Small scale “big data” in the Finnish pharmaceutical product index compilation

Ottawa Group –conference / Eltville, Germany
Kristiina Nieminen
10th May 2017
Content

1. Background and introduction of the data
2. The practices
   1. Define the index compilation strategy
   2. Standardise data collection with metadata
3. The test calculations and the results
   1. Results from current calculation
   2. Index formula tests by Vartia & Suoperä
   3. The chain-drift –test
4. Conclusions
1. Background

- First attempt to utilise the transaction data in year 2000
  - Daily products from selected commodity groups
- Eurostat’s venture on ”Modernisation of price collection and compilation”
  - Recommendations for obtaining and processing the scanner data
  - Facilitates the EU-members in the introduction of scanner-data
- New project in 2014-2016
  - Re-design of data collection >> scanner-data and web-scraping
  - Re-design of the index compilation
- Results of the project
  - Pharmaceutical products data implemented into production in the beginning of year 2017
  - Test calculations with superlative index formulas
1. Introduction of the data

- Source: Pharmaceutical Information Centre
- Pharmaceutical products for eCOICOP-groups
- Medicine prices are regulated
  - No discounts
- All products are identified with VNR-code
  - No relaunches
- Monthly delivery of prices, quantities and descriptive information by product
  - 10 000 individual product in a month, 32 variables
- Aim is to utilise as much of the data as possible
2.1 Practices: The definition of compilation strategy

The purpose for using the index:

1. the characterisation of the commodities >> described in slide 4
2. the reference group of economic actors >> consumers
3. the length of the time periods >> one month

The technical problems of index calculation:

4. the classification applied to the commodities >> COICOP
5. the collection method >> complete microdata collected
6. the appropriate weight structure >> relative value shares of the previous year by commodity

The index calculation methods should be decided by specifying:

7. the index formula >> Log-Laspeyres (elementary aggregates)
8. the strategy for constructing the index series >> Chain method where relative price changes of consecutive months are calculated for each VNR-commodity. These changes are aggregated together with value share weights. Price comparison is made for those commodities that belong to the two year panel data

The special challenges:

9. Quality changes in commodities >> no quality change
10. New and disappearing commodities >> price for disappearing commodities is estimated by calculating the average change by strata >> new commodities are introduced in the next update of panel data
2.2 Practices: The utilisation of metadata in data collection

Take original data and complement it with metadata. Utilise this information in design of data processing.
Pre-analysis report

Source Data: /TKSAS/SASDATA/Tilastot/khi/Import//DWFIN_Prices.csv

Pre-analysis report based on the data description:

### Observation count

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 106</td>
<td></td>
</tr>
</tbody>
</table>

### Key figures for numerical variables

<table>
<thead>
<tr>
<th>Obs variable</th>
<th>variablename in Finnish</th>
<th>obs</th>
<th>missing</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>date</td>
<td>10 106</td>
<td>0</td>
<td>20 910.00</td>
</tr>
<tr>
<td>2</td>
<td>pricenotax</td>
<td>9 998</td>
<td>108</td>
<td>237.03</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>9 998</td>
<td>108</td>
<td>260.74</td>
</tr>
<tr>
<td>10</td>
<td>substitutiongroup</td>
<td>5 582</td>
<td>4 524</td>
<td>968.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Character variable frequencies

<table>
<thead>
<tr>
<th>Obs variable</th>
<th>variablename in Finnish</th>
<th>obs</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>compensation</td>
<td>10 106</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>reimbursementcodes</td>
<td>9 788</td>
<td>318</td>
</tr>
<tr>
<td>3</td>
<td>reimbursementnumber</td>
<td>3 513</td>
<td>6 593</td>
</tr>
<tr>
<td>4</td>
<td>vnr</td>
<td>10 106</td>
<td>0</td>
</tr>
</tbody>
</table>

### Check of classification values

<table>
<thead>
<tr>
<th>reimbursementcodes</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEK. LRPK</td>
<td>38</td>
<td>0.39</td>
<td>38</td>
<td>0.39</td>
</tr>
<tr>
<td>AEK. PK</td>
<td>1372</td>
<td>14.41</td>
<td>1410</td>
<td>14.41</td>
</tr>
<tr>
<td>AEK. PK. YEK</td>
<td>86</td>
<td>0.88</td>
<td>1496</td>
<td>15.28</td>
</tr>
<tr>
<td>EK</td>
<td>4805</td>
<td>49.09</td>
<td>6301</td>
<td>64.37</td>
</tr>
</tbody>
</table>
3.1 Results from current calculation

