On Student's 1908 Article "The Probable Error of a Mean"

S. L. ZABELL

This month marks the 100th anniversary of the appearance of William Sealey Gosset's celebrated article, "The Probable Error of a Mean" (Student 1908a). Gosset's elegant result represented the first in a series of exact, "small-sample" results that were developed by Gosset, Fisher, and others to form a central component of the modern theory of statistical inference. This review celebrates the centenary of Gosset's article by discussing both its background and its impact on statistical theory and practice.

KEY WORDS: Cushny and Peebles experiment; Egon Pearson; Fiducial inference; Friedrich Robert Helmert; Physical randomization; Robustness; Ronald Aylmer Fisher; "Student"; t-distribution; William Sealy Gosset.

1. BACKGROUND

William Sealy Gosset (born June 13, 1876, Canterbury, U.K.; died October 16, 1937, Beaconsfield, U.K.) studied at Winchester College and New College, Oxford, U.K. After being awarded a First in the Mathematical Moderations examination in 1897 and a First-Class Degree in Chemistry in 1899, Gossett was employed by Arthur Guinness, Son & Co., Ltd. in Dublin, Ireland in 1899. Gosset recognized almost immediately the importance of statistical methods in the company's operations, and the problem of small samples in particular. "The circumstances of brewing work, with its variable materials and susceptibility to temperature change and necessarily short series of experiments, are all such as to show up most rapidly the limitations of large sample theory and emphasize the necessity for a correct method of treating small samples. It was thus no accident, but the circumstances of his work, that directed 'Student's' attention to this problem" (McMullen 1939, pp. 205–206). At Guinness, Gosset found an environment that encouraged both practical and theoretical research, as Joan Fisher Box (R. A. Fisher's daughter), describes in her informative account (Box 1987, pp. 46–50).

Gosset's "attention" resulted in a report, "The Application of the 'Law of Error' to the work of the Brewery" dated November 3, 1904. (The report, an internal Guinness document, has never been published. Egon Pearson, while preparing his obituary of Gosset, was given "permission [by Guinness] to see and quote from [the report] and other records available in their Dublin brewery"; see Pearson 1939, p. 213.)

Unable to find the necessary results in the literature, Gosset subsequently arranged to meet Karl Pearson during the latter's vacation in July 1905. The meeting was an important one for Gosset; Guinness had an enlightened policy of permitting technical staff leave for study, and a year later Gosset spent the first two terms of the 1906–1907 academic year in Pearson's Biometric Laboratory at University College London.

Thus it was natural for Gosset to consider the question that he did, and equally natural for him to publish it in *Biometrika*, the journal that Pearson co-founded and edited.

2. "THE PROBABLE ERROR OF A MEAN"

Gosset's article has many interesting features, the first of which arises in its second line: its listing of the author.

2.1 "Student"?

In all, Gosset published 14 papers in *Biometrika* (and 7 others elsewhere) over a 25-year period. All but one appeared under the pseudonym "Student." Why "Student"? The solution to this mystery was likely first revealed in the *Journal of the American Statistical Association* more than three-quarters of a century ago by Harold Hotelling (1930, p. 189) in an article on British statistics:

> American students of statistics have long speculated as to the identity of "Student."... I have heard guesses in this country identifying "Student" with Egon S. Pearson and with the Prince of Wales. He is now so well known in Great Britain that no confidence is violated in revealing that he is W. S. Gosset, a research chemist employed by a large Dublin brewery [note that Guinness is not identified by name]. This concern years ago adopted a rule forbidding its chemists to publish their findings. Gosset pleaded that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, and finally was allowed to publish them, but under a pseudonym, to avoid difficulties with the rest of the staff.

Note that Hotelling's account is quite different than the oft-repeated claim that anonymity was demanded by Guinness so that Guinness's competitors would not realize that Guinness found it useful to employ a statistician. According to Hotelling, the anonymity was designed to hide Gosset's identity because of the brewers *inside* Guinness, not those outside!

The pseudonym "Student" was selected by Christopher Digges La Touche, the Managing Director of Guinness: "It was decided by La Touche that such publication might be made without the brewers' names appearing. They would be merely designated 'Pupil' or 'Student' " (Box 1987, p. 46).

Student (as we shall refer to him from now on) was not the only person to benefit from Guinness's enlightened attitude toward statistical education. Guinness sent Edward Somerfield to work with Fisher at Rothamsted in 1922 and George Story to work with Pearson at University College London in 1928. Both were permitted to publish papers in the outside literature, but once again only under pseudonyms: "Mathetes" (Somerfield) and "Sophister" (Story). This subsequent use of pseudonyms indeed may have been motivated by a desire to maintain a competitive edge. Guinness made extensive use of the t-test after its discovery by Student, but little use was made of it elsewhere (McMullen 1939, p. 206). By the 1920s, there was a large statistical group at Guinness, despite the fact that generally "very few scientists in industry made use of mathematical statistical methods [at that time]" (Pearson 1990, p. 91). So it is quite plausible, as Pearson suggests, that Guinness "probably" insisted on

S. L. Zabell is Professor, Departments of Mathematics and Statistics, Northwestern University, Evanston, IL 60208 (E-mail: *zabell@math.northwestern.edu*). The author thanks David Cox, Persi Diaconis, Steve Portnoy, Jim Reeds, Karen Reeds, and two anonymous referees for their helpful comments on preliminary versions of this manuscript.

