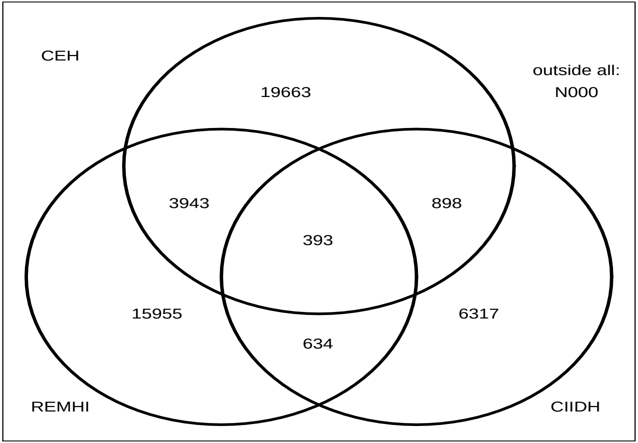
How many were killed in Guatemala, 1978–1996?

Nils Lid Hjort / Stability and Change 2022-2023, CAS

N000 Workshop, CAS, 22/ii/2023

Counting the not counted



Lists REMHI, CEH, CIIDH:
$n_{1,1,1} = 393$, $n_{1,1,0} = 3943$, $n_{1,0,0} = 15955$, $n_{1,0,1} = 634$,
$n_{0,1,1} = 898$, $n_{0,1,0} = 19663$, $n_{0,0,1} = 6317$. How big is N000?

Today (with more to come)

Clearly we need to assume something, to construct estimates of $N_{0,0,0}$ and the total

$$\begin{aligned} N &= N_{1,1,1} + N_{1,1,0} + N_{1,0,1} + N_{0,1,1} + N_{1,0,0} + N_{0,1,0} + N_{0,0,1} + XXX \\ &= N_{\mathrm{counted}} + N_{0,0,0}. \end{aligned}$$

Easiest start assumption is list independence, which means $\mathrm{Pr}(\text{counted in } 1, 2, 3) = pqr$ etc. Then $2^3 - 1 = 7$ probabilities are modelled via 3 parameters.

I will develop a log-likelihood profile method, with $\ell_{\mathrm{prof}}(N)$, giving $\widehat{N}$ and also a full confidence curve $\mathrm{cc}(N)$ – and apply this for Guatemala lists.

The methodology works also for other submodels (and for more than three lists). Wish to build a FIC for N000, a Focused Information Criterion that sorts through candidate models and finds the best.

Two lists: N = N11 + N10 + N01 + how many more?

Multinomial setup, with $(N_{0,0}, N_{0,1}, N_{1,0}, N_{1,1})$ having sum N, and probabilities

$$p_{i,j} = \Pr(X = i, Y = j) \quad \text{for } i, j = 0, 1,$$

1-0 for counted and not counted. Under list independence:

$$p_{0,0} = (1-p)(1-q), \quad p_{0,1} = (1-p)q, \quad p_{1,0} = p(1-q), \quad p_{1,1} = pq.$$

Two quanties aiming for the same pq:

$$\frac{N_{1,1}}{N} \quad \text{and} \quad \frac{N_{1,0}+N_{1,1}}{N}\frac{N_{0,1}+N_{1,1}}{N}.$$

Equating these gives the Petersen estimator (counting fish in Limfjorden, 1896):

$$N^* = \frac{(N_{1,0}+N_{1,1})(N_{0,1}+N_{1,1})}{N_{1,1}} = \frac{N_{1,\cdot}N_{\cdot,1}}{N_{1,1}}.$$

Behaviour of N^*

May work with the four multinomial ratios $\widehat{p}_{i,j} = N_{i,j}/N$:

$$N^{1/2}(\widehat{p}_{i,j} - p_{i,j}) \to_d A_{i,j},$$

a mean-zero four-normal with a clear covariance matrix.
Delta method yields:

$$\frac{N^* - N}{\sqrt{N}} = N^{1/2}(\frac{N^*}{N} - 1) \to_d U = \frac{A_{1,0} + A_{1,1}}{p} + \frac{A_{0,1} + A_{1,1}}{q} - \frac{A_{1,1}}{pq}.$$

We learn

$$N^*/\sqrt{N} - \sqrt{N} \approx_d \mathrm{N}(0, \tau^2), \quad \tau^2 = \frac{(1-p)(1-q)}{pq}.$$

Can construct confidence intervals etc. using this.

Note that p, q small implies high uncertainty (& vice versa).

