Statistics for Astronomers*, Cambridge: Cambridge University Press.

Welch, B. L. (1958) ‘Student’ and Small Sample Theory, *Journal of the American Statistical Association*, **53**, 777–788.

Whittaker, E. & G. Robinson (1924) *Calculus of Observations*, Edinburgh, Blackie.

Wittstein T. (1849) Die Methode der kleinsten Quadrate, Appendix to Wittstein’s *Lehrbuch der Differential-und-Integralrechnung von Louis Navier*, Vol. II: 343-442. Hannover:

Wood, T. B. and F. J. M. Stratton (1910) The Interpretation of Experimental Results, *Journal of Agricultural Science*, **3**, 417-440.

Zabell, S. L. (2008) On Student’s 1908 paper “The probable error of a mean,” with comments by S. M. Stigler, J. Aldrich, A. W. F. Edwards, E. Seneta: Diaconis & E. L. Lehmann and rejoinder from Zabell, *Journal of the American Statistical Association*, **103**, 1–20.

Ziliak, S. T. and D. N. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor: University of Michigan Press.