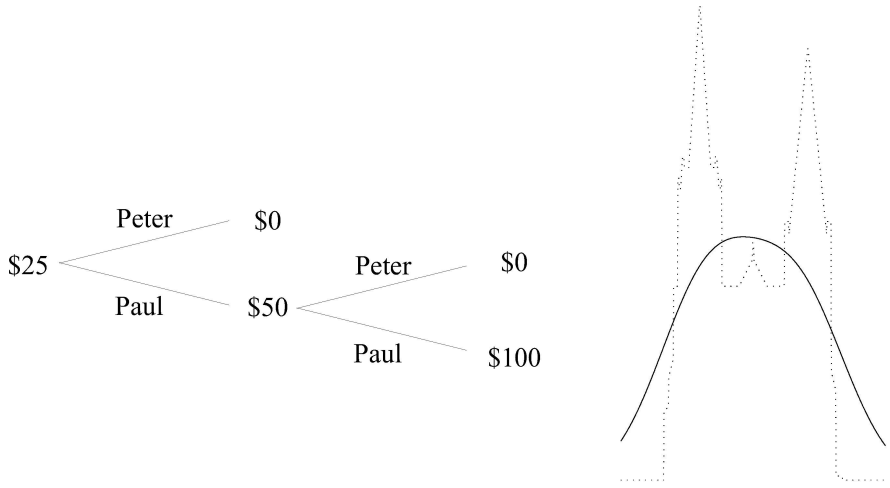
On the nineteenth-century origins of significance testing and p-hacking

Glenn Shafer, Rutgers University



The Game-Theoretic Probability and Finance Project

Working Paper #55

First posted July 18, 2019. Last revised September 14, 2019.

Project web site:
http://www.probabilityandfinance.com

Abstract

This paper examines the development of the Laplacean concept of practical certainty from 1810, when Laplace proved his central limit theorem, to 1925, when Ronald A. Fisher published his *Statistical Methods for Research Workers*.

Many nineteenth-century statisticians had no use for probability. Those who did use probability usually indicated the precision of an estimate by stating its probable error or its modulus, a multiple of the probable error. In the Laplacean (as contrasted with the Gaussian) tradition, it was considered practically certain that a quantity being estimated is within some some specified number k of probable errors (or moduli) of the estimate. In particular, it was considered practically certain that two quantities are different when the estimate of their difference is more than k probable errors (or moduli) from zero.

Laplace's explanations of the applications of his theorem were accessible to only a few mathematicians. But expositions published by Joseph Fourier in 1826 and 1829 made the simplest applications accessible to many statisticians. And whereas Fourier had suggested an error probability of 1 in 20,000, statisticians soon chose less exigent standards. Abuses, including p-hacking, soon emerged and helped discredit Laplace's theory in France to the extent that it was practically forgotten there by the end of the 19th century. But it was survived elsewhere and served as the basis for the work of the British school of biometry launched by Karl Pearson.

The probability that a normally distributed random variable is more than 3 probable errors from its mean is approximately 5%. When Fisher published his *Statistical Methods for Research Workers* in 1925, three probable errors was already a common standard for "likely significance". Fisher's purpose was to show research workers how to use distributions other than the normal — the t distributions, the distribution of the correlation coefficient, etc. So he replaced "three probable errors" with 5%.

The way Fisher and subsequent statisticians have used the word *significant* differs from the way it was used by Karl Pearson and other British statisicians through the first couple decades of the 20th century. For Pearson and his colleagues, a *significant difference* was an observed difference that *signified* a real difference. So while they often said that a difference is likely or very likely to be significant, they never said that it is very significant, and they never used the phrase "levels of significance".

What might this history teach us about proposals to curtail abuses of statistical testing by changing the language (p-value, significance, etc.) used nowadays? The fact similar abuses arose before this language was introduced is an argument for skepticism.

1 Introduction **1**

2 Laplace's theorem **2**
2.1 Laplace's discovery of Laplace's theorem 2
2.2 Direct and inverse probability 3
2.3 Laplacean and Gaussian least squares 4
2.4 Seeing p-hacking 5
2.5 The disappearance of Laplace's theorem in France 6

3 Practical certainty **7**
3.1 *La limite de l'écart* 8
3.2 Tables of the normal distribution 9
3.3 Edgeworthian *significance* 10
3.4 Enter the psychologists 11
3.5 Fisher's *Statistical Methods for Research Workers* 12
3.6 Seeing p-hacking 12
3.7 Who invented the name *p-value*? 15

4 Conclusion **15**

5 What they said: Fourier to Fisher **16**
5.1 Joseph Fourier, 1768–1830 17
5.2 Siméon-Denis Poisson, 1781–1840 18
5.3 Friedrich Wilhelm Bessel, 1784–1846 19
5.4 Thomas Galloway, 1796–1851 21
5.5 Augustin Cournot, 1801–1877 22
5.6 Jules Gavarret, 1809–1890 23
5.7 Wilhelm Lexis, 1837–1914 24
5.8 Francis Edgeworth, 1845–1926 24
5.9 Arthur Schuster, 1851–1934 25
5.10 Karl Pearson, 1857–1936 26
5.11 Gilbert Walker, 1868–1958 27
5.12 Arthur Bowley, 1869–1957 29
5.13 George Udny Yule, 1871–1951 30
5.14 Raymond Pearl, 1879–1940 30
5.15 Truman L. Kelley, 1884–1961 31
5.16 David Brunt, 1886–1965 32
5.17 Ronald A. Fisher, 1890–1962 33
5.18 Morris Viteles, 1898–1996 34

References **35**

1 Introduction

The history of statistical testing and its abuse is sometimes told as a 20th century English-language story. The originality of Ronald A. Fisher's treatment of statistical significance is emphasized, and his methods are contrasted with those of Jerzy Neyman and Egon S. Pearson. Subsequent confusion and misuse of statistical testing is sometimes blamed on textbooks that carelessly hybridize Fisher's theory with the Neyman-Pearson theory [45, Chapter 3], [26, 44, 54].

Fisher and Neyman did profoundly influence the theory and practice of statistical testing, and the interaction between them is a fascinating human and intellectual story [63]. But an exclusively English-language narrative obscures the continuity between early 20th century British work and earlier practice across Europe. Although different words were used, significance testing was already practiced in the first half of the 19th century. Abuses, including p-hacking and facile assumptions about the accuracy of models and the independence of observations, were also widespread. By the end of the 19th century these abuses had contributed to a widespread rejection of the use of probability in statistics.

These 19th-century antecedents are known to historians of statistics, but they are less salient to contemporary statisticians seeking to understand and remedy contemporary confusions and abuses. In this paper, I endeavor to make the broader picture better known by recounting how statistical testing and its abuses developed in the 19th century and how early 20th century British authors drew on existing theory and methods. I provide extensive quotations of some 19th and early 20th century authors, so that readers can judge for themselves the degree of continuity.

Section 2 recounts the trajectory of Laplace's large-sample statistical theory in the 19th century — how it was created, how it differed from Gauss's theory, how it was popularized, and how it was discredited. Laplace's theory was based on the central limit theorem, which he discovered in 1810 and vigorously promoted until his death in 1827. From this theorem, together with the method of least squares but without knowing probabilities for measurement errors, let alone prior probabilities, we can obtain large-sample confidence limits for unknown quantities and hence significance tests for whether the quantities are zero. Gauss's theory of least squares, which emphasized finding the best estimates rather than practical certainty for limits on errors, came to dominate work in astronomy and geodesy, but Laplace's large-sample theory was widely used in the human sciences once Joseph Fourier made it accessible to statisticians. The uses included p-hacking and inferences based on questionable assumptions.

The misuse of Laplace's theory so discredited it in France that it was practically forgotten there by the end of the 19th century. But it was still taught and used elsewhere. Section 3 sketches its transmission into Britain and the United states, from early expositions by De Morgan and Galloway, through the immensely influential work of Karl Pearson, to expositions in the early decades of the 20th century by Pearl and Kelley in the United States and by Yule and Fisher in Britain. I trace how limits for practical certainty were variously ex-

pressed in terms of probable errors, moduli, standard errors, and finally, in Fisher's 1925 *Statistical Methods for Research Workers*, tail probabilities. I also discuss the emergence of the terms *significance* and *p-value*. The use of *significant* as a technical term in statistics derives from its use by Francis Edgeworth in 1885, and some of the confusion associated with the word results from Edgeworth and Pearson using it in a way that is no longer readily understood. The term *p-value* appears in statistics in the 1930s and derives from a much older but less formal use of "the value of P".

Section 4 looks at lessons we might draw from this history. One obvious lesson is that p-hacking and other abuses of statistical testing need not be blamed on particular ways the subject was taught in the 20th century; the same abuses arose already in the 19th century. The difficulties go deeper, and remedies need to go deeper.

Section 5 documents the history set out here with some quotations from 19th- and early 20th-century authors. These authors represent only a fraction of the many who wrote on statistical estimation and testing during this period, and only a few of their words are quoted, but these words may help readers judge for themselves the extent to which the logic and the pitfalls of statistical testing did and did not change over the period.

2 Laplace's theorem

The sum of a large number of independent random variables is approximately normal. This theorem, with any of various regularity conditions that would make it true) is now called the *central limit theorem*, but there is justice in calling it *Laplace's theorem*. Pierre Simon de Laplace proved it in 1810, with his characteristic neglect of regularity conditions, and fully recognized its importance. It was named the central limit theorem (*zentraler Grenzwertsatz der Wahrscheinliehkeitsrechnung*) by Georg Pólya in 1920.

2.1 Laplace's discovery of Laplace's theorem

The integral of e^{-t^2} appeared in probability theory beginning in 1733, with Abraham De Moivre's asymptotic approximation for the sum of an interval of binomial probabilities. But the notion of a probability distribution with a density of the form

$$f(y) = \frac{h}{\sqrt{\pi}} e^{-h^2 y^2} \tag{1}$$

appeared only in 1809, when Carl Friedrich Gauss advanced it as a hypothetical distribution for errors of measurement, awkwardly justified by the argument that the resulting Bayesian posterior for the measured quantity has the arithmetic average of the measurements as its mode. Perhaps partly inspired by Gauss's result, but certainly also inspired by Fourier's emerging theory of the heat equation, Laplace soon afterwards arrived at his theorem, which gave (1)

as the distribution of the arithmetic average of a large number of independent variables, regardless of their individual distribution.[1]

Laplace first applied his theorem to a problem that had long concerned him, that of testing the hypothesis that planetary orbits were distributed by chance. But he quickly realized that he could also use it to justify estimation by least squares. This inspired both his monumental *Théorie analytique des probabilités* (1812) and his more verbal *Essai philosophique sur les probabilités* (1814). He also wrote to colleagues throughout Europe to explain the importance of the theorem, illustrating its power with examples.[2]

2.2 Direct and inverse probability

In the 1770s and 1780s, Laplace had developed the Bayesian theory of statistics, which Thomas Bayes and his friend Richard Price had studied only for the elementary case where we want to find the probability of an event from a sequence of independent trials. In the 1780s, Laplace had tried to use his Bayesian theory to develop a probabilistic theory of errors, but he was stymied by an inability to calculate the distribution of averages or other functions of more than a few observations.

Laplace's 1810 theorem solved the problem in a spectacular way. Not only could he now calculate the distribution of the average of many independent quantities; he could do so without even knowing the distribution of the individual quantities. The Bayesian analysis now also seemed less interesting. The very concentrated normal distribution of the average would dominate any prior distribution, so that the Bayesian analysis would give the same result as a direct argument in the style of Jacob Bernoulli, like the arguments Thomas Simpson and Daniel Bernoulli had earlier used in the theory of errors. Laplace did not disown his Bayesian theory, but he de-emphasized it in his *Théorie analytique*, and in the applications of his theorem he usually just stated the Bernoullian argument [50, 15]. This inattention to the difference between the Bayesian and Bernoullian forms of the argument continued in the Laplacean tradition throughout the 19th century and into the time of Karl Pearson. It allowed mathematical statisticians to communicate with each other easily, regardless of whether (like Cournot) they rejected the Bayesian argument or whether (like Edgeworth and Pearson) they saw the Bernoullian argument merely as a shortcut to getting a Bayesian answer with a roughly uniform prior.

Laplace did not give names to the two methods of argument that I have just called Bayesian and Bernoullian. In 1838 the British mathematician Augustus De Morgan called the two methods *direct probability* and *inverse probability*, respectively [23, 96].[3] These terms were widely used in English for at least a century. Since the 1970s, the two methods have usually been called *frequentist*

[1] Many authors have detailed the interaction between Laplace and Gauss [15, 32, 33, 46, 50, 80, 81, 82, 83].

