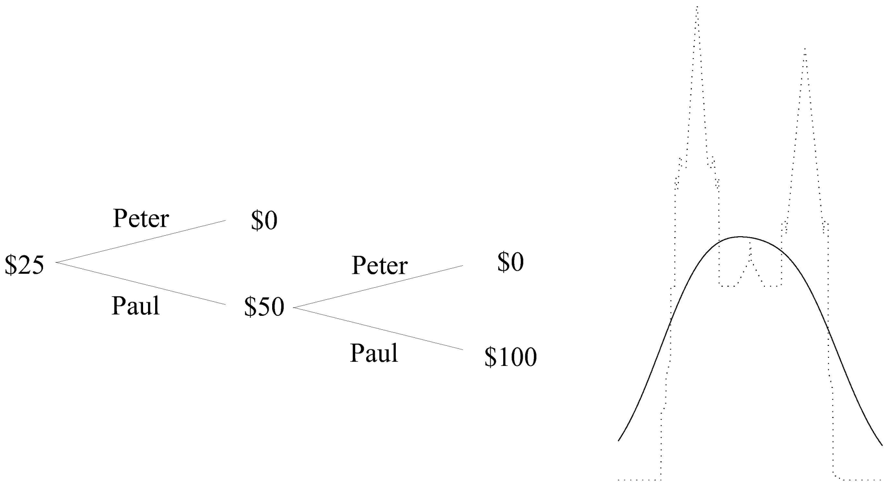
On the nineteenth-century origins of significance testing and p-hacking

Glenn Shafer, Rutgers University



The Game-Theoretic Probability and Finance Project

Working Paper #55

First posted July 18, 2019. Last revised June 11, 2020.

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

Abstract

Although the names *significance test*, *p-value*, and *confidence interval* came into use only in the 20th century, the methods they name were already used and abused in the 19th century. Knowledge of this earlier history can help us evaluate some of the ideas for improving statistical testing and estimation currently being discussed.

This article recounts first the development of statistical testing and estimation after Laplace's discovery of the central limit theorem and then the subsequent transmission of these ideas into the English-language culture of mathematical statistics in the early 20th century. I argue that the earlier history casts doubt on the efficacy of many of the competing proposals for improving on significance tests and p-values and for forestalling abuses. Rather than further complicate the way we now teach statistics, we should leave aside most of the 20th-century embellishments and emphasize exploratory data analysis and the idea of testing probabilities by betting against them.

1 Introduction

At a time when the evaluation of statistical evidence has become ever more important, standard methods for this evaluation are under increasing scrutiny. Some statisticians argue against the continued use of the phrase *statistical significance*; others deplore the way *p-values* are taught and used. Numerous competing proposals for improving the presentation of statistical results are being advanced. For some observers, the entire enterprise of objective judgment based on statistics is threatened by the practice of searching for statistical significance, *p-hacking* as it is now called.

Those addressing this problem draw on experience from many domains of application. History may also have a role to play. Marie-France Bru and Bernard Bru have made the case for history's role with these eloquent words:

> To penetrate to the reasons of things, look at how they have gradually been revealed in the course of time, in their progression and in their ruptures, if any.[1]

By studying the historical record, we may gain more understanding of how the space of possibilities in which we are struggling was constructed and why it has persisted through earlier crises. We may even find discarded paths of thought that might merit renewal.

Some authors have traced current difficulties with significance testing back to the 20th-century statisticians R. A. Fisher and Jerzy Neyman. The development of Fisher's and Neyman's ideas and the conflicts between them form a fascinating and important chapter of intellectual history,[2] and more attention to that history might diminish errors and abuses resulting from university teaching that fails to distinguish clearly between Fisher's ideas about evidence and Neyman's ideas about decision-making.[3] Debates about statistical inference since Fisher and Neyman, concerning objective and subjective probability and the roles of Bayes's rule and likelihood, are also important and in some cases venerable enough to qualify for historical as well as philosophical study.[4] For a full picture, however, we also need to reach further back in history. For in many respects, Fisher's and Neyman's ideas had already been developed in the 19th-century, in traditions launched by Laplace and Gauss.

Nineteenth-century parallels to what has happened in the 20th century and the first two decades of the 21st give us reason to question some diagnoses for current difficulties. Abuses being decried today—searching for significant results, unsupported modeling assumptions for observational data, etc.—arose in the 19th century as well. These abuses cannot be blamed on the terms *statistically significant* and *p-value*, which were not in use in the 19th century.

[1] Translated from [20, pp. 301–302]: Pour pénétrer les raisons des choses, voyons comment elles se sont dévoilées progressivement au cours du temps dans leurs enchaînements et leurs ruptures, s'il s'en trouve.

[2] See especially [78].

[3] See for example [51].

[4] See [82] and its many references.

It also becomes more difficult to hope that an emphasis on confidence intervals can diminish abuses of significance testing when we learn that 19th-century p-hacking grew out of the teaching of confidence intervals.

This history also shows us rupture. The Laplacean tradition was heavily scarred by 19th-century abuses, to the extent that it was ridiculed and largely forgotten, especially in France. Current neglect of the 19th-century history is due in part to the fact that R. A. Fisher and many of his successors felt that they were building on ruins. As Fisher's daughter Joan Fisher Box said this bluntly in 1978 [17, p. 62]:

> The whole field was like an unexplored archeological site, its structure hardly perceptible above the accretions of rubble.

This enduring influence of this view is illustrated by David J. Hand's approving citation of Box's words in 2015 [59, p. 2].

The historical account in this paper begins, in Section 2, by discussing Laplace's large-sample statistical theory in the 19th century—how it was created, how it differed from Gauss's theory of least squares, 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, using the method of least squares but without knowing probabilities for measurement errors or prior probabilities, we can obtain large-sample confidence limits for unknown quantities and hence significance tests for whether the quantities are zero. Gaussian 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 and Siméon-Denis Poisson made it accessible to statisticians. The uses included p-hacking and inferences based on questionable assumptions.

Although the misuse of Laplace's theory so discredited it in France that it was practically forgotten there by the end of the 19th century, it was still taught and used elsewhere. Section 3 sketches its transmission into Britain and the United States: how limits for practical certainty were variously expressed in terms of probable errors, moduli, standard errors, and finally, in Fisher's *Statistical Methods*, tail probabilities, and how the terms *significance* and *p-value* emerged. 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 Karl Pearson using it in a way that is no longer readily understood. The term *p-value* appears in statistics in the 1920s, deriving from the much older "value of P".

Section 4 looks at lessons we might draw from this history. One lesson is that p-hacking and other abuses of statistical testing need not be blamed on 20th-century innovations; the same abuses arose already in the 19th century. The difficulties go deeper, and remedies must go deeper. I argue that we should simplify, not complicate, our teaching of error limits, that we should acknowledge and even emphasize pitfalls, and that we should teach statistical testing in betting terms, reviving a fundamental aspect of probabilistic reasoning that has been progressively minimized for three centuries.

Appendix A documents the history set out here with quotations from 19th- and early 20th-century authors. Those quoted are only a few 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 how the logic and the pitfalls of statistical testing did and did not change over the period.

Appendix B quotes some authors who have distinguished between *Bernoullian* and *Bayesian* statistical methods.

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 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

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 justifying it by the fact that it makes the arithmetic average of the measurements the mode of the posterior probabilities for the quantity measured. Perhaps partly inspired by Gauss's insight, 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.[5]

Laplace first applied his theorem to a problem that had long concerned him, that of testing the hypothesis that planetary orbits were distributed randomly. 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

[5] Many authors have detailed the interplay between Laplace and Gauss: [20, 38, 39, 52, 58, 108, 109, 110, 112].

also wrote to colleagues throughout Europe to explain the importance of his theorem, illustrating its power with examples.[6]

2.2 Direct and inverse probability

In the 1770s and 1780s, Laplace had developed Bayes's rule, which Thomas Bayes and his friend Richard Price had formulated only for the elementary case of independent trials of an event with constant probability. In the 1780s, Laplace had tried to base a theory of errors on his generalization of Bayes's rule, but he had been stymied by an inability to calculate the distribution of averages or other functions of more than a few observations.

Laplace's 1810 theorem solved this problem of calculation in a spectacular way. Not only could Laplace now calculate probabilities for the average of many independent quantities; he could do so without even knowing probabilities for the individual quantities. Bayes's rule also now seemed less interesting. The very concentrated normal distribution of the average would dominate any prior probabilities, so that Bayes's rule 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 proposed for a theory of errors.[7] Laplace did not disown Bayes's rule, but he de-emphasized it in his *Théorie analytique*, and in the applications of his theorem he usually just stated the Bernoullian argument [58, 20]. This inattention to the difference between Bayesian and Bernoullian arguments 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 Antoine Augustin Cournot) they rejected Bayesian arguments or whether (like Edgeworth and Pearson) they saw a Bernoullian argument as a shortcut to a Bayesian answer with roughly uniform prior probabilities.

Laplace did not give names to the two modes of argument that I am calling Bernoullian and Bayesian. In 1838 the British mathematician Augustus De Morgan called them *direct probability* and *inverse probability*, respectively [28, 128],[8] and these terms were widely used in English for over a century. Since the 1970s, the two modes have been called *frequentist* and *Bayesian.* Because *frequentism* also names a view about the meaning of probability, clarity may be served if, as I have done here, we use *Bernoullian* instead of *frequentist* for the non-Bayesian mode of argument.[9]

The 19th-century Bernoullian arguments produced what we now call *confidence intervals* and *two-sided significance tests.* The widespread belief that these concepts were invented only in the 20th century may be due in part to

[6] The scale of this correspondence has only recently become known, with Roger Hahn's publication of Laplace's surviving correspondence by [57]; for a summary, see [20, vol. 2, pp. 455ff].

[7] This was explained very clearly by Cournot in the binomial case; see §A.6. Bienaymé also discussed aspects of this picture; see [61, p. 102].

[8] Fourier had apparently used the corresponding French terms earlier in his teaching [26].

[9] Authors who have used *Bernoullian* in this way include Edgeworth [36], Richard von Mises [123, p. 5], A. P. Dempster [30], and Ian Hacking [56]. See Appendix B.

some authors' reluctance to recognize the merit and coherence of statistical theory not based on 20th-century foundations for probability theory; see §A.23.

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, computational accuracy and efficiency, 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 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 dominated astronomy and geodesy.[10] The Laplacean school, which sought practical certainty from large samples, continued to find adherents in the human sciences.

We can gain insight into 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 [74], Bessel's and Fourier's tables appear to be the first published tables of the normal distribution.[11]

Bessel's table appears in an article on the orbit of Olber's comet, in a passage translated in §A.3. In this passage Bessel explained 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

[10] The story of the triumph of the Gaussian theory in geodesy has been told in an enlightening dissertation by Marie-Françoise Jozeau [66].

[11] 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 [52, p. 79]. Kramp's table 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.

normally distributed variable falls more than α probable errors from its mean:

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

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 that 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 I will call the *modulus*, following 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, that a normal variable will be more than about 2.86783 moduli (or about 4.0557 standard deviations) from its mean.

Fourier's table differs from 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 in France

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 of errors to statistics. In his 1829 report, for example, he gave estimates and error limits for the average age, in Paris during the 18th century, of men and women when they married and when their first son was born [46, Table 64].

Such applications soon made what we now call significance testing popular.[12] 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 [97]. 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 [98]. 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 whose brilliant study of crime in France was commissioned by the Academy of Sciences in 1833 [55, 48].

