Estimating abundance in capture-recapture

The importance of model and estimator choice

Matthew Schofield[1], Bill Link[2], Richard Barker[3], and Heloise Pavanato[1]

[1] Department of Mathematics and Statistics, University of Otago

[2] USGS Patuxent Wildlife Research Center

[3] Division of Science, University of Otago

Thought experiment

- Suppose we discover a coin landed heads y times.
  - The coin was flipped a fixed but unknown number of times: N
  - Cannot assume the coin is fair: probability π
  - The coin is available for us to use
- Question:
  - How to estimate N?

Thought experiment

- Suppose we discover a coin landed heads y times.
  - The coin was flipped a fixed but unknown number of times: N
  - Cannot assume the coin is fair: probability π
  - The coin is available for us to use
- Question:
  - How to estimate N?
- Answer:
  - Use the coin in a secondary experiment: flip M times and see x heads.
    - We'll set $M = y$.
  - Data x provide information to estimate π.

Thought experiment

- Suppose $y = 200$
  - Model: $y \sim \text{binomial}(N, \pi)$
- Also have $M = y = 200$, $x = 100$
  - Model: $x \sim \text{binomial}(y, \pi)$
- Question: what is your estimate of N?

Thought experiment

- We now discover the 200 heads arose from P experiments (with the same coin)
- In experiment j the coin flipped N_j times with y_j heads
- Interest is in estimating $\nu = \sum_{j=1}^{P} N_j$
- Suppose the data came from $P = 10$ experiments
  - $y_1 = y_2 = \ldots = y_P = 20$
    - Model: $y_j \sim \text{binomial}(N_j, \pi), \quad j = 1, \ldots, P$
  - Recall that $M = \sum_j y_j = 200$, $x = 100$.
    - Model: $x \sim \text{binomial}(M, \pi)$
- Question: What is your estimate of ν?

Look in closer

- Start with the pair of (independent) binomials

$$y \sim \mathrm{Bin}(N, \pi), \qquad x \sim \mathrm{Bin}(y, \pi)$$

- N and π are unknown
  - Simplified version of capture-recapture
  - Our thought experiment from the start (with $M = y$)

Conditional maximum likelihood estimator

$$y \sim \mathrm{Bin}(N, \pi), \qquad x \sim \mathrm{Bin}(y, \pi)$$

- Suppose that we had $x = 100$ and $y = 200$
- Condition on y and use x to estimate π
  - $\tilde{\pi} = \frac{x}{y}$
- Condition on $\pi = \tilde{\pi}$ and use y to estimate N
  - $\tilde{N} = \frac{y}{\tilde{\pi}}$
- $\tilde{\pi} = 0.5$
- $\tilde{N} = 400$
- Does this agree with our answer from earlier?

Maximum likelihood estimator

$$y \sim \text{Bin}(N, \pi), \qquad x \sim \text{Bin}(y, \pi)$$

- Suppose that we had $x = 100$ and $y = 200$
- MLE for π depends on y
  - $\hat{\pi} = \frac{x+y}{y+\hat{N}}$
- MLE for N:
  - $\hat{N} = \arg\max_N \frac{N!}{(N-y)!} \hat{\pi}^{x+y} (1-\hat{\pi})^{N-x}$
- $\hat{\pi} = 0.5017$
- $\hat{N} = 398$
- Did anyone have $\hat{N} = 398$?
  - The MLE and conditional estimator differ.

Multiple populations

- Let's suppose we have P populations
  - In the thought experiment these were multiple experiments
- Extend our model (common π):

$$y_j \sim \mathrm{Bin}(N_j, \pi), \qquad x_j \sim \mathrm{Bin}(y_j, \pi), \qquad j = 1, \ldots, P$$

- Interest is in estimation of $\nu = \sum_{j=1}^{P} N_j$

Conditional maximum likelihood estimator

$$y_j \sim \mathrm{Bin}(N_j, \pi), \qquad x_j \sim \mathrm{Bin}(y_j, \pi), \qquad j = 1, \ldots, P$$

