

Counting the uncounted: How many were killed in Guatemala, 1978-1995?

Nils Lid Hjort¹

¹ Department of Mathematics, University of Oslo, Norway

E-mail for correspondence: nils@math.uio.no

Abstract: In various application domains, there is a certain ‘null cell’, inside a multinomial setup, where observations are recorded for the other cells, but where one cannot count the number of occurrences for the null cell. I develop inference theory for assessing such unknown numbers, counting the uncounted, in situations where counts are available for the other cells, via parametric modelling. The methods are used to estimate the number of persons killed in Guatemala during the *Genocidio guatemalteco* years 1978–1995. There are three carefully curated lists of killed people, where the information can be mapped to a Venn diagram with $2^3 = 8$ cells. Summing over the seven observed cells, $R = 47,803$ killed individuals can be identified, but how big is $N_{0,0,0}$, and hence $N = N_{0,0,0} + R$?

Keywords: Capture-recapture; Confidence curves; Counting the unseen; Genocidio guatemalteco; Multinomial modelling; Multiple systems estimation.

1 Who killed the bishop, and who killed 127,000 people?

Figure 1 is a dramatic one. It pertains to people killed in Guatemala, during the tumultuous years 1978–1995 of the *Genocidio guatemalteco*, or *Holocausto silencioso*, recorded via official lists. The three sources in question are the Recovery of Historical Memory (REMHI), Commission for Historical Clarification (CEH), and the International Center for Human Rights Investigations (CIIDH), with acronyms reflecting project names in Spanish. The data, via careful scrutiny of lists from these three organisations, can in Venn diagram terms be translated to $N_{1,1,1} = 393$, $N_{1,1,0} = 3,943$, $N_{1,0,0} = 15,955$, $N_{1,0,1} = 634$, $N_{0,1,1} = 898$, $N_{0,1,0} = 19,663$, $N_{0,0,1} = 6,317$. The total number of confirmed deads, captured via these three lists, is $R = \sum_{(i,j,k) \neq (0,0,0)} N_{i,j,k} = 47,803$.

For glimpses into these horror tales, and to the political, sociological, racist, mass-psychological intricacies which led to these turmoils and the systematic murdering, partly or mostly via U.S.-backed military forces, check Ball

This paper was published as a part of the proceedings of the 40th International Workshop on Statistical Modelling (IWSM), Oslo, Norway, 28 June – 3 July 2026. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

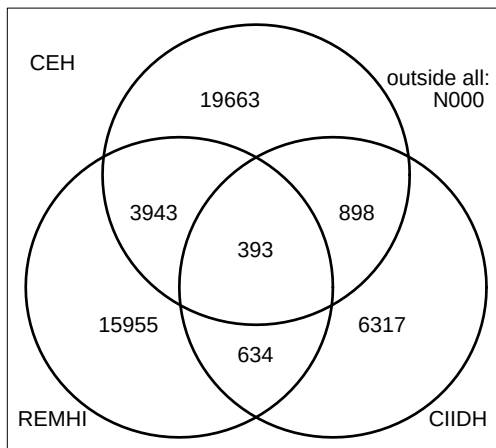


FIGURE 1. Venn diagram for the number of people killed and accounted for, for the three lists REMHI, CEH, CIIDH, in total $R = 47,803$ individuals. But how many are not counted, in the N000 box?

(1999), Lum et al. (2013), and Goldman’s (2010) detailed account *The Art of Political Murder: Who Killed Bishop Gerardi?*. The more mundane but nevertheless important challenge worked with here is the statistical one, to estimate the full number $N = N_{0,0,0} + \dots + N_{1,1,1}$ of individuals killed, hence also in the process the number $N_{0,0,0} = N - R$ of deads not captured on any of the three lists.