[2] The scale of this correspondence has only recently become known, with the publication of Laplace's surviving correspondence by Roger Hahn [49]; for a summary, see [15, p. 414ff].

[3] Fourier had apparently used the corresponding French terms earlier in his teaching [21].

and *Bayesian*. Because *frequentism* also names a view about the meaning of probability, clarity might be served if, as I have done here, we used *Bernoullian* instead.[4]

2.3 Laplacean and Gaussian least squares

Gauss appreciated Laplace's theorem, but he and those who continued his work on least squares tempered their interest in it with a concern about systematic errors, outliers, and other practical problems. Moreover, Gauss eventually gave an alternative justification of least squares that applies when samples are small. This we now call the *Gauss-Markov theorem*: least-squares estimators have the least variance among all unbiased linear estimators when individual errors are unbiased and independent.

The result was two interacting but distinct schools of thought, one Gaussian, the other Laplacean. The Gaussian school soon became dominant in astronomy and geodesy.[5] The Laplacean school, which sought practical certainty from large samples, continued to find adherents in the human sciences.

We can catch a glimpse of how the two schools differed by looking at two tables of tail probabilities for the normal distribution, one published in 1816 by Gauss's disciple Friedrich Wilhelm Bessel, the second published in 1826 by Joseph Fourier, in an exposition of Laplace's theory. Although Christian Kramp had published a table of values of the integral of e^{-t^2} in 1799, Bessel's and Fourier's were the first tables of the normal distribution ever published.[6]

Bessel's table appears in a passage, reproduced in translation in §5.3 below, in an article on the orbit of Olber's comet. The passage explains why the different equations of condition in a least-squares computation should be weighted differently. To this end, he introduced the *probable error* of a continuous variable, which had not been previously defined and used. As the reader may recall, this is the number r such that $P(|X-\mu| < r) = P(|X-\mu| > r)$, where X is the variable and μ is X's mean. When X is normal, $r \approx 0.6745\sigma$, where σ is the standard deviation. For seven different values of α, Bessel's table gives the odds that a normally distributed variable falls more than α probable errors from its

[4] By using *Bernoullian*, I am following examples set by Francis Edgeworth [30], Richard von Mises [92, p. 5], A. P. Dempster [25], and Ian Hacking [48].

[5] The story of the triumph of the Gaussian theory in geodesy has been told in an enlightening dissertation by Marie-Françoise Jozeau [56].

[6] In 1783, in the course of a Bayesian analysis of Bernoulli's binomial problem, Laplace gave a method for calculating values of the incomplete integral of e^{-t^2} and mentioned that a table of these values would be useful [46, p. 79]. The first person to publish such a table was Christian Kramp, in 1799 [60]. Kramp gave values of $\int_\tau^\infty e^{-t^2}\,dt$ for τ from 0.00 to 3.00, in intervals of 0.01. Kramp calculated his table to facilitate the study of refraction, not to facilitate the calculation of probabilities, and because $\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}$, the entries in his table are not probabilities. But we obtain probabilities simply by dividing by $\sqrt{\pi}$. Bessel used Kramp's table to calculate his.

mean:

$\alpha = 1$	$\cdots\cdots$	$1 : 1$
$\alpha = 1.25$	$\cdots\cdots$	$1 : 1.505$
$\alpha = 1.5$	$\cdots\cdots$	$1 : 2.209$
$\alpha = 1.75$	$\cdots\cdots$	$1 : 3.204$
$\alpha = 2$	$\cdots\cdots$	$1 : 4.638$
$\alpha = 3$	$\cdots\cdots$	$1 : 30.51$
$\alpha = 4$	$\cdots\cdots$	$1 : 142.36$

Thus the probability of the variable being more than 4 probable errors from its mean, for example, is approximately $1/(1 + 142.36)$ or 0.007.

Fourier's table appeared in a memoir on the use of probability he included in the 1826 report of the statistics bureau of Paris and its surrounding region. Instead of the probable error, Fourier used as his measure of dispersion the quantity $\sqrt{2}\sigma$, which was called the *modulus* by some later authors. The modulus is a natural choice because the density for a normal variable with mean 0 and modulus 1 is proportional to e^{-t^2}. One modulus is approximately two probable errors. For each of 5 small probabilities (we might call them *significance levels* today), Fourier gave the number ∂ such that a normal random variable will be more than ∂ moduli from its mean with that probability.

∂	P
0.47708	$\frac{1}{2}$
1.38591	$\frac{1}{20}$
1.98495	$\frac{1}{200}$
2.46130	$\frac{1}{2000}$
2.86783	$\frac{1}{20000}$

There is only a probability of 1 in 20,000, for example, a normal variable will be more than about 2.86783 moduli (or about 4.0557 standard deviations) from its mean.

Fourier's table differs in Bessel's in two important ways. First, it uses round numbers for the probabilities rather than for the distance from the mean. It gives the distance from the mean corresponding to a significance level the reader might have in mind. Second, it includes much more extreme values. Whereas Bessel's table extends to only 4 probable errors, Fourier's extends to 2.87 moduli, equivalent to about 6 probable errors and corresponding to a probability two orders of magnitude smaller. Fourier was interested in identifying limits within which we can be practically certain the deviation from the mean will fall.

Fourier decided that 3 moduli is enough for practical certainty. His example was followed by a number of other 19th-century authors.

2.4 Seeing p-hacking

The word "statistics" (*Statistik* in German and *Statistique* in French) was coined to refer to information monarchs might want to know about their kingdoms'

population and wealth. The theory of errors, conceived as a tool for astronomy and geodesy, did not fall under this rubric at the beginning of the 19th century. But Fourier, in his reports for the Paris statistics bureau, applied the Laplacean theory to statistics. In his 1829 report, for example, he gave error limits for the average age, in Paris during the 18th century, of men and women when their first son was born [40, Table 64].

Such applications soon led to what we now call significance testing. In 1824, Siméon-Dénis Poisson, who became the leading expert on Laplace's theory after Laplace's death in 1827 and Fourier's in 1830, published a note observing that the ratio of boys to girls was smaller for illegitimate births than for legitimate births [73]. In 1830, he applied Laplace's theory to decide whether this and other variations in the ratio of boys to girls could have happened by chance [74]. He concluded that the difference was real, and he also found that the ratio was smaller in Paris than in the rest of France, for both legitimate and illegitimate births.

Were Fourier's and Poisson's arguments valid? Were the 505 men and 486 women for whom Fourier was able to find the needed data a random sample? In what sense were the approximately ten million births in France in the decade from 1817 to 1826, which Poisson studied, a random sample from a larger population? Some French statisticians thought Fourier's and Poisson's calculations were ridiculous. Among them was André-Michel Guerry, the statistician who published in 1833 a brilliant summary of the French census commissioned by the Academy of Sciences [47, 41].

Augustin Cournot, twenty years younger than Poisson, was himself a proponent and brilliant expositor of Laplace's theory, but he perceived another problem with its application to the census by Fourier, Poisson, and statisticians who had imitated them. The problem is what we now call *multiple testing* or *p-hacking*. In 1843, safely after Poisson's death, Cournot published his own elegant book on probability, *Exposition de la théorie des chances et des probabilités* [19]. In a passage reproduced in translation in §5.5, Cournot explained that statisticians had been looking for differences in the sex ratio for all sorts of ways of dividing the population: legitimate and illegitimate, by season, by birth order, etc. As the public could not see the extent of the search, they could not evaluate whether a particular apparently remarkable difference might arise by chance.

2.5 The disappearance of Laplace's theorem in France

The uncertainty measured by a p-value may be the least of the uncertainties when we are working with actual data. One of Laplace's favorite examples of the power of his theorem was his estimation of Jupiter's mass relative to the Sun. Combining all the relevant measurements that had been made by that time, he announced bounds on this ratio, bounds on which he claimed one could bet a million to one. Five years after his death, the British astronomer George Biddell Airy showed that the true ratio lay well outside these million to one bounds. Laplace's supreme confidence, whether in his model, his data, or his

calculations, had been misplaced [15, p. 492].

This was only the beginning of the discredit into which Laplace's theory fell. Though a champion of Laplace's theorem, Cournot ridiculed Laplace's Bayesian argument, emphasizing that for many questions it is only possible to justify non-numerical "philosophical probabilities" [19]. Cournot's friend Jules-Irénée Bienaymé further developed Laplace's theory but spent most of his energies combating faulty applications [14, note 27],[52]. The nineteenth century saw an unprecedented flood of data, and many of its collectors and users concluded that it could speak for itself; probability was not needed [48]. By the middle of the century, geodesy, a field dominated by the French before and during Laplace's heyday, had abandoned Laplace's methods, turning instead to the methods developed by Gauss and is followers [56]. Mathematicians and philosophers found many other problems in the Laplacean theory [57, 76]. By the end of the century, the most prominent mathematician in France, Joseph Bertrand, would ridicule Laplace's entire undertaking as a delusion [7]. Its disappearance from French mathematics was so thorough that the leading French mathematicians who worked on the central limit theorem in the early 20th century, Borel, Fréchet, and Lévy, were unaware that Laplace had first proven the theorem until this was brought to their attention by their foreign colleagues.

3 Practical certainty

This section looks at how Laplace's theory and Fourier's criterion for practical certainty evolved in the 19th and into the 20th century. Many of the authors cited here are quoted at greater length in §5.

Throughout this period, authors on the theory of errors, practically without exception, can be classified as either Gaussian or Laplacean. Both schools taught the use of least squares to obtain estimates and used the normal distribution to compute probabilities of error. But the Gaussian authors, considering their models and Laplace's asymptotics too unreliable for extreme conclusions, did not talk about practical certainty, whereas the Laplacean authors usually tried to specify, in one way or another, an interval around the least-squares estimate that will be practically certain to include the true value of the quantity being estimated.

Today we measure a random variable's distance from its mean in terms of its *standard deviation*, and we sometimes call the standard deviation of an estimator its *standard error*. But these English terms appeared only at the end of the 19th century; Karl Pearson introduced *standard deviation* in 1894 [71], and George Udny Yule introduced *standard error* in 1897 [100]. Earlier writers had other names for the standard deviation, but they more often used the modulus or the probable error; see Table 1. The probable error was widely still used in the first decades of the 20th century.[7]

[7]Helen Walker's history, written in 1929 [96], is still a good reference for how various authors used and named the various measures.

Table 1: This table, relating different measures of dispersion for the normal distribution, is taken from page 24 of George Biddell Airy's 1861 book [1]. Airy's *error of mean square* is our *standard deviation*. His *mean error* is the mean of the absolute value.

	Modulus.	Mean Error.	Error of Mean Square.	Probable Error.
In terms of Modulus	1.000000	0.564189	0.707107	0:476948
In terms of Mean Error	1.7724.54	1.000000	1.253314	0.845369
In terms of Error of Mean Square	1.414214	0.797885	1.000000	0.674506
In terms of Probable Error	2.096665	1.182916	1.482567	1.000000

As we have seen, Fourier considered it practically certain that an estimate based on many observations would be within three moduli of the quantity being estimated. In this section, I sketch how this criterion evolved during the 19th century, how Edgeworth translated practical certainty, when a difference between two quantities was being estimated, into a notion of *significance*, and how this notion evolved (or degenerated, we might say). I also discuss the origin of the name *p-value*.

3.1 *La limite de l'écart*

Equating very high probability with moral certainty is ancient. The 16th-century Jesuit Luis Molina even applied it to games of chance [59, pp. 402–403]. But Molina and his fellow scholastics did not gauge degrees of probability numerically. It was Jacob Bernoulli's *Ars conjectandi*, published in 1713, that brought moral certainty into the context of a mathematical theory of probability modeled after calculations of chances for dice. Bernoulli did not advocate for a particular level of probability that would suffice for moral certainty; he thought 0.99 or 0.999 might do but suggested that the level be set by magistrates. So far as I know, Laplace also never specified a level of probability that would suffice for certainty. Fourier may have been the first to do so. As we have seen, Fourier considered a statement certain if the probability of its being wrong is only one in 20,000, and for an interval based on Laplace's theorem, this corresponds to 2.87 moduli, a number that Fourier rounded to 3.

