J. B. S. Haldane's contribution to the bayes factor hypothesis test

Etz, A.; Wagenmakers, E-.J.

DOI
10.1214/16-STS599

Publication date
2017

Document Version
Final published version

Published in
Statistical Science

Statistical Science
2017, Vol. 32, No. 2, 313–329
DOI: 10.1214/16-STS599

J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test

Alexander Etz and Eric-Jan Wagenmakers

Abstract. This article brings attention to some historical developments that gave rise to the Bayes factor for testing a point null hypothesis against a composite alternative. In line with current thinking, we find that the conceptual innovation—to assign prior mass to a general law—is due to a series of three articles by Dorothy Wrinch and Sir Harold Jeffreys (1919, 1921, 1923a). However, our historical investigation also suggests that in 1932, J. B. S. Haldane made an important contribution to the development of the Bayes factor by proposing the use of a mixture prior comprising a point mass and a continuous probability density. Jeffreys was aware of Haldane's work and it may have inspired him to pursue a more concrete statistical implementation for his conceptual ideas. It thus appears that Haldane may have played a much bigger role in the statistical development of the Bayes factor than has hitherto been assumed.

Key words and phrases: History of statistics, induction, evidence, Sir Harold Jeffreys.

1. INTRODUCTION

Bayes factors grade the evidence that the data provide for one statistical model over another. As such, they represent "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378). In addition, Bayes factors can be used for model-averaging (Hoeting et al., 1999) and variable selection (Bayarri et al., 2012). Bayes factors are employed across widely different disciplines such as astrophysics (Lattimer and Steiner, 2014), forensics (Taroni et al., 2014), psychology (Dienes, 2014), economics (Malikov, Kumbhakar and Tsionas, 2015) and ecology (Cuthill and Charleston, in press). Moreover, Bayes factors are a topic of active statistical interest (e.g., Fouskakis, Ntzoufras and Draper, 2015, Holmes et al., 2015, Sparks, Khare and Ghosh, 2015; for a more pessimistic view, see Robert, 2016). These modern applications and developments arguably find their roots in the work of one man: Sir Harold Jeffreys.

Jeffreys is a towering figure in the history of Bayesian statistics. His early writings, together with his coauthor Dorothy Wrinch (Wrinch and Jeffreys, 1919, 1921, 1923a), championed the use of probability theory as a means of induction and laid the conceptual groundwork for the development of the Bayes factor. The insights from this work culminated in the monograph *Scientific Inference* (Jeffreys, 1931), in which Jeffreys gives thorough treatment to how a scientist can use the laws of inverse probability (now known as Bayesian inference) to "learn from experience" (for a review, see Howie, 2002 and an earlier version of this paper available at http://arxiv.org/abs/1511.08180v2).

Among many other notable accomplishments, such as the development of prior distributions that are invariant under transformation and his work in geophysics and astronomy, where he discovered that the Earth's core is liquid, Jeffreys is widely recognized as the inventor of the Bayesian significance test, with seminal papers in 1935 and 1936 (Jeffreys, 1935, 1936a). The centerpiece of these papers is a number, which Jeffreys denotes K, that indicates the ratio of posterior to

Alexander Etz is a graduate research fellow in the Department of Cognitive Sciences, University of California, Irvine, Irvine, California 92617, USA (e-mail: Etz.Alexander@gmail.com). Eric-Jan Wagenmakers is Professor at the Methodology Unit of the Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 VZ Amsterdam, The Netherlands (e-mail: EJ.Wagenmakers@gmail.com).

prior odds; much later, Jeffreys's statistical tests would come to be known as *Bayes factors* (Good, 1958).[1] Once again these works culminated in a comprehensive book, *Theory of Probability* (Jeffreys, 1939).

When the hypotheses in question are simple point hypotheses, the Bayes factor reduces to a likelihood ratio, a method of measuring evidential strength which dates back as far as Johann Lambert in 1760 (Lambert and DiLaura, 2001) and Daniel Bernoulli in 1777 (Kendall et al., 1961; see Edwards, 1974 for a historical review); C. S. Peirce had specifically called it a measure of 'weight of evidence' as far back as 1878 (Peirce, 1878; see Good, 1979). Alan Turing also independently developed likelihood ratio tests using Bayes' theorem, deriving *decibans* to describe the intensity of the evidence, but this approach was again based on the comparison of simple versus simple hypotheses. For example, Turing used decibans when decrypting the Enigma codes to infer the identity of a given letter in German military communications during World War II (Turing, 1941/2012).[2] As Good (1979) notes, Jeffreys's Bayes factor approach to testing hypotheses "is especially 'Bayesian' [because] either [hypothesis] is composite" (p. 393).

Jeffreys states that across his career his "chief interest is in significance tests" (Jeffreys, 1980, p. 452). Moreover, in an (unpublished) interview with Dennis Lindley (DVL) for the Royal Statistical Society on August 25, 1983, when asked "What do you see as your major contribution to probability and statistics?" Jeffreys (HJ) replies,

> HJ: The idea of a significance test, I suppose, putting half the probability into a constant being 0, and distributing the other half over a range of possible values.
> DVL: And that's a very early idea in your work.
> HJ: Well, I came on to it gradually. It was certainly before the first edition of 'Theory of Probability'.
> DVL: Well, of course, it is related to those ideas you were talking about to us a few minutes ago, with Dorothy Wrinch, where you were putting a probability on...
> HJ: Yes, it was there, of course, when the data are counts. It went right back to the beginning.
> ("Transcription of a Conversation between Sir Harold Jeffreys and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys)

That Jeffreys considers his greatest contribution to statistics to be the development of Bayesian significance tests, tests that compare a point null against a distributed (i.e., composite) alternative, is remarkable considering the range of his accomplishments.

In their influential introductory paper on Bayes factors, Kass and Raftery (1995) state,

> In a 1935 paper and in his book *Theory of Probability*, Jeffreys developed a methodology for quantifying the evidence in favor of a scientific theory. The centerpiece was a number, now called the *Bayes factor*, which is the posterior odds of the null hypothesis when the prior probability on the null is one-half. (p. 773, abstract)

Many distinguished statisticians and historians of statistics consider the development of the Bayes factor to be one of Jeffreys's greatest achievements and a landmark contribution to the foundation of Bayesian statistics. In a recent discussion of Jeffreys's contribution to Bayesian inference, Robert, Chopin and Rousseau (2009) recall the importance and novelty of Jeffreys's significance tests,

[1] See Good (1988) and Fienberg (2006) for a historical review. The term 'Bayes factor' comes from Good, who attributes the introduction of the term to Turing, who simply called it the 'factor'.

[2] Turing started his Maths Tripos at King's College in 1931, graduated BA in 1934, and was a Fellow of King's College from 1935–1936. Anthony (A. W. F.) Edwards speculates that Turing might have attended some of Jeffreys's lectures while at Cambridge, where he would have learned about details of Bayes' theorem (Edwards, 2015, personal communication). According to the college's official record of lecture lists, Jeffreys's lectures 'Probability' started in 1935 (or possibly Easter Term 1936), and in the year 1936 they were in the Michaelmas (i.e., Fall) Term. Turing would have had the opportunity of attending them in the Easter Term or the Michaelmas Term in 1936 (Edwards, 2015, personal communication). Jack (I. J.) Good has also provided speculation about their potential connection, "Turing and Jeffreys were both Cambridge dons, so perhaps Turing had seen Jeffreys's use of Bayes factors; but, if he had, the effect on him must have been unconscious for he never mentioned this influence and he was an honest man. He had plenty of time to tell me the influence if he was aware of it, for I was his main statistical assistant for a year" (Good, 1980, p. 26). Later, in an interview with David Banks, Good remarks that "Turing might have seen [Wrinch and Jeffreys's] work, but probably he thought of [his likelihood ratio tests] independently" (Banks, 1996, p. 11). Of course, Turing could have learned about Bayes' theorem from any of the standard probability books at the time, such as Todhunter (1858), but the potential connection is of interest. For more detail on Turing's work on cryptanalysis, see Zabell (2012).

