STANDARD ERRORS IN SYSTEMATIC SAMPLING

Chris Glasbey & Stijn Bierman

Biomathematics and Statistics Scotland
How many acorns are under an oak tree?
Standard estimator of total is

\[ \hat{T} = \frac{N}{m} \sum_{i=1}^{m} y_i \]

where \( N = 36^2, \quad m = 12^2 \)

(NB We are only interested in the total, not in interpolating \( y \)'s)

But what is its variance?

Two frameworks for inference:

1) Design based: Assume unobserved array of \( y \)'s is fixed, and only stochasticity is due to sampling

2) Model based: Assume \( y \)'s are a single realisation from a super-population of possible arrays

But, for systematic sampling, there is no design-based, unbiased estimator of \( \text{var}(\hat{T} - T) \)
1. Design-based inference

If we had sampled at random:

\[ \text{var}(\hat{T} - T) = \left(1 - \frac{m}{N}\right) \frac{N^2}{m} \hat{\sigma}^2, \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{m - 1} \sum_{i=1}^{m} (y_i - \bar{y})^2, \]

is an unbiased estimator with \((m - 1)\) d.f., where \(\sigma^2\) is variance of \(y\) array.
From randomisation, \( \text{var}(\hat{T} - T) = y^T Q y \), with \( y \) arranged as vector

For random sampling

\[
Q = \begin{pmatrix}
8, & -\epsilon, & -\epsilon, & -\epsilon, & -\epsilon, & \ldots \\
-\epsilon, & 8, & -\epsilon, & -\epsilon, & -\epsilon, & \ldots \\
-\epsilon, & -\epsilon, & 8, & -\epsilon, & -\epsilon, & \ldots \\
-\epsilon, & -\epsilon, & -\epsilon, & 8, & -\epsilon, & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots
\end{pmatrix}
\]

\( \epsilon = 0.0062 \)

Whereas for systematic sampling

\[
Q = \begin{pmatrix}
8, & -1, & -1, & 8, & -1, & \ldots \\
-1, & 8, & -1, & -1, & 8, & \ldots \\
-1, & -1, & 8, & -1, & -1, & \ldots \\
8, & -1, & -1, & 8, & -1, & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots
\end{pmatrix}
\]

In the absence of trend, and if “covariances” decay with distance, systematic sampling is most efficient (Bellman, 1977)
For this problem we know the truth, a complete census:

\[ T = 85306 \text{ acorns} \]

(Approximated from Aubry and Debouzie, *Ecology*, 2000.)

So we can conduct a simulation study.
<table>
<thead>
<tr>
<th>sampling design</th>
<th>rmse</th>
<th>d.f.</th>
<th>conf. width</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>random</td>
<td>8900</td>
<td>143</td>
<td>35000</td>
</tr>
</tbody>
</table>
Two other types of design with unbiased estimators of $\text{var}(\hat{T} - T)$:

- **replicated systematic**
  - d.f. = 3

- **stratified**
  - d.f. $\leq \frac{m}{2}$
<table>
<thead>
<tr>
<th>sampling design</th>
<th>rmse</th>
<th>d.f.</th>
<th>conf. width</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-rep systematic</td>
<td>2600</td>
<td>3</td>
<td>15000</td>
</tr>
<tr>
<td>2-sample strata</td>
<td>3600</td>
<td>12</td>
<td>16000</td>
</tr>
<tr>
<td>random</td>
<td>8900</td>
<td>143</td>
<td>35000</td>
</tr>
</tbody>
</table>

We could pretend systematic design was one of other three, to obtain $\text{var}$

If we believe that the systematic design is the most efficient one

But if there are trends it may not be!
For example, if the data were elevations, such as:

<table>
<thead>
<tr>
<th>sampling design</th>
<th>rmse</th>
<th>d.f.</th>
<th>conf. width</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-rep systematic</td>
<td>380</td>
<td>3</td>
<td>2300</td>
</tr>
<tr>
<td>2-sample strata</td>
<td>110</td>
<td>10</td>
<td>500</td>
</tr>
<tr>
<td>random</td>
<td>410</td>
<td>143</td>
<td>1640</td>
</tr>
</tbody>
</table>
2. Model-based inference

Model for trend:

\[ f_{ij} = \beta_1 \exp \left[ -\frac{1}{2} \left( \frac{\log r_{ij} - \beta_4}{\beta_5} \right)^6 \right] \]

\[ r_{ij} = \sqrt{ (i - \beta_2)^2 + (j - \beta_3)^2 } \]

Fit by maximising Poisson quasi-likelihood, \( \sum (y \log f - f) \), and \( \hat{T}_f = \sum \hat{f}_{ij} \)
<table>
<thead>
<tr>
<th>sampling design</th>
<th>rmse $\hat{T}$</th>
<th>rmse $\hat{T}_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>4-rep systematic</td>
<td>2600</td>
<td>2000</td>
</tr>
<tr>
<td>2-sample strata</td>
<td>3600</td>
<td>2800</td>
</tr>
<tr>
<td>random</td>
<td>8900</td>
<td>3600</td>
</tr>
</tbody>
</table>
Isotropic exponential model for autocorrelation of errors, estimated from standardised residuals \( \hat{e} = (y - \hat{f})/\hat{f} \)

\[
\hat{T}_c = E(T \mid \hat{f}, \hat{e})
\]
<table>
<thead>
<tr>
<th>sampling design</th>
<th>rmse $\hat{T}$</th>
<th>rmse $\hat{T}_f$</th>
<th>rmse $\hat{T}_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td>2000</td>
<td>2000</td>
<td>2100</td>
</tr>
<tr>
<td>4-rep systematic</td>
<td>2600</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>2-sample strata</td>
<td>3600</td>
<td>2800</td>
<td>2700</td>
</tr>
<tr>
<td>random</td>
<td>8900</td>
<td>3600</td>
<td>3400</td>
</tr>
</tbody>
</table>

Estimate $\text{var}(\hat{T}_c - T)$ by simulation, using a parametric bootstrap.
3. Summary

For design-based inference with acorn data:

- Systematic sampling is unknowably most precise. In absence of trend, other designs give conservative estimates
- 4-rep systematic and 2-sample stratified designs produce shortest confidence intervals

Model-based inference reduces design effects, but at the cost of subjective assumptions

- In new applications, what is best design, best $\hat{T}$ and $\text{vâr}(\hat{T} - T)$?
- Are there other designs with unbiased $\text{vâr}$?
- How should we deal with trend?