Via log-likelihood profiling

It's fruitful to work with log-likelihood and profiling: results will be
(a) it gives $\widehat{N}$ almost equivalent to Petersen estimator N^*;
(b) there is a useful χ_1^2 recipe;
(c) matters generalise to $k \geq 3$ lists (where $\nexists$ Petersen).

With $N_{0,1}, N_{1,0}, N_{1,1}$ and hence $S = N_{0,1} + N_{1,0} + N_{1,1}$ observed, but $N = S + N_{0,0}$ unknown:

$$L(N,p,q) = \frac{N!}{(N-S)!\,N_{1,0}!\,N_{0,1}!\,N_{1,1}!}\{(1-p)(1-q)\}^{N-S}$$
$$\{(1-p)q\}^{N_{0,1}}\{p(1-q)\}^{N_{1,0}}(pq)^{N_{1,1}}.$$

Taking log, and maximising over p, q:

$$\ell_{\rm prof}(N) = \log(N!) - \log((N-S)!) + NH(\widehat{p}_N) + NH(\widehat{q}_n),$$

in terms of $\widehat{p}_N = N_{1,\cdot}/N$ and $\widehat{q}_N = N_{\cdot,1}/N$, and

$$H(r) = r\log r + (1-r)\log(1-r).$$

A chi-squared theorem for two independent lists

Some analysis, involving approximations, limiting normality, information, etc., and the quantity

$$J = \frac{1 - p_{0,0}}{p_{0,0}} - \frac{p}{1-p} - \frac{q}{1-q} = \frac{pq}{(1-p)(1-q)},$$

leads to

$$D(N_0) = 2\{\ell_{\mathrm{prof,max}} - \ell_{\mathrm{prof}}(N_0)\} \to_d U^2/J \sim \chi^2_1$$

at the true (but still unknown) N_0.

Confidence interval: $\{N_0\colon D(N_0) \le 1.96^2\}$, etc.

Full confidence curve:

$$\mathrm{cc}(N_0) = \Gamma_1(D(N_0)),$$

with $\Gamma_1(\cdot)$ the χ^2_1 c.d.f.

Srebrenica 1995

From Brunborg, Lyngstad, Urdal (2003): ICRC and PHR lists:
$N_{1,1} = 5712$, $N_{1,0} = 1586$, $N_{0,1} = 192$.
Estimate 7543; interval $[7528, 7560]$; $\widehat{p} = 0.967$, $\widehat{q} = 0.783$.

Three lists

First: Assuming list independence:

$$p_{i,j,k} = p_{i,\cdot,\cdot}\, p_{\cdot,j,\cdot}\, p_{\cdot,\cdot,k} \quad \text{for } i,j,k = 0,1.$$

No clear generalisation of the Petersen estimator. But log-likelihood profiling works well:

$$\ell_{\rm prof}(N) = \log(N!) - \log((N-S)!) + N\{H(\widehat p_N) + H(\widehat q_n) + H(\widehat r_N)\},$$

with the same $H(x) = x\log x + (1-x)\log(1-x)$ and

$$\widehat p_N = N_{1,\cdot,\cdot}/N, \quad \widehat q_N = N_{\cdot,1,\cdot}/N, \quad \widehat r_N = N_{\cdot,\cdot,1}/N.$$

Also, a crucial quantity

$$J = \frac{1-p_{0,0,0}}{p_{0,0,0}} - \frac{p}{1-p} - \frac{q}{1-q} - \frac{r}{1-r}$$

is at work. Theorem:

$$D(N_0) = 2\{\ell_{\rm prof,max} - \ell_{\rm prof}(N_0)\} \to_d U^2/J \sim \chi^2_1$$

at the true (but still unknown) N_0.

Fun to do: simulate, estimate, learn

Your fish population: $\{1, \ldots, N\}$.

Go fishing, with mark-release, probabilites p_1, p_2, p_3. This gives subsets A_1, A_2, A_3. Then do all of the above, with quite simple `R` tools

`setdiff`
`intersect`
`union`
`length`

One learns about the importance of p_1, p_2, p_3, the value of fishing even more (!), the somewhat skewed distributions of $\widehat{N}$, etc.

May also put priors into the game.