Cournot, twenty years younger than Poisson, was himself a proponent and brilliant expositor of the Bernoullian version 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 *p-hacking*. In 1843, safely after Poisson's death, Cournot published his own book on probability, *Exposition de la théorie des chances et des probabilités* [24]. In a passage reproduced in translation in §A.6, Cournot observed that statisticians had been searching 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.

[12] As we know, such tests were already being published and debated in the 18th century. The first example usually cited is John Arbuthnot's argument, from boy births outnumbering girl births in London 82 years running, that the ratio must be governed by divine providence rather than chance. A good account of this and subsequent significant tests by Daniel Bernoulli, John Michell, and Laplace is provided by Anders Hald [58]. Barry Gower has provided more philosophical reviews [53, 54].

2.5 The disappearance of Laplace's theorem in France

Uncertainty measured by a p-value is often the least of our uncertainties when we are working with 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 relevant measurements 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 Laplace's death, the British astronomer George Biddell Airy showed that the true ratio lay well outside these bounds. Laplace's supreme confidence, whether in his model, his data, or his calculations, had been misplaced [20, p. 492].

This was only one reason for 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" [24]. Cournot's friend Jules-Irénée Bienaymé further developed Laplace's theory but spent most of his energies combating faulty applications [19, notes 27, 29, 62],[13, 61]. 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 [56]. 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 his followers [66]. Mathematicians and philosophers found many other problems in the Laplacean theory [67, 102]. By the end of the century, the most prominent mathematician in France, Joseph Bertrand, would ridicule Laplace's entire undertaking as a delusion [10]. 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 foreign colleagues.

3 Practical certainty

Now we look at how Laplace's theory and Fourier's criterion for practical certainty evolved in the 19th and into the 20th century. 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 approximate 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 is practically certain to include the true value of the quantity being estimated.

Today we measure a 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 [95], and George

Table 1: Geroge Biddell Airy's table relating different measures of dispersion for the normal distribution [1, p. 24]. Airy's *error of mean square* is the standard deviation. His *mean error* is the mean of the absolute deviation.

	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

Udny Yule introduced *standard error* in 1897 [133]. 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 still widely used in the first decades of the 20th century.[13]

3.1 *La limite de l'écart*

Equating very high probability with moral certainty is an ancient idea. The 16th-century Jesuit Luis Molina even applied it to games of chance [71, 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 settle on 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.[14] 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 1 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 exigent standards for certainty in theoretical discussions but then relaxed them in applications. Siméon-Denis Poisson, in his 1837 book on probability, first mentioned 4 or 5 moduli but relaxed this

[13] Helen Walker's history, written in 1929 [128], is still a good reference for how various authors used and named the various measures.

[14] For Molina and Bernoulli, a *morally certain* thing was one sufficiently probable that it should, by custom or some other norm, be treated as fully certain. For the purposes of this paper we may equate their concept of moral certainty with the later concept of *practical certainty*.

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 §A.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 §A.7.

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 [27] and Galloway's in 1839 [49]. Galloway's may have been more influential among mathematicians, 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 §A.4.

Cournot, in his 1843 book [24], followed Fourier in treating the probability of 1 in 20,000, corresponding to ±2.87 moduli, as practical certainty. He did not round 2.87 to 3. In §35 of the book, he recommended that the limit 2.87 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 deviation.[15] Unlike Poisson and most of the earlier authors, Cournot explicitly 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 taught in the middle of the 20th century; the basic concepts are all there. Cournot's "limite de l'écart" was used in 20th-century French teaching of statistics until after World War II.

Although they used the modulus, Laplace, Fourier, Poisson, Gavarret, Galloway and Cournot did not use the name *modulus* for it. This usage apparently first appeared in English, in George Biddell Airy's 1861 book on the theory of errors [1]; see §A.5.

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* [79], published in 1875, Lexis repeatedly used 3 moduli as his level for practical certainty; see §A.8.

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

[15] This terminology may have already been established in the 1820s. In his 1826 report, Fourier mentions that the difference between an observed and a true mean is called the "erreur ou écart" [45, p. 3].

given by the German astronomer Johann Franz Encke, a follower of Gauss, in an 1832 article on least squares [37]. 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 [84, p. 180]. In 1838 [28], 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 [24], 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 Outlying observations

In the late 19th-century, United States scholars looked to Germany for intellectual leadership, and many of those interested in probability enlisted in the Gaussian school of least squares.[16] On one topic to which they made original contributions, however, the Americans brought Gaussian least squares into closer contact with Laplacean practical certainty. This topic was the rejection of discordant or outlying observations.

The best known suggestions were made by Charles Sanders Peirce in 1852 and William Chauvenet in 1863 [111]. Chauvenet suggested that the observation that deviates most from the average of m observations be rejected if the probability of any particular observation deviating as much is less than $1/2m$. Peirce's and Chauvenet's proposals were strongly criticized by Airy and other European experts on least squares. In 1917, David Brunt reported that their criteria for rejection had not been widely used. (Brunt's assessment is quoted more fully in §A.21.) Consideration of the problem of outliers forces us, however, to think about about how the probability of extreme or otherwise unlikely values increases as we look for them, and this was one of the ways the statistical literature in English began to acknowledge the problem of p-hacking.

3.4 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 [33]. The substance of the paper, as Edgeworth conceded in the discussion after the reading, was largely borrowed from Lexis.[17] Edgeworth's

[16] See, for example, the quotation by Mansfield Merriman in §A.10.

[17] 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

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 §A.9.

An observed difference is significant in the sense of signifying 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 this 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 was anything but consistent, and as his paper rambled on, we do find one occasion where he writes "highly significant". But for the most part he used "significant" to mean "signifying" something causal and not accidental dominated. This usage seems odd to this speaker of American English in 2020. 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.[18] But Karl Pearson and his disciples understood it and adopted it. They used it for thirty years or more. As documented in §A.12, it persisted into the 1920s in Pearson's *Biometrika*. The United States biometrician Raymond Pearl, a student of Pearson's, explained it very clearly in a book he published in 1923; see §A.18.

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

3.5 Enter the American psychologists

Edgeworthian significance disappeared in the 1920s. The word remained, but the meaning shifted. This seems to have been 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 [113], psychologists had begun using mathematical statistics in the 1860s

and was self-taught in mathematics, it would have been natural for him to seek the latest wisdom from a German authority.

[18] Steve Stigler discusses Edgeworth's odd style on pp. 95–97 of [114].

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 *level of significance* in connection with statistical testing; see §A.25.

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 [14, 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" [15]. 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 [116].

In his 1923 textbook *Statistical Method* [68] 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 vague assertion is not quite Edgeworthian, for the comparison of an estimate with its probable errors often leaves us uncertain whether it signifies in the Edgeworthian sense. On page 102, at the beginning of a lengthy passage quoted in §A.20, 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.[19] 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) [116, p. 560].

3.6 Fisher's *Statistical Methods for Research Workers*

In 1925, Fisher published his own statistics manual, his celebrated *Statistical Methods for Research Workers* [42]. The words *significant* and *significance* are prominent in this book, but their Edgeworthian meaning has slipped away 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

[19] Personal communication from Lawrence Hubert, who has examined the Kelley archive at Harvard. See also [7].

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 instead defined it directly in terms of tail probabilities. Two standard deviations was replaced by 5% [115].

None of these features was unprecedented. Edgeworth's *significance* was already fading in some quarters. Yule had emphasized standard errors in his popular 1911 textbook (see §A.15). We saw Fourier mentioning odds of 19 to 1 in 1826, and in 1919 David Brunt used these odds in another context where practical certainty could not be measured in terms of probable errors (see §A.21).

Fisher's tone does not suggest that he had 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 vocabulary less from *Biometrika* than from the U.S. psychologists or others distant from Pearson. Perhaps this included "research workers" at the agricultural experiment station at Rothamsted, where he had already been working for five years.

3.7 Seeing p-hacking in Britain

As soon as we have significance testing, we have p-hacking. We can find Laplace himself varying his method of testing in search of a small p-value. Poisson searched across categories to find apparently significant ratios of male to female births. In the early volumes of *Biometrika*, we can find authors giving lists of estimated differences and their probable errors and then calling attention to one or two cases where the ratio of the estimated difference to its probable error appears to signify. But what makes this p-hacking visible to statisticians? Do we need a mathematical philosopher like Cournot to see it?

If you begin, as many 19th- and early 20th-century British mathematicians did, by assuming that probability is a relation between proposition and evidence, then you may find it paradoxical that the search producing the evidence should matter, and this may make it difficult to see p-hacking. But the historical record reveals at least two cases where the search was too obvious for some British mathematicians to ignore. The first was the problem of rejecting discordant observations (outliers); the second was the search for cycles in time series using harmonic analysis.

Discordant observations. The erudite Edgeworth knew Cournot's work. When he addressed the problem of discordant observations (outliers) in 1887, Edgeworth drew on Cournot's insights to understand the need to take account of the number of observations in deciding whether extreme observations should be considered discordant. Cournot's view was paradoxical, Edgeworth thought, but right ([34, pp. 369–370] quoted in §A.9). The next year, John Venn, in his *Logic of Chance*, acknowledged Cournot's and Edgeworth's point in the course of a discussion of fallacies in probability judgement [120, 3rd edition, pp. 338–339].

I have not found found examples before the 1920s in which Pearson or his disciples cited Cournot's insights concerning p-hacking or Edgeworth's related insights concerning discordant observations. Perhaps they were unaware of these insights and unaware in general of the hazards of p-hacking. But Pearson certainly knew Edgeworth's work, and silence in print about a problem that one cannot solve is not quite evidence of lack of awareness. The alternative hypothesis, that Pearson's circle did talk about p-values being invalidated by search or selection, is supported by a passing reference to the problem by Pearson's son, Egon S. Pearson, in 1925: Egon called it"the old difficulty". A decade later, when Egon and Chidambara Chandrasekaran noted the illegitimacy of choosing a test after seeing the data, the context was precisely Edgeworth's: the rejection of outliers. (See §A.24 for references and fuller quotations.)

Searching for cycles. In the case of cycles, it was the meteorologists who instructed the biometricians.

The notion that cycles in time series might be discovered and used for prediction was very popular at the end of the 19th and beginning of the 20th centuries, when William Stanley Jevons and other respected scholars even conjectured a causal chain from sunspots to business cycles: cycles in sunspots might cause cycles in the weather, hence cycles in agricultural production and other economic activity [85, 70, 47]. 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 [103]. 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 [104] and §A.11.)

Looking for cycles in a time series is a way of searching through the data for something remarkable. One of the first, perhaps the very first, to point out how misleading this particular type of search can be was Gilbert T. Walker, a physicist working as a meteorologist in India; see §A.13. Walker's first critique was published in India, in 1914 [125]. The same point was made in 1919 by the physicist F. J. W. Whipple in the discussion of a paper on cycles in the Greenwich temperature records, read to Royal Meteorological Society by David Brunt [22]; see §§A.21 and A.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 the Royal Statistical Society on cycles in rainfall and wheat prices. Beveridge read his paper, in which he more or less used Schuster's test, on April 25 of that year [12]. 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 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 mildly defended Beveridge against those who found his case entirely implausible. Fisher questioned how strongly periodicities in the weather could be related to periodicities in production, reporting that the total rainfall at Rothamsted accounted for a relatively small amount of the variation in production.

Walker re-entered the story before December 30. 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 [126]; see §§A.13 and A.17. Walker made 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 or more times as large as 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 own conclusions emerged intact.

