CONFIDENCE DISTRIBUTIONS AND RELATED THEMES

Nils Lid Hjort[1] **and Tore Schweder**[2]

[1]**Department of Mathematics and** [2]**Department of Economics**

University of Oslo

Abstract. This is the guest editors' general introduction to a Special Issue of the Journal of Statistical Planning and Inference, dedicated to confidence distributions and related themes. Confidence distributions (CDs) are distributions for parameters of interest, constructed via a statistical model after analysing the data. As such they serve the same purpose for the frequentist statisticians as the posterior distributions for the Bayesians. There have been several attempts in the literature to put up a clear theory for such confidence distributions, from Fisher's fiducial inference and onwards. There are certain obstacles and difficulties involved in these attempts, both conceptually and operationally, which have contributed to the CDs being slow in entering statistical mainstream. Recently there is a renewed surge of interest in CDs and various related themes, however, reflected in both series of new methodological research, advanced applications to substantive sciences, and dissemination and communication via workshops and conferences. The present special issue of the JSPI is a collection of papers emanating from the *Inference With Confidence* workshop in Oslo, May 2015. Several of the papers appearing here were first presented at that workshop. The present collection includes however also new research papers from other scholars in the field.

Key words: confidence curves, confidence distributions, focus parameters, likelihood, meta-analysis, probability

The Journal of Statistical Planning and Inference decided in the autumn of 2015 to arrange for a Special Issue on confidence distribution and related themes. After various efforts, by patient authors, referees, and colleagues, along with the customary revision processes, this has resulted in the current collection of eleven journal articles:

1 Cunen et al. (2017a), on CDs and confidence curves for change points, with applications to mediaeval literature and to fisheries sciences;
2 De Blasi & Schweder (2017), on median bias corrections for fine-tuning CDs;
3 Grünwald (2017), on safe probability, leading also to tools for predictions;
4 Hannig et al. (2017), on fusion learning and inter-laboratory analyses;
5 Lewis (2017), on combining inferences, with application to climate statistics;

Date: June 2017.

6 Lindqvist & Taraldsen (2017), on proper uses of improper distributions;
7 Martin (2017), on generalised inference models;
8 Schweder (2017), with an essay on epistemic probability;
9 Shen et al. (2017), on CDs for predictions, in different setups;
10 Taraldsen & Lindqvist (2017), on conditional fiducial models; and
11 Veronese & Melilli (2017), on CDs and their connections to objective Bayes.

These papers deal with theory and applications for distributional statistical inference, with CDs and fiducial distributions being the central concepts. Quite a few contributions also touch Bayesian angles and connections, however (Cunen et al. (2017a), Grünwald (2017), Lewis (2017), Lindqvist & Taraldsen (2017), Taraldsen & Lindqvist (2017), Veronese & Melilli (2017)). In the present general introduction to the Special Issue, by the guest editors, efforts are made both to explain to the broader statistical audience what confidence distributions (CDs) and confidence curves are; why and how they are steadily becoming more popular, in statistical theory and practice; and to briefly place the eleven papers in a broader context. In our article, which is by itself a gentle introduction to the general CD themes, we also attempt to point to aspects and issues and types of application not already contained in the review paper Xie & Singh (2013) and ensuing discussion.

1. The Holy Grail: frequentist posterior distributions

Suppose data are analysed via some model, and that ψ is a parameter of particular interest. Statisticians have many methods in their toolboxes for conducting inference for ψ, such as reaching a point estimate, assessing its precision, setting up tests, along with p-values when of relevance, finding confidence intervals, comparing the ψ with other parameters from other studies, etc. For the frequentist, constructing a distribution for ψ, given the available information, is more problematic, however, also conceptually.

Somehow it appears to be a strict Bayesian privilege to arrive at an appropriate posterior distribution, say $p(\psi \,|\, \text{data})$ – along with the associated difficulties of carrying out Bayesian work in the first place, involving elicitation of prior distributions and combining these with probability distributions of a different kind. Working out a $p(\psi \,|\, \text{data})$ in the frequentist framework appears to clash with the basic premise that the parameter vector of the model is a fixed but unknown point in the parameter space. This has not stopped scholars from attempting precisely such a feat, called the Holy Grail of parametric statistics by Brad Efron (Efron, 2010). The earliest attempts were by none other than Sir Ronald Fisher, in a series of papers in the 1930ies (Fisher, 1930, 1932, 1933, 1935). Certain obstacles and difficulties were found and pointed to by a number of critical scholars, however, and Fisher did not

quite manage to defend his notion of a fiducial distribution for parameters. Indeed the fiducial ideas have been referred to as 'Fisher's biggest blunder'; see Schweder & Hjort (2016, Ch. 6) for an account of the historical development, and also Grünwald (2017, this issue).

There are however other and partly related notions of how to reach proper frequentist posterior distributions, without priors, and the collective labels for a fair portion of these refined and modernised constructions are *confidence distributions* (CDs) and *confidence curves*. There is a clear surge of interest in these methods and in various related themes, regarding both theory and applications. This is witnessed in books and journal articles and by applied advanced work, and is also reflected in high-level workshops and conferences. The *BFF: Bayes, Frequentist, Fiducial* series of conferences (also referred to as 'Best Friends Forever') is reaching a steadily wider audience, with the current list being Shanghai (2014, 2015), Rutgers, New Jersey (2016), Harvard, Massachusetts (2017), Ann Arbor, Michigan (2018), and Duke and SAMSI, North Carolina (2019). There are also special invited sessions at major conferences, etc., dedicated to CDs and BFF themes. Efron (1998) speculates that Fisher's (alleged) biggest blunder might turn into a big hit for the 21st century; see also Efron & Hastie (2016, Ch. 11).

The present special issue of the JSPI is dedicated to such CDs and the growing list of related topics. The collection of papers and the ensuing organisation of the special issue have grown out of one of these conferences, the *Inference With Confidence* workshop in Oslo in May 2015, organised by the the research group *FocuStat: Focus Driven Statistical Inference With Complex Data*. Some of the papers appearing in this issue were first presented as invited lectures at this workshop. We have also recruited contributions from other scholars in the field, however, in an attempt to exhibit and see discussed a decent range of the more crucial dimensions of CDs and their increasing scope and usefulness, in methodological and applied statistical work.

"The three revolutions in parametric statistical inference are due to Laplace (1774), Gauss and Laplace (1809–1811) and Fisher (1922)", is the clear opening statement in the two books Hald (1998, 2006). Somewhat boldly, Schweder & Hjort (2016, Preface) claim there is an ongoing fourth revolution in statistics, at the start of the current millennium. This fourth revolution has perhaps a less clear focus than the three drastic methodological changes Hald describes, and is arguably more about the *who* and *what* than about the *how*, but we argue there that CDs and confidence curves have a natural place in the world of statistical computation and communication, also with Big Data. "I wish I'd seen a confidence curve earlier", as tweeted J.M. White, who manages a branch of Facebook's Core Data Science team, in April 2017. We

should also make clear that there by necessity are several approaches (partly related and partly competing) to the alleged Holy Grail of reaching posteriors without priors. In addition to the CD theory expounded in Schweder & Hjort (2002, 2003, 2016); Xie & Singh (2013), with roots all the way back to Fisher in the 1930ies, there is generalised fiducial inference, see Hannig et al. (2016) and Hannig et al. (2017, this issue), along with Lindqvist & Taraldsen (2017, this issue) and Taraldsen & Lindqvist (2017, this issue); as well as the theory of inferential models, cf. Martin & Liu (2015) and Martin (2017, this issue). There is bound to be yet other hybrids and connections, and some of these are touched upon in the present collection of journal articles.

2. What are confidence distributions and confidence curves?

There are several ways in which to motivate, define and construct such CDs, along with associated concepts and functions. Suppose the model for the data y is governed by a parameter vector θ, and that the interest parameter ψ is a function $\psi(\theta)$ of the model parameter. A modern definition of a *confidence curve* for ψ, say $\mathrm{cc}(\psi, y)$, see Schweder & Hjort (2002, 2016); Xie & Singh (2013), is as follows. We write Y for the random outcome of the data generating mechanism and $y_{\rm obs}$ for the actually observed data. At the true parameter point $\psi_0 = \psi(\theta_0)$, the random variable $\mathrm{cc}(\psi_0, Y)$ should have a uniform distribution on the unit interval. Then

$$P_{\theta_0}\{\mathrm{cc}(\psi_0, Y) \le \alpha\} = \alpha \quad \text{for all } \alpha. \tag{2.1}$$

Thus confidence intervals, and more generally confidence regions, can be read off, at each desired level; the 90% confidence region is $\{\psi\colon \mathrm{cc}(\psi, y_{\rm obs}) \le 0.90\}$, etc. When α tends to zero the confidence region typically tends to a single point, say $\widehat{\psi}$, an estimator of ψ. In regular cases the $\mathrm{cc}(\psi, y)$ is decreasing to the left of $\widehat{\psi}$ and increasing to the right, in which case the confidence curve $\mathrm{cc}(\psi, y)$ can be uniquely linked to a full confidence distribution $C(\psi, y)$, via

$$\mathrm{cc}(\psi, y) = |1 - 2\,C(\psi, y)| = \begin{cases} 1 - 2\,C(\psi, y) & \text{if } \psi \le \widehat{\psi}, \\ 2\,C(\psi, y) - 1 & \text{if } \psi \ge \widehat{\psi}. \end{cases} \tag{2.2}$$

The confidence name given to these post-data summaries for focus parameters stems from the intimate connection to the familiar confidence intervals. With $C(\psi, y)$ a CD, $[C^{-1}(0.05, y_{\rm obs}), C^{-1}(0.95, y_{\rm obs})]$ becomes an equi-tailed 90% confidence interval, etc. Also, solving $\mathrm{cc}(\psi, y_{\rm obs}) = 0.90$ yields two cut-off points for ψ, precisely those of the 90% confidence interval. Correspondingly one may start with a given set of nested confidence intervals, for all levels α, and convert these into, precisely, a CD.

