Journal of Wine Economics, Volume 6, Number 2, 2011, Pages 252–277

W.S. Gosset and Some Neglected Concepts in Experimental Statistics: Guinnessometrics II*

Stephen T. Ziliak[a]

Error, Sir, creeps in thro' the minute holes, and small crevices, which human nature leaves unguarded.

Laurence Sterne, *Tristram Shandy*

Abstract

Student's exacting theory of errors, both random and real, marked a significant advance over ambiguous reports of plant life and fermentation asserted by chemists from Priestley and Lavoisier down to Pasteur and Johannsen, working at the Carlsberg Laboratory. One reason seems to be that William Sealy Gosset (1876–1937) aka "Student" – he of Student's *t*-table and test of statistical significance – rejected artificial rules about sample size, experimental design, and the level of significance, and took instead an economic approach to the logic of decisions made under uncertainty. In his job as Apprentice Brewer, Head Experimental Brewer, and finally Head Brewer of Guinness, Student produced small samples of experimental barley, malt, and hops, seeking guidance for industrial quality control and maximum expected profit at the large scale brewery. In the process Student invented or inspired half of modern statistics. This article draws on original archival evidence, shedding light on several core yet neglected aspects of Student's methods, that is, Guinnessometrics, not discussed by Ronald A. Fisher (1890–1962). The focus is on Student's small sample, economic approach to real error minimization, particularly in field and laboratory experiments he conducted on barley and malt, 1904 to 1937. Balanced designs of experiments, he found, are more efficient than random and have higher power to detect large and real treatment differences in a series of repeated and independent experiments. Student's world-class achievement poses a challenge to every science. Should statistical methods – such as the choice of sample size, experimental design, and level of significance – follow the purpose of the experiment, rather than the other way around? (JEL classification codes: C10, C90, C93, L66)

* I thank participants at the first conference on Beeronomics (Katholieke Universiteit-Leuven, Belgium, 2009), the editors, and two anonymous referees for helpful comments in preparation of this manuscript. For kind assistance in the archives, and for permission to publish copyrighted material, I thank principals at the Guinness Archives (Diageo), Dublin and at the Special Collections Library, University College London. Any errors are my own.

[a] Roosevelt University, Chicago. 430 S. Michigan Ave, Chicago, IL 60605. e-mail: sziliak@roosevelt.edu.

I. Introduction: Fermentation and Quantification in the History of Science

How does fermentation fare in the history of scientific discovery and method, from quantitative chemistry to econometrics? Fairly well. Organized religion can get people feeling mighty high and precise. But evidently the quantitative study of alcohol creation – of fermentation – can really get people up and moving – moving higher up the ladder of scientific and commercial progress.

Quantitative studies of fermentation have solved some big scientific problems – such as the errors of observations committed daily in science. Few realize that over the past three centuries there are studies of beer and wine standing not at the arid valleys but at the glorious peaks of quantitative method.[1] To see the connection, just think of some story-book figures in the history of science – great experimentalists, such as Priestley, Laplace, Lavoisier, Pasteur, Johannsen, and Gosset aka "Student". Fermentation was for each of these scientists the raw material, the pathway to scientific discovery and commercial innovation.

One can tread further into the history of science and assert that beeronomics and oenonomics have bragging rights.[2] Take, for example, the dissenting minister, political economist, and experimental chemist, Joseph Priestley (1733–1804), who pioneered measurements of – and commercial uses for – oxygen and carbon dioxide, or what he called "fixed air".[3] The discoveries came while Priestley was looking at gases emitting from beer fermented in a public brewery, which happened to adjoin his house in Leeds, in the early 1770s. Priestley explains in his *Memoirs* (1806):[4]

> ... But nothing of a nature foreign to the duties of my profession [as a minister] engaged my attention while I was at Leeds so much as the prosecution of my experiments related to *electricity*, and especially the doctrine of *air*. The last I was led into in consequence of inhabiting a house adjoining to a public brewery, where I at first amused myself with making experiments on the fixed air which I found ready made in the process of fermentation.

Think about that: the tingly, cloudy stuff we fight about in politics and medicine –on policy issues from the amount of fizzy bubbles in cans to the size of

[1] American econometricians have not uniformly condoned the subject: for example, Irving Fisher and Harold Hotelling were both strongly opposed to alcohol and Fisher (1926) clanged a loud bell for Prohibition. Perhaps they did not realize the value added by alcohol to the field of statistics – a field which both men mastered.

[2] "Oenonomics" is the charming title of an Economist article advertising "wine economics" and this *Journal*: http://www.economist.com/blogs/freeexchange/2008/06/oenonomics. See Swinnen and Vandemoortele (2011) for discussion of "beeronomics", past and present.

[3] For example, the introduction of soda water can be traced to Priestley.

[4] Priestley (1806), quoted in Schofield (1966) p. 51; emphasis in original.

the ozone layer and the health of human lungs and bones – isolated by a student of beer, a political economist.

Consider another example, the famous oenological experiments of Lavoisier, conducted in France only a few years later. In 1788, the economist, tax collector, chemist, and pioneering experimental scientist, Antoine Lavoisier (1743–1794), introduced the scientific community to his "principle of the conservation of mass" – the physical accounting principle standing behind the "balanced equation" and "balance sheets" of mass known now to every student of chemistry. Lavoisier, a true economist, discovered the conservation principle in balance sheets he drew up while quantifying chemical substances observed before and after fermentation of grapes into wine.[5] Said Lavoisier in his *Elementary Treatise*:[6]

> One can see that in order to reach a solution ... it is necessary first to know well the analysis and the nature of the substances able to undergo fermentation; *for nothing is created, either in the operations of art, or in those of nature, and one can state as a principle that in every operation there is an equal quantity of material before and after the operation*; that the quality and the quantity of the principles are the same, and that there are nothing but changes, modifications.
>
> It is on this *principle* that the whole art of making experiments on chemistry is founded.

Lavoisier reported estimates of the weight of substances in wine in strikingly novel balance sheets, illustrating his findings in a chemical accounting framework as others before him had not.[7] Priestley's work with beer was a big inspiration to Lavoisier, who coined the term "oxygen" to describe the gas which Priestley had isolated. Still, Lavoisier's computations were rudimentary. He did not know the degree of accuracy afforded by his experiments. Like Priestley before him, Lavoisier lacked the statistical tools necessary for measuring probable uncertainty, both random and real, on for example the weight in pounds of "carbon of the residue of the sugar".[8] His eminent biographer explains that most critical "assessments [of his computations] miss the real problems Lavoisier faced ... The difficulties he encountered ... of knowing how to estimate the quantities of substances he could not directly weigh; of determining what adjustments he could legitimately make when his balance sheets did not work out perfectly, as they almost never did; of deciding what *degree of error was reasonable* in a

[5] Holmes (1985, pp. 394–395). See also Schabas (2006, p. 51), Lavoisier (1788), quoted in Holmes (1985, p. 394, n19).

[6] Lavoisier (1788), quoted in Holmes (1985, p. 394, n19). Compare the Lavoisier "principle" of the "whole art of making experiments" with that of Student (1923, p. 273), quoted in section VII below.

[7] Holmes (1985, pp. 388–389) reproduces figures from the original Lavoisier balance sheets.

[8] See, for example, Priestley's letter of October 19, 1771, to Richard Price (1723–1791) – the same Richard Price who discovered Bayes's essay on conditional probability: Schofield (1966), pp. 89–90. Priestley tells Price how he first tested his theory of oxygen, on live mice, by trapping mice in phials and hanging them face down in "fixed air" he found to exist above vats at the public brewery.

given type of experiment*, when he lacked formal methods for computing expected errors*."[9]

The conservation of mass principle solved a major problem of physical accounting, improving estimates of magnitudes for modified substances and changing the face of chemical, physical, and industrial science. Yet Lavoisier – despite an extended series of collaborations he had with a great mathematician, his friend Laplace – did not ever find a way to control for chance and systematic difference. Lacking a theory of errors, he was left unable to segregate random and real sources of error in studies of fermentation.[10] Lavoisier, despite Laplace, did not compute a single variance or standard deviation, and often he failed to report the number of observations from which an estimate was computed.

