Traditional price index methods in the context of scanner data

UN GWG on Big Data for Official Statistics
Workshop on Scanner Data and Official Statistics
Kigali, Rwanda. 29 April – 1 May 2019

Delivering insight through data for a better Canada
Outline

• Definition and uses of a consumer price index
• Review of traditional price index formula
• Construction of a national consumer price index
• Preprocessing of scanner data for CPI calculation
• Scanner data and traditional price index formula
• Multilateral price index methods: general principle
Consumer price index: definition and uses

• What is the Consumer Price Index (CPI):

  • An indicator of the changes in consumer prices that are experienced by a target population. It measures *average price changes* by comparing, *through time*, the cost of a *fixed basket* of goods and services.

  • The goods and services in the basket must be of unchanging or equivalent quantities and qualities

  • Reflects a *pure price change*, **NOT** a Cost-Of-Living Index

  • Price movements of CPI product categories are *weighted* according to their relative importance in the total expenditures of consumers
Consumer price index: definition and uses

• Uses of the CPI:

  • Central banks: as a measure of inflation, for monetary policy

  • Benefits recipients, workers and unions: to index, escalate or adjust nominal values

  • Governments: budget projection, deflation of nominal values to obtain constant dollar figures

  • Other uses: financial markets and traders

• Not always easy to accommodate all specific requirements of all uses
Review of traditional price index formula

- Index formula for elementary price indices

\[ I_{\text{Carli}}^{0:t} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{p_i^t}{p_i^0} \right) \]

\[ I_{\text{Dutot}}^{0:t} = \frac{1}{n} \sum_{i=1}^{n} p_i^t \]

\[ I_{\text{Jevons}}^{0:t} = \prod_{i=1}^{n} \left( \frac{p_i^t}{p_i^0} \right)^{\frac{1}{n}} \]

<table>
<thead>
<tr>
<th></th>
<th>t=0</th>
<th>t=1</th>
<th>t=2</th>
<th>t=3</th>
<th>t=4</th>
<th>t=5</th>
<th>t=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prices (standardized)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candy 1</td>
<td>$6.00</td>
<td>$6.00</td>
<td>$7.00</td>
<td>$6.00</td>
<td>$6.00</td>
<td>$6.00</td>
<td>$6.60</td>
</tr>
<tr>
<td>Candy 2</td>
<td>$7.00</td>
<td>$7.00</td>
<td>$6.00</td>
<td>$7.00</td>
<td>$7.00</td>
<td>$7.20</td>
<td>$7.70</td>
</tr>
<tr>
<td>Candy 3</td>
<td>$2.00</td>
<td>$3.00</td>
<td>$4.00</td>
<td>$5.00</td>
<td>$2.00</td>
<td>$3.00</td>
<td>$2.20</td>
</tr>
<tr>
<td>Candy 4</td>
<td>$5.00</td>
<td>$5.00</td>
<td>$5.00</td>
<td>$4.00</td>
<td>$5.00</td>
<td>$5.00</td>
<td>$5.50</td>
</tr>
<tr>
<td>Arithmetic Mean prices</td>
<td>$5.00</td>
<td>$5.25</td>
<td>$5.50</td>
<td>$5.50</td>
<td>$5.00</td>
<td>$5.30</td>
<td>$5.50</td>
</tr>
<tr>
<td>Geometric Mean Prices</td>
<td>$4.53</td>
<td>$5.01</td>
<td>$5.38</td>
<td>$5.38</td>
<td>$4.53</td>
<td>$5.05</td>
<td>$4.98</td>
</tr>
</tbody>
</table>

Carli Month-to-Month index:
- 100.0
- 112.5
- 108.9
- 101.8
- 91.2
- 113.2
- 100.0

Carli Chained month-to-month index:
- 100.0
- 112.5
- 122.5
- 124.8
- 113.9
- 128.9
- 129.0