3. General recipes

Suppose a model with parameter vector θ is used for data y and again that $\psi = \psi(\theta)$ is a focus parameter. If $\text{piv}(\psi, y)$ is a function monotone increasing in ψ, with a distribution not depending on the underlying parameter, we term it a pivot. Thus $K(x) = P_\theta\{\text{piv}(\psi, Y) \le x\}$ does not depend on θ, or on ψ, which implies that

$$C(\psi, y) = K(\text{piv}(\psi, y))$$

is a CD. The classical construction of this type is that of Student (1908), namely

$$t = \frac{\mu - \bar{y}}{s/\sqrt{n}}$$

for a normal sample, with $\bar{y}$ and s denoting the sample mean and empirical standard deviation. The ensuing CD for μ becomes

$$C(\mu, \text{data}) = F_\nu(\sqrt{n}(\mu - \bar{y})/s),$$

with F_ν the cumulative distribution function of a t distribution with the relevant degrees of freedom.

In various classical setups for parametric models, there are well-working large-sample approximations for the the behaviour of estimators, deviance functions, etc., and these lead to constructions of CDs and confidence curves. First, if $\widehat{\psi}$ is such that $\sqrt{n}(\widehat{\psi} - \psi) \to_d \text{N}(0, \tau^2)$, and $\widehat{\tau}$ is a consistent estimator for the τ in question, then $\sqrt{n}(\widehat{\psi} - \psi)/\widehat{\tau} \to_d \text{N}(0, 1)$. Writing

$$C_n(\psi, \mathcal{D}_n) = \Phi(\sqrt{n}(\psi - \widehat{\psi})/\widehat{\tau}), \tag{3.1}$$

therefore, with $\mathcal{D}_n$ the data available after n observations, we have $C_n(\psi, \mathcal{D}_n) \to_d \text{unif}$; in particular, the $C_n(\psi, \mathcal{D}_n)$ is asymptotically a pivot in the above sense. Hence such a $C_n(\psi, \mathcal{D}_{n,\text{obs}})$ is a large-sample valid CD, allowing us to write

$$\psi \,|\, \text{data} \approx_d \text{N}(\widehat{\psi}, \widehat{\tau}^2/n), \tag{3.2}$$

in the CD sense. This is akin to a Bayesian posterior distribution for ψ (but without any notion of a prior distribution involved). Also, the associated confidence curve, asymptotically valid, is

$$\text{cc}(\psi, D_{n,\text{obs}}) = |1 - 2\,\Phi(\sqrt{n}(\psi - \widehat{\psi}_{\text{obs}})/\widehat{\tau}_{\text{obs}})|.$$

These first-order large-sample approximations (3.1)–(3.2) are simple and useful but sometimes too coarse. A recipe that typically works better is the following. With $\ell_n(\theta)$ the log-likelihood function, let $\ell_{n,\text{prof}}(\psi) = \max\{\ell_n(\theta) \colon \psi(\theta) = \psi\}$ be the profile, which we then turn into the deviance function

$$\text{dev}_n(\psi) = 2\{\ell_{n,\text{prof}}(\widehat{\psi}) - \ell_{n,\text{prof}}(\psi)\}. \tag{3.3}$$

By the Wilks theorem (see e.g. Schweder & Hjort (2016, Chs. 2-3)), under mild regularity conditions $\mathrm{dev}_n(\psi_0) \to_d \chi^2_1$, at the true value $\psi_0 = \psi(\theta_0)$. Hence $\mathrm{cc}_n(\psi_0, \mathcal{D}_n) = \Gamma_1(\mathrm{dev}_n(\psi_0)) \to_d \mathrm{unif}$, with $\Gamma_1(\cdot)$ denoting the χ^2_1 distribution function, and

$$\mathrm{cc}_n(\psi, \mathcal{D}_{n,\mathrm{obs}}) = \Gamma_1(\mathrm{dev}_n(\psi)) \tag{3.4}$$

is our confidence curve. It can reflect asymmetry and also likelihood multimodality in the underlying distributions, unlike the simpler method of (3.1). Since a confidence curve can be derived from a proper CD, via (2.2), but not always the other way around, the confidence curve is arguably a more fundamental notion or concept than a CD.

There is an extensive literature in probability theory and statistics regarding the many ways of fine-tuning the distributional approximations associated with the first-order normality result (3.1) and the Wilks theorem for (3.3). Key words for such methods include Bartletting, expansions, modified profiles, saddlepointing, bootstrap refinements, prepivoting, etc.; see e.g. Brazzale et al. (2007); Brazzale & Davison (2008); Barndorff-Nielsen & Cox (1994). Many of these methods may then be worked with further to yield fine-tuning instruments for CDs and confidence curves. Some of these translations, from the more traditional setup of assessing accuracy of a certain approximation, or how to correct for a type of bias, are fairly straightforward, leading to good CD recipes. Other such translations, involving perhaps higher-level bootstrapping or modified log-likelihood operations, are non-trivial. Interestingly, some of the more intricate procedures, like Barndorff-Nielsen's 'magic formula', have relatively speaking easier cousins in the CD universe of things, and potentially with easier explanations; see Schweder & Hjort (2016, Ch. 7) for discussion and illustrations.

A confidence curve analysis is often much more informative than providing the prototypical 95% interval or a p-value for an associated hypothesis test. Figure 3.1 displays the confidence curve $\mathrm{cc}(p)$ for the probability p that the world would see a 100 m sprint race in a time of 9.72 seconds or faster, inside the calendar year 2008, with this question asked on January 1 that year. In other words, this is an attempt to quantify how surprised we ought to have been, when we learned that Usain Bolt had set his first world record, in May that year. We have used the general apparatus of extreme value theory to make such a question precise, taking as data the $n = 195$ races (which we were able to track down from various sources) with a result time of 10.00 or better, in the course of the eight calendar years 2000–2007. Theory for extreme values leads to a certain parametric form for the best races, involving parameters (a, σ) (and the model has been shown to fit very well to the sprint data). The $\mathrm{cc}(p)$ given in the figure has come about by (i) expressing p as a function

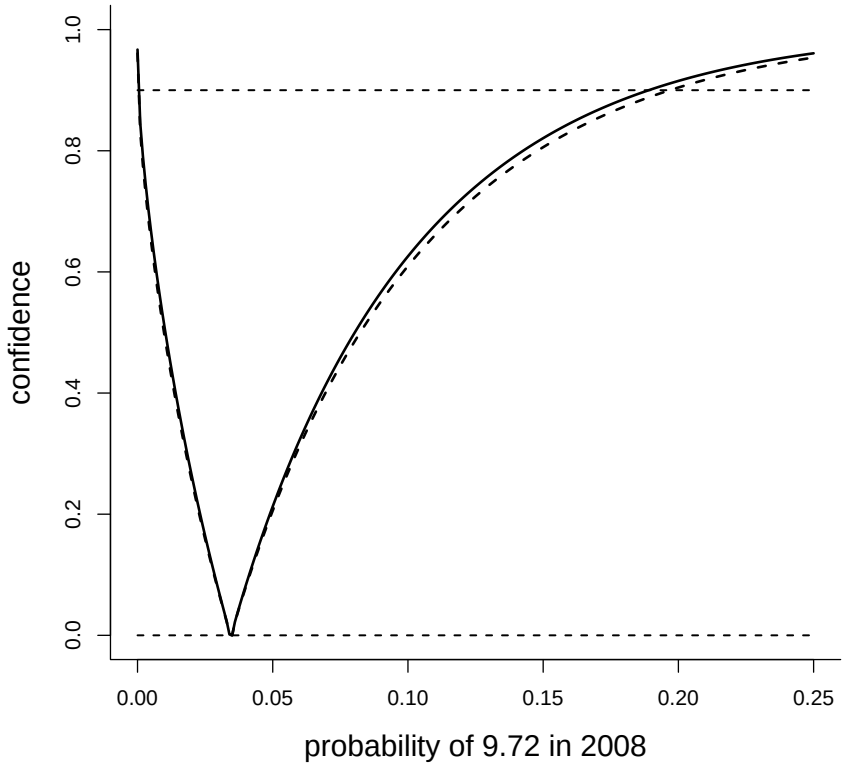

FIGURE 3.1. Confidence curve for the probability p that there would be a 100 m race of 9.72 or better, in the course of 2008, as seen from January 1 that year. The point estimate is 0.034, and the 90% confidence interval is $[0, 0.189]$. The dotted curve is a fine-tuned version of (3.4), via Bartletting.

