Combining Information Across Diverse Sources via Confidence Distributions

focustat
FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

Nils Lid Hjort

Department of Mathematics, University of Oslo

ISI, Rio de Janeiro, July 2015

The problem: Combining information

Suppose ψ is a parameter of interest, with data $y_1, \ldots, y_k$ from sources $1, \ldots, k$ carrying information about ψ. How to combine these pieces of information?

Standard (and simple) example: $y_j \sim \mathrm{N}(\psi, \sigma_j^2)$ are indepenedent, with known or well estimated σ_j. Then

$$\widehat{\psi} = \frac{\sum_{j=1}^k y_j/\sigma_j^2}{\sum_{j=1}^k 1/\sigma_j^2} \sim \mathrm{N}\Big(\psi, \Big(\sum_{j=1}^k 1/\sigma_j^2\Big)^{-1}\Big).$$

Often additional variability among the ψ_j. Would e.g. be interested in assessing both parameters of $\psi \sim \mathrm{N}(\psi_0, \tau^2)$.

We need extended methods and partly new paradigms for handling cases with very different types of information.

Plan

General problem formulation:

Data y_j source j carry information about ψ_j. Wish to assess overall aspects of these ψ_j, perhaps for inference concerning some $\phi(\psi_1, \ldots, \psi_k)$.

- A Confidence distributions.
- B Previous CD combination methods (Singh, Strawderman, Xie, Liu, Liu).
- C A different II-CC-FF paradigm, via steps Independent Inspection, Confidence Conversion, Focused Fusion and confidence-to-likelihood operations.
- D1 Example 1: Effective population size for cod.
- D2 Example 2: Olympic unfairness.
- E Concluding remarks.

A: Confidence distributions

For a parameter ψ, suppose data y give rise to confidence intervals, say $[\psi_{0.05}, \psi_{0.95}]$ at level 0.90, but also for other levels. These are converted into a full distribution of confidence, with

$$[\psi_{0.05}, \psi_{0.95}] = [C^{-1}(0.05, y_{\rm obs}), C^{-1}(0.95, y_{\rm obs})],$$

etc. Here $C(\psi, y)$ is a cdf in ψ, for each y, and

$$C(\psi_0, Y) \sim {\rm unif} \quad \text{at true value } \psi_0.$$

Very useful, also qua graphical summary: the confidence curve

$$ {\rm cc}(\psi) = |1 - 2\, C(\psi, y_{\rm obs})|,$$

with ${\rm cc}(\psi) = 0.90$ giving the two roots $\psi_{0.05}, \psi_{0.95}$, etc.

An extensive theory is available for CDs, cf. Confidence, Likelihood, Probability, Schweder and Hjort (CUP, 2015).

B: Liu, Liu, Singh, Strawderman, Xie et al. methods

Data y_j give rise to a CD $C_j(\psi, y_j)$ for ψ. Under true value, $C_j(\psi, Y_j) \sim \mathrm{unif}$. Hence $\Phi^{-1}(C_j(\psi, Y_j)) \sim \mathrm{N}(0, 1)$, and

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

is a combined CD, if the weights w_j are nonrandom and $\sum_{j=1}^{k} w_j^2 = 1$.

This is a versatile and broadly applicable method, but with some drawbacks: (a) trouble when estimated weights $\widehat{w}_j$ are used; (b) lack of full efficiency. In various cases, there are better CD combination methods, with higher confidence power.

Better (in various cases): sticking to likelihoods and sufficiency.

CD combination via confidence likelihoods

Combining information, for inference about focus parameter $\phi = \phi(\psi_1, \ldots, \psi_k)$: General II-CC-FF paradigm for combination of information sources:

II: Independent Inspection: From data source y_j to estimate and intervals, yielding a CD:

$$y_j \Longrightarrow C_j(\psi_j).$$

CC: Confidence Conversion: From the confidence distribution to a confidence log-likelihood,

$$C_j(\psi_j) \Longrightarrow \ell_{c,j}(\psi_j).$$

FF: Focused Fusion: Use 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.:

$$\ell_c(\psi_1, \ldots, \psi_k) \Longrightarrow \bar{C}_{\mathrm{fusion}}(\phi).$$

FF is also the (focused) Summary of Summaries operation.

Carrying out steps II, CC, FF can be hard work, depending on circumstances. The CC step is sometimes the hardest (conversion of CD to log-likelihood). The simplest method is normal conversion,

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

but more elaborate methods may typically be called for.

Sometimes step II needs to be based on summaries from other work (e.g. from point estimate and a .95 interval to approximate CD).

With raw data and sufficient time for careful modelling, steps II and CC may lead to $\ell_{c,j}(\psi_j)$ directly. Even then having individual CDs for the ψ_j is informative and useful.