In the late 19th century Pasteur's research on bacteria and preservation was published in *Studies on Fermentation* (1879), a book which clarified a few things, but not the errors of observation encountered by Priestley, Lavoisier, and others. "As early as the 1870s, Pasteur was awarded American patents for his methods of manufacturing and preserving beer and wine".[11] Still, like Lavoisier, like most people before Student, Pasteur did not know to what extent random error, inconsistent conditions, and other real yet uncontrolled factors affected the accuracy of his small sample experiments on bacteria in the beer, nor did he have a statistical tool for measuring them.[12]

II. Guinnessometrics: Student's Economic Approach to the Logic of Uncertainty

A general solution to the problem of random error in small sample analysis was given in 1908, by Student.[13] "Student" is the pen name of William Sealy Gosset (1876–1937), an Oxford-trained chemist and experimental scientist who worked his entire adult life as a brewer and business man for the Guinness Brewery,

[9] Holmes (1985), xviii.

[10] The Laplace-Lavoisier collaborations are described by Holmes (1985), for example: pp. 162–201, 491, 558. A tax collector during the Reign of Terror, Lavoisier was charged with treason and executed by guillotine in Paris in 1794.

[11] Geison (1995), p. 266.

[12] Geison (1995), pp. 237–47 and Pasteur (1879), pp. 31–32. Compare Ziliak (2008), p. 208, discussing Gosset's (1908) hops-life regressions which Gosset used at Guinness to estimate the size of hops' contribution to the life and cost of unpasteurized beer. Hearing about these regressions Zellner (2010) replied that "Gosset was 50 years ahead of his time."

[13] Student (1908a), Fisher (1939). Student's (1942) published articles were collected and edited by Pearson and Wishart. See Ziliak (2010, 2008) for additional biographical and brewing-related discussion, and see Ziliak (2011a, 2010) and Harrison (2011) for discussion of the role of randomization, if any, for rational economic decision-making.

Dublin (1899 to 1937) and Park Royal (1935 to 1937).[14] Student was experimenting on three of the chief inputs to Guinness stout – barley, malt, and hops – when he made the discovery leading to what scientists now call Student's *t*-distribution, table, and test of significance.[15]

But Student's contribution to experimental science and the theory of errors extends far beyond Student's t – however permanent and fundamental t is. Between 1904 and 1937, Student innovated – more than two decades before R.A. Fisher – a useful collection of experimental concepts, methods, and attitudes, which were used for doing routine work at cooperating farms and at the Guinness brewery.[16]

As Head Experimental Brewer, a position he held from 1907 to 1935, Student's main charge was to experimentally brew, and to gradually improve, a consistent barrel of Guinness stout, input by input, from barley breeding to malt extract, at efficient economies of scale.[17] Pounding out more than 100 million gallons of stout in annual sales, the problem Student faced at Guinness was economically motivated and non-trivially large. While endeavoring to control product and reduce costs at the large brewery Student was consistently faced with a small number of observations on new barley to try, at $n = 2$, 4, or – if he was lucky – 7.[18] In the process, he – though self-trained in statistics – managed to solve a general problem in the classical theory of errors which had eluded statisticians from Laplace to Pearson.[19]

Less well-known is Student's contribution to experimental design, systematically ignored by Fisher.[20] Student found a method for maximizing the power to detect big economic differences (low Type II error) when the quantitative difference is really there to be detected.[21] Student opposed Fisher's randomized field experiments on grounds that, as Student proved as early as 1911, decisively so in

[14] Guinness maintained a separate laboratory for chemistry but Gosset did not work as a chemist. He was employed in the more prestigious laboratory, for scientific brewers, rising in 1935 from Head Experimental Brewer to Head Brewer. It is interesting for brewers to note that the chemistry laboratory at Guinness was equipped in part with standards, procedures, and instruments suggested by a paid outside consultant, Horace T. Brown (1903), for whom the Horace T. Brown Medal, Institute of Brewing & Distilling (IBD), is named.

[15] Gosset (1904), Student (1908a, 1925).

[16] Ziliak and McCloskey (2008) discuss Student's direct influence on Egon Pearson, Jerzy Neyman, Harold Jeffreys, Walter Shewhart, W. Edwards Deming, and many others.

[17] Beaven (1947); and see Gosset in References below.

[18] Gosset (1904); Student (1908), pp. 13–19.

[19] Fisher (1939), p. 1.

[20] Ziliak (2011a) finds that Student tried and rejected randomized trials as early as 1911. Fisher (1925, 1935) refused to acknowledge Student's rejections, despite repeated requests by Student himself.

[21] Student (1911, 1923, 1938). Ziliak and McCloskey (2008), chp. 20–22, discuss Fisher's decades-long refusal to admit cost, Bayes's rule, and the power of the test into the theory of inference. See also: Jeffreys (1961), pp. 369–397, and Savage (1971).

1923, and again in 1938, balanced designs are more precise, powerful, and efficient compared to random.[22]

Brewers and economists alike have not noticed as much as they might that Student's exacting theory of errors, both random and real, marked a significant advance over ambiguous reports of plant life and fermentation asserted by Priestley and Lavoisier down to Pasteur, Fisher, and Johannsen, working at the Carlsberg Laboratory in Denmark.[23]

The experimental concepts which Student used at the brewery to revolutionize science and brewing are outlined here, basically in order of their development by Student in his job as apprentice brewer (1899–1906), Head Experimental Brewer (1907–1935), and finally Head Brewer of Guinness (1935–1937): (1) net pecuniary advantage and the purpose of the experiment; (2) profitable odds versus a fixed rule for the level of statistical significance; (3) small samples of repeated and independent experiments; (4) random error versus "real" error; and (5) the power and efficiency of "balanced" over "randomized" field experiments in economics. The balance of this article illustrates these concepts with experiments designed and/or evaluated by Student at Guinness's brewery.

III. Note on Methods and Sources

The method here is textual exegesis of primary historical sources, much of which is housed in archives owned and copyrighted by Guinness Archives (Diageo), Dublin, Ireland, and by the Special Collections Library, University College London. The bulk of the discussion is based on unpublished brewing material found in Guinness *Laboratory Reports*, 1898 to 1912. The laboratory reports contain Student's most important theory and results on small samples of repeated experiments on the beer and its inputs (see Gosset, 1904). Other primary sources include: published articles, scientific notebooks, memoranda, and correspondence[24] as well as annual reports, brewers' reports, and financial statements.

[22] Most field experiments in economics published since the 1990s are unaware of the bias and cost introduced to their experiments by artificial randomization: see, for example, Levitt and List (2009). Contrast Ziliak (2011a, 2010), Harrison (2011), and Bruhn and McKenzie (2009).

[23] The science and scientists of the Carlsberg Laboratory, Denmark, are discussed by Holter and Moller (1976). Several of the Guinness "scientific brewers" and managing directors did what was essentially a post-graduate study of barley and beer in Denmark, working with Danish scientists who were evidently in some regards – such as plant breeding and chemistry – years ahead (Dennison and MacDonagh, 1998).

[24] Gosset was a prolific letter writer and many leading scientists – from Karl Pearson to Ronald Fisher – sought his counsel. For example, Gosset (1962) is a five-volume book containing more than 150 letters from Gosset to Fisher, written and exchanged between 1915 and 1934.

IV. Net Pecuniary Advantage and the Purpose of the Experiment

After publication of Fisher's *Statistical Methods for Research Workers* (1925) and, ten years later, *Design of Experiments* (1935), students of biometrics, econometrics and other statistical sciences are normally trained to stare at the smallness of the *p*-value, neglecting – however illogical – the bigness or smallness of the estimated coefficient and thus the economic significance of the real object of interest. I have shown elsewhere that in his influential books and articles Fisher deliberately erased the economic component from Student's methods while claiming to teach the original.[25] Said Fisher: "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 [said Fisher] *are thus formally regarded as significant*."[26]

In *Design*, the best-selling book of experimental statistics in the world, Fisher said: "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."[27] In his influential article, "Arrangement of Field Experiments," Fisher said again:[28]

> It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." ... Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and *ignore entirely all results which fail to reach this level*.

Ziliak and McCloskey (2008) have shown that 8 or 9 of every 10 scientists, articles, editors, and grantors take their principles of significance testing from Fisher, ignoring all results which fail to reach the 5 percent point, considering them, however erroneously, "unimportant", and, at the same time, falsely celebrating results which merely succeed in reaching the 5 percent point yet are insignificant in economic and other substantive regards.

