

Invited reply



Cite this article: Martin R, Balch MS, Ferson S. 2021 Response to the comment Confidence in confidence distributions!. *Proc. R. Soc. A* **477**: 20200579.
<https://doi.org/10.1098/rspa.2020.0579>

Received: 21 July 2020

Accepted: 25 May 2021

Subject Areas:

statistics

Keywords:

false confidence, non-additive beliefs, inferential model, satellite conjunction analysis

Author for correspondence:

Ryan Martin

e-mail: rgmarti3@ncsu.edu

The accompanying comment can be viewed at
<http://doi.org/10.1098/rspa.2019.0781>.

Response to the comment Confidence in confidence distributions!

Ryan Martin¹, Michael S. Balch² and Scott Ferson³

¹North Carolina State University College of Sciences, Statistics, Raleigh, NC 27695, USA

²Alexandria Validation Consulting, LLC, Alexandria, 22309 VA, USA

³Department of Civil Engineering and Industrial Design, University of Liverpool School of Engineering, Liverpool L69 3BX, UK

RM, 0000-0003-0760-9145; MSB, 0000-0003-4018-6471

1. Introduction

Thanks to Drs Céline Cunen, Nils Lid Hjort and Tore Schweder for their interest in our recent contribution [1] concerning the probability dilution phenomenon in satellite conjunction analysis and, more generally, the difficulties associated with representing statistical inference using ordinary or precise probabilities. Our analysis focused primarily on Bayesian uncertainty quantification but, of course, this is not the only probabilistic approach available, so we welcome a confidence distribution-based solution from those who literally wrote the book on confidence distributions [2]. Their illustration reproduces the lack of proper calibration—or *false confidence*—that can emerge when marginalizing a Bayesian posterior distribution and highlights the difference between Bayesian posteriors and confidence distributions (CDs) with respect to the validity property advocated in our paper.

However, the message presented by Cunen *et al.* [3]—that replacing a Bayesian posterior distribution with a CD is all it takes to overcome false confidence—is potentially misleading. Our false confidence theorem applies to all epistemic probability distributions, including CDs; so, contrary to the authors' claim, their proposed CD is at risk of false confidence too. In particular, as is well known, CDs only support reliable inferences on one-sided propositions of the form $(-\infty, a]$ or $[a, +\infty)$. Other sets, including two-sided intervals and their complements, are still subject to the false confidence phenomenon uncovered in Balch *et al.* [1].

Section 2 confirms the presence of false confidence in CDs with a simple example. Section 3 explores one path forward for CDs to be free of false confidence, at least in a certain limited sense. Section 4 emphasizes an important but overlooked point made in Balch *et al.* [1], Sec. 4, namely, that coverage probability control is not enough to ward off false confidence. Adding in the missing consonance feature takes one into the world of imprecise probability, and we show in §5 how this can provide another path forward for advocates of CDs to justifiably claim they are free of false confidence. Section 6 ends the note with some concluding remarks.

2. CDs only offer some C

In the abstract, Cunen *et al.* [3] present their take-home message:

confidence distributions [are] free of the false confidence syndrome.

Simply put, this claim is false; in fact, it cannot be true because our false confidence theorem applies to all epistemic probability distributions, including CDs. The following example confirms this.

To set the scene, a simplified version of the satellite collision problem assumes a bivariate Gaussian observation $Y = (Y_1, Y_2)$ with unknown mean vector $\theta = (\theta_1, \theta_2)$ and known covariance matrix $\sigma^2 I$, representing the true displacement between two satellites at closest approach and the random error in a navigator's prediction of that displacement, respectively. The quantity of interest in this setting is $\delta = \|\theta\| = (\theta_1^2 + \theta_2^2)^{1/2}$, the true distance at closest approach. If that distance is less than the combined size of the two satellites, a collision will occur. Given $y = (y_1, y_2)$, let $C_Y(d)$ denote the distribution function that determines Cunen *et al.*'s CD, i.e.

