

UNIVERSITÀ DEGLI STUDI DI BERGAMO

Facoltà di Economia

Corso di Dottorato di Ricerca in “Marketing per le strategie d’impresa”

Ciclo XXI

**Survey Techniques: an application to prices data for the  
computation of price indexes**

Tutor interno:

Prof.ssa Silvia Biffignandi

Tutor esterno:

Patak Zdenek

Tesi di Dottorato

Daniele TONINELLI

Matricola n. 700217

ANNO ACCADEMICO: 2008/09



# INDEX

INDEX.....	1
Index of figures.....	5
Index of Tables .....	5
Index of Graphs .....	6
1 Introduction: the rules of price in the marketing .....	7
1.1 Dimensions of analysis.....	8
1.1.1 Economic-organizational dimension .....	8
1.1.2 Competitive dimension.....	9
1.1.3 Relation with the demand.....	11
1.1.4 The fourth dimension and the profit margin ratios.....	13
1.2 The management's cognitive requirements.....	14
1.3 Methodological issues .....	16
1.3.1 First statistical solution: the price indexes.....	16
1.3.1.1 The price indexes.....	17
1.3.1.2 The Laspeyres index.....	19
1.3.2 Second statistical solutions: the data quality improvement.....	21
1.4 Conclusions about the aim of this research .....	22
2 Data collection: the Statistics Canada's Wholesale survey .....	25
2.1 Price indexes in Canada.....	25
2.2 The <i>SPPI</i> (Service Producer Price Index) .....	26
2.3 The characteristics of the wholesale survey .....	28
2.3.1 First stage (PSU).....	30
2.3.2 Second stage (SSU) .....	34
2.3.3 The questionnaire .....	35
2.4 First phases of analysis and state of the art .....	37
3 Simulated data generation: the generation process.....	39
3.1 The cells definition .....	41
3.1.1 The NAICS code .....	42
3.1.2 The establishment revenue .....	45

3.1.3	Some adjustments .....	47
3.1.4	Distributions of the units by cells .....	48
3.2	Testing the distribution by cells.....	48
3.2.1	Uniform distribution .....	51
3.2.2	Normal distribution .....	51
3.2.3	Lognormal distribution.....	51
3.2.4	Weibull distribution.....	52
3.2.5	Tests to identify the distribution of each cell .....	52
3.2.5.1	TEST 1: the Kolmogorov-Smirnov statistic (D) .....	53
3.2.5.2	TEST 2: the Anderson-Darling statistic ( $A^2$ ) .....	53
3.2.5.3	TEST 3: the Cramèr-von Mises statistic.....	54
3.2.5.4	An example of distribution tests.....	54
3.2.6	Criteria to identify the distributions.....	57
3.3	Generation of simulated data.....	59
3.3.1	Uniform distribution .....	61
3.3.2	Normal .....	62
3.3.3	Lognormal .....	65
3.3.4	Weibull .....	66
3.3.5	Mixed distribution .....	68
4	Relative efficiency of the methods of generation.....	71
4.1	Bias of the parameters .....	71
4.1.1	Methodology.....	72
4.1.2	Some results.....	74
4.1.2.1	Precision of not-stratified vs stratified selection methods.....	75
4.1.2.2	Simple random sampling vs systematic sampling.....	75
4.1.2.3	Relative bias of the uniformly distributed simulated populations .....	76
4.1.2.4	Bias from the sample reference parameter .....	78
4.1.2.5	The biggest units effect.....	79
4.1.2.6	Precision's improvements with more samples .....	80
4.2	Bias of the different PPS selection methods.....	83
4.2.1	PPS sampling selecting methods .....	84
4.2.1.1	PPS (Probability Proportional to Size) .....	84
4.2.1.2	PPS/Stratified .....	85
4.2.1.3	PPS – Systematic.....	85
4.2.1.4	PPS – Systematic/Stratified .....	86
4.2.1.5	PPS – Sequential random sampling.....	86
4.2.1.6	SPS (Sequential Poisson Sampling) .....	87
4.2.2	Results .....	88
4.3	Some preliminary conclusion about the process of generation .....	93
5	From profit margin ratios to price indexes .....	95
5.1	Micro elementary indexes .....	95
5.2	Aggregate index.....	98
5.2.1	Economic weights .....	98

5.2.2	Sampling weights .....	99
5.2.3	Computation of aggregate indexes .....	100
5.2.3.1	More aggregate indexes.....	102
6	Further researches: data issues and imputation methods application.....	105
6.1	Data quality issues .....	105
6.2	The imputation methods .....	107
6.2.1	Missing data.....	108
6.2.2	Principles of imputation methods .....	112
6.2.3	Classification of imputation methods .....	115
6.2.4	The different kind of imputation methods.....	118
6.2.4.1	Deductive imputation .....	118
6.2.4.2	Mean imputation overall (MO) .....	118
6.2.4.3	Running Mean (RM) .....	119
6.2.4.4	Mean imputation within Classes (MC).....	119
6.2.4.5	Mean of Log variable (MNL) .....	121
6.2.4.6	Random imputation Overall (RO) .....	121
6.2.4.7	Random imputation within Classes (RC) .....	121
6.2.4.8	Hot deck - random selection.....	123
6.2.4.9	Distance function matching – Hot deck-Nearest neighbour.....	124
6.2.4.10	Sequential hot deck.....	127
6.2.4.11	Random hot deck within classes.....	127
6.2.4.12	Flexible Matching Imputation .....	129
6.2.4.13	Cold deck.....	129
6.2.4.14	Establishment trend times the last observed value (UILT) .....	129
6.2.4.15	Sample trend times the last observed value (UIST) .....	130
6.2.4.16	Last observed value for the establishment.....	131
6.2.4.17	Mean and median Ratio variation (MeanR and MedR) .....	131
6.2.4.18	Regression models.....	133
6.2.4.19	Predicted Regression imputation (PR) .....	134
6.2.4.20	Random Regression imputation (RR) .....	135
6.3	Evaluation criteria.....	135
6.3.1	Kalton and Kasprzyk methods.....	136
6.3.1.1	Sample mean .....	136
6.3.1.2	Distribution and variance.....	138
6.3.1.3	Covariance .....	139
6.3.1.4	Standard error estimation.....	140
6.3.2	Other criteria of evaluation.....	140
6.3.3	The West, Butani, Witt and Adkins evaluation criteria.....	141
6.3.3.1	Mean Unit Error .....	141
6.3.3.2	Mean Unit Absolute Error .....	142
6.3.3.3	Relative Error .....	143
6.3.3.4	Relative Absolute Error .....	143
7	Conclusions .....	145

APPENDICES .....	149
APPENDIX # 2.1 - Wholesale price survey questionnaire .....	150
APPENDIX # 3.1 - NAICS ( <i>North American Industry Classification Methods</i> ) .....	156
APPENDIX # 3.2 - Distribution of the sample by cells .....	162
REFERENCES .....	165

## Index of figures

Figure 1.1 – Importance of price policies in the administration of the firm.....	7
Figure 1.2 – Interrelations between the three dimensions of price's analysis.....	12
Figure 1.3 – Profit margin: from purchase to selling price. ....	13
Figure 1.4 – The main object of this research. ....	15
Figure 3.1 – Generation of a random number with a uniform distribution. ....	62
Figure 7.1 – Scheme of the project.....	148

## Index of Tables

Table 3.1 – NAICS (North American Industry Classification Method): 3 and 4 digits Wholesale Trade's classification (2007 version).....	44
Table 3.2 – Distribution of the duplicated population and of the survey sample by NAICS. .....	45
Table 3.3 – Definition of the revenue classes (by deciles).....	46
Table 3.4 – Distribution of duplicated population and sample by revenue classes (deciles). .....	47
Table 3.5 – Distribution tests (all the cells – sample data): Anderson-Darling vs Cramer- von Mises.....	57
Table 3.6 – Distribution tests (all the cells – sample data): Kolmogorov-Smirnov. ....	58
Table 3.7 – Incoherencies of classification defined with Kolmogorov-Smirnov statistics and graphs. ....	58
Table 3.8 – Final classification of the cells. ....	59
Table 4.1 – Bias of the sampling averages from the reference parameters (frame and sample reference parameter). ....	74
Table 4.2 – Percentage bias from the frame reference parameter of the simple random sampling and of the systematic methods (not-stratified and stratified version). ....	75
Table 4.3 – Average bias of the samples' averages from the reference parameters.....	76
Table 4.4 – Bias from the bias of the “mixed” parameters (computed on samples selected from the “mixed” population). ....	77
Table 4.5 – Percentage differences of the sampling averages from the sample reference parameter (wholesale selected sample). ....	78

Table 4.6 – Percentage differences of the sampling averages from the sample and frame reference parameters.....	80
Table 4.7 – Improvements (in percentage points) of the precision of the parameter obtained selecting 1,000 rather than 1 samples to compute the samples’ mean. ....	81
Table 4.8 – Improvements (in percentage points) of the precision of the average obtained using 1,000 rather than 1 selected samples by simulated population.....	82
Table 4.9 – Average percentage differences from the frame and the sample reference parameters by sampling methods and by used population. ....	89
Table 4.10 – Average percentage differences from the frame and the sample reference parameters by used population (all probability proportional to size methods). ....	89
Table 4.11 – Average percentage differences from the reference parameters by population (all probability proportional to size methods, with and without stratification, compared to both the reference parameters). ....	93