2 Theory for multinomials with an unobserved cell

Suppose $(N_0, N_1, \dots, N_k)$ is a multinomial vector, corresponding to N independent experiments, associated with categories $0, 1, \dots, k$, with probabilities $p_0, p_1, \dots, p_k$. We observe $N_1, \dots, N_k$, hence their sum $R = N_1 + \dots + N_k$, but not N_0 for the zero-box, falling outside the other categories. We assume a smooth parametric model is being used for the probabilities, with $p_j = p_j(\theta)$ for $j = 0, 1, \dots, k$, with θ of dimension $r \leq k - 1$, admitting two continuous derivatives. The joint probability distribution is

$$f = \frac{N!}{N_0! N_1! \dots N_k!} p_0(\theta)^{N_0} p_1(\theta)^{N_1} \dots p_k(\theta)^{N_k},$$

and the log-likelihood may be written

$$\begin{aligned}\ell(N, \theta) &= \log(N!) - \log((N - R)!) + (N - R) \log p_0(\theta) \\ &\quad + \sum_{j=1}^k N_j \log p_j(\theta).\end{aligned}$$

For the given model this also leads to the profiled log-likelihood

$$\ell_{\text{prof}}(N) = \max\{\ell(N, \theta) : \text{all } \theta\} = \ell(N, \hat{\theta}_N).$$

Maximising this function gives the maximum likelihood (ML) estimators $(\hat{N}, \hat{\theta})$.

To reach clear and useful approximation statements for inference, for growing N , assume that the model holds, for some true θ_0 , and write N_0 for the true N . The ML estimators $(\hat{N}, \hat{\theta})$ converge towards limits at different rates, and it is fruitful to write $N = N_0\rho$ and $\hat{\rho} = \hat{N}/N_0$. This leads to the approximate log-likelihood

$$\begin{aligned}\ell(\rho, \theta) &= N_0\rho \log(N_0\rho) - (N_0\rho - R) \log(N_0\rho - R) \\ &\quad + (N_0\rho - R) \log p_0(\theta) + \sum_{j=1}^k N_j \log p_j(\theta).\end{aligned}$$

From basic multinomial limits for multinomials, there is joint convergence $N_0^{1/2}\{N_j/N - p_j(\theta_0)\} \rightarrow_d A_j$, say, a zero-mean multinormal vector with $\text{cov}(A_j, A_\ell) = p_j(\theta_0)\{\delta_{j,\ell} - p_\ell(\theta_0)\}$; note that $A_0 + \dots + A_k = 0$. With $\tilde{p}_0 = 1 - R/N_0$, we then have $N_0^{1/2}\{\tilde{p}_0 - p_0(\theta_0)\} \rightarrow_d A_0$.

Letting $u_j(\theta) = \partial \log p_j(\theta) / \partial \theta$, taking the derivative of $\sum_{j=0}^k p_j(\theta) = 1$ leads to $\sum_{j=0}^k p_j(\theta) u_j(\theta) = 0$. Below, write for crispness p_j and u_j for $p_j(\theta_0)$ and $u_j(\theta_0)$. Formulae for the first-order partial derivatives of the log-likelihood $\ell(\rho, \theta)$ can be put up, and when evaluated at (N_0, θ_0) , these become

$$\begin{aligned}U_0 &= N_0(-\log \tilde{p}_0 + \log p_0), \\ V_0 &= N_0\left\{(\tilde{p}_0 - p_0)u_0 + \sum_{j=1}^k (N_j/N_0 - p_j)u_j\right\}.\end{aligned}$$

A key quantity is $M = \sum_{j=0}^k p_j u_j u_j^T$. There is convergence in distribution

$$\frac{1}{N_0^{1/2}} \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \rightarrow_d \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} -(1/p_0)A_0 \\ \sum_{j=0}^k A_j u_j \end{pmatrix} \sim N_{r+1}(0, J),$$

with

$$J = \begin{pmatrix} (1 - p_0)/p_0, & -u_0^T \\ -u_0, & M \end{pmatrix}.$$

This is the Fisher information matrix for this reparametrised model, with parameters (ρ, θ) , evaluated at $(1, \theta_0)$. We may verify that J is also the limit in probability of $-N_0^{-1}H_0$, where H_0 is the $(r+1) \times (r+1)$ Hessian matrix of second order partial derivatives of $\ell(\rho, \theta)$, evaluated at $(1, \theta_0)$. From this, we deduce first that

$$N_0^{1/2}(\hat{N}/N_0 - 1) = (\hat{N} - N_0)/N_0^{1/2} \rightarrow_d N(0, \tau^2), \quad (1)$$

where

$$\tau^2 = \frac{1}{(1-p_0)/p_0 - u_0^T M^{-1} u_0} = \frac{p_0}{1-p_0-p_0\delta}, \quad \text{with } \delta = u_0^T M^{-1} u_0.$$