Carli Direct index on t=0:
- 100.0
- 112.5
- 125.6
- 132.5
- 100.0
- 113.2
- 110.0

Dutot Month-to-Month index:
- 100.0
- 105.0
- 104.8
- 100.0
- 90.9
- 106.0
- 103.8

Dutot Chained month-to-month index:
- 100.0
- 105.0
- 110.0
- 110.0
- 100.0
- 106.0
- 110.0

Dutot Direct index on t=0:
- 100.0
- 105.0
- 110.0
- 110.0
- 100.0
- 106.0
- 110.0

Jevons Month-to-Month index:
- 100.0
- 110.7
- 107.5
- 100.0
- 84.1
- 111.4
- 98.7

Jevons Chained month-to-month index:
- 100.0
- 110.7
- 118.9
- 118.9
- 100.0
- 111.4
- 110.0

Jevons Direct index on t=0:
- 100.0
- 110.7
- 118.9
- 118.9
- 100.0
- 111.4
- 110.0
Review of traditional price index formula

• Index formula for elementary price indices

• Are chained and direct indexes equal?
  • Carli index: Chained ≠ Direct
  • Dutot and Jevons index: Chained = Direct
  • So avoid chained Carli index!

• When prices go back to base level (t=4 and t=0), index level should go back to 100
  • Chained Carli index fails this property!

• By default, G20 countries – except Japan, use, generally a Chained Jevons index rather than a Chained Dutot
  • Dutot only works well for very homogeneous products, which means for EAs that are very narrowly defined and products that have the same unit of measure
## Review of traditional price index formula

- Index formula for aggregate price indices

<table>
<thead>
<tr>
<th>Index name</th>
<th>Laspeyres</th>
<th>Paasche</th>
<th>Fischer</th>
<th>Törnqvist</th>
<th>Lowe</th>
</tr>
</thead>
</table>
| Formula     | \[
I_{L,A}^{0t} = \sum_{i=1}^{n} \frac{p_i^t q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \]
|             | \[
I_{P,A}^{0t} = \sum_{i=1}^{n} \frac{p_i^t q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \]
|             | \[
I_{L,A}^{\infty} = \sum_{i=1}^{n} s_i^0 \left( \frac{p_i^0}{p_i} \right) \]
|             | \[
s_i^0 = \frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \]
|             | \[
I_{P,A}^{\infty} = \sum_{i=1}^{n} \frac{s_i^0 \left( \frac{p_i^t}{p_i} \right)^{-1}}{s_i^i \left( \frac{p_i^0}{p_i} \right)^{-1}} \]
|             | \[
s_i^i = \frac{p_i^t q_i^i}{\sum_{i=1}^{n} p_i^t q_i^i} \]
|             | \[
I_{F,A}^{0t} = \left( I_{L,A}^{0t} \times I_{P,A}^{0t} \right)^{\frac{1}{2}} \]
|             | \[
I_{T,A}^{0t} = \prod_{i=1}^{n} \left( \frac{p_i^t}{p_i^0} \right)^{\frac{1}{2} \left( s_i^0 + s_i^i \right)} \]
|             | \[
I_{L,o,A}^{0t} = \sum_{i=1}^{n} p_i^t q_i^b \]
|             | \[
I_{P,o,A}^{0t} = \sum_{i=1}^{n} \frac{p_i^t}{p_i^0} \]
|             | \[
I_{L,o,A}^{\infty} = \sum_{i=1}^{n} s_i^{ob} \left( \frac{p_i^t}{p_i^0} \right) \]
|             | \[
s_i^{ob} = \frac{p_i^0 q_i^b}{\sum_{i=1}^{n} p_i^0 q_i^b} \]

---

Delivering insight through data for a better Canada
Review of traditional price index formula

