Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 13

4-1-2020

On the Authentic Notion, Relevance, and Solution of the Jeffreys-Lindley Paradox in the Zettabyte Era

Miodrag M. Lovric
Radford University, mlovric@radford.edu

On the Authentic Notion, Relevance, and Solution of the Jeffreys-Lindley Paradox in the Zettabyte Era

Cover Page Footnote

Acknowledgement This research was supported by the Artis College Faculty Research Grant, Radford University, USA.

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP3249
doi: 10.22237/jmasm/1556670180

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

On the Authentic Notion, Relevance, and Solution of the Jeffreys-Lindley Paradox in the Zettabyte Era

Miodrag M. Lovric
Radford University
Radford, Virginia

The Jeffreys-Lindley paradox is the most quoted divergence between the frequentist and Bayesian approaches to statistical inference. It is embedded in the very foundations of statistics and divides frequentist and Bayesian inference in an irreconcilable way. This paradox is the Gordian Knot of statistical inference and Data Science in the Zettabyte Era. If statistical science is ready for revolution confronted by the challenges of massive data sets analysis, the first step is to finally solve this anomaly. For more than sixty years, the Jeffreys-Lindley paradox has been under active discussion and debate. Many solutions have been proposed, none entirely satisfactory. The Jeffreys-Lindley paradox and its extent have been frequently misunderstood by many statisticians and non-statisticians. This paper aims to reassess this paradox, shed new light on it, and indicates how often it occurs in practice when dealing with Big data.

Keywords: Jeffreys-Lindley paradox, Point null hypothesis, p-value, true and false null hypotheses, Fisherian significance testing, Neyman-Pearson hypothesis testing, Bayes factor

Introduction

The current dominant paradigm in statistical testing of a point null hypothesis is inadequate as our expression of uncertainty about the world in the 21st century. For decades, it has produced countless criticisms and recently even methodological crisis in some fields of science and has done serious damage to the image of statistics and statisticians. Within the paradigm, San Andreas fault, the Jeffreys-Lindley (henceforth JL) paradox is deeply embedded, shaking the foundations of statistics and dividing frequentists and Bayesians in an irreconcilable way. In order

doi: 10.22237/jmasm/1556670180 | Accepted: June 4, 2019; Published: April, 1, 2020.
Correspondence: Miodrag M. Lovric, mlovric@radford.edu

to solve accumulated anomalies within the current paradigm and rebuilt healthy foundations of statistical science it is inadequate just to cast out statistical significance (Wasserstein, Schirm, and Lazar, 2019). A method is needed to harmonize frequentist and Bayesian inference in hypothesis testing, and one of the first fundamental steps is the resolution of JL paradox. Regretfully, because of its complexity, statisticians do not interpret this paradox unanimously. They even disagree on whether it sheds a negative light on Bayesian or frequentist inference, as evidenced by the next two quotes.

> "Lindley's paradox has been misunderstood in several places, including by myself in the distant past. It is unfortunate that opposite to Lindley's written words, his 'paradox' has been misunderstood as an 'illness' of Bayes factors and posterior probabilities." Pericchi (2011, p. 20).

> "The Jeffreys-Lindley paradox has played an important role in undermining the credibility of frequentist inference…" Spanos (2013, p, 91).

Hence, before attempting to derive the solution, it is necessary to clarify its meaning.

Brief historical milieu of the JL paradox

The first formal significance test was undertaken by Arbuthnott (1710). From today's perspective, it can be argued that this was an auspicious event in statistics history. However, Arbuthnott opened a Pandora's box foreshadowing the controversies about the role of statistical tests. From one perspective, he correctly analyzed data on the yearly number of male and female christenings in London from 1629 to 1710 and demonstrated that boys were born at a greater rate than girls. This is the first recorded case of confusing statistical with scientific hypotheses, because Arbuthnott equated mere rejection of a null hypothesis with an irrefutable argument for divine providence. Moreover, this approach in testing was without any delay challenged by many (see Hald, 2003, p. 275-285.).

Modern frequentist statistical tests are usually regarded as an anonymous hybrid of two divergent classical statistical paradigms. Fisherian significance testing is founded on a single null hypothesis, p values, inductive reasoning, and drawing conclusions. By contrast, Neyman-Pearson hypothesis testing is

established on two hypotheses: null and an alternative, two types of errors, fixed-level significance statements, making decisions, and deductive reasoning but inductive behavior. These opposing views about the proper manner to conduct a test were never reconciled by their authors, nonetheless, been amalgamated by contemporary authors of statistics textbooks.

In the context of point-null hypothesis testing, Bayesians are forced to choose between the following two broad options.

1. Assign zero as the prior probability of the null hypothesis, $P(H_0) = 0$, because it specifies a single point on the real number line and has Lebesgue measure zero under absolutely continuous distribution. The reasoning is that $H_0 : \theta = \theta_0$ corresponds to a singleton, that is to the set $\{\theta_0\}$ that consists of one single point on the real line. The unfortunate consequence, however, is that the posterior probability of the null hypothesis is always zero and thus, impossible to revise on accumulated information. Therefore, Bayesians would never be influenced by any data and would always reject sharp nulls on a priori grounds, that is, without conducting any test. This standstill can be tackled in at least three following ways:

   First, conclude "the Bayesian approach helps to make clear the logical deficiencies of point null-hypothesis testing. Thus, at least for continuous parameters, we don't test point null hypotheses in the Bayesian approach, and for that matter nor should a frequentist". (Jackman, 2009, 32).

   Second, test point null hypotheses using Bayesian credible intervals (for example, Bolstad, 2007). Similarly, when the prior knowledge is vague and the prior distribution in the neighborhood is reasonable smooth, Lindley (1965, p.61) proposed that credibility of the null hypothetical value can be tested by checking whether or not it belongs to a chosen Bayesian credible interval.

   Third, circumvent this problem using decision theoretical framework. One of the best examples is so-called integrated objective Bayesian estimation and hypothesis testing developed by Bernardo (2011a).

2. As stated by Robert (2007) in order to compete with the traditional methods, "for pragmatic reasons Bayesian toolbox must incorporate

testing devices, if only because users of Statistics have been accustomed to testing as a formulation of their problems" (p. 223). Therefore, the majority of Bayesians follow the procedure initiated by Jeffreys and generate mixed prior distribution, with the positive probability mass assigned to a single point, that is $P(H_0) > 0$. Regardless of the probabilistic arguments scrutinized above, Lindley (2009, p. 184) considered this concept a triumph that provides a general method for the construction of Bayesian tests.

Lindley's original formulation of the paradox

The inharmonious conclusions reached between frequentist tests and Bayesian tests when analyzing sufficiently large samples were famously manifested by Lindley (1957), based on a comparison of the Fisherian significance test and Bayesian posterior probability in case of testing a point null hypothesis $\theta = \theta_0$ within a normal model with known variance σ^2. Lindley (1957) did not envisage alternative hypothesis and p-values, nor the critical values and regions (as did Neyman & Pearson, 1933). The typical Bayesian composition of the prior distribution, initiated by Jeffreys (1939), is to assign probability mass c to the single point indicating by null hypothesis $\theta = \theta_0$ and distributing the remainder, $(1 - c)$ according to the continuous density g(θ) over $\theta \neq \theta_0$. The resulting spike-and-smear prior distribution has the following form

$$P(\theta) = c\delta_{\{\theta=\theta_0\}} + (1-c)g(\theta)_{I\{\theta\neq\theta_0\}} \tag{1}$$

This prior distribution is a fusion of two components: a discrete part (where $\delta_{\{\theta=\theta_0\}}$ represents Dirac mass at θ_0) and a continuous part. Following this idea, Lindley assumed that the prior probability of the null was $P(H_0) = c$, and that the remainder of the prior probability $(1 - c)$ was assigned uniformly to an interval I which included hypothesized value θ_0.

One of the Lindley's motivations was to show vigilance is necessary when using a fixed significance level regardless of the sample size, because "5% in to-day's small sample does not mean the same as 5% in to-morrow's large one" (1957, p. 189). Hence, it was supposed the value of the sample mean was just significant at the α level, that is $\bar{x} = \theta_0 + \lambda_{\alpha/2}\sigma / \sqrt{n}$, where $\lambda_{\alpha/2}$ stands for the upper $\alpha/2$

quantile of the standard normal distribution. He evaluated the posterior probability that $\theta = \theta_0$ as

$$P\left(H_0 \mid \text{ just significant } \overline{x}\right) = \overline{c} = \frac{c\exp\left(-\lambda_{\alpha/2}^2 / 2\right)}{\left\{c\exp\left(-\lambda_{\alpha/2}^2 / 2\right) + \left(1-c\right)\sigma\sqrt{\frac{2\pi}{n}}\right\}} \qquad (2)$$

Lindley concluded as sample size increases, posterior probability of the null hypothesis approaches one. Therefore, for any value of the prior probability c, a value of sample size n can be found such that

I. $\overline{x}$ is statistically significant at the prescribed $\alpha\%$ level (conclusion obtained by traditional significance test), and at the same time

II. the posterior probability that $\theta = \theta_0$ is $(100 - \alpha)\%$ (conclusion reached by Bayesian analysis).

For example, when using traditional 5% significance level we are "95% confident that $\theta \neq \theta_0$, but have 95% belief that $\theta = \theta_0$" (p. 187). He called this conflicting situation the "strong contrast" (p. 190), and the paradox (p. 187).