Though the main concern might be estimating N_0 and $N = N_0 + R$, we note also that $N_0^{1/2}(\hat{\theta} - \theta_0) \rightarrow_d N_r(0, K)$, with

$$K = (M - (1-p_0)u_0 u_0^T / p_0)^{-1} = M^{-1} + dM^{-1}u_0 u_0^T M^{-1},$$

where we write

$$d = r_0/(1-r_0\delta) = p_0/(1-p_0-p_0\delta) \quad \text{and} \quad r_0 = p_0/(1-p_0).$$

Interestingly, had N_0 been observed too, the result would have been $N_r(0, M^{-1})$, so $dM^{-1}u_0 u_0^T M^{-1}$ is the extra amount of uncertainty when estimating θ , caused by N_0 not being counted.

The limit distribution result (1) is informative, by giving the precision of the ML estimator, and can also be used to form approximate confidence intervals. It is easier, and often better, however, to use the associated Wilks type chi-squared methodology. Indeed it can be shown, as for classical parametric likelihood inference, that

$$D(N_0) = 2\{\ell_{\text{prof,max}} - \ell_{\text{prof}}(N_0)\} \rightarrow_d \chi_1^2. \quad (2)$$

This leads to confidence intervals, e.g. $\{N_0: D(N_0) \leq 1.96^2\}$, and a full confidence curve $\text{cc}(N) = \Gamma_1(D(N_0))$, which we can read off directly from the profiled log-likelihood, in effect bypassing the perhaps cumbersome estimation of τ via the the J matrix etc. For pertinent details, see Hjort and Stoltenberg (2026, Ch. 5) and Schweder and Hjort (2016, Ch. 3–4).

3 The three-parameter independent lists model

A structurally simple and sometimes fully adequate model for the cell probabilities is that of *list independence*. The model postulates probabilities for a single person to land in lists One, Two, Three to be p, q, r . With independence, the eight cell probabilities then become $p_{0,0,0} = (1-p)(1-q)(1-r)$,

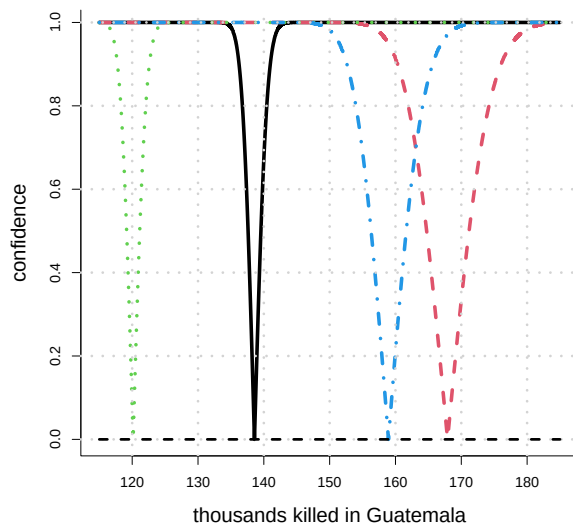


FIGURE 2. Confidence curves for N , the total number of people killed in Guatemala 1978–1996, in thousands, based on list independence, using all three sources (full black curve), and using pairwise analyses. The best estimate, under this assumption, is 138,576, with 95 percent interval [135,794, 141,453].