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481
DOI 10.1198/016214508000000030

anonymous publication because it "did not wish it to be known by rival brewers that they were training some of their scientific staff in statistical theory and its application."

2.2 Statistical Philosophy

Student was initially a Bayesian in philosophical outlook, a view he is likely to have gotten from reading Karl Pearson, rather than textbooks on the theory of errors by Airy (1861) or Merriman (1884), Student's early view of the inferential process is very clearly stated in his correlation coefficient paper (Student 1908b), published in the same year as his t-distribution article:

> A random sample has been obtained from an indefinitely large population and r [the sample correlation coefficient] calculated between two variable characters of the individuals composing the sample. We require the probability that R for the population from which the sample is drawn shall lie between any given limits. It is clear that in order to solve this problem we must know two things: (1) the distribution of values of r derived from samples of a population which has a given R, and (2) the *à priori* probability that R for the population lies between any given limits.

But even at this stage, Student made clear his discomfort with the assignment of a prior, adding that:

> Now (2) can hardly ever be known, so that some arbitrary assumption must in general be made; when we know (1) it will be time enough to discuss what will be the best assumption to make, but meanwhile I may suggest two more or less obvious distributions. The first is that any value is equally likely between $+1$ and -1, and the second that the probability that x is the value is proportional to $1 - x^2$. This I think is more in accordance with ordinary experience; the distribution of *à priori* distribution would then be expressed by the equation $y = \frac{3}{4}(1 - x^2)$.

In 1915 and 1917, one can still find Student—in correspondence with Fisher and Pearson—suggesting the use of the priors mentioned in his 1908 article on correlation (see Pearson 1990, pp. 25, 26, 28). But these appear to have been merely casual suggestions; in his letter to Pearson, Student also noted the "disadvantage of using actual knowledge concerning similar work," and in a later letter to Fisher on April 3, 1922, put the case even more strongly:

> When I was in [Pearson's] lab in 1907 I tried to work out variants of Bayes with a priori probabilities other than [a flat one] but I soon convinced myself that with ordinary sized samples one's a priori hypothesis made a fool of the actual sample. . . and since then have refused to use any other hypothesis than the one which leads to your likelihood [i.e., the flat prior] (where I could deal with the mathematics). Then each piece of evidence can be considered on its own merits (Pearson 1990, pp. 28–29).

Despite this eventual distrust of nonuniform priors (in contrast to Karl Pearson), in his article on the t-distribution, Student's view of the inferential process is clearly Bayesian. He refers to the "usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample" (p. 1), and throughout his examples refers to the odds that one hypothesis holds rather than another: "The chance that the mean of the population of which these experiments are a sample is positive" (p. 20); "the odds are about 666 to 1 that [one of two drugs] is the better soporific" (p. 21); "the odds are about 14 : 1 that kiln-dried corn gives the higher yield" (p. 24).

One particularly interesting remark is Student's statement that "if two observations have been made and we have no other information, it is an even chance that the mean of the (normal) population will lie between them" (Student 1908a, p. 13). This is in effect an interval estimate; Fisher considered it an early example of a fiducial probability (see Fisher 1939, p. 4; Welch 1958, pp. 782–784).

2.3 Derivation of the t-Distribution

Let us turn to the technical content of Gosset's article. Gosset did not give a rigorous mathematical demonstration of his result. Important elements in the proof already existed (buried in the German literature), but the specific question he asked and answered was both new and novel, and his article inspired much later work by demonstrating the feasibility of adjustment for small sample sizes.

Let $X_1, \ldots, X_n$ denote a sequence of independent and identically distributed random variables, $\mu = EX_j$, $\sigma^2 = \operatorname{var} X_j$, and $X_j \sim \mathrm{N}(\mu, \sigma^2)$. Let $\bar{X}$ and S^2 denote the sample mean and sample variance of the X_j. Then the statistic

$$t := \sqrt{n}\frac{\bar{X} - \mu}{S}$$

has *Student's t-distribution on* $n - 1$ *degrees of freedom*; that is, it has the density function

$$\frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})}\left(1 + \frac{x^2}{n-1}\right)^{-n/2}.$$

The usual derivation of the t-distribution proceeds by showing in some order that

- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- $\bar{X}$ and S^2 are independent
- $t \sim t_{n-1}$.

The first of these was well known at the time that Student wrote and could be found in then-standard textbooks on the theory of errors (e.g., Airy 1861, part II, sec. 6, article 43, p. 33; Czuber 1891, pp. 145–147). The first, second, and third sections of Student's article are devoted to deriving successively the last three properties.

Note that, strictly speaking, Student considered the quantity

$$z = (\bar{X} - \mu)\Big/\sqrt{\frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{n}}.$$

The transformation $t = z\sqrt{n-1}$ was introduced by Fisher (1925) and later adopted by Student himself (see Eisenhart 1979).

2.3.1 Distribution of S^2. At the time that Student wrote, the distribution of S^2 was known, but not well known (at least in England); it had been derived some three decades earlier in 1876 by the German geodetic scientist Friedrich Robert Helmert (1843–1917). Helmert had a distinguished career, eventually becoming a professor at the University of Berlin and heading its Geodetic Institute (see Fischer 1973). Helmert, following directly in the tradition of Gauss, had earlier published a textbook on the theory of errors (Helmert 1872); thus his interest in the distribution of S^2 was not unnatural. Helmert's textbook became a classic in the field, remaining in print in second (1907) and third (1924) editions until well into the twentieth century. (For detailed discussion of Helmert's derivation, see Hald 1998, pp. 633–637; 2007, pp. 149–152; Sheynin 1995.)