of (a, σ), (ii) using the log-likelihood function $\ell_n(a, \sigma)$ to arrive at the profile and deviance function for p; and (iii) applying (3.4). The point estimate is $\widehat{p} = 0.034$, and a 90% confidence interval, read off from the figure, is $[0, 0.189]$. The natural skewness of the distributions involved makes this a more appealing method than applying the traditional $\widehat{p} \pm 1.645\,\widehat{\kappa}/\sqrt{n}$, say. The dotted line in Figure 3.1 is what here comes out of using a fine-tuning version of (3.4), namely $\Gamma_1(\mathrm{dev}_n(p)/(1+\widehat{\varepsilon}))$, with $1+\widehat{\varepsilon}$ indicating a Bartlett correction for the distribution of $\mathrm{dev}_n(p)$. In this particular case, $1+\widehat{\varepsilon} = 1.070$, and the curves are nearly identical. For a fuller discussion and the required detail, see Schweder & Hjort (2016, Section 7.4); see also De Blasi & Schweder (2017, this issue), where a novel correction method for fine-tuning of CDs is applied for this Bolt 2008 problem.

4. Risk, performance, optimality, and testing

Different ways of setting confidence intervals for the same parameter, and indeed more generally CDs, entail different performances. What is reasonably to be understood by 'good performance', for a confidence interval or a CD, is less clear than for point estimates or tests, where we are used to assessing root mean squared errors and power curves. Natural classes of loss functions may be put forward, with the risk functions as usual defined as the expected values of these losses, as a function of the the position in the parameter space. Such themes are developed in Schweder & Hjort (2016, Chs. 5, 7, 8). This development may be seen as a natural extension of classical optimality theory, for testing and for point estimation, as with the body of literature on Neyman–Pearson testing, etc.; see e.g. Lehmann (1959); Lehmann & Romano (2005).

Here we are content to quote and then illustrate a certain optimality theorem, which in particular can be put to use in models of the classical exponential structure. Suppose ψ is a focus parameter, and that the log-likelihood function for data can be expressed in the form

$$\ell(\psi, \lambda_1, \ldots, \lambda_k) = B\psi + \sum_{j=1}^{k} A_j \lambda_j - c(\psi, \lambda_1, \ldots, \lambda_k) + h(\mathcal{D}),$$

with nuisance parameters $\lambda_1, \ldots, \lambda_k$, with B and $A_1, \ldots, A_k$ functions of the data $\mathcal{D}$, and appropriate functions $c(\cdot)$ and $h(\cdot)$. In that case, the CD

$$C^*(\psi, \mathcal{D}) = P_\psi\{B \geq B_{\rm obs} \,|\, A_1 = A_{1,\rm obs}, \ldots, A_k = A_{k,\rm obs}\} \tag{4.1}$$

enjoys optimality properties with respect to a large class of loss functions for CDs; see Schweder & Hjort (2016, Ch. 5). That this $C^*(\psi)$ depends only on ψ, and not on the nuisance parameters, is part of the associated theorems.

Table 4.1. Lidocaine data: Death rates for two groups of acute myocardial infarction patients, in six independent studies, with control group associated with (m_0, y_0) and lidocaine treatment group with (m_1, y_1); from Normand (1999). See Figure 4.1.

m_1	m_0	y_1	y_0	z
39	43	2	1	3
44	44	4	4	8
107	110	6	4	10
103	100	7	5	12
110	106	7	3	10
154	146	11	4	15

To illustrate this, consider Table 4.1, summarising the number of deaths y_0 and y_1, with underlying sample sizes m_0 and m_1, in $k = 6$ independent studies, involving

acute myocardial infarction patients. Patients in the treatment group, associated with (m_1, y_1), received the drug lidocaine; the control group, listed under (m_0, y_0), did not; see Normand (1999). These are binomial studies, and modelling and analysis may proceed as in Schweder & Hjort (2016, Ch. 14.6). Since the probabilities are small, we choose a Poisson model for the present illustration. Our model takes

$$y_{j,0} \sim \text{Pois}(e_{j,0}\lambda_{j,0}) \text{ and } y_{j,1} \sim \text{Pois}(e_{j,1}\lambda_{j,1}), \quad \text{with} \quad \lambda_{j,1} = \gamma\lambda_{j,0}, \tag{4.2}$$

with exposure numbers $e_{j,0}$ and $e_{j,1}$ proportional to sample sizes $m_{j,0}$ and $m_{j,1}$, for $j = 1, \ldots, k$. Interest focuses on γ, which signals whether the drug use for these patients led to an increased death risk. The log-likelihood for study j takes the form

$$\begin{aligned}\ell_j &= -e_{j,0}\lambda_{j,0} + y_{j,0}\log\lambda_{j,0} - e_{j,1}\lambda_{j,0}\gamma + y_{j,1}(\log\lambda_{j,0} + \log\gamma) \\ &= y_{j,1}\log\gamma + z_j\log\lambda_{j,0} - e_{j,0}\lambda_{j,0} - e_{j,1}\lambda_{j,0}\gamma,\end{aligned}$$

with $z_j = y_{j,0} + y_{j,1}$. The optimality theorem applies, involving the distribution of $y_{j,1} \,|\, z_j$, which is seen to be a binomial $(z_j, e_{j,1}\gamma/(e_{j,0} + e_{j,1}\gamma))$. The optimal CD for γ, based on study j alone, is hence

$$C_j^*(\gamma, \mathcal{D}_j) = 1 - B(y_{j,1}; z_j, e_{j,1}\gamma/(e_{j,0} + e_{j,1}\gamma)) + \tfrac{1}{2}\, b(y_{j,1}; z_j, e_{j,1}\gamma/(e_{j,0} + e_{j,1}\gamma)),$$

with $\mathcal{D}_j$ signifying the data from source j, and with $B(\cdot; n, p)$ and $b(\cdot; n, p)$ denoting the cumulative and point distribution of a binomial (n, p). Here we are using the beneficial half-correction for discreteness, cf. Schweder & Hjort (2016, Ch. 3.7).

The $k = 6$ confidence curves $\text{cc}_j^*(\gamma, \mathcal{D}_j) = |1 - 2\, C_j^*(\gamma, \mathcal{D}_j)|$ for the risk inflation parameter γ coming out of this are seen in Figure 4.1 (the dashed curves). Also displayed is the overall optimal confidence curve for γ (the fatter, full curve), emerging from studying the combined log-likelihood,

$$\ell = \sum_{j=1}^k \ell_j = B\log\gamma + \sum_{j=1}^k z_j\log\lambda_{j,0} - \sum_{j=1}^k (e_{j,0} + e_{j,1}\gamma)\lambda_{j,0},$$

with $B = \sum_{j=1}^k y_{j,1}$, and where our optimality theorem leads to

$$\begin{aligned}C^*(\gamma, \mathcal{D}) = P_\gamma\{B > B_{\text{obs}} \,|\, z_1 = z_{1,\text{obs}}, \ldots, z_k = z_{k,\text{obs}}\} \\ + \tfrac{1}{2}\, P_\gamma\{B = B_{\text{obs}} \,|\, z_1 = z_{1,\text{obs}}, \ldots, z_k = z_{k,\text{obs}}\},\end{aligned} \tag{4.3}$$

with $\mathcal{D}$ denoting the full dataset. This is evaluated numerically by simulating a large enough number of B, for each γ on a grid of such values, from the distribution of a sum of k binomials with different sets of parameters.

