What Shall we do

With the Ratios?

Sarah GIESSING,
Federal Statistical Office of Germany
Division Mathematical Statistical Methods
Recall: Consistent post-tabular multiplicative noise method (Giessing, 2012)

- Create „seed variable“ in the micro data
- Compute consistent univariate random numbers (Seed) at the table cell level
- Apply consistent multiplicative noise (using Seed) at the table cell level
- Round noisy data (rounding interval ≈ confidence interval)

= Consistency
= Transparency
= Safety
Recall: Seed based multiplicative noise at the table cell level

- For each cell/variable $Y$ do...
  - Determine deviation sense $d$ (using seed)
  - Check primary sensitivity (f.i. by $p\%$-rule)
    - $\mu=2p$; for sensitive cells
    - $\mu=0$; for non-sensitive cells
Recall: Seed based multiplicative noise at the table cell level

- For each cell/variable $Y$ do...
  - Determine deviation sense $d$ (using seed)
  - Check primary sensitivity (f.i. by $p\%$-rule)
    - $\mu=2p$; for sensitive cells
    - $\mu=0$; for non-sensitive cells
  - Draw $z$ from $N(0, \sigma)$ distribution (using seed; $\sigma \approx p/k$)
  - Multiply largest contribution $y_1$ by noise
    - $y_{1\text{post}} := y_1 (1+d (\mu+\text{abs}(z)))$
Recall: Seed based multiplicative noise at the table cell level

- For each cell/variable $Y$ do...
  - Determine deviation sense $d$ (using seed)
  - Check primary sensitivity (f.i. by p%-rule)
    - $\mu=2p$; for sensitive cells
    - $\mu=0$; for non-sensitive cells
  - Draw $z$ from $N(0, \sigma)$ distribution (using seed; $\sigma \approx p/k$)
  - Multiply largest contribution $y_1$ by noise
    - $y_{1,post} =: y_1 (1+d (\mu+\text{abs}(z)))$
  - Exchange true largest contribution by perturbed largest contribution
    - $Y_{post} = Y_{orig} - y_1 + y_{1,post}$
Recall: Seed based multiplicative noise at the table cell level

- For each cell/variable $Y$ do...
  - Determine deviation sense $d$ (using seed)
  - Check primary sensitivity (f.i. by p%-rule)
    - $\mu = 2p$; for sensitive cells
    - $\mu = 0$; for non-sensitive cells
  - Draw $z$ from $N(0, \sigma)$ distribution (using seed; $\sigma \approx p/k$)
  - Multiply largest contribution $y_1$ by noise
    - $y_1^{post} =: y_1 \cdot (1 + d(\mu + \text{abs}(z)))$
  - Exchange true largest contribution by perturbed largest contribution
    - $Y^{post} = Y^{orig} - y_1 + y_1^{post}$
  - Release $Y^{post}$ and its approximate (!) $\zeta_\gamma$-confidence interval
    - $Y^{post} \pm y_1(\mu + \sigma \zeta_\gamma)$
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y^{post} / X^{post}$ from perturbed data

Assume: $X = 1000, x_1 = 300, z_y = z_x = \sigma = 0.05, d_y = -1$
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y^{post} / X^{post}$ from perturbed data

Assume: $X = 1000, x_1 = 300, z_y = z_x = \sigma = 0.05, d_y = -1$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$y_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
</tr>
</tbody>
</table>
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y_{post}/X_{post}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_y = z_x = \sigma = 0.05$, $d_y = -1$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$y_1$</th>
<th>$Y_{post}$</th>
<th>$lb_y$</th>
<th>$ub_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>19.65</td>
<td>19.30</td>
<td>20</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>196.5</td>
<td>193.00</td>
<td>200</td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>1965</td>
<td>1930.00</td>
<td>2000</td>
</tr>
</tbody>
</table>

$\xi_y = 1$
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y_{post} / X_{post}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_y = z_x = \sigma = 0.05$, $d_y = -1$

<table>
<thead>
<tr>
<th>Y</th>
<th>$y_1$</th>
<th>$Y_{post}$</th>
<th>$lb_y$</th>
<th>$ub_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>19.65</td>
<td>19.30</td>
<td>20</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>196.5</td>
<td>193.00</td>
<td>200</td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>1965</td>
<td>1930.00</td>
<td>2000</td>
</tr>
</tbody>
</table>

$d_x = -1$

$d_x = +1$
**What happens to ratios?**

**Example:**

Ratio \( Y/X \) from true data vs.