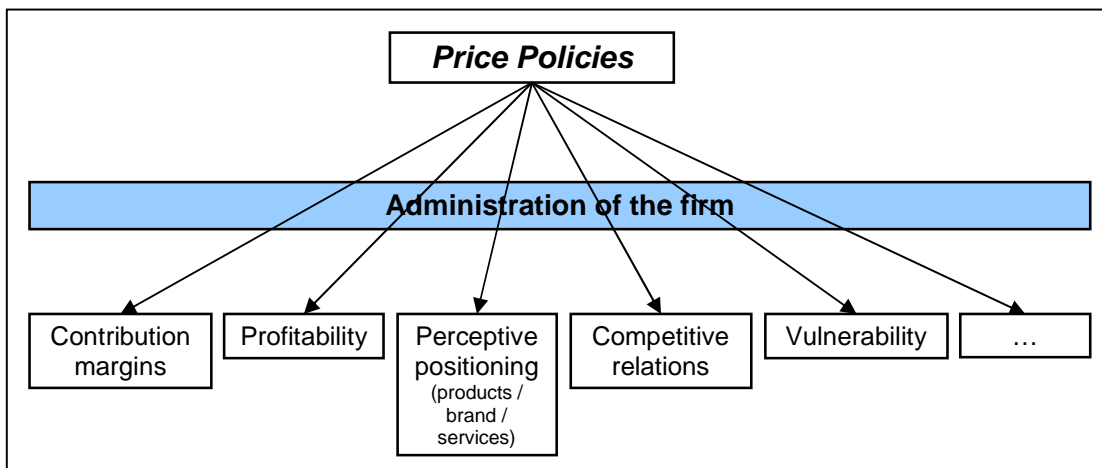
## Index of Graphs

Graph 3.1 – Profit margin ratios distribution (all the sample).....	49
Graph 3.2 – Profit margin ratios distribution (NAICS: 4121).....	50
Graph 3.3 – Profit margin ratios distribution (NAICS: 4162).....	50
Graph 3.4 – <i>Distribution tests (NAICS 4181, Revenue class: 4): graphs and estimated parameters.</i> .....	56
Graph 3.5 – <i>Output tests (NAICS 4181, Revenue class: 4): cumulated distribution and Kolmogorov test.</i> .....	56
Graph 3.6 – Final classification of the cells. ....	59
Graph 3.7 – Distribution of random number $rn_{l,ij}$ generated with standardized normal distribution.....	64
Graph 3.8 – Distribution of profit margin ratios generated with normal distribution. ....	64
Graph 3.9 – Distribution of profit margin ratios generated with lognormal distribution. ....	67
Graph 3.10 – Distribution of profit margin ratios generated with Weibull distribution. ....	67
Graph 3.11 – Distribution of profit margin ratios – “mixed” distribution. ....	69



## 1 Introduction: the rules of price in the marketing

Price is, together with product, placement and promotion, one of the four basic “P” of the marketing. The Policies of price are one of the main issues that the management has to handle with in the administration of the enterprise and it is connected to almost all the main administration phases of the firm. In fact, prices of the products or of the services provided from the firm are connected with the main factors and characteristics that determine the success of the firm itself (Figure 1.1).



**Figure 1.1** – Importance of price policies in the administration of the firm.

First of all, the contribution margins are directly correlated with the prices, but also the profitability depends from the prices we give to our products/services. The price is also important to define the perceptive positioning of the firm, in terms of products’ images, brand strength, ratio cost/quality of provided services, and so on. This means that the price is fundamental to define not only the direct relation with the consumer, but also the

competitive relations with the competitors in the reference market: changing the price we can change the balances of a specific market, giving more strength to our products and/or making them more desirable for our customers. We can improve the image of the product/service or, more in general, of the firm gaining a competitive advantage on the competitors and a better evaluation by the customers. We can give a certain aspect to the image of the products, if we are interested to enter or to conquer a specific sector of the market; moreover, managing the price we can indirectly suggest to the customer an idea of the product's or service's quality level; and so on.

But the price is also a delicate instrument of the management, seen that a wrong action on it could improve the vulnerability of the firm.

## **1.1 Dimensions of analysis**

Costabile (2003) analyzed different models<sup>1</sup> about price and identified three main dimensions of analysis: economic-organizational dimension, competitive dimension, relation with the customer (demand side) dimension. Each of these dimensions will be analyzed in the following paragraphs (respectively, 1.1.1, 1.1.2, 1.1.3).

### **1.1.1 Economic-organizational dimension**

Carù and Cugini (2000) in particular identified the first aspect of the price: the price is the main instrument that management can use to determine the cost and the value of a product, of a service or of a brand for the final consumer. From this point of view it is necessary to consider two main aspects:

- The cost's variables: they are variables indicating the different costs and investments that the firm has to sustain to produce the product or to provide the service to an intermediate or to the final customers.

---

<sup>1</sup> These models were studied by Monroe (2002), Simon (1989), Valdani (1989), Busacca, Costabile and Pasini (1993), Nagle and Holden (1995), Dolan and Simon (1996), Krishnamurthi (2001).

- The differential of cost: it can determine the firm's (or the product's) competitive position in relation to the competitors whose are in the same market.

Thanks to these two kinds of variables, the management can identify a price's lower limit: under this limit there is destruction of the firm's value, and the firm is not able to sustain its production activity, because it's not remunerative.

From an economic and a programmatic/organizational point of view, the price is strategically important, because it's the main variable to define the present and future investments on a product, service or brand. If the management wants to make more investments on a specific product, it's possible to increase the price of the same product, or, in other cases, to make the price of another less strategic product higher (for example when we don't want to influence the sales of the product we want to invest on). So, from the organizational point of view, the price is relevant not only to cover the production costs, but also to plan the sources of the resources that the firm wants to invest in the future. But sometimes the price could be turned under the costs' level for a certain time; this can happen, for example, if the firm wants to fight the entrance of a competitor in the same market.

The result of the management's action on the price from the economic-organizational point of view is the definition of the selling price; this price could be seen directly related to the economical results of the firm's administration. In fact, the final global results of these actions on the prices of the products and services are the generation of the firm's outcomes and the creation of value (not only in terms of firm's revenue, but also, indirectly, in terms of brand image).

### **1.1.2 Competitive dimension**

Valdani (1989, 1995) underlined the competitive dimension of the price. The price's definition, from this point of view, is related to the comparative evaluation of the competitors' price policies.

Two different aspects of this dimension can be underlined:

- *Static approach*: if the price is seen from the cost's structure point of view (as seen in par. 1.1.1), then it is very important, firstly, to define the incidence of the different kinds of cost (production costs, indirect costs of production, promotion's costs, ...) on the total cost of the product/service provided by the firm. But we have also to take into consideration the behavior of our competitors about the prices. So the selling price cannot be fixed without considering the competitors' prices. From this point of view the profit margins on the products/services provided by the firm assume a relevant role, seen that they are the main operate margins for the management.<sup>2</sup>
- *Dynamic approach*: the prices and the profit margins have always to be evaluated and planned having into our mind the competitive dynamics of the market and the always evolving comparison with the competitors. This means that prices have to be decided taking into consideration not only the actual situation of the market and of the prices' system, but also trying to forecast the possible future evolutions. Ad hoc promotions, offensive and defensive manoeuvres on the selling price and other kinds of operation on the price's lever in a long run perspective can be fundamental to determine the success or the failure of the firm. More in general, from the dynamic point of view, is also useful to observe and study the historical price's movements of a specific product, of a more general category, or of a specific market. This study can have the objective, for example, to forecast hypothetical future evolutions of the products' appeal on the customers.

If, as a result of the study of the economic-organizational dimension (par. 1.1.1), we can consider the definition of the price's lower limit, the results of the competitive point of view is the definition of an upper limit. Over this limit the price causes losses of firm's competitiveness, because the price is bigger than the economic value perceived from the customer. For this reason the competitors would be favourite and/or the customers would probably abdicate to buy our product/service.

---

<sup>2</sup> We are underlying the importance of the variable *profit margin* because it will be the main variable that will be studied in the following of this project.

The result of this price's strategies, from a competitive point of view, is, again, the definition of the selling price; nevertheless this time the selling price is not seen as the creation of the business value (like in par. 1.1.1), but it's seen as the value offered to the customer.

### **1.1.3 Relation with the demand**

The third dimension of the price's study is related to the relation with the demand. From this point of view the rule of the customer is the main aspect to take into consideration: in fact the customer, with its evaluation of the characteristics of the product (price included) and with its behavior, gives the final significance and value to the product/service produced or provided by the firm. This evaluation can be expressed in different ways: with the decision of buying the good<sup>3</sup>, with the post-purchase evaluation<sup>4</sup>, and with the dynamic relation between the different operators whose are in the market<sup>5</sup> (Costabile, 2003).

The result of the study of this last aspect is the definition of a selling price comprised between the two limits already defined in par. 1.1.1 and 1.1.2. Into this space of manoeuvre (and using sustainable differentials of prices), the management can handle to plan competitive strategies and to understand which is the better way to fight with the competitors.

The final results of these actions on the price are the firm's share of market, the choice of the price, the results of administration and the offensive or defensive behavior of the firm.

Three last considerations can be done, starting from the conclusive remarks of Costabile (2003):

- From the interaction of the economic-organizational (par. 1.1.1) and competitive (1.1.2) points of view, the competitive manoeuvres based on the price's definition are defined. The results are the more or less highlighted vulnerability of the firm

---

<sup>3</sup> The purchase decision is studied, in particular, in Costabile (1992) and Romani (2000).