> If the hypothesis to be tested is $H_0 : \theta = 0$, against the alternative H_1 that is *the aggregate of other possible values [of* θ*]*, Jeffreys initiates one of the major advances of *Theory of Probability* by rewriting the prior distribution as a mixture of a point mass in $\theta = 0$ and of a generic density π on the range of θ... This is indeed a stepping stone for Bayesian Statistics in that it explicitly recognizes the need to separate the null hypothesis from the alternative hypothesis within the prior, lest the null hypothesis is not properly weighted once it is accepted (p. 157). [emphasis original]

Some commentators on Robert, Chopin and Rousseau (2009) shared their sentiment. Lindley (2009) remarked that Jeffreys's "triumph was a general method for the construction of significance tests, putting a concentration of prior probability on the null value... and evaluating the posterior probability using what we now call Bayes factors" (p. 184). Kass (2009) noted that a "striking high-level feature of Theory of Probability is its championing of posterior probabilities of hypotheses (Bayes factors), which made a huge contribution to epistemology" (p. 180). Moreover, Senn (2009) is similarly impressed with Jeffreys's innovation in assigning a concentration of probability to the null hypothesis, calling it "a touch of genius, necessary to rescue the Laplacian formulation [of induction]" (p. 186).[3]

In discussions of the development of the Bayes factor, as above, most authors focus on the work of Jeffreys, with some mentioning the early related work by Turing and Good. A piece of history that is missing from these discussions and commentaries is the contribution of John Burdon Sanderson (J. B. S.) Haldane (pictured in Fig. 1), whose application of these ideas potentially spurred Jeffreys into making his conceptual ideas about scientific learning more concrete—in the form of the Bayes factor.[4] In a paper entitled "The Bayesian Controversy in Statistical Inference," after discussing some of Jeffreys's early Bayesian developments, Barnard (1967) briefly remarks,

> Another man whose views were closely related to Jeffreys was Haldane, who... proposed a prior having a 'lump' of probability at the null hypothesis with the rest spread out, in connexion [sic] with tests of significance. (p. 238)

Similarly, the *Biographical Memoirs of the Fellows of the Royal Society* includes an entry for Haldane (Pirie, 1966), in which M. S. Bartlett recalls,

> In statistics, [Haldane] combined an objective approach to populations with an occasional cautious use of inverse probability methods, the latter being apparently envisaged in frequency terms... [Haldane's] idea of a concentration of a prior distribution at a particular value was later adopted by Harold Jeffreys, F.R.S. as a basis for a theory of significance tests. (p. 233)

However, we have not seen Haldane's connection to the Bayes factor hypothesis test mentioned in the modern statistics literature, and we are not aware of any in-depth accounts of this particular innovation to date.

Haldane is perhaps best known in the statistics literature by his proposal of a prior distribution suited for estimation of the rate of rare events, which has become known as *Haldane's prior* (Haldane, 1932, 1948).[5] References to Haldane's 1932 paper focus mainly on its proposal of the Haldane prior, and they largely miss his formulation of a mixture prior comprising a point mass and a smoothly distributed alternative—a crucial component in the Bayes factor hypothesis tests that Jeffreys would later develop. Among Haldane's various biographies (e.g., Clark, 1968, Crow, 1992, Lai, 1998, Sarkar, 1992) there is no mention of this development; while they sometimes mention statistics and mathematics among his broad list of interests, they understandably tend to focus on his major advances made in biology and genetics. In fact, this result is not mentioned even in Haldane's own autobiographical account of his career accomplishments (Haldane, 1966).

The primary purpose of this paper is to review the work of Haldane and discuss how it may have spurred Jeffreys into developing his highly influential Bayesian significance tests. We begin by reviewing the developments made by Haldane in his 1932 paper, followed by a review of Jeffreys's earliest work on the topic. We go on to draw parallels between their respective works and speculate on the nature of the connection between the two men.

[3] However, see Senn's recent in-depth discussion at http://tinyurl.com/ow4lahd for a less enthusiastic perspective.

[4] Howie (2002), p. 125, gives a brief account of some ways Haldane might have influenced Jeffreys's thoughts, but does not draw this connection.

[5] Interestingly, Haldane's prior appears to be an instance of Stigler's law of eponymy, since Jeffreys derived it in his book *Scientific Inference* (Jeffreys, 1931, p.194) eight months before Haldane's publication.

2. HALDANE'S CONTRIBUTION: A MIXTURE PRIOR

J. B. S. Haldane was a true polymath; White (1965) called him "probably the most erudite biologist of his generation, and perhaps of the [twentieth] century" (as cited in Crow, 1992, p. 1). Perhaps best known for his pioneering work on mathematical population genetics (alongside Ronald Fisher and Sewall Wright, see Smith, 1992), Haldane is also recognized for making substantial contributions to the fields of physiology, biochemistry, mathematics, cosmology, ethics, religion and (Marxist) philosophy.[6] In addition to this already impressive list of topics, in 1932 Haldane published a paper in which he presents his views regarding the foundations of statistical inference (Haldane, 1932). At the time, this was unusual for Haldane, as most of his academic work before 1932 was primarily concerned with physical sciences.

By 1931, Haldane had been working for twenty years developing a mathematically rigorous account of the chromosomal theory of genetic inheritance; this work began in 1911 (at age 19) when, during a Zoology seminar at New College, Oxford, he announced his discovery of genetic linkage in vertebrates (Haldane, 1966). This would become the basis of one of his most influential research lines, to which he intermittently applied Bayesian analyses.[7] Throughout his career, Haldane would also go on to publish many papers pertaining to classical (non-Bayesian) mathematical statistics, including a presentation of the exact moments of the χ^2 distribution (Haldane, 1937), a discussion of how to transform various statistics so that they are approximately normally distributed (Haldane, 1938), an exploration of the properties of inverse (i.e., negative) binomial sampling (Haldane, 1945), a proposal for a two-sample rank test (Haldane and Smith, 1947) and an investigation of the bias of maximum likelihood estimates (Haldane, 1951). Despite his keen interest in mathematical statistics, Haldane's work pertaining to its foundations are confined to a single paper published in the early 1930s. So, while unfortunate, it is perhaps understandable that the advances he made in 1932 are not widely known: This work is something of an anomaly, buried and forgotten in his sizable corpus.

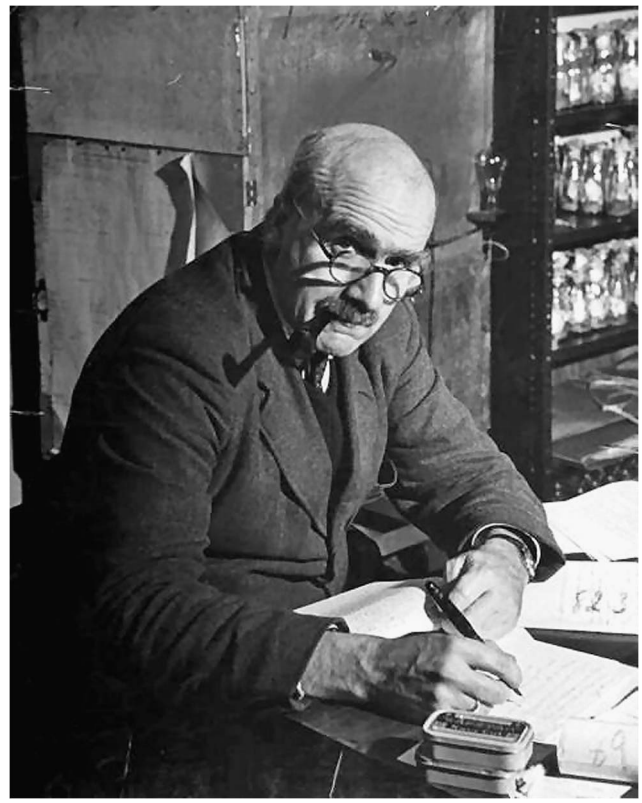

FIG. 1. *John Burdon Sanderson* (*J. B. S.*) *Haldane* (1892–1964) *in* 1941. (*Photograph by Hans Wild*, *LIFE Magazine*.)

Haldane's paper, "A note on inverse probability," was received by the *Mathematical Proceedings of the Cambridge Philosophical Society* on November 19, 1931, and read at the Society meeting on December 7, 1931 (Haldane, 1932). He begins by stating his goal, "Bayes' theorem is based on the assumption that all values of the parameter in the neighbourhood of that observed are equally probable *a priori*. It is the purpose of this paper to examine what more reasonable assumption may be made, and how it will affect the estimate based on the observed sample" (p. 55). Haldane frames his paper as giving Fisher's method of maximum likelihood a foundation through inverse probability (much to Fisher's chagrin, see Fisher, 1932). Haldane gives a short summary of his problem of interest (we will repeatedly quote at length to provide complete context):

> Let us first consider a population of which a proportion x possess the character X, and of

[6] Haldane (1966) provides an abridged autobiography, and Clark (1968) is a more thorough biographical reference. For more details on the wide-reaching legacy of Haldane, see Crow (1992), Lai (1998) and Pirie (1966). Haldane was also a prominent public figure during his time and wrote many essays for the popular press (e.g., Haldane, 1927).

[7] In one such analysis, Haldane (1919) used Bayesian updating of a uniform prior (as was customary in that time) to find the probable error of calculated linkage estimates (proportions). It is unclear from where exactly Haldane learned inverse probability, but over the years he occasionally made references to probabilistic concepts put forth by von Mises (e.g., the idea of a "kollective" from Von Mises, 1931) or to proofs for his formulas given by Todhunter (1865).