$p_{0,0,1} = (1-p)(1-q)r$, and so on, up to $p_{1,1,1} = pqr$. The log-likelihood function becomes

$$\begin{aligned} \ell(N, p, q, r) = & \log(N!) - \log((N-R)!) + A \log p + (N-A) \log(1-p) \\ & + B \log q + (N-B) \log(1-q) + C \log r + (N-C) \log(1-r), \end{aligned}$$

in terms of $A = N_{1,\cdot,\cdot} = N_{1,0,0} + N_{1,0,1} + N_{1,1,0} + N_{1,1,1}$, and similarly $B = N_{\cdot,1,\cdot}$, $C = N_{\cdot,\cdot,1}$. This is easily maximised over p, q, r for each given N , leading to

$$\ell_{\text{prof}}(N) = \log(N!) - \log((N-R)!) + N\{H(\hat{p}_N) + H(\hat{q}_N) + H(\hat{r}_N)\},$$

with $\hat{p}_N = A/N$, $\hat{q}_N = B/N$, $\hat{r}_N = C/N$, involving the function $H(u) = u \log u + (1-u) \log(1-u)$.

Using the general theory of Section 2, related to the precise limiting distribution of (1), some work shows first that the M matrix is diagonal with elements $1/\{p(1-p)\}$, $1/\{q(1-q)\}$, $1/\{r(1-r)\}$, and that

$$\delta = \psi_{0,0,0}^T M^{-1} \psi_{0,0,0} = p/(1-p) + q/(1-q) + r/(1-r).$$

From this we conclude that $(\hat{N} - N)/N^{1/2} \rightarrow_d N(0, \tau^2)$, with

$$\tau^2 = \frac{(1-p)(1-q)(1-r)}{1 - (1-p)(1-q)(1-r)(1+\delta)} = \frac{(1-p)(1-q)(1-r)}{pq + pr + qr - 2pqr}.$$

This result is a 3-list extension of known results in the capture-recapture literature for the 2-list case. It is seen that the variance increases drastically with smaller inclusion probabilities p, q, r .

The simpler case of only two independent sources can indeed be treated similarly, then involving the profiled log-likelihood function

$$\ell_{\text{prof}}(N) = \log(N!) - \log((N-R)!) + N\{H(\hat{p}_N) + H(\hat{q}_N)\}$$

and $\hat{p}_N = A/N$, $\hat{q}_N = B/N$, with $A = N_{1\cdot}$ and $B = N_{\cdot 1}$. One may show that the ML estimator is large-sample equivalent to and in fact very close to the first published estimator for such capture-recapture problems, the Petersen (1896) estimator $N^* = AB/N_{1\cdot 1}$. It was used by him to estimate the number of young plaice immigrating to the Danish Limfjord from the German Sea; see also Hjort and Petersen (1905).

Partly bypassing this $N(0, \tau^2)$ result, we may use the deviance function (2) and the associated Wilks type result we to compute $cc_3(N)$, the confidence curve for N using this 3-parameter model. Figure 2 displays this three-sources based $cc_3(N)$, alongside those obtained using only two sources at the time. The point estimate is $\hat{N}_3 = 138,576$, with 95 percent confidence interval $[135,794, 141,453]$. Point estimates for the inclusion probabilities (p, q, r) are $(0.151, 0.180, 0.059)$, i.e. rather small, implying high chances of a killed individual not being detected; this outside probability $(1-p)(1-q)(1-r)$ is estimated at 0.655. This spells statistical difficulty for assessing the zero cell $N_{0,0,0}$, and indeed also nonrobustness for the $\hat{N}_3$.

4 Better models, with four and five parameters

There are several reasons indicating that the list independence assumption is not holding up for the Guatemala killings, however, as pointed to in Ball (1999) and Lum et al. (2013). That is also born out statistically, as we shall see now, when we construct first a better 4-parameter model and then an even better and eventually satisfactory 5-parameter model. Figure 2, with the difference between the pairwise estimates, is also evidence against independence.