$$C_Y(d) = 1 - H(\sigma^{-2}\|y\|^2; \sigma^{-2}d^2),$$

where $H(\cdot; \lambda)$ denotes a non-central χ^2 distribution function with 2 degrees of freedom and non-centrality parameter λ . With a slight abuse of notation, let $C_Y(A)$ denote the confidence assigned by distribution C_Y to a hypothesis $A \subset (0, \infty)$ about δ . For a hypothesis of the form $A_\delta = (0, \delta]$, $C_Y(A_\delta)$ has a standard uniform distribution, as a function of Y , if the true distance at closest approach happens to be δ ; in fact, this is the defining property of a CD (e.g. [2,4]). This implies that, for any hypothesis $A \subset A_\delta^c$, which is false, the probability that $C_Y(A)$ is large would be relatively small, so there is no false confidence for that hypothesis A . But there are many other potentially false hypotheses, and the above argument says nothing about the behaviour of $C_Y(A)$ for those A 's. For example, consider $A = (0, 1.5] \cup [2.5, \infty)$. If the true distance at closest approach happens to be 2, then this hypothesis is *false*; so, $C_Y(A)$ should tend to be relatively small. To check this, we simulated data Y with true distance at closest approach equal to 2 and several different values of the error standard deviation σ . Figure 1 plots the distribution function of the random variable $C_Y(A)$ for those different σ values. Clearly, this confidence assignment tends to be large, i.e. relatively close to 1, and the practical consequence is that conclusions drawn about the hypothesis A based on the magnitude of $C_Y(A)$ are at risk of being systematically wrong. Therefore, contrary to Cunen *et al.*'s claim, their CD is not free of false confidence in the sense of Balch *et al.* [1].

Two points deserve further emphasis. First, the problematic assertion A identified above is not the only one. As we said, any A that is not fully contained in the interval $A_\delta^c = (\delta, \infty)$ and does not contain δ would be at risk of false confidence, regardless of the true δ value. The A selected for the above illustration was just one of those we found where the effect of false confidence was especially clear. Second, while our illustration of false confidence involves knowledge of the true δ value—otherwise, we would not be able to identify a 'false' hypothesis to investigate in the simulation—the phenomenon itself does not depend on or require a known δ . It is no different than how the CD definition ' $C_Y(A_\delta)$ is uniformly distributed at the true δ value' depends on the true δ value. The point is that a hypothesis A is either true or false, and in applications we use

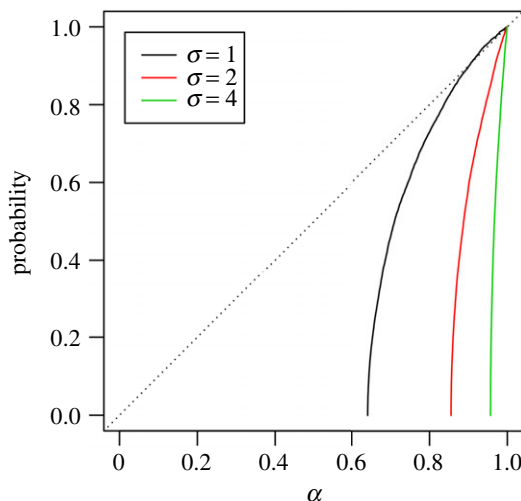


Figure 1. Plots of the distribution function $\alpha \mapsto P_{Y|\theta} \{C_Y(A) \leq \alpha\}$, where $C_Y(A)$ is the confidence assigned to the false hypothesis A by the CD. (Online version in colour.)

data to help make inferences. Since there exist false hypotheses A for which $C_Y(A)$ tends to be large, there is good reason to question those inferences drawn from the CD.

3. CDs are not really Ds

Of course, in the context of satellite conjunction analysis, the satellite navigator really does only care about one-sided sets on δ , in particular, whether the two satellites will get so close together that they actually collide. Our point here is not that the CD solution to conjunction analysis is totally unworkable but, rather, that CDs are not ‘distributions’ in any meaningful sense. The chief advantage of basing inference on a data-dependent distribution is that it ought to provide a complete quantification of uncertainty about the unknown parameters of the statistical model based on the observed data. So once such a distribution has been constructed, it can be used to answer all relevant questions at once. But, according to the false confidence theorem, all epistemic probability distributions are afflicted by false confidence, and hence the distribution’s answers to some questions are not reliable. To avoid this undesirable behaviour, the only option is to quantify uncertainty with something less committal than an ordinary or precise probability distribution.

