Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 119
DOI: 10.11159/icsta22.119

Relative Belief and Combining Evidence

Michael Evans
Dept. of Statistical Sciences
University of Toronto
Toronto, Canada
mevans@utstat.utoronto.ca

Abstract The problem of combining the evidence in several Bayesian inference bases is considered. Evidence is measured in each inference base using the relative belief ratio which gives an unambiguous prescription of whether there is evidence in favour of or against each possible value of an unknown such as a parameter. While there are many possible ways to combine the evidence, the method of linear pooling stands out as it preserves a consensus while others may not. There are constraints on this application, however, if one requires a formal Bayesian justification. In some applications where these restrictions do not hold, the approach can be generalized by allowing for the methodology known as Jeffrey conditionalization.

***Keywords*:** combining priors, statistical evidence, preserving consensus, Jeffrey conditionalization, ancillarity.

1 Introduction

It is common to see phrases like "the evidence in the data suggests …." or "based upon the evidence in the data …" as part of a statistical analysis. This is in spite of the fact that many approaches to formulating statistical theory do not explicitly provide a definition of how evidence is to be measured. For example, a p-value is often considered to be a measure of the evidence against a hypothesis but there are many reasons to suspect that these quantities do not measure evidence appropriately. Indeed p-values do not provide evidence in favour of a hypothesis and, associated with this, is the fact that there is no clear cut-off to even determine when a specific value is to be considered as evidence against.

Given that one can see the purpose of a statistical analysis as providing the evidence in data concerning the answers to questions of interest, this lack of a clear treatment of this concept can be viewed as a defect. There is some recognition of this in the statistical literature, for example, see [1]-[6]. There is in fact a much more extensive discussion of evidence in the philosophy of science literature with [7] and [8] providing representative examples.

In the paper the approach taken to measuring statistical evidence is as described in [5] and the associated issue of how, when there are multiples measures of evidence, these should be combined to provide a single overall measure is discussed. In Section 2 we formulate the basic problem. In Section 3 we discuss the possible solutions to the problem and focus on one of these. In Section 4 conclusions are drawn and further work is discussed. The technical details are not developed here but can be found in [9] on which much of this paper is based.

2 The Problem

It is necessary to first prescribe the ingredients of the statistical problem from which the measure of evidence is to be constructed. Of course, there is the data x which will be assumed to have been collected correctly. In addition, there are two basic components that a statistician needs to choose.

The Model: $M = \{f_\theta : \theta \in \Theta\}$ a collection of conditional probability distributions for the observed data x given θ such that the object of inferential interest is given by $\psi = \Psi(\theta)$, where $\Psi : \Theta \to \Psi$ is onto and to save notation the function and its range have the same symbol.

The Prior: π a probability distribution on Θ.

Both the model and the prior are subjective ingredients in a statistical analysis and so it is necessary to check these via the data to ensure that these are not contradicted by the objective part of the problem, namely, the data. It will be assumed hereafter that this has been done and both ingredients have passed their checks and we refer the reader to [5] as to how model checking and checking for prior-data conflict can be carried out. Note that this doesn't mean that these ingredients are now viewed as being correct, only that they have not been contradicted by the data.

There are two basic problems of inference that any theory of statistics needs to address.

(i) **Estimation**: construct an estimate of the true value of $\psi = \Psi(\theta)$ together with an assessment of the accuracy of the estimate.
(ii) **Hypothesis Assessment**: determine whether there is evidence in favour of or against the hypothesis $H_0: \Psi(\theta) = \psi_0$ together with an assessment of the strength of this evidence.

The solutions to these problems will be determined by the rules of inference adopted and these rules use the components of what will be called here the inference base, namely, $I = (x, M, \pi)$ comprised of the data, model and prior. When interest is in inference about ψ the relevant inference base is given by $I = (x, M_\Psi, \pi_\Psi)$ where $M_\Psi = \{m_\psi : \psi \in \Psi\}$,

$$m_\psi(x) = \int_\Theta f_\theta(x)\pi(\theta \mid \psi)\, d\theta$$

and $\pi(\cdot \mid \psi)$ is the prior on the nuisance parameters. So, the nuisance parameters have been integrated out to leave a model for the data that is now indexed by the parameter of interest ψ.