Ratio \( Y^{post} / X^{post} \) from perturbed data

Assume: \( X = 1000, \ x_1 = 300, \ z_y = z_x = \sigma = 0.05, \ d_y = -1 \)

<table>
<thead>
<tr>
<th>( Y )</th>
<th>( y_1 )</th>
<th>( Y^{post} )</th>
<th>( \xi_y = 1 )</th>
<th>( \xi_y = 1 )</th>
<th>( d_x = -1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>19.65</td>
<td>lb(_y) 19.30</td>
<td>ub(_y) 20</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>196.5</td>
<td>lb(_y) 193.00</td>
<td>ub(_y) 200</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>1965</td>
<td>lb(_y) 1930.00</td>
<td>ub(_y) 2000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( X^{post} )</th>
<th>( \xi_y = 1 )</th>
<th>( \xi_y = 1 )</th>
<th>( d_x = +1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>lb(_x) 985</td>
<td>wh 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ub(_x) 985</td>
<td>wh 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lb(_x) 1015</td>
<td>wh 1030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ub(_x) 1015</td>
<td>wh 1030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lb(_x) 1015</td>
<td>wh 1030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ub(_x) 1015</td>
<td>wh 1030</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y^{post} / X^{post}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_y = z_x = \sigma = 0.05$, $d_y = -1$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$y_1$</th>
<th>$Y/X$</th>
<th>$Y^{post}$</th>
<th>$\xi_y = 1$</th>
<th>$X^{post}$</th>
<th>$\xi_y = 1$</th>
<th>$d_x = -1$</th>
<th>$\xi_y = 1$</th>
<th>$d_x = +1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>0.02</td>
<td>19.65</td>
<td>19.30</td>
<td>985</td>
<td>970</td>
<td>1999</td>
<td>19.93</td>
<td>19.93</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>0.2</td>
<td>196.5</td>
<td>193.00</td>
<td>985</td>
<td>970</td>
<td>19.93</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>2</td>
<td>1965</td>
<td>1930.00</td>
<td>985</td>
<td>970</td>
<td>1.9</td>
<td>1.9</td>
<td></td>
</tr>
</tbody>
</table>

[Table with additional columns and data for $Y^{post}$ and $X^{post}$ with $d_x = -1$ and $d_x = +1$]
What happens to ratios?

Example: Ratio $Y/X$ from true data vs. Ratio $Y_{\text{post}} / X_{\text{post}}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_Y = z_X = \sigma = 0.05$, $d_Y = -1$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$y_1$</th>
<th>$Y/X$</th>
<th>$Y_{\text{post}}$</th>
<th>$\xi_Y = 1$</th>
<th>$Y_{\text{post}}/X_{\text{post}}$</th>
<th>$\xi_Y = 1$</th>
<th>$d_x=-1$</th>
<th>$ul_Y$</th>
<th>$lb_X - ub_X$</th>
<th>$ub_Y$</th>
<th>$lb_Y - ub_Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>0.02</td>
<td>19.65</td>
<td>19.30-20</td>
<td>$0.0199$</td>
<td>193.00-2000</td>
<td>19.95</td>
<td>0.0013</td>
<td></td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>0.2</td>
<td>196.5</td>
<td>193.00-200</td>
<td>$0.1995$</td>
<td>1930.00-2000</td>
<td>1.99</td>
<td>0.0132</td>
<td></td>
<td>0.0130</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>2</td>
<td>1965</td>
<td>1930.00-2000</td>
<td>$1.936$</td>
<td>1930.00-2000</td>
<td>1.936</td>
<td>0.1262</td>
<td></td>
<td>0.1260</td>
<td></td>
</tr>
</tbody>
</table>
Can we do better?

\[ Y/X = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \hat{\beta}, \text{ where} \]

- \( \hat{\beta} \): OLS estimate for \( \beta \) in \( \sum_{i=1}^{n} (y_i - \beta x_i) = 0 \), i.e.

- \( \hat{\beta} \) solution to \( H(\beta) = 0 \), where

\[ H(\beta) = \sum_{i=1}^{n} (y_i - E(y_i|x_i)) \]
Can we do better?

\[ Y/X = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \hat{\beta}, \text{ where} \]

- \( \hat{\beta} \): OLS estimate for \( \beta \) in \( \sum_{i=1}^{n} (y_i - \beta x_i) = 0 \), i.e.
  - \( \hat{\beta} \) solution to \( H(\beta) = 0 \), where
    - \( H(\beta) = \sum_{i=1}^{n} (y_i - E(y_i|x_i)) \)
    - \( = \beta x_i \)
Can we do better?