Lindley (1957) pointed out this disagreement between frequentist and Bayesian results would persist "with almost any prior probability distribution that had a concentration on the null value and no concentration elsewhere" (p. 188). Essentially, the scope of Bayesian testing was restricted by claiming the hypothesized value θ_0 is fundamentally different from any other value of $\theta \neq \theta_0$. This is similar to the suggestion by Edwards et. al. (1963) for Bayesian statisticians "no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence" (p. 235). Lindley alleged testing some special value θ_0 is itself evidence it is likely to be true. To illustrate this claim two not so convincing examples (telepathy and genetic) were given, both based on the count data, although Lindley related the paradox to the continuous parameter testing.

Bartlett's inconsistency

Lindley formulated the paradox without mentioning dependence of the posterior probability on the prior variance. Expression of the posterior probability (2) indicates Lindley was not aware of this dependence. The only variance analyzed was the one of the original population. It was shown when this variance is very large and simultaneously sample size relatively small (hence the standard error $\sigma/\sqrt{n}$ is also very large), the posterior probability can be smaller than p-value and may give very strong evidence against the null hypothesis. Dependence was exposed by Bartlett (1957, p. 533) who corrected a slip in Lindley's analysis by including the extra factor for the uniform density $1/I$, as follows:

$$\overline{c} = \frac{c\exp\left(-\lambda_{\alpha/2}^2/2\right)}{\left\{c\exp\left(-\lambda_{\alpha/2}^2/2\right)+\frac{(1-c)}{I}\sigma\sqrt{\frac{2\pi}{n}}\right\}} \tag{3}$$

As properly revealed by Bartlett, this correction makes the value of posterior probability much more unstable, although "one might be tempted to put I infinity the silly answer $\overline{c}=1$ ensues." (p. 533). The following upsetting fact (upsetting for the Bayes factor) can now be demonstrated: for any fixed data and hence the fixed value of $\lambda_{\alpha/2}$ regardless of its magnitude, posterior probability tends to 1 as I increases. Hence, the evidence in favor of the null hypothesis is becoming increasingly more substantial.

Although this is sometimes called Bartlett's paradox (see, for example, Welsh, 1996, p. 87; LaMont and Wiggins, 2015; Bayarri and Berger, 2013, p. 366), there is nothing paradoxical in the fact that statistical analysis might be easily misused and give nonsensical answers. Hence, this inconsistency of the Bayesian testing using flat priors that can always lead to a non-rejection of any point-null hypothesis should be called more appropriately Bartlett's inconsistency.

The case of Bayes factor

According to Berger (2006) Bayes factor is the "primary tool used in Bayesian inference for hypothesis testing and model selection". It was proposed as an objective" (p. 38) Bayesian answer. "Bayes factor" is a fascinating example of the "Stigler's law of eponymy," which states "no scientific discovery is named after its original discoverer" (Stigler, 1980, p. 147). This name was coined by Irving John

Good (Isadore Jacob Gudak). Good (1958) introduced the term "Bayes factor" (p. 803) and attributed the idea to Turing. It is well-known that Thomas Bayes did not mention anything similar to the Bayes factor in his landmark posthumous essay with the factual title "A Method of Calculating the Exact Probability of All Conclusions founded on Induction", not "An Essay towards solving a Problem in the Doctrine of Chances" (Stigler, 2013, p. 283). Bayes factors are based on the Bayes theorem. However, Bayes did not give "the statement of Bayes theorem, either in its discrete form… or in its continuous form with integration" (Fienberg, 2006, p. 3). Although Stigler (1983, p. 290) hypothesized Bayes theorem was discovered 12 years before Bayes's death by Nicholas Saunderson, it is Bayes who "deserves and gets credit for noticing an interesting but mathematically trivial consequence of the product axiom of probability" (Good, 1965, p. 1). The concept of the Bayes factor was explicitly introduced by Wrinch and Jeffreys (1921, p. 387), not by Good. Finally, in 1931 J. B. S. Haldane made an "important intellectual advancement in the development of the Bayes factor" (Etz and Wagenmakers, 2017, p. 327).

The same irreconcilable conflict between frequentist and Bayesian testing is shared by the Bayes factor. This conclusion can be simply derived by the expression (1) in Bartlett's comment on the Lindley's paradox (1957, p. 533):

$$\frac{\overline{c}}{1-\overline{c}} = \frac{c}{1-c}\left[\frac{I}{\sigma}\sqrt{\frac{n}{2\pi}}\exp\left(-\lambda_{\alpha/2}^2/2\right)\right] \tag{4}$$

Posterior Odds = Prior Odds × Bayes Factor

Obviously, Bartlett could not specifically mention the term "Bayes factors", because this expression was coined a year later, in 1958. We can deduce that for any fixed prior c, and any constant p-value corresponding to a fixed outcome of a significance test λ, Bayes factor in favor of the null hypothesis increases as $\sqrt{n}$ with the sample size and goes to infinity. This means Bayes factor might exceedingly favor null value θ_0 even for datasets extremely inconceivable under H_0.

The Bayes factor for the null hypothesis may be arbitrarily large for sufficiently large sample size, for almost any choice of mixture prior distribution that has a mass on θ and no concentration elsewhere. For example, consider testing a point null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ in case of normal mean with known variance. Using (1) assign the mass π_0 to the null point $\theta = \theta_0$, and spread the remaining mass out on H_1 according to the conjugate prior density

$g(\theta) = N(\mu_0,\sigma_0^2)$, where μ_0 is the prior mean, and σ_0^2 prior variance. As pointed out by Berger and Sellke (1987, p. 112) this prior closely follows Jeffreys recommendation for testing a point null. It can be confirmed (see, for example, Migon, et al., 2014, p. 238) Bayes factor in favor of the null over the alternative may be expressed as

$$B_{01}(x) = \left[\left(\sigma^2 + n\sigma_0^2\right)/\sigma^2\right]^{1/2} \exp\left\{\frac{n}{2}\left[\frac{\left(\bar{x}-\mu_0\right)^2}{\left(\sigma^2+n\sigma_0^2\right)} - \frac{\left(\bar{x}-\theta_0\right)^2}{\sigma^2}\right]\right\} \tag{5}$$

where $\bar{x}$ is the sufficient statistic for θ. To make a comparison between frequentist and Bayesian frameworks, using the exact same conditions as Berger and Sellke (1987) and Berger and Delampaday (1987), we will center prior density $g(\theta)$ over the hypothetical mean value, that is $\mu_0 = \theta_0$, and equate prior variance with the known variance, that is, $\sigma_0^2 = \sigma^2$. Then (5) reduces to

$$B_{01}(x) = \left(1+n\right)^{1/2} \exp\left\{-\frac{z^2}{2}\frac{n}{n+1}\right\} \tag{6}$$

where $z = \dfrac{\sqrt{n}\left(\bar{x}-\theta_0\right)}{\sigma}$ is a familiar classical test statistic. Therefore, using (6), the posterior probability of H_0 is simply calculated as

$$\pi\left(H_0 \mid x\right) = \left[1 + \frac{\left(1-\pi_0\right)}{\pi_0}\frac{1}{B_{01}}\right]^{-1} = \left[1 + \frac{\left(1-\pi_0\right)}{\pi_0}\left(1+n\right)^{-1/2}\exp\left\{\frac{z^2}{2}\frac{n}{n+1}\right\}\right]^{-1} \tag{7}$$

Both the Bayes factor and posterior probability are susceptible to the Lindley's paradox.

Discussion

As the founder of the Bayes factor, Jeffreys (1939) was the first statistician who had noticed the disagreement between *p* values and Bayes factor. In the Appendix B, it was pointed out "at large numbers of observations there is a difference, since the test based on the integral [p value] would sometimes assert significance at

departures which would actually give $K > 1$ [$B_{01}(x) > 1$]. Thus there may be opposite decisions in such cases. But these will be *very rare* [italicized by author]" (p. 435).