Given that there is now a joint probability measure P on (ψ, x), namely, $\pi_\Psi(\psi)m_\psi(x)$, the rules of inference can be stated in the context of a probability model $(\Omega, \mathcal{B}, P)$. Suppose then interest is in whether or not a hidden value $\omega \in \Omega$ is in the event $A \in \mathcal{B}$ and it is observed that $\omega \in C \in \mathcal{B}$. There are three rules employed.

R1 Principle *of conditional probability*: belief about the truth of A, as expressed by $P(A)$, is replaced by $P(A \mid C)$.

R2 Principle *of evidence*: the observation of C is evidence in favour of A when $P(A \mid C) > P(A)$, is evidence against A when $P(A \mid C) < P(A)$ and is evidence neither for nor against A when $P(A \mid C) = P(A)$.

R3 Principle *of relative belief*: when like alternatives $A \in \{A_1, A_2, \ldots\}$ are compared, evidence is ordered quantitatively by the relative belief ratio $RB(A \mid C) = P(A \mid C)/P(A)$.

There is then evidence in favour of (against) A being true when $RB(A \mid C) > (<) 1$ and no evidence either way when $RB(A \mid C) = 1$. In the context where probability measures are given by continuous densities consider a sequence of neighbourhoods A_ε of say ψ which converge nicely to ψ so that

$$\mathrm{RB}_\Psi(\psi \mid x) = lim_{\varepsilon \downarrow 0} RB(A_\varepsilon \mid x) = \pi_\Psi(\psi \mid x)/\pi_\Psi(\psi)$$

holds under weak conditions (continuity and positivity of the prior π_Ψ at ψ) where $\pi_\Psi(\cdot \mid x)$ denotes the posterior density of ψ. The answer to (i) is then given by the estimate $\psi(x) = arg\ sup_\psi\ RB_\Psi(\psi \mid x)$ and the accuracy of the estimate assessed by computing the plausible region $PL_\Psi(x) = \{\psi : RB_\Psi(\psi \mid x) > 1\}$, the set of all values having evidence in their favour, and its posterior content. So, if $Pl_\Psi(x)$ is small with high posterior content then $\psi(x)$ is considered an accurate estimate. For (ii) it is immediate that $RB_\Psi(\psi_0 \mid x)$ indicates whether there is evidence in favor of or against H_0. The strength of this evidence can be assessed by computing the posterior probability $\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x)$. If $RB_\Psi(\psi_0 \mid x) > 1$, then there is strong evidence in favour when this probability is big and if $RB_\Psi(\psi_0 \mid x) < 1$, then there is strong evidence against when this probability is small.

The basic problem of combining evidence can now be stated and this is done for two contexts.

Context I. Suppose there is a single statistical model M for the data x and k distinct priors π_i so there are k inference bases $I_i = (x, M, \pi_i)$ for $i = 1, \ldots, k$. It is assumed that the conditional priors $\pi_i(\cdot \mid \psi)$ on the nuisance parameters are all the same, as is satisfied when $\Psi(\theta) = \theta$. This situation arises when there is a group of analysts who agree on M and perhaps use a default prior for the nuisance parameters, while each member puts forward a prior for Ψ.

Context II. Suppose there are k data sets, models, and priors as given by the inference bases $I_i = (x_i, M_i, \pi_i)$ for $i \in \{1, \ldots, k\}$ and there is a common characteristic of interest $\psi = \Psi(\theta_i)$ with the true value of ψ being the same for each model, as will occur when ψ corresponds to some real-world quantity.

Note that, since ψ references some real-world quantity, in Context II the set of possible values and its true value is the same for each model even though formally the function Ψ may differ between models but we suppress this in the notation.

The simplest step away from Context I is when the data sets differ but all the models are based on the same basic set of candidates for the true probability measure and with the same conditional prior on the nuisance parameters. In such a context it seems obviously correct to simply combine the data sets and use the common model for the combined data set which places the problem within Context I. The result would not necessarily be the same if the data sets were not combined, so it is necessary that such a rule be applied first to the set of inference bases in general.