> which a sample of n members is observed, n being so large that one of the exponential approximations to the binomial probability distribution is valid. Let a be the number of individuals in the sample which possess the character X. (p. 55)

The problem, in slightly archaic notation, is the typical binomial sampling setup. Haldane goes on to say,

> It is an important fact that in almost all scientific problems we have a rough idea of the nature of $f(x)$ [the prior distribution] derived from the past study of similar populations. Thus, if we are considering the proportion of females in the human population of any large area, $f(x)$ is quite small unless x lies between .4 and .6. (p. 55)

Haldane appeals to availability of real scientific information to justify using non-uniform prior distributions. It is stated as a fact that we have information indicating that various regions for x are more probable than others. Haldane goes on to derive familiar posteriors for x starting from a uniform prior (pp. 55–56) before discussing possible implications that deviations from uniform priors have on the posteriors. His description of one such interesting deviation follows:

> If [the slope of the prior probability density], though small compared with $n^{1/2}$ in the neighborhood of $x = a/n$, has an infinity or becomes very large for some other value of x (other than 0 or 1), and if a/n is finite [i.e., $a \neq 0$ and $n \neq \infty$], then the [posterior] probability distribution is approximately Gaussian in the neighborhood of $x = a/n$, but has an infinity or secondary maximum at the other point or points. (pp. 56–57)

In other words, when a prior distribution has a distinct mass of probability at some point between 0 and 1, and a large binomial sample is obtained that contains some a's and some $not\text{-}a$'s, the posterior distribution can be approximated by a Gaussian except for a separate infinity.

Haldane avoids the trouble of an infinite density in the prior[8] by marginalizing across two orthogonal models to obtain a single mixture prior distribution that consists of a point hypothesis and a smoothly distributed alternative. Haldane first clarifies with an example and then solves the problem. We quote at length:

> An illustration from genetics will make the point clear. The plant *Primula sinensis* possesses twelve pairs of chromosomes of approximately equal size. A pair of genes selected at random will lie on different chromosomes in $11/12$ of all cases, giving a proportion $x = .5$ of "cross-overs." In $1/12$ of all cases, they lie on the same chromosome, the values of the cross-over ratio x ranging from 0 to .5 without any very marked preference for any part of this range, except perhaps for a tendency to avoid values very close to .5. $f(x)$ is thus approximately $1/6$ for $0 \leq x < .5$; it has a discontinuity at $x = .5$, such that the probability is $11/12$; while, for $.5 < x \leq 1$, $f(x) = 0$.
>
> Now if a family of 400 seedlings from the cross between [two plants] contains 160 "cross-overs" we have two alternatives. The probability of getting such a family from a plant in which the genes lie in different chromosomes is $11/12^{400}C_{160}2^{-400}$, or 1.185×10^{-5}. The probability of getting it from a plant in which they lie in the same chromosome is
>
> $$\frac{1}{6}^{400}C_{160}\int_0^{.5} x^{160}(1-x)^{240}\,dx.$$
>
> Since this integral is very nearly equal to
>
> $$\int_0^1 x^{160}(1-x)^{240}\,dx, \quad \text{or} \quad \frac{160!\,240!}{401!},$$
>
> this probability is approximately $\frac{1}{6\times 401}$, or 4.56×10^{-4}. Thus, the probability that the family is derived from a plant where the genes lie in different chromosomes and $x = .5$ is .028. Otherwise, the mean value of x is .4, with standard error .0245. The overall mean value, or mathematical expectation, of x is .4028, and the graph of the [posterior probability density] is an approximately normal error curve centred at $x = .4$ with standard deviation .0245, together with an infinity at $x = .5$. (p. 57)

Haldane's passage may be hard to parse since the example is somewhat opaque and the notation is dated.

[8] Incidentally, Jeffreys is struggling to solve a similar problem of infinite densities in a prior distribution around this time (see Jeffreys, 1931, pp. 194–195).

However, the passage is crucial and, therefore, we unpack Haldane's problem as follows. When Haldane speaks of "pairs of genes," he means that there are two different genes that are responsible for different traits, such as genes for stem length and leaf color in *Primula sinensis*. Since the DNA code for particular genes are located in specific parts of specific chromosomes, during reproduction they can randomly have their alleles switch from mother chromosome to father chromosome, which is called "cross-over." We are cross-breeding plants and we want to know the cross-over rate for these genes, which depends on whether the pair of genes are on the same of different chromosomes. For example, if the gene for petal color and the gene for stem length are in different chromosomes, then they would cross-over independently in their respective chromosomes during cell division, and the children should show new mixtures of color and length at a certain rate (50% in Haldane's example). If they are in the same chromosome, it is possible for the two genes to be located in the same segment of the chromosome that crosses over, and because their expression varies together they will show less variety in trait combination on average (i.e., $< 50\%$).

Haldane's example uses this fact to go backwards, from the number of "cross-overs" present in the child plants to infer the chromosomal distance between the genes. We will use θ, rather than Haldane's x, to denote the cross-over rate parameter. If the different genes lie on different chromosomes, they cross-over independently during meiosis, and so there is a 50% probability to see new combinations of these traits for any given offspring. Hence, if traits are on different chromosomes then $\theta = .5$. If they lie on the same chromosome they have a cross-over rate of $\theta < .5$, where the percentage varies based on their relative location on the chromosome. If they are relatively close together on the chromosome, they are likely to cross-over together and we will not see many offspring with new combinations of traits, so the cross-over rate will be closer to $\theta = 0$. If they are relatively far apart on the chromosome, they are less likely to cross-over together, so they will have a cross-over rate closer to $\theta = .5$.

Since there are 12 pairs of chromosomes, there is a natural prior probability assignment for the two competing models: 11/12 pairs of genes selected at random will lie on different chromosomes ($\mathcal{M}_0$) and 1/12 will lie on the same chromosome ($\mathcal{M}_1$); when they are on the same chromosome they could be anywhere on the chromosome, so the distance between them can range from nearly nil to nearly an entire chromosome. To capture this information, Haldane uses a uniform prior from 0 to .5 for θ. When they are on different chromosomes, $\theta = .5$ precisely. Hence, Haldane's mixture prior comprises the prior distributions for θ from the two models

$$\pi_0(\theta) = \delta(.5)$$
$$\pi_1(\theta) = \mathcal{U}(0, .5),$$

where $\delta(\cdot)$ denotes the Dirac delta function, and with prior probabilities (i.e., mixing weights) for the two models of $P(\mathcal{M}_0) = 11/12$ and $P(\mathcal{M}_1) = 1/12$. The marginal prior density for θ can be written as

$$\begin{aligned}\pi(\theta) &= P(\mathcal{M}_0)\pi_0(\theta) + P(\mathcal{M}_1)\pi_1(\theta) \\ &= \frac{11}{12} \times \delta(.5) + \frac{1}{12} \times \mathcal{U}(0, .5),\end{aligned}$$

using the law of total probability. Haldane is left with a mixture of a point mass and a smooth probability density.

Haldane breeds his plants and obtains $n = 400$ offspring, $a = 160$ of which are cross-overs (an event we denote D). The probability of the data, 160 cross-overs out of 400, under $\mathcal{M}_0$ (now known as the marginal likelihood) is

$$P(D \mid \mathcal{M}_0) = \binom{400}{160} (.5)^{400}.$$

The probability of the data under $\mathcal{M}_1$ is

$$P(D \mid \mathcal{M}_1) = 2 \binom{400}{160} \int_0^{.5} \theta^{160}(1-\theta)^{240}\, \mathrm{d}\theta.$$

The probabilities of the data can be used in conjunction with the prior model probabilities to update the prior mixing weights to posterior mixing weights (i.e., posterior model probabilities) by applying Bayes' theorem as follows (for $i = 0, 1$):

$$P(\mathcal{M}_i \mid D) = \frac{P(\mathcal{M}_i)P(D \mid \mathcal{M}_i)}{P(\mathcal{M}_1)P(D \mid \mathcal{M}_1) + P(\mathcal{M}_0)P(D \mid \mathcal{M}_0)}.$$

Using the information found thus far, the posterior model probabilities are $P(\mathcal{M}_1 \mid D) = .972$ and $P(\mathcal{M}_0 \mid D) = .028$. The two conditional prior distributions for θ are also updated to conditional posterior distributions using Bayes' theorem. Under $\mathcal{M}_0$, the prior distribution is a Dirac delta function at $\theta = .5$, which is unchanged by the data. Under $\mathcal{M}_1$, the prior distribution $\mathcal{U}(0, .5)$ is updated to a posterior distribution that is approximately $\mathcal{N}(.4, .0245^2)$. The marginal

posterior density for θ can thus be written as a mixture,

$$\begin{aligned}&\pi(\theta \mid D)\\&\quad = P(\mathcal{M}_0 \mid D)\pi_0(\theta \mid D) + P(\mathcal{M}_1 \mid D)\pi_1(\theta \mid D)\\&\quad = .028 \times \delta(.5) + .972 \times \mathcal{N}\big(.4, .0245^2\big).\end{aligned}$$