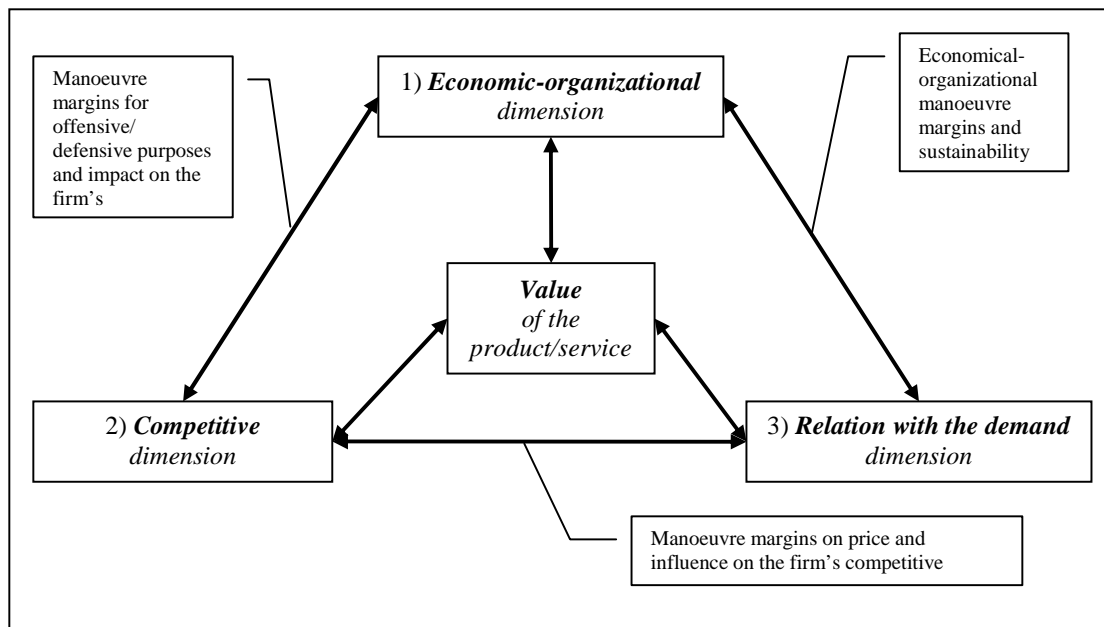
<sup>4</sup> This aspect is discussed in Busacca (1994) and Costabile (1996).

<sup>5</sup> For more details, see Costabile (2001).

and the relative weakness or strength of its products/services. These two aspects are also useful to plan the future defensive rather than offensive plans of the firm.

- From the interaction between the economic-organizational (par. 1.1.1) and relation with the customers (1.1.3) points of view, the management manoeuvre's margins are underlined together with the value's attribution to the product/service by the customer<sup>6</sup> and the following behavior of the customer itself (decision of purchase, customer satisfaction, customer loyalty, and so on).
- The relations with the competitors (1.1.2) and with the demand (1.1.3) define the manoeuvre's margins of the management in the final definition of the price: the price has not to be too high to loose in terms of competitiveness with the competitor's one, nor has to overtake the value attributed by the consumer. From this point of view two fundamental variables are the elasticity of the demand to the price and the degree of differentiation of the product/service.

The three relations are underlined in Figure 1.2.



**Figure 1.2** – Interrelations between the three dimensions of price's analysis (source: Costabile, 2003).

<sup>6</sup> This evaluation is basically based on the ratio: price/attributed value; Costabile (2003).

#### 1.1.4 The fourth dimension and the profit margin ratios

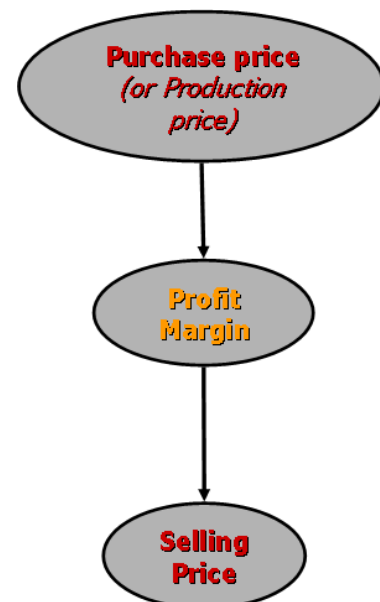
To the previous dimensions of analysis, we can add a different one, that considers the price from a different point of view. We can call this point of view a “decompositive approach”. In fact, if the main objective of the dimensions introduced in the previous part of this chapter was the definition of the price that a product/service will have in a specific market, this fourth approach is based on the decomposition of the selling price in its components.

Each of these components are related to a specific part of the process that brings the product from the earliest stage of production to the final consumer; these costs include material’s costs, workmanship, transformation and other production costs, commercialization’s costs, costs of the personnel, and so on.

But in this research we are interested, particularly, on the profit margin, that is computed as the difference between *selling price* and *purchase price* (or *production price*)<sup>7</sup>.

We already told that this component is very important to study, interpret, understand and try to modify the dynamics of the market, the market’s shares, the success of a product or of an enterprise. Moreover, from what was told in the previous paragraphs (1.1 and 1.1.2 in particular), it is clear that the profit margin is the main lever for the marketing strategies. In fact, acting on the profit margins of one or more product of the firm, the management can:

- Define the actual competitive strategies (defensive or aggressive);
- Effect promotions or discounts on the prices;
- Forecast the future evolution of the prices/profit margins in a specific market or sector;



**Figure 1.3** – Profit margin: from purchase to selling price.

---

<sup>7</sup> With *production price* we mean the total costs that a firm has to sustain during the production phase of a good or of a service, before this last would be put into a market. We usually speak of *purchase price* if a firm purchases and resells a product or a service to another subject (as happens for wholesalers, for example). The *selling price* is the price at which the product or the service is sold to the final consumer or to another intermediary.

- Plan the future attitude of the firm versus the competitor, and the probable reaction to their actions;
- Plan the future investments of the firm and the business politics.

From all this is understandable how is important to know the actual level and the probable evolution of the profit margins along the time, with reference to a specific product, rather than to a specific market or to a more general sector. This study, in fact, is not only useful for the management to fix the final price of a product/service but is also useful for States and public institutions to eventually plan helps for the enterprises or their fiscal and economical politics.

The importance of profit margin is particularly relevant if we refer to the services' sector and, in particular, to the wholesale sector, that will be the main object of our analysis.

## **1.2 The management's cognitive requirements**

All the dimensions introduced in the previous paragraphs<sup>8</sup> are of fundamental importance for the management of a firm to scheme and effect strategic plans.

If we think, in particular, to the prices' definition and control, the cognitive requirements of the management already underlined in the previous part of this chapter are two: the knowledge of the competitive dynamics of the market and the freedom degree of action.

Regarding the first point (the knowledge of the competitive dynamics of a specific market), it involves, for the management, the study and the comprehension of the cognitive and behavioural dynamics of the market of reference, with reference to the customers, to the competitors and to the suppliers, and the dynamics observed between these same three subjects. The knowledge of the competitors, in particular, assumes a big relevance. It involves the knowledge of the direct competitors and of their strategies on one hand, and the knowledge of the specific market (or of a more general level) on the other hand. This

---

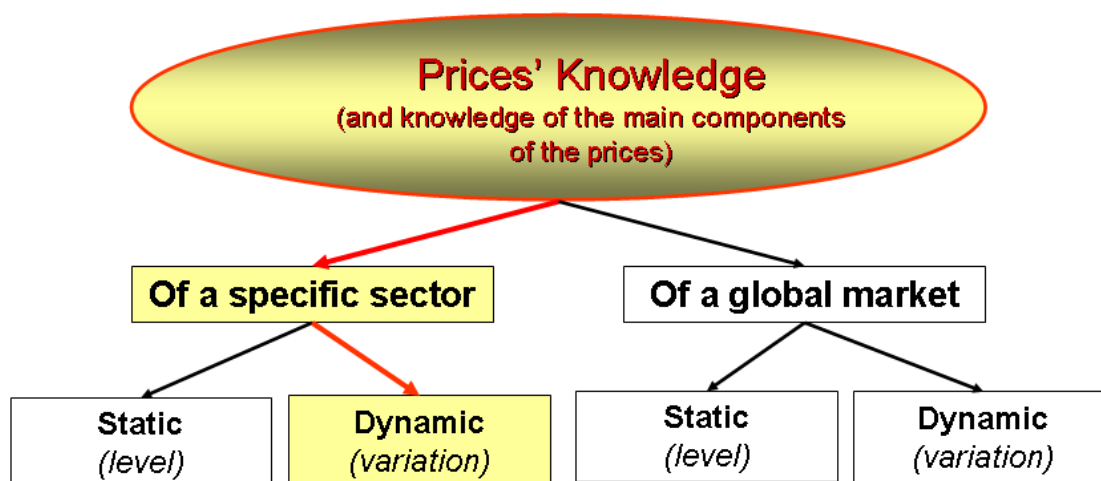
<sup>8</sup> The economic-organizational (paragraph 1.1.1), the competitive (1.1.2), the relation with the demand (1.1.3) and the "decompositive" vision (1.1.4).



knowledge is made through the study of the value of the products (as they are actually perceived by the consumers), of the factors whose determine the perceptions, of the interactions and relations between price's policies on one hand and product's policies on the other one, of the price's policies in comparison to the behavioural dynamics of the competitors (in terms of actions and also in terms of reactions), and so on. The direct consequence of this deepen studies are firm's strategic plans and actions for the future, focalized on facing the competitors with operating strategies, keeping always in mind the customer satisfaction.

This first aspect (the knowledge of the competitive dynamics of the market) is strictly related to the second one, the freedom degree of action of the management in the prices' definition. From this point of view it's necessary to define the level of selling price, it is useful to control its variation across the time, and so on. For all these purposes it's evident the importance of the study of the profit margins' relevance and evolutions. And that's why we need to apply an analysis of the different components of the price and, in particular, of the profit margins.

The Figure 1.4 shows, in the field we introduced before, our specific interest's area.



**Figure 1.4** – The main object of this research (underlined in grey).

From the study of the prices and of their components, we can get a better knowledge of our market/sector of activity. This can be done taking into consideration, above all, the

dynamics evolution of both prices and their components across time. In our research we are interested in the study of prices in the services' sector (and, in particular, in the wholesalers' field). We are interested, in particular, in the study of a variation; this means that we use a dynamic approach: the change is evaluable using a ratio between the data at two times. The variable object of study is the most important and influenceable (by the management) component of the prices: the profit margin.

### **1.3 Methodological issues**

The management's cognitive requirements introduced in the previous paragraph (1.2) require some solutions to approach to two main methodological and implementation's challenges: the methodology that is used to measure the prices' variation across time and some solutions to improve the quality level of the data.