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 address 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 [41, p. 94–95], he derived the adjustment required when one variable is chosen from several for a regression. 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 [43].

There are, however, no cautions about p-hacking in Fisher's first edition of *Statistical Methods*. Why did he 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*, published in 1936 (p. xii), we finally see a recipe of sorts for dealing with selection, gingerly offered: perhaps a very high level of significance, such as 0.1 per cent, should be used; see §A.22. 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 mathematical theory.[20]

[20] I am indebted to John Aldrich for calling my attention to Fisher's 1925 article and this

3.8 Who invented the name *p-value*?

No one. The term simply evolved from the 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 Roman upper case, in Fourier, Poisson, Gavarret, and Cournot. Beginning at least in his 1900 article on χ^2 [96], 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". Yule and Fisher followed Pearson's example throughout their careers [134, 40, 44].

By the 1920s, some authors outside Britain had occasionally and casually turned *value of P* into *P value*. The earliest examples I have seen are in articles by the North American geneticist John W. MacArthur in 1926 [80, pp. 397, 400], the United States biostatistician Persis Putnam in 1927 [100, pp. 672–673], and the Chinese statistician C. P. Sun in 1928 [117, p. 67].[21] Later authors who used *P value* or *P-value* casually include John Wishart [131, p. 304] and his associate H. O. Hirschfeld in 1937 [62, p. 68][22] and W. Edwards Deming in 1943 [29, p. 30]. None of these authors gave any indication that they thought they were coining a novel usage. I have not yet seen any use of *P value* in the social sciences before 1970; we do not see it in any of the articles in [86].

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, italics and roman, text and mathematical font, with and without the hyphen.

4 Conclusion

In 1949 [132, 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.

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 to correct these problems. Some propose to shift the level

preface.

[21] I am indebted to Sander Greenland for calling my attention to these articles.

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

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.

What lessons should we draw from the durability of these basic ideas and of their abuses? Everyone will answer this question for themselves, but the durability does suggest that superficial changes in terminology will not be helpful. The quick emergence of significance testing and p-values from Laplace's 19th-century confidence intervals further suggests that urging people to return to confidence intervals and stay there may be equally futile.

The failure of 20th-century embellishments to forestall misunderstanding and abuses also suggests that we might well simplify our elementary teaching by emphasizing Fourier's and Cournot's Bernoullian calculation of limits on error based on the central limit theorem, treating p-values for other tests as an intuitive extension of this basic example. This simplification would allow space in the curriculum to emphasize the inevitability of p-hacking and the necessity of treating some calculations as merely exploratory analysis. It would also allow some space to teach about testing by betting.

4.1 Can we return to practical certainty?

In March 2019, hundreds of statisticians lent their support to a commentary in *Nature* entitled "Retire statistical significance" [2]. What should replace it? How is a scientist or journalist inexpert in statistical mathematics to interpret a p-value?

Here is a fanciful thought experiment. Suppose we all (all the teachers and textbook writers in statistics) reach a consensus to return to Edgeworthian significance. And suppose we signal this change by replacing *significant* with *signifying*. When a test is fully planned in advance,

- a p-value of 0.05 (about three probable errors or two standard deviations) is likely to signify;
- a p-value of 0.005 (about four probable errors or two moduli) is practically certain to signify;
- a p-value of 0.00006 (about six probable errors or four standard deviations) definitely signifies.

What problems would this pose? The obvious problem is the "likely" in "likely to signify". Shall we give it a Bayesian interpretation with a uniform prior, as Edgeworth did? A Bernoullian interpretation as Cournot and Jerzy Neyman would? Here the fancied consensus splinters.

In my view, we do not particularly need Edgeworth's *significant* or *signify* in the teaching of elementary statistics, but there is much to be said for returning to the simple language of Bernoullian practical certainty as it was used by Cournot in 1843. Leaving aside any appeal to an axiomatic theory of probability, we could teach students how to find reasonable limits on error, and then use these limits to obtain confidence intervals, just as Cournot, Neyman, and countless statisticians between them did. Then, following Gavarret, we could translate this into a two-sided test for a difference between two treatments: one treatment is practically certain to be better than the other when its observed advantage is greater than the limit of possible error; otherwise we are not authorized to pronounce in favor of either treatment; see A.7.

When Jacob Bernoulli first undertook to base judgements of *probability* in civic, moral, and business matters on calculations like those in games of chance, he did not envision counting chances for and against every civil, moral, and business question. He certainly did not suppose that every proposition would have probabilities for and against adding to one. On the contrary, he thought chances could be counted only for some propositions, and he hoped that arguments that compared and combined these counts could then cast light on other propositions. This might not produce probabilities that follow Kolmogorov's axioms, but it might sometimes lead to practical certainty [105]. Gavarret's explanation of how to use a "limit of possible errors" stands squarely in this Bernoullian tradition, which may allow us to conclude that a proposition is practically certain without assigning it a numerical probability.

Critics of the current pedagogy of elementary mathematical statistics have complained that it has become a confusing mixture of competing theories and ideas, mixing Fisherian, Bayesian, and decision-theoretic viewpoints [32, 51, 63]. To a large extent, these competing theories are different ways of forcing the judgements of practical certainty taught by Laplace, Fourier, and their 19th-century successors into the 20th-century's measure-theoretic framework for mathematical probability. Must we do this?

4.2 Can we return to exploratory data analysis?

Beginning in the late 1940s, when the center of research in mathematical statistics had shifted to the United States, there was a great deal of work on formal methods for *multiple comparisons*—a name appropriate when a plan is made in advance to test many different hypotheses and significance levels are adjusted accordingly [107, 82]. John Tukey, one of the leaders in this work [9], was acutely aware that it was not always relevant to practice. In practice, scientists often cannot plan in advance what they will think of doing later. One way of dealing with this problem is to drop the ambition of arriving at practical certainties in a particular study, treating the study as purely exploratory. With this in mind, Tukey coined the term *exploratory data analysis*, to be distinguished from *confirmatory data analysis.* Promising results from an exploratory study were not to be taken seriously as science until they were confirmed by later studies.

The sociological realities of academia have not been kind to exploratory data analysis. It has not proven to be a path to publication and prestige, and it is now often thought of as merely a collection of techniques for the visual and tabular presentation of data. Some contributors to recent discussions of significance testing in *The American Statistician* have suggested putting new emphasis on the exploratory/confirmatory contrast [129]. If we were to simplify the discussion of testing by returning to Gavarret's simple framework, there might be room for this in elementary statistics instruction.

4.3 Can we return to betting?

At its core, probability theory is a calculus of betting, but in the three centuries since Bernoulli, mathematical statistics has been increasingly formalized and taught in ways that hide this fact. My own proposal for fundamental change in statistical testing is to bring betting back to the center of the picture.

To do this, we must shift from the Bayesian statement that

> assuming model B, the observations make hypothesis A unlikely

and the Bernoullian statement that

> assuming B, the observations are unlikely if A is true

to a statement about a bet:

> a bet that was fair according to B has paid off handsomely (multiplied the money it risked by a large factor) if A is true.

When we do this, we may say that we are testing the hypothesis A by betting against it. This has two obvious advantages. First, it makes very salient the need to state the bet in advance; no one is impressed when you identify after the fact a bet that would have been successful. Second, it makes very salient the remaining uncertainty; no one forgets that success in betting, no matter how striking, may have been accidental.

As I argue in [106], testing by betting can also help us give more honest accounts of opportunistic searches for significance. A honest Bayesian or Bernoullian account of such a search requires the specification of a complete strategy for the search. What would you have done if the first test had come out differently, etc.? If the search is opportunistic, such a plan can only be conjectured after the fact and can hardly ever be convincing. When we test by trying to multiply the money we risk, no grand strategy is required; we can change direction opportunistically so long as each successive bet is honest (made before we look) and risks only the net capital resulting from the preceding bet.

5 Acknowledgements

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

Appendices

Appendix A What they said

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

In translations, I have sometimes 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.

A.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 [90]. 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, but there is no doubt that he edited them and wrote the mathematical memoirs that appear at the beginning of the 1826 and 1829 ones [19, 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 [45, pp. xxi–xxii].

Fourier in English

To complete this discussion, we must find the probability that H, the quantity sought, is between proposed limits $\mathrm{A} + \mathrm{D}$ and $\mathrm{A} - \mathrm{D}$. 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 $\mathrm{A} + g\partial$ and $\mathrm{A} - g\partial$. Here A is the average result of a large number m of particular values $a, b, c, d, \dots, n$; ∂ 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 $a^2, b^2, c^2, d^2, \dots, n^2$ 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 $6 + 0.246$.

The French original

Pour compléter cette discussion , il faut déterminer quelle probabilité il y a que la quantité cherchée H est comprise entre des limites proposées $\mathrm{A} + \mathrm{D}$ et $\mathrm{A} - \mathrm{D}$. A est le résultat moyen que l'on a trouvé, H est la valeur fixe que donnerait un nombre infini d'observations , et D est une quantité proposée que l'on ajoute à la valeur A ou que l'on en retranche. La table suivante fait connaître la probabilité P d'une erreur positive ou négative plus grande que D; et cette quantité D est le produit de g par un facteur proposé ∂.

∂	P
0,47708	$\frac{1}{2}$
1,38591	$\frac{1}{20}$
1,98495	$\frac{1}{200}$
2,46130	$\frac{1}{2000}$
2,86783	$\frac{1}{20000}$

Chacun des nombres de la colonne P fait connaître quelle probabilité il y a que la valeur exacte H, qui est l'objet de la recherche , est comprise entre les limites $\mathrm{A} + g\partial$ et $\mathrm{A} - g\partial$. A est le résultat moyen d'un grand nombre m de valeurs particulières $a, b, c, d, \ldots, n$; ∂ est un facteur donné; g est la racine carrée du quotient que l'on trouve en divisant par m le double de la différence de la valeur moyenne des carrés $a^2, b^2, c^2, d^2, \ldots, n^2$ au carré A^2 du résultat moyen. On voit par cette table que la probabilité d'une erreur plus grande que le produit de g et 0,47708, c'est-à-dire, plus grande qu'environ la moitié de g, est $\frac{1}{2}$. II y a 1 contre 1 ou 1 sur 2 à parier que l'erreur commise ne surpassera pas le produit de g par 0,47708, et il y a autant à parier que l'erreur surpassera ce produit.

La probabilité d'une erreur plus grande que le produit de g par 1,38591 est beaucoup plus petite que la précédente; elle n'est que $\frac{1}{20}$. II y a 19 sur 20 à parier que l'erreur du résultat moyen ne surpassera pas ce second produit.

La probabilité d'une erreur plus grande que la précédente devient extrêmement petite, à mesure que le facteur ∂ augmente. Elle n'est plus que $\frac{1}{200}$ lorsque ∂ approche de 2. La probabilité tombe ensuite en dessous de $\frac{1}{2000}$. Enfin, il y a beaucoup plus de vingt mille à parier contre 1 que l'erreur du résultat moyen sera au-dessous du triple de la valeur trouvée pour g. Ainsi , dans l'exemple cité art. VI, où l'on a 6 pour le résultat moyen, on peut regarder comme certain que cette valeur 6 n'est pas en défaut d'une quantité triple de la fraction 0,082 que la règle a donnée pour la valeur de g.

La grandeur cherchée H est donc comprise entre $6 - 0,246$ et $6 + 0,246$.