Moreover, Haldane then uses the law of total probability to arrive at a model-averaged prediction for θ, as follows: $E(\theta) = .5 \times .028 + \frac{160}{400} \times .972 = .4028$. This appears to be the first concrete application of *Bayesian model averaging* (Hoeting et al., 1999).[9]

Haldane uses a mixture prior distribution to solve another challenging problem.[10] We quote Haldane:

> We now come to the case where $a = 0\ldots$ Unless we have reasons to the contrary, we can no longer assume that $x = 0$ is an infinitely improbable solution... In the case of many classes both of logical and physical objects, we can point to a finite, though undeterminable, probability that $x = 0$. Thus, there are good, but inconclusive, reasons for the beliefs that there is no even number greater than 2 that cannot be expressed as the sum of 2 primes, and that no hydrogen atoms have an atomic weight between 1.9 and 2.1. In addition, we know of very large samples of each containing no members possessing these properties. Let us suppose then, that k is the *a priori* probability that $x = 0$, and that the *a priori* probability that it has a positive [nonzero] value is expressed by $f(x)$, where $\underset{\epsilon \to 0}{Lt} \int_\epsilon^1 f(x)\,dx = 1 - k$. (pp. 58–59)

There is an explicit (albeit data-driven, by the sound of it) interest in a special value of the parameter, but under a continuous prior distribution the probability of any point is zero. To solve this problem, Haldane again uses a mixture prior; in modern notation, k denotes the prior probability $P(\mathcal{M}_0)$ of the point-mass component of the mixture, a Dirac delta function at $\theta = 0$, with a second component that is a continuous function of θ with prior probability $P(\mathcal{M}_1) = 1 - k$.

In a straightforward application of Bayes' theorem, Haldane finds that "the probability, after observing the sample, that $x = 0$ is

$$\frac{k}{k + \int_0^1 (1-x)^n f(x)\,dx}.$$

"If $f(x)$ is constant this is $(n+1)k/(nk+1)\ldots$ This is so even if $f(x)$ has a logarithmic infinity at $x = 0\ldots$ Hence, as n tends to infinity the probability that $x = 0$ tends to unity, however small be the value of k" (Haldane, 1932), p. 59. Haldane again goes on to perform Bayesian model-averaging to find "the probability that the next individual observed will have the character X" (p. 59).

In sum, in 1932 Haldane publishes his views on the foundations of statistical inference, and proposed to use a two-component mixture prior comprising a point mass and smoothly distributed alternative. He went on to apply this mixture prior to concrete problems involving genetic linkage, and in doing so also performed the first instances of Bayesian model-averaging. To assess the importance and originality of Haldane's contribution, it is essential to discuss the related earlier work of Dorothy Wrinch and Harold Jeffreys (pictured in Fig. 2 and Fig. 3, respectively), and the later work by Jeffreys alone.

3. WRINCH AND JEFFREYS'S DEVELOPMENT OF THE BAYES FACTOR BEFORE HALDANE (1932)

Jeffreys was interested in the philosophy of science and induction from the beginning of his career. He learned of inverse probability at a young age (circa 1905, when he would have been 14 or 15 years old) from reading his father's copy of Todhunter (1858) (Exhibit H204, St John's College Library, Papers of Sir Harold Jeffreys). To Jeffreys, Toddhunter explained "inverse probability absolutely clearly and I [Jeffreys] never saw there was any difficulty about it" (Transcribed by AE from the audio-cassette in Exhibit H204, St John's College Library, Papers of Sir Harold Jeffreys). His views were refined around the year 1915 while studying Karl Pearson's influential *Grammar of Science* (Pearson, 1892), which led Jeffreys to see inverse probability as the method that "seemed... the sensible way of expressing common sense" ("Transcription of a Conversation between Sir Harold Jeffreys

[9] Robert, Chopin and Rousseau (2009), p. 166, point out that Jeffreys's *Theory of Probability* (Jeffreys, 1939) "includes the seeds" of model averaging. In fact, the seeds appear to go back to Wrinch and Jeffreys (1921), p. 387, where they briefly note that if future observation q_2 is implied by law p (i.e., $P(q_2 \mid q_1, p) = 1$), "the probability of a further inference from the law is not appreciably higher than that of the law itself" since the second term in the sum $P(q_2 \mid q_1) = P(p \mid q_1)P(q_2 \mid q_1, p) + P(\sim p \mid q_1)P(q_2 \mid q_1, \sim p)$ is usually "the product of two small factors" (p. 387).

[10] The way Haldane (1932) sets up the problem shows a great concurrence of thought with Jeffreys, who was tackling a similar problem at the time in his book *Scientific Inference* (Jeffreys, 1931, pp. 194–195).

FIG. 2. *Dorothy Wrinch* (1894–1976) *circa* 1920–1923. (*Photograph by Sir Harold Jeffreys, included by permission of the Master and Fellows of St John's College, Cambridge.*)

and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys). This became a theme that permeated his early work, done in conjunction with Dorothy Wrinch, which sought to put probability theory on a firm footing for use in scientific induction.[11]

Their work was motivated by that of Broad (1918), who showed[12] that when one applies Laplace's principle of insufficient reason—assigning equal probability to all possible states of nature—to finite populations, one is lead to an inductive pathology: A general law could virtually never achieve a high probability. This was a result that flew in the face of the common scientific view of the time that "an inference drawn from a simple scientific law may have a very high probability, not far from unity" (Wrinch and Jeffreys, 1921, p. 380). Jeffreys (1980) later recalled Broad's result,

> Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property φ, and we do not know r, the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find all of them with φ, the posterior probability that all n are φ's is $(m+1)/(n+1)$.

[11] More background for the material in this section can be found in Aldrich (2005) and Howie (2002). See Howie (2002), Chapter 4, and Senechal (2012), Chapter 8, for details on the personal relationship between Wrinch and Jeffreys. Many of the ideas presented by Wrinch and Jeffreys are similar in spirit to those of W. E. Johnson; as a student, Wrinch attended Johnson's early lectures on advanced logic at Cambridge (see Howie, 2002, p. 86). Jeffreys would later emphasize Wrinch's important contributions their work on scientific induction, "I should like to put on record my appreciation of the substantial contribution she [Wrinch] made to this work [(Wrinch and Jeffreys, 1919, 1921, 1923a, 1923b)], which is the basis of all my later work on scientific inference" (Hodgkin and Jeffreys, 1976, p. 564). See Senechal (2012) for more on Wrinch's far-reaching scientific influence.

[12] For an in-depth discussion about the law of succession, including Broad's insights, see Zabell (1989). Zabell points out that Broad's result had been derived earlier by others, including Prevost and LHuilier (1799), Ostrogradski (1848) and Terrot (1853). Interested readers should see Zabell (1989), p. 286, as well as the mathematical appendix beginning on p. 309.

> A general rule would never acquire a high probability until nearly the whole of the class had been sampled. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919–1923. (p. 452)

The consequence of Wrinch and Jeffreys's series of papers was the derivation of two key results. First, they found a solution to Broad's quandary in the assignment of a finite initial probability, independent of the population's size, to the general law itself, which allows a general law to achieve a high probability without needing to go so far as to sample nearly the entire population. Second, they derived the odds form of Bayes' theorem (Wrinch and Jeffreys, 1921, p. 387): "If p denote the most probable law at any stage, and q an additional experimental fact, [and h the knowledge we had before the experiment,]" (p. 386) a new form of Bayes' theorem can be written as

$$\frac{P(p \mid q.h)}{P(\sim p \mid q.h)} = \frac{P(q \mid p.h)}{P(q \mid\sim p.h)} \cdot \frac{P(p \mid h)}{P(\sim p \mid h)}.$$

At the time, the conception of Bayes' rule in terms of odds was novel. Good (1988) remarked, "[the above equation] has been mentioned several times in the literature without citing Wrinch and Jeffreys. Because it is so important I think proper credit should be given" (p. 390). We agree with Good; the importance of this innovation should not be understated, because it lays the foundation for the future of Bayesian hypothesis testing. A simple rearrangement of terms highlights that the Bayes factor is the ratio of posterior odds to prior odds,

$$\underbrace{\frac{P(q \mid p,h)}{P(q \mid\sim p,h)}}_{\text{Bayes factor}} = \underbrace{\frac{P(p \mid q,h)}{P(\sim p \mid q,h)}}_{\text{Posterior odds}} \Big/ \underbrace{\frac{P(p \mid h)}{P(\sim p \mid h)}}_{\text{Prior odds}}.$$

Wrinch and Jeffreys went on to show how the Bayes factor—which is the amount by which the data, q, shift the balance of probabilities for p versus $\sim p$—can form the basis of a philosophy of scientific learning (Ly, Verhagen and Wagenmakers, 2016).

4. JEFFREYS'S DEVELOPMENT OF THE BAYES FACTOR AFTER HALDANE (1932)

Remarkably, the Bayes factor remained a conceptual development until Jeffreys published two seminal papers in 1935 and 1936. In 1935, Jeffreys published his first significance tests in the *Mathematical Proceedings of the Cambridge Philosophical Society*, with the title, "Some Tests of Significance, Treated by the Theory of Probability" (Jeffreys, 1935). This was shortly after his published dispute with Fisher (for an account of the controversy, see Aldrich, 2005, Howie, 2002, Lane, 1980).

