# Appendix: the Jeffreys-Lindley paradox and its relevance to statistical testing

Frank Lad*

## 1 Introduction

The purpose of this technical appendix is to present a formal statement of the Jeffreys-Lindley paradox, which I alluded to in the main text of this lecture in Section 7.1, and to explain subsequent developments in Bayesian statistical methodology that resolve the paradox. The resolution requires a change in statistical methodology from the long-standing but misleading practice of reporting the "significance" of a test-statistic that is meant to test the validity of a statistical hypothesis.

This appendix is not meant to be a primer on subjectivist statistical methods, much less a serious introduction. In the main text of this lecture I made reference to the article of Lad (2003) and the book of Lad (1996) that provide materials for those purposes. Nonetheless, to make this appendix self-contained, I shall begin Section 2 with a brief statement of Bayes' Theorem as it is formulated in discrete and continuous contexts. This shall include a simple example of its application that can be understood only through a subjective interpretation of probability. Section 3 shall present a formal statement and proof of the Jeffreys-Lindley paradox. Developments of statistical literature that have proceeded since the paradox was first discussed are described briefly in Section 4. Finally, in Section 5 I shall describe a detailed application of new methods of presenting evidence that is relevant to a practical question of interest.

## 2 Bayes' Theorem and its application

In pondering an observable feature of Nature, a scientist eventually surmises a limited array of hypotheses regarding its development. Suppose we list them and denote them by the symbols $H_1, H_2, ..., H_N$. Presumably, each of these hypotheses has some degree of credibility to its proponent, or else it would not merit listing. Since the list is meant to be exhaustive of the ideas that we entertain, a scientist's total belief in the reasons for the phenomena would be distributed among these N hypotheses. Norming the total belief to a unit of 1, relative beliefs in these hypotheses are expressed by the scientist's probabilities $P(H_1), P(H_2), ..., P(H_N)$, which sum to 1.

Empirical science is oriented towards collecting new experimental or historical observations that can inform us about the truth of these hypotheses. Suppose we denote the numerical data that summarises the outcome of an experiment by a letter D. Bayes' Theorem provides a logical mathematical framework for calibrating how the observation of D should be used to inform us about a new understanding of the relative probabilities of the hypotheses given this new data. Technically, it yields numerical values for the conditional probabilities

$$P(H_1|D), P(H_2|D), ..., P(H_N|D)$$

[*] Frank Lad is a consulting statistician and a research associate in the Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand/Aotearoa. This is an appendix to an invited lecture presented at the Conference on Science and Democracy 2, Palazzo Serra di Cassano, Napoli, 12-14 June, 2003.

that are based on their initial information sources *augmented by the new observed data*, D. The numerical differences between the probability distribution $P(H_1), P(H_2), ..., P(H_N)$ and the conditional distribution $P(H_1|D), P(H_2|D), ..., P(H_N|D)$ display what one can rightfully learn about our assessment of the hypotheses from this data. This assessment, in the long and the short of statistical theory, is the goal of scientific experimentation and data analysis. Because these distributions can represent uncertain opinions before and after observing the data, D, the former is commonly called a "prior" distribution for the hypotheses, while the latter is called the "posterior" distribution.

The formulation of Bayes' theorem requires that we assess the probability of observing the experimental data result that we do, given each of the relevant hypotheses. These assessments are represented by the numbers

$$P(D|H_1), P(D|H_2), ..., P(D|H_N) \quad .$$

The relative sizes of these numbers for the various possible hypotheses based on the same data are commonly referred to as the "likelihoods" for the hypotheses.

Using these probabilities for the observed data on the basis of each of the tendered hypotheses, Bayes' theorem allows the computation of each of the posterior probabilities $P(H_1|D)$, $P(H_2|D), \ldots, P(H_N|D)$ according to the computational formula

$$P(H_i|D) \quad = \quad \frac{P(D|H_i)P(H_i)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2) + \ldots + P(D|H_N)P(H_N)} \tag{1}$$

for each hypothesis $H_i$ .