Later authors sometimes set a very exigent standard for certainty in theoretical discussions but then relaxed it in applications. Siméon-Denis Poisson, in his 1837 book on probability, first mentioned 4 or 5 moduli but relaxed this to 2 moduli when he turned to examples, even writing at one point that the

probability of an estimate being within 2 moduli, 0.9953, is very close to certainty; see §5.2. Jules Gavarret, in his 1840 book on medical statistics, cited Poisson's authority in deciding that 2 moduli gives "a probability after which any therapeutic fact can and should be accepted without dispute"; see §5.6.

Laplace's theory was brought into English in the 1830s by Augustus De Morgan and Thomas Galloway, both of whom published books containing proofs of Laplace's theorem. De Morgan's book appeared in 1837 [22] and Galloway's in 1839 [42]. Galloway's was probably more influential, because whereas De Morgan followed Laplace directly, Galloway followed Poisson's simplified and clearer proof. Both used the modulus. So far as I know, De Morgan did not single out a particular number of moduli, but Galloway did mention 3 moduli; see §§5.4.

Cournot, in his 1843 book [19], followed Fourier in treating the probability of 1 in 20,000, corresponding to ± 2.87 moduli, as practical certainty, but he did not round 2.87 to 3. In §35, he recommended that this limit be held in mind not only because it corresponds to a value of P equal to 1 in 20,000 but also because it comes very close to 6 probable errors. In §69, he called 2.87 moduli the "limite extrême de l'écart" — the extreme limit of the deviation. Unlike Poisson and most of the earlier authors, Cournot rejected Laplace's Bayesian theory, accepting only a Bernoullian interpretation of the bounds given by Laplace's mathematics. This makes his 1843 book very close to mathematical statistics as it was practiced in the middle of the 20th century; the basic concepts are all there.

None of the authors I have been discussing, from Fourier to Cournot, used *modulus* for the quantity I have been calling by this name. It seems that this usage first appeared in an 1861 book on the theory of errors by George Biddell Airy, the British Astronomer Royal from 1835 to 1881. As customary in the Gaussian tradition, Airy did not discuss practical certainty. When discussing the law of error, on page 17, he did observe that "after the Magnitude of Error amounts to $2.0 \times$ Modulus, the Frequency of Error becomes practically insensible", but here he was referring to the density of the normal distribution, not to its tail probabilities.

Wilhelm Lexis, a prominent economist and statistician, kept the Laplacean tradition alive in Germany. Although he taught and wrote in German, Lexis had studied for ten years in France, and the Laplacean aspiration to find practical certainty by multiplying observations was natural to his area of study. In his introduction to population statistics *Einleitung in die Theorie der Bevölkerungsstatistik* [64], published in 1875, Lexis repeatedly used 3 moduli as his level for practical certainty; see §5.7.

3.2 Tables of the normal distribution

As mentioned in §2.3, Bessel had calculated his small table of normal probabilities using Kramp's table of the incomplete integral of e^{-t^2}. By the 1830s, it became common for books on probability to provide much more extensive tables of normal probabilities. The first such table, also calculated from Kramp's, was

given by the German astronomer Johann Franz Encke, a follower of Gauss, in an 1832 article on least squares [31]. Encke tabulated the values of

$$\frac{2}{\sqrt{\pi}} \int_0^\tau e^{-t^2} dt \tag{2}$$

to seven decimal places for values of τ (the number of moduli) from 0.00 to 2.00, in intervals of 0.01. He also gave a similar table in terms of the probable error. Encke's article was translated into English and printed with its tables in *Taylor's Scientific Memoirs*, 1841, Vol. II, pp. 317–669 [65, p. 180]. In 1838 [23], De Morgan extended Encke's tables from 2 to 3 moduli. Galloway, in his 1939 book, also gave a table going up to 3 moduli.

In his 1843 book [19], Cournot gave a table of (2) for values of τ from 0.00 to 3.00, with a variable number of decimal places. He also gave the value for $\tau = 3.00$ to 10 places, for $\tau = 4.00$ to 13 places, and for $\tau = 5.00$ to 17 places.

3.3 Edgeworthian *significance*

A British economist and statistician, Edgeworth was by all accounts the first to use the English word *significant* in connection with statistical testing. He did this in a paper he read at the Jubilee meeting of the Statistical Society of London in 1885 [27]. The substance of the paper, as Edgeworth conceded in the discussion after the reading, was largely borrowed from Lexis.[8] Edgeworth's originality lie in translation; where Lexis discerned a real difference being *praktisch gewiss* (practically certain), Edgeworth discerned an apparent difference *signifying* a real difference; see §5.8.

An observed difference is significant in Edgeworth's sense if and only if two conditions are satisfied: there is a real difference, and the observed difference is large enough to suggest it. Either both conditions are satisfied, or not. We may not be sure. So when using the word in Edgeworth's sense, we may say that a difference is "perhaps significant", "likely significant", "probably significant", "definitely significant", or "certainly significant". We may also say flatly that a difference is "not significant"; if it is less than a probable error from zero, then it does not signify a real difference even if there is one. We will not say that a difference is "barely significant" or "just significant", because if it does not definitely signify, then it may not signify at all. Nor will we use phrases like "highly significant", "very significant", and "more significant". It is the likelihood of signifying, not signifying itself, that is a matter of degree.

Edgeworth's use of *significant* seems very strained to this speaker of American English. Perhaps it was natural for Edgeworth's social class in his time and place, or perhaps it was merely one of his quirky turns of phrase.[9] But British statisticians understood it and adopted it. They used it for thirty years

[8] Had Edgeworth studied mathematics at university, as Karl Pearson did, he might have learned the Laplacean theory from Galloway's book. But as he had been trained as a classicist and was self-taught in mathematics, it would have been natural for him to seek the latest wisdom from a German authority.

[9] Steve Stigler discusses Edgeworth's odd style on pp. 95–97 of [85].

or more. As we will see in §5, it persisted into the 1920s in Karl Pearson's *Biometrika*, and the United States biometrician Raymond Pearl, a student of Pearson's, explained it very clearly in a book he published in 1923.

Pearson diverged from Edgeworth in one respect; he measured the deviation of an estimate from the quantity estimated using the probable error rather than the modulus. The most common practice among the biometricians, reported by Pearl and followed by *Biometrika*, was that 3 probable errors (about two standard deviations) indicated likely significance and 6 probable errors (about 4 standard deviations) indicated definite significance.

3.4 Enter the psychologists

Edgeworthian significance disappeared in the 1920s. The word remained, but the meaning shifted. This was a gradual process, unnoticed by many. It seems that many statisticians outside Pearson's international circle of biometricians picked up the word *significance* without grasping its Edgeworthian interpretation, which must have been as unexpected for many ears then as it is for mine now.

One field where we can see this happening is psychology. As Steve Stigler has noted [84], psychologists had begun using mathematical statistics in the 1860s and had developed their own methods long before Pearson created biometry. In the late 1910s and late 1920s, we see young United States psychologists using *significant* and *significance* in relatively vague and certainly non-Edgeworthian ways. In 1922, for example, we find Morris Viteles, who later became very prominent in industrial and organizational psychology, writing that test results were "greatly significant" and "highly significant". He may also have been the first to use the phrase "level of significance" in connection with statistical testing; see §5.18.

The first use of the non-Edgeworthian term *statistical significance* in its modern sense that I have found is in a 1916 article by another young and eventually very prominent U.S. psychologist, Edwin Boring [10, p. 315]. Boring understood the vocabulary of the British biometricians reasonably well, but he soon concluded that the assumptions underlying the Laplacean method (e.g., independence of observations and a common meaning for a parameter in different individuals or groups) were usually not satisfied in his work. His most often cited criticism of the method was a 1919 article entitled "Mathematical vs. scientific significance" [11]. He carried on a years-long debate on the use of statistics in psychology with Truman Kelley, at the time one of psychology's most prominent experts on statistical methods [87].

In his 1923 textbook *Statistical Method* [58] Kelley wrote, "If these two relationships do not exactly hold, the significance of the discrepancy can be determined by formulas giving probable errors..." (p. 99). This statement, though vague, is certainly not Edgeworthian. On page 102, at the beginning of a lengthy passage quoted in §5.15, Kelley made a similar statement: "The significance of any measure is to be judged by comparison with its probable error." This passage is also of interest because it shows how Kelley was shifting

his readers from the probable error to the standard deviation, and because it shows how to perform a one-sided test.

Shortly before Kelley competed his book, he had spent a sabbatical year in London with Pearson.[10] Perhaps he also met Fisher at that time. When he sent Fisher a copy of the text, Fisher responded that it was "quite the most useful and comprehensive book of the kind yet written" (1924, January 12) [87, p. 560].

3.5 Fisher's *Statistical Methods for Research Workers*

In 1925, Fisher published his own statistics manual, his celebrated *Statistical Methods for Research Workers* [36]. Features of the book relevant to this paper are documented in §5.17. The most salient are these:

- The words *significant* and *significance* are prominent, but their Edgeworthian meaning has disappeared in favor of a meaning that allows degrees of significance.
- Probable error has given way completely to standard deviation.
- The main purpose of the book was to provide tables for the many distributions that Fisher had studied, including Student's t and the distribution of the correlation coefficient, and to teach research workers how to use these tables. Because these distributions were not normal and sometimes not symmetric, "significance" could not be defined in terms of standard deviations. Fisher defined it instead in the only possible way, in terms of tail probabilities. Two standard deviations was replaced by 5% [86].

The third of these features appears to be the most original. Yule had used the standard deviation for judging the likelihood of Edgeworthian significance in 1911 (see §5.13), and we just saw the prominence of the standard deviation in Kelley's 1924 book.

Fisher's tone does not suggest that he has deliberated about rejecting the Edgeworthian meaning of *significant*. He was never part of Pearson's circle, and by 1925 he was certainly not looking to Pearson's work for guidance. It seems likely that he drew his understanding of significance tests less from a careful reading of *Biometrika* than from the usage of the U.S. psychologists or others distant from Pearson. Perhaps this included colleagues at the agricultural experiment station at Rothamsted, where he had already been working for five years.

3.6 Seeing p-hacking

The British statistician's were slow to come to grips, at least in print, with the problem of p-hacking. Cournot's critique of p-hacking in English was mentioned

[10] Personal communication from Lawrence Hubert, who has examined the Kelley archive at Harvard. See also [5].

by Francis Edgeworth in 1887 [28] and John Venn in 1888 [90, third edition, pp. 338–339]; they reported that they did not understand it.

To my knowledge, the problem of p-hacking was first acknowledged by the British in the context of searches for cycles in meteorological and economic data. The notion that such cycles might be discovered and used for prediction was very popular at the end of the 19th- and beginning of the 20th centuries, when respected scholars even conjectured a possible connection between sunspots and business cycles: cycles in sunspots might cause cycles in the weather, hence cycles in agricultural production and other economic measurements [66].

A statistical test for whether an apparent cycle in a time series is real was suggested in 1898 by the British physicist Arthur Schuster, in the article in which he introduced the name *periodogram* for a graph showing the estimated intensity of different frequencies in the series' Fourier transform [77]. Schuster eventually explained his test in a very simple way: the probability that a particular estimated intensity will be h or more times as large as its expected value is e^{-h}. (Being the sum of the squares of two normally distributed variables, it will have an exponential distribution; see [78] and §5.9.)

The enterprise of looking for cycles in data is, by definition, a search through the data for something remarkable, just the kind of search for statistical significance criticized by Cournot, derided in the 1970s as *data mining* and derided today as *p-hacking*. One of the first, perhaps the very first, to point out how misleading it can be was Gilbert T. Walker, a physicist working as a meterologist in India ; see §5.11. Walker's first critique was published in India, in 1914 [93]. It was also pointed out in 1919, by the physicist F. J. W. Whipple in the discussion of paper on cycles in the Greenwich temperature records, read to Royal Meteorological Society by David Brunt in 1919 [17]; see §5.16.

The British biometricians may have overlooked Walker's 1914 critique and Whipple's 1919 comments. But there is no doubt that the p-hacking issue raised by Schuster's test came to their attention in 1922, after the prominent civil servant and scholar William Beveridge read a paper to Royal Statistical Society on cycles in rainfall and wheat prices. Beveridge read his paper, in which he used methods he asserted to agree with Schuster's test, on April 25 of that year [9]. Yule was one of the discussants. None of the Society members who commented on the paper were fully convinced by Beveridge's conclusions, but he evidently stirred their interest. In its issues for August 19 and August 26, (vol. 110, pp. 265, 289), *Nature* reported that the fall meeting of the British Association at Hull would include a special session on "Weather Cycles in Relation to Agriculture and Industrial Fluctuations", to be held on September 7. The session was to be sponsored jointly by three sections of the association, Economic Sciences and Statistics, Mathematics and Physics, and Agriculture, and it was to feature discussion by Beveridge, Fisher, and Yule. In its December 30 issue (vol. 110, pp. 889–890), *Nature* summarized the discussion. Beveridge made his case for relating periodicities in the weather to those of wheat prices. Yule offered a mild defense of Beveridge against discussants who found his case totally implausible. Fisher questioned how strongly peiodicities in the weather would be directly related to periodicities in production, reporting that the total

rainfall at Rothamsted accounted for a relatively small amount of the variation in production.