So Student could have just cited Helmert's article for the distribution of S^2. Being unaware of it, however, he instead proceeded to compute the first four moments of S^2 and, noting that,

suitably scaled, these agreed with those of chi-squared (a member of the four-parameter Pearson family), correctly conjectured that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. To quote Student (p. 5), "hence it is probable that the curve found represents the theoretical distribution of s^2; so that although we have no actual proof, we shall assume it to do so in what follows."

2.3.2 Joint Distribution of $\bar{X}$ and S^2. Let x_j denote the observed value of X_j. Helmert's proof begins by making the change of variable $e_j = x_j - \bar{x}$ and computing the joint density

$$f(e_1, \ldots, e_{n-1}, \bar{x}) = \frac{n}{\sqrt{2\pi\sigma^2}^n} \exp\left\{-\frac{\sum_{j=1}^n e_j^2 + n\bar{x}^2}{2\sigma^2}\right\}$$

(see Uspensky 1937, pp. 333–336 for a proof). This in turn immediately implies the independence of $\bar{X}$ and S^2. Lacking this fact, Student instead showed that $\bar{X}$ and S^2 were uncorrelated and then proceeded as if they were in fact independent. Fisher (1939, pp. 3–4) considered this the most serious gap in Student's argument:

A second theoretical point was treated even more brusquely. The sampling distribution of s is in reality completely independent of that of $\bar{x}$. In order to derive the distribution of the significance test

$$t = \frac{\bar{x} - \mu}{s\sqrt{n}},$$

where μ is the true mean of the population sampled, it was necessary that this independence should be established. "Student" perceived this necessity, and any capable analyst could have shown him the demonstration he needed. As it was, he satisfied himself with showing somewhat laboriously that the distribution of s^2 was uncorrelated with both $\bar{x}$ and with $\bar{x}^2$. This was the most striking gap in his argument, for in truth it was not merely the distribution of s found by Helmert, but the exact simultaneous distribution of s and $\bar{x}$ that "Student" needed to develop his test.

Suppose that the summands X_j have any distribution with a third moment. If $\mu_3 = E[(X_j - \mu)^3]$, then

$$\operatorname{cov}(\bar{X}, S^2) = \frac{\mu_3}{n}$$

(see, e.g., Zhang 2007, pp. 159–160). This simple equation sheds light on Student's observation; the fact that $\bar{X}$ and S^2 are uncorrelated simply reflects the absence of skewness in the case of the normal, and indeed holds whenever μ_3 vanishes. The *independence* of $\bar{X}$ and S^2, in contrast, is much more restrictive; it is both a necessary *and* sufficient condition for normality, as first noted by Geary (1936) (also see Lukacs 1942).

Deriving the joint distribution of $\bar{X}$ and S^2 has had a recurrent fascination for statisticians since the work of Fisher (1915). Kruskal (1946), for example, found an interesting inductive proof inspired by an earlier computation of Helmert; this in turn inspired Stigler (1984) to produce an elegant reformulation of Kruskal's proof. (For discussion of Stigler's proof, see Zehna 1991 and the ensuing debate in the "Letters" section of *The American Statistician*, 1992, vol. 46, pp. 70–77.)

2.3.3 Distribution of t. Given these results, it was a straightforward exercise in integration for Student to derive the distribution of t. Perversely, here too he had been (partly) anticipated, by Jakob Lüroth (1876) and Francis Ysidro Edgeworth (1883), both of whom had derived the t-distribution as a posterior distribution for the population mean using a flat prior, Lüroth for the more general case of linear regression and Edgeworth in the one-sample case considered by Student (Welch 1958, p. 779).

It is not entirely surprising that Student (as well as *Biometrika* and its readership) were unfamiliar with Helmert's work. In general, there was a considerable gulf between continental and English statistics both at the time that Student wrote and for some time after. The reasons for and extent of this divide could easily be the subject of a lengthy article in itself.

Karl Pearson (the editor of *Biometrika*) eventually did note (in 1931!) Helmert's priority in an editorial in *Biometrika* (Pearson 1931, p. 416):

The familiar distribution of the standard deviations of samples from an indefinitely large normal population appears to have been discovered on several occasions—without doubt independently—and by later writers attributed to various investigators. There can, we think, be no doubt that the original discoverer was Helmert and that he much antidates other claimants.

What was the reason for such persistent neglect? Pearson could give no good explanation, noting that Helmert's article had appeared "in a journal which would be rarely consulted by statisticians," but conceding at the same time that "Helmert's result has been reproduced in German textbooks on the theory of observations" (e.g., Czuber 1891). Pearson suggested that the formula be called "Helmert's equation."

2.4 "Practical Test of the Foregoing Equation"

One of the most interesting aspects of Student's article is his use of simulation to investigate the small-sample behavior of z $(= t/\sqrt{n-1})$. He noted the following:

Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3,000 criminals, from a paper by W. R. Macdonell (*Biometrika*, vol. I, p. 219). The measurements were written out on 3,000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3,000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation of each sample determined. The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample, giving us the z of Section III.