## 2.1 Uniquely subjective applications of Bayes' Theorem

Bayes' theorem is commonly associated with a subjective interpretation of probability because there are many important practical examples of its use in contexts that defy a non-subjective interpretation. Consider briefly the situation of statistical methods used to audit bookkeeping records. The financial records of a business for a fiscal year are presented to an auditor. These "books" contain some number of errors. That is a fact of completed history. The auditor would like to know what the value of this number of errors is. Based on the auditor's limited knowledge of this company and knowledge of other companies, the auditor is uncertain about the value of this number, and may express a personally assessed probability distribution for its value. There can be nothing "objective" about this distribution, nor can it have a physical interpretation, because the fact of the matter is already settled. There exists some specific number of errors in the books. The auditor merely does not know what the value is. In order to reduce the amount of uncertainty about this number, a sample of transactions may be selected for examination. The likelihood of errors in the sample depends on the unknown number of errors in the books. The likelihood probabilities (conditional on the various possibilities for this unknown number) can be assessed routinely in a way that would be agreed by knowledgeable auditors. Conditional on the results of the audit of the sample, Bayes' theorem can be used to compute a posterior distribution for the unknown number of errors conditioned on this data. This is despite the fact that there is no objective interpretation for what such computations could mean.

## 2.2 Bayes' theorem in the context of a continuum

In many statistical contexts, the list of hypotheses is represented as a continuum of possibilities, typically denoted by the Greek letter $\theta$, which is presumed to lie within some interval, denoted by $\Theta$. In such a context, prior opinions are representable as a probability density over $\theta$, viz., $f(\theta)$. Statistical data is symbolised by a vector of possible observations, viz., $\mathbf{X}_n$ observed to equal

$\mathbf{x}_n = (x_1, x_2, ..., x_n)$. The likelihood of the hypotheses relative to this data is then representable via the likelihood function $f(\mathbf{x}_n|\theta)$ for the various possible values of $\theta$ in the possible continuum. In this context, Bayes' theorem is represented as saying

$$f(\theta|\mathbf{x}_n) \;=\; \frac{f(\mathbf{x}_n|\theta)\ f(\theta)}{\int_\Theta\ f(\mathbf{x}_n|\theta)\ f(\theta)\ d\theta} \quad .$$

Finally, in the context of hypothesis testing, some particular value of $\theta$ within the continuum is recognised as a particularly appropriate possibility that merits probability mass on its own. Suppose that value is denoted by $\theta_0$, and is accorded with the probability $P(\theta = \theta_0) = p$. When the remaining probability is distributed over the rest of the domain $\Theta$ via a density $f(\theta)$, the resulting posterior probability for $\theta = \theta_0$ would be computed via the representation

$$P(\theta = \theta_0|\mathbf{x}_n) \;=\; \frac{f(\mathbf{x}_n|\theta_0)\ p}{f(\mathbf{x}_n|\theta_0)\ p\ +\ (1{-}p)\ \int_\Theta\ f(\mathbf{x}_n|\theta)\ f(\theta)\ d\theta} \quad .$$

This is the context that is relevant to the formal statement of the Jeffreys-Lindley paradox.

## 3 The original formulation of the paradox

In a brief article, Lindley (1957) stated and proved the following result, using language that is still widely used by most statisticians today:

**Result:** Suppose $X_1, X_2, ...., X_n, ...$ compose a sequence of independent and identically distributed normal random variables with the distribution $X_i|\theta \sim N(\theta, \sigma^2)$. Here $\theta$ is an unknown location parameter of interest, and $\sigma^2$ is a known constant representing the variance of the random measurements about the unknown $\theta$. Prior opinion about $\theta$ is a mixture of a point mass $p$ at a specified value of $\theta$, call it $\theta_0$, and the remaining weight $1 - p$ distributed uniformly over an interval of width $I$, centered at $\theta_0$. The well-known critical values of the sample mean, $\bar{X}$, based on $n$ observations that would imply the rejection of a null hypothesis $H_0 : \theta = \theta_0$ in favour of the two sided alternative $H_A : \theta \neq \theta_0$, at significance level $\alpha = P(\text{type I error})$, are $\bar{X} = \theta_0 \pm (\sigma/\sqrt{n})z_{\alpha/2}$. Here $z_{\alpha/2}$ denotes the ordinate demarking the $\alpha/2$ tail area for a standard normal density. The posterior probability that the null-hypothesis is true conditional upon the observation of data that achieves this significance level approaches 1 as the sample size, $n$, becomes arbitrarily large.