Illustration 1: Classic meta-analysis.

II: Independent Inspection: Statistical work with data source y_j leads to $\widehat{\psi}_j \sim \mathrm{N}(\psi_j, \sigma_j^2)$; $C_j(\psi_j) = \Phi((\psi_j - \widehat{\psi}_j)/\sigma_j)$.

CC: Confidence Conversion: From $C_j(\psi_j)$ to $\ell_{c,j}(\psi_j) = -\frac{1}{2}(\psi_j - \widehat{\psi}_j)^2/\sigma_j^2$.

FF: Focused Fusion: With a common mean parameter across studies: Summing $\ell_{c,j}(\psi_j)$ leads to classic answer

$$\widehat{\psi} = \frac{\sum_{j=1}^k \widehat{\psi}_j/\sigma_j^2}{\sum_{j=1}^k 1/\sigma_j^2} \sim \mathrm{N}\Big(\psi, \Big(\sum_{j=1}^k 1/\sigma_j^2\Big)^{-1}\Big).$$

With ψ_j varying as $\mathrm{N}(\psi_0, \tau^2)$: then $\widehat{\psi}_j \sim \mathrm{N}(\psi_0, \tau^2 + \sigma_j^2)$. CD for τ:

$$C(\tau) = \mathrm{Pr}_\tau\{Q_k(\tau) \geq Q_{k,\mathrm{obs}}(\tau)\} = 1 - \Gamma_{k-1}(Q_{k,\mathrm{obs}}(\tau)),$$

with $Q_k(\tau) = \sum_{j=1}^k \{\widehat{\psi}_j - \bar{\psi}(\tau)\}^2/(\tau^2 + \sigma_j^2)$. There is a positive confidence probability for $\tau = 0$. CD for ψ_0: based on t-bootstrapping and

$$t = (\bar{\psi}(\widehat{\tau}) - \psi)/\kappa(\widehat{\tau}).$$

Illustration 2: Let $Y_j \sim \mathrm{Gamma}(a_j, \theta)$, with known shape a_j.

II: Independent Inspection: Optimal CD for θ based in Y_j is $C_j(\theta) = G(\theta y_j, a_j, 1)$.

CC: Confidence Conversion: From $C_j(\theta)$ to $\ell_{c,j}(\psi_j) = -\theta y_j + a_j \log \theta$.

FF: Focused Fusion: Summing confidence log-likelihoods, $\bar{C}_{\mathrm{fusion}}(\theta) = G(\theta \sum_{j=1}^{k} y_j, \sum_{j=1}^{k} a_j, 1)$. This is the optimal CD for θ, and has higher CD performance than the Singh, Strawderman, Xie type

$$\widetilde{C}(\theta) = \Phi\Big(\sum_{j=1}^{k} w_j \Phi^{-1}(C_j(\theta)\Big),$$

even for the optimally selected w_j.

Crucially, the II-CC-FF strategy is very general and can be used with very different data sources (e.g. hard and soft and big and small data). The potential of the II-CC-FF paradigm lies in its use for much more challenging applications (where each of II, CC, FF might be hard).

D1: Effective population size ratio for cod

A certain population of cod is studied. Of interest is both actual population size N and effective population size N_e (the size of a hypothetical stable population, with the same genetic variability as the full population, and where each individual has a binomially distributed number of reproducing offspring). The biological focus parameter in this study is $\phi = N_e/N$.

Steps II-CC for N: A CD for N, with confidence log-likelihood: A certain analysis leads to confidence log-likelihood

$$\ell_c(N) = -\tfrac{1}{2}(N - 1847)^2/534^2.$$

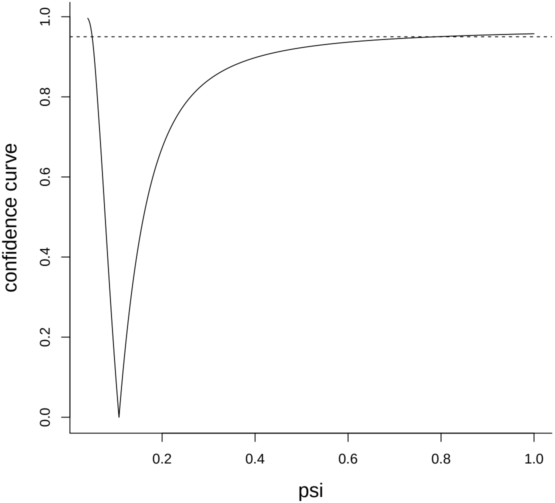
Steps II-CC for N_e: A CD for N_e, with confidence log-likelihood: This is harder, via genetic analyses, etc., but yields confidence log-likelihood

$$\ell_{c,e}(N_e) = -\tfrac{1}{2}(N_e^b - 198^b)/s^2$$

for certain estimated transformation parameters (b, s).