<table>
<thead>
<tr>
<th>Item</th>
<th>Period 0</th>
<th>Price ($)</th>
<th>Quantity</th>
<th>Expenditure ($)</th>
<th>Expenditure shares</th>
<th>Price relatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>White fresh bread</td>
<td></td>
<td>2.90</td>
<td>2,000</td>
<td>5,800</td>
<td>0.3932</td>
<td>1.0000</td>
</tr>
<tr>
<td>Apples</td>
<td></td>
<td>5.50</td>
<td>500</td>
<td>2,750</td>
<td>0.1864</td>
<td>1.0000</td>
</tr>
<tr>
<td>Beer</td>
<td></td>
<td>8.00</td>
<td>200</td>
<td>1,600</td>
<td>0.1085</td>
<td>1.0000</td>
</tr>
<tr>
<td>LCD TV</td>
<td></td>
<td>1,200.00</td>
<td>2</td>
<td>2,400</td>
<td>0.1627</td>
<td>1.0000</td>
</tr>
<tr>
<td>Jeans</td>
<td></td>
<td>55.00</td>
<td>40</td>
<td>2,200</td>
<td>0.1492</td>
<td>1.0000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>14,750</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Period t</th>
<th>Price ($)</th>
<th>Quantity</th>
<th>Expenditure ($)</th>
<th>Expenditure shares</th>
<th>Price relatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>White fresh bread</td>
<td></td>
<td>3.00</td>
<td>2,000</td>
<td>6,000</td>
<td>0.4220</td>
<td>1.0345</td>
</tr>
<tr>
<td>Apples</td>
<td></td>
<td>4.50</td>
<td>450</td>
<td>2,025</td>
<td>0.1424</td>
<td>0.8182</td>
</tr>
<tr>
<td>Beer</td>
<td></td>
<td>8.40</td>
<td>130</td>
<td>1,092</td>
<td>0.0768</td>
<td>1.0500</td>
</tr>
<tr>
<td>LCD TV</td>
<td></td>
<td>1,100.00</td>
<td>3</td>
<td>3,300</td>
<td>0.2321</td>
<td>0.9167</td>
</tr>
<tr>
<td>Jeans</td>
<td></td>
<td>50.00</td>
<td>30</td>
<td>1,800</td>
<td>0.1256</td>
<td>1.0909</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>14,217</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

Laspeyres

\[
= (0.3932 \times 1.0345) + (0.1864 \times 0.8182) + (0.1085 \times 1.0500) + (0.1627 \times 0.9167) + (0.1492 \times 1.0909) \times 100
= 98.51
\]

Paasche

\[
= \frac{1}{((0.4220 / 1.0345) + (0.1424 / 0.8182) + (0.0768 / 1.0500) + (0.2321 / 0.9167) + (0.1256 / 1.0909)) \times 100}
= 97.62
\]
Review of traditional price index formula

<table>
<thead>
<tr>
<th>Item</th>
<th>Price ($)</th>
<th>Quantity</th>
<th>Expenditure ($)</th>
<th>Expenditure shares</th>
<th>Price relatives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period 0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White fresh bread</td>
<td>2.50</td>
<td>2000</td>
<td>5800</td>
<td>0.3932</td>
<td>1.0000</td>
</tr>
<tr>
<td>Apples</td>
<td>5.50</td>
<td>500</td>
<td>2750</td>
<td>0.1864</td>
<td>1.0000</td>
</tr>
<tr>
<td>Beer</td>
<td>8.00</td>
<td>200</td>
<td>1600</td>
<td>0.1085</td>
<td>1.0000</td>
</tr>
<tr>
<td>LCD TV</td>
<td>1200.00</td>
<td>2</td>
<td>2400</td>
<td>0.1627</td>
<td>1.0000</td>
</tr>
<tr>
<td>Jeans</td>
<td>55.00</td>
<td>40</td>
<td>2200</td>
<td>0.1492</td>
<td>1.0000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>14750</strong></td>
<td><strong>1.0000</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Price ($)</th>
<th>Quantity</th>
<th>Expenditure ($)</th>
<th>Expenditure shares</th>
<th>Price relatives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period t</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White fresh bread</td>
<td>3.00</td>
<td>2000</td>
<td>6000</td>
<td>0.4220</td>
<td>1.0345</td>
</tr>
<tr>
<td>Apples</td>
<td>4.50</td>
<td>450</td>
<td>2025</td>
<td>0.1424</td>
<td>0.8162</td>
</tr>
<tr>
<td>Beer</td>
<td>8.40</td>
<td>130</td>
<td>1092</td>
<td>0.0758</td>
<td>1.0500</td>
</tr>
<tr>
<td>LCD TV</td>
<td>1100.00</td>
<td>3</td>
<td>3300</td>
<td>0.2321</td>
<td>0.9167</td>
</tr>
<tr>
<td>Jeans</td>
<td>60.00</td>
<td>30</td>
<td>1800</td>
<td>0.1266</td>
<td>1.0909</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>14217</strong></td>
<td><strong>1.0000</strong></td>
<td></td>
</tr>
</tbody>
</table>