Few realize that Student and Fisher battled over "significance" for two decades, largely, it seems, because Fisher did not mention the debate in accessible articles and books which, as Ziliak, McCloskey, and others have shown, continue to exert a large influence. Gosset aka Student rejected arbitrary rules about statistical significance as found in Student's tables – Fisher's arbitrary 5 percent rule

[25] Ziliak and McCloskey (2008), "How Economics Stays That Way: The Textbooks and the Referees", pp. 106–122; McCloskey and Ziliak (1996), pp. 99–101.
[26] Fisher (1925a [1941]), p. 42, italics supplied.
[27] Fisher (1935), p. 16.
[28] Fisher (1926), p. 504, italics supplied.

included. As Student told Karl Pearson in an important letter of 1905, reprinted in *Biometrika* by Egon Pearson:[29]

> When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery" (Gosset, 1904), I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [in mathematics, such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment*.

Yet Student was a businessman, not an academic, and behind-the-scenes he was in no position to off-set the influence of Fisher's various campaigns.[30] Student never deviated from the central point, making benefit and cost primary to experimental design, to interpretation of regression coefficients, and to assessment of mean differences and of the error of the whole experiment. Student's expected loss function approach was (well before Savage and Wald) well-suited to the principles he pursued for Guinness at the bottom line. In stark contrast to Fisher's "null" purpose of experimentation, the brewer said again and again, such as nearly twenty years after his original letter to Karl Pearson, that:[31]

> The object of testing varieties of cereals [barley] is to find out which will pay the farmer best. This may depend on quality, but in general it is an increase of yield which is profitable, and since yield is very variable from year to year and from farm to farm it is a difficult matter upon which to obtain conclusive evidence.

Yet Fisher, it has been shown, dismissed out of hand any talk of economic value or motive. Despite Student's advice, Fisher pushed the artificial 5 percent rule onto unsuspecting scientists, from the eminent mathematical economist Harold Hotelling to, for many years, Milton Friedman (who late in life, inspired by Savage, became a Bayesian).[32] Indeed, in Student's last published article,

[29] Gosset, c. April 1905, quoted in E.S. Pearson (1939), pp. 215–216; italics supplied; see also: Gosset (1905), Guinness Archives, file GDB/BRO/1102.

[30] Levitt and List (2009), p. 4, n5, claim surprisingly that "Gossett" [sic] published as "Student" because the "Guiness" [sic] "brewery did not allow employees to publish their research" (p. 4). Guinness brewers were not allowed to use their real name – and they were not allowed to mention Guinness or beer – in print, true. But they were otherwise encouraged to publish, and did. Between 1907 and 1938, for example, Gosset himself published 14 articles in *Biometrika* alone while rising in rank from Apprentice to Head Brewer (Ziliak (2008), Student (1942)).

[31] Student (1923), p. 271, paragraph one.

[32] Fisher's impact on the development of econometrics and on statistics in general is discussed by Ziliak and McCloskey (2008), chp. 20–22, *passim*.

Student, now clearly irritated by Fisher's anti-economic approach, wrote in *Biometrika*:

> I personally choose the method which is most likely to be profitable when designing the experiment rather than use Prof. Fisher's system of a posteriori choice* which has always seemed to me to savour rather too much of "heads I win, tails you lose".[33]

Student's economic approach to statistics did not disappear altogether.[34] It re-appeared in, for example, foundations endorsed by Pearson (1990), Jeffreys (1961), Zellner (2005), and Savage (1954): "In principle," Savage said, "if a number of experiments are available to a person, he has but to choose one whose set of derived acts has the greatest value to him, due account being taken of the cost of observation."[35] Recently the economic approach to the theory of errors has gained official legitimacy, thanks in part to a March 22nd, 2011 U.S. Supreme Court decision which unanimously rejected use of a bright-line rule for statistical significance.[36] But it has a long way to go as Fisherian methods continue to assume priority in the scientific bureaucracy.

V. Profitable Odds versus Statistical Significance

Student rejected artificial rules about significance from the beginning of his inquiries at the Brewery – at least four years before he published the first *t*-table and small sample test of significance. In November, 1904, Gosset – he would not be known as Student until three years later – discussed his first break-through on the economic meaning of statistical significance, in an internal report titled "The Application of the 'Law of Error' to the Work of the Brewery".[37] The Apprentice Brewer said:[38]

> Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment. What the odds should be depends –
>
> 1. On the degree of accuracy which the nature of the experiment allows, and
> 2. On the importance of the issues at stake.

[33] Student (1938), p. 370. His sarcasm about "*" is directed at Fisher's *Statistical Methods for Research Workers*, § 24.1, wherein against the brewer's advice Fisher advocates a 5 percent rule of statistical significance.
[34] See, for example, Pearson (1990) and Jeffreys (1961).
[35] Savage (1954), p. 116.
[36] Ziliak (2011b), McCloskey and Ziliak (2010); see also, for example: Zellner (2005), Press (2003).
[37] Gosset (1904), p. 3; Student (1907) is Student's first publication, notably to students of beer, "On the Error of Counting [Yeast Cells] with a Haemacytometer".
[38] Gosset (1904), p. 3.

In 1905 he told Karl Pearson, fellow brewers, and the Guinness Board of Directors that judgments depend on the estimated net pecuniary advantage forthcoming from an experiment, supposing that the findings of said experiment are pursued at or for the large-scale Brewery. Judgments also depend, he said in 1904, on the odds of being wrong, on the nature of the experiment, and on the importance of the issues at stake.

Comparing the level of saccharine content in a series of malt extracts which he and others' mixed in the Experimental Brewery with that found in malts being used in the Main Brewery, Gosset brought attention to a positive correlation he found between "the square root of the number of observations" – that is, the number of calculated differences in saccharine content between Experimental and Main Brewery malts – and the level of statistical significance. Other things equal, he said "the greater the number of observations of which means are taken [the larger the sample size of extract differences], the smaller the [probable or standard] error" of the estimates.[39] "And the curve which represents their frequency of error," he showed in a graph and plot drawing, "becomes taller and narrower."[40]

Prior to Gosset the relation between sample size and the level of statistical significance was rarely explored. For example, while looking at biometric samples with up to thousands of observations, Karl Pearson declared that a result departing by more than three standard deviations is "definitely significant."[41] The normal tables assumed very large samples. Yet Gosset, self-trained in statistics, found by experiment that at such large samples nearly everything is *statistically* "significant" – though not, in Gosset's terms, economically or scientifically "important". And, likewise, Gosset found that a small number of observations can be profitable, though not statistically significant in Pearson's conventional sense. Regardless, Gosset did not have the luxury of large samples. One of his earliest experiments employed a sample size of $n = 2$, which helps to explain why in the original 1908 article Gosset calculated a z statistic for $n = 2$.[42]

His 1904 article is worth exploring a bit further – especially for the econometrician and real-world firm that wants to earn more with less. Guinness malt was produced in Gosset's time primarily from Irish and English barley stock – Old Irish, Prentice, Plumage Archer, and Spratt Archer were effective varieties. Malt extract was measured by "degrees saccharine" per barrel of 168

[39] Gosset (1904), p. 5.
[40] Ibid, p. 7.
[41] *K.P. Lectures Volume I* [Gosset's Classroom Notebook], p. 13, 1906–7, Pearson Papers, Gosset file, UCL.
[42] Gosset (1904), p. 7; Student (1908a), p. 23. Student's t was called "z" until Student (1925), pp. 105–106. Student's original z table starts at $n = 4$ and stops at $n = 10$ (Student (1908a), p. 19). His illustration of the z test for $n = 2$ appears in the text as a separate calculation.

pounds malt weight.[43] An extract in the neighborhood of 133° saccharine gave the targeted level of alcohol for Guinness's beer. A much higher degree of saccharine would affect the stability and life of the beer, but it also increases alcohol content – which in turn increases the excise tax which Guinness owes to the British government, which – sad to say – ups the price of Dad's pint.[44] If, on the other hand, the alcohol content comes in too low, if the degree of saccharine is insufficient, customers would riot, or switch to Beamish and Beck's. In Gosset's view, ± .5 degrees saccharine was a difference or error in malt extract which Guinness and its customers could swallow. "It might be maintained," he said, "that malt extract "should be [estimated] within .5 of the true result with a probability of 10 to 1".[45] Using the mean differences of saccharine values of extracts, between the Main and Experimental breweries, Gosset calculated the odds of observing the stipulated accuracy for small and then large numbers of extracts.[46] He found that:

