Sampling of the Consumer Price Index

Patrick Sillard

INSEE – France

May 2014

Plan

Plan

The sample

- An article
- belongs to a given shop
- belongs to a given type of product (variety)
- the shop belongs to a given city
- the city belongs to a given type of city and to a territory

Why should we sample and what do we need for doing so ?

Because

- we only need to collect some products to reach a certain level of accuracy
- we only need to collect in some places and sending price collectors everywhere would have a very heavy cost

We need

- to choose the dimension where to sample
- in order to sample, we must know the universe on the dimension of sampling

The Laspeyres Aggregation

A is a partition of the households consumption ; a is a set included in this partition (for example a city for a class of products).

$$I = \sum_{a \in A} w_a I_a$$

where I_a is the price index of a and w_a is the weight of a in the total consumption. In other words :

$$\sum_{a \in A} w_a = 1$$

Sample Estimation

$$I = \sum_{a \in A} w_a I_a$$

We may estimate I...

by

$$\hat{I} = \sum_{a \in \mathscr{A}} \omega_a I_a$$

where $\mathscr{A}$ is a sample drawn in A and ω_a is a sample weight computed according to

1. the sample process (probability of inclusion of a into $\mathscr{A}$) ;
2. the proper weight w_a that a should have in the real value of I.

In which dimension should we sample ?

Dimension of sampling

The geographical dimension because :

- we know the universe,
- it is directly related to collection costs,
- we may define geographic strata in such a way that the sample precision is optimised because w_a is correlated with geographic strata.

Plan

Sampling variance minimization

$\hat{I} = \sum_{a \in \mathscr{A}} \omega_a I_a$ is an estimate

of the sum of the variable $y_a = w_a I_a$. This variable is approximately proportional to w_a. Therefore, the probability of inclusion should be proportional to w_a.

Stratification

Stratification is an additional way to improve the accuracy.

The geographic sample

- We define first some geographic strata corresponding to the crossing of 7 geographic areas and 4 types of cities according to their size (Paris / 100 000 $< pop$ / 20 000 $< pop <$ 100 000 / $pop <$ 20 000)
- In each strata h, we compute the number of cities n_h we are going to survey according to :

$$n_h = \mathcal{N} \times W_h$$

  where $\mathcal{N}$ is the total number of cities we want to survey and W_h is the weight of the h stratum ($W_h = \sum_{a \in h} w_a$).
- we draw a sample of size n_h in the stratum h. Each city a included in strata h has a probability of inclusion proportional, in this stratum, to w_a.
- Doing so, we can show that the variance is minimal, conditional to the total number $\mathcal{N}$ of surveyed cities.

A few words about the estimate $\hat{I}$ of I

$$\hat{I} = \sum_{a \in \mathscr{A}} \omega_a I_a$$

is an unbiased estimate of

$$I = \sum_{a \in A} w_a I_a$$

if and only if,

$$\omega_a = \frac{w_a}{\pi_a} \text{ and } \sum_{a \in \mathscr{A}} \pi_a = \mathcal{N}$$

where $\pi_a = \Pr(a \in \mathscr{A})$ is the probability of inclusion of a into the sample $\mathscr{A}$. This π_a is directly linked with the sample design of $\mathscr{A}$.

How to select the cities within a given stratum (1) ?

On the probability of inclusion

As seen before, we need to adopt a sample design that makes π_a proportional to a variable w_a corresponding to the weight of a in the theoretical index I.

The weight w_a

corresponds theoretically to the household final expenditure value for the part a of consumption. Considering cities, we may approximate this by demography. We may also improve this with Household Budget Survey from which we can compute the weight of city a (or for the type of city a belongs to) according to the location of purchase.

How to select the cities within a given stratum (2) ?

If we have no previous sample

The we can draw a sample according to a systematic sample design within each stratum h with a given size n_h.

If we have no previous sample and we don't want to change everything

We may draw a sample with a total probability of inclusion proportional to w_a and with a conditional probability of inclusion maximum for a subset of cities that belongs to previous sample.

How to select the cities within a given stratum (3) ?

In case of France and for the 2015 CPI rebasement, we have decided to adopt the last approach : if $\mathscr{X}$ is the new sample and $\mathscr{I}$ is the old one. Let $\pi_a^{\mathscr{I}}$ be the probability of inclusion of a in $\mathscr{I}$. We try to set $\Pr(a \in \mathscr{X} | a \in \mathscr{I})$ in such a way that -1) it is maximum and -2) the total probability of inclusion of a into $\mathscr{X}$ is a given number $\pi_a^{\mathscr{X}}$. One can show that :

- if $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$, then :

$$\begin{cases} \Pr(a \in \mathscr{X} | a \in \mathscr{I}) & = & \pi_a^{\mathscr{X}} \Big/ \pi_a^{\mathscr{I}} \\ \Pr(a \in \mathscr{X} | a \notin \mathscr{I}) & = & 0 \end{cases}$$

- if $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, then :

$$\begin{cases} \Pr(a \in \mathscr{X} | a \in \mathscr{I}) & = & 1 \\ \Pr(a \in \mathscr{X} | a \notin \mathscr{I}) & = & (\pi_a^{\mathscr{X}} - \pi_a^{\mathscr{I}}) \Big/ (1 - \pi_a^{\mathscr{I}}) \end{cases}$$

How to select the cities within a given stratum (4) ?