It has sometimes been suggested that this represents the earliest instance of the use of simulation in statistics (see, e.g., Pearson 1939, p. 223; Teichrow 1965); but Stigler (1991) pointed to several earlier examples in the nineteenth century (apart from Buffon's classic needle experiment in the eighteenth). Nevertheless, it is certain that Student's simulation is one of a small number performed before the appearance of L. H. C. Tippett's table of random numbers in 1927. After Tippett (1925, p. 378) and Church (1926, pp. 324–325, 332–333) had encountered difficulties in carrying out random sampling experiments for simulation studies reported in *Biometrika*, Karl Pearson suggested that Tippett prepare a table of random numbers to facilitate such experiments. The result was Tippett's *Tracts for Computers* no. XV, published in 1927 (see Pearson 1990, pp. 88–95). Note that the term "computer" in Tippett's title refers to *human* computers, a usage that continued through World War II.

It is a pity Student does not expand on how his cards were "very thoroughly shuffled." Although his results appear to have been satisfactory, true physical randomization is often very difficult to achieve, and in the hands of the inexperienced often is not achieved. The 1970 draft lottery debacle is a celebrated instance (see Fienberg 1971). Rowlett (1998, pp. 53–54) provided an amusing description of some of the difficulties that U.S. codemakers encountered during the 1930s when they tried to mix thousands of cards for a codebook.

2.5 "Illustrations of Method"

Technometrics states in its "Information for Authors" that "every article shall include adequate justification of the application of the technique, preferably by means of an actual application to a problem in the physical, chemical, or engineering sciences." Student, given his strong practical bent, needed no such directive; his article contains no fewer than four applications to real data. (It should be noted that all of these involve one-sample t-tests.)

The first of Student's four examples concerned a famous experiment investigating whether different optical isomers might have different sleep-inducing effects. Student (1908a, p. 20) noted that:

> As an instance of the kind of use which may be made of the tables, I take the following figures from a table by A. R. Cushny and A. R. Peebles in the *Journal of Physiology* for 1904 [sic], showing the different effects of the optical isomers of hyoscyamine hydrobromide in producing sleep. The sleep of 10 patients was measured without hypnotic and after treatment (1) with D. hyoscyamine hydrobromide, (2) with L. hyoscyamine hydrobromide. The average number of hours' sleep gained by the use of the drug is tabulated below.

This is a celebrated data set (Cushny and Peebles 1905) in part because Fisher used Student's tabular presentation of the data as an example in his classic textbook *Statistical Methods for Research Workers* (1925; hereinafter referred to as SMRW, sec. 24). Student stated that the drugs in question were the levo and dextro forms of hyoscyamine hydrobromide, a statement accepted by Fisher in the earlier editions of his book. However, Dr. Isidor Greenwald of the New York University and Bellevue Hospital Medical College pointed out to Fisher in a letter (dated December 12, 1934) that the two drugs corresponding to the data in question were L-hyoscyamine hydrobromate and L-hyoscine hydrobromate—not optical isomers at all! Fisher—who disliked major surgery in his books—dealt with this unpleasantness by replacing the names of the drugs in the two columns by "A" and "B" in later editions of SMRW! The episode is a useful cautionary tale in the use of secondary sources for information and data, especially when the original data are readily available. (The correspondence was given in full in McMullen 1970, and extracts were quoted in Pearson 1990, p. 54.)

But the problems with Student's (and Fisher's) analysis extend far beyond mere mislabeling. The "A" and "B" measurements are averages based on samples varying in size from 3 to 6, so the measurements have no σ in common for S to estimate! (For detailed discussion of both the original Cushny and Peebles experiment and data and Student's analysis of it, see Preece 1982; Senn and Richardson 1994; Senn 2002, pp. 1–7.)

2.6 Robustness of the t-Distribution

Of course, no real distribution is exactly normal, and so the practical applicability of the t-distribution must depend on its relative *robustness* to departures from normality. What is needed, furthermore, is not merely an asymptotic result, but some connection to our everyday experience. Some forms of limiting behavior can take a very long time to set in. The central limit theorem vindicates its name in part because normality sets in both early and often.

Student was sensitive to this issue. In his Introduction, he concedes that his conclusions "are not strictly applicable to populations known not to be normally distributed," but asserts that "it appears probable that the deviation from normality must be very extreme to lead to serious error." This claim was later supported by empirical studies in Section VI (p. 19): "I believe that the tables at the end of the present paper may be used in estimating the degree of certainty arrived at by the mean of a few experiments, in the case of most laboratory or biological work where the distributions are as a rule of a 'cocked hat' type and so sufficiently nearly normal."

The (surprising?) robustness of the t has been perhaps one of its greatest practical features to emerge in the years since 1908. There have been many studies of the distribution of the t-statistic under departures from normality (one example in *JASA* is Efron 1969). It is interesting that one natural question, when the t-statistic is asymptotically normal as $n \to \infty$, remained open for nearly a quarter of century after its answer was suggested by Logan, Mallows, Rice, and Shepp (1973). Giné, Götze, and Mason (1996) were ultimately able to prove that the necessary and sufficient condition on the summands X_j of the t-statistic to ensure convergence to the normal is that $EX_j = 0$ and X_j lie in the domain of attraction of the normal distribution. Such results appear in the literature as part of the theory of *self-normalized* (or "studentized") sums.