Walker enters the story before December 30, however. In its October 14 issue (vol. 110, pp. 511–512), *Nature* published a letter to the editor from Walker, along with a response by Beveridge [94]. Walker makes the point that he had made already in 1914: if a statistician is going to test the largest estimated intensity from a periodogram that shows estimated intensities for k different frequencies, the probability of this greatest estimated intensity being h times the expected value for any given intensity is $e^{-h/k}$ rather than e^{-h}. With this adjustment, Walker did not find Beveridge's cycles to be convincing. Beveridge conceded the conceptual point but adjusted the assumptions in Walker's analysis so that his conclusions emerged intact nonetheless.

What was Fisher to say about this? The whole periodogram story being a mess, he probably did not have much to contribute, and nothing would have been gained by entering a controversy between two such powerful individuals. In due time, however, Fisher did make some effort to deal with the problem of selecting the most significant test. In 1925, when he finally published his thoughts on the Rothamsted data on rainfall and crop production [35, p. 94–95], he derived the adjustment required when one variable is chosen from several for a regression.[11] In 1929, he showed how Schuster's and Walker's criteria for testing intensities in a periodogram can be adjusted to account for the fact that the variance must be calculated from the data [37].

There are, however, no cautions about p-hacking in the 1925 edition of *Statistical Methods for Research Workers*. Why did Fisher omit the topic? The obvious answer is that the topic is impossibly difficult for a book that offers research workers with limited mathematical training recipes for statistical testing. Perhaps too difficult for anyone. As Beveridge's response to Walker's 1922 letter suggests, adjusting p-values for selection is often topic for debate, not for recipes.

In the preface to the sixth edition of *Statistical Methods for Research Workers*, published in 1936 (p. xii), we finally see a recipe for dealing with selection, gingerly offered:

> I am indebted to Dr W. E. Deming for the extension of the table of z to the 0.1 per cent. level of significance. Such high levels of significance are especially useful when the test we make is the most favourable out of a number which *a priori* might equally well have been chosen. Colcord and L. S. Deming have published a slightly full Table in the *Indian Journal of Statistics* (1936).

In this same preface, Fisher refutes critics who had asked why he did not provide mathematical derivations for his recipes in the book. The book, he explains, is for research workers, not people doing the theory of statistics.[12]

[11] I am indebted to John Aldrich for calling my attention to this article.

[12] Again, I am indebted to John Aldrich for calling my attention to this preface.

3.7 Who invented the name *p-value*?

The answer, it seems, is that no one did. The term simply evolved from the casual use of the letter P to denote the probability that an estimated quantity or difference will fall inside or outside given limits.

We already see this use of P, in majuscule and roman, not in mathematical font, in Fourier, Poisson, Gavarret, and Cournot. Beginning at least in his 1900 article on χ-square [72], Karl Pearson similarly wrote P for the probability of a result more extreme than the observed value of a test statistic and referred to it as "the value of P". R. A. Fisher followed Pearson's example throughout his career [34, 38].

By the late 1930s, some authors had casually turned *value of P* into *P-value*. The earliest examples I have seen are in articles by John Wishart [98, p. 304] and his associate H. O. Hirschfeld [53, p. 68][13] and in a book by W. Edwards Deming [24, p. 30]. The expression *P-value* was not widely used in the social sciences before 1970; we do not see it in any of the articles in [67].

Today the use of *P-value* is widespread, but there is no consensus on the font for the letter P. We see lower and upper case, italicized and roman, text and mathematical font, with and without the hyphen.

4 Conclusion

In 1949 [99, p. 90], the accomplished British statistician John Wishart wrote,

> If one were asked to say what has been the distinctively British contribution to the theories of probability and mathematical statistics during the present century, the answer, I fancy, would be found, not so much in the formulation of a satisfactory theory of probability, including the nature of inference, as in the fashioning of significance test tools to guide the practical experimenter.

The history reviewed in this paper confirms Wishart's judgement. The notions of testing and estimation used in mathematical statistics even today were in place already in the 19th century.

Unfortunately, there is also a parallel with respect to the misuse and abuse of these basic concepts. The inappropriate models and inferences that led to the collapse of the Laplacean tradition in France in the second half of the 19th century are rampant today, inspiring loss of confidence and hand-wringing. We see a blizzard of proposals for how to correct these problems. Some propose to shift the level required for calling attention to a p-value (replace Fisher's 0.05 with Poisson's and Gavarret's 0.005); some propose to change the words we use (eliminate *p-value* or *statistical significance*); others propose various more or less complicated ways of complementing the statement of a p-value.

[13] Hirschfeld later anglicized his name to H. O. Hartley

In March 2019, 854 statisticians published an editorial in *Nature* entitled "Retire statistical significance" [2]. What should replace it? How is a scientist or journalist inexpert in Laplacean methods to interpret a p-value?

Here is a fanciful thought experiment. Suppose we all (all the teachers of statistics and textbook writers) reach a consensus to return to Karl Pearson's version of Edgeworthian significance. And suppose we signal this change by replacing *significant* with *signifying*. A p-value of 0.05 (three probable errors or two standard deviations) is likely to signify; a tail probability of 0.00006 (six probable errors or four standard deviations) definitely signifies. What problems would this pose? The obvious problem is the "likely" in "likely to signify". Are we willing to give it a Bayesian interpretation with a uniform prior, as Edgeworth did? A Bernoullian interpretation as Cournot and Neyman would? Here the fancied consensus splinters.

In my view, we cannot stanch the abuse of Laplacean methods by changing their vocabulary. We need a more fundamental change in order to convey both the conclusion of a statistical study and its uncertainty. My own proposal for a fundamental change is to shift from the Bayesian statement that

> conclusion A is likely assuming model B

and the Bernoullian statement that

> the observations are unlikely unless A and B are true

to a statement about a bet:

> A bet that was fair according to B has paid off handsomely (multiplied the money it risked) unless A is true.

One advantage of emphasizing betting is the salience, even to the untutored, of the possibility that even the longest shot can succeed by chance. As I argue in [79], it can also help us give an honest account of searches that can so easily degenerate into p-hacking. A conventional probability account for a search, whether Bernoullian or Bayesian, requires the specification of a complete strategy for the search. What would you have done if the first test had come out in various ways, etc.? When we test by trying to multiply the money we risk, no fixed strategy is required; we can change direction opportunistically so long as we our next bet always risks only the net capital resulting from the last bet.

5 What they said: Fourier to Fisher

In this section, I provide more detail about how a number of 19th and early 20th century authors wrote about Laplace's theorem, error limits, and practical certainty. Aside from Bessel, these authors worked in the Laplacean tradition. I consider them in order of their birth.

I have chosen these authors, somewhat arbitrarily, to illustrate how the Laplacean tradition evolved as it was transmitted into English. A more com-

prehensive review would include authors working in a wider variety of applications and in other European countries, including Italy, Russia, Belgium, and Denmark.

In the translations, I have shifted the authors' notation to current practice in American English, italicizing symbols, indicating limits of integration with subscript and superscript, writing 0.9953 instead of 0,9953 or 0·9953, writing h^2 instead of hh, etc.

5.1 Joseph Fourier, 1768–1830

Joseph Fourier is most renowned for his mathematical analysis of the diffusion of heat, but he was also a revolutionary and a politician, an impassioned participant in the French revolution and an administrator under Napoleon. After Napoleon's final defeat and the return of a royalist regime in 1815, Fourier was briefly left with neither a position nor a pension, but the royalist Chabrol de Volvic, who had been his student, rescued him from impoverishment with an appointment to the census bureau of the Paris region [68]. The appointment left him time for his mathematical research, but he faithfully attended to his duties at the census, issuing masterful reports in 1821, 1823, 1826, and 1829. Fourier's name was not included in the reports, which were issued under the auspices of the census bureau, but there is no doubt that Fourier at least edited them all and wrote the mathematical memoirs that appear at the beginning of the 1826 and 1829 ones [14, p. 198].

Given independent observations $y_1, \ldots, y_m$ and their average $\overline{y}$, Fourier estimated what I am calling the modulus by

$$g := \sqrt{\frac{2}{m}\left(\frac{\sum_{i=1}^{m} y_i^2}{m} - \overline{y}^2\right)}$$

This is consistent with modern practice; the modulus is $\sqrt{2}$ times $\overline{y}$'s standard deviation, and g is $\sqrt{2}$ times the maximum likelihood estimate of this standard deviation.

The passage from Fourier's 1826 memoir translated here includes a table of significance levels, which may look more familiar when we add a third column translating units of g into units of $\overline{y}$'s standard error $g/\sqrt{2}$:

units of g	P	units of $g/\sqrt{2}$
0.47708	$\frac{1}{2}$	0.67
1.38591	$\frac{1}{20}$	1.96
1.98495	$\frac{1}{200}$	2.81
2.46130	$\frac{1}{2000}$	3.48
2.86783	$\frac{1}{20000}$	4.06

Fourier wrote that it is "a 19 out of 20 bet" that the error will not exceed $1.38591g$. This is the familiar 95% confidence interval obtained using 1.96 standard errors. He also concludes that the error certainly will not exceed $3g$.

The following passage constitutes §XI of the 1826 memoir [39, pp. xxi–xxii].

Fourier in English

To complete this discussion, we must find the probability that H, the quantity sought, is between proposed limits A + D and A − D. Here A is the average result we have found, H is the fixed value that an infinite number of observations would give, and D is a proposed quantity that we add to or subtract from the value A. The following table gives the probability P of a positive or negative error greater than D; this quantity D is the product of g and a proposed factor ∂.

∂	P
0.47708	$\frac{1}{2}$
1.38591	$\frac{1}{20}$
1.98495	$\frac{1}{200}$
2.46130	$\frac{1}{2000}$
2.86783	$\frac{1}{20000}$

Each number in the P column tells the probability that the exact value H, the object of the research, is between $A + g\partial$ and $A - g\partial$. Here A is the average result of a large number m of particular values $a, b, c, d, \ldots, n$; ∂ is a given factor; g is the square root of the quotient found by dividing by m twice the difference between the average of the squares $a^2, b^2, c^2, d^2, \ldots, n^2$ and the square A^2 of the average result. We see from the table that the probability of an error greater than the product of g and 0.47708, i.e. greater than about half of g, is $\frac{1}{2}$. It is a 50–50 or 1 out of 2 bet that the error committed will not exceed the product of g and 0.47708, and we can bet just as much that the error will exceed this product.

The probability of an error greater than the product of g and 1.38591 is much less; it is only $\frac{1}{20}$. It is a 19 out of 20 bet that the error of the average result will not exceed this second product.

The probability of an even greater error becomes extremely small as the factor ∂ increases. It is only $\frac{1}{200}$ when ∂ approaches 2. The probability then falls below $\frac{1}{2000}$. Finally one can bet much more than twenty thousand to one that the error of the average result will be less than triple the value found for g. So in the example cited in Article VI, where the average result was 6, we can consider it certain that the value 6 is not wrong by a quantity three times the fraction 0.082 that the rule gave for the value of g.

The quantity sought, H, is therefore between $6 - 0.246$ and $6 + 0.246$.

5.2 Siméon-Denis Poisson, 1781–1840

Poisson advanced Laplace's theory substantially. Beginning in the 1820s, he simplified the proof of Laplace's theorem, making it accessible to many more mathematicians [50, §17.3]. In 1830, he gave straightforward instructions for

calculating limits of practical certainty for the difference between two estimated quantities [74].[14] Finally, in 1837, he pulled together all his theoretical and applied results on probability in an impressive treatise *Recherches sur la probabilité des jugements* [75].

Like Fourier, Poisson discussed limits in terms of numbers of moduli. When writing theory, he required 3, 4, or even 5 moduli for practical certainty [75, §§80, 87, and 96]. But when analyzing data, he calculated less exigent limits. In §89, when dealing with Buffon's data, he gave limits and odds corresponding to 2 moduli. In §111, he reduced this to 1.92 moduli, corresponding to a bet at odds 150 to 1.