Finally, we sample according to the following design :

1. If $a \in \mathscr{I}$, two cases may occur :
   1. if $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$, then a is selected in $\mathscr{X}$ with a probability equal to $\pi_a^{\mathscr{X}} / \pi_a^{\mathscr{I}}$;
   2. if $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, then a is selected in $\mathscr{X}$.
2. If $a \notin \mathscr{I}$, two cases may occur :
   1. if $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$, then a is not selected in $\mathscr{X}$;
   2. if $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, then a is selected in $\mathscr{X}$ with a probability equal to $(\pi_a^{\mathscr{X}} - \pi_a^{\mathscr{I}})/(1 - \pi_a^{\mathscr{I}})$.

Doing so. . .

we maximize the probability of inclusion in the new sample for a city that was in the previous sample $\mathscr{I}$.

At this stage. . .

We have a sample $\mathscr{A}$ of cities.

Plan

The sample of varieties

Each consumption segment (COICOP group - the consumption is divided into 300 levels at the finest disaggregation level) is divided into types of products, called varieties. A variety must be understood as a **representative** of the COICOP group. We follow 1 000 varieties in France. For all of them, we can estimate the Households expenditures according to National Account data. We therefore have a weight for each of variety. Let w_v be the weight of variety v ($\sum_v w_v = 1$).

From variety to price observations (1)

We also know from professional information the share of each type of shop in the total expense of variety v. We follow 12 types of shops in the French CPI. Let $\alpha_{v,j}$ the share (marginal distribution at the national level) of the type of shop j ($\sum_{j=1}^{12} \alpha_{v,j} = 1$).
At the end, we get

$$w_{a,v} = w_v \times \frac{w_a}{\pi_a}$$

is the weight that should be adopted for a micro-aggregate $I_{a,v}$ computed with price observation referring to variety v in city a.
The share of type of shop j in the number of observations made for variety v in city a should be equal to $\alpha_{v,j}$ (the same for any city).

From variety to price observations (2)

Assuming that we must do $n_{a,v}$ observations of variety v in the city a, then the number of price observation we should do in the type of shop j (for example supermarkets) is :

$$n_{a,v,j} = n_{a,v} \times \alpha_{v,j}$$

Yearly organisation

We give, to the price collector involved in the city a, a goal to follow $n_{a,v,j}$ products in city a, for variety v in the type of shop j while we are defining the sample of products for the year Y, at the end of the year $Y-1$. the price collector must identify the products within some shops and he will come again to these shops every months during year Y to observe the prices.

Plan

Remaining questions

Questions

- What is the right number of observations per variety (n_v) ?
- What is the right number of observations in a city a for a given variety v ($n_{a,v}$) ?

Answers

Assumptions : what we know

- the time the price collector use to make a given observation (depending on the type of product)
- the variance of observations σ_v^2 for each variety

What we do : we compute n_v and $n_{a,v}$ through

- a minimization of variance
- under the constraint that the total cost of collection (overall time spent by price collectors) is the actual CPI cost of collection

Variance optimization

For that, we need :

- An expression of the variance of a micro-index $\mathrm{var}(\hat{I}_{a,v})$
- From this, we can compute the variance of :
  - $\mathrm{var}(\hat{I}_v) = \mathrm{var}\left(\sum_{a \in \mathscr{A}} \frac{w_a}{\pi_a} \hat{I}_{a,v}\right)$
  - and then $\mathrm{var}(\hat{I}) = \mathrm{var}\left(\sum_v w_v \hat{I}_v\right)$

The variance of $\hat{I}_{a,v}$ $(n \equiv n_{a,v})$ – (1)

For a Dutot index : $\hat{I}_{a,v} = \left(n^{-1} \sum_{i \in \mathscr{S}_{a,v}} p_i^t\right) \Big/ \left(n^{-1} \sum_{i \in \mathscr{S}_{a,v}} p_i^0\right)$

$$\mathrm{var}(\hat{I}_{a,v}) = \left(\frac{1}{\hat{\bar{P}}^0}\right)^2 \frac{1}{n(n-1)} \sum_{i \in \mathscr{S}_{a,v}} \left(p_i^t - \hat{I}_{a,v} p_i^0\right)^2 \equiv \frac{\widehat{\sigma^2_{a,v}}}{n}$$

with $\hat{\bar{P}}^0 = n^{-1} \sum_{i \in \mathscr{S}_{a,v}} p_i^0$ and $\mathscr{S}_{a,v}$ is the sample of products in (a, v).