Odds in favour of smaller error than .5 [are with:]	
2 observations	4:1
3 "	7:1
4 "	12:1
5 "	19:1
82 "	practically infinite

Thus, Gosset concluded, "In order to get the accuracy we require [that is, 10 to 1 odds with .5 accuracy], we must, therefore, take the mean of [at least] four determinations."[47]

The Guinness Board cheered. The Apprentice Brewer found an economical way to assess the behavior of population parameters, using very small samples, and he created a small sample table of significance to calculate the probable errors of his experiments. These economic and statistical advances were achieved

[43] The formula is: Malt extract = ([Specific gravity of the wort] – 1000) × 4.67. Page 2 of *Laboratory Report, Vol. VII*, No. 5, Oct. 25, 1906, "The Relationship Between Laboratory and Brewery Extracts, Introduction and Part I," by Alan Jackson (with the assistance of W.S. Gosset), Guinness Archives, Diageo.

[44] The excise tax on beer charged by the British government increased exponentially with alcohol level during the First World War. Just before the war, in April, 1914 the "beer duty" was 3 s. 9 1/2 d. per barrel of 1027 gravity beer, 5 s. 9 1/2 d. for 1041 gravity, and 7 s. 9 d. for 1055 gravity. By November, 1914 the duty per barrel for those same gravities shot up to 11 s. 3 1/2 d, 17s. 1 1/2 d, and 23 s. 0 d. By April, 1919 duty per barrel stood at 34 s., 52 s., and 70 s. – more than 1,000 percent above pre-war levels (McMullen (1950), p. 234, Table 2: "The Beer Duty, 1914–1949"). See Nye (2007) for discussion of beer duty in previous centuries.

[45] Gosset (1904), p. 7.

[46] Ibid, p. 7.

[47] Ibid, p. 8.

four years before he completed and published Student's distribution, table, and test.

Discovering the odds of a low "real" error, not merely a low random error, was Gosset's theoretical and practical focus in the laboratory from the beginning, but especially after he designed and evaluated an agricultural field experiment for Mercer and Hall, in 1911, described below.[48]

VI. The Power and Efficiency of Balanced over Randomized Field Experiments in Economics

Real error is, in Student's method, what remains after controlling for both systematic and random sources of error in what is ideally a series of independent and repeated experiments. The classic example of systematic error in barley yield experiments is the differential fertility gradient in agriculture – that is, the diminishing marginal productivity of soil as one travels from one end of the field to another.[49] Fisher's randomized blocks do not control for differential soil fertility and other systematic factors, biasing results and causing supply side inefficiency.[50] Student found through repeated experimentation and in international cooperation with others – such as Pearson, Neyman, Beaven, Mercer, Hall, Hudson, and many others – that "balanced" designs of experiments are more precise, powerful, and efficient. The reason is precisely because balancing is the best way to minimize the confounding effect which differential fertility, systematic flood plains, bird attacks, and the like have on output. Randomization, by contrast, biases the level of the chief output of interest, that is, the mean difference in yield between treatment and control groups.

In 1937 Student was in full-blown battle with Fisher over the issues of randomization, significance testing, and the purpose of experiments. Student wrote to his collaborator and friend, Egon S. Pearson – son of Karl, co-author with Neyman, editor of *Biometrika*, and chair of the Statistics Department at University College London – using no uncertain terms:[51]

> Obviously the important thing [Student said] ... is to have a low real error, not to have a [statistically] "significant" result at a particular station. The latter seems to me to be nearly valueless in itself. ... Experiments at a single station [that is, tests of statistical significance on a single set of data] *are* almost valueless. ... You want to be able to say not only "We have significant evidence that if farmers in general do this

[48] Student (1911).
[49] Student (1911, 1923).
[50] See Ziliak (2011a) for discussion of Student and post-Student experimental philosophy not considered by field experimentalists such as Levitt and List (2009), Duflo *et al.* (2006), Banerjee and Duflo (2011), and Karlan and List (2007), all of whom claim to take their cues from Fisher.
[51] Student (1937), quoted in Pearson (1939), p. 244; emphasis in original.

they will make money by it", but also "we have found it so in nineteen cases out of twenty and we are finding out why it doesn't work in the twentieth." To do that you have to be as sure as possible which is the 20th – your real error must be small.

Why did Student prefer balanced over randomized designs? Suppose a rational brewer is determined to experiment on barley. And suppose for simplicity that the brewer is a price taker in a competitive market who seeks to maximize output Q* at long-run efficient economies of scale. Other things equal, higher output of Guinness stout implies higher demand for barley – and thus a greater need for precision, efficiency, and robustness on the large scale.

Imagine a square or rectangular agricultural field arranged in n blocks and k treatments to be compared in (for example) one-way or more ANOVA. To lay out the $n \times k$ field, both real and random sources of fluctuation must be balanced. Artificial randomization of design, one option, entails use of a table of random numbers, shuffled cards, dice, or other randomizing device to determine the layout of treatments and controls, the planting of old and new varieties of barley, for example. In general treatments and/or varieties are selected and assigned to experimental units from the set of $(k!)^{n-1}$ possible patterns.[52]

Likewise in a series of such experiments over time and space, the analyst might assign treatments and/or varieties to plots of land sowed by Farmer Jo but not by Farmer Julian, or on these blocks of Jo's plot but not on Jo's other blocks in the same plot, and vice versa – meaning that one can randomly select plots or blocks to receive a given treatment and/or variety. Regardless, only one pattern will be selected from the sample space for experimental trial, given Q*.

The first known attempt to compare the precision and value of random versus balanced designs appears to be Student (1911), in his appendix to Mercer's and Hall's "The Experimental Error of Field Trials". Student was recruited by A.D. Hall, the director of research at Rothamsted in the era before Fisher, to design and analyze the Mercer-Hall experiment.

Student titled his appendix, "Note on a Method of Arranging Plots so as to Utilize a Given Area of Land to the Best Advantage in Testing Two Varieties".[53] He found through repeated experimentation on Mercer's and Hall's mangolds (a close relative of beets and Swiss chards) that standard deviations of yield differences shrank more and more, the smaller and smaller is the plot size. He found that the closer in space that the competing varieties and/or treatments and controls are sown together – the more precise the standard deviations. In an admittedly crude way, Student proved in 1911 with Mercer's and Hall's mangolds data that randomization is inferior to deliberate balancing. Student spent three

[52] Pearson (1938), p. 163.
[53] Student (1911).

decades refining his theory and further proving the statistical and economic advantages of balanced designs.[54]

On farm land in any part of the world, differential fertility gradients cut across a heterogeneous field, creating a systematic, non-random hierarchy of good and bad growing conditions which cannot be randomized artificially without experiencing a cost. The point that Student made and Fisher refused to acknowledge is that deliberate randomization gives less assurance that treatments and controls will have equal access to good and bad soil, to good and bad growing conditions.[55] Randomization-based "validity" is lovely to contemplate in theory. Yet compared with balanced designs, random designs have been found to be imprecise, expensive and comparatively powerless.[56]

Balanced designs are deliberate or systematic arrangements of treatments and controls made by the analyst conscious of real ground and/or other confounding sources of fluctuation in the output of interest. "Balancing" means creation of symmetry in all of the important sources of error, random and systematic, good soil and bad.