Fisher

\[
(98.51 \times 97.62)^{1/2} = 98.06
\]

Törnqvist is best calculated by first taking the logs of the index formula

\[
\begin{align*}
&= \frac{1}{2} \left(0.3932 + 0.4220\right) \times \ln(1.0345) \\
&+ \frac{1}{2} \left(0.1864 + 0.1424\right) \times \ln(0.8162) \\
&+ \frac{1}{2} \left(0.1085 + 0.0768\right) \times \ln(1.0500) \\
&+ \frac{1}{2} \left(0.1627 + 0.2321\right) \times \ln(0.9167) \\
&+ \frac{1}{2} \left(0.1492 + 0.1266\right) \times \ln(1.0909) \\
&- 0.0199
\end{align*}
\]

and then taking the exponent multiplied by 100

\[
e^{0.0199} \times 100 = 98.04
\]
Construction of a national CPI

- **Define the scope of the index**
  - Product coverage
  - Target population

- **Classifications**
  - Product classification
  - Geography classification

- **Source of expenditure weights and the frequency of their update**
  - Survey of Household Spending
  - CPI basket
  - Frequency of basket update

- **Sampling strategy**
  - Outlet sample
  - Product sample
  - Collection pattern for each product

- **Price collection**
  - In the field, by price interviewers
  - Using administrative data
  - Internet, online
  - Scanner data, etc.

- **Data editing and quality control of micro-data**
Construction of a national CPI

• CPI is built up from price indices for elementary aggregates (EA)
• Elementary aggregates are pairings of lowest level product classes and lowest level geography classes
  • Banana in geo strata 1, banana in geo strata 2 are two different EAs
  • Banana in Canada is not an EA
  • Some product classes may not be available in some geo classes
  • Canada: 695 lowest level product classes, 19 lowest level geo classes

• EAs price indices estimated by direct price observation or by imputation

• Elementary price indices:
  • In general, Jevons index formula is used in most countries
  • Prices for goods and services of same quantity and same quality observed over time

• Aggregate level indices:
  • Consumption expenditures, to give relative importance to product/geo classes
  • Expenditures must be price-updated: the expenditure value of each category is multiplied by its monthly price change. Same quantities valued at current month’s prices
  • In general, Lowe index aggregation formula is used in most countries
Preprocessing of scanner data for CPI calculation

• **Typical structure of scanner data**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Format</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Numeric or date</td>
<td>20180104, 01/04/2018</td>
</tr>
<tr>
<td>Store ID</td>
<td>Text or numeric</td>
<td>Store_0001</td>
</tr>
<tr>
<td>Store address</td>
<td>Text</td>
<td>123 ABC street, Region, postal code</td>
</tr>
<tr>
<td>Province/Region</td>
<td>Text or numeric</td>
<td>Ontario</td>
</tr>
<tr>
<td>Product identifier (UPC, SKU)</td>
<td>Numeric or text</td>
<td>ABC_0001</td>
</tr>
<tr>
<td>Retailer classification</td>
<td>Text or numeric</td>
<td>Grocery – Dairy – Cheese – Entertainment Cheese</td>
</tr>
<tr>
<td>Description</td>
<td>Text</td>
<td>‘Tasty’ Brand Brie Cheese 200g</td>
</tr>
<tr>
<td>Quantity</td>
<td>Numeric</td>
<td>61 units</td>
</tr>
<tr>
<td>Turnover</td>
<td>Numeric or currency</td>
<td>$501</td>
</tr>
</tbody>
</table>