To verify Jeffreys claim, a program was developed in R, and the source code is attached in Appendix A. The results of the simulation are displayed in Table 1. The program is based on testing average height (in cm) of women in New Zealand under assumption that the population variance is known to be 9. It calculates a number of cases for which Bayes factor supports true and false null hypotheses, respectively, that were previously rejected using traditional significance testing. It also evaluates posterior probabilities of the null hypothesis in the following normal conjugate model testing scenario:

$H_0 : \mu = 170$; $\sigma^2 = 9$; "objective priors" equal 0.5;

Prior mean = hypothesized value = 170; Prior variance = known variance ($\sigma_0^2 = \sigma^2$)

Table 1 is based on 1,000 trials (iterations) for each cell and displays frequencies of cases for which classical tests produced significant results, yet Bayes factors and posterior probabilities inclined to support the null hypothesis (with values larger than 1 and 0.5, respectively). To illustrate our argument, that Jeffreys claim is incorrect, in case of false null hypothesis, a value in the vicinity of the null was taken as the true parameter value ($\mu = 170.007$).

Table 1. Number of opposite conclusions made by a significance test and Bayesian analysis in 1,000 trials per each cell.

		Sample size *n*				
		$\theta = \theta_0$		$\theta \neq \theta_0$; $\mu = 170.007$		
z	*p*-value	100,000	1,000,000	100,000	800,000	1,000,000
1.959	0.050	53	46	103	488	550
2.326	0.020	22	14	47	343	425
2.576	0.010	11	9	23	257	332
2.807	0.005	5	3	14	173	237

As Table 1 confirms, Jeffreys' statement is inaccurate, because he was unaware of the conflict between the Bayes factor and frequentist p-value. Discrepancies do not happen sometimes, and their occurrences are not very rare: discrepancies occur regularly, provided the sample size is large enough. Furthermore, his remark did not point to the Lindley's paradox in its strict sense,

but to any disagreement between frequentist and Bayesian testing result. This table also demonstrates that the similar claim made by Edwards et. al. (1963, p. 235) "that results of [Bayesian and classical testing] procedures will usually agree" is unsubstantiated.

The most striking result in Table 1 is the number of opposite conclusions (237 out of 1,000) for significance level 0.005 and the sample size one million. This disturbing result indicates that the recommendation given by Benjamin and 72 other eminent statisticians (2018) to lower the impact of the reproducibility crisis in science by lowering threshold for defining statistical significance for new discoveries to 0.005 is unfortunately unproductive and cannot reconcile frequentist and Bayesian inference. Surely, we applaud their effort, but a different elucidation has to be adopted in order to solve all accumulated anomalies in the deepest foundations of statistical science. At least, all simulation results are easy to reproduce, because each simulation starts with the same seed.

The notion of the JL paradox may be summarized as follows: In a Gaussian model $N(\theta, \sigma^2)$ with known variance σ^2, when testing a point null hypothesis $\theta = \theta_0$ there is a general disagreement between frequentist and Bayesian conclusions when an analysis is based on sufficiently large samples. Particularly, for any fixed value of significant frequentist test statistic z (and therefore any fixed significant p-value), for any fixed prior probability of the null hypothesis strictly larger than 0, $P(H_0) > 0$, and for almost any choice of prior distribution that has a concentration on the null value θ_0, posterior probability of the null hypothesis $P(H_0 \mid x)$ tends to 1, and at the same time Bayes factor tends to infinity, with increasing sample size. This also means that the posterior probability that $\theta \neq \theta_0$ tends to zero. In its strict sense, JL paradox is attained when the p-value is significant at α% (say 0.005) and at the same time the posterior probability that $\theta \neq \theta_0$ reaches the same level (0.005). In other words, the frequentist test will reject the point null hypothesis and Bayesian test will support it.

Table 2. Number of wrong conclusions in case of true null hypothesis (1,000 iterations)

	Significance test		Bayes factor : Evidence against H_0		
n	p-value = 0.05	p-value = 0.005	Slight	Substantial	Strong
20	45	3	74	13	1
30	55	5	59	14	1
50	44	6	40	15	0
100	55	4	25	9	0
500	47	8	14	6	0
10,000	53	2	2	1	0
50,000	52	2	1	0	0

It is remarkable to note that Lindley did not relate the paradox to the veracity of the null hypothesis, i.e. whether the null hypothesis was true or false. In relation to this, the following two disappointing facts can be established.

1. If the null hypothesis is true, significance test will reject the null only when Type I error is committed, which according to the Neyman-Pearson model can be expected in a relatively small proportion of cases depending on the preselected significance level α%. In these instances, frequentist test will lead to the wrong conclusion and Bayes factor will be asymptotically correct. However, small samples can mislead Bayes factor to wrongly reject the true null hypothesis since the value of Bayes factor depends on the significant value of the classical test statistic, as shown in (6). In other words, the Bayes factor is also susceptible to the type I error. Table 2 illustrates this situation for different sample sizes and is established on the classification of different values of Bayes factor given by Kass and Raftery (1995, p. 777).

   Based on Table 2, for moderate and relatively large sample sizes, Bayes factor has lower Type I error rate, when "substantial" or "strong" evidence against H_0 are used. Ironically, the most astonishing results are obtained for the class "slight". It is worth mentioning that with small samples, so classified Bayes factor may wrongly reject true null in more instances than a significance test. Finally, when sufficiently large sample is reached, Bayes factor always support true null, i.e. Type I error cannot be committed. On the other hand, a significant test is repeatedly (proportionally to the number of applications) prone to Type I error, regardless of the sample size.

2. When the point null value does not hold, JL paradox implies that Bayes factor becomes increasingly misleading by supporting false null. Contrariwise, with sufficiently large sample, frequentist test will detect that the null is false since the test statistic will converge almost surely to infinity.

Obviously, the problem is that when we are confronted with a single conflicting result—Bayes factor supports H_0 and p-value is significant—it is not possible to deduce the true cause. We can only treat this inversely, by running simulation assuming that null takes different values. Hence, the last section of this

paper will try to answer the most important question: how often JL paradox occur when testing large samples if the null is true and if the nulls are false. In other words, it will answer a rather provocative question stated as a title of the recent article written by Aris Spanos (2013): who should be afraid of the JL paradox?

According to Robert (2014) "[T]here is obviously no mathematical issue with the paradox—otherwise it would have been readily dismissed" (p. 218). Consider the root of the JL paradox. Lindley required (1957, p. 187) $\bar{x} \to \theta_0$ as n increases. Similarly, Sprenger (2013, p. 736) asked the following question: "Why is that this result [Bayes factor] diverges so remarkably from the frequentist finding of significant evidence against the null?" His opinion is that if the p-value has to remain constant when the sample size increases without bound, the sample mean has to converge to the hypothesized null value, favoring it over the alternatives. It can be argued this occurs because of the unnatural composition of the Jeffreys mixed prior model. Dirac's mass at θ_0, $\delta_{\{\theta=\theta_0\}}$ behaves like a black hole: it exhibits such a strong gravitational pull that absorbs any evidence, no matter how strong, against the null hypothesis. The point at which the gravitational pull of the Dirac's mass becomes so great to exclude any evidence against H_0, can be called Jeffreys event horizon. This horizon will be always reached; the only question is for what sample size.

However, according to the Law of Large Numbers the sample mean almost surely converges to the true value of the parameter, not to the hypothesized value. Consider this as the fundamental statistical inaccuracy as the hidden root of the JL paradox, and it is caused by the presence of the Dirac's mass. From this perspective, JL paradox could be understood as a rather artificial process, devoid of reality, in which a sample mean in the long run converges to the hypothesized value of the parameter. This is happening because Lindley required that with an increasingly large sample, the sample mean has to be just significant. The interesting question is how is it possible to occur? By carefully examining his expression for the significant sample mean $\bar{x} = \theta_0 + \lambda_{\alpha/2}\sigma / \sqrt{n}$, we can see that θ_0, σ, and $\lambda_{\alpha/2}$ are constants, and that $\bar{x}$ has to be adjusted relative to n. Using the same testing setup as before, $\theta_0 = 170$, $\sigma = 3$, and $\lambda_{\alpha/2} = 1.96$ we obtain Table 3.

It is now obvious what is the driving mechanism of the JL paradox. As the sample size increases, the sample mean in each subsequent sample has to take a specific value so that it converges to the hypothesized value ($\bar{x} \to 170$). Of course, this makes sense only when the point null hypothesis is true. However, when H_0 is false, by preserving the same significance level, as the sample size increases the sample mean is forced to approach wrong hypothesized value. Undoubtedly, this is

not justifiable, because the sample mean is an unbiased estimator. In these cases significant sample mean will lead to the correct rejection by a classical test.