By any measure (citation, theoretical extension, practical use), Student's two 1908 articles initially attracted surprisingly little notice. Egon Pearson (1939, p. 225), who was certainly in a position to judge, observed that "one of the curious things that must strike us now about these two papers of Gosset's... is the small influence that their publication had for a number of years on current statistical literature and practice." But eventually Student "acquired a new champion of exceptional brilliance and enormous energy" (Lehmann 1999, p. 419).

3. RONALD AYLMER FISHER

When Isidor Greenwald wrote Fisher in 1934 about the Cushny–Peebles gaffe in SMRW, he concluded by remarking "I am greatly surprised that this error should not have been corrected long ago" (Pearson 1990, p. 54). When Student wrote back to Fisher a week later acknowledging the error, he remarked in a postscript that "of course it is not surprising that no one discovered the blunder for in the pre Fisher days no one paid the slightest attention to the paper." The history of Student and the history of Fisher are inextricably linked. Fisher not only championed Student's work, but Student exerted a profound influence on the nature and direction of Fisher's research for nearly two decades.

3.1 Fisher Meets Student

Fisher's supervisor at Gonville and Caius (Fisher's college at Cambridge), the astronomer Frederick J. M. Stratton, had initially put Fisher in touch with Student, suggesting he send Student a copy of his first paper (Fisher 1912). Mathematicians, especially young and ambitious ones fresh out of school, are always looking out for crisply defined problems on which they can bring their prowess to bear. R. A. Fisher was no exception. Providing a rigorous mathematical derivation of the t-distribution was a natural challenge, and in 1912 Fisher wrote to Student, giving a complete and rigorous derivation of the t-distribution, exploiting his facility in n-dimensional geometry. Student forwarded the proof to Pearson, suggesting that

the result might be the subject of a note in *Biometrika*, but Pearson demurred, responding that he was unable to follow Fisher's proof (Pearson 1968, pp. 446–447; 1990, pp. 47–48). Fisher likely never learned this; in his obituary notice of Student, Fisher (1939, p. 5) stated that "in his correspondence with me [Student] did not even suggest that the completed proofs should be published."

This was merely a temporary setback. Fisher was later able to find a geometric proof for the distribution of the correlation coefficient as well, and the resulting article published in *Biometrika* was his first important contribution to statistics. Fisher (1915, p. 507) wrote, referring to Student's derivation of the t:

> This result, although arrived at by empirical methods, was established almost beyond reasonable doubt in the first of "Student's" papers. It is, however, of interest to notice that the form establishes itself instantly, when the distribution of the sample is viewed geometrically.

The formal proof itself was given later by Fisher (1923).

3.2 Ubiquity of the t

This was only the tip of the iceberg. On April 3, 1922, Student wrote to Fisher asking about the distribution of a regression coefficient, and followed this up on April 9 with a similar question about partial correlation and regression coefficients. Fisher replied by early May that the t-distribution supplied the answer to both of Student's questions as well as the case of the difference of two means (McMullen 1970, letters 5–7; Pearson 1990, pp. 48–49; Box 1981, pp. 62–63). These results were reported in part by Fisher (1922), summarized in his article "Applications of 'Student's' Distribution" (Fisher 1925), formed the basis of his address to the International Congress of Mathematicians at Toronto in 1924 (Fisher 1924), and set out for practical use in *Statistical Methods for Research Workers*.

This explosion in the potential uses of the t-distribution led to new tables, published jointly by Fisher and Student in *Metron*, based in part on new expansions derived by Fisher (see Eisenhart 1979; Box 1981, pp. 63–64). It was precisely at this point that the passage from Student's initial use of "z" to Fisher's "t" took place.

3.3 Robustness Revisited

Student, as we have seen, was certainly alert to the issue of the sensitivity of his approach to possible departures from normality. Clearly, such concerns would apply with equal force to Fisher's methods, vigorously championed in SMRW.

3.3.1 Egon Pearson's Review of SMRW. When the second edition of SMRW appeared in 1928, Egon Pearson raised this issue in a review in *Nature*, angering Fisher. Pearson (1929, pp. 866–867) wrote:

> There is one criticism, however, which must be made from the statistical point of view. A large number of tests are developed upon the assumption that the population sampled is of normal form. That this is the case may be gathered from a very careful reading of the text, but the point is not sufficiently emphasised. It does not appear reasonable to lay stress on the exactness of tests, when no means whatever are given of appreciating how rapidly they become inexact as the population samples diverges from normality. That the tests, for example, connected with the analysis of variance are far more dependent on normality than those involving "Student's" z (or t) distribution is almost certain, but no clear indication of the need for caution in their application is given to the worker. It would seem wiser in the long run, even in a text-book, to admit the incompleteness of the theory in this direction, rather than risk giving the reader the impression that the solution of all his problems have been achieved. The author's contributions to the development of normal theory will stand by themselves, both for their direct practical values and as an important preliminary to the wider extension of theory, without any suggestion of undue completeness.

Ironically, Pearson's concerns had been strongly influenced by Student. Student had written to Pearson on May 11, 1926, raising the issue of the robustness of the rapidly burgeoning number of tests based on normal distribution theory. The existence of Tippett's tables made possible a systematic study of the issue, a study that Pearson proceeded to carry out between 1928 and 1931 (Pearson 1990, sec. 6.4).