**Proof** :

$$\begin{aligned}
P \quad (H_0|\bar{X} = \bar{x} = \theta_0 + (\sigma/\sqrt{n})z_{\alpha/2}) \quad &= \quad \frac{P(H_0)P(data|H_0)}{P(H_0)P(data|H_0)) + P(H_A)P(data|H_A)} \\
&= \frac{p\ \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\ e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta_0)]^2\}}}{p\ \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\ e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta_0)]^2\}} + (1-p)\int_{\theta_0-I/2}^{\theta_0+I/2} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\ e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta)]^2\}}\ \frac{1}{I}\ d\theta} \\
&= \frac{p\ e^{\{-(1/2)z^2_{\alpha/2}\}}}{p\ e^{\{-(1/2)z^2_{\alpha/2}\}} + \frac{(1-p)}{I}\int_{\theta_0-I/2}^{\theta_0+I/2} e^{\{(-1/2)[(\sqrt{n}/\sigma)(\theta-\bar{x})]^2\}} d\theta} \\
&\geq \frac{p\ e^{\{-(1/2)z^2_{\alpha/2}\}}}{p\ e^{\{-(1/2)z^2_{\alpha/2}\}} + \frac{(1-p)}{I}\frac{\sqrt{2\pi}\sigma}{\sqrt{n}}} \quad \to \quad 1 \quad as \quad n \quad \to \quad \infty
\end{aligned}$$

The second line merely fills in details appropriate to the components of Bayes' Theorem. In the third line, the common factor $\frac{\sqrt{n}}{\sqrt{2\pi}\sigma}$ is cancelled in all terms of the numerator and denominator, and the conditioning value of $\bar{x}$ is substituted in the numerator and first term of the denominator. In the fourth line the integral is replaced by the larger value of that integral if its domain of integration were the entire real line, which would be the proportionality constant appropriate to a normal density for $\theta$ centered at the conditioning value of $\bar{x}$ with variance $\sigma^2/n$.

The limiting result follows simply since $\sqrt{n}$ appears only in the second term of the denominator of the final line.

Thus arises the paradox, that upon the observation of data that the long-standard significance testing procedures would pronounce as yielding a result that is "significantly different" from a prescribed value at level $\alpha$ would yield a posterior level of belief in the prescribed value that can be close to certainty.

The paradox founders on the fact that even according to the principles of classical hypothesis testing, the choice of an appropriate significance level for a test (the size of type I error that would be allowed) should depend on the amount of type II error that must be incurred with it. To presume that a specific level of significance has the same meaning no matter what is the sample size, i.e. no matter what level of type II error is incurred, is a fundamental error that has riddled the practice of so-called objective statistical methods for years, even to the present day.

## 4 Subsequent developments

In the era when his article was written, Lindley's discussion of the result was largely oriented toward motivating the method of analysis that lies behind it, explaining the assessment of the prior distribution and its use in Bayes' Theorem. During the course of the following thirty years these arguments supporting the practice of Bayesian statistics became persuasive enough that a sizeable number of statisticians began to think in this way. Nonetheless, conventional statistical practice had come to enshrine $P$(type I error) with such unique importance that the reporting of a p-value associated with a test statistic became the almost universal form of reporting test results. It was even required by many applied scientific, medical and business journals. The p-value of a test is the smallest value of $\alpha$ at which an observed test statistic would specify rejection of the null hypothesis. The weight of Lindley's paradoxical result, along with related analyses by Jeffreys (1948), Good (1958) and Edwards, Lindman and Savage (1963) began to burden proponents of the Bayesian movement who wished to reorient statistical practice to sound principles of inferential evidence.