A.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 [58, §17.3]. In 1830, he gave straightforward instructions for calculating limits of practical certainty for the difference between two proportions [98].[23] Finally, in 1837, he pulled together his theoretical and applied results on probability in an impressive treatise, *Recherches sur la probabilité des jugements* [99].

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 [99, §§80, 87, and 96]. But when analyzing data, he used 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 times 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 the unknown probability has not changed, for the estimated modulus of the difference $\frac{m'}{\mu'} - \frac{m}{\mu}$ and for 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.

> Si donc on a pris pour u un nombre tel que trois ou quatre, qui rende la probabilit'e $\tilde{\omega}$ très approchante de la certitude (n° 80), et si, néamoins, l'observation donne pour $\frac{m'}{\mu'} - \frac{m}{\mu}$ ou $\frac{n'}{\mu'} - \frac{n}{\mu}$ des valeurs qui s'écartent notablement de ces limites,on sera fondé à en conclure, avec une très grande probabilité, que les chances inconnu des évévements E et F ont changé, dans l'intervalle des deux séries d'épreuves, ou même pendant ces épreuves.

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

[23] In his second memoir on mathematical statistics, in 1829 [46], 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 the case where this function is simply the difference between two proportions.

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 its modulus was 0.00335. He states that if we use 2 moduli,

> . . . we will also have
>
> $$P = 0.9953,$$
>
> 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.

> . . . on aura aussi
>
> $$P = 0.9953,$$
>
> pour la probabilité, très approchante de la certitude, que l'inconnue R_5 et la fraction 0,6094 ne diffèrent pas de 0,0067, i'une de l'autre.

A.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 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 [11]. 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 $\pm\alpha \times$ (probable error) for values of α up to 4. The odds are more than 140 to 1 that it fall within $\pm 4 \times$ (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 recommends estimating the probable error 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:

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

The German original

Der Erfolg der Bestimmung der endlichen Elemente aus diesen Bedingungsgleichungen, noch mehr aber die Schätzung ihrer wahrscheinlichen, aus den Beobachtungsfehlern entstehenden Unsicherheit, hängt hauptsächlich von der richtigen Würdigung der Bedingungsgleichungen ab. Es war daher nothwendig, über diesen Gegenstand eine eigene Untersuchung anzustellen, deren Resultat ich bereits seit einigen Jahren mit Vortheil benutzt habe.

Nach der von Gauss gegebenen Theorie der kleinsten Quadrate ist die Wahrscheinlichkeit, einen Fehler Δ zu begehen,

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

(*Theoria mot. corp. coel. P. 212.*), wo h von der Genauigkeit der Beobachtungen abhängt. Mittelst dieses Ausdrucks kann man leicht aus einer vorhandenen Reihe von Beobachtungen den wahrscheinlichen Fehler einer einzelnen bestimmen, unter der Voraussetzung, dass die wirklich vorkommenden Fehler von allen beständigen Einwirkungen frei, und nur durch die Unvollkommenheiten der Instrumente und Sinne erzeugt sind. Man hat nämlich, desto näher, je grösser die Anzahl der Beobachtungen ist, das arithmetische Mittel aus allen Fehlern, sämmtlich mit gleichem Zeichen genommen, welches wir ϵ nennen wollen,

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

und auch die Quadratwurzel aus dem arithmetischen Mittel der Quadrate der Fehler, welche wir durch ϵ' bezeichnen wollen, aus der Gleichung

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

Je zahlreicher nämlich eine vorhandene Beobachtungsreihe ist, mit desto mehr Rechte wird man annehmen können, dass die Fehler darin so vorkommen, wie es die Gausssche Theorie erfordert; das aus der Vergleichung einer sehr zahlreichen Reihe mit einer ihr so gut als möglich entsprechenden Theorie folgende ϵ oder ϵ', wird nun den wahrscheinlichen Fehler einer Beobachtung, den ich durch ϵ'' bezeichnen werde, geben. Ich verstehe unter dieser Benennung die Grenze, die eine Anzahl kleinerer Fehler von einer gleichen Anzahl grösserer trennt, so dass es wahrscheinlicher ist, eine Beobachtung innerhalb jeder weiteren Grenze von der Wahrheit abirren zu sehen, als ausserhalb derselben.

Durch die Auflösung der Gleichung

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

findet man $x = 0,4769364 = h\epsilon''$, so dass man hat

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

Die Wahrscheinlichkeit eines Fehlers, kleiner als $\alpha\epsilon''$, verhält sich zu der eines grössern, wie der Werth des Integrals $\int e^{-t^2} dt$ von $t = 0$ bis $t = \alpha \times 0.4769364$, zu dem Werthe desselben Integrals von $t = \alpha \times 0,4769364$ bis $t = \infty$ genommen. Für einige Werthe von α findet man, aus den bekannten Tafeln dieses Integrals:

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

A.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* [49, 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 [94, pp. 177, 180].

When tabulating normal probabilities, Galloway wrote τ for the number of moduli and Θ for the corresponding probability—the probability that a quantity following the error law is within τ moduli. His table stopped at $\tau = 3$; and he seems to have agreed with Fourier that this was the limit of practical certainty.

In his discussion of Bernoulli's theorem on page 144, for example, he pointed 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 $\Theta = \cdot 9999779$".

A.5 George Biddell Airy, 1801–1892

A prolific mathematician, Airy was the British Astronomer Royal from 1835 to 1881. His book on the theory of errors, entitled *On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations* and first published in 1861, might be considered the first manual on the Gaussian theory of errors published in English. Curiously, however, he relied on Laplace for this basic theory and did not mention Gauss's name. He included this statement in the preface to the first edition (p. vi):

> No novelty, I believe, of fundamental character, will be found in these pages. At the same time I may state that the work has been written without reference to or distinct recollection of any other treatise (excepting only Laplace's *Théorie des Probabilités*); and the methods of treating the different problems may therefore differ in some small degrees from those commonly employed.

Further editions appeared in 1875 and 1879.

On p. 15 of the book, Airy introduced the name *modulus* for the constant c in the expression

$$\frac{1}{c\sqrt{\pi}}.e^{-\frac{x^2}{c^2}}.\delta x$$

for the probability that an error falls between x and $x + \delta x$.

As customary in the Gaussian tradition, Airy did not discuss practical certainty. When discussing the law of error, on p. 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.

A.6 Augustin Cournot, 1801–1877

In 1833, Cournot published a translation into French of John Herschel's *Treatise on Astronomy*, which had appeared in English that same year. In an appendix to the translation, he discussed the application of probability to astronomical observations [81]. Here we find this statement about practical certainty [25, vol. XI.2, p. 686].

> A probability of 1000 to 1 is almost considered equivalent to certainty, and one can hardly make the same judgement about a probability of 12 to 1.

Already in 1828, Cournot had already began working on the ideas that became his *Exposition de la théorie des chances et des probabilités*, but he finally completed and published the book only in 1843.[24] Below I translate two passages from the *Exposition*, a brief passage from §95 discussing the agreement between Bernoullian and Bayesian methods, and an extended passage from §111 criticizing the p-hacking of the census. Cournot discussed p-hacking 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 [18] and Michel Armatte [3, 4].

Cournot on Bayes, from [24, §95]

When the numbers . . . are very large, . . . , the result from Bayes's rule no longer differs noticeably from the calculation that Bernoulli's theorem would give. This is the way it should be, because the truth of Bernoulli's theorem is independent of any hypothesis concerning the initial choice of the urn. In this case it is not (as many authors seem to have imagined) Bernoulli's rule that becomes exact by approximating Bayes's rule; it is Bayes's rule that becomes exact, or acquires an objective value that it did not have before, by coming together with Bernoulli's rule.

The French original. Quand les nombres . . . sont très grands, . . . le résultat trouvé par la régle de Bayes ne diffère plus sensiblement de calcul que donnerait le théorème de Bernoulli. Il faut bien qu'il en sont ainsi, puisque la vérité du théorème de Bernoulli est indépendante de toute hypothèse sur le tirage préalable de l'urne. Ce n'est point dans ce cas (comme beaucoup d'auteurs ont paru se le figurer) la régle de Bernoulli qui devient exacte en se rapprochant de la règle de Bayes; c'est la règle de Bayes qui devient exacte, ou acquiert une valeur objective qu'elle n'avait pas; en se confondant avec la règle de Bernoulli.

Cournot on p-hacking, from [24, §111]

. . . 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

[24] For Cournot 1828 article on probability, together with commentary by Bernard Bru and Thierry Martin, see pp. 442–453 of Volume XI-1 of Cournot's complete works [25]. See also the edition of the *Exposition* that appeared in 1984, with introduction and notes by Bru, as Volume I of the complete works.

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.

The French original

...Il est clair que rien ne limite le nombre des faces sous lesquelles on peut considérer les événements naturels ou les faits sociaux auxquels s'appliquent les recherches de statistique, ni, par suite, le nombre des caractères d'après lesquels on peut les distribuer en plusieurs groupes ou catégories distinctes. Supposons, pour prendre un exemple, qu'il s'agisse de déterminer, d'après un grand nombre d'observations recueillies dans un pays tel que la France, la chance d'une naissance masculine qui, en général, comme on le sait, surpass 1/2: on pourra distinguer d'abord les naissances légitimes des naissances hors mariage; et comme on trouvera, en opérant sur de grands nombres, une différence très-sensible entre les valeurs du rapport du nombre des naissance masculines au nombre total des naissances, selon qu'il agit d'enfants légitimes ou naturels, on en conclura avec une probabilité très-grande que la chance d'une naissance masculine, dans la catégoire des naissances légitime, surpasse sensiblement la chance du même événement, dans la catégoire des naissances hors mariage. On pourra distinguer encore les naissances dans les campagnes des naissances dans les villes, et l'on arrivera à une conclusion analogue. Ces deux classifications s'offrent si naturellement à l'esprit, qu'elles ont été un objet d'épreuve pour tous les statisticiens.

Maintenant il est clair qu'on pourrait aussi classer les naissances d'après l'ordre de primogéniture, d'après l'âge, la profession, la fortune, la religion des parents; qu'on pourrait distinguer les premières noces des secondes, les naissances survenues dans telle saison de l'année, des naissances survenues dans une autre saison; en un mot, qu'on pourrait tirer d'une foule de circonstances accessoires au fait même de la naissance, des caractères, en nombre indéfini, qui serviraient de base à autant de systèmes de distribution catégorique. Il est pareillement évident que, tandis que le nombre des coupes augmente ainsi sans limite, il est *à priori* de plus en plus probable que, par le seul effet du hasard, l'une des coupes au moins offrira, pour le rapport du nombre des naissances masculines au nombre total des naissancees, dans les deux catégories opposées, des valueurs sensiblement différentes. En conséquence, ainsi que nous l'avons déjà expliqué, pour le statisticien qui se livre à un travail de dépouillement et de comparaison, la probabilité qu'un écart de grandeur donnée n'est pas imputable aux anomalies du hasard, prendra des valeurs très-différentes, selon qu'il aura essayé un plus ou moins grand nombre de coupes avant de tomber sur l'écart observé. Comme on suppose toujours qu'il a opéré sur de grands nombres, cette probabilité ...n'en aura pas moins, dans chaque système d'essais, une valeur objective, en ce sens qu'elle sera proportionnelle au nombre de paris que l'expérimentateur gagnerait effectivement, s'il répétait un grand nombre de fois le même pari, toujours à la suite d'autant d'essais parfaitement semblables, et si l'on possédait d'ailleurs un *criterium* certain pour distinguer les cas où il se trompe des cas où il rencontre juste.