The main interest for the analysis of the lidocaine dataset is the assessment of the risk inflation, if present, i.e. the degree to which the treatment for these patients leads to increased risk of death. In our Poisson model (4.2), this is measured via the parameter γ. The perhaps most traditional statistical approach is to test the null

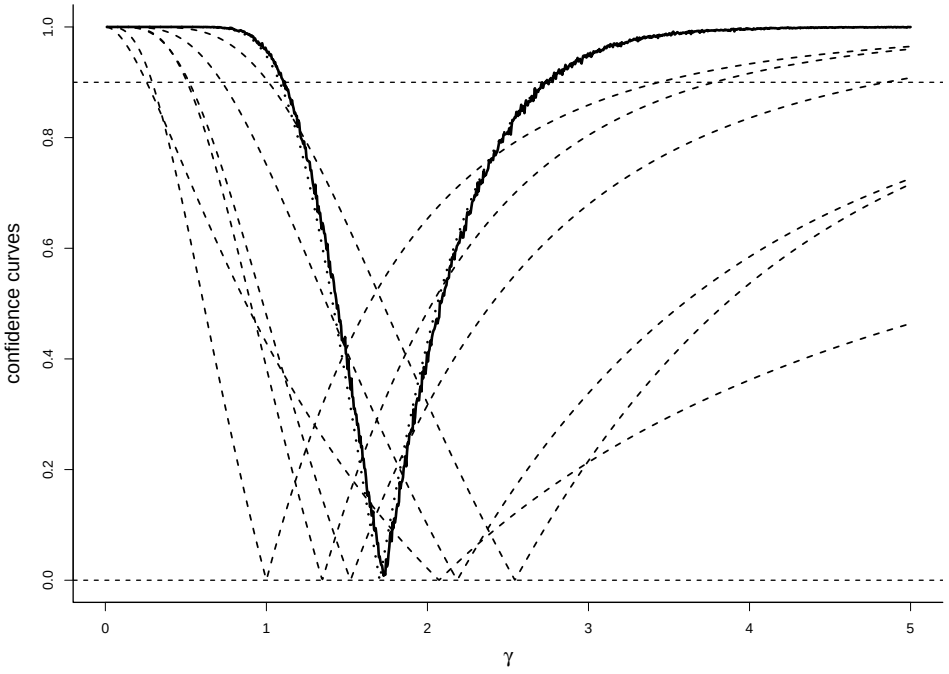

FIGURE 4.1. The dashed lines are the confidence curves for the risk inflation parameter γ from each of the six studies, from the model (4.2) with the lidocaine data of Table 4.1. The thick black curve is the optimal combined confidence curve, while the virtually identical dashed curve is the combined confidence curve based on the II-CC-FF methods of Section 5, without using the Poisson model properties per se.

hypothesis $H_0 \colon \gamma \leq 1$ versus the alternative that $\gamma > 1$. As Figure 4.1 reveals, there is often more information in conducting a full confidence curve analysis than in executing a test with its traditional yes-or-no answer at a certain level of significance, like the ubiquitous 0.05. The $\mathrm{cc}^*(\gamma, \mathcal{D})$ reveals not merely the overall point estimate 1.732, but the 0.95 interval $[1.023, 3.027]$, along with all other intervals; also, the configram clearly reveals the relative influence of each of the $k = 6$ separate information sources. The p-value can also be read off, from $p = C^*(1, \mathcal{D})$, the epistemic confidence that $\gamma \leq 1$; the value is 0.021.

As another illustration of this general point, about how CD analyses and plots often convey more statistical information than simple accept-or-reject answers from carrying out a test, consider Figure 4.2, with the left panel showing the increase in expected lifelength for women born in Norway (full curve), Sweden (dashed curve), Denmark (dotted curve), for the years 1960, 1970, 1980, 1990, 2000, 2010, 2015, from the website `worldlifeexpectancy.com/history-of-life-expectancy`. The growth in expected lifelength is amazingly linear, for this span of calendar time, and we view the data as three linear regressions, say $y_{i,j} = \alpha_i + \beta_i x_j + \varepsilon_{i,j}$ for

countries $i = 1, 2, 3$ and calendar years x_j represented by $j = 1, \ldots, 7$, and with error terms modelled as independent and $\varepsilon_{i,j} \sim \mathrm{N}(0, \sigma_i^2)$. We may query whether the regression slope coefficient β is the same for the three Scandinavian countries. Rather than merely testing the hypothesis H_0 that $\beta_1 = \beta_2 = \beta_3$, which would be standard (the point estimates are 0.140, 0.162, 0.144, with considerable overlap in their 0.95 confidence intervals), we address the question by modelling these three β coefficients as coming from a background $\mathrm{N}(\beta_0, \tau^2)$ model; hence H_0 is the same as $\tau = 0$. Using methods of Schweder & Hjort (2016, Ch. 13), we can derive and compute full CDs $C(\tau, \mathcal{D})$ for the spread parameter, displayed in the right panel. For the three male regressions (not shown here), the CD has a big point-mass 0.603 at $\tau = 0$; there is hence no reason to reject H_0, and confidence intervals at all reasonable levels start at zero (a 90% interval is $[0, C^{-1}(0.90, \mathcal{D})] = [0, 0.060]$). For the female regressions, however, there are noticeable differences in the three slopes underlying what is seen in the left panel; the p-value is $C(0, \mathcal{D}) = 0.021$, and a 90% interval is $[C^{-1}(0.05, \mathcal{D}), C^{-1}(0.95, \mathcal{D})] = [0.003, 0.053]$.

5. Data fusion via CDs

Meta-analysis is a well-developed area of theoretical and applied statistics, having to do with the comparison, assessment and perhaps ranking of different parameters across similar studies. Typical applications include analyses of different schools, or hospitals, or sport teams, or departments of statistics. Over the past few years these topics and methods have been expanded further, to account for the need to fuse together information from potentially very different types of sources, also in connection with the Data Science exploitation of Big Data. It is also important in various application areas to combine Bayesian with frequentist information, as discussed in Liu et al. (2015) and Lewis (2017, this issue); also, Grünwald (2017, this issue) touches on ways in which to handle multiple priors.

Suppose in general that data source y_j carries information about parameter ψ_j, for sources $j = 1, \ldots, k$. We wish to assess overall aspects of these ψ_j, perhaps aiming for inference concerning one of more functions $\phi(\psi_1, \ldots, \psi_k)$. Let us first assume that the ψ_j parameter is the same, across studies, and that the separate studies have led to CDs $C_j(\psi, y_j)$. A class of methods for combining these is as follows; see Singh et al. (2005); Xie & Singh (2013); Liu et al. (2014) and further references therein. Under the true value, $C_j(\psi, Y_j) \sim$ unif, from which follows $\Phi^{-1}(C_j(\psi, Y_j)) \sim \mathrm{N}(0, 1)$. With weights w_j nonrandom and satisfying $\sum_{j=1}^k w_j^2 = 1$, therefore,

$$\bar{C}(\psi, \mathcal{D}) = \Phi\Big(\sum_{j=1}^k w_j \Phi^{-1}(C_j(\psi, Y_j))\Big),$$

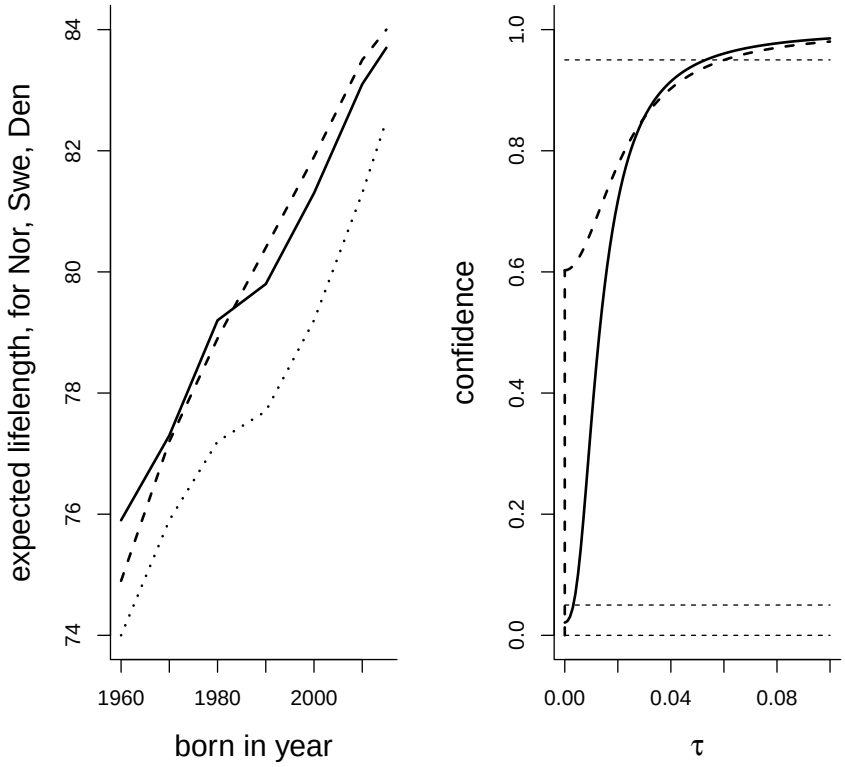

FIGURE 4.2. Left panel: The expected mean lifelength for women born in Norway (full curve), Sweden (dashed curve), Denmark (dotted curve), in calendar years 1960, 1970, 1980, 1990, 200, 2010, 2015. Right panel: The CD $C(\tau, \mathcal{D})$ for the spread parameter in the model $\beta_1, \beta_2, \beta_3 \sim \mathrm{N}(\beta_0, \tau^2)$ for the three regression slope parameters; for men (dashed curve, starting at 0.603 at zero) and for women (full curve, starting at 0.021 at zero). 95% intervals for τ are $[0, 0.060]$ for men and $[0.003, 0.053]$ for women.

with $\mathcal{D}$ the full dataset, is a CD for the common interest parameter ψ. Other start ingredients than the normal could also be put to use, but with less amenable convolutions and inversions. This is a versatile and broadly applicable method, but with some drawbacks. There are difficulties when estimated weights $\widehat{w}_j$ are used, and there is lack of full efficiency. In various cases, there are better CD combination methods, with higher confidence power; see the discussion in Cunen & Hjort (2016).