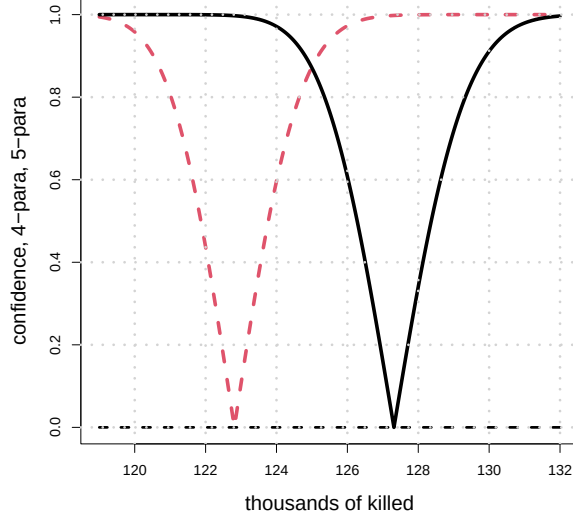


FIGURE 3. Confidence curves for N , the number of persons killed in Guatemala, in thousands, using the 4-parametric (red, dashed) and 5-parametric (black, full) models. Note that the N scale used here is shorter than for Figure 2. The point estimates are 122,812 and 127,314, with 95 percent intervals [120,100, 125,634] and [124,341, 130,415].

Consider first the four-parameter model

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

where the γ is a parameter associated with cell 001, modifying independence in that direction, and s is the factor required to give sum 1 over the eight cells. We can then numerically maximise the log-likelihood $\ell(N, p, q, r, \gamma)$, and learn that the Karl Pearson type statistic

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

is much smaller than for the simpler three-parameter independence model; here $\hat{N}_{0,0,0} = \hat{N} - R$ is used for $N_{0,0,0}$. Checking the numerics we find that the K is reduced from 487.05 with the 3-parametric model to 220.31 for the 4-parametric model; also, the log-likelihood max is increased with as much as 156.22. The ML estimates are $\hat{N}_3 = 138,576$ and $\hat{N}_4 = 122,812$. It turns out that an even better model is as follows, with five parameters. It uses push or modification factors γ_1 and γ_2 , placed for cells 001 and 111,

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

and again with s the required factor securing these eight probabilities having sum 1. There are actually 28 such models, placing γ_1 and γ_2 different places, so to speak, but the one above, involving such pushes away from independence for cells 001 and 111, is seen to be the clearly best one, in terms of log-likelihood increase and smaller value of the Pearson K . The push parameters (γ_1, γ_2) are estimated at $(1.85, 2.31)$; in particular, there is a higher chance for a killed person than merely pqr of landing in each of the three lists. In yet other words, if you have been detected once, the chance is bigger that you will be detected by one or both of the others. The informative summary Table 1 gives the expected or predicted numbers $\hat{E}_{i,j,k} = \hat{N}\hat{p}_{i,j,k}$, for the three models, as exp3, exp4, exp5, and then the Pearson residuals $P_{i,j,k} = (N_{i,j,k} - E_{i,j,k})/E_{i,j,k}^{1/2}$, as pear3, pear4, pear5. For the null cell, I use $\hat{N}_{0,0,0} = \hat{N}\hat{p}_{0,0,0}$. As we see, there are clear discrepancies between data and the 3- and 4-parameter models, whereas the 5-parameter model passes muster and is judged to deliver the most precise estimates. From Figure 3, an estimated $\hat{N}_5 = 127,314$ individuals were killed, with confidence interval $[120,100, 125,634]$.

5 Concluding remarks

Srebrenica 1995. A partly similar but statistically somewhat easier tale is that of counting the killed in Srebrenica, 1995. There are two lists, not three, but the inclusion probabilities p, q are higher than for the Guatemala case. See Hjort and Stoltenberg (2026, Story #65) for details, with point estimate $\hat{N} = 7,543$ for the number killed.