Mais ordinairement ces essais par lesquels l'expérimentateur a passé ne laissent pas de traces; le public ne connaît que le résultat qui a paru mériter de lui être signalé; et en conséquence, une personne étrangère au travail d'essais qui a mis ce résultat en évidence, manquera absolument de règle fixe pour parier que le résultat est ou non imputable aux anomalies du hasard. On ne saurait assigner approximativement la valeur du rapport du nombre des jugements erronés qu'elle portera, au nombre des jugements portés, même en supposant très-grand le nombre des jugements semblables, portés dans des circonstances identiques. En un mot, la probabilité que nous avons appelée Π, et qui correspond à l'écart δ, perdra, pour la personne étrangère aux essais qui ont manifesté cet écart, toute consistance objective; et, selon l'idée que cette personne se fera de la *valeur intrinsèque* du caractère qui a servi de base à la division catégorique correspondante, elle devra porter des jugements différents, la grandeur de l'écart signalé restant la même.

A.7 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 [50, 64]. 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, Gavarret appealed to Poisson's authority to support the 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,[25] 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:
>
> $$\mathrm{P} = 0.9953; \text{that is to say, betting odds of 212 to 1.}$$

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

> Mais pour rendre ces formules immédiatement applicable aux questions dont nous nous occupens, nous devons leur faire subir une transformation très simple. A cet effect, rappelons nous le principe général établi (pag. 39), à savoir que: du moment ou un observateur est arrivé à un haut degré de probabilité relativement à l'existence d'un fait, il peut l'employer comme s'il était absolument certain. Convenons donc d'une probabilité à partir de laquelle tout fait thérapeutique puisse et doive être admis sans contestation. Cette probabilité doit satisfaire à deux conditions importantes: l'une, d'être assez élevée pour ne laisser aucun doute doute dans les esprits; l'autre, de ne pas exiger un nombre trop grand d'observations, pour que les rapports fournis par la statistique recueillie se rapprochent convenablement de la chance moyenne cherchée. Le choix de cette probabilité, dont on puisse se contenter, eut été fort délicat à faire; mais heureusement nou pouvons nous appuyer à cet égard sur une autorité dont personne sans doute n'essayera de contester l'importance. Quand M. Poisson a exposé dans son ouvrages les règles qui devaient présider à la recherche des erreurs possibles dans les jugements du jury, la probabilité la plus élevée qu'il ait donné à ses propositions, pour se croire fondé à les considérer comme à l'abri de toute objection raisonnable, est:
>
> $$\mathrm{P} = 0.9953;\text{c'est à dire 212 à parier contre 1.}$$

Today we teach students to estimate the standard error for the difference between two estimated proportions $\hat{p}_1$ and $\hat{p}_2$ by the formula

$$\text{standard error } (\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where n_i, for $i = 1, 2$, is the number of individuals from which the proportion $\hat{p}_i$ is estimated. Gavarett used the same formula, with the additional factor $\sqrt{2}$, to estimate the modulus. He calculated 2 moduli using this formula for a number of examples. Here are two examples he gave on pp. 157–158:

- If one treatment results in 100 deaths out of 500 patients, while a second results in 150 deaths out of 500 patients, then we have an observed difference of 0.1, with a limit of possible errors 0.07694. Here Gavarret concludes:

  > The difference in calculated death rates is greater than this *limit of possible errors* in the *a posteriori* conclusion. So we must recognize that the first treatment really is better than the second.
  >
  > La *difference* entre les mortalités obtenues est supérieure à cette *limite des erreurs possibles* dans la conclusion *à posteriori*, nous devons donc reconnaître qu'en réalité la première médication est supérieure à la seconde.

- If one treatment results in 100 deaths out of 500 patients, while a second results in 130 deaths out of 500 patients, then we have an observed difference of 0.06, with a limit of possible errors 0.07508. Here Gavarret concludes:

  > Evidently, because the difference between the two death rates is less than this *limit of possible errors* in the *a posteriori* conclusion, we must recognize that this variation in the results tells us nothing, and that we are not authorized to prefer one of the two methods to the other.
  >
  > Évidemment, puisque la *différence* entre les deux mortalités moyennes est inférieure à cette *limite des erreurs possible* dans la conclusion *à posteriori*, nous devons reconnaître, que cette variation dans les résultats ne nous enseigne rien, que nous ne sommes pas autorisés à préférer une des deux méthodes à l'autre..

A.8 Wilhelm Lexis, 1837–1914

A prominent German economist and statistician, Lexis published his introduction to population statistics *Einleitung in die Theorie der Bevölkerungsstatistik* [79], 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 ....

In the original:

> Für die Zwecke der Statistik dürfte es jedoch geeigneter sein, für u eine so grosse Zahl zu nehmen, dass F_u der Einheit sehr nahe kommt, also eine Wahrscheinlicbkeit ausdrückt, die in der Praxis der Gewissheit gleich geachtet werden kann. Es genügt, wie oben,

> u gleich 3 zu setzen, und man erhalt alsdann die Wahrscheinlichkeit $F_3 = 0,999978$ für die Grenzgleichung ....)

A.9 Francis Edgeworth, 1845–1926

As noted in §3.4, 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 [33]. 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.

> 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* [35, §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.

The following passage appears in Edgeworth's 1887 article on the rejection of discordant observations [34, pp. 369–370]:

> There is something paradoxical in Cournot's proposition that a certain deviation from the Mean in the case of Departmental returns of the proportion between male and female births is significant and indicative of a difference in kind, provided that we select at random a single French Department; but that the same deviation may be accidental if it is the maximum of the respective returns for several Departments. There is something plausible in De Morgan's implied assertion that the deficiency of seven in the first 608 digits of the constant π is theoretically not accidental; because the deviation from the Mean 61 amounts to twice the Modulus of that probability curve which represents the frequency of deviation for any assigned digit. I submit, however, that Cournot is right, and that De Morgan, if he is serious in the passage referred to, has committed a slight inadvertence. When we select out of the ten digits the one whose deviation from the Mean is greatest, we ought to estimate the improbability of this deviation occurring by accident, not with De Morgan as $1 - \theta(1 \cdot 63)$, corresponding to odds of about 45 to 1 against the observed event having occurred by accident; but as $1 - \theta^{10}(1 \cdot 63)$, corresponding to odds of about 5 to 1 against an accidental origination.

A.10 Mansfield Merriman, 1848–1925

A prominent American mathematician and engineer, Merriman taught at Yale and Lehigh and worked at different times in his country's Corps of Engineers and Coast and Geodetic Survey. He enters our story as an authority on Gauss's theory of least squares. In one of his earliest publications on the subject, he wrote [83, p. 195]:

> To Gauss is also due the development of the algorithm of the method, the formulae for probable error, the determination of weights, the method of correlatives, and many other features of the subject, as well as numerous practical applications with which his writings abound. Very few branches. of science owe so large a proportion of subject matter to the labours of one man.

A.11 Arthur Schuster, 1851–1934

Schuster was born in Germany but in 1870 he 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 [103, 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 [104, 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.

A.12 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.[26]

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 [130, 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 · 0229 mm.), is less than twice the probable error of the determination.”

- In the second issue, from Oswald H. Latter [76, 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

[26] The early decades of *Biometrika* can be searched conveniently at the Biodiversity Library.

regarding differences of at least $2 \cdot 5$ times their probable errors as significant" [60, p. 458].

- In a 1912 article co-authored by Pearson himself [8, 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 [119, 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 [23, 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 [118, 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.

A.13 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 [125], 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 [126, p. 511] and in an article in the *Quarterly Journal of the Royal Meteorological Society* in 1925 [127].

Walker's 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,[27] 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

[27] 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.

A.14 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 [16], 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 [33] 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.

A.15 George Udny Yule, 1871–1951

After studying mathematical physics, Yule became a statistician as an assistant to Pearson. He later quarreled with Pearson and went his own way on a number of points. In 1897 [133], he introduced the name *standard error* for the standard deviation of an estimate. The first edition of his *Theory of Statistics* [134] appeared in 1911.

On page 262 of the first edition, we find this nod to Edgeworth's *significant*:

> ... 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**.

Yule never or hardly ever used the Edgeworthian *significant*, however. We find the word repeatedly in the book, but usually merely to mean "important".

A.16 Francis J. W. Whipple, 1876–1943

The following passage appears on p. 336 of the discussion of Brunt's paper on the Greenwich temperature records.

Mr. F. J. W. WHIPPLE called attention to a difficulty which was not always faced when the “reality” of periods was under discussion. In CaptḂrunt’s paper it was stated that if the amplitude of one of the harmonic components was as great as 06∘·4 F, the odds were “19 to 1 in favour of the reality” of the corresponding period, and yet the author arrived at the conclusion that such periods had little physical significance. The difficulty lay in the enunciation of the statement with regard to the probability of the occurrence of an amplitude of a specified magnitude. It might be illustrated by an example from the card-table: the chances of three aces occurring in a particular hand was very small, but the chance of three aces occurring in some hand in the course of an evening’s play was very high. In the case under consideration, if a particular period was named in advance and the corresponding amplitude turned out to be large, then the probability that this was not merely casual would be high; but if all possible periods were investigated then there was good reason to expect that some of the computed amplitudes would be large. It was stated that the standard (root-mean-square) value of the amplitudes computed from the author’s data, regarded as distributed by chance, would be 2 · 3 units, and as a matter of fact the sum of the squares of the amplitudes of the first forty-nine harmonics was given as 380, so that the root-mean-square was 2 · 8, not much in excess of the 2 · 3. Accordingly there was little reason to suppose that the periods were significant. Any agreement between the periods found for one element and another, for example. for temperatures at different places, would indicate a correlation between the elements and would merit further investigation. Periodogram analysis might prove a useful though laborious method for discovering such correlations.

A.17 William Beveridge, 1879–1963

When Beveridge published his harmonic analysis of hundreds of years of wheat prices in 1922, he was already well known because of his work in the British civil service before and during the First World War. In 1919, he had become director of the London School of Economics. During the Second World War, he led a committee that produced the *Beveridge Report*, which became a blueprint for Britain’s postwar welfare system.

Beveridge’s response to Gilbert Walker’s 1922 letter to the Editor of *Nature* was printed by the journal immediately after the letter.

Beveridge’s response. DR. WALKER’s note contains, I think, a valid and valuable criticism of the procedure commonly adopted hitherto in comparing individual intensities with the average intensity in harmonic analysis. It would lead me, now to modify in several ways my general discussion of the “test of intensity” (pp. 412–424 of my paper in the Journal of the Royal Statistical Society). I was particularly careful, however, in that paper to avoid laying stress on intensity as such. The net result of Dr. Walker's calculations is not to weaken but to confirm my main thesis: that a number of real periodicities exist in European wheat prices from 1550 to 1850.

According to these calculations, the chance of my getting by pure luck between five and forty years one intensity as great as $3a$ is $0 \cdot 926$, but the chance of my getting seven such intensities is $0 \cdot 0127$, and that of getting eleven is $0 \cdot 00003$. Actually I have, between five and forty years, fifteen intensities above $3a (= 17{\cdot}69)$; the odds are therefore 80 to 1 that at least nine of these intensities, and $33,000$ to 1 that at least five of them, are not due to luck. Obviously every such intensity does, in the circumstances, present a *prima facie* case for further inquiry, the object of the inquiry being to determine which of the 15 intensities have the strongest probabilities of being due to real periods.