An example of a theoretical discussion is found in §87, where Poisson considered the problem of testing whether the unknown probability of an event E has changed between the time two samples are taken. There are μ observations in the first sample; E happens in n of them, and its opposite $\mathrm{F} = \mathrm{E}^c$ happens in $m = \mu - n$ of them. For the second sample, he uses analogous symbols μ', n', and m'. He gives formulas, under the assumption that unknown probability has not changed, for the estimated modulus of the difference $\frac{m'}{\mu'} - \frac{m}{\mu}$ and a formula the probability that this difference will be within u moduli of 0. Then he writes,

> So if we had chosen a number like three or four for u, making the probability $\tilde{\omega}$ very close to certainty (n° 80), and yet observation gives values for $\frac{m'}{\mu'} - \frac{m}{\mu}$ or $\frac{n'}{\mu'} - \frac{n}{\mu}$ that are substantially outside these limits, we will have grounds to conclude, with very high probability, that the unknown probabilities of the events E and F have changed in the interval between the two series of trials, or even during the trials.

The closest Poisson came to identifying ± 2 moduli with practical certainty may have been in §135 of the book, where he considered the 42,300 criminal trials in France during the years 1825 through 1830. The defendant was convicted in 25,777 of these trials. So his estimate of the average probability of conviction, which he called R_5, was $(42300/25777) \approx 0.6094$. His estimate of the modulus was 0.00335. Then he stated that if one uses 2 moduli,

> . . . we will also have
>
> $$\mathrm{P} = 0.9953,$$
>
> for the probability, very close to certainty, that the unknown R_5 and the fraction 0.6094 will not differ from each other by more than 0.0067.

5.3 Friedrich Wilhelm Bessel, 1784–1846

Having determined the position of over 50,000 stars, Friedrich Wilhelm Bessel was renowned as an astronomer. In the course of his work, he developed and

[14] In his second memoir on mathematical statistics, in 1829 [40], Fourier had explained how to calculate limits on a function of several estimated quantities, but he had not spelled out how his formulas specialize to this problem.

popularized Gauss's theory of errors. He believed that systematic errors are often more important than random errors, and his influence helped establish the emphasis on perfecting instruments and computational methods that pushed the Germans ahead of the French in astronomy and geodesy by the middle of the 19th century.

The passage translated here is §10 (pp. 141–142) of Bessel's study of the orbit of Olber's comet [8]. Published in 1816, it includes the first known tabulation of significance levels for the normal distribution. The table shows the odds that an error will fall within $\pm\alpha$(probable error) for values of α up to 4. The odds are more than 140 to 1 that it fall within ± 4(probable error). (Four probable errors is about two moduli or 2.8 standard deviations.) But Bessel does not pause over whether this should be regarded as practical certainty. The point of the table is not to show what is required for practical certainty but to show why different observations (or equations) must be weighted differently in order to arrive at the best estimates of unknowns.

Bessel is credited with inventing the notion of a probable error. In the translated passage he estimates the probable error directly from the observations, taking it to be the median of the observed absolute errors.

Bessel in English

Success in determining the final values of quantities from these equations of condition, and even more so the estimation of their likely uncertainty arising from errors in the observations, depends principally on the proper weighting of the equations of condition. It was, therefore, necessary to make a study of this question, the result of which I have already used with advantage for some years.

According to Gauss's least-squares theory, the probability of making an error Δ is

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

(*Theoria mot. corp. coel. P. 212.*), where h depends on the precision of the observations. By means of this expression one can easily determine the probable error of a single observation from an actual set of observations, under the assumption that the errors that actually occur are free from all systematic influences, and are produced only by the imperfections of the instruments and senses. Indeed, the greater the number of observations, the closer we come to the arithmetic mean of all errors, taken together with the same sign, which we shall call ϵ,

$$= 2\int_0^\infty \phi(\Delta)\Delta d\Delta = \frac{1}{h\sqrt{\pi}};$$

and also to the square root of the arithmetic mean of the squares of the errors, which we will denote by ϵ, from the equation

$$\epsilon'^2 = 2\int_0^\infty \phi(\Delta)\Delta^2 d\Delta = \frac{1}{2h^2}.$$

The greater the number of actual observations, the more we are entitled to assume that these errors occur as the Gaussian theory requires, so that from the coincidence of the ϵ and ϵ' obtained from a very large number of observations with the best possible corresponding values from the theory, we now obtain the probable error of an observation, which we will denote ϵ''. This designates the boundary drawn between a number of smaller errors and an equal number of larger ones, so that it is more likely that an observation falls within any wider limit as outside it.

Solving the equation

$$\int_0^x d^{-t^2} dt = \int_x^\infty e^{-t^2} dt,$$

we find that $x = 0,4769364 = h\epsilon''$, so that

$$\epsilon'' = \alpha \times 0.8453\epsilon' = 0.6745\epsilon.$$

The probability of an error smaller than $\alpha\epsilon''$ is to the probability of one larger as the value of the integral $\int e^{-t^2} dt$ from $t = 0$ to $t = \alpha \times 0.4769364$ is the the value of the same integral from $t = \alpha \times 0,4769364$ to $t = \infty$. From the known table of this integral we find the following for several values of α:

$\alpha = 1$	$\cdots\cdots$	$1 : 1$
$\alpha = 1.25$	$\cdots\cdots$	$1 : 1.505$
$\alpha = 1.5$	$\cdots\cdots$	$1 : 2.209$
$\alpha = 1.75$	$\cdots\cdots$	$1 : 3.204$
$\alpha = 2$	$\cdots\cdots$	$1 : 4.638$
$\alpha = 3$	$\cdots\cdots$	$1 : 30.51$
$\alpha = 4$	$\cdots\cdots$	$1 : 142.36$

5.4 Thomas Galloway, 1796–1851

The British mathematician Thomas Galloway wrote on astronomy but worked as an actuary beginning in 1833. His *Treatise on Probability* [42, p. 144], published as a book in 1839, first appeared as the article on probability in the 7th edition of the Encyclopedia Britannica.

In the preface of his *Treatise* (page xi), Galloway explained that, "In the investigation of the most probable mean value of a quantity, to be determined in magnitude or position, from a series of observations liable to error, and the determination of the limits of probable uncertainty, I have followed the very general and elegant analysis of Poisson." Because Poisson was much clearer than Laplace, Galloway played an important role in making the mathematics of Laplace's asymptotic theory understood in Britain. One indication of the influence of Galloway's *Treatise* is that Karl Pearson recommended it to his readers in the book on the philosophy of science that he published in 1892, before he began his research in statistics [70, pp. 177, 180].

in his table of normal probabilities, Galloway used Θ for a quantity following the error law to be within τ moduli of zero for values of τ from 0 to 3. When

discussing particular problems, he treated $\tau = 3$ as practical certainty. In his discussion of Bernoulli's theorem on page 144, for example, he points out that Θ "approaches nearer and nearer to certainty" as τ increases, adding that "it may be seen, by referring to the table, that it is only necessary to have $\tau = 3$ in order to have $\Theta = \cdot 9999779$".

5.5 Augustin Cournot, 1801–1877

Here I translate a passage in Cournot's 1843 book where he criticizes the practice of p-hacking the census [19, S111] . Cournot discussed the issues further in §§102 and 112–114, concluding that judgement about the meaningfulness of such observed differences is ultimately a matter of philosophical (non-numerical) probability. His critique of p-hacking has been discussed by Bernard Bru [13] and Michel Armatte [3, 4].

In France, Cournot's friend Jules-Irénée Bienaymé carried on Cournot's criticism of the abuses of Laplace's theorem.

Cournot in English

. . . Clearly nothing limits the number of the aspects under which we can consider the natural and social facts to which statistical research is applied nor, consequently, the number of variables according to which we can distribute them into different groups or distinct categories. Suppose, for example, that we want to determine, on the basis of a large number of observations collected in a country like France, the chance of a masculine birth. We know that in general it exceeds 1/2. We can first distinguish between legitimate births and those outside marriage, and as we will find, with large numbers of observations, a very appreciable difference between the values of the ratio of masculine births to total births, depending on whether the births are legitimate or illegitimate, we will conclude with very high probability that the chance of a masculine birth in the category of legitimate births is appreciably higher than the chance of the event in the category of births outside marriage. We can further distinguish between births in the countryside and births in the city, and we will arrive at a similar conclusion. These two classifications come to mind so naturally that they have been an object for examination for all statisticians.

Now it is clear that we could also classify births according to their order in the family, according to the age, profession, wealth, and religion of the parents; that we could distinguish first marriages from second marriages, births in one season of the year from those in another; in a word, that we could draw from a host of circumstances incidental to the fact of the birth, of which there are indefinitely many, producing just as many groupings into categories. It is likewise obvious that as the number of groupings thus grows without limit, it is more and more likely *a priori* that merely as a result of chance at least one of the groupings will produce, for the ratio of the number of masculine births to the total number of births, values appreciably different in the two distinct categories. Consequently, as we have already explained, for a statistician who undertakes a thorough

investigation, the probability of a deviation of given size not being attributable to chance will have very different values depending on whether he has tried more or fewer groupings before coming upon the observed deviation. As we are always assuming that he is using a large number of observations, this probability will nevertheless have an objective value in each system of groupings tried, inasmuch as it will be proportional to the number of bets that the experimenter would surely win if he repeated the same bet many times, always after trying just as many perfectly similar groupings, providing also that we had an infallible *criterium* for distinguishing the cases where he is wrong from those where he is right.

But usually the groupings that the experimenter went through leave no trace; the public only sees the result that seemed to merit being brought to its attention. Consequently, an individual unacquainted with the system of groupings that preceded the result will have absolutely no fixed rule for betting on whether the result can be attributed to chance. There is no way to give an approximate value to the ratio of erroneous to total judgments a rule would produce, even supposing that a very large number of similar judgments were made in identical circumstances. In a word, for an individual unacquainted with the groupings tried before the deviation δ was obtained, the probability corresponding to that deviation, which we have called Π, loses all objective substance and will necessarily carry varying significance for a given magnitude of the deviation, depending on what notion the individual has about the *intrinsic importance* of the variable that served as the basis for the corresponding grouping into categories.

5.6 Jules Gavarret, 1809–1890

In 1840, Jules Gavarret published *Principes généraux de statistique médicale*, the first book on the use of probability to evaluate medical therapies [43, 55]. For Gavarret, introducing probability into medicine was a way of bringing medicine up to the level of the most exact sciences, which also, according to Laplace and Poisson, rested ultimately only on probabilities (p. 39). On page 257, he appealed to Poisson's authority to support his choice of 2 moduli as the level of probability sufficient for practical certainty:

> But to make these formulas immediately applicable to the questions we are concerned with, we must transform them in a very simple way. To this end, recall the general principle established on page 39, namely that once an observer has arrived at a high degree of probability for the existence of a fact, he may use the fact as if he were absolutely certain of it. Let us therefore agree on a probability after which any therapeutic fact can and should be accepted without dispute. This probability must satisfy two important conditions: one, to be sufficiently high to leave no doubt in people's minds; the other, not to require too large a number of observations in order for the ratios provided by the statistics we have collected to properly approximate the average chance we are estimating. The choice of

such a probability, one that can and should satisfy us, would have been very delicate; but fortunately we can rely in this matter on an authority whose importance no one, surely, will try to dispute. When M. Poisson set out in his book the rules which should govern the search for possible errors in the judgements of juries,[15] the highest probability that he would give to his propositions, in order to consider himself justified in considering them as free from any reasonable objection, is:

$$\mathrm{P} = 0.9953; \text{ that is to say, betting odds of 212 to 1.}$$

5.7 Wilhelm Lexis, 1837–1914

A prominent German economist and statistician, Lexis published his introduction to population statistics *Einleitung in die Theorie der Bevölkerungsstatistik* [64], in 1875.

On p. 98, he gave this small table for normal probabilities, F_u being the probability of a quantity estimated being within u moduli of the estimate.

u	F_u	u	F_u
0,10	0,11246	1,50	0,966105
0,20	0,22270	2,00	0,995322
0,30	0,32863	2,50	0,999593
0,40	0,42839	3,00	0,999977909
0,50	0,52050	4,00	0,999999985
1,00	0,84270	5,00	0,999999999998

Lexis consistently used Fourier's criterion of three moduli for practical certainty. This passage, from p. 100, is typical of the explanations he gives for choosing u to be 3:

> For the purposes of statistics it should however be more appropriate to take u so large that F_u comes very near to one and therefore expresses a probability that can be considered in practice equal to certainty. It suffices, as before, to set u equal to 3, and we then obtain the probability $F_3 = 0,999978$ for the limit equation ....