In clearly structured cases, as with several of the simpler meta-analysis setups, one can work with the full likelihood of the observed data, and deduce good CDs for interest parameters, see Schweder & Hjort (2016, Ch. 13). This does sometimes require the full set of raw data, however, which is often a too tall order. General

ways of dealing with data fusion with CDs are discussed and applied in Liu et al. (2015) and Hannig et al. (2017, this issue). Here we describe a more general setup for carrying out data fusion, via CDs, which we call the II-CC-FF paradigm; see Cunen & Hjort (2016). It is a more broadly applicable formulation of likelihood synthesis ideas first proposed, developed and applied in Schweder & Hjort (1996, 1997), in the specific context of population dynamics models for whale abundance.

II *Independent Inspection:* From data source y_j to estimate and confidence analysis, yielding a CD $C_j(\psi_j, y_j)$; $y_j \Longrightarrow C_j(\psi_j, y_j)$.

CC *Confidence Conversion:* From the CD to a confidence log-likelihood, $\ell_{c,j}(\psi_j)$; $C_j(\psi_j, y_j) \Longrightarrow \ell_{c,j}(\psi_j)$.

FF *Focused Fusion:* Using the combined confidence log-likelihood $\ell_c = \sum_{j=1}^k \ell_{c,j}(\psi_j)$ to construct a CD for the given focus $\phi = \phi(\psi_1, \ldots, \psi_k)$, perhaps via profiling, median-Bartletting, etc.; $\ell_c(\psi_1, \ldots, \psi_k) \Longrightarrow \bar{C}_{\text{fusion}}(\phi, \mathcal{D})$, with $\mathcal{D}$ denoting the combined dataset.

The FF step, which may also be described as the Summary of Summaries operation, will typically involve log-likelihood profiling and operations like (3.4), perhaps along with fine-tuning operations for increased accuracy. Sometimes the CC step is the more difficult one, since a clear translation from confidence to likelihood often would involve details of sampling design and protocol, etc. Under mild conditions, however, the *normal conversion* works well, which is

$$\ell_{c,j}(\psi_j) = -\tfrac{1}{2}\,\Gamma_1^{-1}(\mathrm{cc}_j(\psi_j, y_j)) = -\tfrac{1}{2}\{\Phi^{-1}(C_j(\psi_j, y_j))\}^2,$$

cf. (3.4).

For an illustration, let us go back to the lidocaine story of Table 4.1 and Figure 4.1, for which we have already displayed the optimal meta-analysis confidence curve (4.3) for the risk inflation parameter γ. We may however attempt the II-CC-FF recipe, which leads to a $\bar{C}_{\text{fusion}}(\gamma, \mathcal{D})$ just from converting the $k = 6$ individual $\mathrm{cc}_j^*(\gamma, \mathcal{D}_j)$ curves, using normal conversion, but without using the raw data per se, or any further knowledge of the underlying Poisson nature details of the modelling of the data. Amazingly, this FF fusion curve is almost indistinguishable from the $C^*(\gamma, \mathcal{D})$.

6. CDs in semi- and nonparametric situations

The CDs and confidence curves may also be constructed in non- and semiparametric situations. By arguments above, as long as there is an estimator $\widehat{\psi}$ for the required interest parameter ψ, with an associated limit distribution (typically normal), we may construct a CD for ψ based on that estimator. The empirical

likelihood may also be worked with to produce nonparametric CDs, in broad classes of situations, as developed and illustrated in Schweder & Hjort (2016, Ch. 11).

In some cases a more exact analysis is possible. A case in point is the following, where inference is required for the quantiles $\mu_p = F^{-1}(p)$ of a continuous and increasing distribution function, based on i.i.d. data $y_1, \ldots, y_n$. From the fact that the vector of ordered observations $y_{(i)}$ has the same distribution as that of $F^{-1}(u_{(i)})$, where the $u_{(i)}$ are the ordered sample from a uniform distribution, we can compute

$$s_n(a,b) = P\{Y_{(a)} \le \mu_p \le Y_{(b)}\} = P\{U_{(a)} \le p \le U_{(b)}\}$$

for each pair (a,b); see Schweder & Hjort (2016, Ch. 11). This can then be used to compute and display confidence curves $\mathrm{cc}(\mu_p, y)$ for each p of interest, as a nested sequence of confidence intervals. An illustration is given in Figure 6.1, where we give the full confidence curves for the 0.1, 0.3, 0.5, 0.7, 0.9 deciles for the birthweight distributions of boys and girls, born in Oslo, 2001–2008. The $\mathrm{cc}(\mu_p, y)$ curves tend to be slimmer where there is more data, i.e. around the median on this occasion.

In nonparametric situations there are often parameters which cannot be estimated at the usual $\sqrt{n}$ rate. Kim & Pollard (1990) give an overview of classes of cases for which the estimator $\widehat{\psi}$ for the focus parameter ψ in question exhibits cube-root convergence in distribution, i.e. $n^{1/3}(\widehat{\psi} - \psi) \to_d L$ for the appropriate (and non-normal) limit L. With appropriate extra efforts, involving the limit distribution and a consistent estimator for its variance, say $\widehat{\tau}$, one may construct CDs of the type $K(n^{1/3}(\psi - \widehat{\psi})/\widehat{\tau})$, perhaps along with further fine-tuning.

7. Robust CDs for parametric models

The standard theory for parametric models evolves around the use of likelihood methods. This is also at least partly the case for the theory and applications of CDs and confidence curves (Xie & Singh, 2013; Schweder & Hjort, 2016). The basic concepts and recipes are however not limited to likelihoods per se, and various robust alternatives may be worked with. To illustrate such general ideas and tools, suppose independent observations $y_1, \ldots, y_n$ stem from an unknown density g, and that one wishes to fit the data to a parametric model, say $f_\theta = f(\cdot, \theta)$, with θ a p-dimensional parameter. Consider

$$d_a(g, f_\theta) = \int \{f_\theta^{1+a} - (1 + 1/a) g f_\theta^a + (1/a) g^{1+a}\} \, \mathrm{d}y,$$

for a positive tuning parameter a. This is a divergence (nonnegative, and zero only if $g = f_\theta$), and for $a \to 0$ one finds the Kullback–Leibler divergence $\int g \log(g/f_\theta) \, \mathrm{d}y$ associated with the maximum likelihood method. The BHHJ method, from Basu et al. (1998); Jones et al. (2001), estimates θ by minimising an empirical version of