Compilation of elementary indices
• According to the strategy definition (slide 5)
  • Two year panel
  • Paired comparison of the prices of base and comparison periods
  • relative change in prices is estimated for each commodity
  • Laspeyres used in aggregation
• Results:
  • over-the-counter medicine prices have grown by almost 12.5 per cent between 2009/1 and 2016/12
  • comparison between new index series and the published index series tells another story
3.1 Results from current calculation
3.2 Index formula tests by Vartia & Suoperä

• Tests were accomplished in joint-work of professor Yrjö Vartia and methodologist Antti Suoperä

• Most popular index numbers were analysed
  – At first comparison between old and new weights: Laspeyreyes, Paasche etc.
  >> so called Fisher-Five-tined fork
  – Then superlative index formulas: Fisher, Törnqvist, Stuvel, Diewert, Sato & Vartia, and Montgomery & Vartia

• Aim was to treat new and disappearing commodities in systematic and simple way

• Before calculations data was split in two groups:
  – 5S – commodities with larger relative change in values
  – 5N – commodities where values stay constant
3.2 Index formula tests by Vartia & Suoperä

The Six-tined fork represented by Vartia and Suoperä

- Laspeyres
- Log-Laspeyres
- Harm Laspeyres
- Palgrave
- Log-Paasche
- Paasche

Graph showing the index formula tests from 2014,7 to 2015,6.
3.2 Index formula tests by Vartia & Suoperä

Results from the tests of superlative index formula by Vartia and Suoperä
3.3 The test of chain-drift

- **Aim** was to analyse existence of the chain-drift and to construct new method that eliminates the chain drift phenomenon

- **Following strategies were used:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>Sample strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Törnqvist (1)</td>
<td>$t_{Base}^{t/0} = \exp \left{ \frac{1}{2}(w_i^0 + w_i^t) \log \left( \frac{p_i^t}{p_i^0} \right) \right}$</td>
<td>commodity set ${a_1, a_2, ..., a_n}$ excluding new and disappearing commodities</td>
</tr>
<tr>
<td>Chain Törnqvist (2)</td>
<td>$t_{Chain}^{t/(t-1)} = \exp \left{ \frac{1}{2}(w_i^{t-1} + w_i^t) \log \left( \frac{p_i^t}{p_i^{t-1}} \right) \right}$</td>
<td>commodity set ${a_1, a_2, ..., a_n}$ excluding new and disappearing commodities</td>
</tr>
<tr>
<td>Chain Törnqvist (3)</td>
<td>$t_{Proper\ Chain}^{t/(t-1)} = \exp \left{ \frac{1}{2}(w_i^{t-1} + w_i^t) \log \left( \frac{p_i^t}{p_i^{t-1}} \right) \right}$</td>
<td>Maximum number of matched pairs in base and observation periods</td>
</tr>
<tr>
<td>Mixed Törnqvist (4)</td>
<td>In next row, below</td>
<td>All commodities except new and disappearing (base Törnqvist) + new and disappearing (price ratio)</td>
</tr>
<tr>
<td></td>
<td>$t_{Mixed}^{2/1} = \exp \left{ \frac{1}{2}(w_{Base}^1 + w_{Base}^2) \log t_{Base}^{2/1} + \frac{1}{2}(w_{N&amp;D}^1 + w_{N&amp;D}^2) \log t_{Chain,N&amp;D}^{2/1} \right}$</td>
<td></td>
</tr>
</tbody>
</table>
3.3 Existence of chain-drift -test

Comparison between alternative methods used with Törnqvist index formula for over-the-counter medicines, 2010-2016
Conclusions

A lot of experience and competence achieved

When complete datasets (e.g. scanner-data) are available
• new approaches in CPI compilation may be taken
• accuracy and reliability of CPI is improved
• superlative index formulas produce more accurate index series
  • chain-drift must be controlled

Pharmaceutical products were implemented into CPI-production in the beginning of year 2017

Finland continues the tests with new data sources:
1) the daily products data obtained from the major retail chain,
2) the alcoholic beverages obtained from monopoly owner and
3) the hardware store data obtained by web-scraping
Thank you for your attention

Kristiina Nieminen / Statistics Finland, CPI-team
Kristiina.nieminen@stat.fi