The two topics are introduced, respectively, in paragraphs 1.3.1 and 1.3.2.

#### **1.3.1 First statistical solution: the price indexes**

When we are studying the behaviour of prices across time we face with two different kinds of problems: the first one is how to measure the prices' variation along the time, and the second one is how to synthesize the collected data about prices.

To solve the first problem there is a specific statistical solution, that is the computation of price indexes (they will be shortly introduced in par. 1.3.1.1).

As methodological aspect we have to face the problem on how to synthesize the product/service prices and weight them. A widely used solution to measure the variation of a group of prices across the time is synthesizing them using the Laspeyres index (it will be introduced in par. 1.3.1.2).

### 1.3.1.1 The price indexes

A price index<sup>9</sup> computes the relative variation between two different times or two different spatial contexts, considering economical phenomena like: prices, quantities, volumes.

If we want to measure the variation between two times (say  $t$  and  $0$ ) of the price  $x$  of a specific product, we can use the simple index formula:

$${}_0I_t = \frac{x_t}{x_0} \cdot 100$$

where:

- $x_t$  is the price of the product at time  $t$  (the time  $t$  is also called reference or actual time;  $t = 1, 2, \dots$ );
- $x_0$  is the price of the product at time  $0$  (the time  $0$  is also called base time),
- ${}_0I_t$  is the price index that expresses the percentage variation of the price observed between the base time and the reference/actual time.

We have to underline that this kind of index is a fixed-base index, because the base time doesn't change across time.

Nevertheless, there is another version of the simple index called *moving base*: in this version the base time varies constantly, from time to time. The most common version of the moving base index is the following:

$${}_0I_t = \frac{x_t}{x_{t-1}} \cdot 100.$$

As shown in this last version of the formula, the base time is the time directly antecedent the reference/actual time. If the prices are monthly data, for example, this index measures the percentage change of the price of the considered product month by month.

If we are working with monthly data, the following version is particularly useful if we want to exclude the seasonal effect in the variation of the price:

$${}_0I_t = \frac{x_t}{x_{t-12}} \cdot 100.$$

---

<sup>9</sup> For more details about price indexes, see Predetti (2006).

This index compares the price of the actual time  $t$  with the price observed 12 months before.

The price indexes have some common properties that make them very useful and easy to interpret.

- they measure a relative variation;
- they are always positive;
- they are pure (without a unit of measurement and independent from that).

A price index can be easily interpreted, using the following rules:

- If  ${}_0I_t > 100$ , in the time  $t$ , in comparison with the time 0, there is a positive variation in the price of the considered product equal to:  $({}_0I_t - 100)\%$ .
- If  ${}_0I_t = 100$ , in the time  $t$ , compared with the time 0, there isn't any variation in the price of the considered product.
- If  ${}_0I_t < 100$ , going from time 0 to time  $t$ , a negative variation in the price of the considered product is observed; this variation is, in percentage terms, equal to:  $({}_0I_t - 100)\%$ .

For example, if we have an average price of a certain product  $i$  equal to 25.5 \$ in 2004 and an actual average price (referred to 2009, same product  $i$ ) of 14.5 \$, the relative variation between the two years, measured with a simple price index with fixed base, is:

$$\begin{aligned}
 {}_0I_t &= \frac{x_t}{x_0} \cdot 100 = \\
 &= {}_{2004}I_{2009} = \frac{x_{2009}}{x_{2004}} \cdot 100 = \frac{14.5\$}{25.5\$} \cdot 100 = 0.5686 \cdot 100 = 56.86.
 \end{aligned}$$

This means that in 5 years there was a negative percentage variation of the price of - 43.14%, in fact:

$$(56.86 - 100)\% = -43.14\%.$$

### 1.3.1.2 The Laspeyres index

If the simple price index is useful to measure the variation of the price of a single item; to measure the variation of the prices of a group of goods or services<sup>10</sup> other kind of complex indexes are suggested by the literature.

For the computation of these indexes is very important to taking into consideration the economic weights of the different products/services whose are in the considered group. One way to do this is using, as a weighting system, the values of the considered products in the two compared times.

On this last principle is based the construction of the most common complex indexes, used to measure a synthesis of the change observed on more than one product between two times. They are the Laspeyres index, the Paasche index and the Fisher index.

The Paasche index needs the computation of the values of all the products in the actual time and in the base time (it's a kind of moving base index). This could be not easily done, because often we have data about prices of all products, but not about the exchanged quantities.

The Fisher index is a synthesis of the Paasche's and Laspeyres' index, so it requests the previous computation of both the other indexes.

The most widely used index is the Laspeyres index<sup>11</sup>: it is a fixed-base index, so it needs only the knowledge of the values (quantities and prices) of the observed group of products in the base time 0. This means that is less expansive to compute and that data are not so difficult to retrieve.

The Laspeyres index of prices between time 0 and  $t$  ( ${}_0^p I_t^L$ ) uses, as weights, the values of the considered products at time 0:

$${}_0^p I_t^L = \frac{\sum_{h=1}^n \frac{p_{th}}{p_{0h}} p_{0h} q_0}{\sum_{h=1}^n p_{0h} q_0} = \frac{\sum_{h=1}^n p_{th} q_0}{\sum_{h=1}^n p_{0h} q_0}$$

where:

---

<sup>10</sup> These goods or services can also be of a different kind.

<sup>11</sup> See also Laspeyres (1884).

- ${}_h p_t$  is the price of the  $h^{\text{th}}$  product ( $h = 1, 2, \dots, n$ ) at the actual time  $t$ ;
- ${}_h p_0$  is the price of the  $h^{\text{th}}$  product ( $h = 1, 2, \dots, n$ ) at the base time  $0$ ;
- ${}_h q_0$  is the quantity of the  $h^{\text{th}}$  product ( $h = 1, 2, \dots, n$ ) at the base time  $0$ ;
- ${}_h p_0 {}_h q_0$  is the value of the  $h^{\text{th}}$  product ( $h = 1, 2, \dots, n$ ) at the base time  $0$ ;
- $n$  is the number of considered products.

The Laspeyres index can be seen (first part of the formula) as an arithmetic weighted average of simple price indexes of  $n$  products (or services). The weights of the average are the values of the  $n$  products/services observed at base time  $({}_h p_0 {}_h q_0)^{12}$ .

The Laspeyres index can also be seen (last term of the formula) as the average of the prices of the two times weighted with the quantities of the base time.

From the way the index is computed, we can notice the first two characteristics of the index: it has a fixed base and the weighting system is also fixed in time  $0$  too.

The index measures the relative variation across time of the cost of the group of  $n$  goods/services, fixed in base time. It measures the variation of prices, while the quantities are considered constant.

Nevertheless it's possible to give a second interpretation, because the index represents the ratio of the virtual aggregate at time  $t$  (once the quantities are fixed: they are the quantities of time  $0$ ) on the real aggregate value observed at time  $0$ .

The Laspeyres index, despite is widely used, is not a perfect index. The main defect is that from time to time the index tends to loose its representativeness, because the exchanged quantities of the different goods/services vary from time to time (we can talk about "attrition" of the index). A solution to this problem could be the use of Laspeyres index with variable weights.

The Laspeyres index used in this project to measure the changes across time of the profit margin ratios is a particular version of the Laspeyres index that will be introduced in chapter 5 (par. 5.2.3).

---

<sup>12</sup>  $h = 1, 2, \dots, n$ .

### 1.3.2 Second statistical solutions: the data quality improvement

In the beginning of paragraph 1.3 a second issue was introduced, that is the quality improvement of collected data. This is a common challenge we have to handle with when we are working on surveys of any kind.

The main aspects we had to face along the first part of this project were three: the generation of simulated data, the testing of the most efficient sampling method and the test of strategy to handle with the general problems related to the presence of missing data and of outliers value. These three aspects will be shortly introduced in the following.

1. We had to find an efficient system of data generation: this necessity was due to the data quality improvement process. In fact the data needed to be deeply studied to solve issues related to the presence of outliers and missing values. Moreover the number of collected data was considered not enough to bring ahead the first testing phase of the methodology (that should have a more general impact). We had also to consider the confidentiality issues about not already released data. For all this reason it was not possible to work with original collected data; but it was extremely important also to go on with the simulation analysis that was done while it was waiting for the final version of the dataset. It is, then, understandable how was important to generate a simulated population that would be as close as possible to the observed data, to go on with the project as soon as possible. In this way the results would be ready (and useful) not only once the real data would be available and released, but also for other waves of the same survey or for other projects. Therefore the fact of working with simulated data while we were waiting for the final version of the dataset didn't affect the right sequence of the operations, but was finalized to make the testing process in a faster way and to get more generally valid results.
2. We intended to test the more efficient sampling method: using simulated data we can test the relative efficiency of used sampling method in comparison to other available methods to understand if it's possible to improve the quality of collected data optimizing the data collection methodology.

3. The third aspect regards specifically the solution of some common problems we face when we are working with survey's data. The two main problems are the presence of outlier values and of missing data. To face these challenges an efficient and reliable system of outlier detection on one hand and the application of imputation methodology on the other hand are extremely useful. Our research is particularly focused on this last aspect: we want to find an efficient way to impute the missing data. Moreover the test of imputation methods would be also useful to understand the impact of the biggest units on the survey's estimates. The imputation methods will be introduced in this first part of the work (par. 6.2, 6.2.3 and 6.2.4), and the deepened study of the imputation strategy on the collected data will be developed in further researches (see also chapters 6 and 7).

## **1.4 Conclusions about the aim of this research**

If we want to study the prices' data and, in particular, how prices of a specific sector change along time, we have to face with some methodological issues and we have to answer to some interesting questions. We need to study, in particular, the way to improve the data quality and how to measure the change of the prices along the time.