In his opening statement to the 1935 article, Jeffreys states his main goal:

> It often happens that when two sets of data obtained by observation give slightly different estimates of the true value we wish to know whether the difference is significant. The usual procedure is to say that it is significant if it exceeds a certain rather arbitrary multiple of the standard error; but this is not very satisfactory, and it seems worth while to see whether any precise criterion can be obtained by a thorough application of the theory of probability. (p. 203)

Jeffreys calls his procedures "significance tests" and surely means for this work to contrast with that of Fisher's. Even though Jeffreys does not mention Fisher directly, Jeffreys does allude to their earlier dispute by reaffirming that his probabilities express "no opinion about the frequency of the truth of [the hypothesis] among any real or imaginary populations" (presumably addressing one of Fisher's 1934 objections to Jeffreys's solution to the theory of errors) and that assigning equal probabilities to propositions "is simply the formal way of saying that we do not know whether it is true or not of the actual populations under consideration at the moment" (Jeffreys, 1935, p. 203).

Jeffreys strives to use the theory of probability to find a satisfactory rule to determine whether differences between observations should be considered "significant," and he begins with a novel approach to test a difference of proportions as present in a contingency table. Jeffreys first posits the existence of two "large, but not infinite, populations" and that these "have been sampled in respect to a certain property" (Jeffreys, 1935, p. 203). He continues,[13]

> One gives $[x_0]$ specimens with the property, $[y_0]$ without; the other gives $[x_1]$ and $[y_1]$ respectively. The question is, whether the difference between $[x_0/y_0]$ and $[x_1/y_1]$

[13] To facilitate comparisons to Haldane's section we have translated Jeffreys's unconventional notation to a more modern form. See Ly, Verhagen and Wagenmakers (2016), Appendix D, for details about translating Jeffreys's notation.

FIG. 3. *Sir Harold Jeffreys* (1891–1989) *in* 1928. (*Photographer unknown, included by permission of the Master and Fellows of St John's College, Cambridge.*)

> gives any ground for inferring a difference between the corresponding ratios in the complete populations. Let us suppose that in the first population the fraction of the whole possessing the property is $[\theta_0]$, in the second $[\theta_1]$. Then we are really being asked whether $[\theta_0 = \theta_1]$; and further, if $[\theta_0 = \theta_1]$, what is the posterior probability distribution among values of $[\theta_0]$; but, if $[\theta_0 \neq \theta_1]$, what is the distribution among values of $[\theta_0]$ and $[\theta_1]$. (p. 203)

Take $\mathcal{M}_0$ to represent the proposition that $\theta_0 = \theta_1$, so that $\mathcal{M}_1$ represents $\theta_0 \neq \theta_1$. Jeffreys takes the prior probability of both $\mathcal{M}_0$ and $\mathcal{M}_1$ as .5. This problem can be restated so that $\mathcal{M}_0$ represents the difference between θ_0 and θ_1, namely, $\theta_0 - \theta_1 = 0$. This is Jeffreys's point null hypothesis which is assigned a finite prior probability of .5. If the null is true, and $\theta_0 = \theta_1$, then Jeffreys assigns θ_0 a uniform prior distribution in the range $0 - 1$. The remaining half of the prior probability is assigned to $\mathcal{M}_1$, which specifies θ_0 and θ_1 have their probabilities "uniformly and independently distributed" in the range $0 - 1$ (p. 204). Thus,

$$\pi_0(\theta_0) = \mathcal{U}(0, 1)$$
$$\pi_1(\theta_0) = \mathcal{U}(0, 1)$$

$$\pi_1(\theta_1) = \mathcal{U}(0, 1)$$

and $\pi_1(\theta_0, \theta_1) = \pi_1(\theta_0)\pi_1(\theta_1)$. Subsequently,

$$P(\mathcal{M}_0, \theta_0) = P(\mathcal{M}_0)\pi_0(\theta_0)$$
$$P(\mathcal{M}_1, \theta_0, \theta_1) = P(\mathcal{M}_1)\pi_1(\theta_0, \theta_1),$$

which follows from the product rule of probability, namely,

$$P(p, q) = P(p)P(q \mid p).$$

Jeffreys then derives the likelihood functions for D (i.e., the data, or compositions of the two samples) on the null and the alternative. Under the null hypothesis, $\mathcal{M}_0$, the probability of D is

$$f(D \mid \theta_0, \mathcal{M}_0) = \frac{(x_0 + y_0)!}{x_0!y_0!}\frac{(x_1 + y_1)!}{x_1!y_1!} \cdot \theta_0^{x_0}(1 - \theta_0)^{y_0}\theta_0^{x_1}(1 - \theta_0)^{y_1}.$$

Under the alternative hypothesis, $\mathcal{M}_1$, the probability of D is

$$f(D \mid \theta_0, \theta_1, \mathcal{M}_1) = \frac{(x_0 + y_0)!}{x_0!y_0!}\frac{(x_1 + y_1)!}{x_1!y_1!} \cdot \theta_0^{x_0}(1 - \theta_0)^{y_0}\theta_1^{x_1}(1 - \theta_1)^{y_1}.$$

In other words, the probability of both sets of data ($x_0 + y_0$ and $x_1 + y_1$), on $\mathcal{M}_0$ or $\mathcal{M}_1$, is equal to the product of their respective binomial likelihood functions and constants of proportionality. Since the prior distributions for θ_0 and θ_1 are uniform and the prior probabilities of $\mathcal{M}_0$ and $\mathcal{M}_1$ are equal, the posterior distributions of θ_0 and (θ_0, θ_1) are proportional to their respective likelihood functions above. Thus, the posterior distributions for θ_0 and (θ_0, θ_1) under the null and alternative are, respectively,

$$\pi_0(\theta_0 \mid D) \propto \theta_0^{x_0+x_1}(1 - \theta_0)^{y_0+y_1},$$

and

$$\pi_1(\theta_0, \theta_1 \mid D) \propto \theta_0^{x_0}(1 - \theta_0)^{y_0}\theta_1^{x_1}(1 - \theta_1)^{y_1}.$$

Now Jeffreys has two contending posterior distributions, and to find the posterior probability of $\mathcal{M}_1$ he integrates $P(\mathcal{M}_0, \theta_0 \mid D)$ with respect to θ_0 and integrates $P(\mathcal{M}_1, \theta_0, \theta_1 \mid D)$ with respect to both θ_0 and θ_1. Using the identity $\int_0^1 \theta_0^{x_0}(1 - \theta_0)^{y_0}\,\mathrm{d}\theta_0 = (x_0!y_0!)/(x_0 + y_0 + 1)!$, he derives the relative posterior probabilities of $\mathcal{M}_0$ and $\mathcal{M}_1$. For $\mathcal{M}_0$, it is simply the previous identity with the additional terms of x_1 and y_1 added to the respective factorial terms,

$$P(\mathcal{M}_0 \mid D) \propto \frac{(x_0 + x_1)!(y_0 + y_1)!}{(x_0 + x_1 + y_0 + y_1 + 1)!}.$$

To obtain the posterior probability of $\mathcal{M}_1$, the product of the integrals with respect to each of θ_0 and θ_1 is needed,

$$P(\mathcal{M}_1 \mid D) \propto \frac{x_0!y_0!}{(x_0 + y_0 + 1)!}\frac{x_1!y_1!}{(x_1 + y_1 + 1)!}.$$

Their ratio gives the posterior odds, $P(\mathcal{M}_0 \mid D)/P(\mathcal{M}_1 \mid D)$, and is the solution to Jeffreys's significance test. When this ratio is greater than 1, the data support $\mathcal{M}_0$ over $\mathcal{M}_1$ and vice versa.

Jeffreys ends his paper with a discussion of the implications of his approach.[14] One crucial point is that if the prior odds are different from unity then the final calculations from Jeffreys's tests yield the marginal likelihoods of the two hypotheses, $P(D \mid \mathcal{M}_0)$ and $P(D \mid \mathcal{M}_1)$, whose ratio gives the Bayes factor. We quote at length,

> We have in each case considered the existence and the nonexistence of a real difference between the two quantities estimated as two equivalent alternatives, each with prior probability 1/2. This is a common case, but not general. If, however, the prior probabilities are unequal the only difference is that the expression obtained for $[P(\mathcal{M}_0 \mid D)/P(\mathcal{M}_1 \mid D)]$ now represents $[\frac{P(\mathcal{M}_0|D)}{P(\mathcal{M}_1|D)}/\frac{P(\mathcal{M}_0)}{P(\mathcal{M}_1)}]$ [NB: the Bayes factor, which is the posterior odds divided by prior odds]. Thus, if the estimated ratio exceeds 1, the proposition $[\mathcal{M}_0]$ is rendered more probable by the observations, and if it is less than 1, $[\mathcal{M}_0]$ is less probable than before. It still remains true that there is a critical value of the observed difference, such that smaller values reduce the probability of a real difference. The usual practice [NB: alluding to Fisher] is to say that a difference becomes significant at some rather arbitrary multiple of the standard error; the present method enables us to say what that value should be. If, however, the difference examined is one that previous considerations make unlikely to exist, then we are entitled to ask for a greater increase of the probability before we accept it and, therefore, for a larger ratio of the difference to its standard error. (p. 221)

[14] In this paper, Jeffreys gives many more derivations of new types of significance tests, but one example suffices to convey the general principle.