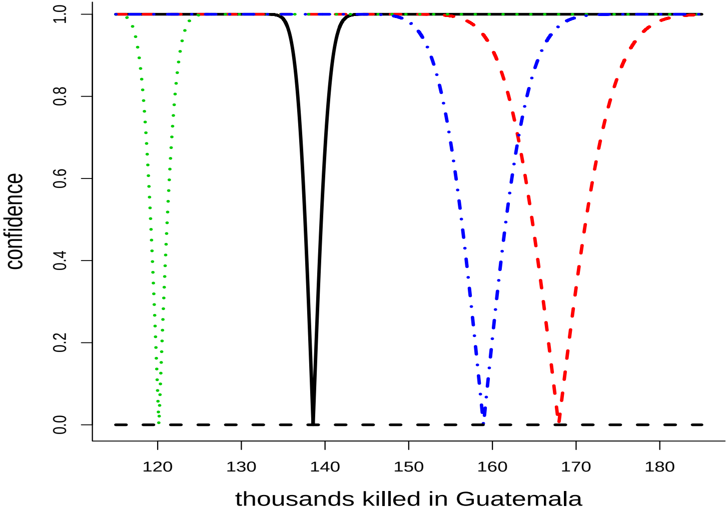
Guatemala

From Lum, Price, Banks (2013): Lists REMHI, CEH, CIIDH:

$n_{1,1,1} = 393$, $n_{1,1,0} = 3943$, $n_{1,0,0} = 15955$, $n_{1,0,1} = 634$, $n_{0,1,1} = 898$, $n_{0,1,0} = 19663$, $n_{0,0,1} = 6317$.

Using list independence (first): total estimate 138,576; 95 percent interval 135,794 to 141,453; low detection rates $(\widehat{p}, \widehat{q}, \widehat{r}) = (0.151, 0.179, 0.069)$.

Can do two lists at a time and the three lists jointly (looking for biases?).



With list independence assumption: Three two-sources curves, three-sources $\mathrm{cc}(N)$ in the middle.

With dependence among the lists

The log-likelihood profile machinery still works, for any $p_{i,j,k}(\theta)$; need $\dim(\theta) \le 6$. A class of four-parameter models:

$$
\begin{aligned}
p_{0,0,0} &= (1-p)(1-q)(1-r)/s \\
p_{0,0,1} &= (1-p)(1-q)r\,\gamma/s \\
p_{0,1,0} &= (1-p)q(1-r)/s \\
p_{0,1,1} &= (1-p)qr/s \\
p_{1,0,0} &= p(1-q)(1-r)/s \\
p_{1,0,1} &= p(1-q)/s \\
p_{1,1,0} &= pq(1-r)/s \\
p_{1,1,1} &= pqr/s
\end{aligned}
$$

where the γ is a parameter associated with cell 001, modifying independence in that direction; s is the factor to give sum 1.

This is the best of 8 similar choices. Then a clear leap in log-likelihood, and much better Pearson statistic

$$K = \sum_{i,j,k} (N_{i,j,k} - \widehat{N}\widehat{p}_{i,j,k})^2 / (\widehat{N}\widehat{p}_{i,j,k}).$$

A five-parameter model

Starting with independence equations, then modifying, in two directions:

$$\begin{aligned}
p_{0,0,0} &= (1-p)(1-q)(1-r)/s \\
p_{0,0,1} &= (1-p)(1-q)r\,\gamma_1/s \\
p_{0,1,0} &= (1-p)q(1-r)/s \\
p_{0,1,1} &= (1-p)qr/s \\
p_{1,0,0} &= p(1-q)(1-r)/s \\
p_{1,0,1} &= p(1-q)r/s \\
p_{1,1,0} &= pq(1-r)/s \\
p_{1,1,1} &= pqr\,\gamma_2/s
\end{aligned}$$

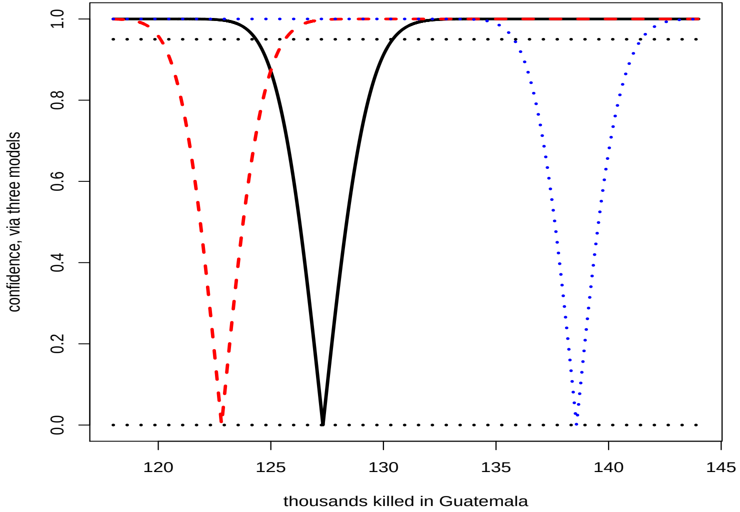
with s scale to get sum $p_{0,0,0} + \cdots + p_{1,1,1} = 1$.