The questions that this work originated brought to study some other methodological issues related firstly to find an efficient data generation and selection method and, then, to test the application of imputation methods.

In this thesis the first part of the research will be explained and the second part will be introduced.

In particular, in chapter 3 the simulated data generation process is presented; the different kinds of simulated distributions are tested, together with some sample selection method (and PPS method in particular) in chapter 4. The methodology for the computation of the index is presented in chapter 5. The imputation methodology and the main imputation methods are introduced in chapter 6.



This research project took place during a work term at *Statistics Canada*. The project's objective was to address various methodological challenges associated with price indices. Among these are: sampling methods determination and imputation.

For that purpose, simulated data were derived from the preliminary wholesale price index survey data. This survey has been developed by *Prices Division* of Statistics Canada in 2006. The structure, the main characteristics and the aim of the wholesale price index survey will be introduced in the following chapter 2.

I was also involved in a study of the preliminary collected data, after the first wave of this survey (started in 2006 with the first data collection) from January the 30<sup>th</sup>, 2008 to March the 7<sup>th</sup>, 2008 and from June the 2<sup>nd</sup>, 2008 to September the 26<sup>th</sup>, 2008.

The project took place in the Statistics Canada's offices situated in the R.H. Coat building, 100 Tunney's Pasture Driveway, Ottawa, Ontario (Canada). I worked within the *BSMD* (*Business Survey Methods Division*) team directed by Sylvie Gauthier, under the supervision of Zdenek Patak.

The views expressed in this thesis are those of the author and do not necessarily reflect the official position of Statistics Canada. All the potential and hypothetical remaining errors and oversights are those of the author.



## 2 Data collection: the Statistics Canada's Wholesale survey

### 2.1 Price indexes in Canada

The first experiences in producing price indexes for the Canadian economy were in the second part of the 17<sup>th</sup> century with the *Wholesale Price Index* (WPI, 1967) and the *Industrial Producer Price Index* (IPPI)<sup>13</sup>. Only in 1913 was computed the first *Consumer Price Index* (CPI)<sup>14</sup>.

The services' sector is a quite new field of study, but its importance grew very fast in the last years. This is clear if we think that in 1961 the employees of services in Canada was 55% of total employees (and represented the 57% of *GDP*<sup>15</sup>), while in 2003 the employees in the services in Canada covered the 75% of the working force and the 68% of the *GPD*.<sup>16</sup>

This fast increasing importance of the services suggested to measure the change of prices in the services' producer area using the price indexes' methodology. For this reason Statistics Canada proposed, in 2004, to measure the changes of prices through the computation of price indexes of 83 business service commodities regarding the Service Producer's area.

The final goal is the computation, using the different indexes, of an overall Service Producer Price Index (SPPI), that would be a synthetic measure of the change of the prices of the producers in the services' sector. This index will be presented in the following par.

2.2.

---

<sup>13</sup> Patak and Rais (2005).

<sup>14</sup> For more details, see "*The Consumer Price Index reference paper*" (1995).

<sup>15</sup> The *Gross Domestic Product*.

<sup>16</sup> Source: Patak and Rais (2005).

## 2.2 The SPPI (Service Producer Price Index)

The SPPI was developed by Statistics Canada starting from 2004. The index belongs to the PPI (Producer Price Indexes) family<sup>17</sup>.

The target population of the SPPI project is made by all the business industries operating in Canada in the sector of services.

The aim of the SPPI is to measure the change, in terms of added value, of the services across time. *“Value added can be defined as the value of outputs less the value of intermediate inputs used to produce those outputs”* (Patak and Rais, 2005).

This change is measured in terms of ratio: the comparison between an actual time and the base time is done through a ratio (this happens, more in general, for all the price indexes; see chapter 1).

The final index has the function not only to *“deflate nominal measures of goods or services to obtain measures of quantity in the services’ sector, but is also useful for government’s departments and private companies [that] use price indexes for policy and economic decision-making”* (Patak and Rais, 2005).

The SPPI index is furthermore an indispensable instrument to measure inflation and its evolution, a deflator useful to convert the National Account and the added value of each industry (splitting the price and volume components) and to index-link the legal contracts. This brings also to get better indicator of productivity and of the growth of the economy, and it’s useful to underline past and present economic trends.

The index is computed every year. The detail’s level of the monthly index is the 5 NAICS (North American Industry Classification Methods) digits<sup>18</sup>.

The data collected are also useful to estimate a national index for the whole considered industry; this last index could be useful to measure prices movements in the Canadian economy, especially about important industry and commodity groups.

The results can be used to develop estimates of outputs and valuation of imports.

---

<sup>17</sup> For more details about the PPI methodology, that is the foundation of the SPPI methodology, see the *“Producer Price Index Manual”* (International Monetary Fund, 2004).

<sup>18</sup> For more detailed information about the NAICS code, see par 3.1.1.

For the computation of the SPPI some of the most representative industries within the 83 classified by the National Account are selected. For this selection, parameters like the following are considered: complexity, relevance, contribution to the services' industry, coverage of the subsidiary services.

The SPPI project pertains to nine specific areas, and the most recent group of industries for which an index was planned to be produced are<sup>19</sup>:

- The wholesale services,
- The retail services,
- The for-hire trucking,
- The property and casualty (P&C) insurance.

These services are all part of the PPI, so they have many elements in common (from methodology to data source, from users to distribution of the outputs, and so on...).

This research project was based on simulated data derived from the wholesale price index survey. *“This sector comprises establishments primarily engaged in wholesaling merchandise and providing related logistics, marketing and support services. The wholesaling process is generally an intermediate step in the distribution of merchandise; many wholesalers are therefore organized to sell merchandise in large quantities to retailers, and business and institutional clients. However, some wholesalers, in particular those that supply non-consumer capital goods, sell merchandise in single units to final users”* (source: <http://stds.statcan.gc.ca/naics-scian/2007/cs-rc-eng.asp?criteria=41>).<sup>20</sup>

---

<sup>19</sup> Source: Patak and Rais (2005).

<sup>20</sup> *“This sector recognizes two main types of wholesalers, that are wholesale merchants and wholesale agents and brokers.*

1. **Wholesale Merchants.** Wholesale merchants buy and sell merchandise on their own account, that is, they take title to the goods they sell. They generally operate from warehouse or office locations and they may ship from their own inventory or arrange for the shipment of goods directly from the supplier to the client. In addition to the sale of goods, they may provide, or arrange for the provision of logistics, marketing and support services, such as packaging and labelling, inventory management, shipping, handling of warranty claims, in-store or co-op promotions, and product training. Dealers of machinery and equipment, such as dealers of farm machinery and heavy-duty trucks, also fall within this category. Wholesale merchants are known by a variety of trade

So the wholesaling activity consists in buying and selling of goods on one's own account or in engaging in the buying and/or selling goods owned by others on a commission or fee basis.

The wholesale is an economic area studied through the *wholesale survey*. The main characteristics of this survey will be introduced in the next paragraph (par. 2.3).

### 2.3 The characteristics of the wholesale survey

The wholesale survey was developed by Price Division from Statistics Canada and implemented for the first time in 2006. The aim of the wholesale survey is the computation of the monthly SPPI index for each group of the wholesale services industry<sup>21</sup>.

The wholesale survey, in particular, regards wholesale businesses in the motor vehicle, building supplies and machinery and equipment trade groups. Secondary activities included in wholesale services are: breaking of bulk (the goods are purchased in large lots and resold in smaller quantities), warehousing (i.e. holding on inventory), inventory management (that include actions like receiving, unpacking, checking, sorting, organizing, pricing, storing and tracking materials), shipping, in-store or co-op promotions (that means marketing strategies to motivate and encourage the end consumer trading by marketing via a retailer or to encourage retailers to order products carried by wholesaler), handling of warranty

---

*designations depending on their relationship with suppliers or customers, or the distribution method they employ. Examples include wholesale merchants, wholesale distributors, drop shippers, rack-jobbers, import-export merchants, buying groups, dealer-owned cooperatives and banner wholesalers. The first eight subsectors of wholesale trade comprise wholesale merchants. The grouping of these establishments into industry groups and industries is based on the merchandise line or lines supplied by the wholesaler.*

2. **Wholesale Agents and Brokers.** *Wholesale agents and brokers buy and sell merchandise owned by others on a fee or commission basis. They do not take title to the goods they buy or sell, and they generally operate at or from an office location. Wholesale agents and brokers are known by a variety of trade designations including import-export agents, wholesale commission agents, wholesale brokers, and manufacturer's representatives and agents."*

Source: <http://stds.statcan.gc.ca/naics-scian/2007/cs-rc-eng.asp?criteria=41>. For more information about industries' classification, see also the Statistics Canada's web site: [www.statcan.gc.ca](http://www.statcan.gc.ca).

<sup>21</sup> For further information about the SPPI and the Wholesale Price Survey, see also Barzyk (2008).

claims, product training or training of sales staff, marketing services, repair and maintenance, leasing and renting and other activities.<sup>22</sup>

The target population is composed by most of the activities classified with a NAICS code between 411 and 419 at 5-digit level<sup>23</sup>.

The frame used for the selection of the sample is the *Business Register* of Canada<sup>24</sup>. This is an excellent source because is the most complete list of businesses available. It also contains the main variables useful to describe and classify the frame's units (stratification and size variables). The BR is updated for deaths and births and other classification variables (like: industries, activity or classification, size measure and so on); it is updated at least once a year, but many units are updated more frequently. The BR is maintained and updated by the *BRD (Business Register Division)* of Statistics Canada. To update the information contained in the database also tax records and other administrative sources are often used.

The frame is stratified by industry line using the NAICS code or, sometimes, by province.

The data useful to compute indexes are collected through a survey based on a two-stage sample: the first stage (see par. 2.3.1) is focused on the selection of the business units (also called establishments<sup>25</sup>), the second stage (see par. 2.3.2) is based on the selection of items belonging to each business unit selected in the first stage.

