Integrating confidence intervals, likelihoods and confidence distributions

Tor Schweder[*], Nils L. Hjort
University of Oslo, Norway

Abstract

Independent likelihoods are integrated by multiplication. We argue that confidence distributions should also be integrated by their related likelihoods (confidence likelihoods), and that confidence intervals should be integrated by first estimating a related approximate confidence distribution, and then integrate their confidence likelihoods.

1 Introduction

With confidence distribution we mean a nested family of confidence regions $R_\alpha(X)$. $\alpha \in [0,1]$ is the confidence level and $\theta \in \Theta$ is the parameter in the model for the data X, assumed continuously distributed. The confidence distribution might be cast in terms of the *confidence curve* $cc(\theta; X) : \Theta \to [0,1]$ such that $R_\alpha(X) = \{\theta : cc(\theta; X) \leq \alpha\}$. Confidence curves have the properties

- $min_\theta cc(\theta; x) = 0$ for all outcome of the data x
- $cc(\theta_0; X) \sim U[0,1]$, i.e. the uniform distribution on the unit interval, when θ_0 is the true value of the parameter.

By the first property, it is assumed that there are non-empty confidence regions at all levels of confidence. The second property makes the level sets of the confidence curves confidence regions. Indeed, $P_{\theta_0}(R_\alpha(X) \ni \theta_0) = P_{\theta_0}(cc(\theta_0; X) \leq \alpha) = \alpha$.

When the confidence curve is flat at zero, every confidence set contains the whole parameter space. The confidence curve $cc(\theta) \equiv 0$ is consequently non-informative.

When $\Theta = \mathbf{R}$ and the confidence regions are left-open intervals $< -\infty, r_\alpha(X)]$, the confidence curve is the cumulative distribution function $H(\theta; X)$ with confidence α-quantiles $r_\alpha(X) = H^{-1}(\alpha; X))$. When the confidence regions are equi-tailed confidence intervals $< r_{\alpha/2}, r_{1-\alpha/2} >$, suppressing the dependence on the data X, the confidence curve is $cc(\theta) = |1 - 2H(\theta)|$. The confidence regions could also be ellipsoids, say for a mean vector θ of dimension p, $R_\alpha = \{\theta : (\theta - \hat{\theta})'\Sigma^{-1}(\theta - \hat{\theta}) \leq \Gamma_p^{-1}(\alpha)$

[*]Corresponding author: Tore Schweder, email: tore.schweder@econ.uio.no

where $\hat{\theta} \sim N(\theta, \Sigma)$ and Γ_p is the cdf of the χ^2_p distribution. In this case $cc(\theta) = \Gamma_p((\theta - \hat{\theta})'\Sigma^{-1}(\theta - \hat{\theta}))$. Other forms are possible.

As an example of a more unfamiliar confidence distribution take the simplest form of the Fieller-Cressie problem (Schweder and Hjort, 2013) of $\theta = \mu_1/\mu_2$ when $X_i \sim N(\mu_i, \sigma_i^2)$ and independent. The variances are assumed known. The profile deviance is

$$D(\theta) = \frac{(\theta X_2 - X_1)^2}{\theta^2\sigma_2^2 + \sigma_1^2}.$$

Since $D(\theta; X) \sim \chi^2_1$, $cc(\theta; X) = \Gamma_1(D(\theta; X))$. The observed confidence curve has a peak of $p = \Gamma_1(\frac{x_1\sigma_2^2 + x_2\sigma_1^2}{\sigma_1^2\sigma_2^2})$ at $\theta = -x_2\sigma_1^2/(x_1\sigma_2^2)$, asymptotes at $q = \Gamma_1(x_2^2/\sigma_2^2)$ at $\pm\infty$, and a minimum of 0 at $\theta = \hat{\theta} = x_1/x_2$. The confidence region R_α is thus the entire real line for $\alpha \geq p$, two half-open intervals stretching out to infinity at both ends for $q \leq \alpha < p$, and a finite interval for $\alpha < q$.