Here, Jeffreys has also laid out the benefits his tests possess over Fisher's tests: The critical standard error, where the ratio supports $\mathcal{M}_0$, is not fixed at an arbitrary value but is determined by the amount of data (Jeffreys, 1935, p. 205); in Jeffreys's test, larger sample sizes increase the critical standard error, and thus increase the barrier for 'rejection' at a given threshold (note that this anticipates Lindley's paradox, Lindley, 1957). Furthermore, if the phenomenon has unfavorable prior odds against its existence we may reasonably require more evidence (i.e., a larger Bayes factor) before we are reasonably confident in its existence. That is to say, extraordinary claims require extraordinary evidence.

In a complementary paper, Jeffreys (1936a) expanded on the implications of this method:

> To put the matter in other words, if an observed difference is found to be on order [of one standard error from the null], then on the hypothesis that there is no real difference this is what would be expected; but if there was a real difference that might have been anywhere within a range m it is a remarkable coincidence that it should have happened to be in just this particular stretch near zero. On the other hand if the observed difference is several times its standard error it is very unlikely to have occurred if there was no real difference, but it is as likely as ever to have occurred if there was a real difference. In this case beyond a certain value of x [a distance from the null] the more remarkable coincidence is for the hypothesis of no real difference... The theory merely develops these elementary considerations quantitatively... (p. 417)

In sum, the key developments in Jeffreys's significance tests are that a point null is assigned finite prior probability and that this point null is tested against a distributed (composite) alternative hypothesis. The posterior odds of the models are computed from the data, with the updating factor now known as the Bayes factor.

5. THE CONNECTION BETWEEN JEFFREYS AND HALDANE

We are now in a position to draw comparisons between the developments of Haldane and Jeffreys. The methods show a striking similarity in that they both set up competing models with orthogonal parameters, each with a finite prior model probability, and through the calculation of each model's marginal likelihood find posterior model probabilities and posterior distributions for the parameters. Where the methods differ is primarily in the focus of their inference. Haldane focuses on the overall mixture posterior distribution for the parameter, $\pi(\theta \mid D)$, marginalized across the competing models. In Haldane's example, this means to focus on estimating the cross-over rate parameter, using relevant real-world knowledge of the problem to construct a mixture probability distribution. It would have been but a short step for Haldane to find the ratio of the posterior and prior model odds as Jeffreys did, since the prior and posterior model probabilities are crucial in constructing the mixture distributions, but that aspect of the problem was not Haldane's primary interest.

Jeffreys's focus is nearly the opposite of Haldane's. Jeffreys uses the models to instantiate competing scientific theories, and his focus is on making an inference about which theory is more probable. In contrast to Haldane, Jeffreys isolates the value of the Bayes factor as the basis of scientific learning and statistical inference, and he takes the posterior distributions for the parameters within the models as a mere secondary interest. If one does happen to be interested in estimating a parameter in the presence of model uncertainty, Jeffreys recognizes that one should ideally form a mixture posterior distribution for the parameter (as Haldane does), but that "nobody is likely to use it. In practice, for sheer convenience, we shall work with a single hypothesis and choose the most probable" (Jeffreys, 1935, p. 222).

Clearly, there was a great concurrence of thought between Haldane and Jeffreys during this time period. A natural question is whether the men knew of each others' work on the topic. In his paper, Haldane (1932) makes no reference to any of Jeffreys's previous works (recall, Haldane's stated goal was to extend Fisher's method of maximum likelihood and discuss the use of nonuniform priors). Jeffreys's first book, *Scientific Inference* (Jeffreys, 1931), came out a mere eight months before Haldane submitted his paper, and it was not widely read by Jeffreys's contemporaries; around the time the revised first edition was released in 1937, Jeffreys remarked to Fisher in a letter (on June 5, 1937) that he "should really have liked to scrap the whole lot and do it again, but at the present rate it looked as if the thing would take about 60 years to sell out" (Bennett, 1990, p. 164).[15] So it would be no surprise if

[15]Indeed, whereas 224 copies were sold in Great Britain and Ireland in 1931, only twenty copies in total were sold in 1932, sixty-

Haldane had not come across Jeffreys's book. In fact, had Haldane known of *Scientific Inference*, he would have surely recognized that Jeffreys had given the same topics a thorough conceptual treatment. For example, Haldane might have cited Wrinch and Jeffreys (1919), who detached the theory of inverse probability from uniform distributions over a decade before; or Haldane might have recognized the complete form of the Haldane prior presented by Jeffreys (1931), p. 194. Therefore, according to the evidence in the literature, one might reasonably conclude that by 1931/1932, Haldane did not know of Jeffreys's work. What about Jeffreys, is there evidence in the literature that he knew of Haldane's work while working on his significance tests?

According to Jeffreys (1977), in the footnote in a piece looking back on his career, he did recognize the similarity of his work to Haldane's:

> The essential point [in solving Broad's quandary] is that when we consider a general law we are supposing that it may possibly be true, and we express this by concentrating a positive (nonzero) fraction of the initial probability in it. Before my work on significance tests, the point had been made by J. B. S. Haldane (1932). (p. 95)

Jeffreys also acknowledges Haldane in a footnote of his first edition of *Theory of Probability* (1939, retained in subsequent editions) when he remarks, "This paper [(Haldane, 1932)] contained the use of... the concentration of a finite fraction of the prior probability in a particular value, *which later became the basis of my significance tests*"[emphasis added] (p. 114, footnote). So it is clear that Jeffreys, at least some time later, recognized the similarity of his and Haldane's thinking.

Jeffreys's awareness of Haldane's work must go back even further. Jeffreys certainly knew of Haldane's work by 1932, because he wrote a critical commentary on Haldane's paper (which Jeffreys read at *The Cambridge Philosophical Society* on November 7, 1932), pointing out, *inter alia*, that Haldane's proposed prior distribution largely "agrees with a previous guess" of his, but only after a slight correction to its form (Jeffreys, 1933, p. 85).[16] However, when publishing his seminal significance test papers in 1935 and 1936 (Jeffreys, 1935, 1936a), Jeffreys does not mention Haldane's work. It may appear to a reader of those papers as if Jeffreys's tests are entirely his own invention; indeed, they are the natural way to apply his early work with Wrinch on the philosophy of induction to quantitative tests of general laws. Perhaps Jeffreys simply forgot, or did not realize the significance of Haldane's paper on his own thinking at the time. But in another paper published at that time, Jeffreys (1936b), in pointing out how he and Wrinch solved Broad's quandary, remarked,

> but we [Wrinch and Jeffreys] pointed out that agreement with general belief is obtained if we take the prior probability of a simple law to be finite, whereas on the natural extension of the previous theory it is infinitesimal. Similarly, for the case of sampling J. B. S. Haldane and I have pointed out that general laws can be established with reasonable probabilities if their prior probabilities are moderate and independent of the whole number of members of the class sampled. (p. 344)

So it seems that Jeffreys did recognize the importance and similarity of Haldane's work in the very same years he was publishing his own work on significance tests (1935–1936). Why Jeffreys did not cite Haldane's work directly in connection with his own development of significance tests is not clear just from looking at the literature. There is, however, some potential clarity to be gained by examining the personal relationship between the two men.

Haldane and Jeffreys were both in Cambridge from 1922–1932 while working on these problems (Haldane at Trinity College and Jeffreys at St John's), after which Haldane left for University College London. In an (unpublished) interview with George Barnard (recall we quoted Barnard above as being one of the few to recognize Haldane's innovative "lump" of prior probability), after Jeffreys (HJ) denies having known Fisher while they were both at Cambridge, Barnard (GB) asks,

> GB: But Haldane was in Cambridge, was he?
> HJ: Yes.

nine in 1933, thirty-five in 1934, fourteen in 1935 and twenty-two in 1936 (Exhibits D399-400, St John's College Library, Papers of Sir Harold Jeffreys).

[16]Notably, Jeffreys discusses Haldane's assignment of separate prior probability k to $\theta = 0$, suggesting that for completeness one should also give consideration to $\theta = 1$. See Howie (2002), pp. 121–126, for more detail on the reactions to Haldane's paper by both Jeffreys and Fisher; Fisher (1932) was particularly pointed in his commentary.