Context I is a major simplification over Context II. Context I has been considered previously but from the point-of-view of combining priors as, for example, in [10]-[19]. The approach here is different in that interest is in combining the evidence measures, as given by the relevant relative belief ratios, rather than priors as the proper expression of the evidence is the goal of a statistical analysis. A solution for Context I is obtained based upon it having desirable properties. While a general solution for Context II is not currently available, it will be argued that there is a natural generalization of the solution for Context I that can be applied to many of these problems

3 Combining Evidence

The rules for combining evidence are stated here initially when interest is in θ. The proofs of all statements in this section can be found in [9]. Let $\alpha = (\alpha_1, \ldots, \alpha_k) \in S_k$ the $(k-1)$-dimensional simplex for some $k \geq 2$ and, for now, suppose that α is given. While general combination rules could be considered, attention is restricted here to the power means of densities

$$\pi_{t,\alpha}(\theta) \propto \left(\sum_{i=1}^{k} \alpha_i \, \pi_i^t(\theta) \right)^{1/t}$$

for $t \in \mathbb{R}$. These priors are all proper provided $t \leq 1$, but otherwise this needs to be checked.

These rules for combining priors lead to rules for combining the individual relative belief ratios (see 1 below). From many points of view *linear pooling*, which corresponds to $t = 1$, represents the most logical way to combine the evidence measures in Context I and the properties justifying this conclusion are summarized below.

1. A combination rule for the priors immediately leads to a combination rule for the relative belief ratios $RB_i(\theta \mid x) = \pi_i(\theta \mid x)/\pi_i(\theta)$, namely, $RB_{t,\alpha}(\theta \mid x) = \pi_{t,\alpha}(\theta \mid x)/\pi_{t,\alpha}(\theta)$ using the posterior and prior based on the t-th combination rule. It follows that, with m_i denoting the i-th prior predictive based on the model M and prior π_i and $m_{t,\alpha}$ denoting the prior predictive based on the prior $\pi_{t,\alpha}$,

$$RB_{t,\alpha}(\theta \mid x) = \frac{m_{1,\alpha}(x)}{m_{t,\alpha}(x)} RB_{1,\alpha}(\theta \mid x) = \frac{m_{1,\alpha}(x)}{m_{t,\alpha}(x)} \sum_{i=1}^{k} \frac{\alpha_i m_i(x)}{m_{1,\alpha}(x)} RB_{i,\Psi}(\psi \mid x).$$

So, the t-th combination rule is a constant times the linear pooling rule. This implies that

$$arg\ sup_\theta RB_{t,\alpha}(\theta \mid x) = arg\ sup_\theta RB_{1,\alpha}(\theta \mid x)$$

and

$$\{\theta: RB_{t,\alpha}(\theta \mid x) \leq RB_{t,\alpha}(\theta_0 \mid x)\} = \{\psi: RB_{1,\alpha}(\theta \mid x) \leq RB_{1,\alpha}(\theta_0 \mid x)\}.$$

Therefore, the estimate of θ and the ordering of the θ values are determined by linear pooling.

2. It follows that $\cap_{i=1}^{k} Pl_i(x) \subset Pl_{1,\alpha}(x)$, or all those values where the statisticians agree that there is evidence in favour, also have evidence in their favour under linear pooling.

3. More generally, linear pooling *preserves consensus*, namely, if $RB_i(\theta \mid x) \geq (\leq)1$ for $i = 1, \ldots, k$ and at least one $RB_i(\theta \mid x) > (<)1$, then $RB_{1,\alpha}(\theta \mid x) > (<)1$ while other rules do not necessarily satisfy this, see [9].

4. If (i, θ, x) has prior $\alpha_i \pi_i(\theta) f_\theta(x)$, then the relative belief ratio is $RB_{1,\alpha}(\theta \mid x)$ and $\alpha_i m_i(x)/m_{1,\alpha}(x)$ is the posterior probability of i. This means that linear pooling is formally Bayes. Here α_i is interpreted as the convenor's prior belief concerning the reliability of the i-th statistician's inferences. Of course, it is possible to simply weight the statisticians equally. There are also many other approaches discussed in the literature for the determination of the α_i, see [20]-[22].