If the confounding variable in the environment exhibits a systematic spatial or temporal correlation with respect to the experimental output the arrangement chosen by pure randomization is going to bias results; coefficients will be wrong and power, inadequate. This is as true of medicine as it is of social policy and the new development economics.[57] In barley trials, for example, a significance test based on artificial randomization only does not control for a major source of error and will give in general less valid results. Early examples of balanced designs in crop yield trials are chessboard, checkerboard, Knight's move, Latin Square, and "ABBA" – otherwise known as "the half-drill strip" method.[58]

Take the Knight's move, for example, balancing an 8×8 chessboard-type design. Suppose a brewer is comparing yield of 8 different varieties of barley in the field – as Student did in his classic 1923 *Biometrika* article – the 8 varieties matching the 8 rows and columns of an actual chessboard. How shall the different treatments be planted? "Knight's move" says to balance the distribution of real sources of error (that is the diminishing marginal productivity of the land) by allocating seeds of a unique variety as one would a Knight's piece in chess – plant Variety A in a block that is two up and one over from a block occupied by A; Variety B, again, like the Knight's move in chess, should be assigned to a block that is one down and two over from a block occupied by one of its own kind, and so forth, for each variety

[54] In: Student (1923, 1938), Gosset (1936); see also: Pearson (1938, 1990), Neyman and Pearson (1938), and Jeffreys (1961), discussed in Ziliak (2010).
[55] Discussed in Ziliak (2011a) and Ziliak and McCloskey (2008), chps. 20–22.
[56] Ziliak (2010).
[57] Rothman *et al*. (2008); Bruhn and Mackenzie (2009); Heckman and Vytlacil (2007); Ziliak (2010).
[58] Student (1923); Gosset (1936); Beaven (1947).

and permutation of the experiment, given the chosen $n \times k$ design and the number of experiments (farms) in the series.

Consider the simplest case, comparing yields of two different barleys, barley A (the standard variety) and barley B the new.

In the 8×8 chessboard layout the experimental field has $n = 64$ blocks in which one may randomly or deliberately grow variety A or variety B. (In a more complicated strategy, blocks may be further subdivided, such that individual blocks can grow seeds from A and B. We will stick to the simple event that each block gets a unique "treatment".) A random assignment of A and B to blocks may produce, for example, the following pattern:

A A A A A A B B
A A A A A A B A
...etc. (i)
→ Direction of increase in soil fertility (higher yielding soil)

Another random draw may produce blocks of this sort:

B B B B B A A B
B B B B B A A A
...etc. (ii)
→ Direction of increase in soil fertility (higher yielding soil)

How precise are the estimated differences in average yields, A-B, or B-A, if fertility on the left side of the field is systematically lower than fertility on the right? Layouts such as (i) and (ii) – though random – produce biased mean squared errors and parameter estimates with respect to a major source of fluctuation – differential soil fertility. In example (i) the As are bunched up and growing in the very worst soil; thus the yield of the Bs will be artificially high, and the real treatment difference, A-B, will be undetermined.

Student found again and again that deliberate balancing – though adding to the "apparent" error, that is, to Type I error in ANOVA terms, actually *reduces* the real error of the experiment – minimizing Type 2 error and errors from fixed effects, such as non-random soil heterogeneity.[59]

Examples (i) and (ii) suggest that whenever there is a systematically variant fertility slope (or other temporal or secular source of local and fixed effect) which cannot be artificially randomized, the systematic source of fluctuation cannot be ignored without cost: differences in yield will be correlated by local and adjacent fertility slopes – ground effects – any temporal or spatial or real difference which

[59] Student (1938), pp. 364–372.

can't be randomized[60]. Random layouts analyzed with Student's test of significance will yield on average more biased differences, A-B and B-A, and less ability to detect a true difference when the difference is large.

By 1923 Gosset's solution for dealing with systematic sources of variation between A and B became (grammarians have to admit) perfectly balanced: he dubbed his balanced design of choice, "ABBA".[61] The ABBA layout is:

$$\begin{array}{l} \text{A B B A A B B A} \\ \text{A B B A A B B A} \\ \text{A B B A A B B A} \\ \text{...etc.} \end{array} \qquad \text{(iii)}$$

The ABBA design minimizes bias caused by differential soil fertility. Given the built-in symmetry of ABBA, no matter the magnitude of differential fertility gradients, the A's and B's are equally likely to be grown on good and bad soil. Random throws of seed do not have this virtue, biasing mean yield differences, A–B.[62]

Yet ABBA brings additional statistical and economic advantages, too. On the supply side, with ABBA the ease and cost of sowing and harvesting and calculating basic statistics on yield is plot-wise and block-wise reduced. Compare the rows and columns of ABBA with the random rows and columns in (i) and (ii) above and it's easy to appreciate Student's sensitivity to supply side economic conditions.

With ABBA there is no need for chaotic tractor driving while planting seed in blocks randomly dispersed; and thus with ABBA there is a lot less measurement error and loss of material at harvest and counting time.[63] Imagine harvesting and counting up the mean difference in yield of strip A minus strip B, block by block, in the ABBA field versus the randomized and one can appreciate further still the efficiency of Student's balanced solution. As Student told Fisher in a letter of 1923 previously mentioned, "There must be essential similarity to ordinary [in this case, farming] practice". (Pearson shows how to adjust the ANOVA and Student's test of significance to accommodate the ABBA structure.[64]) After all, "[t]he randomized treatment pattern is sometimes extremely difficult to apply with ordinary agricultural implements, and he [Student] knew from a wide

[60] As Heckman and Vytlacil (2007), p. 4836, put it: "Randomization is a metaphor and not an ideal or "gold standard". Heckman's research on selection bias in social policy experiments is a good example – another reason to balance both random and real sources of error.
[61] Student (1938), pp. 364–378.
[62] Unfortunately the new development economists have not grasped this point, despite their stated goal of raising income and reducing poverty; for example, Banerjee and Duflo (2011).
[63] See Beaven (1947), pp. 275–295, for detailed advice on how to prepare, sow, and harvest the ABBA ("half-drill strip") field experiment.
[64] Pearson (1938), pp. 163–164.

correspondence how often experimenters were troubled or discouraged by the statement that without randomization, conclusions were invalid".[65]

Fisher, for his part, rejected Student's ABBA and other balanced designs (see, for example, Fisher and Yates (1938), which fails to mention Student's methods, though published less than a year after Student died). Student's last article – which he worked on during the final months and days of his life and until the day he died – was a direct response to Fisher:[66]

> It is of course perfectly true that in the long run, taking all possible arrangements, exactly as many misleading conclusions will be drawn as are allowed for in the tables [Student's tables], and anyone prepared to spend a blameless life in repeating an experiment would doubtless confirm this; nevertheless it would be pedantic to continue with an arrangement of plots known before hand to be likely to lead to a misleading conclusion. ...
>
> In short, there is a dilemma – either you must occasionally make experiments which you know beforehand are likely to give misleading results or you must give up the strict applicability of the tables; assuming the latter choice, why not avoid as many misleading results as possible by balancing the arrangements? ... To sum up, lack of randomness may be a source of serious blunders to careless or ignorant experimenters, but when, as is usual, there is a fertility slope, *balanced arrangements tend to give mean values of higher precision compared with artificial arrangements*

What about variance? What affect does balancing have on variance and thus on the level of statistical significance?

"The consequence is that balanced arrangements more often fail to describe small departures from the 'null' hypothesis as significant than do random, though they make up for this by ascribing significance more often when the differences are large."[67]

VII. Pearson's Illustration of the Higher Power

The intuition behind the higher power of ABBA[68] and other balanced designs to detect a large and real treatment difference was given by Student in

[65] Pearson (1938), p. 177.
[66] Student (1938), p. 366.
[67] Student (1938), p. 367.
[68] Student's and Beaven's ABBA design is, formally speaking, *chiasmus* – one of the most powerful and influential design patterns in the history of language, music, religion, and science. What is chiasmus beyond the symmetric Greek symbol for *chi*, X, from which the term derives? Lanham (1991), p. 33, defines chiasmus as "The ABBA pattern of mirror inversion". Unaware of Student's ABBA, the classical rhetorician explains: "Chiasmus seems to set up a natural internal dynamic that draws the parts closer together ... The ABBA form," he notes, "seems to exhaust the possibilities of argument, as when

1911.[69] "Now if we are comparing two varieties it is clearly of advantage to arrange the plots in such a way that the yields of both varieties shall be affected as far as possible by the same causes to as nearly as possible an equal extent".[70] He called this part of his theory, to repeat, the principle of "maximum contiguity", a principle he put to work again and again, such as when he illustrated the higher precision and lower cost associated with a small-sample study of biological twins, to determine the growth trajectory of children fed with pasteurized milk, unpasteurized milk, and no milk at all, in "The Lanarkshire Milk Experiment".[71] The power of balanced designs to detect real differences can be seen if one imagines doing as Student did, trying to maximize the correlation of adjacently growing varieties and/or treatments, the As and Bs, just as one might when studying the effect of differential milk consumption on identical twins.