Neyman (1941) saw the connection between his confidence intervals and the fiducial distributions of Fisher (1930). He was less interested in the likelihood, $L(\theta; x) = |\frac{d}{dx}H(\theta; x)| = f(x; \theta)$ obtained from the cumulative confidence/fiducial distribution function that Fisher noted in the case of a one-dimensional statistic X and parameter θ. A likelihood, or an approximate likelihood, obtained from a confidence distribution is called a *confidence likelihood* (Schweder and Hjort, 2013). Let $\ell_c(\theta)$ be a confidence log likelihood, and $D_c(\theta) = 2(\max_t(\ell(t) - \ell(\theta))$ the confidence deviance. Approximate confidence deviances might be obtained from confidence curves by the χ^2 quantile transformation

$$D_c(\theta) = \Gamma_p^{-1}(cc(\theta))$$

Wilks' theorem (Schweder and Hjort, 2013) provides a rational for this construction. When X is a large sample from a regular distribution of parameter (θ, τ), where θ is the p-dimensional parameter of interest and τ is a nuisance parameter, the profile deviance D^{prof} evaluated at θ_0 is asymptotically χ^2_p distributed. Thus $cc(\theta) = \Gamma_p(D^{prof}(\theta))$ is asymptotically a confidence curve, and $D_c(\theta) = \Gamma_p^{-1}(cc(\theta))$ is the profile deviance.

Efron (1993) proposed $L_i(\theta) = \exp(-\frac{1}{2}\Phi^{-1}(H(\theta))^2)$, called the implied likelihood, as the natural likelihood obtained from a cumulative distribution H for the parameter. The implied likelihood is the same as our confidence likelihood.

With limited data or an irregular model, the χ^2 approximation might be too inaccurate. It is good practice to investigate the distribution of the deviance used for constructing the confidence curve by simulation or otherwise, and use a better probability transform, if available. It is also wise to investigate to what extent the nuisance parameter τ is a problem. By Wilks' theorem the profile deviance evaluated at θ_0 has asymptotically a χ^2 distribution independent of τ in the regular case. Data are however always of limited size, and the profile deviance might have to be adjusted when the χ^2 approximation is too inaccurate, say by methods discussed in Brazzale and Davison (2008). Bootstrapping and utilizing the normal transformation model might also be a way forward (Shweder and Hjort, 2013).

Since the profile deviance is exactly χ_1^2-distributed in the simple Fieller-Cressy model for the quotient of two normal means, it is recovered by $\Gamma_1^{-1}(cc(\theta))$. The same happens in linear normal models when the variance is known. For large data the profile likelihood is recovered when the model is smooth, at least approximately.

Here we shall assume that the χ^2 approximation is sufficiently accurate, and that the nuisance parameter is not a serious problem.

2 Integrating confidence distributions

Given k independent confidence distributions for the same parameter θ, how should they be integrated to a combined confidence distribution for θ?

As an example, assume the parameter of interest to be the spatial density of points, e.g. whales, in an area of the ocean. Whales are counted by line-transect sampling. for each transect leg the spatial density is estimated as the ratio of the number of observed whales to the estimated size of the effectively observed area along the transect leg. For each transect leg there are thus a Fieller-Cressie confidence curve, assuming for simplicity normal distribution and independence of both numerators and denominators. How should these confidence curves be integrated?

Our proposal is to integrate the independent confidence curves by their confidence likelihoods. The confidence deviance of confidence curve cc_i is $\Gamma_1^{-1}(cc_i(\theta))$. They sum to twice the the combined confidence log likelihood. The proposed combined confidence curve is thus

$$cc_{int}(\theta) = G\{\sum_i^k \Gamma_1^{-1}(cc_i(\theta)) - min_t(\sum_i^k \Gamma_1^{-1}(cc_i(t))\}, \qquad (1)$$

where G is the distribution function of the combined pseudo-deviance in the argument.

In case the the pseudo-deviance is a profile deviance, it should nearly have a χ_1^2 distribution by Wilk's theorem. The distribution of the combined pseudo-deviance must usually be estimated by simulation.

In case of the Fieller-Cressie example of combining $k = 10$ confidence curves, with $\theta_0 = 1$ and unit mean and variance in both numerators and denominators, the combined pseudo-deviance has the distribution of $\chi_1^2/0.58$, as found by simulation. For simulated data, the first of the 10 confidence curves is shown in black, and the combined confidence curve is in read, see Figure 1.