5. Linear pooling of priors does not preserve independence of events, which has been viewed as a negative property, but this is not an issue when combining evidence since, if $RB_i(\theta \mid x) = 1$ for $i = 1, \ldots, k$, then $RB_{1,\alpha}(\theta \mid x) = 1$.

6. The inferences are consistent as $n \to \infty$ and the prior that assigns the highest weight to the true value of the parameter will have the largest posterior weight for large n.

7. It follows from a result proved in [23] that linear pooling is marginalization consistent, namely,

$$RB_{1,\alpha,\Psi}(\psi \mid x) = E_{\pi(\cdot \mid \psi)}(RB_{1,\alpha}(\theta \mid x)$$

and it is the only combination rule that satisfies this property. This says that, if the relative belief ratio for θ is averaged using the conditional prior on the nuisance parameters, then the relative belief ratio for the parameter of interest is obtained. This implies that properties 1-6 also hold for $RB_{1,\alpha,\Psi}(\psi \mid x)$.

There seems to be little reason to doubt that combining via linear pooling is the correct way to proceed in Context I.

Context II is not as straight-forward and discussion is restricted here to situations where there is one set of data, although the models and priors can differ among the statisticians. The natural generalization of linear pooling is then to use the combination rule for parameter of interest ψ given by

$$RB^*_{1,\alpha,\Psi}(\psi \mid x) = \sum_{i=1}^{k} \frac{\alpha_i\, m_i(x)}{m_{1,\alpha}(x)} RB_{i,\Psi}(\psi \mid x)$$

with the posterior given by the combination

$$\pi^*_{1,\alpha,\Psi}(\psi \mid x) = \sum\nolimits_{i=1}^{k} \frac{\alpha_i m_i(x)}{m_{1,\alpha}(x)} \pi_{i,\Psi}(\psi \mid x).$$

In general, however, these rules are not formally Bayesian. In particular $RB^*_{1,\alpha,\Psi}(\psi \mid x)$ is not the ratio of the posterior $\pi^*_{1,\alpha,\Psi}(\psi \mid x)$ to the prior $\pi^*_{1,\alpha,\Psi}(\psi)$.

The form of $RB^*_{1,\alpha,\Psi}(\psi \mid x)$ can however be justified using the idea of Jeffrey conditionalization, see [24] and [25]. The value of $RB_{i,\Psi}(\psi \mid x)$ is formally Bayesian for the i-th statistician and, if the convener takes $\alpha_i m_i(x)$ to be their prior probability of (i, x), then, having observed the data x, their posterior probability of i becomes $\alpha_i m_i(x)/m_{1,\alpha}(x)$. The convenor then uses these probabilities to combine the measures of evidence and to form their overall posterior. So really there are several separate Bayesian updates taking place.

Indeed $RB^*_{1,\alpha,\Psi}$ has all the nice properties of $RB_{1,\alpha,\Psi}$, although it is not formally Bayes, but there is one concern that needs to be addressed. In Context I the weights $\alpha_i m_i(x)/m_{1,\alpha}(x)$ all depend on the data through a minimal sufficient statistic (mss) for the common model $\{m_\psi : \psi \in \Psi\}$ and also can be considered as conditional weights given the value of any ancillary statistic for this model. So, in Context I the i-th weight is determined by the convenor's initial prior probability α_i but also by how well the i-th model does at predicitng the observed value of this mss. In Context II it is generally not the case that $\alpha_i m_i(x)/m_{1,\alpha}(x)$ depends on the data through a common mss and ancillary statistics play a role in the prediction. This raises the issue as to whether or not these weights are directly comparable.

There are situations, however, where this issue can be addressed. For example, suppose for each model the data can be split as $x \leftrightarrow (L(x), A(x))$ where $A(x)$ is ancillary for each model. In this case it makes more sense to use the weights given by

$$\frac{\alpha_i m_i(x \mid A(x))}{m_{1,\alpha} a(x \mid A(x))}$$

as now all the inference bases are being considered based on how well they are predicting $L(x)$. This situation arises with group models and we provide a simple example here that is more extensively discussed in [9].