• **Potential advantages of scanner data:**
  • Full enumeration of products sold during a given time period
    • Universe of products purchased by consumers
  
  • Quantity and turnover information
    • Actual average transaction price paid by consumers

  Large scale increase in CPI product sample size
  Lower amount of resources for price collection

• **Possible drawback: Risk of over-coverage**
  • Purchases made by businesses are included!
  • Is that really an issue?
Preprocessing of scanner data for CPI calculation

• **How is a product defined?**
  - A set of homogeneous items
  - Associated with: Global Trade Item Number (GTIN), Universal Product Code (UPC) or retailer assigned codes, Stock Keeping Unit (SKU)?
  - What if GTINs change frequently with very small or no changes in product characteristics: relaunches?

• **Data aggregation**
  - Individual transactions during a time period need to be aggregated
    - Total quantities, total turnover and average prices
  - Doing this aggregation at outlet (location) level or at retailer level?
    - Do prices differ across stores of the same retailer?
    - No, in general; but a market intelligence research recommended
  - Doing this aggregation at regional level?
    - National or regional pricing?
    - Do purchasing patterns vary by region? Regional weights may be needed!
  - For monthly CPI, how many weeks to include in the aggregation?
    - Depends on the CPI production cycle
    - Ideally, all days of the month
    - Practically, 2 or 3 weeks.
Scanner data and traditional price index formula

• Ideally, we would use a method that:
  • Uses census of products
  • Weights prices at the product (and product group) level
  • and automated processes (less resources)

• ILO/IMF Consumer Price Index manual recommends ‘superlative’ indexes (e.g. Fisher, Törnqvist) as the ideal CPI target

• Can we apply these methods directly to scanner data?

• Could use ‘direct’ or ‘chained’ weighted bilateral indexes

• However, dynamic nature of transactions data can make these methods perform poorly
Scanner data and traditional price index formula

- ‘Direct’ bilateral indexes suffer from a ‘matching’ problem
Scanner data and traditional price index formula

- Consumers responsive to sales: price and quantity bouncing can cause problems for chained indexes
Scanner data and traditional price index formula

- Chained bilateral indexes suffer from a ‘chain drift’ problem
Multilateral price index methods: general principle

• Many National Statistical Institutes (NSI) continue to use a geometric mean at elementary aggregate level
  
  • No weight used at this level
  • Mostly supermarket products
  • Belgium, Canada, Denmark, Iceland, Netherlands, Norway, Sweden, Switzerland

• A few NSI have implemented multilateral index methods for some of their CPI components:
  
  • Statistics Netherlands (CBS): Mobile phone and department store products
  • Statistics New Zealand (SNZ): Audio visual and household appliance products
  • Australia Bureau of Statistics (ABS): Food expenditure classes
Multilateral price index methods: general principle

- Bilateral index methods compare prices between two time periods

- Multilateral index methods:
  - Price comparisons across multiple (three or more) time periods
  - Historically used in constructing spatial price indexes (comparison of price levels between countries)
  - Use all matched products between any two months
  - Are “average” of multiple bilateral indexes:

    - Example:
      \[
      I_{jl} = \left( \prod_{k=0}^{T} I_{jk}I_{kl} \right)^{1/(T+1)}
      \]
      - Weight products by their economic importance (turnover)
      - Are free of ‘chain drift’
Questions?

Thank you!