Table 3. Convergence of just significant sample means in case of JL paradox

$\overline{x}$	θ_0	N	σ	$\lambda_{\alpha/2}$
175.880000	170	1	3	1.96
171.859419	170	10	3	1.96
171.314808	170	20	3	1.96
170.588000	170	100	3	1.96
170.185942	170	1,000	3	1.96
170.058800	170	10,000	3	1.96
170.018594	170	100,000	3	1.96
170.005880	170	1,000,000	3	1.96

However, if Bayesians in most applications do not treat point null hypothesis as a single point, but as a hazily defined small region (Edwards et al., 1963, p. 214; Kadane, 1984, p.54; Bernardo, 2011b, p. 301), Bayes factor by supporting the wrong null when effect size is scientifically irrelevant, may provide some protection to researchers. In contrast, as observed by Hodges and Lehmann (1954, p. 261) "whenever the available data are extensive, the [significance] tests may become embarrassingly powerful" and reject trivially significant point null hypothesis. We argue that these imprecisions and difficulties should be avoided by making a clear distinction between sharp and interval null hypotheses in the formulation of the problems. Otherwise, a blind equalization of these two forms of null hypotheses will lead to further inconsistencies.

Different views and attempts to solve JL paradox

There were many attempts in the literature to resolve detrimental consequences imposed by the JL paradox, including (Shafer, 1982; Bernardo, 1980; Robert, 1993; Sprenger, 2013; and Naaman, 2016). An excellent review from a non-statistician perspective is given in Cousins (2014), from a frequentist perspective in Spanos (2013), and from a Bayesian in Robert (2014). Unfortunately, after 56 years of intensive discussions and debates since JL paradox was enunciated, Gelman and Shalizi (2012, p. 22), assert that the final verdict is that "the Jeffreys–Lindley paradox… *is really a problem without a solution*" [italicized by author].

Some tried to uncover a flaw in Lindley's argument. Bartlett (1957, p. 534) suggested that in the uniform priors settings the sample size should be chosen in

such a way to make $\sqrt{n}$ proportional to 1/*I*. Likewise, Bernardo (1980, p. 613) argues that Lindley did not include the factor $\sqrt{n}$, thus compensating for the different dimensionalities of sharp H_0 and infinitely diffuse H_1. The most convincing explanation for making this adjustment was made by Cox (2006, p. 106). He strongly advocates modification of the Jeffreys "mixed prior distribution paradigm" by showing that the fixed prior density over H_1, $g(\theta)$, should not be independent of the sample size *n*. He argues that $g(\theta)$ should usually be taken in the form $g\left\{\left(\theta-\theta_0\right)\sqrt{n}\right\}\sqrt{n}$. The relation between the p-value and the posterior odds would be independent of *n*. However, Bernardo (1999, p. 102) objects this is a rather artificial solution. A similar fix of JL paradox is suggested by Naaman (2016, p. 1526), by "allowing the significance level to decrease with the number of observations in the study". However, JL paradox strictly requires that the same significant level is used throughout the process, hence this resolution can be only classified as contradictio in adiecto. In spite of that, this suggestion is offered in Wikipedia as a reconciliation of the Bayesian and frequentist approaches to force a frequentist test to support a ridiculous claim that in some imaginary city boys were born at the same rate as girls.

There are numerous possible resolutions with a goal to avoid JL paradox by finding alternatives to the standard Bayes factor. For example, Shafer (1982) proposed using the theory of belief functions. Robert (1993) advocated noninformative answer by imposing dependence between the prior probability of the null hypothesis and the prior variance under the alternative hypothesis that leads to the same decisions as the p value. Claiming that both frequentist significance tests and subjective Bayesian inference failed to resolve the JL paradox, Sprenger (2013) recommended using Bayesian Reference Criterion (see, for example, Bernardo, 1999) that gives a sensible treatment of the paradox. Many other variants of the Bayes factors have been proposed to overpower problems related to the usage of improper priors, like intrinsic Bayes factor (Berger and Pericchi, 1996), fractional Bayes factors (O'Hagan, 1995), etc. Another important reference is Li et al. (2014), who proposed a new Bayesian test statistic based on the difference between the two deviances averaged over the posterior distribution. This test is immune to JL paradox and constructed in a decision theoretical framework.

The most impressive approach in modern Bayesian analysis that, inter alia, dismisses JL paradox, is the "Integrated objective Bayesian estimation and hypothesis testing," (Bernardo, 2011a). Since the Jeffreys pioneering book, Bayesians have usually used two fundamentally different types of priors, one

category for estimation and radically different (and often polemic) for testing point-null hypotheses. Bernardo's all-encompassing approach facilitates objective decision-making by using prior distributions that only depend on the assumed model and the quantity of interest. Particularly, the same prior distribution may be used for point estimation, region estimation, and point-null hypotheses testing.

However, Bernardo's usage of loss functions may not be directly relevant to inference problems. In Fisher's words (Fisher, 1935, pp. 25-26) "In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretense, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence". A similar attitude was shared by Lindley (in Bernardo, 1999, p. 122), that "hypothesis testing is, in principle…the calculation of $P(H_0 \mid x)$, for data $\boldsymbol{x}$. It is part of our total expression of uncertainty about the world…it has no element of decision-making in it." Furthermore, Bernardo's approach suffers from certain still unresolved problems and is not fully accepted among Bayesians (see, for example, Bernardo (2011a, pp. 25-50) for opposing views.

Consider some different views and misapprehensions of the JL paradox. One idea is to show how extremely complex and confusing the nature of this paradox is. It is curious that in the vast literature where JL paradox is discussed, many authors reduce its importance or interpret it quite differently.

1. The concept of JL paradox is not identical to Bartlett's inconsistency. Even if we consider Jeffreys-Lindley-Bartlett paradox as a single entity with two dimensions (Lindley's and Bartlett's), solving one dimension is just a local solution, not the global one. This equivalence was incorrectly alleged in Robert's early paper (1993, p. 601): "[T]he Jeffreys-Lindley paradox, namely the fact that a point null hypothesis will always be accepted when the variance of a conjugate prior goes to infinity…" As we have already discussed, what Robert emphasizes by this, is just a troublesome behavior of the Bayes factor that could be misused by unethical researchers to "prove" any null hypothesis. Specifically, from (5) it is clear that for any fixed $\overline{x}$ and prior probability of H_0, $B_{01}(x) \longrightarrow \infty$ as prior variance (set by a researcher), σ_0^2, increases ad infinitum. Simultaneously, regardless of the data, for any fixed prior, π_0, the posterior probability (6) of the null hypothesis can be made as close to one as desired. As Bernardo (1980, p. 607) rightfully observes, this is "rather disturbing, for a large prior variance has been traditionally accepted as a description of vague initial

knowledge." Lindley's posterior probability given in (3) exhibits the same deficiency.

There are three interesting points related to the abovementioned Robert's paper with the title "A note on Jeffreys-Lindley paradox".

a) If one substitutes all occurrences of the words "Jeffreys-Lindley paradox" with "Bartlett's inconsistency" his analysis would be correct,

b) No one has noticed this slip, although many have quoted it, and

c) Subsequently, Robert denied (Rousseau and Robert, 2011, p. 137) the usefulness of his "solution of the Jeffreys-Lindley paradox", claiming that it is flawed from the measure theoretic-angle. A simple proof that Robert grossly missed the topic is the fact that Lindley was not at all aware of the dependence of the posterior probability on the prior variance; he even made an omission by excluding the width of the interval from the posterior probability $\overline{c}$.

The JL paradox is an enormously complex and slippery issue. Robert (2007, 2014) revisited the JL paradox to correct the initial viewpoint and stated there is a "dual interpretation" (2014, p. x). Unfortunately, the earlier 2007 reference stimulated authors to the same misinterpretation. For example, Villa and Walker (2017, p. 12290) claimed Lindley "shows that, for point null hypothesis testing, there may be a concern with the objective Bayesian approach. In the specific example used, if the prior for the location parameter, in the alternative model to the parameter being zero, has infinite variance, then the Bayesian will always select the null model, regardless of the observed data." As pointed above, it is clear that Lindley never showed that. Analogous imprecision is also shared by Moreno (2011, p, 41). Similarly, Baskurt and Evans (2013, p. 579) presented their solution of the JL paradox, but their discussion was based on Bartlett's inconsistency. It seems that the solution of JL paradox should be equated with the solution of the problem of squaring the circle.

2. According to Good (1982, p. 342) , "Jeffreys paradox is closely related to the fact, not mentioned by Jeffreys, that a user of tail-area probabilities [a frequentist] can cheat, and can reach arbitrarily small tail-area probabilities, if he is allowed to use optional stopping, even when the null hypothesis is true." However, we have just proved that the similar endless opportunities for cheating exist also within the Bayes factor framework.