Example *Linear regression.*

Suppose that the data is (x_i, y_i) for $i = 1, \ldots, n$ and there are two analysts where both propose a simple regression model $y = X\beta + \sigma z$ where $X = (1_n \quad x) \in \mathbb{R}^{n\times 2}$with $1_n \perp x$ and $||x|| = 1, \beta = (\beta_1, \beta_2)' \in \mathbb{R}^2$ and $\sigma > 0$ are unknown and z is a sample of n from a $N(0,1)$ for analyst 1 and is a sample of n from a $t_\lambda/\sqrt{(\lambda-2)/\lambda}$ distribution for analyst 2 for some value $\lambda > 2$, where t_λ denotes the t distribution on λ degrees of freedom. In both models σ^2 is the variance of a y_i. Letting $b = (X'X)^{-1}X'y$ be the least squares estimate of β and $s^2 = ||y - Xb||^2$, then $y \leftrightarrow (L(y), A(y))$ where $L(y) = (b, s^2)$ and $A(y) = (y - Xb)/s$ is ancillary for both models. Further suppose that the quantity of inferential interest is the slope parameter $\psi = \Psi(\beta_1, \beta_2, \sigma^2) = \beta_2$.

For the prior suppose both analysts agree on $\beta \mid \sigma^2 \sim N_2(0, \tau_0^2\sigma^2 I)$ and $1/\sigma^2 \sim \text{gamma}(\alpha_1, \alpha_2)$. Note that the prior mean for β equal to 0 may entail subtracting a known, fixed constant vector from y so this, and the assumption that $1_n \perp x$, may entail some preprocessing of the data. The prior distribution of the quantity of interest is then $\beta_2 \sim \tau_0 \sqrt{\alpha_2/\alpha_1}\, t_{2\alpha_1}$.

Consider now a numerical example drawn from [26] where the response variable is income in U.S. dollars per capita (deflated), and the predictor variable is investment in dollars per capita (deflated) for the United States for the years 1922–1941. The data are provided in Table 1. The data vector y was replaced by $y - X(340,3)'$ as this centered the observations about 0. The hyperparameters were determined by an elicitation procedure, see [9], and this led to the values $\tau_0 = 0.54, \alpha_1 = 4.05, \alpha_2 = 140.39$.

Table 2 presents the weights that result when different standardized (to have unit variance) t_λ error distributions are considered to be combined with the results from a $N(0,1)$ error assumption. Presumably this arises when one analyst is concerned that tails longer than the normal are appropriate. As can be seen the normal error assumption dominates except for $\lambda = 100$ when the inferences don't differ by much in any case. This is not surprising as various residual plots don't indicate any issue with the normality assumption for these data. These weights were computed using importance sampling and were found to be robust to the prior by repeating the computations after making small changes to the hyperparameters. The approach taken in this example is easily generalized to more general linear regression models including situations where the priors change.

4 Conclusions

The problem of how to combine evidence has been considered for a Bayesian context where each analyst proposes a model and prior for the same data. Linear opinion pooling is seen as the natural way to make such a combination at least when the inference bases only differ in the priors on the parameter of interest. This has been shown to have appropriate properties such as preserving a consensus with respect to the evidence and, when combining evidence is considered as opposed to just combining priors, behaves appropriately when considering independent events. In certain contexts the idea can be extended in a logical way based on the idea underlying Jeffrey conditionalization. There are restrictions as in the end the posterior weights have to be seen to be comparable and focused on that aspect of the data which is relevant for inference about the unknowns. Asymptotically the approach behaves correctly.

This does not cover all contexts where one might want to combine evidence as when there are different data sets and different models. Generally, it may be that the only aspect in common among the models is the characteristic of interest Ψ and then it is not clear how we should combine and this warrants further investigation.