In that inquiry the actual height of the intensity in any case (the "test of intensity") is only one and not necessarily the most important point for consideration. As Dr. Walker shows, an intensity in my periodogram of nearly seven times the average might well be due to pure luck (the odds being only 20 to 1 against it). On the other hand, a much lower intensity might represent a true and perfectly regular but weak periodicity, just as a quite small correlation co-efficient may prove a real though weak connexion, if the number of cases compared is very large. Indication of the same period in each half of a sequence when analysed separately (the "test of continuity") and in independent sequences (the "test of agreement with other records") are often more important criteria of reality than is the height of the intensity itself. The former test, at least, should never be neglected; it has led me to relegate to my fourth class as merely "possible," several periods, such as those near 11, 17, and 24 years, indicated by high intensities in the whole sequence, but failing in either the first or the second half.

Ultimately, of my fifteen intensities between 5 and 40 years, I have treated only nine (at $5 \cdot 100$, $5 \cdot 671$, $5 \cdot 960$, $8 \cdot 050$, $9 \cdot 750$, $12 \cdot 840$, $15 \cdot 225$, $19 \cdot 900$, and $35 \cdot 500$ years respectively) as certainly or probably due to real periodicities, because they show in all cases perfect or fair continuity and in most an agreement with other records. The smallest of these fifteen intensities ($21{\cdot}72$ at $7{\cdot}417$ years) in fact equals not $3a$ but $3 \cdot 683a$. If with this revised figure, the probabilities are calculated in the way suggested by Dr. Walker, the odds that at least nine of the fifteen intensities are not due to luck work out at more than 2000 to 1, while the odds in favour of seven at least are $14 \cdot 000$ to 1.

This remarkable result, which seems to establish beyond all reasonable doubt the reign of periodicities in wheat prices, is not affected by the fact that of the fifteen intensities only four are so high that any one of the four, if it occurred alone and had to be judged by height alone, would have odds of more than 20 to 1 in its favour. Each intensity does not occur alone. Every period, moreover, to which I attach importance rests on more evidence than mere height in my periodogram.

With reference to the last paragraph but one of Dr. Walker's note, on the relation of my synthetic curve and the rainfall, I should like to emphasise the point made in my paper (pp. 449–450) that the synthetic curve as now drawn represents only a first approximation of the roughest possible character; the correlation co-efficient of $0 \cdot 38$ between it and the rainfall from 1850 to 1921 is sufficient to demonstrate some connexion between the wheat price cycles and the rainfall, but is in no sense to be treated as a measure of the degree of connexion. In constructing the synthetic curve, for instance, the periodicities have all been treated as of equal importance; inspection shows that weighting according to the intensities would almost certainly give a better fit and so a higher co-efficient of correlation. In many other ways a more accurate determination of the cycles is required. How high a correlation might ultimately be obtained as the result of this, it is impossible now to say, but it might easily prove to be very high indeed. Unfortunately, I have no resources for carrying my own investigations further for the present; I can only hope that others may be better placed.

W. H. BEVERIDGE.

Aftermath. By the 1940s, British mathematicians had reached a consensus that the cycles detected by harmonic analysis of time series had little meaning or value for prediction. An important marker was an extensive paper read to the Royal Statistical Society in 1945 by Maurice G. Kendall. Kendall concluded the discussion of his paper by saying that, "the reason why people continually discover cycles in all kinds of time series, is that they are looking for them" [69, pp.140–141]. In another discussion of cycles at the Royal Statistical Society in 1946, we see this report on a comment by Harold Jeffreys (*Supplement to the Journal of the Royal Statistical Society*, Vol. 8, No. 1, p. 90): "DR. HAROLD JEFFREYS said that he had no experience in detecting empirical periodicities in geophysical data. He had a good deal of experience of failing to find evidence for them."

A.18 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* [91].

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.

A.19 Sergei Bernstein, 1880–1968

Born in Odessa, Bernstein[28] studied in Paris and Göttingen. His dissertation, submitted to the Sorbonne in Paris in 1904, solved Hilbert's 19th problem. He taught at the university at Kharkov from 1907 until 1932, when he moved to the Academy of Sciences in Leningrad; later he taught in Moscow.

In 1927, Bernstein published his course on probability as a book [5]. We can call the book Laplacean tradition, for it discusses practical certainty, and it often appeals to Laplace's theorem. Bernstein mentions Lexis, Charlier, and Bowley but relies heavily on the authority of Andrei Markov, whose own probability course was published in Russian in 1900, with further Russian editions in 1908, 1913, and 1924 and a German edition in 1912.

Bernstein's terminology for practical certainty influenced later developments. The Russian word уверенный can be translated into English as *confident* or *sure*, and Bernstein frequently uses практически уверенный for *practically certain.* On the first page of his book, he contrasts уверенность (confidence) based on probabilities with абсолютная достоверность (absolute certainty). On p. 233, he gives a prediction interval for a binomial outcome as follows:

> ...уже при $t = 4$, $2\Phi(t) = 0.999936$, и соответствующее неравенство
>
> $$\left|\frac{m-np}{\sqrt{npq}}\right| < 4$$
>
> обычно считают практически достоверным. Отсюда следует, например, что, если мы произведем 2 500 бросаний монеты, то можно быть практически уверенным, что число т появлений „орла" будет удовлетворять неравенству
>
> $$|m - 1250| < 4 \times 25,$$

[28] In Russian, С. Н. Бернштейн. Modern transliterations render Бернштейн as *Bernshtein.* But he is usually *Bernstein* in the mathematical literature.

> т.-е.
>
> $$1150 < m < 1350;$$

In English:

> ... already for $t = 4$, $2\Phi(t) = 0.999936$, and the corresponding inequality
>
> $$\left|\frac{m - np}{\sqrt{npq}}\right| < 4$$
>
> is usually considered practically certain. It follows, for example, that if we make 2,500 coin tosses, then we can be almost sure that the number m of heads will satisfy the inequality
>
> $$|m - 1250| < 4 \times 25,$$
>
> or
>
> $$1150 < m < 1350;$$

Here I have translated практически уверенным as *almost sure*.

On p. 274, we find similar language in the description of what we now call a confidence interval:

> ... благодаря тому, что $p - p'$ следует закону Гаусса, после определения $p' = \frac{m}{n}$ можно быть практически уверенным, что
>
> $$p' - 4\sigma < p < p' + 4\sigma;$$

In English

> ... because $p - p'$ follows Gauss's law, after observing $p' = \frac{m}{n}$ you can be practically certain that
>
> $$p' - 4\sigma < p < p' + 4\sigma;$$

Here, for the sake of variety, I have translated практически уверенным as *practically certain*.

Bernstein was Bayesian, but he does not draw the Bayesian/Bernoullian contrast in this elementary book, and he sounds Bernollian when using Laplace's theorem. Jerzy Neyman attended Bernstein's probability course at Kharkov in 1915–1916, and as this was his main training in probability as a student, it may be unsurprising that he was somewhat uncertain about the role of prior probabilities when he began working in theoretical statistics and that he used the Polish *ufnośń* and the English *confidence* when he began calculating intervals in applied statistics; see [78, p. 43] and §A.23.

A.20 Truman L. Kelley, 1884–1961

The following passage is drawn from pp. 102–103 of Kelley's 1924 book *Statistical Method* [68].

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.

A.21 David Brunt, 1886–1965

In 1917, the Welsh meteorologist David Brunt published a book on the theory of errors, *The Combination of Observations* [21], which included a chapter on the periodogram. The book was squarely in the Gaussian tradition, the book did not mention Laplace and 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 §3.7, Brunt set 19 to 1 as the odds for practical certainty in his periodogram analysis of Greenwich temperature records in 1919 [22, p. 328].

On pp. 131–132 of his book, Brunt summarized the controversy concerning the rejection of observations as follows:

> The whole question of the possibility of rejecting observations on the ground of theoretical discussion based on residuals only, has given rise to a considerable amount of controversy. Bessel opposed the rejection of any observation unless the observer was satisfied that the external conditions produced some unusual source of error not present in the other observations of the series. Peirce's criterion was at an early date subjected to very severe criticism. Airy claimed that it was defective in foundation, and illusive in its results. He maintained that, so long as the observer was satisfied that the same sources of error were at work, though in varying degrees, throughout a series of observations, the computer should have no right to reject any observation by a discussion based solely on antecedent probability. An observation should be rejected only when a thorough examination showed that the causes of error normally at work were not sufficient to produce the error in the doubtful observation. Airy also cited a case where the rejection of the observations having large residuals led to poor results. . . .
>
> Though many of the arguments of Airy and others against the use of mathematical criteria such as Peirce's have been shown to be based on faulty premises, the fact remains that none of these criteria have ever come into general use.

Brunt concludes, however, by subscribing to a simple rule for "a moderate number of observations": reject observations for which the residual exceeds five probable errors and take a close look at any others whose residual exceeds 3.5

probable errors. This may have been influential; in 1931, Harold Jeffreys reported that rejecting observations with residuals greater five probable errors was “a common astronomical practice” [65, p. 83].

A.22 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* [42], published in 1925 and in many subsequent editions.

Shelving the probable error

As we saw in §3.6, 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).

Preface to the 6th edition (1936). As John Aldrich has pointed out to me, Fisher discusses the nature of his book and touches on the issue of p-hacking in the preface to the 6th edition of *Statistical Methods for Research Workers*, published in 1936. We find this argument on pp. ix–x:

> Those critics who would like to have seen the inclusion of mathematical proofs of the more important propositions of the underlying theory, must still be referred to the technical publications given in the list of sources. There they will encounter exactly those difficulties

> which it would be undesirable to import into the present work; and will perceive that modern statistics could not have been developed without the elaboration of a system of ideas, logical and mathematical, which, however fascinating in themselves, cannot be regarded as a necessary part of the equipment of every research worker.
>
> To present "elementary proofs," of the kind which do not involve these ideas, would be really to justify the censure of a second school of critics, who, rightly feeling that a fallacious proof is worse than none, are eager to decry any attempt to "teach people to run before they can walk." The actual scope of the present volume really exempts it from this criticism, which, besides, in an age of technical co-operation, has seldom much force. The practical application of general theorems is a different art from their establishment by mathematical proof. It requires fully as deep an understanding of their meaning, and is, moreover, useful to many to whom the other is unnecessary.

And then we find this paragraph on p. xii:

> 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 fuller Table in the *Indian Journal of Statistics* (1936).

A.23 Jerzy Neyman, 1894–1981

As mentioned in §4.1, Cournot defined what we now call a confidence interval in his 1843 *Exposition*, and such intervals were widely used in the 19th century. Yet the notion of a confidence interval is now widely attributed to Jerzy Neyman. How did this come about?

Neyman introduced the terms *confidence coefficient* and *confidence interval* in English, in 1934 [87]. Already, beginning in 1930, Fisher had published examples of probability intervals for parameters obtained without the use of prior probabilities. Fisher called his intervals *fiducial*, and few of his readers, then or now, have been able to find a discern a consistent theory behind them. Confronted with this reality and Fisher's stature, Neyman tried to credit Fisher with priority while presenting his own definition as a clarification; he explained that Fisher's papers…

> …have been misunderstood and the validity of statements they contain formally questioned. This I think is due largely to the very condensed form of explaining ideas used by R. A. Fisher, and perhaps also to a somewhat difficult method of attacking the problem.