3. JL paradox is not "the fact that in testing a point null hypothesis for a fixed prior, and posterior cutoff points… as the sample size goes to infinity, $P(H_0) \longrightarrow 1$." (Gill, 2015, p. 228). It is the posterior probability of H_0 that tends to 1, not prior. In Gill's notation (from page 218), this should be corrected to $P(H_0 \mid \text{data}) \longrightarrow 1$.

4. The moral of the JL paradox is not "that if you pick a stupid prior, you can get a stupid posterior" (Christensen et al., 2011, p. 60). As pointed out by Lindley, almost any prior distribution concentrated on the hypothetical value θ_0 inevitably leads to the paradox. Nevertheless, Christensen is more unambiguous in (2005, p. 123): "Bayesian tests can go seriously wrong if you pick inappropriate prior distributions."

5. JL paradox does not indicate that "if the convention of applying a significance level of 0.01 or 0.05 is followed, then Lindley's `paradox' shows that with growing sample size, any hypothesis will be rejected." (Keuzenkamp, 2000, p. 54).

6. The following misconception is similar to the previous one. It amounts to the claim that "the large n sample problem was initially raised by Lindley" by pointing out that "there is always a large enough sample size n for which any simple null hypothesis $H_0 : \mu = \mu_0$ will be rejected by a frequentist α-significance level test" (Spanos, 2014, p. 646, and similarly in 2013, p, 73). First, large n problem was initially discussed by Berkson (1938), not by Lindley. Second, Lindley based his paradox explicitly on selecting sample means that were just significant at the α percentage point, regardless of the sample size. Finally, Lindley showed that for all these significant means, regardless of the significance level, for sufficiently large samples, posterior probability can be calculated that would support the null hypothesis. Hence, there is always a large enough sample size n for

which any simple null hypothesis $H_0 : \mu = \mu_0$ will be supported by the Bayes factor, not rejected by a frequentist test. To illustrate this, consider a normal model with conjugate priors. If a frequentist test based on just five elements produced a significant value that is ten standard errors away from the hypothesized mean ($z = 10$), than for the huge sample size $n = 10^{49}$, Bayes factor (1,928.75) and posterior of the null (0.999), would strongly favor null hypothesis over the alternative.

7. JL paradox does not describe a situation "when the p-value is very close to zero but the probability of correlation being true is very close to zero as well" (Zhu et al., 2012, p. 41).

How often JL paradox may occur in applied Statistics?

To answer this question, we will present the results of simulations obtained by a program developed in R. Before that, let us summarize some main conclusions we have reached so far, for the fixed level of significance, in Figure 1.

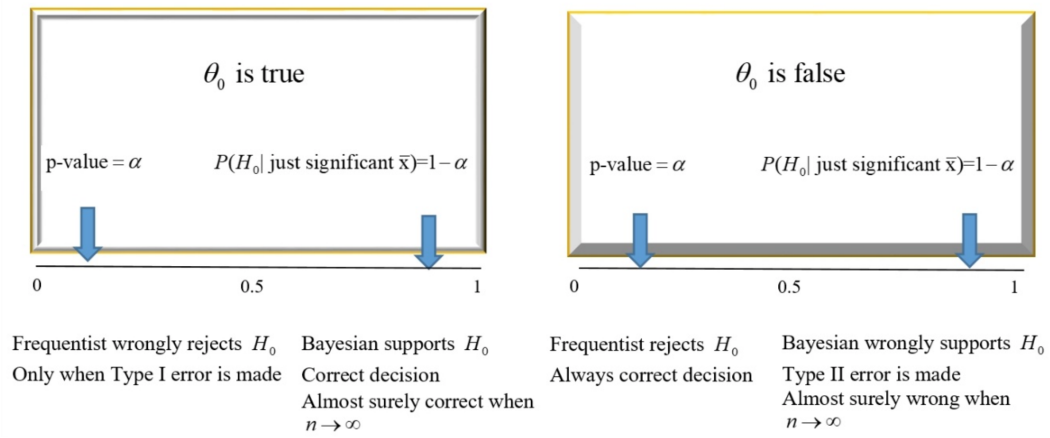

Figure 1. Implications of JL Paradox

Figure 1 shows the conflicting conclusions derived using a frequentist test and Bayesian testing when a sample mean is taken to be just significant, and sample size increases. When this condition is imposed, one of the most striking implications is that Bayes factor will always support null hypothesis, be it true or

false, even for data that are extremely implausible under the null hypothesis. As per our previous example, this would happen for a very large frequentist test statistic $Z = 10$, when the sample size is $n = 10^{49}$. In this case Bayes factor (1,928.75) and posterior of the null (0.999), would strongly favor null hypothesis over the alternative. To find the answer what would be appropriate scientific conclusion in this conflicting case (frequentist or Bayesian), let us quote Jeffreys opinion. While commenting differences between $z \leq 2$ and $z \geq 3$. (1980, p. 453) he stated that "…differences up to twice standard error usually disappear when more or better observations become available, and that those of *three or more times usually persist* [italicized by author]." What would Jeffreys think about persistence of 26 standard errors difference away from the hypothesized value, with a huge sample size 10^{299} where Bayesian analysis (Bayes factor = 1,616 and $P\left(H_0 \mid \bar{x}\right) = 0.999$) exceedingly support the null value?

As portrayed in Figure 1, we considered only the cases of JL paradox in its strict sense. Consequently, as a first step, we have found minimum sample size that is required for JL paradox to materialize, depending on the significance level. These values are displayed in Table 4, for both cases (original Lindley's set-up and normal conjugate priors). For example, when the level of significance is fixed at 5% level, the smallest sample size to produce JL paradox in Lindley's set-up is 105,685 and with normal conjugate priors 16,816 (Lindley made several slips in his 1957 paper when calculating similar minimum sample sizes; on page 190 bottom, correct sample size to reach the strong contrast in his paradox is 105,685, not 10,000; on page 191 top, number of trials in an experiment to raise our belief that telepathy did not exist to 95% is not 1,600, but 16,910).

Table 4. Minimum sample size to induce of JL paradox in its original denotation

		Minimum *n*	
α level	$P(H_0 \mid$ just significant $\bar{x})$	Lindley's set-up	Normal conjugate priors
0.050	0.950	105,685	16,816
0.040	0.960	245,701	39,098
0.030	0.970	728,954	116,011
0.020	0.980	3,380,074	537,945
0.010	0.990	46,875,786	2,195,961
0.005	0.995	657,481,111	104,625,626
0.001	0.999	315,983,097,898	50,212,131,719

Table 5. Frequency of occurrences of JL paradox in its original sense for normal conjugate priors

		Sample size n							
		$H_0 : \mu = 170$				$H_1 : \mu \neq 170$			
z	p-value	100,000	800,000	10,000,000	250,000,000	Δ = 0.01 100,000	Δ = 0.007 800,000	Δ = 0.002 10,000,000	Δ = 0.0004 250,000,000
1.959	0.050	31	45	53	44	107	308	417	506
2.054	0.040	18	42	45	36	54	237	354	454
2.170	0.030	-	18	31	28	-	141	284	394
2.326	0.020	-	4	16	14	-	35	198	310
2.576	0.010	-	-	1	5	-	-	30	169
2.807	0.005	-	-	-	3	-	-	-	38

Simulation results based on 1,000 iterations are presented in Table 5 for various fixed level of significance, several values of n, for a true null hypothesis, and also for several false null hypotheses, where Δ denotes the distance between the true alternative and the false null hypothesis. For instance, when $\Delta = 0.01$ the true alternative value was taken to be 170.01. The implication of the JL paradox is evident. The most remarkable empirical result is obtained for the samples $n = 250$ million units and the shift value 0.0004; JL paradox occurs in more than 50% of trials. This outcome conveys a clear alarming message: JL paradox happens frequently and persistently in the Zettabyte Epoch, where statistical analysis and data science are routinely applied on enormous datasets.

A graphical comparison of the number of times JL paradox occurs depending on the power of a frequentist test for significance level 0.05 is given in Figure 2.

The bottom curve that is superimposed shows empirical frequencies when H_0 is true, with the sample size chosen to correspond to $\Delta = 0.01$. It is easy to notice several regularities: 1) for the same power, frequencies are considerably larger for the false null hypothesized value closer to the true parameter values, 2) the shape of the empirical frequency curves are similar, with the largest number for power between 0.6 and 0.8, and 3) as anticipated JL paradox occurs much more often for false null hypotheses. However, it would be imprudent just to compare frequencies and to conclude that Figure 1 and Figure 2 when observed jointly convey a sinister message to the Bayesians that they should be afraid of JL paradox. This naturally depends on the proportion of the true null hypotheses. Nevertheless, Figure 2 reiterates one of the most important commandments to the Bayesian statisticians: "for Bayesian statisticians, however, no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence." (Edwards et. al., 1963, p.235).