The variance of $\hat{I}_{a,v}$ $(n \equiv n_{a,v})$ – (2)

For a Jevons index : $\hat{I} = \left(\prod_{i\in\mathscr{S}_{,}} p_i^t\right)^{1/n} \Big/ \left(\prod_{i\in\mathscr{S}_{a,v}} p_i^0\right)^{1/n}$

$$\mathrm{var}(\hat{I}_{a,v}) = \frac{\hat{I}_{a,v}^2}{n(n-1)} \sum_{i\in\mathscr{S}_{a,v}} \left[\ln\left(\frac{p_i^t}{p_i^0}\right) - \ln \hat{I}_{a,v}\right]^2 \equiv \frac{\widehat{\sigma_{a,v}^2}}{n}$$

where $\mathscr{S}_{a,v}$ is the sample of products in (a, v).

At the end

The number of observation $n_v^{\mathscr{A}}$ requested for a variety v is :

$$n_v^{\mathscr{A}} = \frac{C}{c_v} \times \frac{\kappa_v^{\mathscr{A}} \sqrt{c_v}}{\sum\limits_v \kappa_v^{\mathscr{A}} \sqrt{c_v}}$$

where $\kappa_v^{\mathscr{A}} = \sum_{a \in A} w_{av} \sigma_{av} / \pi_a$; C : total cost ; c_v cost for variety v.

The number of observation $n_{a,v}^{\mathscr{A}}$ requested for a variety v in a city a is :

$$n_{a,v}^{\mathscr{A}} = \frac{w_{a,v} \sigma_{a,v} / \pi_a}{\sum\limits_{a \in \mathscr{A}} w_{a,v} \sigma_{a,v} / \pi_a} \times n_v^{\mathscr{A}}$$

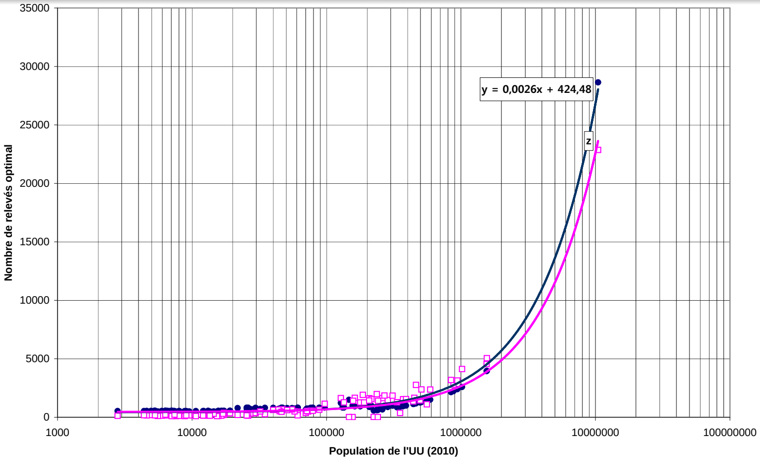
Weight of geographic strata

Geographic Stratum	Weight (in %)			
	demography (1)	Food (2)	(1) normalised without rural	(2) normalised without rural
A	16.7	18.1	21.5	19.5
B2	24.6	25.6	31.7	27.6
B1	5.3	7	6.8	7.6
C	13.6	18.6	17.5	20.1
D	17.4	23.4	22.4	25.2
Rural	22.5	7.4		
Total	100	100	100	100

Average duration of price observation

type	average duration without walk	normalised inc. walk	Number of shop per obs.
Durables g.	102s	1.30	0.27
Clothing	60s	0.86	0.19
Food	49s	0.53	0.09
Manufactured g.	78s	1.16	0.27
Services	25s	1.73	0.61

Link between the optimal number of price observations and the size of the city



Number of price observations per type of products

Secteur	Base 1998 (total)	Base 1998 (kept c.)	Base 2015
AL	35 568	31 801	49 380
BD	6 544	6 074	5 652
HA	21 033	20 451	11 449
MA	29 939	27 683	29 019
SE	19 683	18 119	21 375
Total	112 767	104 128	116 875

Note : AL=Food products, BD=durables, HA=clothing, MA=manufactured goods, SE=services ; “kept c.” means cities included in base 1998 also included in base 2015.

Plan

1. What should be sampled ?
2. The geographic sample
3. The sample of products
4. Sample optimization
5. Conclusion

Conclusion

It is possible to :

- fix all the parameters of the sample, knowing the weight of the cities
- having an idea of the variance of observation (0.1 perc. point on the yearly increase std ; possible to reach 0.02)
- having an idea of the elementary cost of observation (per type of observation)

But... for some types of product, it is not applicable

- the case of products for which the purchases are highly concentrated spatially
- the tariffs