Table 1: Haavelmo's data on income and investment from [26]

Year	Income	Investment	Year	Income	Investment
1922	433	39	1932	372	22
1923	483	60	1933	381	17
1924	470	42	1934	419	27
1925	486	52	1935	449	33
1926	494	47	1936	511	48
1927	498	51	1937	520	51
1928	511	45	1938	477	33
1929	534	60	1939	517	46
1930	478	39	1940	548	54
1931	440	41	1941	629	100

Table 2: Weights for normal and (standardized) t_λ errors in the Example.

	λ =100	λ=50	λ=20	λ=10	λ=5	λ=3
$N(0,1)$	0.556	0.612	0.766	0.928	0.998	1.00
t_λ	0.444	0.388	0.234	0.072	0.002	0.000

References

[1] A. Birnbaum, On the foundations of statistical inference (with discussion). J. Amer. Stat. Assoc., 57, 269-332, 1962.
[2] G. Shafer, A Mathematical Theory of Evidence. Princeton University Press: Princeton, NJ, USA, 1976.
[3] R. Royall, Statistical Evidence: A Likelihood Paradigm. Chapman and Hall/CRC Press, 1997.
[4] M. Aitkin, Statistical Inference: an Integrated Bayesian/Likelihood Approach, Chapman and Hall/CRC Press 2010.
[5] M. Evans, Measuring Statistical Evidence Using Relative Belief. Chapman and Hall/CRC Press, 2015.
[6] R. Morey, J.-W. Romeijn and J. Rouder, The philosophy of Bayes factors and the quantification of statistical evidence. J. Mathematical Psychology, 72, 6-18, 2016.

[7] K. Popper, The Logic of Scientific Discovery. Harper Torchbooks, 1968.
[8] W. Salmon, Confirmation. Scientific American, 228, 5, 75-81, 1973.
[9] M. Evans and Y. Guo, Combining Evidence. arXiv:2202.02922, 2022.
[10] M. Stone, The opinion pool. Annals of Mathematical Statistics, 32,1339-1342, 1981.
[11] R. L. Winkler, The consensus of subjective probability distributions. Management Science, 15, 2, B-61-B-75, 1968.
[12] C. Genest, A characterization theorem for externally Bayesian groups. Annals of Statistics, 12, 2, 1100-1105, 1984.
[13] C. Genest, A conflict between two axioms for combining subjective distributions. J. of the Royal Statistical Society. Series B, 46 (3), 403-405, 1984.
[14] C. Genest, Pooling operators with the marginalization property. Canadian Journal of Statistics, 12, 2, 153-163, 1984.
[15] C. Genest and J.V. Zidek, Combining probability distributions: a critique and an annotated bibliography. Statistical Science, 1, 114-135, 1986.
[16] R.T. Clemen and R.L. Winkler, Combining probability distributions from experts in risk analysis. Risk Analysis, 19 (2), 187-203, 1999.
[17] A. O'Hagan, C.E. Buck., A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley and T. Rakow, Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons Ltd, 2006.
[18] S. French, Aggregating expert judgement. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A, Matemáticas, 03, 105 (1), 181-206, Springer, 2011.
[19] C. Farr, F. Ruggeri and K. Mengersen, Prior and posterior linear pooling for combining expert opinions: uses and impact on Bayesian networks—the case of the wayfinding model. Entropy, 20, 209. doi:10.3390/e20030209, 2018.
[20] M.H. DeGroot, Reaching a consensus. J. of the American Statistical Association, 69, 118-121, 1974.
[21] K. Lehrer, K. and C. Wagner Rational Consensus in Science and Society. D. Reidel Publishing Co, 1981.
[22] C. Genest and K.J. McConway, Allocating the weights in the linear opinion pool. J. of Forecasting, 9, 53-73, 1990.
[23] K.J. McConway, Marginalization and linear opinion pools. J. of the American Stat. Assoc., 76:374, 410-414, 1981.
[24] R. Jeffrey, The Logic of Decision, New York: McGraw-Hill, 1965.
[25] P. Diaconis and S. Zabell, Updating subjective probability. Journal of the American Statistical Association, 77, 380, 822-830, 1982.
[26] A. Zellner, An Introduction to Bayesian Inference in Econometrics. Wiley Classics, 1996.