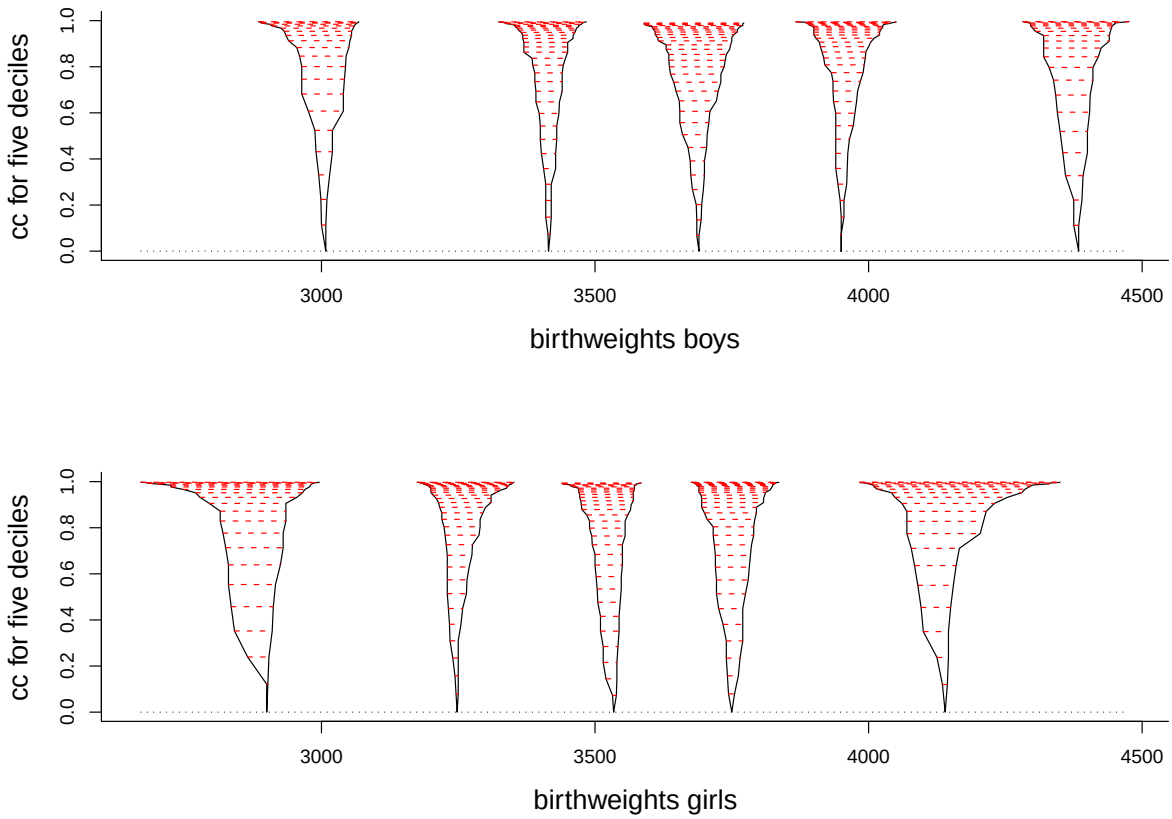

FIGURE 6.1. Confidence curves $\mathrm{cc}(q)$ for deciles 0.1, 0.3, 0.5, 0.7, 0.9 of birthweight distributions, for boys ($n = 548$) and girls ($n = 480$) born in Oslo 2001–2008.

d_a, namely $H_n(\theta) = \int f_\theta^{1+a}\,\mathrm{d}y - (1+a/n)\,n^{-1}\sum_{i=1}^n f(y_i,\theta)^a$. Setting the derivatives equal to zero, the BHHJ estimator is also the solution to the equations

$$n^{-1}\sum_{i=1}^n f(y_i,\theta)^a u(y_i,\theta) = \int f_\theta^{1+a} u_\theta\,\mathrm{d}y,$$

where $u_\theta(y) = u(y,\theta) = \partial \log f(y,\theta)/\partial\theta$ is the score function for the model. Contributions from data points with low probability under the model thus get weighted down. The method is a successful robustification of the maximum likelihood strategy (also in regression setups and other models more elaborate than the i.i.d. situation considered here), earning bounded influence functions at the expense of a very mild loss of efficiency under perfect model conditions, if a is small.

The present point we wish to make is that the criterion function H_n, used to find the BHHJ estimator and its approximate multinormal distribution, can also be

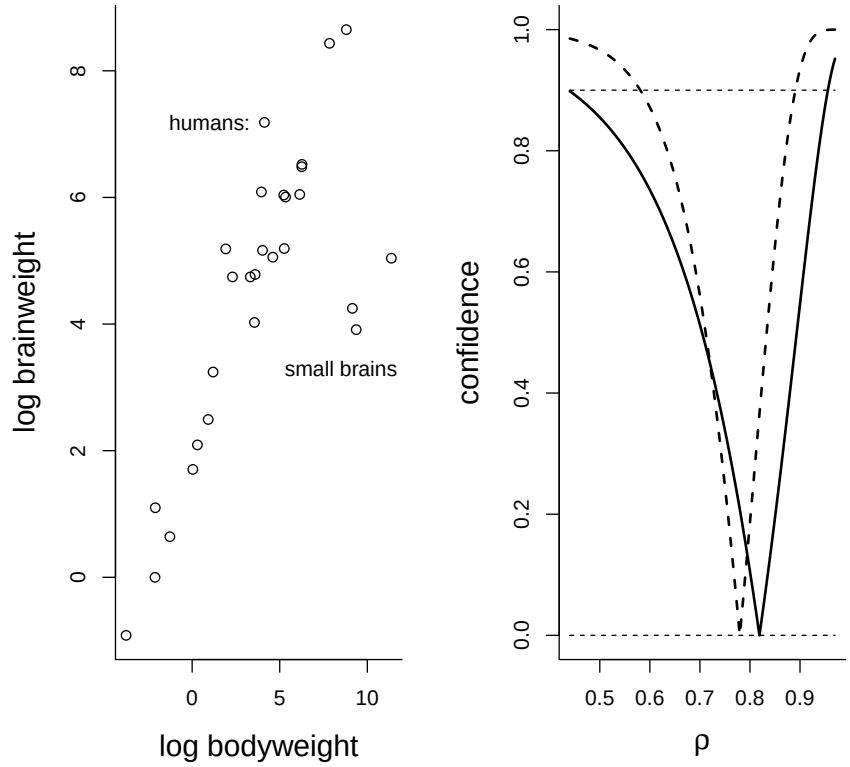

FIGURE 7.1. Left panel: with average x_0 bodyweight (in kg) and average brainweight y_0 (in g), for 28 species of land animals, the plot gives $(x, y) = (\log x_0, \log y_0)$. Right panel: two confidence curves for the correlation coefficient ρ, based on maximum likelihood (estimate 0.779) and one using the robust BHHJ method (estimate 0.819).

profiled, leading to confidence curves for focus parameters. With $\psi = \psi(\theta)$ such a focus parameter, the BHHJ estimator is $\widehat{\psi} = \psi(\widehat{\theta})$, and we form $H_{n,\text{prof}}(\psi) = \min\{H_n(\theta)\colon \psi(\theta) = \psi\}$ and then the associated deviance function,

$$D_n(\psi) = 2n\{H_{n,\text{prof}}(\psi) - H_{n,\text{prof}}(\widehat{\psi})\} = 2n\{H_{n,\text{prof}}(\psi) - H_{n,\min}\}.$$

With arguments along the lines of Schweder & Hjort (2016, Ch. 2.4, Appendix A.6), one may establish that $D_n(\psi_0) \to_d k\chi^2_1$, at the appropriate least false parameter value $\psi_0 = \psi(\theta_0)$, with θ_0 minimising the distance $d_a(g, f_\theta)$ from the true g to the parametric model. Here k is a certain extra factor which may be estimated consistently from the data. This leads to the robust confidence curve $\text{cc}(\psi, \mathcal{D}_n) = \Gamma_1(D_n(\psi)/\widehat{k})$ (again with $\mathcal{D}_n$ denoting the dataset), in generalisation of (3.4).

This machinery works also for multidimensional data. Figure 7.1 relates to an illustration of this, where we have studied the dataset *Animals* in R, with (x_0, y_0)

equal to average bodyweight (in kg) and average brainweight (in g) for $n = 28$ species of land animals. On the log-and-log scale of $(x, y) = (\log x_0, \log y_0)$, intriguingly, the points nearly form a linear regression structure; the deviants, from this perspective, are the big-brained humans, and the big-bodied small-brained Brachiosaurus, Triceratops, and Diplodocus (left panel). Our chosen focus, for this illustration, is the correlation coefficient ρ. The estimate is 0.779, based on all 28 species, but a much higher 0.960 if we remove the three small-brained just mentioned. We fit the five-parametric binormal model to the data, first using maximum likelihood analysis, then the BHHJ method with $a = 0.105$; this value makes data pairs an average distance away from the centre, as measured by the Mahalanobis distance, be downweighted 10% (and pairs further away from the centre will be downweighted more). This value also ensures good robustness. The two confidence curves are displayed in the right panel; the maximum likelihood version points to $\widehat{\rho} = 0.779$ whereas the BHHJ method has $\widehat{\rho}_a = 0.819$. The robust 90% confidence interval is $[0.441, 0.955]$. Importantly, these two confidence curves do not assume that the binormal model holds. In this particular application the robust BHHJ method leads to a somewhat broad confidence curve, since the method attempts to fit a somewhat non-homogeneous dataset to a single binormal density. For larger values of a, the BHHJ estimation method will indirectly downweight the three outliers more fully, and the correlation estimate will come closer to 0.960.

8. Bayes versus CDs

The Holy Grail of statistics Brad Efron alludes to is to enjoy the Bayesian omelet without breaking the Bayesian eggs (Efron, 2010). It was the non-existence of a non-informative prior which led Fisher to fiducial distributions. That a CD is 'posterior' without any prior is its main selling points.

Bayes' formula is of course true, and the Bayesian posterior is the correct updating of a trustworthy prior. But problems arise when there is no trust in the chosen prior, or when there are more than one legitimate priors.