In 1987, both the *Journal of the American Statistical Association* and a journal of the Institute of Mathematical Statistics, *Statistical Science*, published extensive readable papers by Berger and Sellke (1987) and Berger and Dalampady (1987) showing that the standard practice of publishing "p-values" typically misconstrues the evidence contained in the data that is relevant to tested hypotheses. This occurs for much the same reason that the Jeffreys-Lindley paradox arises, the ignoring of type II error. These articles appeared with discussion by several of the most eminent theoretical and applied statisticians of the day.

Nonetheless, the mass marketing of statistical practice in required university courses, which promotes the idea of supposed scientific "objectivity" that does not rely on scientists' intuitions, beliefs or values, has chosen to ignore such results. As one example from many, the popular text of Moore and McCabe (2003) uses 828 pages of text supporting the methodology of p-values without even mentioning arguments surrounding the distortion it provokes. Many applied journals still require the presentation of statistical evidence in the form of p-values.

Rather than repeat the readable discussions that can be found in the international journals mentioned, I will conclude with a simple practical example that displays the improved style of analysis and reporting motivated by subjective Bayesian methods.

## 5 Posterior probability of differences of a meaningful size

To conclude, I shall report briefly on a computation by Deely (2003) appearing in a readable article that is completely devoted to the exposure and explanation of the methodology that I summarise here. The computation pertains to data reported by Ford et al. (1988) that measures blood cholesterol levels in 119 black American men and 130 black American women. The particular measurements that concern us here are made on the presence of high density lipoproteins (HDL cholesterol) in a subject, measured in milligrams per deciliter of blood. As it happens, HDL cholesterol is considered to be a beneficial component of blood cholesterol, whereas low density lipoproteins are considered a risk factor for several medical problems. Among other features of their article, Ford et al. reported the p-value statistic relevant to testing a null hyothesis that mean HDL levels in male subjects is equal to the mean levels in female subjects. The published data are summarised in Table 1.

**Table 1: HDL Statistics for Black Male and Female Subjects**

| | men | women | p-value |
|---|---|---|---|
| mean | 55.2 | 59.2 | .0621 |
| stdev | 16.5 | 17.2 | |
| n | 116 | 130 | |

Standard practice would be to conclude that "the results are not significant because the p-value is not as low as .05." If it were as small as .01 the difference would be reported as "highly significant," while a p-value of .001 or smaller would be termed "extremely significant". Experience with the practical consequences of the Jeffreys-Lindley paradox, that even extreme "statistical significance" can be achieved on the basis of sample mean statistics that are virtually identical, has provoked a concern with "practical significance" as an issue that is distinct from so-called "statistical significance" which is based merely on assessment of type I error.

The Bayesian alternative to assessing the content of these same statistics is laid out simply and in detail in the article of Deely. For this example it is developed as follows. Since the target populations of black adults from whom the samples were taken were sharply defined, the measurements among males and among females are widely regarded exchangeably within each group. More explicitly, measurements within each group are assessed with a mixture normal distribution, with conceivably different means for each. Denoting these means by $\mu_M$ and $\mu_F$, what is of interest is to compute the posterior probability $P(\mu_M \geq \mu_F + b \mid \text{data on males and females})$. Interest focuses directly on various possible values of the difference "b", denominated in units of mg/dl, by which male mean HDL might exceed female mean HDL measurements.

The computation of this posterior probability relies on a specification of an assessed prior probability distribution for this difference. For the situation under consideration, prior to measuring the data for males and females there would have been reason to suspect that HDL levels among males might exceed, equal or be lower than those of females. Moreover, if differences are actually to be found, there was little information available on how large they might be. To some extent, differences of virtually any size could be believable. In such a situation, a common practice is to assess the difference with an improper, uniform distribution over all possible values. Though unrealistic as an assessment of prior opinion relative to extreme differences, such a prior distribution is a benchmark with which computations based on more informative prior distributions can be and are compared. In fact, the computational results are virtually identical to those using the exact finite interval uniform prior specified in Lindley's result, with I = 50 or even 100. For the purpose of this Appendix, I shall report posterior probabilities only for this prior distribution. Deely also discusses and reports more details based on a more extensive assessment of prior information.