Step FF for the ratio: A CD for $\phi = N_e/N$. This is achieved via log-likelihood profiling and median-Bartletting,

$$\ell_{\rm prof}(\phi) = \max\{\ell_c(N) + \ell_{c,e}(N_e)\colon N_e/N = \phi\}.$$



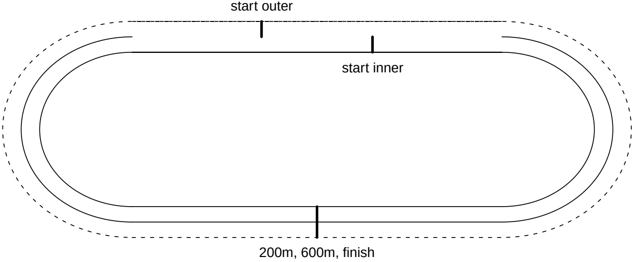
D2: The Olympic unfairness of the 1000 m

Olympic speedskaters run the 1000 m in less than 70 seconds (speed more than 50 km/h). They skate two and a half laps, in pairs, with a draw determining inner/outer. Acceleration matters ($mv^2/r_1 > mv^2/r_2$ with $r_1 = 25$ m and $r_2 = 29$ m), and so does fatigue at end of race.

Start in inner lane: three inners, two outers.

Start in outer lane: two inners, three outers.

I shall estimate the Olympic unfairness parameter d, the difference between outer and inner, for top skaters.



In the Olympics: only one race. In the annual World Sprint Championships: they race 500 m and 1000 m both Saturday and Sunday, and they switch start lanes.

The six best men, from Calgary, January 2012, Saturday and Sunday, with 'i' and 'o' start lanes, and passing times:

			200 m	600 m	1000 m		200 m	600 m	1000 m
1	S. Groothuis	i	16.61	41.48	1:07.50	o	16.50	41.10	1:06.96
2	Kyou-Hyuk Lee	i	16.19	41.12	1:08.01	o	16.31	40.94	1:07.99
3	T.-B. Mo	o	16.57	41.67	1:07.99	i	16.27	41.54	1:07.99
4	M. Poutala	i	16.48	41.50	1:08.20	o	16.47	41.55	1:08.34
5	S. Davis	o	16.80	41.52	1:07.25	i	17.02	41.72	1:07.11
6	D. Lobkov	i	16.31	41.29	1:08.10	o	16.35	41.26	1:08.40

I need a model for (Sat, Sun) results (Y_1, Y_2), utilising passing times $u_{i,1}, v_{i,1}$ for Sat race and $u_{i,2}, v_{i,2}$ for Sun race, along with

$$z_{i,1} = \begin{cases} -1 & \text{if no. } i \text{ starts in inner on Saturday,} \\ 1 & \text{if no. } i \text{ starts in outer on Saturday,} \end{cases}$$

$$z_{i,2} = \begin{cases} -1 & \text{if no. } i \text{ starts in inner on Sunday,} \\ 1 & \text{if no. } i \text{ starts in outer on Sunday.} \end{cases}$$

to get hold of d.

My model for (Sat, Sun) results, for skater i:

$$\begin{aligned} Y_{i,1} &= a_1 + bu_{i,1} + cv_{i,1} + \tfrac{1}{2}dz_{i,1} + \delta_i + \varepsilon_{i,1}, \\ Y_{i,2} &= a_2 + bu_{i,2} + cv_{i,2} + \tfrac{1}{2}dz_{i,2} + \delta_i + \varepsilon_{i,2}. \end{aligned}$$

Here $u_{i,1}, u_{i,2}$ are 200 m passing time, $v_{i,1}, v_{i,2}$ are 600 m passing time; δ_i follows the skater, with $\delta_i \sim \mathrm{N}(0, \kappa^2)$ across skaters; and $\varepsilon_{i,1}, \varepsilon_{i,2}$ are independent $\mathrm{N}(0, \sigma^2)$. The inter-skater correlation is $\rho = \kappa^2/(\sigma^2 + \kappa^2)$.