Student acted as a mediator between Fisher and Pearson after Pearson's (unsigned) review appeared in *Nature* on June 8, 1939; in the end, *Nature* published letters responding to the review by both Student (on July 20) and Fisher (on August 3) (see Pearson 1990, sec. 6.5). One comment from Fisher is of particular interest because of the reaction that it elicited from Student. Fisher wrote:

> I have never known difficulty to arise in biological work from imperfect normality of the variation, often though I have examined data for this particular cause of difficulty; nor is there, I believe, any case to the contrary in the literature.

Student made the following comment to Pearson in a letter on September 25 (Pearson 1990, p. 99):

> Fisher is only talking through his hat when he talks of his experience; it isn't so very extensive and I bet he hasn't often put the matter to the test; how could he?

How ironic coming from the person who in so many ways inspired Fisher's development of statistical methods based on the normal distribution!

3.3.2 Fisher's Randomization Test. The third chapter of Fisher's classic book *The Design of Experiments* (Fisher 1935a) is devoted to an analysis of Charles Darwin's *Zea mays* data. In brief, these data comprise the heights of 15 pairs of plants, one member of the pair being a cross-fertilized plant, the other a self-fertilized plant. In Section 17 ("Student's" t test) the data are summarized in Table 3 as 15 differences in eighths of an inch, and used to illustrate use of the t-test.

Later on, in Section 21, noting that some statisticians had stressed the assumption of normality in the hypothesis being tested, Fisher advanced an early example of a nonparametric test:

> [T]he physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, afford the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.

Suppose that there are n paired differences (here $n = 15$). Let x_i be the height of a cross-fertilized plant, y_i be the height of a self-fertilized plant, $d_i = |x_i - y_i|$, and

$$S_n = \epsilon_1 d_1 + \cdots + \epsilon_n d_n$$

for some choice of $\epsilon_i = \pm 1$. Fisher's (typically clever) idea was to view the observed value of S_n as one of 2^n possible such values, and to see whether the observed value was unusual compared with the other possible values. Formally, this can be considered a test of the null that the observed value S_n has been randomly selected from the set of all possible sums.

In the case of Darwin's data, $n = 15$, $2^n = 32{,}728$, the observed value of S_n is 314, and there are 1,726 possible values of S_n such that $|S_n| > 314$; thus the level of significance for the randomization test is $1{,}726/32{,}728 = .05267$. In contrast, the level of significance computed based on the t-test is .0497. The p value for the t-test turns out to be an excellent approximation to the p value for the randomization test! Diaconis and Holmes

(1994) explored the use of Gray codes to expedite the computation of randomization distributions and discuss Fisher's *Zea mays* example as an illustration; their figure 2.1 shows that the randomization distribution for this set of data is bell-shaped (indeed remarkably so), and their figure 2.2 shows that both the normal and t provide excellent approximations to the randomization distribution.

3.4 Fiducial Inference

The t-distribution has served as a touchstone for fiducial inference throughout its history. Fisher, in *The Design of Experiments* (1935a, sec. 62), credited the biologist E. J. Maskell (who worked at Rothamsted in the early 1920s) with the observation that one could use the percentiles of the t-distribution to give interval estimates for μ corresponding to any desired level of significance. Such use eventually morphed into Fisher's concept of a *fiducial distribution*, although Maskell's precise role in its birth is a matter of dispute (compare Edwards 1995, pp. 800–801 and Aldrich 2000, p. 158). In any event, Fisher regarded the use of the t a pivotal quantity to be a basic example of the fiducial argument (see Fisher 1935b, pp. 391–394).

Fisher revisited this use of the t in his final book, *Statistical Methods and Scientific Inference* (Fisher 1956, chap. 4, sec. 2). In the intervening two decades, however, his view of the fiducial argument had radically changed; his defense of fiducial inference now centered around the idea of a *recognizable subset*. In the case of the t, for example, Fisher claimed that in considering the inequality

$$\mu < \bar{x} - \frac{1}{\sqrt{N}} t s,$$

"there is no possibility of recognizing any sub-set of cases, within the general set, for which any different value of the probability should hold" (ibid., p. 84). No mathematical proof of this assertion was given, however, and Buehler and Feddersen (1963)—perhaps somewhat surprisingly—were able to show that recognizable subsets do in fact exist in this case (see also Brown 1967). (For subsequent defences of the fiducial argument addressing this issue, see Yates 1964 and Barnard 1995.)

Jeffreys (1931, 1937, 1939) discussed Bayesian derivation of the t as a posterior distribution for the mean, and explored the reasons why Fisher's derivation of the fiducial distribution of the mean comes to the same thing. Note that Edgeworth (1883), Burnside (1923), and Jeffreys arrived at t-posteriors with different degrees of freedom because they each used (in effect) different priors for σ.

4. SOME LITERATURE

An enormous amount has been written about Student, his work, and his 1908 article. The following is intended as a brief tour through this thicket.

Of the appreciations that came out at the time of Student's death, those of Fisher (1939), McMullen (1939), and Pearson (1939) are of particular interest; the last of these gave a detailed and lengthy overview of the totality of Student's work. Egon Pearson's statistical biography of Student (Pearson 1990), edited by Plackett and Barnard after Pearson's death in 1980, contains a wealth of further information and quotes extensively from correspondence Student had with both Karl and Egon Pearson. Fisher's biography by his daughter Joan Fisher Box (1978) is also useful.

