Lack-of-Balance and R-indicators as Measures of Utility in Statistical Disclosure Control

Michael Carlson, Dep. of Statistics, Stockholm University
Ingegerd Jansson, Statistics Sweden
Peter Lundquist, Statistics Sweden
The idea

- Statistical disclosure control (SDC) methods typically alter the original data in some way. “Error” is introduced to the data, adding to the Total Survey Error (TSE).

- Recent research on Lack-of-Balance and R-indicators as measures of the usefulness of a sample with non-response (Schouten, Cobben and Bethlehem, 2009; Särndal, 2010).

- Is it possible to utilize these ideas to assess the usefulness/utility of a protected data set?

- Can we monitor the behavior in a data collecting process?
Overall response rate

A probability sample \( s \) is drawn from the population \( U \).

The inclusion probability of unit \( k \) is \( \pi_k = \Pr(k \in s) \) with the design weight \( d_k = 1/\pi_k \).

The weights may be calibrated \( \omega_k = d_k m_k \).

The response set is \( r; \quad r \subseteq s \subseteq U \).

The overall response rate is \( P = \sum_r d_k / \sum_s d_k \).
Relative difference RDF

\( y = \) Income; register variable and known for \( s \).

\( x = \) auxiliary vector (the \( x \)-vector) of dimension 24:

(Age x Gender, High education, Swedish origin, Region, Married, Children)

\[
RDF(\hat{Y}_{CAL}) = 100 \times (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}
\]

\[
RDF(\hat{Y}_{EXP}) = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}
\]

where

\[
\hat{Y}_{FUL} = \sum_s d_k y_k
\]

\[
\hat{Y}_{CAL} = \sum_r d_k m_k y_k \quad m_k = \left(\sum_s d_k x_k\right)' \left(\sum_r d_k x_k x_k'\right)^{-1} x_k
\]

\[
\hat{Y}_{EXP} = \left(\sum_s d_k\right) \sum_r d_k y_k / \left(\sum_r d_k\right)
\]
Balance indicators 1

Define \( \mathbf{D} = \bar{x}_r - \bar{x}_s = (D_1, \ldots, D_j, \ldots, D_J)' \)

Under perfect balance, \( \mathbf{D} = 0 \), though normally \( \mathbf{D} \neq 0 \)

A univariate measure of imbalance is defined by the quadratic form

\[
\mathbf{D}'\Sigma_s^{-1}\mathbf{D} = (\bar{x}_r - \bar{x}_s)'\Sigma_s^{-1}(\bar{x}_r - \bar{x}_s)
\]

where

\[
\bar{x}_r = \frac{\sum_{k \in r} d_k x_k}{\sum_{k \in r} d_k}, \quad \bar{x}_s = \frac{\sum_{k \in s} d_k x_k}{\sum_{k \in s} d_k}, \quad \Sigma_s = \frac{\sum_{k \in s} d_k x_k x_k'}{\sum_{k \in s} d_k}
\]

Increased mean differences \( D_j \) tend to increase \( \mathbf{D}'\Sigma_s^{-1}\mathbf{D} \)
Balance indicators 2

- Balance indicator: \[ BI = 1 - \sqrt[Q-1]{D\Sigma_s^{-1}D} \]

- Distance measure \[ dist = [(\bar{x}_r - \bar{x}_{s-r})'\Sigma_s^{-1}(\bar{x}_r - \bar{x}_{s-r})]^{1/2} \]

where \[ \bar{x}_{s-r} = \sum_{k \in s-r} d_k x_k / \sum_{k \in s-r} d_k \]

\[ \bar{x}_r = \sum_{k \in r} d_k x_k / \sum_{k \in r} d_k, \quad \Sigma_s = \sum_{k \in s} d_k x_k x_k' / \sum_{k \in s} d_k \]
Risk assessment and record suppression

- μ-Argus approach
- The individual risk per combination $k$ of the key variables is computed (Benedetti and Franconi, 1998)

**Individual risk:**

$$ r_k = E(F_k^{-1} | f_k) = \sum_{k \in S} h^{-1} \text{Pr}(F_k = h | f_k) $$

$$ \hat{r}_k = \frac{\hat{p}_k^{f_k}}{f_k} 2 F_1(f_k; f_k; f_k + 1; 1 - \hat{p}_k) \quad \hat{p}_k = \frac{f_k}{\sum_{i:k(i)=k} w_i} $$

- Re-identification rate:

$$ \hat{\xi} = \frac{1}{n} \sum_{k=1}^{K} f_k \hat{r}_k $$

- Select a threshold value on the re-identification rate and suppress records (crude) that contribute most to $\hat{\xi}$
Illustration based on survey data 1

Synthetic data set created from real data

- $N = 182,200$, $n = 9,110$ (SRS)
- $dk = N/n = 20$

The variable of interest is $y_k = \text{income}$ and the auxiliary vector is available for all objects in the sample $S$.