With much data the CD will tend to be close to the Bayesian posterior, by various Bernshteĭn-von Mises type theorems (see e.g. Hjort et al. (2010, Introduction)). This might also happen in some cases with moderate and small data, particularly when a Jeffreys prior is used. A case of the latter is seen in Lewis (2017, this issue), where he develops both a Byesian posterior and a CD for the climate sensitivity. They are seen to be indistinguishable.

A marginal posterior distribution might be misleading, as illustrated by the so called length problem: With independent $Y_i \sim \mathrm{N}(\mu_i, 1)$ for $i = 1, \ldots, m$, the marginal posterior for $\psi = \|\mu\| = (\sum_{j=1}^m \mu_j^2)^{1/2}$ based on a flat (Jeffreys) prior for

the m mean parameters is biased in the frequentist sense that its credibility intervals will not have correct coverage probabilities (Schweder & Hjort, 2016, Section 9.4). The distribution is actually shifted to the right relative to ψ and more so the larger m is. This is also a problem for the marginal of Fisher's joint fiducial distribution, which is not a CD. A similar bias inherent in Bayes setups is noted for change-point assessments in higher dimensions, in Cunen et al. (2017a, this issue).

Potential bias seems not to be a concern for most Bayesians. When your prior is to be updated from new data, you get the posterior you get, and performance in repeated applications is seen as irrelevant. Frequently the model is complex and the model parameter of substantial dimensions, however, as in the length problem. As a more realistic example consider the parameter θ_1 of interest to Sims (2012) in his Nobel Memorial Prize in Economic Sciences acceptance lecture, where it is also argued that $\theta_1 \geq 0$ on a priori grounds. The model is a linear simultaneous equations model for macroeconomic data. The chosen prior for coefficients, including θ_1, is flat. Since the unrestricted maximum likelihood $\widehat{\theta}_1$ is negative the posterior is shifted to the right of the CD for θ_1. The latter has actually a point mass of 0.90 at zero (when the restriction is $\theta_1 \geq 0$), while Sims's posterior has all its mass on the positive values; see Schweder & Hjort (2016, Section 14.4).

Bayesian methods are very often used. It is thus a bit odd that performance in repeated applications is mostly neglected. Bias and other frequentist properties are however of concern to some Bayesians. The invariance of the posterior based on Jeffreys priors, to transformations of the model parameter, will, as noted above, make the posterior nearly or exactly a CD. In cases with a parameter ψ of interest, of lower dimension than the model parameter, the objective Bayesian uses a reference prior (Berger & Bernardo, 1992; Berger & Sun, 2008) tailored to ψ. This is parallel to confidence inference where new calculations are needed for each ψ. The posterior based on a reference prior aims at having correct coverage probabilities for its credibility regions in repeated applications. The CD has the same aim – it is actually its defining property. The realised posterior and also the CD are understood as epistemic probability distributions for ψ (Schweder, 2017, this issue). To be a CD might actually be the gold standard for an epistemic probability distribution for a parameter of interest, at least for the objective Bayesian; cf. Veronese & Melilli (2017, this issue) and also Grünwald (2017, this issue). Fraser (2011) actually suggests that Bayes posterior distributions are just quick and dirty confidence distributions.

We note that CDs may be constructed not only for parameters of models, but also for not-yet-seen random variables, as in prediction contexts. There are again similarities with Bayesian approaches; see Schweder & Hjort (2016, Ch. 12) and Shen et al. (2017, this issue).

An important virtue of the Bayesian approach is its coherence. Ordinary probability calculus applies to posterior distributions. For CDs some probability calculations yield new CDs, while others do not. When the prior has been set up, the challenge is to calculate the well-defined posterior, say by Markov Chain Monte Carlo. The technical virtuosity of current days Bayesians is really impressive, and has led to sensible analyses of complex data in many areas. The machinery for confidence inference is by far less developed. Significant applications in science are still rather few (but see Cunen et al. (2017b), where the main findings were communicated to the Scientific Committee of the International Whaling Commission via confidence curves). Software for CDs needs further development in good packages in order for the dissemination to gain momentum.

Robert (2013) points out that a CD is in essence just a representation of a nested family of confidence regions, and as such not particularly novel, per se. The emphasis on CDs as distributions on par with Bayesian posteriors might however be a rather novel insight, distributions that "provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)" (Cox, 2013). There is also scope for novel and streadily more impressive uses of CDs for data fusion, when information sources are more diverse than in the typical meta-analysis setups; see Xie & Singh (2013); Liu et al. (2015); Cunen & Hjort (2016).

9. The present collection of papers

Articles appearing in the present Special Issue have been mentioned above, in the relevant contexts. Here we offer just a few more comments to help readers navigate through these contributions and to see connections between them.

Each of the contributions Hannig et al. (2017), Lindqvist & Taraldsen (2017), Taraldsen & Lindqvist (2017) deal with fiducial and generalised fiducial inference questions, also with relevance for the eternal comparison with Bayesian constructions. Articles De Blasi & Schweder (2017), Veronese & Melilli (2017) are partly concerned with fine-tuning mechanisms for the constructions of CDs, with further connections to so-called objective Bayes. Several contributions are involved with the important topic of combining information across diverse sources, sometimes called data fusion: Hannig et al. (2017), Grünwald (2017), Cunen et al. (2017a), Lewis (2017). The latter paper is also a well-argued contribution to the always hot topic of climate research, where there typically are very different sources of information. One of the challenges, worked with by Lewis, is that of combining summaries reached by Bayesian and frequentist perspectives; see also Liu et al. (2015) and Cunen & Hjort (2016).

When constructing CDs, and more generally machineries aspiring to deliver posteriors without priors, one is often close to the more fundamental issues and ideas of how probability can or should be defined and interpreted, cf. again the half-eternal half-disagreements between Bayesians and frequentists. This is also touched on in the essay Schweder (2017), somewhat indirectly in Hannig et al. (2017), Martin (2017), and by Grünwald (2017). In certain application areas it might be natural to interpret confidence in the language of epistemic probabilities, as argued by Schweder (2017); see in this connection also Helland (2018).

The Shen et al. (2017) and Cunen et al. (2017a) articles are occupied with respectively prediction issues, e.g. for time series models, and with estimating and assessing change-points and regime-shifts, in settings with discrete data. Applications in the latter paper involve finding when Author B took over for Author A, in the world's first ever novel (from 1460), and searching for a regime-shift in a complex fisheries model. That paper also contains novel goodness-of-fit tests for checking whether a probability distribution has remained constant over a stretch of time.

10. Concluding remarks

We started out discussing the Holy Grail of parametric inference (Efron, 2010), that of reaching well-defined posteriors for interest parameters without putting up priors. We argue that the CDs are the answer, or part of the answer. In particular, in classes of clear-cut situations, in exponential class type models where the broad optimality theory of Schweder & Hjort (2016, Chs. 5, 7, 8) applies, the optimal CD provides what a rational statistician ought to believe about the unknown parameter, given the model and the data.

There are of course many remaining issues and obstacles for our profession to work with and perhaps slowly sort out, through the statistical symbiotic machineries of good theory and solid practice. Let us mention some of these.

The serious study of CDs and indeed related themes often enough touches the fundamental issues of what probability is, or ought to be. This has of course been discussed inside and outside academics since around 1665 (see the engaging account by Hacking (1975)), and perhaps also the modern statisticians and data scientists need to accept that there are several valid notions, living if not always comfortably side-by-side: The clear aleatory probability; the subjective used by strands of Bayesians; the epistemic; and yet further cousins and hybrids, like Dempster–Shafer belief functions (Dempster, 2008; Martin, 2017). There is still a need for a better axiomatic theory for epistemic probability, and its connections to likelihood theory and related issues.

On the technical side there is scope for important work along the lines of further fine-tuning of approximate CDs to deliver more accurate coverage, which in the language of CDs, and of this article, may be described as calibrating the CD such that $C(\psi(\theta_0), Y)$ has a distribution close to the unifom, where $\psi = \psi(\theta)$ is the focus parameter and $\psi_0 = \psi(\theta_0)$ is the implied true focus value. The basics of a list of relevant techniques is in Schweder & Hjort (2016, Chs. 7-8), but there is more to do, also in situations with multimodal log-likelihoods, with growing dimension p compared to sample size n, etc. A class of alternatives to the profiling involved in (3.4) is via the operation of integrating out other parameters, which may work well also from the frequentist viewpoint (Berger et al., 1999).

Several of the issues that bothered Fisher's contemporaries, when they hesitated to embrace his fiducial inference ideas in the 1930ies and 1940ies, have to do with the usual probability machinery not being applicable in general. The distribution $H(|\mu|) = C(|\mu|) - C(-|\mu|)$, for example, is usually not a CD for $|\mu|$ when C is a CD for μ; see Schweder & Hjort (2013) for this and similar examples. Schweder & Hjort (2016, Ch. 6) refer to various attempts to figure out when a distribution obtained by ordinary probability calculus from a CD is itself a CD. These theories are far from complete. It is perhaps more fruitful to study when good approximate CDs can be obtained by ordinary calculus than to try to develop a calculus for fiducial distributions and CDs.