- Suppose that we had $P = 10$, $x_1 = \ldots = x_P = 10$ and $y_1 = \ldots = y_P = 20$
- Condition on $y_1, \ldots, y_P$ and use $x_1, \ldots, x_P$ to estimate π
  - $\tilde{\pi} = \frac{\sum_j x_j}{\sum_j y_j}$
- Condition on $\pi = \tilde{\pi}$ and use $y_1, \ldots, y_P$ to estimate $\nu = \sum_j N_j$
  - $\tilde{N} = \frac{\sum_j y_j}{\tilde{\pi}}$
- $\tilde{\pi} = 0.5$
- $\tilde{\nu} = 400$
- Identical to earlier estimator
  - Totals $\sum_j y_j$ and $\sum_j x_j$ are unchanged

Maximum likelihood estimator

$$y_j \sim \mathrm{Bin}(N_j, \pi), \qquad x_j \sim \mathrm{Bin}(y_j, \pi), \qquad j = 1, \ldots, P$$

- Suppose that we had $P = 10$, $x_1 = \ldots = x_P = 10$ and $y_1 = \ldots = y_P = 20$
- MLEs:
  - $\hat{\pi} = 0.513$
  - $\hat{\nu} = 385$
- Differs from earlier estimator
  - $P = 1$: $\hat{\nu} = 398$
  - $P = 10$: $\hat{\nu} = 385$
  - $P = 25$: $\hat{\nu} = 362$

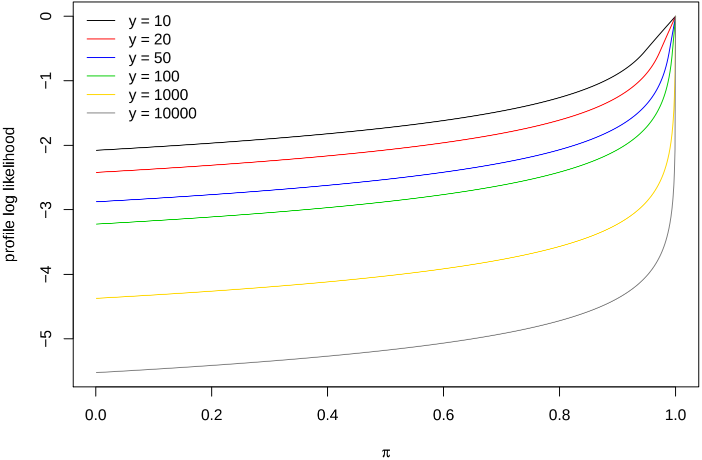
Model for y

- $y \sim \mathrm{Bin}(N, \pi)$
  - One observation, two unknowns
- Over-specified model!

Model for y

- $y \sim \mathrm{Bin}(N, \pi)$
  - One observation, two unknowns
- Over-specified model!
- It has a unique MLE
  - $\hat{N} = y$, $\hat{\pi} = 1$

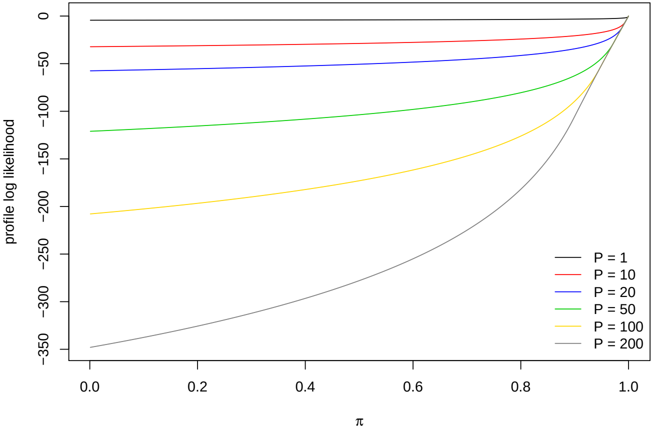
Profile log likelihood of π from data y



- $y \sim \mathrm{Bin}(N, \pi)$ is providing weak information about π

Profile log likelihood of π from data $y_1, \ldots, y_P$

- Let $\sum_{j=1}^{P} y_j = 1000$ and vary number of populations P (we set $y_1 = \ldots = y_P$)