CDs, properly caveated, could fall under this ‘something less committal’ umbrella. The truth is, CDs are limited objects; they only support reliable inferences on one-sided hypotheses about scalar parameters:

distributions derived from a CD by ordinary calculus do however not automatically inherit the property of being a CD ... even in dimension 1 ([5], p. 58).

Also,

joint CDs should not be sought, we think, since they might easily lead the statistician astray ([5], p. 59).

This is why Cunen *et al.* [3] do not start with a joint CD for θ in the satellite collision example and derive CD for δ . Instead, they directly construct a CD for δ . This also explains the phenomenon observed in the example from §2: that problematic hypothesis corresponds to a one-sided set on a new parameter $\delta' = |\delta - 2|$, which is a not-one-to-one function of δ , and the corresponding

marginal distribution for δ' is not necessarily a CD. If CDs lack the status of a distribution, as Cunen *et al.* and many others acknowledge [6], then they are less committal than an ordinary or precise probability. And the way to ensure that false confidence is avoided, at least in a certain limited sense, is for CDs' restriction to one-sided hypotheses to be clearly stated and consistently enforced. This is preferable to making claims like 'confidence in confidence distributions!' that hide the risks that might very well lead the statistician astray.

The downside to this less committal perspective on CDs is that they lose much of their appeal. Given a CD for one variable, if the proposition of interest is a one-sided hypothesis concerning a second variable, then one must derive a completely new CD about that second variable. And there is no clear *general* path for doing that—at least not using the ordinary/precise probability calculus; but see §5 below.

Some authors have relaxed the requirement that CDs be proper probability distributions. For example, Schweder & Hjort ([5], p. 61) suggest that a CD could be improper, i.e. assign infinite total mass to the parameter space, and Thornton & Xie ([7], Def. 3) replace a single CD with a lower and upper CD pair. But neither of these proposed relaxations addresses the issue in question. First, an improper CD does not provide probabilistic uncertainty quantification because, without propriety, the resulting $C_Y(A)$ values could be arbitrarily large, even infinite, so interpretation is unclear. Second, the sole purpose of lower/upper CDs in the latter reference is to obtain confidence intervals in the context of discrete data; the authors make no proposal for using their lower/upper CD pair for assigning lower/upper probabilities to quantify uncertainty. Moreover, it was shown recently [8] that a particular imprecise probability model—a 'confidence-box'—encoded in the lower/upper CD pair derived for binomial inference in Thornton & Xie ([7], Example 2) is also afflicted by false confidence. So, the false confidence phenomenon is rather subtle; it cannot be avoided simply by achieving a coverage probability property or by assigning beliefs in an arbitrary but non-additive way. The next two sections explain this in more detail.

4. Coverage probability is not enough

Several statements in Cunen *et al.* [3] suggest that belief assignments made by CDs inherit the reliability properties of those made by simple confidence regions. This claim is also false. The authors misunderstood the proof offered in §4 of Balch *et al.* [1], which demonstrates that simple confidence regions are free from false confidence. However, as stated in the footnote accompanying that section, the reliability of simple confidence regions is due to the combination of coverage probability *and* consonance. Simple confidence regions are trivially consonant and thereby enjoy the reliability properties that hold for all consonant confidence structures, as proved in Denœux & Li [9]. However, if you build a *non-consonant* structure out of multiple confidence regions—for example, a CD—the broad reliability guarantees that held for the individual constituent confidence regions do not necessarily hold for the total structure, as has been conclusively demonstrated in §2 of this note and elsewhere. Otherwise, as proved in Balch [10], it *would* be possible to propagate CDs meaningfully. But because coverage probability is such a weak criterion, in and of itself, the scheme outlined in Balch [10] is only useful when applied to *consonant* confidence structures.