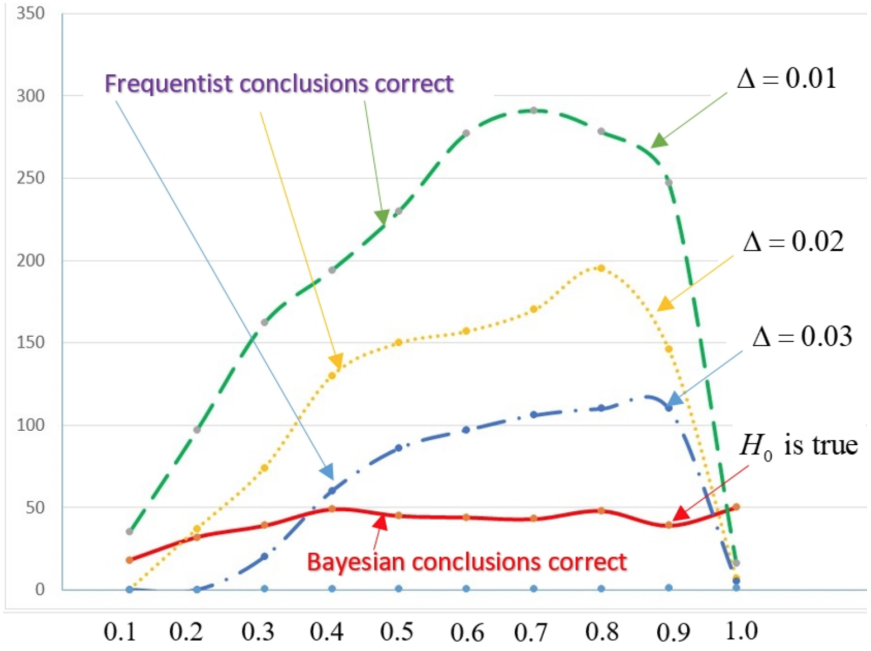

Figure 2. Frequency of occurrences of JL paradox depending on the power of the frequentist test

Conclusion

> "We [statisticians] will all be Bayesians in 2020, and then then we can be a united profession" A. Lindley (Smith, 1995, p. 317).
>
> "I have lamented that Bayesian statisticians do not stick closely enough to the pattern laid down by Bayes himself: if they would only do as he did and publish posthumously we should all be saved a lot of trouble" (Kendall, 1968, p. 185).

Probably the most appropriate characterization of JL paradox is given by Bernardo (2011b, p. 302): "unappealing behaviour of posterior probabilities based on sharp, non-regular priors—generally known as Lindley's paradox— is always present in the conventional Bayesian approach to sharp hypothesis testing". The JL paradox is induced by the assignment of a probability mass (Dirac's mass) to a single point that has Lebesgue measure zero in the Jeffreys mixed prior model. This is contrary to Lindley's assertion (1957) "paradox arises because the significance level

argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve" (p. 189-190) This is not the case, because the JL paradox does not occur in one-sided testing when it is possible to reconcile measures of evidence between the frequentist and Bayesian paradigm. As proved by Casella and Berger (1987, p. 106) for many categories "of reasonable prior distributions the infimum of the Bayesian posterior probability of H_0 is equal to the p-value, or even strictly lower bound on the p-value". Why there is no general conflict in one-sided testing? Because in this case, Bayesians do not need Jeffreys mixed prior model!

Lindley formulated the paradox to show it is not always appropriate to use the same prescribed significance level. There is nothing unanticipated in his recommendation. For example, Fisher (1973) shared the same opinion: "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (p. 44-45).

Using simulation, a conjecture by Jefferys (1990) that JL paradox is not a "pathological result of little practical interest" (p. 154), but instead a general phenomenon was confirmed. Hence, this paradox is not of "questionable relevance" (Berger & Delampady, 1987, p. 322). With zettabyte datasets, the number of conflicting conclusions between frequentist and Bayesian measures escalates. As an illustration, consider the radically different answers among Bayesian statisticians in Bernardo (2011a, pp. 25-50), when analyzing a huge sample (n = 104,490,000 with 52,263,471 successes) obtained while testing for extra sensorial perception. The very small p-value, 0.0003, suggests compelling evidence against the point null hypothesis ($\theta_0 = 0.5$). Bernardo (2011a, p. 19) reached the same conclusion based on the minimum likelihood ratio against the null 1,400. However, the Bayes factor obtained using a mixed prior distribution with a uniform prior under the alternative amounts to 12, and thus supports H_0 (Jefferys, 1990, p. 159).

If the next "generation of statisticians must build tools for massive data sets" (Laan & Rose, 2010, p. 38), then this generation should not leave fundamental paradoxes unresolved. Otherwise, it is highly likely that other journals will follow the example of BASP (Basic and Applied Social Psychology) by banning "NHSTP" (Null Hypothesis Significance Testing Procedures), and lead us to a deeper crisis. After all, BASP editors did not put the veto only on the statements about significant differences, but also on p-values, t-values, and confidence intervals (Trafimow and Marks, 2015). To regain confidence in statistical testing within the scientific community we need to reconcile measures of evidence between frequentist and Bayesian approaches. This was brilliantly indicated in the next two

Good's sentiments: "The resolution of inconsistencies will always be an essential method in science" (1952, p. 107), and "a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning" (1980, p. 489). In the case of JL paradox there are at least seven options:

1. Abandon p-values as requested by Berger and Delampady (1987, p. 330), since there exists "irreconcilability of traditional measures and evidence" (Berger and Sellke, 1987). In addition, a) with increasing sample size p-value consumers will reject almost any sharp null hypothesis, and thus frequently highlight trivial findings, and b) conclusions based on p-values are always susceptible to type I and II errors,

2. In the light of recent Robert's article "The expected demise of the Bayes factor", put the kibosh on the standard Bayes factor because with large enough samples it will almost always support any point null value; ultimately, as stated by Bernardo (2011a, p. 56), "Bayes factors have no direct foundational meaning to a Bayesian: only posterior probabilities have a proper Bayesian interpretation.",

3. Dismiss JL paradox by relying on Bernardo's integrated objective Bayesian estimation and hypothesis testing (or similar),

4. Calibrate p-values or posterior probabilities,

5. Develop a different Bayes factor that is not prone to JL paradox,

6. Keep on searching for "the statistical holy grail: prior distributions reflecting ignorance" (Fienberg, 2006, p.5),

7. Develop a new paradigm of Bayesian testing like in Kamary et. al. (2014), or

8. Abandon testing point-null hypothesis following the ideas in Rao and Lovric (2016), and many other statisticians.

The first two suggestions are not supportable, because they do not settle a dispute by mutual concession. The Jeffreys-Lindley paradox need not be feared, once it disengaged from applied statistics suggestions 3 through 7. This leaves suggestion 8 as the most tenable.

Acknowledgement

This research was supported by the Artis College Faculty Research Grant, Radford University, USA.

References

Arbuthnott, J. (1710). An Argument for Divine Providence, taken from the constant Regularity observ'd in the Births of both Sexes. *Philosophical Transactions of the Royal Society of London, 27*(328), 186–190. doi: 10.1098/rstl.1710.0011

Bartlett, M. S. (1957). A Comment on D. V. Lindley's Statistical Paradox. *Biometrika, 44*(3/4), 533-534. doi: 10.1093/biomet/44.3-4.533

Baskurt, Z, and Evans, M. (2013). Hypothesis Assessment and Inequalities for Bayes Factors and Relative Belief Ratios. *Bayesian Analysis, 8*(3), 569–590. doi: 10.1214/13-ba824

Bayarri, M. J and Berger, J. O. (2013). Hypothesis testing and model uncertainty. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens (Eds.), *Bayesian Theory and Applications* (pp. 484-498). Oxford, UK: Oxford University Press. doi: 10.1093/acprof:oso/9780199695607.003.0018

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. doi: 10.1038/s41562-017-0189-z

Berger, J. O. and Delampady, M. (1987). Testing Precise Hypotheses. *Statistical Science, 2*(3), 317-335. doi: 10.1214/ss/1177013238

Berger, J. O. and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association, 82*(397), 112-122. doi: 10.2307/2289131

Berger, J. O. (2006). Bayes Factors. In S. Kotz, N. Balakrishnan, C. B. Read, B. Vidakovic (Eds.), *Encyclopedia of Statistical Sciences* (2nd ed.)*, 1* (pp. 378-386). NY: John Wiley & Sons. doi: 10.1002/0471667196.ess0985.pub2

Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of American Statistical Association, 91*(433), 109-122. doi: 10.1080/01621459.1996.10476668

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association. 33*(203), 526-542. doi: 10.1080/01621459.1938.10502329

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds), *Bayesian Statistics* (pp. 605-647). Valencia: Valencia University Press.