He also credited Markov with having considered, in the context of least squares, the problem of finding the narrowest confidence intervals for a given confidence coefficient. This is a evidently a reference to Gauss's theorem, according to which the arithmetic mean has the smallest variance among linear unbiased estimators. We may conjecture that Neyman learned the erroneous attribution to Markov from Bernstein; as a result, the theorem is now called the *Gauss-Markov theorem.*

Beyond granting credit to Fisher and Markov, Neyman also pointed out that confidence intervals were already widely used in practice. He wrote (p. 562),

> ...the methods of estimating, particularly in the case of large samples, resulting from the work of Fisher, are often precisely the same as those which are already in common use. Thus the new solution of the problems of estimation consists mainly in a rigorous justification of what has been generally considered correct more or less on intuitive grounds.

Neyman was evidently unaware that Laplace, Cournot, and Bienaymé had explained the large-sample agreement between the inverse and direct methods for obtaining probability intervals. Perhaps he was still unaware of this 19th century work in the 1950s, but by that time he had discovered the theorem in an article by von Mises in 1919 [122] and had also discerned it, perhaps between the lines, in Bernstein's 1927 book. He and his student Lucien Le Cam then dubbed the theorem the *Bernstein-von Mises theorem* [77].

Fisher was not happy with Neyman's appreciation of his work, and Neyman was forced to present his theory as different from Fisher's; see [78, Chapter 6]. In 1941 [89], Neyman explained that he had in fact been using confidence intervals in Poland beginning in 1930, before he had known about Fisher's work. His Polish name for them had been *przedział ufności.* Both the English *confidence* and the Polish *ufność* refer to the attitude of a person who is sure of something. Neyman's teacher Bernstein may also have had some influence here; *ufność* is a reasonable translation into Polish of Bernstein's Russian уверенность.

In 1937, Neyman encountered difficulties when he tried to publish a more extensive theoretical study of confidence intervals in the *Journal of the Royal Statistical Society.* As he explained many years later to his biographer Constance Reid [101, p. 139], one referee (later known to be Yule, who could scarcely have seen anything novel in the notion of a confidence interval) was unfavorable, the other (later known to be A. C. Aitken) was favorable. Aitken advised the editor that the paper might benefit from using the foundation for mathematical probability recently published by Kolmogorov [73]. Neyman had not previously known about Kolmogorov's measure-theoretic axiomatization for probability, but he made it central to a successful revision [88]. Now there was prestigious mathematical backing for Neyman's contention that he was providing a previously absent theoretical basis for an established practice. He made the point this way, on pp. 346–347 of the 1937 article:

> If we look through a number of recent statistical publications, we shall find that it is exceedingly rare that the values of unique estimates are given without the $\pm \mathrm{S_T}$. We shall find also that the

> comments on the values of T are largely dependent on those of $\mathrm{S_T}$. This shows that what the statisticians have really in mind in problems of estimation is not the idea of a unique estimate but that of two estimates having the form, say
>
> $$\underline{\theta} - k_1\mathrm{S_T} \qquad \text{and} \qquad \overline{\theta} + k_2\mathrm{S_T},$$
>
> where k_1 and k_2 are certain constants, indicating the limits between which the true value of θ presumably falls.
>
> In this way the practical work, which is frequently in advance of the theory, brings us to consider the theoretical problem of estimating the parameter θ by means of the interval $(\underline{\theta}, \overline{\theta})$, extending from $\underline{\theta}$ to $\overline{\theta}$. These limits will be called the lower and upper estimates of θ respectively. . . .

When Russian mathematicians later discussed Fisher's work, they translated *fiducial limits* as доверительные границы; like *fiducial*, the adjective доверительный evokes the notion of trust. Bernstein rejected Fisher's reasoning in favor of the classical (Bayesian) argument [6], whereas Kolmogorov favored Neyman's theory [72]. Kolmogorov's view carried the day, but the Russians retained the Fisherian adjective; a confidence interval is now a доверительный интервал in Russian. Perhaps *confidence* could not be translated back into уверенность, because that word was already taken by the classical theory; see §A.19.

A.24 Egon S. Pearson, 1895–1980

When Pearson, Karl Pearson's son, joined his father's department at the University of London in 1921, he had already encountered the theory of errors and statistical methods at Cambridge, where he had interacted with Eddington and Yule. Passing comments in his early articles suggest that by the early 1920s he and his fellow British statisticians were all too aware that p-values are invalidated when tests are selected on the basis of data.

In 1925, for example, we find the following comments in an article in which Pearson tried to evaluate Bayes's theorem empirically [93, p. 435]:

> . . . If, to take a different example, in our statistical experience, we only apply the χ^2 Test for Goodness of Fit to cases where from inspection the fit looks unsatisfactory, then clearly we must be very careful what deductions we draw. For we might obtain in nearly 100% of tests, fits with $P(\chi^2)$ less than $0 \cdot 1$, where yet in every case the sample had been drawn from the supposed population.
>
> It is of course the old difficulty; once or twice in a thousand times an event will occur against which the odds are 999 to 1, but when it does occur we are inclined to think that something is wrong, focusing our attention on the single event and forgetting the 999 times when we have observed that it did not happen.

A decade later, in 1936, we find this passage in Pearson's article with Chandra Sekar on the rejection of outliers [92, p. 317]:

> ... In conclusion, since it is sometimes held that the appropriate test can be chosen after examining the data in the sample, a final word of caution is necessary. To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the hypothesis is true.

A.25 Morris Viteles, 1898–1996

In a brief article on intelligence testing published by Viteles in 1922 [121], 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 ± .091, much below the level of significance.

and

> ... reduces the co-efficient of correlation of these two tests to plus 0.37 ± .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.

Appendix B Bernoullian and Bayesian

On p. 4, I cited some authors who have used the adjective *Bernoullian* rather than *frequentist* to designate statistical methods that follow Jacob Bernoulli's example rather than that of Thomas Bayes. Here are some quotations from those authors.

- Francis Edgeworth used *Bernoullian* in this way in 1918, contrasting "the *direct* problem associated with the name of Bernoulli" with "the *inverse* problem associated with the name of Bayes" [36].
- Richard von Mises made a similar remark in German in 1919 ([123], page 5): "Man kann die beiden großen Problemgruppen ... als den Bernoullischen und den Bayesschen Ideenkreis charakterisieren." In English: "We can call the two large groups of problems the Bernoullian and Bayesian circles of ideas."

- A. P. Dempster advocated the usage in 1966 [30]. In 1968 [31], in a review of three volumes of collected papers by Jerzy Neyman and E. S. Pearson, Dempster wrote

  > Neyman and Pearson rode roughshod over the elaborate but shaky logical structure of Fisher, and started a movement which pushed the Bernoullian approach to a high-water mark from which, I believe, it is now returning to a more normal equilibrium with the Bayesian view.

- Ian Hacking used *Bernoullian* repeatedly in his 1990 book, *The Taming of Chance* [56]. Writing about Poisson's interest in changes in the chance of conviction by a jury, he wrote (page 97):

  > Laplace had two ways in which to address such questions. One is Bernoullian, and attends to relative frequencies; the other is Bayesian, and is usually now interpreted in terms of degrees of belief. Laplace almost invited his readers not to notice the difference.

This usage would recognize Bernoulli as the first to state a theory of direct statistical estimation, just as Bayes was the first to state Bayes's formula. It would also allow us to contrast Bernoullian and Bayesian methods without asserting anything about how probabilities are to be interpreted. Using *frequentist* for both an interpretation of probability and a method of inference is a source of conceptual and historical confusion. It obscures, for example, the fact that von Mises, long recognized as the leading proponent of "the frequency theory of probability", always contended that Bayes's formula provides the correct method of statistical inference [124].

References

[1] George Biddell Airy. *On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations.* Macmillan, London, 1861. 2nd edition 1875, 3rd edition 1879. 9, 10

[2] Valentin Amrhein, Sander Greenland, and Blake McShane et al. Retire statistical significance. *Nature*, 567:305–307, 2019. 18

[3] Michel Armatte. Discussion de l'article de D. Denis [32]. *Journal de la Société Française de Statistique*, 145(4):27–36, 2004. 30

[4] Michel Armatte. Contribution à l'histoire des tests laplaciens. *Mathematics and social sciences*, 44(176):117–133, 2006. 30

[5] С. Н. Бернштейн (Sergei Bernstein). Теория вероятностей *(Theory of Probability).* Государственное Издательство (State Publishing House), Москва (Moscow), 1927. 46

[6] С. Н. Бернштейн (Sergei Bernstein). О «доверительных» вероятностях Фишера. Известия АН СССР, Серия математическая *(Izv. Akad. Nauk SSSR Ser. Mat.)*, 5(2):85–94, 1941. Translated as "On the Fisherian "confidence probabilities"" on pp. 112–121 of *Probability and Statistics: Russian Papers of the Soviet Period, selected and translated by Oscar Sheynin*, NG Verlag, Berlin, 2005. 54

[7] David R. Bellhouse. Karl Pearson's influence in the United States. *International Statistical Review*, 77(1):51–63, 2009. 13

[8] 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. 39

[9] Yoav Benjamini and Henry Braun. John W. Tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6):1576–1594, 2002. 19

[10] Joseph Bertrand. *Calcul des Probabilités*. Gauthier-Villars, Paris, 1889. Second edition 1907. 8

[11] 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. 25

[12] William Beveridge. Wheat prices and rainfall in western Europe (with discussion). *Journal of the Royal Statistical Society*, 85(3):412–478, May 1922. 15

[13] Jules Bienaymé. Mémoire sur la probabilité des résultats moyens des observations; démonstration directe de la règle de Laplace [présenté le 12 mai 1834 à l' Académie des sciences]. *Mémoires présentés par divers savants à l'A cadémie royale des sciences de l'Institut de France*, 5:513–558, 1838. 8

[14] Edwin G. Boring. The number of observations on which a limen may be based. *The American Journal of Psychology*, 27(3):315–319, 1916. 13

[15] Edwin G. Boring. Mathematical vs. scientific significance. *Psychological Bulletin*, 16(10):335–338, 1919. 13

[16] Arthur Lyon Bowley. *Elements of Statistics*. King, Westminster, 1901. Later editions appeared in 1902, 1907, 1920, 1925, and 1937. 41

[17] Joan Fisher Box. *R. A. Fisher: the life of a scientist*. Wiley, New York, 1978. 2

[18] Benard Bru. Remarques sur l'article de D. Denis [32]. *Journal de la Société Française de Statistique*, 145(4):37–38, 2004. 30

[19] 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. 8, 21

[20] Marie-France Bru and Bernard Bru. *Les jeux de l'infini et du hasard*. Presses universitaires de Franche-Comté, Besançon, France, 2018. Two volumes. 1, 3, 4, 8

[21] David Brunt. *The The Combination of Observations*. Cambridge University Press, 1917. 2nd edition in 1931. 48

[22] David Brunt. A periodogram analysis of the Greenwich temperature records. *Quarterly Journal of the Meteorological Society*, 45(192):323–338, 1919. 15, 49

[23] Beatrice M. Cave and Karl Pearson. Numerical illustrations of the variate difference correlation method. *Biometrika*, 10:340–355, 1914/1915. 39

[24] Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (Bernard Bru, editor) of [25]. 7, 8, 10, 11, 30

[25] Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–2010. The volumes are numbered I through XI, but VI and XI are double volumes. 29, 30, 58

[26] Andrew I. Dale. *A History of Inverse Probability From Thomas Bayes to Karl Pearson*. Springer, New York, second edition, 1999. 4