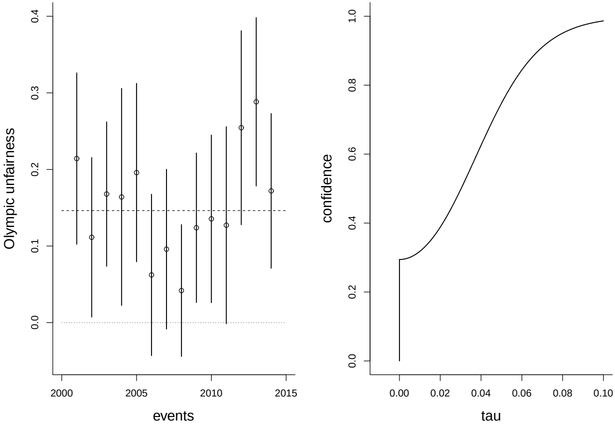
Crucially, outer lane start means adding $\frac{1}{2}d$, inner lane start means adding $-\frac{1}{2}d$, so d is overall difference due to start lane. Fairness means d should be very close to zero.

The model has seven parameters, and I need full analysis of dataset from each World Sprint Championships event to get hold of a CD for the focus parameter d.

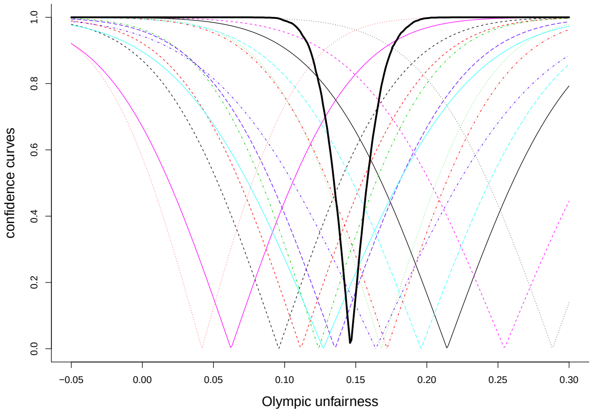
From full analysis of World Sprint events 2014, ..., 2001 (seven parameters in each model), I get hold of

$$\widehat{d}_j \sim \mathrm{N}(d_j, \sigma_j^2),$$

and I then use $d_j \sim \mathrm{N}(d_0, \tau^2)$. Full CD analyses are then available for d_0 and for τ.



Confidence curves $\mathrm{cc}(d_j)$ for the fourteen unfairness parameters, over 2014 to 2001. The overall estimate 0.14 seconds (advantage inner-starter) is very significant, and big enough to make medals change necks.



Conclusion: The skaters need to run twice. (I've told the ISU.)

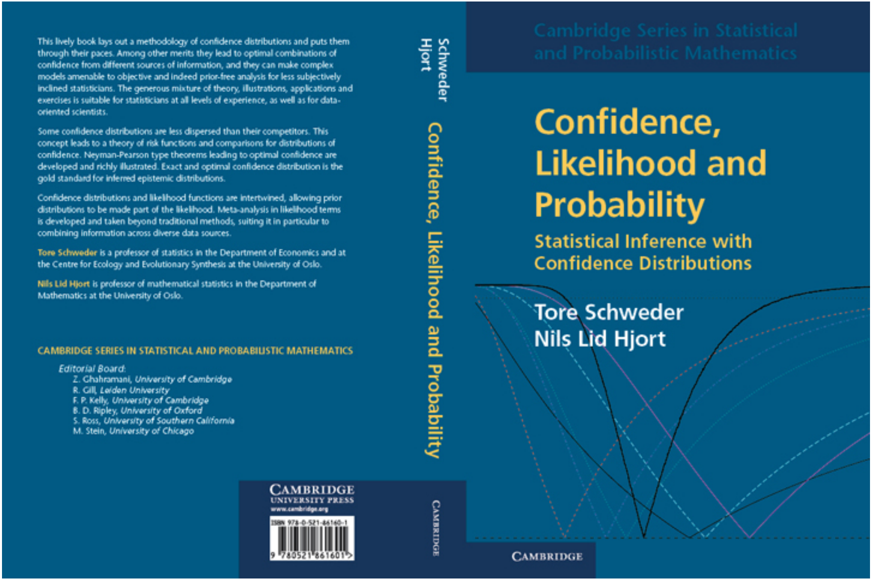
E: Concluding remarks (and further questions)

a. If we have the raw data, and have the time and resources to do all the full analyses ourselves, then we would find the $C_j(\psi_j)$ in Step II = Independent Inspection. In real world we would often only be able to find a point estimate and a 95% interval for the ψ_j. We may still squeeze an approximate CD out of this.

b. Step CC = Confidence Conversion is often tricky. There is no one-to-one correspondence between log-likelihoods and CDs. Data protocol matters. See CLP (2015).

c. Step FF = Focused Fusion may be accomplished by profiling the combined confidence log-likelihood, followed by fine-tuning (Bartletting, median correction, abc bootstrapping).

d. We see a good potential for the II-CC-FF scheme in harder applications involving hard and soft data, as well as with big and small data. Such applications will be worked with inside the FocuStat research programme 2014–2018.



More material: CLP, Cambridge University Press, 2015.