Bernardo, J. M. (1999). Nested Hypothesis Testing: The Bayesian Reference Criterion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds), *Bayesian Statistics 6* (pp. 101-130). Oxford: Oxford University Press.

Bernardo, J. M. (2011a). Integrated objective Bayesian estimation and hypothesis testing. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics 9* (pp. 1-68). Oxford: Oxford: University Press. doi: 10.1093/acprof:oso/9780199694587.003.0001

Bernardo, J. M. (2011b). Modern Bayesian Inference: Foundations and Objective Method. In P. Bandyopadhyay and M. R. Forster (Eds.), *Philosophy of Statistics* (pp. 263–306). Amsterdam: Elsevier. doi: 10.1016/b978-0-444-51862-0.50008-3

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics* (2nd ed.), NY: Wiley Interscience.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association, 82*(397), 106-111. doi: 10.1080/01621459.1987.10478396

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121-126. doi: 10.1198/000313005x20871

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press. doi: 10.1201/9781439894798

Cousins, R. D. (2014). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese, 194*(2), 395–432. doi: 10.1007/s11229-014-0525-z

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press. doi: 10.1017/cbo9780511813559

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193-242. doi: 10.1037/h0044139

Etz, A. and Wagenmakers, E. J. (2017). J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test. *Statistical Science, 32*(2), 313-329. doi: 10.1214/16-sts599

Fienberg, S. E. (2006). When Did Bayesian Inference Become "Bayesian"? *Bayesian Analysis, 1*(1), 1–40. doi: 10.1214/06-ba101

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference* (3rd ed.). London: Collins Macmillan.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Gelman, A. and Shalizi, C. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8–38. doi: 10.1111/j.2044-8317.2011.02037.x

Gill, J. (2015). *Bayesian Methods: A Social and Behavioral Sciences Approach* (3rd ed.). Chapman & Hall/CRC.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(1), 107-114. doi: 10.1111/j.2517-6161.1952.tb00104.x

Good, I. J. (1958). Significance Tests in Parallel and in Series. *Journal of the American Statistical Association, 53*(284), 799-813. doi: 10.1080/01621459.1958.10501480

Good, I. J. (1965). A List of Properties of Bayes-Turing Factors. *NSA Technical Journal, 10*(2), 1-6.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. *Trabajos de Estadistica Y de Investigacion Operativa, 31*(1), 489-519. doi: 10.1007/bf02888365

Good, I. (1982). [Lindley's Paradox] Comment. *Journal of the American Statistical Association, 77*(378) 342-344. doi: 10.2307/2287248

Hald, A. (2003). *A History of Probability and Statistics and Their Applications before 1750*. NY: John Wiley & Sons.

Hodges, J. L. and Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B., 16*(2), 262–268. doi: 10.1111/j.2517-6161.1954.tb00169.x

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences.* NY: John Wiley & Sons. doi: 10.1002/9780470686621

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration, 4(*2), 153-169.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: The Clarendon Press.

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

Kadane, J. B. (1984). *Robustness of Bayesian analyses*. Amsterdam: North-Holland.

Kamary, K., Mengersen, K., Robert, C. P. and Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. arXiv e-print arXiv:1412.2044. Retrieved from https://arxiv.org/abs/1412.2044.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795. doi: 10.1080/01621459.1995.10476572

Kendall, M. G. (1968). On the Future of Statistics – A Second Look. *Journal of the Royal Statistical Society. Series A (General), 131*(2), 182-204. doi: 10.2307/2343841

Keuzenkamp, H. A. (2000). *Probability, Econometrics and Truth.* Cambridge University Press. doi: 10.1017/cbo9780511493300

Laan, M. and Rose, S. (2010, September 1). Statistics Ready for a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets. *Amstat News*, *399*, 38-39. Retrieved from https://magazine.amstat.org/blog/2010/09/01/statrevolution/

LaMont, C. H. and Wiggins P. A. (2015). Information-based inference for singular models and finite sample sizes: A frequentist information criterion. arXiv preprint arXiv:1506.05855v5. Retrieved from https://arxiv.org/abs/1506.05855

Li, Y., Zeng, T. and Yu, J. (2014). A new approach to Bayesian hypothesis testing. *Journal of Econometrics, 178*(Part 3), 602–612. doi: 10.1016/j.jeconom.2013.08.035

Lindley, D. V. (1957) A Statistical Paradox. *Biometrika, 44*(1/2),187-192. doi: 10.1093/biomet/44.1-2.187

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference*. Cambridge University Press. doi: 10.1017/cbo9780511662973

Lindley, D. V. (2009). [Harold Jeffreys's Theory of Probability Revisited]: Comment. *Statistical Science, 24*(2), 183-184. doi: 10.1214/09-sts284f

Migon, H. S., Gamerman, D. and Louzada, F. (2014). *Statistical Inference: An Integrated Approach* (2nd ed.). CRC Press. doi: 10.1201/b17229

Moreno, E. (2011). [Integrated objective Bayesian estimation and hypothesis testing]. Discussion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics 9* (pp. 40-42). Oxford: Oxford: University Press.

Naaman, M. (2016). Almost sure hypothesis testing and a resolution of the Jeffreys-Lindley paradox. *Electronic Journal of Statistics, 10*(1), 1526-1550. doi: 10.1214/16-ejs1146

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A. 231*(694–706), 289–337. doi: 10.1098/rsta.1933.0009

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 99-138. doi: 10.1111/j.2517-6161.1995.tb02017.x

Pericchi, L. (2011). [Integrated objective Bayesian estimation and hypothesis testing]. Discussion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics 9* (pp. 25-30). Oxford: Oxford: University Press.

Rao, C. R. and Lovric, M. M. (2016). Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective. *Journal of Modern Applied Statistical Methods, 15*(2), 2-21. doi: 10.22237/jmasm/1478001660

Robert, C. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica, 3*, 601–608.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). NY: Springer Verlag.

Robert, C. (2014). On the Jeffreys-Lindley Paradox. *Philosophy of Science, 81*(2), 216–232. doi: 10.1086/675729

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology, 72*, 33–37. doi: 10.1016/j.jmp.2015.08.002

Rousseau, J. and Robert, C. P. (2011). [On moment priors for Bayesian model choice]. Discussion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.), *Bayesian Statistics 9* (pp. 136-137). Oxford: Oxford: University Press.

Shafer, G. (1982). Lindley's Paradox. *Journal of the American Statistical Association, 77*, 325-351. doi: 10.1080/01621459.1982.10477809

Smith, A. (1995). A Conversation with Dennis Lindley. *Statistical Science, 10*(3), 305-319. doi: 10.1214/ss/1177009940

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science, 80*(1), 73–93. doi: 10.1086/668875

Spanos, A. (2014). Recurring controversies about P values and confidence intervals revisited. *Ecology*, *95* (3), 645–651. doi: 10.1890/13-1291.1

Sprenger, J. (2013). Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science, 80*(5), 733–744. doi: 10.1086/673730

Stigler, S. M. (1980). Stigler's Law of Eponymy. In T. F. Gieryn and R. K. Merton (Eds.), Science and Social Structure: A Festschrift for Robert K. Merton. *Transactions of the New York Academy of Sciences, 2*(39), 147–158. doi: 10.1111/j.2164-0947.1980.tb02775.x

Stigler, S. M. (1983). Who Discovered Bayes' Theorem? *American Statistician, 37*(4), 290–296. doi: 10.2307/2682766

Stigler, S. M. (2013). The True Title of Bayes's Essay. *Statistical Science, 28*(3), 283–288. doi: 10.1214/13-sts438

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2. doi: 10.1080/01973533.2015.1012991.

Villa, C. and Walker, S. (2017). On the Mathematics of the Jeffreys–Lindley Paradox. *Communications in Statistics - Theory and Methods, 46*(24), 12290-12298. doi: 10.1080/03610926.2017.1295073

Wasserstein, R. W., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". *The American Statistician, 73*(Sup1), 1-19, doi: 10.1080/00031305.2019.1583913

Welsh, A. H. (1996). *Aspects of Statistical Inference*. NY: John Wiley & Sons. doi: 10.1002/9781118165423

Wrinch, D. and Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine, 42*(249), 369-390. doi: 10.1080/14786442108633773

Zhu, Q., Shyu, M. and Chen, S (2012). Discriminative Learning-Assisted Video Semantic Concept Classification. In F. Y. Shih (Ed.), *Multimedia Security Watermarking, Steganography, and Forensics* (pp. 31-49). CRC Press.