The survey has a four week long data collection period. For each week we compute the indicators based on the cumulative response set

$$r^{(\text{week 1})} \subseteq r^{(\text{week 2})} \subseteq r^{(\text{week 3})} \subseteq r^{(\text{week 4})}$$
Illustration based on survey data 2

- **Key variables:**
  - Gender(2), Children(2), Region(10), Age(80), Married(2)
  - Total no. of possible combinations = 6,400

- For each week we compute the individual risks based on the $\mu$-Argus approach and identify the risk records/cells

- **Threshold = 3.5 % re-identification rate $\hat{\xi}$**

<table>
<thead>
<tr>
<th>Data collection</th>
<th>$m_r$</th>
<th>Max $\hat{r}_k$</th>
<th>Max $\hat{\xi}$</th>
<th>Unsafe records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>1487</td>
<td>0.1869</td>
<td>2.82 %</td>
<td>2 **</td>
</tr>
<tr>
<td>Week 2</td>
<td>4067</td>
<td>0.1627</td>
<td>3.54 %</td>
<td>11</td>
</tr>
<tr>
<td>Week 3</td>
<td>4863</td>
<td>0.1650</td>
<td>3.59 %</td>
<td>175</td>
</tr>
<tr>
<td>Week 4</td>
<td>5484</td>
<td>0.1645</td>
<td>3.64 %</td>
<td>525</td>
</tr>
</tbody>
</table>
Results *before* protection/suppression

- Response set and rate (in per cent); balance indicator and distance measure; relative differences. The computations are based on the auxiliary $x$-vector.

| Data collection | $m_r$ | $100 \times P$ | $BI$ | $dist_{r|nr}$ | $RDF(\hat{Y}_{EXP})$ | $RDF(\hat{Y}_{CAL})$ |
|-----------------|-------|----------------|------|---------------|---------------------|---------------------|
| Week 1          | 1487  | 16.3           | 0.835| 0.446         | 6.92                | 1.50                |
| Week 2          | 4067  | 44.6           | 0.796| 0.409         | 8.84                | 2.86                |
| Week 3          | 4863  | 53.4           | 0.791| 0.418         | 9.14                | 3.58                |
| Week 4          | 5484  | 60.2           | 0.789| 0.431         | 8.34                | 3.62                |
Results after protection/suppression

- Response set and rate (in per cent); balance indicator and distance measure; relative differences. The computations are based on the auxiliary x-vector.

| Data collection | $m_{rc}$ | $100 \times P$ | $BI$ | $dist_{r|nr}$ | $RDF(\hat{\gamma}_{\text{EXP}})$ | $RDF(\hat{\gamma}_{\text{CAL}})$ |
|-----------------|---------|----------------|------|---------------|-------------------------------|-------------------------------|
| Week 1          | 1485    | 16.3           | 0.837| 0.442         | 7.06                          | 1.52                          |
| Week 2          | 4056    | 44.5           | 0.802| 0.398         | 9.13                          | 3.08                          |
| Week 3          | 4688    | 51.5           | 0.853| 0.295         | 13.21                         | 8.11                          |
| Week 4          | 4959    | 54.4           | 0.881| 0.240         | 19.81                         | 18.05                         |
Results

**Response rate (100×P)**

- **Week**: 1, 2, 3, 4
- **Response rate (%)**: 0, 10, 20, 30, 40, 50, 60, 70
- **Graph**:
  - **Unsafe** (red line)
  - **Safe** (green line)

**Relative difference (RDF)**

- **Week**: 1, 2, 3, 4
- **Relative difference (%)**: 0, 5, 10, 15, 20, 25
- **Graph**:
  - **Exp.; Safe** (green line, solid)
  - **Exp.; Unsafe** (red line, dashed)
  - **Cal.; Safe** (green line, dashed)
  - **Cal.; Unsafe** (red line, solid)
Results

Balance indicator BI

Distance measure
Conclusions and open questions

- How can the proposed strategy help in decision making in a statistics production system? What criteria or thresholds should be used?
- Suppression can be viewed as “non-response”; but what is the relevance of this idea and the results shown here for other methods of protection?
- The uncertainty introduced by SDC in relation to other sources of uncertainty (TSE paradigm)?
- The contradictory results need to be explored and better understood. More research is needed!