TABLE 1. Estimates of $N_{0,0,0}$, with expected numbers $E_{i,j,k}$ for the other seven categories, along with Pearson residuals $P_{i,j,k} = (N_{i,j,k} - E_{i,j,k})/E_{i,j,k}^{1/2}$, for the 3-, 4-, 5-parameter models. The Pearson type statistic $K = \sum_{i,j,k} P_{i,j,k}^2$ takes the values 487.05, 220.31, 6.91 for the 3-, 4-, 5-parameter models, the latter being the clear and adequate winner.

	obs	exp3	exp4	exp5	pear3	pear4	pear5
000		90773.343	75009.236	79511.180	-0.001	-0.001	-0.001
001	6317	5740.222	6316.999	6317.008	7.612	0.000	-0.001
010	19663	19880.337	19606.796	19712.903	-1.541	0.401	-0.355
011	898	1257.170	954.011	847.908	-10.129	-1.813	1.720
100	15955	16144.568	15819.072	15904.699	-1.491	1.080	0.398
101	634	1020.932	769.711	684.106	-12.109	-4.891	-1.915
110	3943	3535.834	4134.975	3943.191	6.847	-2.985	-0.003
111	393	223.595	201.196	393.002	11.329	13.522	-0.001
					487.052	220.308	6.914

FIC. Bartolucci and Lupparelli (2008) have developed a Focused Information Criterion, in the style of Claeskens and Hjort (2008), applicable to such multinomial data with missing observations. There is a need to generalise this further, building on methods from this paper. A broader FIC can be built, e.g. for the purpose of sorting through and ranking the 28 5-parameter candidate models of the type used in Section 3, via (i) putting up a fixed wide model, assumed to contain an approximation to the true data-generating model, and (ii) establishing and estimating biases, variances, mean squared errors, for the resulting N estimators. This requires careful extensions of the material of Section 2, to setups where the parametric models used are outside the correct data-generating process. Ongoing work in such directions will be reported on elsewhere.

Bayes. Methods developed and worked with in this paper have been frequentist in spirit, but parallel Bayesian methods can be constructed too, via appropriate priors for either N or $N_{0,0,0}$ themselves, or for the parameters of the models used. One may e.g. use information from Ball (1999) to elicit priors for the inclusion probabilities p, q, r , both for the simpler 3-parameter model of Section 2, and for the more complex ones of Section 3.

Further methodological advances. There is a need to develop and refine models and methods for assessing null cells in multinomial or Poisson setups, as this paper can be seen as an invitation to. See in this connection also models in Hjort (2026) for negatively correlated binomials and Poissons.

Other uses of the methodology. Examples abound in various application domains, where one needs to ‘count the uncounted’. How many people are cheating on their taxes; how many illegal immigrants are there; how many words did Shakespeare know (apart from those he actually used); how many

unseen sharks are there in a certain region. And, sadly, one may continue to hunt, as does Patrick Ball and the *Human Rights Data Analysis Group*, for the number of killed people, in wars and in conflict zones.

Acknowledgments: I am grateful for many fruitful discussions on themes related to this article with Céline Cunen, and also for information and comments from Patrick Ball.

References

- Ball, P. (1999). Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis. *American Academy for the Advancement of Science*, Washington
- Bartolucci, F. and Lupporelli, M. (2008). Focused Information Criterion for capture-recapture models for closed populations. *Scandinavian Journal of Statistics*, **9**, 658–664.
- Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.
- Goldman, F. (2010). *The Art of Political Murder: Who Killed Bishop Gerardi?* New York: Atlantic Books.
- Hjort, J. and Petersen, C.G.J. (1905). Short review of the results of the international fisheries investigation, volume 3. *Conseil Permanent International Pour l'Exploration de la Mer, Rapport et Proces-Verbaux des Reunions*, København.
- Hjort, N.L. (2026). Modelling pairs of Poissons and binomials with negative correlation. Statistical Research Report, Department of Mathematics, University of Oslo.
- Hjort, N.L. and Stoltenberg, E.Aa. (2026). *Statistical Inference: 600 Exercises, 100 Stories*. Cambridge: Cambridge University Press (in process).
- Lum, K., Price, M. E., and Banks, D. (2013). Applications of multiple systems estimation in human rights research. *American Statistician*, **24**, 191–200.
- Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, **6**, 5–84.
- Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability: Statistical Inference via Confidence Distributions*. Cambridge: Cambridge University Press.