Student's *Collected Papers* (Pearson and Wishart 1942) has long been out of print, but most of his important papers appeared in *Biometrika*, and thus are available online through *JSTOR*. Student's correspondence with Fisher (McMullen 1970) has, unfortunately, appeared only in privately published form. Excerpts from and discussion of the correspondence have been given by Box (1981) and Pearson (1990, chap. 5).

For a modern technical discussion of Student's work, one can consult Hald's comprehensive and meticuluously documented historical studies (Hald 1998, 2007). Welch (1958) provided an interesting view at mid-century of Student's work, published in *JASA* to mark the 50th anniversary of Student's article. It will be interesting to see what *JASA* publishes 50 years from now!

5. CONCLUSION

Student's article is truly remarkable for its richness. It simultaneously heralded the advent of small-sample distributional studies in statistics, used simulation in a serious way to investigate such distributions, and investigated the robustness of its results against modest departures from normality. It was a singular accomplishment, especially so given the limited formal background in mathematics and statistics that Student had. But it took the genius and drive of a Fisher to give Student's work general currency. The contribution of Fisher to our profession is widely appreciated; the extent of Student's contribution perhaps not as much.

[Received October 2007. Revised November 2007.]

REFERENCES

Airy, G. B. (1861), *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*, Cambridge and London: Macmillan and Co.

Aldrich, J. (2000), Fisher's "'Inverse Probability' of 1930," *International Statistical Review*, 68, 155–172.

Barnard, G. A. (1995), "Pivotal Models and the Fiducial Argument," *The American Statistician*, 63, 309–323.

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

——— (1981), "Gosset, Fisher, and the t Distribution," *The American Statistician*, 35, 61–66.

——— (1987), "Guinness, Gosset, Fisher, and Small Samples," *Statistical Science*, 2, 45–52.

Brown, L. (1967), "The Conditional Level of Student's t Test," *The Annals of Mathematical Statistics*, 38, 1068–1071.

Buehler, R. J., and Feddersen, A. P. (1963), "Note on a Conditional Property of Student's t," *The Annals of Mathematical Statistics*, 34, 1098–1100.

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

Church, A. E. R. (1926), "On the Means and Squared Standard Deviations of Small Samples From Any Population," *Biometrika*, 18, 321–394.

Cushny, A. R., and Peebles, A. R. (1905), "The Action of Optical Isomers. II: Hyoscines," *Journal of Physiology*, 32, 501–510.

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

Diaconis, P., and Holmes, S. (1994), "Gray Codes for Randomization Procedures," *Statistics and Computing*, 4, 287–302.

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

Edwards, A. W. F. (1995), "Fiducial Inference and the Fundamental Theorem of Natural Selection," *Biometrics*, 51, 799–809.

Efron, B. (1969), "Student's t-Test Under Symmetry Conditions," *Journal of the American Statistical Association*, 64, 1278–1302.

Eisenhart, C. (1979), "On the Transition From 'Student's' z to 'Student's' t," *The American Statistician*, 33, 6–10.

Fienberg, S. E. (1971), "Randomization and Social Affairs: The 1970 Draft Lottery," *Science*, 171, 255–261.