The wholesale is a panel survey: the same group of establishments and the same group of products are followed along the subsequent waves of the survey.

---

<sup>22</sup> Source: wholesale price report questionnaire.

<sup>23</sup> See Appendix 3.1 for the list of targeted NAICS. For further information about the standard classification structure in the Wholesale Price Survey, see also Barzyk (2008).

<sup>24</sup> For more information about the BR, see: Bérard et al. (2005), Castonguay et al. (2000), Colledge (1987), Cuthill (1990), Gagnè (2004).

<sup>25</sup> "Establishment is the level at which the accounting data required to measure production is available (principal inputs, revenues, salaries and wages). The establishment, as a statistical unit, is defined as the most homogeneous unit of production for which the business maintains accounting records from which it is possible to assemble all the data elements required to compile the full structure of the gross value of production (total sales or shipments, and inventories), the cost of materials and services, and labour and capital used in production" (source: <http://www.statcan.gc.ca/concepts/definitions/estab-etabl-eng.htm>).

The survey is conducted every three months through a paper questionnaire (introduced in par. 2.3.3): the collected data are the prices of the three previous months, and the results are organized by quarters (four quarters).

### 2.3.1 First stage (PSU)

In the first stage, the *Primary Survey Units (PSU)* are selected. As PSU, the Business Units are considered, and for wholesale SPPI they are identified with establishments.

The PSU units are selected with a PPS (Probability Proportional to Size) sampling selection method: the units of the frame have a probability of being selected proportional to their size<sup>26</sup>. This means that larger units have a higher probability of being included in the sample; this is done because they are considered to have a bigger influence in determining the prices' movements; these units are usually included for a long time in the sample, once selected; the smaller units, on the other hand, should be re-sampled at a frequency that is usually suggested by industry turnover and by the perceived response burden of the respondent units themselves.

The variable considered to define the size of a unit is the annual revenue observed in the previous year<sup>27</sup>.

If the predefined sampling method is the PPS, nevertheless sometimes cut-off sampling is preferred; this means that there is a selection of the largest units of the target population, whose represent a fixed percentage of some population size variables<sup>28</sup>.

The units of the population, by an iterative process, are classified in three secondary size strata: TN, TA, TS.

1. *TN (Take None)* units: the smallest units in terms of contribution to the revenue of the primary stratus. They are removed from the sample before the computation of the probabilities of inclusion.

---

<sup>26</sup> For more details about the Probability Proportional to Size sample selection methods, see Kish (1965, 1987), and Särndal et al. (1992).

<sup>27</sup> In our case the survey was carried out in 2006, so the revenue data are referred to 2005.

<sup>28</sup> This kind of sampling is preferred when the distribution of a size variable is highly skewed and one wants to obtain a predefined coverage of some size measure of a variable strictly related to the estimate we are interested in.



2. *TS (Take Some)* units: they have a probability of inclusion less than one.
3. *TA (Take All)* units: they have a probability of inclusion equal or greater than 1.

For the selection of the sample, a random number  $\varepsilon_i$  with a uniform distribution between 0 and 1 is generated for each unit. The  $i^{\text{th}}$  unit of the frame population is selected if the probability of inclusion  $\pi_i$  is greater than  $\varepsilon_i$ .

To compute the  $\pi_i$ s, an auxiliary variable  $x$  closely correlated with the variable of interest and known for all the units of the sample is considered. This variable is the size of each unit in terms of revenue; for this reason the units with a larger size have usually a larger probability of inclusion.

For example, if  $n$  is the expected sample size and  $x_i$  is the revenue of the  $i^{\text{th}}$  establishment:

$$\pi_i = n \cdot \frac{x_i}{X},$$

where:

- $X = \sum_{i=1}^N x_i$  and
- $N$  is the total number of units of the frame population.

Considering the kind of selection method, the sample size  $n$  is a random number (and it cannot be planned in advance); this could be considered a problem.

The sample selection scheme was developed by the BSMD division of Statistics Canada<sup>29</sup>. For the selection of the first-stage's units in particular, the BSMD uses the SPS (Sequential Poisson Sampling<sup>30</sup>) method. This sampling selection method belongs to the PPS's family.

---

<sup>29</sup> The methodologists might be involved in various areas of a SPPI survey (source: Statistics Canada's web site, [www.statcan.gc.ca](http://www.statcan.gc.ca)):

- the construction of the frame (identifying the units of interest);
- the choice of the most appropriate sample design (based on the goals of the survey, on the expected precision and timeliness);
- the identification of patterns in the collected data, with the development of cross-sectional and longitudinal edits finalized to reduce the number of outliers and with an analytical module that helps to choose the longitudinal or the cross-sectional approach (the latter could be useful to build

The TA units are easily identified if the Poisson Sampling is used, given that it makes possible to identify units with a probability of selection greater than 1<sup>31</sup>. These last, called natural-take all units, should be inserted in the sample for sure. In the survey about business is often useful to have natural TA, in the sample, because they are usually more representative than the others.

Nevertheless, the final group of TA units is not only made of the natural TA. In fact, the probability of inclusion of the units with  $\pi_i > 1$  after the first selection is set to 1 and, in the second step of selection, the probability of inclusion of the other units ( $\pi_i'$ ) is adjusted to respect the following condition:

$$E(n_s) = n = \sum_i \pi_i'.$$

After the adjustment, we can get more units with a probability inclusion equal or bigger than 1 (that is, we get more TA units): these units are added to the group of natural-take all units. This is an iterative process that continues until no more units with a probability inclusion equal or bigger than 1 are found.<sup>32</sup>

One advantage of the Poisson Sampling is that it is uncomplicated and of easy application and has the implicit definition of the TA units. For this reason the variance could be inflated, even if a fixed sample size is guaranteed.

---

confidence region to monitor and adjust data for the presence of observation falling outside of it and to improve the overall data quality);

- the detection and the manual or automatic correction of further outliers;
- the manual intervention to treat the non-conforming observations that increase the variability of estimates;
- the imputation's studies: an imputation method is choose between several options of imputing schemes (for example cross-sectional or longitudinal imputation);
- the computation of imputation's variance (a relevant quality measure of the different imputation methods);
- the development of an estimation module to compute a specific variance formulation with benchmarking and calibration measures whose are also useful for further research.

<sup>30</sup> For more information about Sequential Poisson Sampling, see: Ohlsson (1990, 1998).

<sup>31</sup> This usually happens when:

$$\pi_i = n \cdot \frac{x_i}{X}.$$

See Sarndal et al. (1992) for details.

<sup>32</sup> There are other kind of TA selection methods; see Hidirolou (1986) and Lavalée and Hidirolou (1988) for details.

Similarly to what we have seen for PPS methods, Sequential Poisson sampling computes the inclusion probability  $\pi_i$  of the  $i^{\text{th}}$  unit as a function of the relative contribution to some auxiliary variable  $x_i$  (usually the annual revenue of the unit). Then a random number ( $\varepsilon_i$ ) is generated for each unit of the frame population.

A function considering both  $\pi_i$  and  $\varepsilon_i$  is used to generate another random number ( $\xi_i$ )<sup>33</sup>. To select the units of the sample the  $n$  units with the smallest  $\xi_i$  are selected.

The absence of an existing index required to use a coefficient of variation to determine the appropriate sample size of units whose had to being selected. Nevertheless, due the 2006's one was the first run of the survey, it was necessary to determine the sample size as a function of availability resources. In the following waves of the survey, the data collected during this first implementation of the survey are useful to determine the sample size.

The sample is also allocated across trade groups using the revenue as measure of the size. The choice to have only one stratification variable was taken to avoid the risk to have (with much more strata) a small number of units for each stratum. In particular it was used an  $x$ -optimal allocation based on the stratum revenue, as an auxiliary variable; this last variable is considered a good indicator of the level of influence that a given stratum has on the national index. The disadvantage of this method is that sometimes  $x$ -optimal allocation allocates more sampling units to the stratum than the number of population units in the same stratum<sup>34</sup>. Based on a minimum sample and on cost's criteria,  $x$ -allocation is computed by minimizing stratum and overall variances. But other two aspects are to be taken into consideration determining the number of units in each stratum: the attrition in response for each stratum and the sufficient representation by the units belonging to a stratum (for this reason sometimes a re-adjustment is necessary).

---

<sup>33</sup>  $\xi_i = \frac{n\varepsilon_i}{\pi_i}$ .

<sup>34</sup> In these cases an adjustment for over-allocation was applied, according to Cochran (1977).

Once selected, the sample coming from the BR is periodically compared, to test its coherence and to address response burden, with the corresponding activity (or revenue) survey conducted by Distributive Trades Division of Statistics Canada.

Actually the targeted sample size is of about 3,900 establishments and there is a minimum of 40 sample units allocated for each NAICS. This might eventually be refined, based on future needs.

The expected response rate of the survey is 80% or higher.

### **2.3.2 Second stage (SSU)**

The second-stage units (*Secondary Survey Units* or *SSU*) are goods/services selected from the ones produced/provided and sold by the business units selected at the first stage. This is also a fixed basket of goods (panel survey).

At this level, a sampling proportional to size selection method would request the knowledge of all the sales of each establishment's product. However, these data are usually unknown and difficult to collect, because their knowledge requires a use of extensive resources and a big amount of time. Moreover a probability selection of the products is almost impossible, seen the unavailability of an exhaustive list of all goods sold by all the selected establishments.

For these reasons, the selection of the SSU is based on a cut-off approach. This means that a judgmental sample is made. In fact, the respondents are asked to report in a questionnaire the prices of three products. In particular, the prices of the three most representative items, in terms of sales, of every business unit are observed (that is the prices of the three best selling products). This is done because the three most sold products are considered representative of the general movements of the prices from a month to the other one, in a specific sector<sup>35</sup>.