5. Coverage + consonance = no false confidence

There is a relatively straightforward fix to the shortcoming of coverage probabilities and CDs. One can simply recast the CD in a possibilistic framework. In the CD literature, there is an object called the *confidence curve*, dating back to Birnbaum [11]. In Cunen *et al.* [3], this is given by

$$c_y(d) = |2C_y(d) - 1|.$$

The confidence curve is somewhat mysterious because the above manipulation is not a meaningful one in the theory of probability. But it turns out that it is a fundamental structure

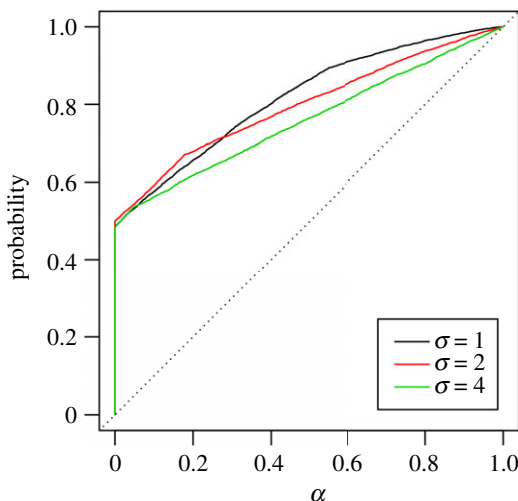


Figure 2. Plots of the distribution function $\alpha \mapsto P_{Y|\theta}\{b_Y(A) \leq \alpha\}$, where $b_Y(A)$ is the confidence assigned to the false hypothesis A by (5.1). (Online version in colour.)

in the theory of imprecise probability. Indeed,

$$p_Y(d) = 1 - c_Y(d)$$

is the so-called *plausibility contour* associated with a *consonant belief function* (e.g. [12,13]); see, also, the seminal work of Dempster [14–16] and the connections to possibility theory [17] and imprecise probability more generally (e.g. [18]). In a possibilistic framework, the confidence assignments are carried out differently. For a hypothesis A about δ , the degree of belief or support for a proposition or hypothesis is computed as

$$b_Y(A) = 1 - \sup_{d \in A^c} p_Y(d). \quad (5.1)$$

For illustration, we repeat the simulation study described in §2 above, and plot the distribution of $b_Y(A)$, as a function of data Y , for that same problematic hypothesis A . Figure 2 displays this distribution function for three different values of σ . The key difference compared with figure 1 is that this distribution function is above the diagonal line, indicating that $b_Y(A)$ tends to be small. This means the support for the false hypothesis is relatively small, hence no false confidence.

6. Conclusion

Contrary to claims in Cunen *et al.* [3], CDs are not completely free of false confidence, not if they are understood as probability distributions. However, all it takes to remove the CD's false confidence affliction is to change perspectives, to recognize that the complement of the corresponding confidence curve is a plausibility contour, and work in an imprecise rather than precise probability framework. This perspective also makes it possible to construct a multi-parameter CD but, again, interpreted as an imprecise probability with different rules for marginalization compared with precise probability. Connections between confidence—and frequentist inference more generally—and imprecise probability can be found in Balch [10] and Martin [19–22], but more work is needed. Also, the solution based on (5.1) is exactly that presented in Martin & Liu ([23], Sec. 4.3) based on an alternative approach, the so-called inferential model (IM) framework, which works directly in the domain of imprecise probability through the use of random sets. To our knowledge, the only general approach to distributional inference without false confidence is the IM framework; see Martin & Liu [24,25] and Martin [26].

Finally, while it is clearly desirable to avoid false confidence if possible, it is also interesting to better understand what kind of hypotheses are afflicted by false confidence. Currently, theory establishes the existence of problematic hypotheses, and a few have been constructed in specific examples. A more precise characterization of those problematic hypotheses would shed more light on the limitations of and risks associated with the use of probability as a tool for uncertainty quantification in statistical inference.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. R.M. acknowledges support from the NSF (DMS-1811802) and S.F. acknowledges support from the EPSRC (EP/R006768/1).