An interesting research direction is that of matching good CDs, in particular those known to be optimal, with corresponding priors, i.e. for so-called objective Bayes. This is also touched on in Veronese & Melilli (2017, this issue). There are situations, e.g. those with bounded parameter spaces, where optimal CDs apparently have no matching prior. An instance of this is our disagreement with what Sims (2012) claimed in his Bayesian-flavoured Nobel Memorial Prize acceptance speech; see Schweder & Hjort (2016, Ch. 14.4).

An important objection from the Bayesian camp is that CDs are usually not well-defined when the model has been arrived at via a preliminary model selection step. Further decisions are usually needed to guide the calculations. Robert (2013) calls this 'ad hockery'. But is it more ad hockery than choosing the prior in the absence of solid prior information? Also, progress can be expected regarding working out good refinements for CDs after model selection, using the machinery developed in Hjort & Claeskens (2003) and Hjort (2014) for frequentist model averaging.

We have seen how p-values often can be seen as special components of a bigger CD picture, and these connections can be worked out more fully, both for enhanced interpretation and for better assessments of well-understood hypotheses; cf. the still

ongoing debate on the uses and many misuses of p-values (Wasserstein & Lazar, 2016).

The CDs and confidence curves should find more use in the contemporary world of Big Data, Data Science and Machine Learning, also for conveying summary information about the most pertinent issues, based on often complex and massive background data. Instrumental here is also the task of combining and fusing together information across very diverse sources, where what we describe above as the II-CC-FF paradigm ought to be harnessed further.

There is already a body of literature and results on the performance and optimality of classes of CDs, cf. again Schweder & Hjort (2016, Ch. 5). Aspects of this theory ought to be extended from CDs to confidence curves, as there are natural cases where the $\mathrm{cc}(\psi, y)$ is the more fundamental notion of confidence; cf. the Fieller problem, situations with multimodal log-likelihoods, etc. There is similarly a need for more work and better insights for confidence curves in higher dimensions. Finally we point to the correlated worlds of CDs, inferential models and generalised fiducial inference, where there is a need to sort out better when the approaches agree, and where they might not.

Acknowledgements

As guest editors we are first of all grateful for the high-quality and patience-demanding efforts of all the authors, also during the required rounds of revisions and cross-referencing. We are also indebted to the Norwegian Research Council for its funding of the FocuStat project (Focus Driven Statistical Inference With Complex Data) at the Department of Mathematics, University of Oslo; this was also instrumental for the group being able to host a series of international workshops in Oslo, including *Inference With Confidence* in May 2015. Thanks are in particular due to Céline Cunen, whose efforts in connection with the full process of handling and sharpening the contributions to the present Special Issue were crucially helpful, along with the help and insights from a select group of conscientious referees. We finally acknowledge with gratitude the active help of Holger Dette and the other editors of the JSPI in creating a high-quality special issue.

References

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. London: Champmann & Hall.

Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. C. (1998). Robust and efficient estimation by minimising a densithy power divergence. *Biometrika* **85**, 549–559.

Berger, J. O. & Bernardo, J. M. (1992). On the development of reference priors [with discussion and a rejoinder]. In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds. Oxford: Oxford University Press, pp. 35–60.

Berger, J. O., Liseo, B. & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–26.

Berger, J. O. & Sun, D. (2008). Objective priors for the bivariate normal model. *Annals of Statistics* **36**, 963–982.

Brazzale, A. R. & Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science* **23**, 465–484.

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge: Cambridge University Press.

Cox, D. R. (2013). Discussion of M. Xie and K. Singh's paper, 'Confidence distributions, the frequentist estimator of a parameter: a review'. *International Statistical Review* **81**, 40–41.

Cunen, C., Hermansen, G. & Hjort, N. L. (2017a). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Cunen, C. & Hjort, N. L. (2016). Combining information across diverse sources: The II-CC-FF paradigm. In *Proceedings from the Joint Statistical Meeting, Chicago 2016*. American Statistical Association, pp. 138–153.

Cunen, C., Walløe, L. & Hjort, N. L. (2017b). Decline in energy storage in Antarctic Minke whales during the JARPA period: Assessment via the Focused Information Criterion (FIC). *Reports of the Scientific Committee of the International Whaling Commision SC/67A/EM/04*.

De Blasi, P. & Schweder, T. (2017). Confidence distributions from likelihoods by median bias correction. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Dempster, A. P. (2008). The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning* **48**, 365–377.

Efron, B. (1998). R.A. Fisher in the 21st century [with discussion and a rejoinder]. *Statistical Science* **13**, 95–122.

Efron, B. (2010). The future of indirect evidence. *Statistical Science* **25**, 145–157.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26**, 528–535.

Fisher, R. A. (1932). Inverse probability and the use of Likelihood. *Proceedings of the Cambridge Philosophical Society* **28**, 257–261.

Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society, Series A* **139**, 343–348.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **6**, 391–398.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? [with discussion and a rejoinder]. *Statistial Science* **26**, 249–316.

Grünwald, P. (2017). Safe probability. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. New York: Wiley.

Hald, A. (2006). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935*. Sources and Studies in the History of Mathematics and Physical Sciences. Berlin: Springer.

Hannig, J., Feng, Q., Iyer, H., Wang, C. M. & Liu, X. (2017). Fusion learning for inter-laboratory comparisons. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Hannig, J., Iyer, H., Lai, C. S. & Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association* **111**, 1346–1361.

Helland, I. (2018). *Epistemic Processes*. Berlin: Springer-Verlag.

Hjort, N. L. (2014). Discussion of Efron's 'Estimation and accuracy after model selection'. *Journal of the American Statistical Association* **110**, 1017–1020.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators [with discussion]. *Journal of the American Statistical Association* **98**, 879–899.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Jones, M. C., Hjort, N. L., Harris, I. R. & Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

Kim, J. & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics* **18**, 191–219.

Lehmann, E. & Romano, J. P. (2005). *Testing Statistical Hypotheses [3rd ed.]*. New York: Wiley.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.

Lewis, N. H. (2017). Combining independent bayesian posteriors into a confidence distribution, with application to estimating climate sensitivity. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Lindqvist, B. H. & Taraldsen, G. (2017). On the proper treatment of improper distributions. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Liu, D., Liu, R. Y. & Xie, M. (2014). Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events. *Journal of the American Statistical Association* **109**, 1450–1465.

Liu, D., Liu, R. Y. & Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association* **110**, 326–340.

Martin, R. (2017). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Martin, R. & Liu, C. (2015). *Inferential Models: Reasoning with Uncertainty*. Toronto: CRC Press.

Normand, S.-L. T. (1999). Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321–359.

Robert, C. (2013). Discussion of M. Xie and K. Singh's paper, 'Confidence distributions, the frequentist estimator of a parameter: a review'. *International Statistical Review* **81**, 52–56.

Schweder, T. (2017). Confidence is epistemic probability for empirical science. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Schweder, T. & Hjort, N. L. (1996). Bayesian synthesis or likelihood synthesis – what does Borel's paradox say? *Reports of the International Whaling Commission* **46**, 475–479.

Schweder, T. & Hjort, N. L. (1997). Indirect and direct likelihoods and their synthesis – with an appendix on minke whale dynamics. Tech. rep., Department of Mathematics, University of Oslo.

Schweder, T. & Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309–322.

Schweder, T. & Hjort, N. L. (2003). Frequentist analogues of priors and posteriors. In *Econometrics and the Philosophy of Economics: Theory-Data Confrontation in Economics*, B. Stigum, ed. Princeton University Press, pp. 285–217.

Schweder, T. & Hjort, N. L. (2013). Discussion of M. Xie and K. Singh's paper, 'Confidence distributions, the frequentist estimator of a parameter: a review'. *International Statistical Review* **81**, 56–68.

Schweder, T. & Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge: Cambridge University Press.

Shen, J., Liu, R. & Xie, M.-g. (2017). Prediction with confidence: A general framework for prediction. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Sims, C. A. (2012). Statistical modeling of monetary policy and its effects [Nobel Memorial Prize in Economic Sciences Lecture]. *American Economic Review* **102**, 1187–1205.

Singh, K., Xie, M. & Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Annals of Statistics* **33**, 159–183.

Student (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

Taraldsen, G. & Lindqvist, B. H. (2017). Conditional fiducial models. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Veronese, P. & Melilli, E. (2017). Fiducial, confidence and objective bayesian posterior distributions for a multidimensional parameter. *Journal of Statistical Planning and Inference* **xx**, xx–xx.

Wasserstein, R. W. & Lazar, N. A. (2016). The ASA's statement on p-value: context, process, and purpose. *American Statistician* **70**, 129–133.

Xie, M. & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review [with discussion and a rejoinder]. *International Statistical Review* **81**, 3–39.