There are no proofs that three products only are really representative of all the universe of establishment's products, but this number was chosen also thinking that a small sample size

---

<sup>35</sup> The number of selected items for each unit varies, from a SPPI survey to the other one, considering the response burden and the survey's costs.

would not represent a problem in term of response burden. However, the presence of a judgmental sample means also that we cannot estimate the variance neither the bias of the computed index. Some discussion and studies about the second stage's sample representativeness is actually in progress.

### **2.3.3 The questionnaire**

The questionnaire to collect data was tested for the first times in October and November, 2005. The questionnaire, self filled, is proposed quarterly to the selected PSU sample.

The objective of the survey is to collect data on a monthly base about prices of representative products and services transactions. From this point of view, a price could be intended as both a representation of the completion of a service or a proxy measure for the completed transaction. The respondent involved in the survey has to fill this questionnaire with the prices object of study.<sup>36</sup>

The monthly data are collected on a quarterly basis; this means that three month's data are collected in the same questionnaire.

The observed prices' movements must reflect the average of all transactions happened over the months considered in the questionnaire. For this reason, the respondent is requested to continue reporting, each quarter, about the selected products, and to replace one (or more) of them if it (ore they) become obsolete.

The first part of the questionnaire asks some information about the establishment (legal name, business name, contact, address and so on).

In the second part, some information about the establishment activity (i.e. the description of wholesaling service) are collected; furthermore the respondent have to choose a list of the

---

<sup>36</sup> For more details see the QAF (Quality Assessment Framework) ([www.statcan.gc.ca](http://www.statcan.gc.ca)).

three products that are considered most representative of its business, considering the amount of sales<sup>37</sup> in the past fiscal year.

In the third part of the questionnaire, some information about the prices of the last three months and about each of the three selected products is requested:

- if the product is imported (and at what percentage level),
- the average selling and purchase prices (dollars per unit for each month of the considered quarter)<sup>38</sup>,
- the possible reason for price's change (the respondent can choose between "change in supplier", "change in product", "change in service offered", "inflation" or can specify other reasons of the change).

The main collected data used for the elaboration of the index are purchase and selling prices of each selected service/product (SSU) and the products' characteristics (specifically for those products whose are changed from a month to the following one).

The respondent could be furthermore contacted to follow-up the participation to the survey or to clarify the reason of a change, if needed. By means of a combination of mail and telephone contacts (or follow-up), some other information about the chosen products, their description, characteristics and specifications are also obtained. The collection of these last kinds of data is essential to monitoring services' specifications connected with the service's price. This also allows the comparison of the collected prices month by month (the quality and quantity of the products considered must be the same over time). After this contact the price registered may be adjusted for the observed quality change of the product. This operation has to be done every time is necessary, in fact prices collected must always be representative of the current period production. Therefore, if there are new products or new

---

<sup>37</sup> I.e. the volume of dollars sold.

<sup>38</sup> Both the selling and the acquisition price are intended as strictly real transaction prices or, at least, equal to the transaction price. More in the details, both prices are intended as an average of monthly price. Therefore the recorded price reflects an average of actual transactions and incorporates the prices of all features found in an actual transaction. The prices collected should be also not lagged, that is representative of the current month.

product's futures, there should be a necessary substitution of products (of the items become obsolete) or an update of the models, finalized to reflect the recent changes. The self filled worksheet is considered useful for the respondents to initiate a new item into the index.

In the last part of the questionnaire a certification of the respondent is requested: he/she has also to provide information about the name and the contacts of a person from which further information can be obtained. Finally the respondent is asked about the time spent to complete the questionnaire and if he/she wants to receive the next quarter's questionnaire in a pre-filled version.

The detailed version of the questionnaire is available on the Statistics Canada website: [http://www.statcan.gc.ca/imdb-bmdi/instrument/5106\\_Q1\\_V1-eng.pdf](http://www.statcan.gc.ca/imdb-bmdi/instrument/5106_Q1_V1-eng.pdf).<sup>39</sup>

## **2.4 First phases of analysis and state of the art**

The collected preliminary data was provided from the Price Division of Statistics Canada. The original dataset contains data of the 2006 wholesale survey. The total number of surveyed units is 42,081.

The BSMD (Business Survey Methods Division) provided the following datasets: the frame dataset (used for the selection of the sample), the system of weights, and a dataset about the detailed status of every sampled unit.

The first phase of the analysis was the integration of the different datasets. The total number of analyzed variables is 124. Some re-iterations of the analysis of all the variables were made to check for inconsistencies or other issues in the preliminary collected data: the main challenges were some incoherencies, the presence of some missing values, the presence of outliers (identifiable with the implementation of various outliers' detection methods).

---

<sup>39</sup> The questionnaire is also available in Appendix 2.1.

The first step of this study was the generation of a simulated population based on the preliminary data, that will be discussed in the following chapter (chap. 3).

The preliminary collected data are being reviewed by the Price Division. The first release is expected during 2009.

Generally speaking, the research's project is very interesting, from the methodological point of view, because this work, even if it's made on a simulated population, can give a strong contribution to the optimization of the price indexes' computation. In fact, some of the conclusions we proposed to gain, should be useful not only to be implemented in the computation of other kind of indexes, but, more in general, they can be applied also to other fields of research.



### 3 Simulated data generation: the generation process

The main goal of the project is to apply the different imputation methods to the preliminary data collected by the wholesale survey, to understand which method is more appropriate.

One of the most relevant questions is the influence of the frame's biggest units.

The variable object of study is the profit margin ratio.

We define the “profit margin” in the following way:

$$PM_{j,i}^t = p_{j,i}^S - p_{j,i}^P$$

where:

- $p_{j,i}^S$  is the selling price observed in the month  $t$  on the  $j^{\text{th}}$  enterprise ( $j = 1, 2, \dots, n$ ) for the  $i^{\text{th}}$  product ( $i = 1, 2, 3$ );
- $p_{j,i}^P$  is the analogue purchase price observed in the same month for the same enterprise  $j$  and product  $i$ ;

All prices are computed as monthly average of transaction prices.

The profit margin ratio,  $R_{j,i}^t$ , is the ratio of the profit margin computed for a specific time  $t$  on the analogue profit margin computed for the previous time ( $t-1$ ):

$$R_{j,i}^t = \frac{PM_{j,i}^t}{PM_{j,i}^{t-1}}.$$

This last variable is very important because the price index's computation is based on the profit margin ratios computed for each observed product of each establishment unit.

In the wholesale sample, for each establishment, the prices about three different products considered in each quarter are registered.

In our analysis we hypothesize that each product is an independent product, and not linked to a specific establishment. Also it was decided to do not consider the sequence of the profit margin ratios along the different reference months. This means that the 3 months of data in the first quarter are combined and considered as they were one group of different units (that is as they were data about a single month). If, for example, we would have to manage with three different establishment only, we have the data about 3 product registered in each of the four quarter of the year. The total number of units is 36 (3 products \* 4 quarters \* 3 establishments), that in the working dataset are considered as independent units. This was decided to have enough units for each cell. In our opinion, the presence, in each cell, of the same product observed in different times will not affect the validity of the study, but could give more strong confirmation to the distribution of the variable object of study in the cell itself.

Before collapsing together all the data (independently from the time of observation), their seasonal connotation was studied to check if some seasonality in the profit margins could be found in the first year of observation (2006). For this reason the mean of price relatives were studied. In this paper detailed results of this phase of the analysis are not presented. Nevertheless, the most relevant conclusion is that the different seasonality patterns of the profit margin ratios are mainly related to the category of the considered products and, secondarily, to the size of the establishment. This was considered a suggestion useful to choose the variables used to define the working cells.

Considering the limited preliminary data available at the time of this project, it was decided to work with simulated data. In fact:

- The preliminary data are being reviewed by the Prices Division, due to quality issues (mainly related to missing data, incoherences and outliers); the first phase of this analysis (and, in particular, the process of simulated data generation) was carried over excluding these outliers.
- Furthermore there was the need to work with a number of data larger than the one available from the sample. So it was decided to generate a simulated population of price margin ratios for a bigger number of units. The frame population was

considered the biggest number of units; this strategy would also make possible to work from the earliest stage of the survey (that is the selection of the sample).

The main problem was that we didn't have any information about the profit margin ratios of the whole population; for this reason it was necessary to find a way to use the data collected with the sample (and the computed profit margin ratio) referring them to the entire frame population.

It was then necessary to generate profit margin ratios for the entire population that would have the same distribution of the sample observed through the 2006 survey.

So the starting points to generate the data were considered on one hand the frame population (all the wholesale establishments included in the *Business Register* whose are used to select the survey's sample) and the sample data about profit margin ratios on the other hand. The dataset made by simulated data needed to be built in a way that was as close as possible to the target population (the frame population used to select the sample) and with a distribution (regarding the variable of interest) as close as possible to the one observed on the sample selected for the survey, that could provide real data about prices and profit margin ratios.

For this reason it was decided to work at a ratios' micro-level generation (that is working at the level of single cells). This level was considered useful to gain, at the same time, detailed results that would be as close as possible to the observed data.

But at this point it was necessary to find the most appropriate variables to define the working cells: the study of the seasonality of the profit margin ratios suggested to use variables representative of the specific activity and of the size of the business units.

### **3.1 The cells definition**

The definition of the cells (i.e. the stratification of the sample and of the generated data) was made considering two different variables with two fundamental characteristics of each

establishment: the sector of activity (the “NAICS” code) and the size of the establishment (the variable “Establishment Revenue”).

Both of these variables are available on the *Business Register* (*BR*, the frame dataset introduced in paragraph 2.3) that includes all the industries of the Canadian economy (so the data are available about all the establishments of the whole population). As seen, these data are monthly updated by the *Business Register Section*, so the risk to work with not updated or incorrect data is small.

### **3.1.1 The NAICS code**

The “NAICS” (*“North American Industry Classification Methods”*) code is the official classification used for Industry in North America. In our study it was decided to give a strictly link with the kind of activity of the wholesalers, defining the studied cell. This was also done because the different products were considered similar, working with cell built in this way. The 2007 version of the classification was used.