> GB: Because he joined in a bit I think in some of the... [NB: interrupted]
> HJ: Yes, well Haldane did anticipate some things of mine. I have a reference to him somewhere.
> GB: But the contact was via the papers [NB: Haldane, 1932, Jeffreys, 1933] rather than direct personal contact.
> HJ: Well I knew Haldane, of course. Very well.
> GB: Oh, ah.
> HJ: Can you imagine him being in India?
> (Transcribed by AE from the audio-cassette in Exhibit H204, St John's College Library, Papers of Sir Harold Jeffreys)

So it would seem that Jeffreys and Haldane knew each other personally very well in their time at Cambridge. We cannot be sure of the extent to which they knew of each others' professional work at that time, because we can find no surviving direct correspondence between Jeffreys and Haldane during the 1920s to early 1930s; we are unable to do more than speculate about what topics they might have discussed in their time together at Cambridge. However, there is some preserved correspondence from the 1940s where they discuss, among other topics, materialist versus realist philosophy, scientific inference, geophysics, and Marxist politics.[17] It stands to reason that Haldane and Jeffreys would have discussed similar topics in their time together at Cambridge.

This personal relationship opens up the possibility that Jeffreys and Haldane discussed their work in detail and consciously chose not to cite each other in their published works in the 1930s. Perhaps Haldane brought up the genetics problems he was working on and Jeffreys suggested the idea to Haldane to set up orthogonal models as he did, an application which Jeffreys thought of but had not yet published on. Haldane goes on to publish his paper, and both men, realizing the question of credit is somewhat murky, decide to ignore the matter. This would explain why Haldane never references these developments later in his career. Admittedly, this account leaves a few points unresolved. It does not appear that Haldane used this method in any of his later empirical investigations; if Haldane and Jeffreys together developed this method in order to apply it to one of Haldane's genetics problems, why did Haldane never go on to actually apply it? And if Jeffreys had thought of how to apply this revolutionary method of inference, why was it not included in his book? Under this account, Jeffreys would have had to think of this development in the few months between when he wrote and published *Scientific Inference* and when Haldane began to write his paper. Moreover, in his interview with Barnard and in his later published works, Jeffreys readily notes that Haldane anticipated some of his own ideas (although, it is not always entirely clear to which ideas Jeffreys refers). Did Jeffreys begin to feel guilty that he would be getting credit for ideas that Haldane helped develop?

Today one might be surprised to hear that two people could become good friends while potentially never discussing their work with each other; however, Jeffreys has a history of remaining unaware of his close probabilist contemporaries' work. It is well known that Jeffreys and fellow Cambridge probabilist Frank Ramsey were good friends while at Cambridge and they never discussed their work with each other either. Ramsey was highly regarded in Cambridge at the time, and as Jeffreys recalls in his interview with Lindley, "I knew Frank Ramsey well and visited him in his last illness but somehow or other neither of us knew that the other was working on probability theory" ("Transcription of a Conversation between Sir Harold Jeffreys and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys).[18] When one realizes that the friendship between Jeffreys and Ramsey began with their shared interest in psychoanalysis (Howie, 2002, p. 117), perhaps it makes sense that they would not get around to discussing their work on such an obscure topic as probability theory.

6. CONCLUDING COMMENTS

Around 1920, Dorothy Wrinch and Harold Jeffreys were the first to note that in order for induction to be

[17] This correspondence is available online thanks to the UCL Wellcome Library: http://goo.gl/9qHBF6. Interestingly, near the end of this correspondence (undated, but from some time between 2 October 1942 and 29 March 1943) Haldane says to Jeffreys that he felt "emboldened by your [Jeffreys's] kindness to my views on inverse probability."

[18] Jeffreys was also unaware of the work of Italian probabilist Bruno de Finetti, another of his contemporary probabilists. In their interview, Lindley asks Jeffreys if he and de Finetti ever made contact. Jeffreys replies that, not only had he and de Finetti never been in contact, "I've only just heard his name... I've never seen anything that he's done... I'm afraid I've just never come across him" ("Transcription of a Conversation between Sir Harold Jeffreys and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys).

possible it is essential that general laws be assigned finite initial probability. This argument was revolutionary in that it went against their contemporaries' blind adherence to Laplace's principle of indifference. It is this brilliant insight of Wrinch and Jeffreys that forms the conceptual basis of the Bayes factor. Later Jeffreys would develop concrete Bayes factors in order to test a point null against a smoothly distributed alternative to evaluate whether the data justify changing the form of a general law. For these reasons, we believe that Wrinch and Jeffreys should together be credited as the progenitors of the concept of the Bayes factor, with Jeffreys the first to put the Bayes factor to use in real problems of inference.

However, our historical review suggests that in 1931 J. B. S. Haldane made an important intellectual advancement in the development of the Bayes factor. We speculate that it was the specific nature of the linkage problem in genetics that caused Haldane to serendipitously adopt a mixture prior comprising a point mass and smooth distribution; it does not appear as if Haldane derived his result from a principled philosophical stance on induction, in contrast to Jeffreys, but merely through a pragmatic attempt at utilizing non-standard (i.e., nonuniform) distributions with inverse probability. And yet, Haldane never went on to employ this method to any of his future applications, so we cannot discount the possibility that he was simply momentarily inspired to address the foundations of statistical inference—as a polymath is wont to do. Then, having lost interest, he never follows up on this work, thereby dooming his important development to obscurity. We may never know his true motivations. Nevertheless, Haldane's work likely formed the impetus for Jeffreys to make his conceptual ideas concrete, leading to his thorough development of Bayes factors in the following years.

The personal relationship between Haldane and Jeffreys further complicates the story behind these developments. The two men were in close contact during the period when they developed their ideas, and the extent to which they knew of each others' work on essentially the same problem is unclear. Haldane and Jeffreys had closely converging ideas, as is seen by the similarity of their work in the 1930s, and both were statistical pioneers whose influence is still felt today. We hope this historical investigation will bring Haldane some well-deserved credit for his impact on the development of the Bayes factor.

ACKNOWLEDGMENTS

We are especially grateful to Kathryn McKee for helping us access the material in the Papers of Sir Harold Jeffreys special collection at the St John's College Library; the quoted material is included by permission of the Master and Fellows of St John's College, Cambridge. We would like to thank Anthony (A. W. F.) Edwards for keeping our Bernoullis straight, as well as looking up lecture lists from the 1930s to check on Turing's and Jeffreys's potential interactions. We would like to thank Christian Robert and Stephen Stigler for helpful comments and suggested references. We also thank Alexander Ly for critical comments on an early draft and for fruitful discussions about the philosophy of Sir Harold Jeffreys. The first author (AE) is grateful to Rogier Kievit and Anne-Laura Van Harmelen for their hospitality during his visit to Cambridge. Finally, we thank the anonymous reviewers and editor for valuable comments and criticism. This work was supported by the National Science Foundation Graduate Research Fellowship Program #DGE-1321846, NSF grant #1534472 from the Methods, Measurements, and Statistics panel, and the ERC grant "Bayes or Bust."

REFERENCES

ALDRICH, J. (2005). The statistical education of Harold Jeffreys. *Int. Stat. Rev.* **73** 289–307.

BANKS, D. L. (1996). A conversation with I. J. Good. *Statist. Sci.* **11** 1–19. MR1437124

BARNARD, G. A. (1967). The Bayesian controversy in statistical inference. *J. Inst. Actuar.* **93** 229–269.

BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035

BENNETT, J. H. (1990). *Statistical Inference and Analysis*: *Selected Correspondence of R. A. Fisher*. The Clarendon Press, Oxford Univ. Press, New York.

BERGER, J. O. (2006). Bayes factors. In *Encyclopedia of Statistical Sciences*, *Vol.* 1, 2nd ed. (S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic and N. L. Johnson, eds.) 378–386. Wiley, Hoboken, NJ.

BROAD, C. D. (1918). On the relation between induction and probability (Part I). *Mind* **27** 389–404.

CLARK, R. W. (1968). *JBS*: *The Life and Work of J. B. S. Haldane*. Coward-McCann, Inc., New York.

CROW, J. F. (1992). Centennial: JBS Haldane, 1892–1964. *Genetics* **130** 1.

CUTHILL, J. H. and CHARLESTON, M. (in press). Wing patterning genes and coevolution of Müllerian mimicry in *Heliconius* butterflies: Support from phylogeography, co-phylogeny and divergence times. *Evolution*.

DIENES, Z. (2014). Using Bayes to get the most out of non-significant results. *Front. Psychol.* **5**, 781.

EDWARDS, A. W. F. (1974). The history of likelihood. *Int. Stat. Rev.* **42** 9–15. MR0353514

FIENBERG, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Anal.* **1** 1–40. MR2227361

FISHER, R. A. (1932). Inverse probability and the use of likelihood. *Math. Proc. Cambridge Philos. Soc.* **28** 257–261.

FISHER, R. A. (1934). Probability likelihood and quantity of information in the logic of uncertain inference. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **146** 1–8.