In "Some Aspects of the Problem of Randomization: II. An Illustration of Student's Inquiry Into the Effect of 'Balancing' in Agricultural Experiment," Egon S. Pearson (1938) – another skeptic not mentioned by Fisher and the randomization school – clarified and confirmed Student's practical theory of balancing.[72] Said Pearson:[73]

> In co-operative experiments undertaken at a number of centres, in which as he [that is Gosset aka Student] emphasized he was chiefly interested, it is of primary concern to study the difference between two (or more) "treatments" under the varying conditions existing in a number of localities.

Samuel Johnson destroyed an aspiring author with, 'Your manuscript is both good and original; but the part that is good is not original, and the part that is original is not good'" (p. 33). Good, original, original, good: the ABBA layout. James Joyce, another famous Dubliner in Student's day, wrote chiasmus in his novella *Portrait of the Artist as a Young Man*. Other examples of chiasmus are by John F. Kennedy ("Ask not what your country can do for you; ask what you can do for your country") and by Matthew 23:11–12 ("Whoever exalts himself will be humbled, and whoever humbles himself will be exalted"). In science, supply and demand and the double helix are two notable examples of chiasmus.

[69] Compare Beaven (1947), pp. 273–5.

[70] Student (1911), p. 128.

[71] "Maximum contiguity" appears importantly in Student (1923), p. 278, converting Neyman away from randomization and in support of balance. In the Lanarkshire Milk Experiment Student (1931a, p. 405), applied contiguity to experimental design and found that "50 pairs of [identical twins] would give more reliable results than the 20,000" child sample, neither balanced nor random, actually studied in the large-scale social experiment funded by the Scotland Department of Health. "[I]t would be possible to obtain much greater certainty" in the measured difference of growth in height and weight of children drinking raw versus pasteurized milk "at an expenditure of perhaps 1–2% of the money and less than 5% of the trouble." Likewise, Karlan and List (2007, p. 1777), could have learned more about the economics of charitable giving – for less – using a variant of Student's method. Instead the *AER* article studied n = 50,083 primarily white, male, pro-Al Gore donors to public radio, neither random nor balanced.

[72] Box 10, brown folder, Egon Pearson Papers, University College London, Special Collections Library.

[73] Pearson (1938), p. 177.

For treatments and/or varieties A and B, Student's idea is to estimate from the ABBA experiment:

$$x_A = m_{A+}\delta_A$$

and

$$\begin{gathered} x_B = m_{B+}\delta_B \\ \text{and thus}: x_A - x_B = (m_A - m_B) + \delta_A - \delta_B = (m_A - m_B) + \Delta_{AB} \end{gathered} \tag{iv}$$

where x_i is the yield from the ith block or plot, m_i is the yield in the ith block or plot to which a treatment has been applied (an unchanging value no matter the treatment) and δ_i is the real treatment in the block or plot.[74]

Students of Heckman, Friedman, and Zellner, for example, will not be surprised by what follows from Student's and Pearson's set up, which strives to achieve real error minimization. The comparative advantage of Student's and Pearson's ABBA design in repeated trials is: (1) ABBA enables explicit control of the *m's* – the difference in growing conditions or other fixed factor whose influence you are trying to minimize, and (2) ABBA enables more control of the variance of Student's $\Delta_{AB's}$ –the real treatment effects (or causes if you will) on yield, within and between farm fields.

> It has been said from an experiment conducted by this method no valid conclusion can be drawn, but even if this were so, it would not affect a series of such experiments.[75] Each is independent of all the others, and it is not necessary to randomize a series which is already random, for, as Lincoln said, "you can't unscramble an egg". Hence, the tendency of deliberate randomizing is to increase the error.[76]

Using a simple workhorse formula, Student showed that the ABBA layout – arranging varieties or treatments A and B close together in strips – reduces the standard deviation of yield differences by maximizing ρ – the correlation between yields of the competing treatments and/or varieties, A and B. The formula he used as the basis for measuring the variance of mean differences, A-B,

[74] Pearson (1938), pp. 163–164.

[75] Beaven (1947), p. 293, reported after 50 years of experimentation on barley using his and Student's methods that selection of a new cereal takes "about ten years" (p. 293) of repeated and balanced experimentation. By the early 1920s three different varieties of barley, selected and proved by Beaven and Student, were grown on "well over five million acres of land" (Beaven, xiv). When the great experimental maltster and barley farmer, Edwin S. Beaven, died (in 1941) Guinness acquired ownership and continued to produce at the famous farm and malt house in Warminster, UK (http://www.warminster-malt.co.uk/history.php). Contrast the new field experiments in economics, neither repeated nor balanced, yet full of advice for going concerns (Herberich, Levitt and List (2009), Levitt and List (2009), Banerjee and Duflo (2011), and by now hundreds of others).

[76] Gosset (1936), p. 118.

he got in 1905 directly from Karl Pearson, during a July visit to Pearson's summer house:[77]

$$\sigma^2_{A-B} = \sigma^2_A + \sigma^2_B - 2\rho_{AB}\sigma_A\sigma_B \quad (v)$$

where σ^2 is variance and ρ is the Pearson correlation coefficient.[78]

Given the systematic way that the sun, wind, water, and other environmental features – such as rabbit holes and fertility gradients – affect growth potential in a given block or sub-block of the agricultural plot, the spatial closeness and symmetry of ABBA maximizes the size of ρ – exactly what the analyst wants when high power, efficiency, and equal balance of confounding errors are goals.

The higher the correlation ρ between yields A and B the lower is the variance of their differences A-B and B-A. Thus compared to random the ABBA design gets more statistically significant results when the differences between A and B are truthfully large – the power to detect is high when the effect size is large – exactly what the firm – such as a large scale brewery – wants when precision and profit are goals.

Fisher's randomization – and the new field experiments – ignore the fundamental importance of the correlation coefficient, ρ; assuming independent and identically distributed observations in imaginary replications, artificial randomization seeks only to minimize $\sigma^2{}_A$ and $\sigma^2{}_B$. Yet plot by plot, as Student said:[79]

> The art of designing all experiments lies even more in arranging matters so that ρ [the correlation coefficient] is as large as possible than in reducing $\sigma^2{}_x$ and $\sigma^2{}_y$ [the variance].
>
> The peculiar difficulties of the problem lie in the fact that the soil in which the experiments are carried out is nowhere really uniform; however little it may vary from eye to eye, it is found to vary not only from acre to acre but from yard to yard, and

[77] Karl Pearson, quoted by Gosset (1905), Guinness Archives; Reprinted: Pearson (1939), p. 212.
[78] Though unaware of the Student-Pearson model for estimating the variance of real treatment effects in yield trials, Friedman (1953) used the same basic set-up to simulate the effect of a counter-cyclical expenditure (spent by a central government in search of full-employment) on the variance of aggregate income. "Let $X(t)$ represent income at time t in the absence of the specified full-employment policy. The full-employment policy," Friedman said, "may be regarded as having effects that add to or subtract from income." Searching for something like Student's real treatment effect, $\delta_A - \delta_B$, Friedman continued: "Let $Y(t)$ represent the amount so added or subtracted from $X(t)$, so that: $Z(t) = X(t) + Y(t)$ represents income at time t. ... What is the optimum size of $\sigma^2{}_y$?" Friedman asked. "By a well-known statistical theorem $\sigma^2{}_z = \sigma^2{}_x + \sigma^2{}_y + 2r_{xy}\ \sigma_x\ \sigma_y$ where r_{xy} is the correlation coefficient between X and Y" (Friedman, pp. 122–123). Other things equal, the effect of the "stabilizing" action depends in large part on the magnitude of the correlation coefficient, just as Friedman and others after Student and Pearson have demonstrated in a vast body of literature ignored by new field experimentalists, notably, Levitt and List (2009), Banerjee and Duflo (2011), and Karlan and List (2007).
[79] Student (1923), p. 273.

> even from inch to inch. This variation is anything but random [Student observes], so the ordinary formulae for combining errors of observation which are based on randomness are even less applicable than usual.