- y provides increasing information about π as P increases

Poisson models

- An alternate model: Poisson rather than binomial
  - $N \sim \text{Poisson}(\lambda)$
  - Marginalize over N
- The MLE from Poisson model is equivalent to conditional estimator (Cormack)
- Result extends to multiple populations P

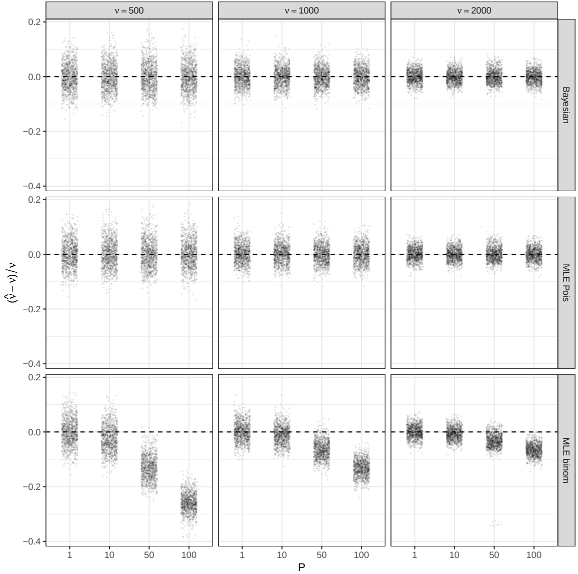
What about Bayes?

- The motivation was inconsistency between MLE and Bayes estimates
  - MRDS example
- Prior choice is important
  - We consider scale prior for N: $f(N) \propto N^{-1}$ (Link)
- Identical posteriors for binomial and Poisson models
  - For specific prior choice

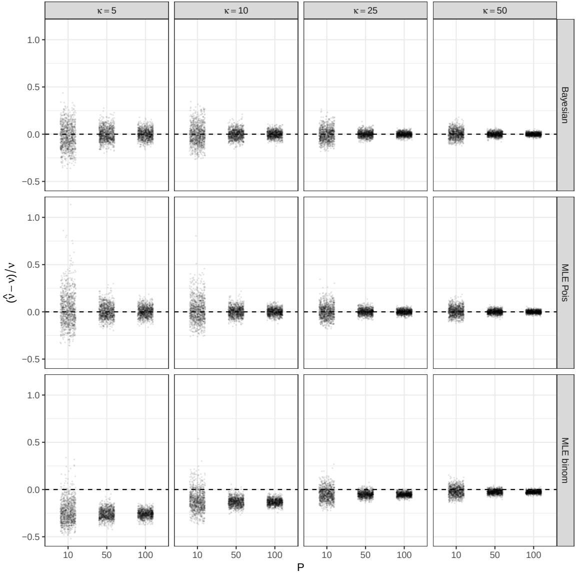
Beyond the pair of binomials

- Results generalize to more realistic mark-recapture models
  - The term $y \sim \text{Bin}(N, \pi)$ essentially remains unchanged
  - More complex model for x
- Simulate and fit closed population model $\mathrm{M_t}$
  - $K = 5$ sampling periods
  - Simulation 1:
    - Vary $P = 1, 10, 50, 100$ and $\nu = 500, 1000, 2000$
  - Simulation 2: consider $\kappa = N_1 = \ldots = N_P$
    - Vary $P = 10, 50, 100$ and $\kappa = 5, 10, 25, 50$
- Fixed $N_1, \ldots, N_P$
  - The true model is multinomial/binomial

Simulation 1



Simulation 2



Discussion

- This work was motivated by a real example
  - Estimator sensitivity that was unexpectedly 'extreme' for a moderately sized sample
- What I haven't talked about:
  - What is known about asymptotic behaviour of these estimators (e.g. Fewster & Jupp)
  - Use notions of ancillary to help explain results
  - Frame the problem in terms of nuisance parameters
  - Connections to REML in mixed effect models
- Summar:
  - MLE estimation performs poorly (as P increases)
  - Important to understand estimator behaviour in finite samples