FOUSKAKIS, D., NTZOUFRAS, I. and DRAPER, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Anal.* **10** 75–107. MR3420898

GOOD, I. J. (1958). Significance tests in parallel and in series. *J. Amer. Statist. Assoc.* **53** 799–813. MR0103560

GOOD, I. J. (1979). Studies in the history of probability and statistics. XXXVII. A. M. Turing's statistical work in World War II. *Biometrika* **66** 393–396. MR0548210

GOOD, I. J. (1980). The contributions of Jeffreys to Bayesian statistics. In *Bayesian Analysis in Econometrics and Statistics*: *Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 21–34. North-Holland, Amsterdam.

GOOD, I. J. (1988). The interface between statistics and philosophy of science. *Statist. Sci.* **3** 386–412. MR0996411

HALDANE, J. B. S. (1919). The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series. *J. Genet.* **8** 291–297.

HALDANE, J. B. S. (1927). *Possible Worlds*: *And Other Essays* **52**. Chatto & Windus.

HALDANE, J. B. S. (1932). A note on inverse probability. *Math. Proc. Cambridge Philos. Soc.* **28** 55–61.

HALDANE, J. B. S. (1937). The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit, when expectations are small. *Biometrika* **29** 133–143. MR0478438

HALDANE, J. B. S. (1938). The approximate normalization of a class of frequency distributions. *Biometrika* **29** 392–404.

HALDANE, J. B. S. (1945). On a method of estimating frequencies. *Biometrika* **33** 222–225. MR0019900

HALDANE, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35** 297–300.

HALDANE, J. B. S. (1951). A class of efficient estimates of a parameter. *Bull. Inst. Internat. Statist.* **23** 231–248. MR0088845

HALDANE, J. B. S. (1966). An autobiography in brief. *Perspect. Biol. Med.* **9** 476–481.

HALDANE, J. B. S. and SMITH, C. A. B. (1947). A simple exact test for birth-order effect. *Ann. Eugen.* **14** 117–124.

HODGKIN, D. C. and JEFFREYS, H. (1976). Obituary. *Nature* **260** 564.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

HOLMES, C. C., CARON, F., GRIFFIN, J. E. and STEPHENS, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal.* **10** 297–320. MR3420884

HOWIE, D. (2002). *Interpreting Probability*: *Controversies and Developments in the Early Twentieth Century*. Cambridge Univ. Press, Cambridge. MR2362916

JEFFREYS, H. (1931). *Scientific Inference*, 1st ed. Cambridge Univ. Press, Cambridge.

JEFFREYS, H. (1933). On the prior probability in the theory of sampling. *Math. Proc. Cambridge Philos. Soc.* **29** 83–87.

JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. *Math. Proc. Cambridge Philos. Soc.* **31** 203–222.

JEFFREYS, H. (1936a). Further significance tests. *Math. Proc. Cambridge Philos. Soc.* **32** 416–445.

JEFFREYS, H. (1936b). XXVIII. On some criticisms of the theory of probability. *The London*, *Edinburgh*, *and Dublin Philosophical Magazine and Journal of Science* **22** 337–359.

JEFFREYS, H. (1939). *Theory of Probability*. Oxford Univ. Press, Oxford. MR0000924

JEFFREYS, H. (1977). Probability theory in geophysics. *IMA J. Appl. Math.* **19** 87–96. MR0434330

JEFFREYS, H. (1980). Some general points in probability theory. In *Bayesian Analysis in Econometrics and Statistics*: *Essays in Honor of Harold Jeffreys* (A. Zellner, ed.) 451–453. North-Holland, Amsterdam.

KASS, R. E. (2009). Comment: The importance of Jeffreys's legacy. *Statist. Sci.* **24** 179–182. MR2655844

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KENDALL, M. G., BERNOULLI, D., ALLEN, C. G. and EULER, L. (1961). Studies in the history of probability and statistics: XI. Daniel Bernoulli on maximum likelihood. *Biometrika* **48**.

LAI, D. C. (1998). John Burdon Sanderson Haldane (1892–1964) polymath beneath the firedamp: The story of JBS Haldane. *Bull. Anesth. History* **16** 3–7.

LAMBERT, J. H. and DILAURA, D. L. (2001). *Photometry*, *or*, *on the Measure and Gradations of Light*, *Colors*, *and Shade*: *Translation from the Latin of Photometria*, *Sive*, *de Mensura et Gradibus Luminis*, *Colorum et Umbrae*. Illuminating Engineering Society of North America, New York.

LANE, D. A. (1980). Fisher, Jeffreys, and the nature of probability. In *R. A. Fisher*: *An Appreciation*. *Lecture Notes in Statist.* **1** 148–160. Springer, New York. MR0578901

LATTIMER, J. M. and STEINER, A. W. (2014). Neutron star masses and radii from quiescent low-mass X-ray binaries. *Astrophys. J.* **784**, 123.

LINDLEY, D. V. (1957). Binomial sampling schemes and the concept of information. *Biometrika* **44** 179–186. MR0087273

LINDLEY, D. (2009). Comment [MR2655841]. *Statist. Sci.* **24** 183–184. MR2655845

LY, A., VERHAGEN, J. and WAGENMAKERS, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *J. Math. Psych.* **72** 19–32. MR3506022

MALIKOV, E., KUMBHAKAR, S. C. and TSIONAS, E. G. (2015). Bayesian approach to disentangling technical and environmental productivity. *Econometrics* **3** 443–465.

OSTROGRADSKI, M. V. (1848). On a problem concerning probabilities. *St. Petersburg Acad. Sci.* **6** 321–346.

PEARSON, K. (1892). *The Grammar of Science* **17**. Walter Scott.

PEIRCE, C. S. (1878). The probability of induction. *Pop. Sci. Mon.* **12** 705–718.

PIRIE, N. W. (1966). John Burdon Sanderson Haldane. 1892–1964. *Biogr. Mem. Fellows R. Soc.* **12** 219–249.

PREVOST, P. and LHUILIER, S. A. (1799). Sur les probabilités. *Mémoires de L'Academie Royale de Berlin* **1796** 117–142.

ROBERT, C. P. (2016). The expected demise of the Bayes factor. *J. Math. Psych.* **72** 33–37. MR3506023

ROBERT, C., CHOPIN, N. and ROUSSEAU, J. (2009). Harold Jeffreys's theory of probability revisited. *Statist. Sci.* **24** 141–172. MR2655841

SARKAR, S. (1992). A centenary reassessment of JBS Haldane, 1892–1964. *BioScience* **42** 777–785.

SENECHAL, M. (2012). *I Died for Beauty*: *Dorothy Wrinch and the Cultures of Science*. Oxford Univ. Press, Oxford.

SENN, S. (2009). Comment on "Harold Jeffreys's theory of probability revisited". *Statist. Sci.* **24** 185–186.

SMITH, J. M. (1992). JBS Haldane. In *The Founders of Evolutionary Genetics* 37–51. Springer, New York.

SPARKS, D. K., KHARE, K. and GHOSH, M. (2015). Necessary and sufficient conditions for high-dimensional posterior consistency under g-priors. *Bayesian Anal.* **10** 627–664. MR3420818

TARONI, F., MARQUIS, R., SCHMITTBUHL, M., BIEDERMANN, A., THIÉRY, A. and BOZZA, S. (2014). Bayes factor for investigative assessment of selected handwriting features. *Forensic Sci. Int.* **242** 266–273.

TERROT, C. (1853). Summation of a compound series, and its application to a problem in probabilities. *Trans. R. Soc. Edinb.* **20** 541–545.

TODHUNTER, I. (1858). *Algebra for the Use of Colleges and Schools*: *With Numerous Examples*. Cambridge Univ. Press, Cambridge.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability*: *From the Time of Pascal to That of Laplace/by I. Todhunter*. Macmillan and Company.

TURING, A. M. (1941/2012). The Applications of Probability to Cryptography. UK National Archives, HW 25/37.

VON MISES, R. (1931). *Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theorestischen Physik*. Franz Deuticke.

WHITE, M. J. D. (1965). JBS Haldane. *Genetics* **52** 1.

WRINCH, D. and JEFFREYS, H. (1919). On some aspects of the theory of probability. *Philos. Mag.* **38** 715–731.

WRINCH, D. and JEFFREYS, H. (1921). On certain fundamental principles of scientific inquiry. *Philos. Mag.* **42** 369–390.

WRINCH, D. and JEFFREYS, H. (1923a). On certain fundamental principles of scientific inquiry (second paper). *Philos. Mag.* **45** 368–374.

WRINCH, D. and JEFFREYS, H. (1923b). The theory of mensuration. *Philos. Mag.* **46**.

ZABELL, S. (1989). The rule of succession. *Erkenntnis* **31** 283–321.

ZABELL, S. (2012). Commentary on Alan M. Turing: The applications of probability to cryptography. *Cryptologia* **36** 191–214.