5.8 Francis Edgeworth, 1845–1926

As noted in §3.3, Edgeworth introduced the English word *significant* into statistical testing in a paper read at the Jubilee meeting of the Statistical Society of London in 1885 [27]. In this paper, Edgeworth refers the reader to Quetelet, Galton, and Jevons for details on the law of error, but he uses the modulus and Fourier's criterion, repeated by Lexis, of thrice the modulus. Here are a few key quotations.

[15] This is a reference to §135 of Poisson's book.

> From p. 182: The science of Means comprises two main problems: 1. To find how far the difference between any proposed Means is accidental or indicative of a law? 2. To find what is the best kind of Mean; whether for the purpose contemplated by the first problem, the elimination of chance, or other purposes? ... The first problem investigates how far the difference between the average above stated and the results usually obtained in similar experience where pure chance reigns is a significant difference; indicative of the working of a law other than chance, or merely accidental. ...
>
> ... out of a set of (say) N statistical numbers which fulfil the law of error, we take one at random, it is exceedingly improbable that it will differ from the Mean to the extent of twice, and *à fortiori* thrice, the modulus.
>
> From p. 188: ... we shall find that the observed difference between the proposed Means, namely about 2 (inches) far exceeds thrice the modulus of that curve, namely 0*2. The difference therefore "comes by cause."

In his report on the discussion of the paper (p. 217), the president of the session reported that when pressed by the Italian statistician Luigi Perozzo on whether his paper contained anything new, Edgeworth had said that "he did not know that he had offered any new remarks, but perhaps they would be new to some readers. He had borrowed a great deal from Professor Lexis."

Edgeworth again used *significant* in his article on probability in the 11th edition of the *Encyclopedia Britannica* [29, §137]. There he explained that the method discussed in his 1885 paper was a way of deciding whether a difference is real without resorting to a complete inverse (Bayesian) analysis.

> This application of probabilities not to the actual data but to a selected part thereof, this economy of the inverse method, is widely practised in miscellaneous statistics, where the object is to determine whether the discrepancy between two sets of observation is accidental or significant of a real difference.

5.9 Arthur Schuster, 1851–1934

Schuster was born in Germany but in 1870 the followed his parents to England, where he became a prominent physicist. He is best remembered for coining the term the term *periodogram* and analyzing it statistically.

Schuster introduced the word *periodogram* in an 1898 article on the evidence for a 26-day cycle in the weather. In this article [77, p. 18], he described an intensity that would be exceeded only one time in 23 as one that would not "justify us ... to consider a real periodicity as proved, although we might be encouraged to continue the investigation by taking an increased number of events into account." In a 1906 article on the periodicity of sunspots [78, p. 79], he was more exigent:

> ... The probability of an intensity greater than h times the average value is e^h, and we may perhaps begin to suspect a real periodicity when this value is 1 in 200. This gives 5·3 as the value of h and 80,000 as the smallest value of the intensity which invites further discussion. When h has the value 8, the probability of an intensity greater than h times the expectancy is 1 in 3,000 and we may begin to be more confident that there is some definite cause at work to bring up the periodogram to that value. The intensity in that case is 120,000. When h is 16, the chances of being misled by accident is only one in a million.

5.10 Karl Pearson, 1857–1936

Karl Pearson is remembered as the driving force behind the British school of biometry at the beginning of the 20th century. One of his roles was editor, from its founding in 1901 until his death in 1936, of the journal *Biometrika*. The early issues of the journal provide a convenient view on how he and his followers talked about statistical testing at the beginning of the century.

In *Biometrika*'s first volume, we find "perhaps significant", "more probably significant", and "certainly significant". Here are some additional instances of the Edgeworthian "significant":

- In the very first issue, from Pearson's close collaborator W. F. R. Weldon [97, p. 119]: "With probable errors of the order indicated by Tables I. and II., it is unlikely that any of these differences are significant. Even in the case of the last pair of entries the difference, although it is considerable ($0 \cdot 0229$ mm.), is less than twice the probable error of the determination."

- In the second issue, from Oswald H. Latter [62, p. 167]: "To test whether any deviation is significant, M_r is taken as the mean of the whole race of Cuckoos and M_s the mean of Cuckoo's eggs found in the nest of any one species of foster-parent: the standard deviation (σ_s) of such eggs is also ascertainied. The value of $M_r - M_s$ is then compared with that of $0 \cdot 67449\sqrt{\frac{\sigma_r^2}{n_1} + \frac{\sigma_s^2}{n_2}}$, where n_1 = total number of Cuckoo's eggs and n_2 = the number of Cuckoo's eggs in the nests of the species in question, which is the probable error of $M_r - M_s$ due to random sampling. If the value of $M_r - M_s$ be not at least 1.5 to 3 times as great as the value of the other expression the difference of M_r and M_s is not definitely significant."

- In volume 7, for 1909/1910, the American James Arthur Harris (1880–1930) wrote "... I follow the rather common example of statisticians in regarding differences of at least $2 \cdot 5$ times their probable errors as significant" [51, p. 458].

- In a 1912 article co-authored by Pearson himself [6, p. 301]: "Hence the difference is more than three times the probable error and likely to be significant."

These quotations indicate that Pearson and his school consistently used *significant* in the Edgeworthian sense, and that they still measured the likelihood of significance with the probable error rather than the standard error. A difference of more than three probable errors was judged definitely significant, a difference of less than two was thought unlikely to be significant.

In later years, however, we see some non-Edgeworthian uses of *significance* creep into *Biometrika*. Here are some examples.

- In the volume for 1908/1909, J. F. Tocher [89, p. 163], writes "it is possible that a locality may exhibit a difference or differences almost or just significant for one or more colour classes. . . ".
- In the volume for 1914/1915, in an article on the variate difference method co-authored by Pearson himself [18, p. 347]: "Stripped therefore of the common time factor the *Synthetic Index* will be seen to be no very appropriate measure of trade, business activity, and spare money for savings and luxuries. With *Post, Stamp Duties and Savings*, it has probably only a spurious relationship, expenditure on railways has little influence, that on luxuries is very slightly significant, or indeed in the case of tobacco negative."
- In the volume for 1918/1919, in an article on psychophysics Godfrey H. Thomson [88, p. 219]: "The difference is therefore three times its probable error and is just significant."

The subtle nuances of Edgeworth's *significant* were definitively lost in the 1920s. Perhaps they were too subtle to survive. But they did survive long enough for the word to become embedded in mathematical statistics, with all its confusing awkwardness and stubborn permanence.

5.11 Gilbert Walker, 1868–1958

Walker was already an accomplished applied mathematician when he accepted an appointment to the British meteorological office in India. By 1914 [93], he had published a memoir under that office's auspices deploring multiple testing in the statistical interpretation of Schuster's periodograms. This publication may have escaped the notice of his colleagues back in Britain, but he made his point well known in a letter to the editor of *Nature* in 1922 [94, p. 511] and in an article in the *Quarterly Journal of the Royal Meteorological Society* in 1925 [95].

Here is the full text of the 1922 letter to the Editor of *Nature*, entitled "On periodicities":

> THE recent paper by Sir William Beveridge on "Wheat Prices and Rainfall" (Journal of the Royal Statistical Society, vol. 85, pp. 412–478, 1922) raises a rather important question of principle which is involved not only in discussions over the existence of periodicities, but also over relationships between different variables.

Before Schuster's papers on the periodogram it was customary for a period to be accepted as real provided that it had an amplitude comparable with that of the original figures under analysis; and he revolutionised the treatment of the subject by showing that if the squares of the intensities of the various periodic terms are plotted in a periodogram, and if the data are those of an entirely chance distribution, then the average value of an ordinate being a, the probability that a particular ordinate will equal or exceed ka is e^{-k}. Sir William Beveridge is accordingly perfectly justified in taking Schuster's sunspot period of 11·125 years, or Brückner's 34·8 year period, and deciding that these periods probably occur in his wheat prices if the corresponding intensities are three or four times the average. But he, like many other investigators, goes a stage further, and after picking out the largest from a large number of intensities he applies the same criterion as if no selection had occurred. It is, how ever, clear that if we have a hundred intensities the average of which, a, is derived from a number of random figures, then the probable value of the largest of these chance intensities will not be a but will be considerably greater, and it is only when the largest amplitude actually derived materially exceeds the theoretical chance value thus obtained that reality can be inferred.

Taking the periodicities of wheat prices on pp. 457–459 between 5 years and 40 years,[16] I estimate that the "width of a line" ranges from 0·1 year for a 5 years' period, through 0·5 at 12 years to 4 years at 33 years; and accordingly that the number of independent periods between 5 years and 40 is in this case about 51. The value of a, the average intensity, being 5·898, it is easily seen that the chance of all the 51 random intensities being less than $3a$ is $(1 - e^{-3})^{51}$, or 0·074, so that the chance of at least one intensity greater than $3a$ is 0·926, not e^{-3} or 0.050, as is habitually assumed. Instead of the chance of an occurrence of $3a$ "making a *prima facie* case for enquiry" (p. 424), the odds are 12 to 1 in favour of its production by mere chance. The chance of at least two intensities above $3a$ is 0·728, of three it is 0·470, of four 0·248, of five 0·109, of six 0·0403, of seven 0·0127, of nine 0·00085, and of eleven 0·00003. Thus it is not until six intensities over $3a$ are found that the chance of production by pure luck is less than 1 in 20. It is also easily found that if the chance of all the 51 intensities being less than na is to be 19/20, n is 6·9; i.e. the greatest intensity for wheat price fluctuations must be 41, not 18, before the probability of its being due to luck is reduced to 1/20; and if the likelihood is to be 1/100 we must have $n = 8 \cdot 5$, the corresponding wheat-price intensity being 50. Of intensities greater

[16] Footnote by Walker: Sir William Beveridge points out on pp. 423–424 that amplitudes for periods of less than 5 years are inevitably diminished, while those above 31 are diminished by the process employed for eliminating secular trend: I calculate that the intensity at 35 years should be multiplied by $(0 \cdot 87)^{-2}$ or 1·3, and that at 54 by 3·8.

than 41 Sir William Beveridge found four, and greater than 50 only two.

At first sight it might seem that the agreement between Sir William Beveridge's forecasted synthesis rainfall curve and the actual rainfall was too great to be explained by a few harmonic terms; but the correlation co-efficient of $0 \cdot 38$ (see p. 475) indicates that while $0 \cdot 38$ of the rainfall variations are accounted for, only $(0 \cdot 38)^2$, or about a seventh, of the independent factors which control these variations have been ascertained.

As pointed out in a paper "On the Criterion for the Reality of Relationships or Periodicities," in the Indian Meteorological Memoirs (vol. 21, No. 9, 1914), the same principle is valid when discussing relationships. If we are examining the effect of rainfall on temperature and ascertain that the correlation coefficient between the rainfall and temperature of the same month in a particular English county is four times the probable error, we may infer that the effect is highly probable. But if we work out the co-efficients of that temperature with a hundred factors taken at random, e.g. with the monthly rainfall of Tashkend 5·8 years previously, and pick out the largest co-efficient, it would be wrong to compare it with the average co-efficient produced by mere chance; as shown in the paper referred to, the probable value of the largest of 100 co-efficients is $4 \cdot 01$ times as great as the probable value of one taken at random.

GILBERT T. WALKER.

Meteorological Office, Simla, August 24.

5.12 Arthur Bowley, 1869–1957

Bowley was a professor of economics at the London School of Economics. Like his fellow economist Edgeworth, he was not part of Pearson's biometric circle. He published several textbooks on statistics, beginning with the first edition of his *Elements of Statistics* in 1901 [12], where he acknowledged a debt to Edgeworth, both for Edgeworth's publications and for personal instruction. As we see on page 6 of the book, he adopted Edgeworth's criterion of 3 moduli for practical certainty:

> Without the aid of statistical method, the averages obtained show mere numbers from which no logical deductions can be made. With the help of this knowledge, it can be seen whether the change from year to year is significant or accidental; whether the figures show a progressive or periodic change; whether they obey any law or not.

On page 313, he cites Edgeworth [27] as authority for the proposition that an apparent difference of 3 moduli signifies a real difference.