It may happen that some of the individual confidence curves come out as non-informative. They are then identically zero, and do not contribute to the combined confidence curve. This good property is not shared with Bayesian integrative methods. Nor with the method proposed by Singh et al. (2005) based on adding the normal scores (or some other quantile transforms) of the individual confidence distributions. The problem here is that there are no such thing as a non-informative cumulative confidence distribution function or posterior Bayesian distribution.

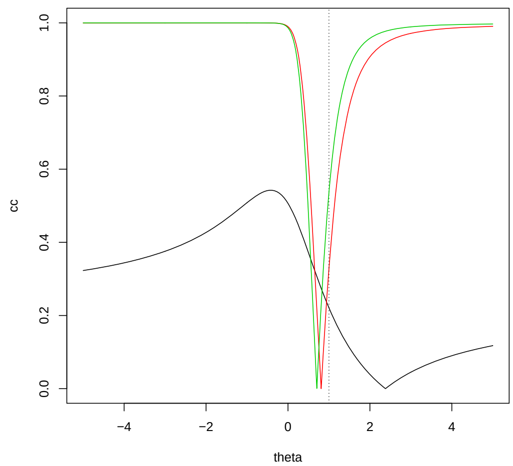

Figure 1: Fieller-Cressie confidence curves. The black curve is for the first of $k = 10$ ratios to merge. The read curve is the combined curve (1). The green curve is obtained from the summed numerators and denominators.

The combined confidence curve is invariant to monotonic transformations. This property is shared with the method of Singh et al. (2005). But not with the Bayesian proposal of normalizing the product of the individual posterior densities.

The individual confidence curves might be in conflict with each other. It will then make sense to introduce a random component representing the distribution of θ across the k cases. Let $g(\theta; \psi)$ be the density of the mixing distribution, parametrized by ψ. Instead of adding the confidence log likelihoods, the method should be to add the confidence log likelihoods of the g-mixed models. The g-mixed confidence likelihood is $\tilde{L}_i(\theta) = \int \exp(-0.5\Gamma_1^{-1}(cc_i(\theta + t)))g(t; \psi)dt$. If the mixing parameter ψ is unknown, it could be estimated by maximizing the sum of the g-mixed confidence log likelihoods.

3 Integrating confidence intervals

Scientific reporting is often briefed to a point estimate $\hat{\psi}$ and a 95% confidence interval (ψ_1, ψ_2) for the parameter ψ. To merge this information with that from another confidence interval and point estimate, or indeed other information, a likelihood based on the information would be helpful.

The more information there is about how the confidence interval is established, the better a confidence likelihood based on the confidence interval is grounded. We will discus 2 possible models summarizing the available background information, or perhaps being just assumed models if no background information is available. The idea is to extrapolate from the point estimate and the confidence interval to a confidence distribution, in view of the model. The confidence distribution is then turned into a likelihood by the quantile χ^2_1-transform.

Assume the point estimate to be median unbiased and tail-symmetric, i.e. that $\hat{\psi}$ is the confidence median. With degree of confidence $1-\alpha$, the confidence limits are respectively the $\alpha/2$ and $1-\alpha/2$ confidence quantiles

A confidence interval is symmetric when $\hat{\psi} - \psi_1 = \psi_2 - \hat{\psi}$, and is otherwise skewed. Most confidence intervals obtained by statistical software are symmetric. They are often based on large sample theory giving that $\hat{\psi}$ is approximately normally distributed with standard error s estimated from the Hessian of the log likelihood function. The related confidence curve is thus

$$cc(\psi) = |1 - 2\Phi\left((\psi - \hat{\psi})/s\right)|,$$

withe confidence deviance $\Gamma_1^{-1}(cc(\psi)) = \left((\psi - \hat{\psi})/s\right)^2$.