Table 2 presents the relevant posterior probabilities based on this "flat" prior distribution and the same data for which the p-value was reported in Table 1. I shall comment on the results below.

**Table 2: Posterior Probabilities for Differences in Mean HDL Measurements among Black Male and Female Subjects (to 3 decimal places)**

| $b$ | $P(\mu_M \geq \mu_F + b \mid \text{data on males and females})$ |
|---|---|
| 0 | .909 |
| 1 | .748 |
| 2 | .500 |
| 3 | .252 |
| 4 | .091 |
| 5 | .024 |
| 6 | .004 |
| 7 | .000 |

Comparing Tables 1 and 2, it is worthy of comment firstly that whereas the "insignificant p-value" of .0621 is purportedly supposed to suggest that there is no difference between mean HDL in the two populations, the first row of Table 2 specifies a posterior probability that $\mu_M$ exceeds $\mu_F$, based on the same data, with the value .909. Contrary to the conclusion from the p-value, it is highly probable that mean male HDL levels exceed those of females. This comparison highlights the regular misinterpretation of comparatively high p-values (those above .05 for example) as supporting a high probability for no difference between the two populations.

A second comment is to notice that the array of posterior probabilities that $\mu_M$ exceeds $\mu_F$ by at least the amount "b", for various values of "b" expresses precisely the type of information that common practice would require to distinguish between "statistical significance" and "practical significance". The results in Table 2 show that whereas mean HDL levels for men are probably (posterior to this data) greater than for women, they are also probably not greater by an amount that is large enough to worry about. Based on mean levels of both groups on the order of some 57 mg/dl or so, a difference in mean of 5 or 6 units would amount to a difference of 10% between the two groups. But Table 2 identifies a posterior probability for a difference of this great being less than .024.

The substance of this Appendix is to show explicitly how and why the Jeffreys-Lindley paradox arises, to show why the presentation merely of p-values is so misleading, and to show that the subjective Bayesian statistical methodology derives the information decision makers really desire in a natural way.

## References

**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**, 112-122. Discussions by J.W. Pratt, I.J. Good, J.M. Dickey, S.B. Vardeman, C.W. Morris in the same issue, pp. 123-139.

**Berger, J.O. and Delampady, M.** (1987) Testing precise hypotheses, *Statistical Science*, **2**, 317-335. Discussions by D.R. Cox, M.L. Eaton, Arnold Zellner, M.J. Bayarri, J. Casella and R. Berger in the same issue, pp. 335-352.

**Deely, J.** (2003) Comparing two groups or treatments - a Bayesian approach, in *Applied Bayesian Statistical Studies in Biology and Medicine*, Di Bacco, M., D'Amore, G., and Scalfari, F. (eds.) Boston: Kluwer Academic Publishers.

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

**Ford, E., Cooper, R., Simmons, B., Katz, S. and Patel, R.** Sex differences in high density lipoprotein cholesterol in urban blacks, *American Journal of Epidemiology*, **127**, 753-761.

**Good, I.J.** (1958) Significance tests in parallel and in series, *Journal of the American Statistical Association*, **53**, 799-813.

**Jeffreys, H.** (1948) *Theory of Probability*, Second edition, Oxford: Clarendon Press.

**Lad, F.** (1996) *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley.

**Lad, F.** (2003) An example of the subjectivist statistical method for learning from data: Why do whales strand when they do? *Applied Bayesian Statistical Studies in Biology and Medicine*, edited by M. DiBacco and G. D'Amore and F. Scalfari, Boston: Kluwer, 2003, Chapter 9: 153-188.

**Lindley, D.V.** (1957) A statistical paradox, *Biometrika*, **44**, 187-192.

**Moore, D.S. and McCabe, G.P.** (2003) *Introduction to the Practice of Statistics*, Fourth Edition, New York: W.H. Freeman and Co., 828 pp.