> . . . the modulus of a difference is most useful in comparing two groups selected as having certain qualities. Thus Professor Edgeworth discusses whether an ascertained difference of 2 inches between the

> average heights of a large number of criminals and that of the general population is significant; and finding that the modulus for the difference between two random groups is only 0.08, holds that there is a cause of the difference in the method of selection; that is, that criminality and low stature are found together. We might apply the same principle to the investigation of the existence of a period in any figures; for if the modulus of the figures was c, the modulus for the difference between the averages of two random samples of 20 months each would be $c\sqrt{\frac{1}{20} + \frac{1}{20}}$; if the difference between the averages of the figures for 20 Decembers and 20 Junes was 3 times this quantity the existence of a period would be established.

5.13 George Udny Yule, 1871–1951

After studying mathematical physics, Yule became a statistician as an assistant to Pearson. But he later quarreled with Pearson and went his own way on a number of points. The first edition of his *Theory of Statistics* [101] appeared in 1911. On page 262, we find this explanation of significant differences:

> . . . if we observe a different proportion in one sample from that which we have observed in another, the question again arises whether this difference may be due to fluctuations of simple sampling alone, or whether it indicates a difference between the conditions subsisting in the universes from which the two samples were drawn: in the latter case the difference is often said to be **significant**. These questions can be answered, though only more or less roughly at present, by comparing the observed difference with the standard-deviation of simple sampling. We know roughly that the great bulk at least of the fluctuations of sampling lie within a range of ± three times the standard-deviation; and if an observed difference from a theoretical result greatly exceeds these limits it cannot be ascribed to a fluctuation of "simple sampling" as defined in §8: it may therefore be significant. The "standard-deviation of simple sampling" being the basis of all such work, it is convenient to refer to it by a shorter name. The observed proportions of A's in given samples being regarded as differing by larger or smaller errors from the true proportion in a very large sample from the same material, the "standard-deviation of simple sampling" may be regarded as a measure of the magnitude of such errors, and may be called accordingly the **standard error**.

Here *significant* is Edgeworthian, but modulus and probable error have given way to standard error.

5.14 Raymond Pearl, 1879–1940

The American biologist Raymond Pearl, who spent most of his career at Johns Hopkins, studied with Karl Pearson for a year in 1906. Like Yule and many

other of Pearson's disciples, he eventually quarreled with the master, but he fondly acknowledged Pearson in the preface to the textbook he published 1923, *Introduction to Medical Biometry and Statistics* [69].

Perusing the occurrences of *significant* in this textbook, we might conclude that Pearl has studied and learned the Edgeworthian way of using the word but does not quite find it natural. It is, he tells us a conventional way of talking:

On page 214:

> ...Is a difference six times its probable error likely to arise from chance alone, or does it represent a really significant difference?
>
> There has grown up a certain conventional way of interpreting probable errors, which is accepted by many workers. It has been practically a universal custom among biometric workers to say that a difference (or a constant) which is smaller than twice its probable error is probably not significant, whereas a difference (or constant) which is three or more times its probable error is either "certainly," or at least "almost certainly," significant.

On page 217:

> From this table it is seen that a deviation of four times the probable error will arise by chance less often than once in a hundred trials. When one gets a difference as great or greater than this he may conclude with reasonable certainty that it did not arise by chance alone, but has significant meaning.

If we want to quibble, we can argue that Pearl has not mastered the jargon perfectly. The antecedent of "it" in the first quoted sentence is the difference six times its probable error. Edgeworth would say that this observed difference *probably is* a significant difference, not that it *represents* one.

5.15 Truman L. Kelley, 1884–1961

The following passage is drawn from pp. 102–103 of Kelley's 1924 book *Statistical Method* [58].

Kelley's words

The normal curve assists in establishing the degree of confidence which may be placed in statistical findings. The significance of any measure is to be judged by comparison with its probable error. If a child makes a score of 80 on a certain test and if the probable error of the score is 5, we may estimate the chances of the child's true ability being as much as 100. We assume that the distribution of the child's performances would follow a normal curve. Note that the assumption is not that the talents of children in general follow a normal distribution. This latter might be less reasonable than the one we are called upon to make. Moreover, so little difference in probabilities, except for extreme deviates, is ordinarily consequent to differences in forms of distribution, that the

assumption of normality is little likely to result in serious error for such problems as the present one. For extreme deviates it generally does not matter so far as any practical deductions are concerned whether the chances are 1 in 1000 or ten times as great. Nor for smaller deviates does it make any particular difference whether the chances are 400 in 1000 or 410 in 1000. Should such differences as mentioned be significant in any particular problem, no assumption should be made, but the type of the curve should be experimentally determined.

For the problem in hand: If the P. E . is 5 the standard error is $\left(\frac{5}{.6745}\right) = 7.413$. The difference between the scores that we are concerned with is $(100 - 80) = 20$, which is $\left(\frac{20}{7.413}\right) = 2.698$ standard errors. The K-W Table, or more conveniently for this problem Sheppard's Tables, may be used to find the area in the tail below the point which is 2.698 standard deviations below the mean. The tables give .0035. To interpret this we should postulate the person's true ability as being 100 and his various performances distributing themselves in a normal distribution, with standard deviation equal to 7.413 around this mean. Then .0035 of the area of the curve will lie below the point 80. Accordingly if his true ability is 100, only 35 times in 10000, or 3.5 times in 1000, would a score as low or lower than 80 be expected. With such figures a person could accept the proposition that the child's ability was not as great as 100 with about as much certainty as he can start across a business street expecting not to be hit by an automobile. It is, in other words, just such a conclusion as one is justified in acting upon.

5.16 David Brunt, 1886–1965

In 1917, the Welsh meteorologist David Brunt published a book the theory of errors, *The Combination of Observations* [16], which included a chapter on the periodogram. True to the Gaussian tradition, the book did not set a standard for practical certainty. Brunt explained Schuster's probabilistic treatment of the Fourier coefficient, giving the following table and explanation on p. 200:

κ	$e^{-\kappa}$	κ	$e^{-\kappa}$
1	·3679	6	$2 \cdot 4 \times 10^{-3}$
2	·1353	8	$3 \cdot 35 \times 10^{-4}$
3	·0498	10	$4 \cdot 54 \times 10^{-5}$
4	·0183	12	$6 \cdot 14 \times 10^{-6}$
5	·00674	16	$1 \cdot 13 \times 10^{-7}$

> This table may be interpreted thus:—The chance of obtaining for the square of a Fourier coefficient a value greater than three times its expectancy or mean value is ·0498, or about 1 in 20. So that, if on analysing a series of observations we obtain a coefficient whose square is more than three times the expectancy, we can state that the probability that it is produced by a chance distribution of the quantities analysed is $\frac{1}{20}$. If the square of the Fourier coefficient be 12 times its expectancy, the probability that it is produced by a chance distribution is 1 in 160,000.

But, as noted in §3.6, Brunt set 19 to 1 as the odds for practical certainty in his periodogram analysis of Greenwich temperature records in 1919 [17, p. 328].

5.17 Ronald A. Fisher, 1890–1962

Fisher is celebrated as the most accomplished mathematical statistician of the 20th century. He laid out his understanding of "tests of significance", by no means his most important contribution, in his monograph *Statistical Methods for Research Workers* [36], published in 1925 and in many subsequent editions.

Shelving the probable error

As we saw in §3.5, the most novel aspect of the 1925 book was that it tabulated values of the tail probability P not only for the normal distribution and Pearson's χ^2 but also for a number of other distributions that can be used when the assumption that individual observations have a normal distribution is taken seriously, including Student's t and the distribution of the correlation coefficient. As Fisher explains in the following passage, this led him to abandon measurement in terms of the probable error in favor of measurement in terms of tail probabilities, and in particular to replace the criterion of two probable errors by the criterion of 5%.

> Pp. 47–48: "The value of the deviation beyond which half the observations lie is called the quartile distance, and bears to the standard deviation the ratio ·67449. It is therefore a common practice to calculate the standard error and then, multiplying it by this factor, to obtain the probable error. The probable error is thus about two-thirds of the standard error, and as a test of significance a deviation of three times the probable error is effectively equivalent to one of twice the standard error. The common use of the probable error is its only recommendation; when any critical test is required the deviation must be expressed in terms of the standard error in using the probability integral table."

The end of Edgeworthian signifying

Here are some examples the book's non-Edgeworthian use of "significant".

- P. 21: "The table illustrates the general fact that the significance in the normal distribution of deviations exceeding four times the standard deviation is extremely pronounced."
- P. 123: "This suggests the possibility that if we had fitted a more complex regression line to the data the probable errors would be further reduced to an extent which would put the significance of b beyond doubt."
- Pp. 158–159: "Taking the four definite levels of significance, represented by P = ·10, ·05, ·02, and ·01, the table shows for each value of n, from 1

to 20, and thence by larger intervals to 100, the corresponding values of r."

- P. 90: "the significance will become more and more pronounced as the sample is increased in size..."
- P. 47, with reference to the table for the normal distribution: "The value for which P = ·05, or 1 in 20, is $1 \cdot 96$ or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available.
- Pp. 81–82: "The expected values are calculated from the observed total, so that the four classes must agree in their sum, and if three classes are filled in arbitrarily the fourth is therefore determinate, hence $n = 3$, $\chi^2 = 10.87$, the chance of exceeding which value is between .01 and .02; if we take P = .05 as the limit of significant deviation, we shall say that in this case the deviations from expectation are clearly significant."
- Pp. 102–102: "If, therefore, we know the standard deviation of a population, we can calculate the standard deviation of the mean of a random sample of any size, and so test whether or not it differs significantly from any fixed value. If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance; this is roughly equivalent to the corresponding limit P = ·05, already used for the χ^2 distribution."
- P. 158: "very much exaggerating the significance."
- P. 161: "The values given in Table V. (A) for $n = 25$, and $n = 30$, give a sufficient indication of the level of significance attained by this observation.

It is also notable that we find the term "statistical significance" (page 218).

5.18 Morris Viteles, 1898–1996

In a brief article on intelligence testing published by Viteles in 1922 [91], we find "greatly significant", "particularly significant", and "high enough to be of considerable significance". We also see the first use of "level of significance" that I have found. Viteles states

> ...reduces the co-efficient of correlation ... to plus $.21 \pm .091$, much below the level of significance.

and

> . . . reduces the co-efficient of correlation of these two tests to plus $0.37 \pm .080$, also below the level of significance.

Here the level of significance is evidently six probable errors. Viteles, who spent most of his career at the University of Pennsylvania, had not benefited from a year with Pearson, but he became a prominent figure in industrial and organizational psychology.

Acknowledgements

This paper would not have been possible without insights freely shared by Bernard Bru and John Aldrich. I have also benefited from conversations with many other colleagues, especially Michel Armatte, Hans Fischer, Steve Goodman, Prakash Gorroochurn, Sander Greenland, Steve Stigler, Volodya Vovk, and Sandy Zabell.