[27] 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. 10

[28] Augustus De Morgan. *An Essay on Probabilities, and on their application to Life Contingencies and Insurance Offices*. Longman, Orme, Brown, Green & Longmans, London, 1838. 4, 11

[29] W. Edwards Deming. *Statistical Adjustment of Data*. Wiley, New York, 1943. 17

[30] Arthur P. Dempster. Further examples of inconsistencies in the fiducial argument. *Annals of Mathematical Statistics*, 34(3):884–891, 1966. 4, 56

[31] Arthur P. Dempster. Crosscurrents in statistics; Review of *The Selected Papers*, by E. S. Pearson, *Joint Statistical Papers*, by Jerzy Neyman and E. S. Pearson, and *A Selection of Early Statistical Papers*, by J. Neyman. *Science*, 160:661–663, 1968. 56

[32] 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. 19, 56, 57

[33] Francis Edgeworth. Methods of statistics. *Journal of the Statistical Society of London*, Jubilee Volume:181–217, 1885. 11, 36, 41

[34] Francis Edgeworth. On discordant observations. *Philosophical Magazine Series 5*, 23(143):364–375, 1887. 14, 37

[35] Francis Edgeworth. Probability. In *Encyclopædia Britannica*, volume 22. Cooper, 11th edition, 1911. 36

[36] Francis Edgeworth. Mathematical representation of statistics: A reply. *Journal of the Royal Statistical Society*, 81(2):322–333, 1918. 4, 55

[37] Johann Franz Encke. Über die Methode der kleinsten Quadrate. *Astronomisches Jahrbuch für 1834. Der Sammlung Berliner astronomischer Jahrbücher*, 59:249–312, 1832. 11

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

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

[40] Ronald A. Fisher. Applications of "Student's" distribution. *Metron*, 5(3):90–104, 1925. 17

[41] 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. 16

[42] Ronald A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 1925. The thirteenth edition appeared in 1958. 13, 50

[43] 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. 16

[44] Ronald A. Fisher. Combining independent tests of significance. *The American Statistician*, 2(5):30, 1948. 17

[45] 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. 10, 22

[46] 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. 7, 24

[47] Walter A. Friedman. *Fortune Tellers: The Story of America's First Economic Forecasters.* Princeton, 2014. 15

[48] Michael Friendly. A.-M. Guerry's *Moral Statistics of France*: Challenges for multivariable spatial analysis. *Statistical Science*, 22(3):368–399, 2007. 7

[49] Thomas Galloway. *Treatise on Probability.* Black, Edinburgh, 1839. 10, 28

[50] Jules Gavarret. *Principes généraux de statistique médicale, ou développement des règles qui doivent présider à son emploi.* Bechet, Paris, 1840. 33

[51] 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, 19

[52] Prakash Gorroochurn. *Classic Topics on the History of Modern Mathematical Statistics from Laplace to More Recent Times.* Wiley, New York, 2016. 3, 5

[53] Barry Gower. Astronomy and probability: Forbes versus Michell on the distribution of stars. *Annals of Science*, 39:142–160, 1982. 7

[54] Barry Gower. Planets and probability: Daniel Bernoulli on the inclinations of the planetary orbits. *Studies in the History and Philosophy of Science*, 18(4):441–454, 1987. 7

[55] André-Michel Guerry. *Essai sur la statistique morale de la France.* Crochard, Paris, 1833. 7

[56] Ian Hacking. *The Taming of Chance.* Cambridge University Press, New York, 1990. 4, 8, 56

[57] Roger Hahn. *Correspondance de Pierre Simon Laplace (1749–1827).* Brepols, Turnhout, Belgium, 2013. Two volumes. 4

[58] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York, 1998. 3, 4, 7, 24

[59] David J. Hand. From evidence to understanding: A commentary on Fisher (1922) 'On the mathematical foundations of theoretical statistics'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2039), 2015. 2

[60] J. Arthur Harris. On the selective elimination occurring during the development of the fruits of staphylea. *Biometrika*, 7:452–504, 1909/1910. 39

[61] Christopher Charles Heyde and Eugene Seneta. *I. J. Bienaymé: Statistical Theory Anticipated*. Springer, New York, 1977. 4, 8

[62] 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. 17

[63] 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. 19

[64] Edward Huth. Jules Gavarret's *Principes Généraux de Statistique Médicale*. *Journal of the Royal Society of Medicine*, 101:205–212, 2008. 33

[65] Harold Jeffreys. *Scientific Inference*. Cambridge, London, 1931. 50

[66] 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. 5, 8

[67] Andreas Kamlah. The decline of the Laplacian theory of probability: A study of Stumpf, von Kries, and Meinong. In Krüger et al. [75], pages 91–116. 8

[68] Truman L. Kelley. *Statistical Method*. Macmillan, New York, 1923. 13, 47

[69] Maurice G. Kendall. On the analysis of oscillatory time-series. *Journal of the Royal Statistical Society*, 108(1/2):93–141, 1945. 45

[70] Judy L. Klein. *Statistical Visions in Time: A History of Time Series Analysis, 1662–1938*. Cambridge, 1997. 15

[71] Sven K. Knebel. *Wille, Würfel und Wahrscheinlichkeit: Das System der moralischen Notwendigkeit in der Jesuitenscholastik 1550–1700*. Meiner, Berlin, 2000. 9

[72] Andrei Kolmogorov. Определение центра рассеивания и меры точности по ограниченному числу наблюдений. Известия АН СССР, Серия математическая *(Izv. Akad. Nauk SSSR Ser. Mat.)*, 6(1-2):3–32, 1942. Translated as "Determining the center of scattering and the measure of precision given a restricted number of observations" on pp. 200–231 of *Probability and Statistics: Russian Papers of the Soviet Period, selected and translated by Oscar Sheynin*, NG Verlag, Berlin, 2005. 54

[73] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933. An English translation by Nathan Morrison appeared under the title *Foundations of the Theory of Probability* (Chelsea, New York) in 1950, with a second edition in 1956. 53

[74] Christian Kramp. *Analyse des Réfractions Astronomiques et Terrestres.* Schwikkert, Leipsic, 1799. 5

[75] Lorenz Krüger, Lorraine J. Daston, and Michael Heidelberger, editors. *The Probabilistic Revolution. Volume 1: Ideas in History.* MIT, Cambridge, Massachusetts, 1987. 61, 64

[76] 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. 38

[77] Lucien Le Cam. On the asymptotic theory of estimation and testing hypotheses. In Jerzy Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 129–156. University of California Press, Berkeley, 1956. 53

[78] Erich L. Lehmann. *Fisher, Neyman, and the Creation of Classical Statistics.* Springer, New York, 2011. 1, 47, 53

[79] Wilhelm Lexis. *Einleitung in die Theorie der Bevölkerungsstatistik.* Trübner, Strassburg, 1875. 10, 35

[80] John W. MacArthur. Linkage studies with the tomato. *Genetics*, 11(4):387–405, 1926. 17

[81] Thierry Martin. Les premiers écrits probabilistes de Cournot (1825-1828). In Jose Maria Arribas et al., editor, *Historia de la probabilidad y la estadística VI*, pages 173–186. AHEPE-UNED, Madrid, 2012. 29

[82] Deborah Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars.* Cambridge, 2018. 1, 19

[83] Mansfield Merriman. *Elements of the Method of Least Squares.* Macmillan, London, 1874. 37

[84] 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. 11

[85] Mary Morgan. *The History of Econometric Ideas.* Cambridge University Press, Cambridge, 1990. 15

[86] Denton E. Morrison and Ramon E. Henkel. *The Significance Test Controversy — A Reader.* Aldine, Chicago, 1970. 17

[87] Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934. 52

[88] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937. 53

[89] Jerzy Neyman. Fiducial argument and the theory of confidence intervals. *Biometrika*, 32(2):128–150, 1941. 53

[90] 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. 21

[91] Raymond Pearl. *Introduction to Medical Biometry Statistics.* Saunders, Philadelphia and London, 1923. 45

[92] E. S. Pearson and C. Chandra Sekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28(3/4):308–320, 1936. 55

[93] Egon S. Pearson. Bayes' theorem, examined in the light of experimental sampling. *Biometrika*, 17(3/4):388–442, 1925. 54

[94] Karl Pearson. *The Grammar of Science.* Scott, London, 1892. 28

[95] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894. 8

[96] 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. 17

[97] Siméon-Denis Poisson. Observations relatives au nombre de naissances des deux sexes. *Annuaire le bureau des longitudes pour 1825*, pages 98–99, 1824. 7

[98] 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. 7, 24

[99] 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. 24

[100] Persis Putnam. Sex differences in pulmonary tuberculosis deaths. *The American Journal of Hygiene*, 7(6):663–705, 1927. 17

[101] Constance Reid. *Neyman*. Springer, 1998. 53

[102] Ivo Schneider. Laplace and thereafter: The status of probability calculus in the nineteenth century. In Krüger et al. [75], pages 191–214. 8

[103] 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. 15, 37

[104] 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. 15, 38

[105] Glenn Shafer. Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19:309–370, 1978. 19

[106] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistics Society, Series A*, To appear. For a pre-publication version, see Working Paper 54, `www.probabilityandfinance.com`. 20

[107] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995. 19

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

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

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

[111] Stephen M. Stigler. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, 68(344):872–879, 1973. 11

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

[113] Stephen M. Stigler. A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, 101(1):60–70, 1992. 12

[114] Stephen M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999. 12

[115] Stephen M. Stigler. Fisher and the 5% level. *Chance*, 21:12–21, 2008. 14

[116] 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. 13

[117] C. P. Sun. On the examination of final digits by experiments in artificial sampling. *Biometrika*, 20A(1/2):64–68, 1928. 17

[118] Godfrey H. Thomson. The criterion of goodness of fit of psychophysical curves. *Biometrika*, 12:216–230, 1918/1919. 39

[119] J. F. Tocher. Pigmentation survey of school children in Scotland. *Biometrika*, 6:130–235, 1908/1909. 39

[120] 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. 14

[121] Morris S. Viteles. A comparison of three tests of "general intelligence". *Journal of Applied Psychology*, 6(4):391–402, 1922. 55

[122] Richard von Mises. Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 4:1–97, 1919. 53

[123] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919. 4, 55

[124] Richard von Mises. On the correct use of Bayes' formula. *Annals of Mathematical Statistics*, 13(2):156–165, 1942. 56

[125] Gilbert T. Walker. On the criterion for the reality of relationships or periodicities. *Indian Meteorological Memoirs*, 21(9), 1914. 15, 39

[126] Gilbert T. Walker. Periodicities. (Letter to the Editor. *Nature*, 110(2763):511, 1922. 16, 39

[127] Gilbert T. Walker. On periodicity. *Quarterly Journal of the Royal Meteorological Society*, 51(216):337–346, 1925. 39

[128] Helen M. Walker. *Studies in the History of Statistical Method.* Williams & Wilkins, Baltimore, 1929. 4, 9

[129] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, 2019. 20

[130] W. F. R. Weldon. A first study of natural selection in clausilia laminata (montagu). *Biometrika*, 1(1):109–124, 1901. 38

[131] John Wishart. Field experiments of factorial design. *The Journal of Agricultural Science*, 28(2):299–306, 1938. 17

[132] 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. 17

[133] George Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60:812–854, 1897. 9, 42

[134] George Udny Yule. *An Introduction to the Theory of Statistics*. Griffin, London, first edition, 1911. 17, 42