Thus it is quite misleading when Levitt and List assert that "Gossett [sic] understood randomization and its importance to good experimental design and proper statistical inference."[80]

When estimating how Δ_{AB} – the real treatment difference – varies from one set of conditions to another (for example, from one farm to another) one is free to assume the validity of Student's table of *t* and test of significance. Randomness – not randomization – is all that one needs to justify use of Student's table.

In *Theory of Probability* (1961), "§4.9 Artificial Randomization," the great Bayesian experimentalist Harold Jeffreys agreed with Student, and strongly disagreed with Fisher. When fertility contours are present (and uniformity trials showed that they always were) "there is an appreciable chance that [the differences in soil] may lead to differences that would be wrongly interpreted as varietal [as relating to the barley rather than to the fixed features of the soil; in medicine think of the pill and the different abilities of hearts]."[81] "Fisher proceeds ... to *make it* into a random error."[82] But:[83]

> Here is the first principle [Jeffreys said]: we must not try to randomize a systematic effect that is known to be considerable in relation to what we are trying to find ... The [balanced] method of analysis deliberately sacrifices some accuracy in estimation for the sake of convenience in analysis. The question is whether this loss is enough to matter, and we are considering again the efficiency of an estimate. But this must be considered in relation to the purpose of the experiment in the first place.

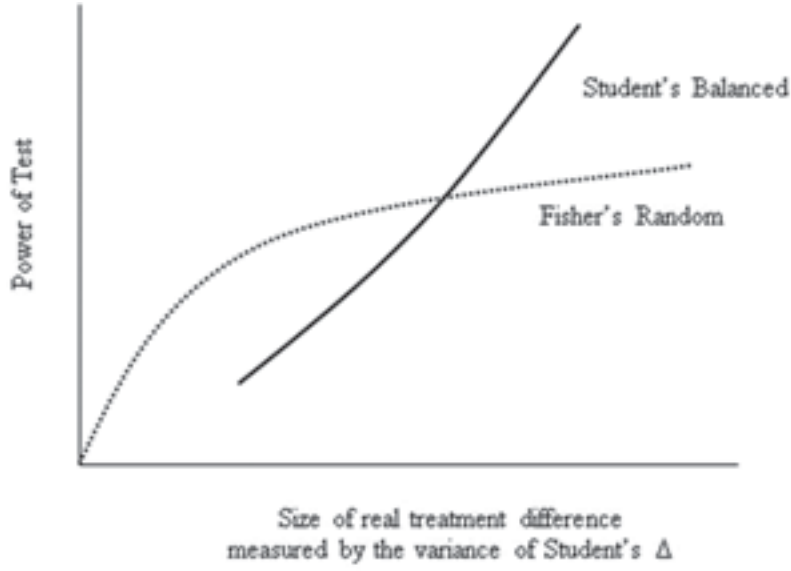
Thus a well-designed field experiment in economics strives for efficiency, and for the power to detect a minimally important difference, with a low real error. Fisher-randomization and significance, measured by the p-value, does not. Said Jeffreys again, citing Student (1923, 1938) as the source of his ideas:[84]

> There will in general be varietal differences; we have to decide whether they are large enough to interest a farmer, who would not go to the expense of changing his methods unless there was a fairly substantial gain in prospect. There is, therefore, a minimum difference that is worth asserting.

[80] Levitt and List (2009), p. 4.
[81] Jeffreys (1961), p. 242.
[82] Ibid, p. 243; italics in original. The "it" is, in this case, the non-random distribution of soil productivity.
[83] Ibid, p. 243.
[84] Jeffreys (1961), p. 244. Jeffreys considered Student's methods to be in basic agreement with his own Bayesian approach (Jeffreys (1961), pp. 379, 393, 369–400). Ziliak (2008) and Ziliak and McCloskey (2008) describe Student's early and lasting endorsement of Bayes's theorem in practical work.

Figure 1

When Real Treatment Effects are Large, Power Curves Cross, Yielding Advantage to Balanced Over Random Designs



Sources: Student (1938), pp. 372-378; E.S. Pearson (1938), p. 177.

And to detect a minimum important difference, Student (1938) discovered in his last article – and Pearson's simulations later confirmed – "a definite advantage that seemed to be gained from balancing".[85] Exactly as Student expected, Pearson found that when treatment and/or varietal differences grow large, the power curves of balanced and random designs cross, lending advantage of detection to balanced designs (see Figure 1).

VIII. Beeronomics: Putting Problems Before Methods

Fermentation provides a valuable pathway to scientific discovery – among the greatest discoveries in the world. "The prominent place that Lavoisier accorded fermentation" was, the historian of science Frederic Holmes surmised, "a reflection of his keen interest in the process".[86] Maybe so. In his *Memoirs*, Priestley hints that the "process" of fermentation "amused" him, too – extra reason to stay near the beer. Student, we have seen, applied hawkish attention to cost and to the

[85] Pearson (1938), p. 177.
[86] Holmes (1985), p. 9.

magnitude of correlations at each stage of the process, estimating econometrically, over time and space, the value-added of each input to the beer. Additional research is needed to explain *why* fermentation plus the economic approach proves an exceptionally good stimulant for clarifying the theory of errors. Animal and insect cross-breeders, for instance – despite the undoubted greatness of Darwin, Mendel, and Dobzhansky – cannot claim as much.

But it would seem that the problem facing beeronomics today is not a choice between economic versus non-economic approaches to the theory of errors. After Student's achievements at Guinness, after the Matrixx v. Siracusano Supreme Court decision, that part seems obvious. The problem is how to incentivize science so that concrete problems and contextual knowledge, not abstract rules about method, can once again lead the way, using as Student did a plurality of methods.

The legacy of Student is cautionary on top of inspiring: Student started with concrete problems and ended up with general methods. Today, following a model cast as if in stone by his younger friend Fisher, applied statisticians tend to do the opposite, putting artificial rules about Student's methods above the purpose, problem, profit, and context. Small sample theory was an economic, not a mathematical, decision. Cost minimization was central to Student's profit-seeking experimental mission. He wished to know, for example, the minimum number of observations necessary to claim a finding of such and such magnitude on a sliding scale of odds. If the odds should be greater, out of fear, say, of assuming too much of the Type I kind of risk, how small a number of "n" would be necessary in repeated experiments to have confidence in the general applicability of the result?

The most famous result of Student's experimental method is Student's *t*-table. But the real end of Student's inquiry was taste, quality control, and minimally efficient sample sizes for experimental Guinness – not to achieve statistical significance at the .05 level or, worse yet, boast about an artificially randomized experiment. Student's world class achievement raises a basic question about statistical science after Fisher: to what extent should scientists start with mechanical methods rather than with the nature of the economic or biometric problem to be solved? How much do we lose in beer, cereal, and money, by reversing the priority of problem and method?

References

Banerjee, A. and Duflo, E. (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Beaven, E.S. (1947). *Barley: Fifty Years of Observation and Experiment*. London: Duckworth.

Brown, H.T., (ed.) (1903). *Transactions of the Guinness Research Laboratory, Vol. I, Part I*. Dublin: Arthur Guinness, Son and Co., Ltd.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: randomization in practice in development economics. *American Economic Journal: Applied Economics*, 4, 200–232.

Dennison, S.R. and MacDonagh, O. (1998). *Guinness 1886–1939*. Cork: Cork University Press.

Duflo, E., Glennerster, R. and Kremer, M. (2006). *Using Randomization in Development Economics Research: A Toolkit*. December 12, 2006, J-PAL Poverty Action Lab, MIT.

Fisher, I. (1926). *Prohibition at its Worst*. New York: The Macmillan Company.

Fisher, R.A. (1925a [1941]). *Statistical Methods for Research Workers*. New York: G.E. Stechart and Co.

Fisher, R.A. (1926). Arrangement of Field Experiments. *Journal of Ministry of Agriculture*, 23, 503–513.

Fisher, R.A. (1935). *The Design of Experiments*, Edinburgh: Oliver & Boyd. Reprinted in eight editions and in at least four different languages.

Fisher, R.A. (1939). Student. *Annals of Eugenics*, 9, 1–9.

Fisher, R.A. and Yates, F. (1938 [1963]). *Statistical Tables for Biological, Agricultural and Medical Research*, Edinburgh: Oliver and Boyd. Sixth edition.