References

[1] George Biddell Airy. *On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations*. Macmillan, London, 1st edition, 1861. 8

[2] Valentin Amrhein, Sander Greenland, and Blake McShane et al. Retire statistical significance. *Nature*, 567:305–307, 2019. 16

[3] Michel Armatte. Discussion de l'article de D. Denis [26]. *Journal de la Société Française de Statistique*, 145(4):27–36, 2004. 22

[4] Michel Armatte. Contribution à l'histoire des tests laplaciens. *Mathematics and social sciences*, 44(176):117–133, 2006. 22

[5] David R. Bellhouse. Karl Pearson's influence in the United States. *International Statistical Review*, 77(1):51–63, 2009. 12

[6] R. Crewdson Benington and Karl Pearson. A study of the Negro skull with special reference to the Congo and Gaboon crania. *Biometrika*, 8(3/4):292–339, 1912. 26

[7] Joseph Bertrand. *Calcul des Probabilités*. Gauthier-Villars, Paris, 1889. Second edition 1907. 7

[8] Friedrich Wilhem Bessel. Untersuchungen über die Bahn des Olberschen Kometen. *Abhandlungen der mathematischen Klasse der Königlichen-Preussischen Akademie der Wissenschaften aus den Jahren 1812–1813*, pages 119–160, 1816. 20

[9] William Beveridge. Wheat prices and rainfall in western Europe (with discussion). *Journal of the Royal Statistical Society*, 85(3):412–478, May 1922. 13

[10] Edwin G. Boring. The number of observations on which a limen may be based. *The American Journal of Psychology*, 27(3):315–319, 1916. 11

[11] Edwin G. Boring. Mathematical vs. scientific significance. *Psychological Bulletin*, 16(10):335–338, 1919. 11

[12] Arthur Lyon Bowley. *Elements of Statistics*. King, Westminster, 1901. Later editions appeared in 1902, 1907, 1920, 1925, and 1937. 29

[13] Benard Bru. Remarques sur l'article de D. Denis [26]. *Journal de la Société Française de Statistique*, 145(4):37–38, 2004. 22

[14] Bernard Bru, Marie-France Bru, and Oliver Bienaymé. La statistique critiquée par le calcul des probabilités: Deux manuscrits inédits d'Irenée Jules Bienaymé. *Revue d'histoire des mathématiques*, 3:137–239, 1997. 7, 17

[15] Marie-France Bru and Bernard Bru. *Les jeux de l'infini et du hasard*. Presses universitaires de Franche-Comté, Besançon, France, 2018. Two volumes. 3, 7

[16] David Brunt. *The The Combination of Observations*. Cambridge University Press, 1917. 32

[17] David Brunt. A periodogram analysis of the Greenwich temperature records. *Quarterly Journal of the Meteorological Society*, 45(192):323–338, 1919. 13, 33

[18] Beatrice M. Cave and Karl Pearson. Numerical illustrations of the variate difference correlation method. *Biometrika*, 10:340–355, 1914/1915. 27

[19] Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [20]. 6, 7, 9, 10, 22

[20] Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The volumes are numbered I through XI, but VI and XI are double volumes. 36

[21] Andrew I. Dale. *A History of Inverse Probability From Thomas Bayes to Karl Pearson*. Springer, New York, second edition, 1999. 3

[22] Augustus De Morgan. A treatise on the theory of probabilities. In Edward Smedley, Hugh James Rose, and Henry John Rose, editors, *Encyclopaedia Metropolitana*, volume 2, pages 393–490. Griffin, London, 1837. 9

[23] Augustus De Morgan. *An Essay on Probabilities, and on their application to Life Contingencies and Insurance Offices*. Longman, Orme, Brown, Green & Longmans, London, 1838. 3, 10

[24] W. Edwards Deming. *Statistical Adjustment of Data*. Wiley, New York, 1943. 15

[25] Arthur P. Dempster. Further examples of inconsistencies in the fiducial argument. *Annals of Mathematical Statistics*, 34(3):884–891, 1966. 4

[26] Daniel J. Denis. The modern hypothesis testing hybrid: R. A. Fisher's fading influence. *Journal de la Société Française de Statistique*, 145(4):5–26, 2004. 1, 35, 36

[27] Francis Edgeworth. Methods of statistics. *Journal of the Statistical Society of London*, Jubilee Volume:181–217, 1885. 10, 24, 29

[28] Francis Edgeworth. On discordant observations. *Philosophical Magazine Series 5*, 23(143):364–375, 1887. 13

[29] Francis Edgeworth. Probability. In *Encyclopædia Britannica*, volume 22. Cooper, 11th edition, 1911. 25

[30] Francis Edgeworth. Mathematical representation of statistics: A reply. *Journal of the Royal Statistical Society*, 81(2):322–333, 1918. 4

[31] Johann Franz Encke. Über die Methode der kleinsten Quadrate. *Astronomisches Jahrbuch für 1834. Der Sammlung Berliner astronomischer Jahrbücher*, 59:249–312, 1832. 10

[32] Richard William Farebrother. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900*. Springer, New York, 1999. 3

[33] Hans Fischer. *A History of the Central Limit Theorem from Classical to Modern Probability Theory*. Springer, New York, 2011. 3

[34] Ronald A. Fisher. Applications of "Student's" distribution. *Metron*, 5(3):90–104, 1925. 15

[35] Ronald A. Fisher. The influence of rainfall on the yield of wheat at Rothamsted. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:89–142, 1925. 14

[36] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925. The thirteenth edition appeared in 1958. 12, 33

[37] Ronald A. Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 125(796):54–59, 1929. 14

[38] Ronald A. Fisher. Combining independent tests of significance. *The American Statistician*, 2(5):30, 1948. 15

[39] Joseph Fourier. Mémoire sur les résultats moyens déduits d'un grand nombre d'observations. In Joseph Fourier, editor, *Recherches statistiques sur la ville de Paris et le département de la Seine*, pages ix–xxxi. Imprimerie royale, Paris, 1826. 18

[40] Joseph Fourier. Second mémoire sur les résultats moyens et sur les erreurs des mesures. In Joseph Fourier, editor, *Recherches statistiques sur la ville de Paris et le département de la Seine*, pages ix–xlviii. Imprimerie royale, Paris, 1829. 6, 19

[41] Michael Friendly. A.-M. Guerry's *Moral Statistics of France*: Challenges for multivariable spatial analysis. *Statistical Science*, 22(3):368–399, 2007. 6

[42] Thomas Galloway. *Treatise on Probability*. Black, Edinburgh, 1839. 9, 21

[43] Jules Gavarret. *Principes généraux de statistique médicale, ou développement des règles qui doivent présider à son emploi*. Bechet, Paris, 1840. 23

[44] Gerd Gigerenzer. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018. 1

[45] Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Krüger. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge, 1989. 1

[46] Prakash Gorroochurn. *Classic Topics on the History of Modern Mathematical Statistics from Laplace to More Recent Times*. Wiley, New York, 2016. 3, 4

[47] André-Michel Guerry. *Essai sur la statistique morale de la France*. Crochard, Paris, 1833. 6

[48] Ian Hacking. *The Taming of Chance*. Cambridge University Press, New York, 1990. 4, 7

[49] Roger Hahn. *Correspondance de Pierre Simon Laplace (1749–1827)*. Brepols, Turnhout, Belgium, 2013. Two volumes. 3

[50] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Wiley, New York, 1998. 3, 18

[51] J. Arthur Harris. On the selective elimination occurring during the development of the fruits of staphylea. *Biometrika*, 7:452–504, 1909/1910. 26

[52] Christopher Charles Heyde and Eugene Seneta. *I. J. Bienaymé: Statistical Theory Anticipated*. Springer, New York, 1977. 7

[53] H. O. Hirschfeld. The distribution of the ratio of covariance estimates in two samples drawn from normal bivariate distributions. *Biometrika*, 29(1/2):65–79, 1937. 15

[54] Richard A. Hurlbert, Stuart H. Levine and Jessica Utts. Coup de grâce for a tough old bull: "statistically significant" expires. *The American Statistician*, 73:sup1:352–357, 2019. 1

[55] Edward Huth. Jules Gavarret's *Principes Généraux de Statistique Médicale*. *Journal of the Royal Society of Medicine*, 101:205–212, 2008. 23

[56] Marie-Françoise Jozeau. *Géodésie au XIXème Siècle: De l'hégémonie française à l'hégémonie allemande. Regards belges*. PhD thesis, Université Denis Diderot Paris VII, Paris, 1997. 4, 7

[57] Andreas Kamlah. The decline of the Laplacian theory of probability: A study of Stumpf, von Kries, and Meinong. In Krüger et al. [61], pages 91–116. 7

[58] Truman L. Kelley. *Statistical Method*. Macmillan, New York, 1923. 11, 31

[59] Sven K. Knebel. *Wille, Würfel und Wahrscheinlichkeit: Das System der moralischen Notwendigkeit in der Jesuitenscholastik 1550–1700*. Meiner, Berlin, 2000. 8

[60] Christian Kramp. *Analyse des Réfractions Astronomiques et Terrestres*. Schwikkert, Leipsic, 1799. 4

[61] Lorenz Krüger, Lorraine J. Daston, and Michael Heidelberger, editors. *The Probabilistic Revolution. Volume 1: Ideas in History*. MIT, Cambridge, Massachusetts, 1987. 39, 40

[62] Oswald H. Latter. The egg of cuculus canorus. An enquiry into the dimensions of the cuckoo's egg and the relation of the variations to the size of the eggs of the foster-parent, with notes on coloration, &c. *Biometrika*, 1(2):164–176, 1902. 26

[63] Erich L. Lehmann. *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, New York, 2011. 1

[64] Wilhelm Lexis. *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strassburg, 1875. 9, 24

[65] Mansfield Merriman. Least squares: A list of writings relating to the method, with historical and critical notes. *Transactions of the Connecticut Academy of Arts and Sciences*, 4:151–232, 1877. 10

[66] Mary Morgan. *The History of Econometric Ideas*. Cambridge University Press, Cambridge, 1990. 13

[67] Denton E. Morrison and Ramon E. Henkel. *The Significance Test Controversy — A Reader*. Aldine, Chicago, 1970. 15

[68] Marie-Vic Ozouf-Marignier. Administration, statistique, aménagement du territoire: l'itinéraire du Préfet Chabrol de Volvic (1773–1843). *Revue d'histoire moderne et contemporaine*, 44(1):19–39, 1997. 17

[69] Raymond Pearl. *Introduction to Medical Biometry Statistics*. Saunders, Philadelphia and London, 1923. 31

[70] Karl Pearson. *The Grammar of Science*. Scott, London, 1892. 21

[71] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894. 7

[72] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50:157–175, 1900. 15

[73] Siméon-Denis Poisson. Observations relatives au nombre de naissances des deux sexes. *Annuaire le bureau des longitudes pour 1825*, pages 98–99, 1824. 6

[74] Siméon-Denis Poisson. Mémoire sur la proportion des naissances des filles et des garçons. *Mémoires de l'Académie royale des sciences*, IX:239–308, 1830. 6, 19

[75] Siméon-Denis Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédés des règles générales du calcul des probabilités*. Bachelier, Paris, 1837. 19

[76] Ivo Schneider. Laplace and thereafter: The status of probability calculus in the nineteenth century. In Krüger et al. [61], pages 191–214. 7

[77] Arthur Schuster. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestial Magnetism*, III(1):13–41, 1898. 13, 25

[78] Arthur Schuster. II. On the periodicities of sunspots. *Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 206(402–412):69–100, 1906. 13, 25

[79] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication, 2019. Working Paper 54, `www.probabilityandfinance.com`. 16

[80] Oscar Sheynin. Laplace's theory of errors. *Archive for History of Exact Sciences*, 17(1):1–61, 1977. 3

[81] Oscar Sheynin. C.F. Gauss and the theory of errors. *Archive for History of Exact Sciences*, 20(1):21–72, 1979. 3

[82] Oscar Sheynin. Early history of the theory of probability. *Archive for History of Exact Sciences*, 46(3):253–283, 1994. 3

[83] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA, 1986. 3

[84] Stephen M. Stigler. A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, 101(1):60–70, 1992. 11

[85] Stephen M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999. 10

[86] Stephen M. Stigler. Fisher and the 5% level. *Chance*, 21:12–21, 2008. 12

[87] Dale Stout. A question of statistical inference: E. G. Boring, T. L. Kelley, and the probable error. *The American Journal of Psychology*, 102(4):549–562, 1989. 11, 12

[88] Godfrey H. Thomson. The criterion of goodness of fit of psychophysical curves. *Biometrika*, 12:216–230, 1918/1919. 27

[89] J. F. Tocher. Pigmentation survey of school children in Scotland. *Biometrika*, 6:130–235, 1908/1909. 27

[90] John Venn. *The Logic of Chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan, London, 1866. 13

[91] Morris S. Viteles. A comparison of three tests of "general intelligence". *Journal of Applied Psychology*, 6(4):391–402, 1922. 34

[92] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919. 4

[93] Gilbert T. Walker. On the criterion for the reality of relationships or periodicities. *Indian Meteorological Memoirs*, 21(9), 1914. 13, 27

[94] Gilbert T. Walker. Periodicities. (Letter to the Editor. *Nature*, 110(2763):511, 1922. 14, 27

[95] Gilbert T. Walker. On periodicity. *Quarterly Journal of the Royal Meteorological Society*, 51(216):337–346, 1925. 27

[96] Helen M. Walker. *Studies in the History of Statistical Method*. Williams & Wilkins, Baltimore, 1929. 3, 7

[97] W. F. R. Weldon. A first study of natural selection in clausilia laminata (montagu). *Biometrika*, 1(1):109–124, 1901. 26

[98] John Wishart. Field experiments of factorial design. *The Journal of Agricultural Science*, 28(2):299–306, 1938. 15

[99] John Wishart. Test of homogeneity of regression coefficients, and its application in the analysis of covariance. In *Colloques internationaux XIII, Le calcul des probabilités et ses applications, Lyon, 28 juin au 3 juillet 1948*, pages 93–99. CNRS, Paris, 1949. 15

[100] George Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60:812–854, 1897. 7

[101] George Udny Yule. *An Introduction to the Theory of Statistics*. Griffin, London, first edition, 1911. 30