- (Chipperfield and Yu, 2011):
  - Instead of solving $H(\beta) = 0$:
  - **Solve** $H(\beta) = E^*$ and release the solution $\hat{\beta}^*$ where
    
    $$E^* = u \cdot e = u \cdot \max_i |y_i - \beta x_i|$$

- Estimate $\hat{\beta} = Y/X$. Then
  - $H(\beta) = E^*$ (i.e. $\sum_{i=1}^{n} \{y_i - \beta x_i \} = E^*$) yields solution $\hat{\beta}^*$:
    
    $$\hat{\beta}^* = \frac{\sum_{i=1}^{n} y_i - u \cdot e}{\sum_{i=1}^{n} x_i} = \frac{Y}{X} - \frac{u \cdot e}{X} = \hat{\beta} - u \frac{e}{X}$$

- Draw noise $u$ from $N(0, \sigma_r)$
- Release $\hat{\beta}^*$ and its approximate $\zeta_\gamma$–confidence interval
  - $\hat{\beta}^* \pm \sigma_r \zeta_\gamma \frac{e}{X}$
We can do better!

Example: Ratio $Y/X$ from true data vs. Ratio $Y_{post}/X_{post}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_y = z_x = \sigma = \sigma_r = 0.05$, $d_y = -1$, $u = 1*0.05=0.05$

<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
<th>$y_1$</th>
<th>$Y/X$</th>
<th>$Y_{post}$</th>
<th>$\xi_y = 1$</th>
<th>$X_{post}$</th>
<th>$\xi_y = 1$</th>
<th>$d_x = -1$</th>
<th>$Y_{post}/X_{post}$</th>
<th>$\frac{lb_y}{ub_y} - \frac{lb_y}{ub_y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>0.02</td>
<td>19.65</td>
<td>19.30</td>
<td>20</td>
<td>985</td>
<td>970</td>
<td>1000</td>
<td>0.0199</td>
<td>0.0013</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>0.2</td>
<td>196.5</td>
<td>193.00</td>
<td>200</td>
<td>985</td>
<td>970</td>
<td>1000</td>
<td>0.1995</td>
<td>0.0132</td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>2</td>
<td>1965</td>
<td>1930.00</td>
<td>2000</td>
<td>985</td>
<td>970</td>
<td>1000</td>
<td>1.99</td>
<td>0.1319</td>
</tr>
</tbody>
</table>

\[
\hat{\beta}^* = 2\sigma_r \frac{e}{x} = 0.019850\quad 0.0003
\]

© Federal Statistical Office of Germany, Division Mathematical Statistical Methods
We can do better!

Example: Ratio $Y/X$ from true data vs. Ratio $Y^{post} / X^{post}$ from perturbed data

Assume: $X = 1000$, $x_1 = 300$, $z_y = z_x = \sigma = \sigma_r = 0.05$, $d_y = -1$, $u = 1\times0.05 = 0.05$

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$y_1$</th>
<th>$Y/X$</th>
<th>$Y^{post}$</th>
<th>$\xi_y = 1$</th>
<th>$\xi_y = 1$</th>
<th>$d_x = -1$</th>
<th>$\frac{ub_Y}{lb_X} - \frac{lb_Y}{ub_X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>7</td>
<td>0.02</td>
<td>19.65</td>
<td>19.30</td>
<td>20</td>
<td>0.0199</td>
<td>0.0013</td>
</tr>
<tr>
<td>200</td>
<td>70</td>
<td>0.2</td>
<td>196.5</td>
<td>193.00</td>
<td>200</td>
<td>0.1995</td>
<td>0.0132</td>
</tr>
<tr>
<td>2000</td>
<td>700</td>
<td>2</td>
<td>1965</td>
<td>1930.00</td>
<td>2000</td>
<td>1.99</td>
<td>0.1319</td>
</tr>
</tbody>
</table>

$\hat{\beta}^* = \frac{e}{2\sigma_r x}$

$e$

3

30

300
Summary: Seed based multiplicative noise for a ratio $Y/X$

For each cell/ratio $Y/X$ where $X$ and $Y$ are non-sensitive do…

- Determine deviation sense $d$ (using seed)
- Draw $u$ from $\mathcal{N}(0, \sigma_r)$ distribution (using seed; suitable $\sigma_r$)
Summary: Seed based multiplicative noise for a ratio \( Y/X \)