Fischer, W. (1973), "Helmert, Friedrich Robert," in *Dictionary of Scientific Biography*, Vol. 7, New York: Scribners, pp. 239–241.
Fisher, R. A. (1912), "On an Absolute Criterion for Fitting Frequency Curves," *Messenger of Mathematics*, 41, 507–521.
——— (1915), "Frequency Distribution of the Values of the Correlation Coefficient in Samples From an Indefinitely Large Population," *Biometrika*, 10, 507–521.
——— (1922), "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), "On a Distribution Yielding the Error Functions of Several Well Known Statistics," in *Proceedings of the International Congress of Mathematics, Toronto*, Vol. 2, pp. 805–813.
——— (1925), *Statistical Methods for Research Workers*, Edinburgh and London: Oliver & Boyd.
——— (1935a), *The Design of Experiments*, Edinburgh and London: Oliver & Boyd.
——— (1935b), "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, 6, 391–398.
——— (1939), "Student," *Annals of Eugenics*, 9, 1–9.
——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh and London: Oliver & Boyd.
Geary, R. C. (1936), "The Distribution of "Students" Ratio for Non-Normal Samples," *Supplement to the Journal of the Royal Statistical Society*, 3, 178–184.
Giné, E., Götze, F., and Mason, D. M. (1996), "When Is the Student t-Statistic Asymptotically Standard Normal?" *The Annals of Probability*, 25, 1514–1531.
Hald, A. (1998), *A History of Mathematical Statistics From 1750 to 1930*, New York: Wiley.
——— (2007), *A History of Parametric Statistical Inference From Bernoulli to Fisher, 1713–1935*, New York: Springer-Verlag.
Helmert, F. R. (1872), *Die Ausgleichungsrechnung nach der Methode der kleinsten Quadrate*, Leipzig: Teubner.
——— (1876), "Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit im Zusammenhange stehende Fragen," *Zeitschrift für Mathematik und Physik*, 21, 192–218.
Hotelling, H. (1930), "British Statistics and Statisticians Today," *Journal of the American Statistical Association*, 25, 186–190.
Jeffreys, H. (1931), *Scientific Inference*, Cambridge, U.K.: Cambridge University Press.
——— (1937), "On the Relation Between Direct and Inverse Methods in Statistics," *Proceedings of the Royal Society of London*, Ser. A, 60, 325–348.
——— (1939), *Theory of Probability*, Oxford, U.K.: Clarendon Press.
Kruskal, W. (1946), "Helmert's Distribution," *The American Mathematical Monthly*, 53, 435–438.
Lehmann, E. L. (1999), "'Student' and Small-Sample Theory," *Statistical Science*, 14, 418–426.
Logan, B. F., Mallows, C. L., Rice, S. D., and Shepp, L. A. (1973), "Limit Distributions of Self-Normalized Sums," *The Annals of Probability*, 1, 788–809.
Lukacs, E. (1942), "A Characterization of the Normal Distribution," *The Annals of Mathematical Statistics*, 13, 91–93.
Lüroth, J. (1876), "Vergleichung von zwei Werten des wahrscheinlichen Fehlers," *Astron. Nachr.*, 87, 209–220.
McMullen, L. (1939), "'Student' as a Man," *Biometrika*, 30, 205–210.
——— (ed.) (1970), *Letters From W. S. Gosset to R. A. Fisher*, privately published. (Original manuscript letters preserved in the University College London Archive; reference code GB 0103 MS ADD 274.)
Merriman, M. (ed.) (1884), *A Textbook on the Method of Least Squares*, New York: Wiley.
Pearson, E. (1929), Review of *R. A. Fisher: Statistical Methods for Research Workers* (2nd ed.), *Nature*, 30, 866–867.
——— (1939), "'Student' as a Statistician," *Biometrika*, 30, 210–250.
——— (1968), "Studies in the History of Probability and Statistics, XX: Some Early Correspondence Between W. S. Gosset, R. A. Fisher, and Karl Pearson, With Notes and Comments," *Biometrika*, 55, 445–457.
——— (1990), *'Student': A Statistical Biography of William Sealy Gosset*, Oxford, U.K.: Clarendon Press.
Pearson, E. S., and Wishart, J. (eds.) (1942), *"Student's" Collected Papers*, London: Biometrika, University College London.
Pearson, K. (1931), "Historical Note on the Distribution of the Standard Deviations of Samples of Any Size Drawn From an Indefinitely Large Normal Parent Population," *Biometrika*, 23, 416–418.
Pfanzagl, J., and Sheynin, O. (1996), "Studies in the History of Probability and Statistics, XLIV: A Forerunner of the t-Distribution," *Biometrika*, 83, 891–898.
Preece, D. A. (1982), "t Is for Trouble (and Textbooks): A Critique of Some Examples of the Paired-Samples t-Test," *Statistician*, 31, 169–195.
Rowlett, F. (1998), *The Story of Magic: Memoirs of an American Cryptologic Pioneer*, Walnut Creek, CA: Aegean Park Press.
Senn, S. (2002), *Cross-Over Trials in Clinical Research* (2nd ed.), New York: Wiley.
Senn, S., and Richardson, W. (1994), "The First t-Test," *Statistics in Medicine*, 13, 785–803.
Sheynin, O. B. (1995), "Helmert's Work in the Theory of Errors," *Archive for History of Exact Sciences*, 49, 73–104.
Stigler, S. M. (1984), "Kruskal's Proof of the Joint Distribution of $\overline{X}$ and s^2," *The American Statistician*, 38, 134–135.
——— (1991), "Stochastic Simulation in the Nineteenth Century," *Statistical Science*, 6, 89–97.
Student (1908a), "The Probable Error of a Mean," *Biometrika*, 6, 1–25.
——— (1908b), "Probable Error of a Correlation Coefficient," *Biometrika*, 6, 302–310.
Teichroew, D. (1965), "A History of Distribution Sampling Prior to the Era of the Computer and Its Relevance to Simulation," *Journal of the American Statistical Association*, 60, 27–49.
Tippett, L. H. C. (1925), "On the Extreme Individuals and the Range of Samples Taken From a Normal Population," *Biometrika*, 17, 364–387.
——— (1927), *Random Sampling Numbers*, Cambridge, U.K.: Cambridge University Press.
Uspensky, J. V. (1937), *Introduction to Mathematical Probability*, New York: McGraw-Hill.
Welch, B. L. (1958), "'Student' and Small Sample Theory," *Journal of the American Statistical Association*, 53, 777–788.
Yates, F. (1964), "Fiducial Probability, Recognisable Sub-Sets and Behrens' Test," *Biometrics*, 20, 343–360.
Zabell, S. L. (1992), "R. A. Fisher and Fiducial Argument," *Statistical Science*, 7, 369–387.
Zehna, P. W. (1991), "On Proving That $\overline{X}$ and s^2 Are Independent," *The American Statistician*, 45, 121–122.
Zhang, L. (2007), "Sample Mean and Sample Variance: Their Covariance and Their (In)dependence," *The American Statistician*, 61, 159–160.

Comment

Stephen M. STIGLER

Sandy Zabell's excellent account of Student's 1908 article is exemplary in all respects. I will add only a brief comment on the question of the role of Gosset in the historical development of statistics.

Stephen M. Stigler is the Ernest DeWitt Burton Distinguished Service Professor of Statistics, Department of Statistics, University of Chicago, Chicago, IL 60637 (E-mail: *stigler@uchicago.edu*).

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481
DOI 10.1198/016214508000000049