The best cell for modification 1, with γ_1, is 001; and the best cell for modification 2, with γ_2, is 111.

So with modification parameters γ_1 placed at cell 001 and γ_2 placed at cell 111, I have a quite good model, with data fitting the model well (when it comes to the seven observed cells in the Venn diagram; can never check the 000 box).

The modifications amount to upward pushes at these two cells, with $\widehat{\gamma}_1 = 1.85$ and $\widehat{\gamma}_2 = 2.32$.

	obs3	obs5	expect3	expect5	pearson3	pearson5
n000	90772	79522	90772.223	79521.883	-0.001	0.000
n100	15955	15955	16144.571	15905.047	-1.492	0.396
n010	19663	19663	19880.329	19713.270	-1.541	-0.358
n001	6317	6317	5740.255	6317.001	7.612	0.000
n110	3943	3943	3535.877	3942.820	6.847	0.003
n101	634	634	1020.951	684.094	-12.110	-1.915
n011	898	898	1257.193	847.890	-10.130	1.721
n111	393	393	223.602	392.995	11.328	0.000



3-para: 138,576, with 135,794 to 141,453 (width 5,659)
4-para: 122,812, with 120,100 to 125,634 (width 5,534)
5-para: 127,314, with 124,341 to 130,415 (width 6,074)
Ball (1999): 132,174 (with a standard error of 6,568 ?).

Things To Do: bigger models, more sources

Looking for biases.

Inventing and using other models for the

$$p_{i,j,k}(\theta) = \mathrm{Pr}(X = i, Y = j, Z = k) \quad \text{for } i, j, k = 0, 1.$$

As long as $2^3 - 1 = 7$ probabilities in terms of θ of dimension 6 or lower, we're in business and can do log-likelihood profiling etc. Can search systematically (or 'logically') through

$$p_{i,j,k}(\theta) = p^{\mathrm{ind}}_{i,j,k} \exp(d_1 e_{i,j,k} + d_2 f_{i,j,k})/\mathrm{sum}.$$

Insights $\Longrightarrow$ covariates, or priors; will be helpful.

Yes, we can attack situations with $k \geq 4$ lists (did Patrick say '$k = 42$'?), but then need more care, for both modelling; principles giving shorter lists of candidate models; and clever algorithms for identifying and travelling through the most important ones.

Bayesian versions.

Things To Do: a FIC for N

Wish to develop a FIC for N ... which is (more or less) the same problem as constructing a FIC for $p_{0,0,0}$. Different models give different $\widehat{p}_{0,0,0}$ and $\widehat{N}$; a FIC will sort these via biases and variances to find the best model, for a given dataset.

Need a somewhat deeper theory than what I find the literature, regarding log-likelihood profiling and $\widehat{N}$. Basic lemma:

$$Z_{N_0}(d) = \ell_{\rm prof}(N_0 + dN_0^{1/2}) - \ell_{\rm prof}(N_0) \rightarrow_d Z(d) = dU - \tfrac{1}{2}Jd^2,$$

as N_0 increases, where $U \sim \mathrm{N}(0, K)$, and K is the same as J only under model conditions. So model-robust inference needs to use

$$D(N_0) = 2\{\ell_{\rm prof,max} - \ell_{\rm prof}(N_0)\} \rightarrow_d U^2/J = (K/J)\,\chi_1^2,$$

and these things enter my envisaged FIC-to-be.

(Some) references

P Ball. Lots of papers and reports.

H Brunborg, TH Lyngstad, H Urdal (2003). Accounting for genocide: How many were killed in Srebrenica? European Journal of Population.

G Claeskens, NL Hjort (2008). *Model Selection and Model Averaging.* CUP.

C Cunen, NL Hjort (2022). Combining information across diverse sources: newine the II-CC-FF paradigm. *Scandinavian Journal of Statistics.*

C Cunen, NL Hjort, HM Nygård (2020). Statistical Sightings of Better Angels. *Journal of Peace Research.*

NL Hjort (2019). Your Mother is Alive with Probability One Half. *FocuStat Blog xii.*

NL Hjort, EAa Stoltenberg (2023). *Statistical Inference: 666 Exercises, 66 Stories (and Solutions to All).* CUP.

M Jullum, NL Hjort (2017). Parametric or nonparametric? The FIC approach. *Statistica Sinica*.

K Lum, ME Price, D Banks (2013). Applications of multiple systems estimation in human rights research. *American Statistician*.

T Schweder, NL Hjort (2016). *Confidence, Likelihood, Probability.* CUP.