References

- Balch MS, Martin R, Ferson S. 2019 Satellite conjunction analysis and the false confidence theorem. *Proc. R. Soc. A* **475**, 20180565. (doi:10.1098/rspa.2018.0565)
- Schweder T, Hjort NL. 2016 *Confidence, likelihood, probability*, vol. 41. Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY: Cambridge University Press.
- Cunen C, Hjort NL, Schweder T. 2020 Confidence in confidence distributions! *Proc. R. Soc. A* **476**, 20190781. (doi:10.1098/rspa.2019.0781)
- Xie M-G, Singh K. 2013 Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int. Stat. Rev.* **81**, 3–39. (doi:10.1111/insr.12000)
- Schweder T, Hjort NL. 2013 Discussion: ‘Confidence distribution, the frequentist distribution estimator of a parameter: a review’ [mr3047496]. *Int. Stat. Rev.* **81**, 56–68. (doi:10.1111/insr.12004)
- Fraser DAS. 2011 Rejoinder: ‘Is Bayes posterior just quick and dirty confidence?’ *Stat. Sci.* **26**, 329–331.
- Thornton S, Xie M-G. 2020 Bridging Bayesian, frequentist and fiducial (BFF) inferences using confidence distribution. (<http://arxiv.org/abs/2012.04464>)
- Balch MS. 2020 New two-sided confidence intervals for binomial inference derived using Walley’s imprecise posterior likelihood as a test statistic. *Int. J. Approx. Reason.* **123**, 77–98. (doi:10.1016/j.ijar.2020.05.005)
- Dencœux T, Li S. 2018 Frequency-calibrated belief functions: review and new insights. *Int. J. Approx. Reason.* **92**, 232–254. (doi:10.1016/j.ijar.2017.10.013)
- Balch MS. 2012 Mathematical foundations for a theory of confidence structures. *Int. J. Approx. Reason.* **53**, 1003–1019. (doi:10.1016/j.ijar.2012.05.006)
- Birnbaum A. 1961 Confidence curves: an omnibus technique for estimation and testing statistical hypotheses. *J. Am. Stat. Assoc.* **56**, 246–249. (doi:10.1080/01621459.1961.10482107)
- Shafer G. 1976 *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Shafer G. 1987 Belief functions and possibility measures. In *The Analysis of Fuzzy Information* (ed. JC Bezdek), vol. 1: Mathematics and Logic, pp. 51–84. New York, NY: CRC.
- Dempster AP. 1966 New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Stat.* **37**, 355–374. (doi:10.1214/aoms/1177699517)
- Dempster AP. 1967 Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Stat.* **38**, 325–339. (doi:10.1214/aoms/1177698950)
- Dempster AP. 1968 Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statist.* **39**, 957–966. (doi:10.1214/aoms/1177698328)
- Dubois D, Prade H. 1988 *Possibility theory*. New York, NY: Plenum Press.
- Augustin T, Coolen FPA, de Cooman G, Troffaes MCM (eds). 2014 *Introduction to imprecise probabilities*. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons, Ltd.
- Martin R. 2015 Plausibility functions and exact frequentist inference. *J. Am. Stat. Assoc.* **110**, 1552–1561. (doi:10.1080/01621459.2014.983232)
- Martin R. 2017 A statistical inference course based on p -values. *Am. Stat.* **71**, 128–136. (doi:10.1080/00031305.2016.1208629)
- Martin R. 2018 On an inferential model construction using generalized associations. *J. Stat. Plann. Inference* **195**, 105–115. (doi:10.1016/j.jspi.2016.11.006)
- Martin R. 2021 An imprecise-probabilistic characterization of frequentist statistical inference. See <https://researchers.one/articles/21.01.00002>.

23. Martin R, Liu C. 2015 Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Am. Stat. Assoc.* **110**, 1621–1631. (doi:10.1080/01621459.2014.985827)
24. Martin R, Liu C. 2013 Inferential models: a framework for prior-free posterior probabilistic inference. *J. Am. Stat. Assoc.* **108**, 301–313. (doi:10.1080/01621459.2012.747960)
25. Martin R, Liu C. 2015 *Inferential models: reasoning with uncertainty*, vol. 147. Monographs on Statistics and Applied Probability. Boca Raton, FL: CRC Press.
26. Martin R. 2019 False confidence, non-additive beliefs, and valid statistical inference. *Int. J. Approx. Reason.* **113**, 39–73. (doi:10.1016/j.ijar.2019.06.005)