Asymmetric confidence intervals might be obtained by a monotonic transformation h of a normally distributed estimate. The related confidence curve is then $cc(\psi) = |1-2\Phi\left((h(\psi) - h(\hat{\psi}))/s\right)|$, with related confidence likelihood. When ψ is population size or some other positive parameter, the log normal model might apply. The confidence deviance is then $D_c(\psi) = \left((\log(\psi) - \log(\hat{\psi}))/s\right)^2$. Positive parameters might also be power transformed $h(\psi; a) = \frac{1}{a}\psi^a$. The power a and the scale s must then be estimated by solving the two equations $\psi_1^a - \hat{\psi}^a = -sz$, $\psi_2^a - \hat{\psi}^a = sz$ where z is the upper $\alpha/2$ quantile in the normal distribution. For the correlation coefficient Fisher found the $h = \arcsin$ function to provide a nearly normal pivot. When ψ is a probability the logit function $h(\psi) = \log(\psi) - log(1-\psi)$ might work.

Effective population size for cod The effective population size N_e of a given stable population is the size of a hypothetical stable population where each individual has binomially distributed number of reproducing offsprings. The hypothetical population maintains the same genetic variability as the actual population over the generations. In a study of a cod population a point estimate of 198, and a 95% confidence interval of (106, 1423) was found for N_e. They were obtained from genetic data by a relatively complex method involving jackknifing. A point estimate of 1847, and a 95% confidence interval of (800, 2893) was also found for the actual size N_a of the population. The latter confidence interval was based on $\hat{N}_a \sim N(N_a, s)$. The results for N_e and N_a are stochastically independent.

From these data a confidence interval is sought for the ratio $\psi = N_e/N_a$. This ratio can be used to estimate the variance in number of reproductive offsprings, which is hard to estimate by other more direct methods. The confidence deviance for N_a is simply $D_c(N_a) = \left(\frac{N_a - 1847}{534}\right)^2$, since we find $s = 534$. For N_e we chose the linear normal model in the power transformed parameter. Solving the two equations we find $a = -0.992$ and $s = .00233$. Using $a = -1$ the confidence Deviance is $D_c(N_e) = \left(\frac{N_e^{-1} - 198^{-1}}{0.00233}\right)^2$.

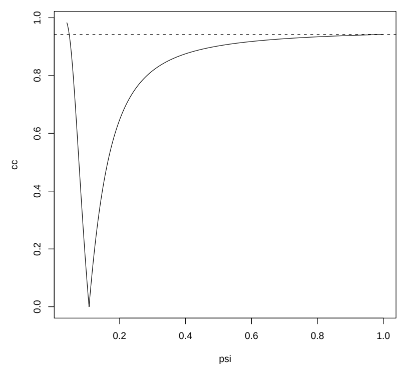

Figure 2: Confidence curve for the ratio $\psi = N_e/N_a$ based on confidence intervals for N_e and N_a

Except for an additive constant, the combined confidence deviance is $D(N_a, N_e) = D_c(N_a) + D_c(N_e)$, leading to the profile confidence deviance $D_c(\psi) = B(\psi) - \min_\psi B(\psi)$ where $B(\psi) = \min_{N_a} D(N_a, \psi N_a)$. Since by definition $N_e \leq N_a$, $\psi \in (0, 1]$. The resulting confidence curve for ψ, $\Gamma_1(D_c(\psi))$ is shown in Figure 2. The 95% confidence interval is (0.0467, 1]. The confidence curve reaches its right hand maximum $cc(1) = 0.94$. Tail-symmetric confidence intervals with non-trivial upper bound are thus only possible for degree of confidence less than 0.94.

References

[1] Brazzale, A. R. and Davison, A. C. (2008) “Accurate parametric inference for small samples”. *Statistical Science*, 23:465-484.

[2] Efron, B. (1993) “Bayes and likelihood calculations from confidence intervals”. *Biometrika*, 80:3-26.

[3] Fisher, R.A. (1930) “Inverse probability”. *Proceedings of the Cambridge Philosophical Society*, 26:528-535.

[4] Neyman, J. (1941) “Fiducial argument and the theory of confidence intervals”. *Biometrika*, 32:128-150.

[5] Schweder, T. and Hjort, N.L. 2013. *Confidence Likelihood, Probability.* Cambridge Univerity Press (forthcoming)

[6] Singh, K., Xie,M., and Strawderman,W. E. (2005) “Combining information from independent sources through confidence distributions”. *Annals of Statistics*, 33:159-183.