For each cell/ratio \( Y/X \) where \( X \) and \( Y \) are non-sensitive do…

- Determine deviation sense \( d \) (using seed)
- Draw \( u \) from \( N(0, \sigma_r) \) distribution (using seed; suitable \( \sigma_r \))
- Identify most influential observation \( e = \max \left\{ \left| y_i - \frac{Y}{X} x_i \right| \right\} \)
- Compute perturbed ratio:
  - \( \hat{\beta}^* = \frac{Y}{X} - u \frac{e}{X} \)
Summary: Seed based multiplicative noise for a ratio \( Y/X \)

For each cell/ratio \( Y/X \) where \( X \) and \( Y \) are non-sensitive do...

- Determine deviation sense \( d \) (using seed)
- Draw \( u \) from \( \mathcal{N}(0, \sigma_r) \) distribution (using seed; suitable \( \sigma_r \) )
- Identify most influential observation \( e = \max_i \left| y_i - \frac{Y}{X} x_i \right| \)
- Compute perturbed ratio:
  \[ \beta^* = \frac{Y}{X} - u \frac{e}{X} \]
- Compute width of the \( \zeta_\gamma \)-confidence interval \[ 2\sigma_r \xi_{\gamma} \frac{e}{X} \]
- Compute rounding basis \( B \), considering this width and the desired precision of the ratio
- Release the estimate \( \beta^* \) after rounding
Rounding step

Recall: Rounding for $Y^{post}$

- Calculate rounding basis $B$:
  - $B = 10 \times \text{Round}(\log_{10}(2y_1(\mu + \sigma \zeta \gamma)))$
- Round $Y^{post}$ to adjacent multiple of $B$

Example:
- $B = 10 \times \text{Round}(\log_{10}(14 \times 0.05)) = 1$
- Round $Y^{post} = 19.65$ to 20
Rounding step for perturbed ratio

Round $\hat{\beta}^*$, considering desired precision $d$:

- Calculate transformed „confidence interval“ rounding basis $B$:
  
  \[ B = 10^{\text{round} \left( \log_{10} \left( 2\sigma_r e^{10^d} \right) \right)} \]

- Transform estimate: $10^d \hat{\beta}^*$

- Round transformed estimate to adjacent multiple of $B$ and divide by $10^d$

- Example:
  
  \[ B = 10^{\text{round} \left( \log_{10} \left( 2 \cdot 0.05 \cdot \frac{3}{1000} \cdot 10^4 \right) \right)} = 1 \]

  - Round $10^d \hat{\beta}^*$ = 198.5 to 199
  - Backtransformation: $199/10^4 = 0.199$

- For the data of the example:

<table>
<thead>
<tr>
<th>$\hat{\beta}^*$</th>
<th>$d$</th>
<th>$10^d \hat{\beta}^*$</th>
<th>round($\hat{\beta}^*$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01985</td>
<td>4</td>
<td>198.5</td>
<td>0.0199</td>
</tr>
<tr>
<td>0.19850</td>
<td>4</td>
<td>1985</td>
<td>0.199*</td>
</tr>
<tr>
<td>1.98500</td>
<td>2</td>
<td>19850</td>
<td>1.99</td>
</tr>
</tbody>
</table>
Secondary Disclosure Risk Issues

Users/Intruders approximately know those intervals:

I. \( \hat{\beta}^* - \sigma_r \xi_y \frac{e}{X} \leq \frac{Y}{X} \leq \hat{\beta}^* + \sigma_r \xi_y \frac{e}{X} \)

II. \( X^* - \sigma \xi_y x_1 \leq X \leq X^* + \sigma \xi_y x_1 \)

III. \( Y^* - \sigma \xi_y y_1 \leq Y \leq Y^* + \sigma \xi_y y_1 \)

They might

- Use II to calculate new interval for true \( Y \) from I
- Use III to calculate new interval for true \( X \) from I

Therefore: we should show that it is

- possible to choose \( \sigma_r \) such that the new intervals for \( Y \) and \( X \) are (very likely to be) at least as wide as the intervals II and III

Has been proven in the paper for non-sensitive \((X,Y)\)
Conclusion, Future Work

In the context of a flexible table server relying on post tabular stochastic noise as SDC method:

- Noise variance of ratios can be reduced when the server offers perturbed ratio data (instead of calculating ratios from the perturbed enumerator and denominator data).
- Future work to address the special case where $X$ is the number of observations, i.e. the case
  - $Y/X = \text{mean}(Y)$
Thanks for your attention