Appendix A

This is the original code for the program that was used in this paper to compare number of occurrences of the Jeffreys-Lindley paradox in case the null hypothesis is true or false.

```
# This analysis is based on the normal conjugate model with a known variance
#
# Written by M. Lovric March 16, 2019
###############################################################################
rm(list = ls())            # clear memory
set.seed(12345)         # set the same seed for all comparisons
# BF_interpret function interprets values of Bayes factor according to the
# paper "Bayes factors", by Robert Kass & Adrian Raftery (1995),
#                         JASA, Vol. 90, No. 430. pp. 773-795.
#
# BF in this program is BF_0_1, not BF_1_0 hence the inverse values are taken
#
BF_interpret <- function(BF){
    if (1/BF > 150){
      res <- "Very strong evidence against Ho."
```

```
    }else if(1/BF > 20){
      res <-  "We have strong evidence against Ho."
    }else if(1/BF > 3){
      res <-  "We have positive evidence against Ho."
    }else if(1/BF > 1){
      res <-  "Not worth more than a bare mention evidence against Ho."
    }else{
      res <- "supports Ho."
    }
  }
# Set alpha level
alpha_level = 0.05  # Initial significance level
sigma_2 <- 9          # Known variance

######################################################################
# First initialize vectors for the comparison at 0.05 level
Z_vector <- c()                                  ## Vector that contains z-values
BF_Wrongly_Support_H_0 <- c()    ## Bayes factor wrongly supports false null hypothesis
Posterior_H_0_Wrong <- c()
p_value_Correct <- c()   ## vector of  p-value that correctly reject false null  hypothesis

# now initialize vectors for comparison at 0.04 level
Z_vector_004 <- c()
BF_Wrongly_Support_H_0_004 <- c()
Posterior_H_0_Wrong_004 <- c()
p_value_Correct_004 <- c()

# now initialize vectors for comparison at 0.03 level
Z_vector_003 <- c()
BF_Wrongly_Support_H_0_003 <- c()
Posterior_H_0_Wrong_003 <- c()
p_value_Correct_003 <- c()  ## p-value correctly rejects false null hypotheis

# now initialize vectors for comparison at 0.02 level
Z_vector_002 <- c()
BF_Wrongly_Support_H_0_002 <- c()
Posterior_H_0_Wrong_002 <- c()
p_value_Correct_002 <- c()
```

```
# now initialize vectors for comparison at 0.01 level
Z_vector_001 <- c()
BF_Wrongly_Support_H_0_001 <- c()
Posterior_H_0_Wrong_001 <- c()
p_value_Correct_001 <- c()

# Finally, initialize vectors for comparison at 0.005 level recommended by Benjamin et. al.
Z_vector_0005 <- c()
BF_Wrongly_Support_H_0_0005 <- c()    ## Bayes factor wrongly supports false null hypothesis
Posterior_H_0_Wrong_0005 <- c()
p_value_Correct_0005 <- c()
# priors for H_0 and H_1
pi_H_0 <- 0.5   #  assign “objective” priors for the point null and alternative
pi_H_1 <- 1 - pi_H_0

n= 250000000  # sample size

N = 1000  # number of iterations (“trials”)

Theta_0 = 170  # hypothesized value of the parameter (mean)

######### a loop for a false or true null hypothesis
for (k in 1:N){
        NormalRandomSample <- rnorm(n, mean = 170, sd = sqrt(9)) # take a random sample of size
n from a normal
        x_bar = mean(NormalRandomSample)
        Z <- (sqrt(n)*(x_bar - Theta_0))/sqrt(sigma_2) ;  # calculate the value of the Z test
statistic
        p_value <- 2*pnorm(-abs(Z)) ;
        BF <- (1 + n)^(1/2) *exp((-0.5*Z^2)*n/(n+1))   ;  # calculate Bayes factor for a normal
conjugate prior case
        # next find the posterior probability of the null hypothesis

        Post_H0_two_sided <- (1 + ((1 - pi_H_0)/pi_H_0)*(1 + n)^(-1/2)*
                                exp((0.5*Z^2)*n/(n+1)))^(-1);
        print(k)  # print the number of iterations on the screen
```

```
        if (Post_H0_two_sided > 0.95  & p_value < 0.05)  {     # Impose a condition for strict
JLP, when a comparison is made at 0.05 level of significance
        print ("F O U N D   JLP")  # Print on the screen when a JLP is found
        BF_Wrongly_Support_H_0 <- c(BF_Wrongly_Support_H_0, BF);
        Posterior_H_0_Wrong <- c(Posterior_H_0_Wrong,
                              Post_H0_two_sided);
        p_value_Correct <- c(p_value_Correct, p_value);
        Z_vector <- c(Z_vector, Z)
        }
        if (Post_H0_two_sided > 0.96  & p_value < 0.04)  { # the same for 0.04 significance
          print ("F O U N D   JLP at alpha = 0.04")
          BF_Wrongly_Support_H_0_004 <- c(BF_Wrongly_Support_H_0_004, BF);
          Posterior_H_0_Wrong_004 <- c(Posterior_H_0_Wrong_004,
                                    Post_H0_two_sided);
          p_value_Correct_004 <- c(p_value_Correct_004, p_value);
          Z_vector_004 <- c(Z_vector_004, Z)
        }
        if (Post_H0_two_sided > 0.97  & p_value < 0.03)  { #  the same for 0.03 significance
          print ("F O U N D   an occurrence of JLP! At alpha = 0.03")
          BF_Wrongly_Support_H_0_003 <- c(BF_Wrongly_Support_H_0_003, BF);
          Posterior_H_0_Wrong_003 <- c(Posterior_H_0_Wrong_003,
                                    Post_H0_two_sided);
          p_value_Correct_003 <- c(p_value_Correct_003, p_value);
          Z_vector_003 <- c(Z_vector_003, Z)
        }
        if (Post_H0_two_sided > 0.98  & p_value < 0.02)  { #  the same for 0.02 significance
          print ("F O U N D   JLP")
          BF_Wrongly_Support_H_0_002 <- c(BF_Wrongly_Support_H_0_002, BF);
          Posterior_H_0_Wrong_002 <- c(Posterior_H_0_Wrong_002,
                                    Post_H0_two_sided);
          p_value_Correct_002 <- c(p_value_Correct_002, p_value);
          Z_vector_002 <- c(Z_vector_002, Z)
        }
        if (Post_H0_two_sided > 0.99  & p_value < 0.01)  { #  the same for 0.01 significance
          print ("F O U N D   JLP for alpha = 0.01")
          BF_Wrongly_Support_H_0_001 <- c(BF_Wrongly_Support_H_0_001, BF);
          Posterior_H_0_Wrong_001 <- c(Posterior_H_0_Wrong_001,
                                    Post_H0_two_sided);
```

```
          p_value_Correct_001 <- c(p_value_Correct_001, p_value);
          Z_vector_001 <- c(Z_vector_001, Z)
        }
        if (Post_H0_two_sided > 0.995  & p_value < 0.005)  { #  the same for 0.005 significance
          print ("F O U N D   JLP at 0.005 level")
          BF_Wrongly_Support_H_0_0005 <- c(BF_Wrongly_Support_H_0_0005, BF);
          Posterior_H_0_Wrong_0005 <- c(Posterior_H_0_Wrong_0005,
                                  Post_H0_two_sided);
          p_value_Correct_0005 <- c(p_value_Correct_0005, p_value);
          Z_vector_0005 <- c(Z_vector_0005, Z)
        }
    }
# Finally, store simulation results in data frame objects
Output005 <- data.frame(BF_Wrongly_Support_H_0, Posterior_H_0_Wrong, Z_vector, p_value_Correct)
edit(Output005)
length(Output005$BF_Wrongly_Support_H_0)

Output004   <-   data.frame(BF_Wrongly_Support_H_0_004,   Posterior_H_0_Wrong_004,   Z_vector_004,
p_value_Correct_004)
edit(Output004)
length(Output004$BF_Wrongly_Support_H_0_004)

Output003   <-   data.frame(BF_Wrongly_Support_H_0_003,   Posterior_H_0_Wrong_003,   Z_vector_003,
p_value_Correct_003)
edit(Output003)
length(Output003$BF_Wrongly_Support_H_0_003)

Output002   <-   data.frame(BF_Wrongly_Support_H_0_002,   Posterior_H_0_Wrong_002,   Z_vector_002,
p_value_Correct_002)
edit(Output002)
length(Output002$BF_Wrongly_Support_H_0_002)

Output001   <-   data.frame(BF_Wrongly_Support_H_0_001,   Posterior_H_0_Wrong_001,   Z_vector_001,
p_value_Correct_001)
edit(Output001)
length(Output001$BF_Wrongly_Support_H_0_001)
```

```
Output0005  <-data.frame(BF_Wrongly_Support_H_0_0005,  Posterior_H_0_Wrong_0005,  Z_vector_0005,
p_value_Correct_0005)
edit(Output0005)
length(Output0005$BF_Wrongly_Support_H_0_0005)d type II conflicts:", length(BF_Critical)/5)
```