Friedman, M. (1953). The effects of a full-employment policy on economic stability: a formal analysis. In: Friedman, M., *Essays in Positive Economics*. Chicago: University of Chicago Press, 117–132.

Geison, G.L. (1995). *The Private Science of Louis Pasteur*. Princeton: Princeton University Press.

Gosset, W.S. [see "Student", below] (1904). The Application of the 'Law of Error' to the Work of the Brewery. *Laboratory Report*, 8, Arthur Guinness & Son, Ltd., Diageo, Guinness Archives, 3–16 and unnumbered appendix.

Gosset, W.S. (1905). Letter from W.S. Gosset to K. Pearson, Guinness Archives, GDB/BRO/1102 (partially reprinted in Pearson (1939), 215–216).

Gosset, W.S. (1908). The present position of our knowledge of the connection between life and hops in the experimental brewery. *Laboratory Report,* 10, Arthur Guinness & Son, Ltd., Diageo, Guinness Archives, 137–150.

Gosset, W.S. (1909). The brewing of the experimental hop farm hops, 1907 Crop (Part II), Together with a note on the present method of hop analysis. *Laboratory Report,* 10, Arthur Guinness & Son, Ltd., Diageo, Guinness Archives, 202–220.

Gosset, W.S. (1936). Co-operation in large-scale experiments. *Supplement to the Journal of the Royal Statistical Society*, 3, 115–36.

Gosset, W.S. (1962). *Letters of William Sealy Gosset to R.A. Fisher. Vols. 1–5,* Eckhart Library, University of Chicago. Private circulation.

Harrison, G. (2011). Randomization and its discontents. *Journal of African Economies*, 20, 626–652.

Heckman, J.J. and Vytlacil, E.J. (2007). Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation. In: Heckman, J.J. and Leamer, E. (eds.), *Handbook of Econometrics* 6B. Amsterdam: Elsevier, 4836–5143.

Herberich, D.H., Levitt, S.D. and List, J.A. (2009). Can field experiments return agricultural economics to the glory days? *American Journal of Agricultural Economics*, 91, 1259–1265.

Holmes, F.L. (1985). *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity*. Madison: University of Wisconsin Press.

Holter, H. and Møller, K.M. (eds.) (1976). *The Carlsberg Laboratory, 1876–1976*, Copenhagen: Rhodos and the Carlsberg Foundation.

Jeffreys, H. 1939 [1961]. *Theory of Probability*, London: Oxford University Press.

Karlan, D. and List, J. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, 97, 1774–1793.

Lanham, R.A. (1991). *A Handlist of Rhetorical Terms*. Los Angeles: University of California Press.

Leonard, A. (2009). Celebrate the history of statistics: drink a Guinness. How a master brewer forged new ground in the quantitative progress of science. *Salon*, September 28. http://mobile.salon.com/tech/htww/2009/09/28/guinessometrics/index.html

Levitt, S.D. and List, J.A. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, 53, 1–18.

McCloskey, D.N. and Ziliak, S.T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34, 97–114.

McCloskey, D.N. and Ziliak, S.T. (2010). *Brief of Amici Curiae By Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents* (Vol. No. 09–1156, Matrixx v. Siracusano, p. 22), Washington DC: Supreme Court of the United States. Edward Labaton *et al*., Counsel of Record (ed.)

McMullen, L. (1950). The sources and nature of statistical information in special fields of statistics. *Journal of the Royal Statistical Society, Series A (General)*, 113, 230–237.

Mercer, W.B. and Hall, A.D. (1911). The experimental error of yield trials. *Journal of Agricultural Science*, 4, 107–127.

Neyman, J. and Pearson, E.S. (1938). Note on some points on 'Student's' paper on 'comparison between balanced and random arrangements of field plots. *Biometrika*, 29, 379–88.

Nye, J. (2007). *War, Wine, and Taxes*. Princeton: Princeton University Press.

Pasteur, L. (1879). *Studies of Fermentation: The Diseases of Beer – Their Causes and the Means of Preventing Them*. London: Macmillan. (Faulkner, F. and Robb, D. Constable, transl.)

Pearson Papers (Containing files on K. Pearson, E.S. Pearson, W.S. Gosset, R.A. Fisher, and J. Neyman), University College London (UCL), Special Collections Library.

Pearson, E.S. (1938). Some aspects of the problem of randomization: II. An illustration of "Student's" inquiry into the effect of "balancing" in agricultural experiment. *Biometrika*, 30, 159–179.

Pearson, E.S. (1939). 'Student' as statistician. *Biometrika*, 30, 210–50.

Pearson, E.S. (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, 445–57.

Pearson, E.S. (1990) [posthumous]. *'Student': A Statistical Biography of William Sealy Gosset*. Oxford: Clarendon Press. Plackett, R.L. (ed.), with the assistance of Barnard, G.

Press, S.J. (2003). *Subjective and Objective Bayesian Statistics*. New York: Wiley.

Priestley, J. (1806). *Memoirs of Dr. Joseph Priestley, to the Year 1795, Written by Himself*. London: J. Johnson.

Rothman, K.J., Greenland, S. and Lash, T.L. (2008). *Modern Epidemiology*, Philadelphia: Lippincott, Williams & Wilkins.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Dover.

Savage, L.J. (1971). [1976 posthumous]. On re-reading R.A. Fisher. *Annals of Statistics*, 4, 441–500.

Schabas, M. (2006). *The Natural Origins of Economics*. Chicago: University of Chicago Press.

Schofield, R.E. (1966). *A Scientific Autobiography of Joseph Priestley, 1733–1804: Selected Scientific Correspondence, with Commentary*. Cambridge, MA: The MIT Press.

Stewart, G.G. (2009). The Horace Brown Medal Lecture: forty years of brewing research. *Journal of the Institute of Brewing*, 115, 3–29.

Student [see also: Gosset, W.S.] (1907). On the error of counting with a haemacytometer. *Biometrika*, 5, 351–60.
Student (1908a). The probable error of a mean. *Biometrika*, 6, 1–24.
Student (1908b). The probable error of a correlation coefficient. *Biometrika*, 6, 300–310.
Student (1911). Appendix to Mercer and Hall's paper on 'the experimental error of field trials'. *Journal of Agricultural Science*, 4, 128–131.
Student (1923). On testing varieties of cereals. *Biometrika*, 15, 271–293.
Student (1925). New tables for testing the significance of observations. *Metron*, 5, 105–108.
Student (1931a). The Lanarkshire milk experiment. *Biometrika*, 23, 398–406.
Student (1931b). On the '*z*' test. *Biometrika*, 23, 407–8.
Student (1931c). Yield Trials. *Bailliere's Encyclopedia of Scientific Agriculture,* 1342–1360; Reprinted: pp. 150–168 in Pearson, E.S. and Wishart, J. (eds.) (1942). *Student's Collected Papers*. London: Biometrika Office.
Student (1938, posthumous). Comparison between balanced and random arrangements of field plots. *Biometrika*, 29, 363–78.
Student (1942, posthumous). *Student's Collected Papers*. London: Biometrika Office. Pearson E.S. and Wishart, J. (eds.).
Swinnen, J. and Vandemoortele, T. (2011). Beeronomics: the economics of beer and brewing. In: Swinnen, J.F.M. (ed.), *The Economics of Beer*. Oxford: Oxford University Press, 335–355.
Zellner, A. (2005). *Statistics, Econometrics, and Forecasting*. The Stone Lectures in Economics. Cambridge, UK: Cambridge University Press.
Zellner, A. (2010). Personal communication. University of Chicago, Quadrangle Club.
Ziliak, S.T. (2008). Guinnessometrics: the economic foundation of Student's' *t*. *Journal of Economic Perspectives*, 22, 199–216.
Ziliak, S.T. (2010). The *validus medicus* and a new gold standard. *The Lancet*, 376, 324–325.
Ziliak, S.T. (2011a). Field experiments in economics: comment on an article by Levitt and List. *CREATES Research Paper No. 2011–25*, Aarhus: Center for Research in Econometric Analysis of Times Series, Aarhus University, Denmark.
Ziliak, S.T. (2011b). Matrixx v. Siracusano and Student v. Fisher: statistical significance on trial. *Significance*, 8, 131–134.
Ziliak, S.T. and McCloskey, D.N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.