Different kinds of NAICS’ details (4, 5 or 6 digits) were tested (observing the distribution of the units of the sample) before deciding which one was the best for this research.

The final NAICS version used to classify the establishments of the population is the 4-digits level; this was decided mainly to avoid the risk to do not have units enough for each cell.

In the Table 3.1 the codification of the three and four digits NAICS codes for the Wholesale Trade sector is shown<sup>40</sup>. Using the 4 digits NAICS, for the population of trade wholesalers we are studying, a total number of 25 classes is considered.

---

<sup>40</sup> A more detailed explanation of 2007 Wholesale trade’s NAICS codes is available in Appendix 3.1.

<b>41</b>	<b>Wholesale Trade</b>
<b>411</b>	<b>Farm Product Wholesaler-Distributors</b> <small>CAN</small>
4111	Farm Product Wholesaler-Distributors <small>CAN</small>
<b>412</b>	<b>Petroleum Product Wholesaler-Distributors</b> <small>CAN</small>
4121	Petroleum Product Wholesaler-Distributors <small>CAN</small>
<b>413</b>	<b>Food, Beverage and Tobacco Wholesaler-Distributors</b> <small>CAN</small>
4131	Food Wholesaler-Distributors <small>CAN</small>
4132	Beverage Wholesaler-Distributors <small>CAN</small>
4133	Cigarette and Tobacco Product Wholesaler-Distributors <small>CAN</small>
<b>414</b>	<b>Personal and Household Goods Wholesaler-Distributors</b> <small>CAN</small>
4141	Textile, Clothing and Footwear Wholesaler-Distributors <small>CAN</small>
4142	Home Entertainment Equipment and Household Appliance Wholesaler-Distributors <small>CAN</small>
4143	Home Furnishings Wholesaler-Distributors <small>CAN</small>
4144	Personal Goods Wholesaler-Distributors <small>CAN</small>
4145	Pharmaceuticals, Toiletries, Cosmetics and Sundries Wholesaler-Distributors <small>CAN</small>
<b>415</b>	<b>Motor Vehicle and Parts Wholesaler-Distributors</b> <small>CAN</small>
4151	Motor Vehicle Wholesaler-Distributors <small>CAN</small>
4152	New Motor Vehicle Parts and Accessories Wholesaler-Distributors <small>CAN</small>
4153	Used Motor Vehicle Parts and Accessories Wholesaler-Distributors <small>CAN</small>
<b>416</b>	<b>Building Material and Supplies Wholesaler-Distributors</b> <small>CAN</small>
4161	Electrical, Plumbing, Heating and Air-Conditioning Equipment and Supplies Wholesaler-Distributors <small>CAN</small>
4162	Metal Service Centres <small>CAN</small>
4163	Lumber, Millwork, Hardware and Other Building Supplies Wholesaler-Distributors <small>CAN</small>
<b>417</b>	<b>Machinery, Equipment and Supplies Wholesaler-Distributors</b> <small>CAN</small>
4171	Farm, Lawn and Garden Machinery and Equipment Wholesaler-Distributors <small>CAN</small>
4172	Construction, Forestry, Mining, and Industrial Machinery, Equipment and Supplies Wholesaler-Distributors <small>CAN</small>
4173	Computer and Communications Equipment and Supplies Wholesaler-Distributors <small>CAN</small>
4179	Other Machinery, Equipment and Supplies Wholesaler-Distributors <small>CAN</small>

*Continues on the next page...*

...continues from the previous page

<b>418</b>	<b>Miscellaneous Wholesaler-Distributors</b> <sup>CAN</sup>
4181	Recyclable Material Wholesaler-Distributors <sup>CAN</sup>
4182	Paper, Paper Product and Disposable Plastic Product Wholesaler-Distributors <sup>CAN</sup>
4183	Agricultural Supplies Wholesaler-Distributors <sup>CAN</sup>
4184	Chemical (except Agricultural) and Allied Product Wholesaler-Distributors <sup>CAN</sup>
4189	Other Miscellaneous Wholesaler-Distributors <sup>CAN</sup>
<b>419</b>	<b>Wholesale Electronic Markets, and Agents and Brokers</b> <sup>US</sup>
4191	Wholesale Electronic Markets, and Agents and Brokers <sup>US</sup>

**Table 3.1** – NAICS (North American Industry Classification Method): 3 and 4 digits Wholesale Trade's classification (2007 version).<sup>41</sup>

In the Table 3.2 the distribution by 4 digits NAICS code of the sample's establishments is compared with the distribution of the entire duplicated population<sup>42</sup> in terms of percentage on the total number of units. The last column shows the differences between the sample's and the population's percentages by NAICS. If, in a specific activity code, the sample shows an over-coverage or an under-coverage the sign of the difference is, respectively, positive or negative. For example, -0.7% (NAICS: 4111) means that the percentage of the units in the 4111 code is proportionally less represented in the sample of 0.7 percentage points (2.3% in the duplicated population, 1.6% in the sample). The 4143 NAICS, for example, is more represented in the sample than in the population of 3.4 percentage points (in fact the percentage weight in the population is 2.5%, while the analogue weight in the sample is 5.9%). The differences shown in the table are probably a consequence of the sample selection strategy, that is base on a PPS (Probability Proportional to Size) sample selection method (in the table the size of the units is not considered).

<sup>41</sup> The superscript symbols at the end of NAICS class titles used to signify comparability are:

- CAN Canadian industry only;
- MEX Canadian and Mexican industries are comparable;
- US Canadian and United States industries are comparable;
- [Blank] [No superscript symbol] Canadian, Mexican and United States industries are comparable.

<sup>42</sup> The duplicate population (that will be considered the frame population) is defined in par. 3.3.

	Duplicated population	Sample	
NAICS	%	%	Diff. (Sample-Pop. in % points)
4111	2.3	1.6	-0.7
4121	1.8	0.5	-1.2
4131	10.3	5.8	-4.5
4132	1.1	1.3	0.3
4133	0.1	0.2	0.1
4141	4.6	5.9	1.3
4142	1.0	2.4	1.4
4143	2.5	5.9	3.4
4144	5.0	7.5	2.4
4145	2.6	6.7	4.1
4151	2.5	2.7	0.3
4152	4.4	6.5	2.1
4153	0.8	0.9	0.1
4161	4.8	6.1	1.3
4162	1.7	5.3	3.6
4163	8.2	8.9	0.7
4171	3.1	2.0	-1.0
4172	9.1	7.4	-1.8
4173	5.0	5.3	0.3
4179	9.0	9.9	0.9
4181	2.5	1.8	-0.8
4182	1.8	0.4	-1.4
4183	2.4	1.6	-0.8
4184	2.2	0.9	-1.3
4189	11.2	2.4	-8.8
	<b>100.0</b>	<b>100.0</b>	<b>0.0</b>

**Table 3.2** – Distribution of the duplicated population<sup>43</sup> and of the survey sample by NAICS.

### 3.1.2 The establishment revenue

The best variable to give an idea of the size of each establishment considered in the study is considered the establishment revenue.

<sup>43</sup> The duplicate population (that will be considered, in the following, the frame population) is defined in par. 3.3.

To define the size classes and to study the distribution of the profit margin ratios by them, it was decided to recode the establishment revenues. The classes are defined basing on the distributions by deciles in every NAICS' class.

Starting from the sample data, for each NAICS it is defined the distribution of the variable "revenue" by deciles. So firstly, in each NAICS, the units are ordered by establishment revenue. Then the units of the same NAICS are attributed to the different classes according their belonging to the decile of the cell's distribution.

So, for example, the first class (class "1") of a NAICS includes units with a revenue lower than the first decile (identified for that NAICS), class "2" comprises the units with a revenue included between first decile (included) and second decile (excluded), and so on; the tenth class ("10") comprehends the units with a revenue higher than the class' tenth decile. The Table 3.3 could be useful to understand the recoding process.

Class	Units with revenue...	
	...higher than...	...lower than...
1		1 <sup>st</sup> decile
2	1 <sup>st</sup> decile	2 <sup>nd</sup> decile
3	2 <sup>nd</sup> decile	3 <sup>rd</sup> decile
4	3 <sup>rd</sup> decile	4 <sup>th</sup> decile
...	...	...
9	9 <sup>th</sup> decile	10 <sup>th</sup> decile
10	10 <sup>th</sup> decile	

**Table 3.3** – Definition of the revenue classes (by deciles).

In the Table 3.4 the distribution by revenue classes of the sample and of the duplicate population<sup>44</sup> is shown. The differences in the last column show if a revenue class is under-represented (negative sign) or over-represented (positive sign) by the sample in comparison with the frame population. The high level of under-representation (proportionally to the weight of the class in the duplicated population) of the first class in the sample (-78.9%) is due to the preliminary exclusion from the survey of many small units, usually classified as "Take None" (or "TN") units<sup>45</sup>.

<sup>44</sup> The duplicate population is defined in par. 3.3.

<sup>45</sup> The *TN* / *Take None* units are units excluded from the survey (about the definition of TN Units, see par. 2.3.1).



REV. CLASS	Duplicate population		Sample		Diff. (%) <i>Smpl-Pop</i>
	%	% Cum.	%	% Cum.	
1	87.8	87.8	8.9	8.9	-78.9
2	4.3	92.1	9.9	18.8	5.6
3	2.5	94.6	9.9	28.7	7.4
4	1.7	96.3	10.2	38.9	8.5
5	0.9	97.2	10.1	49.0	9.2
6	0.7	97.9	9.9	58.9	9.2
7	0.7	98.6	9.8	68.7	9.1
8	0.5	99.1	10.3	78.9	9.7
9	0.5	99.5	9.9	88.9	9.5
10	0.5	100.0	11.1	100.0	10.7
	<b>100.0</b>		<b>100.0</b>		

**Table 3.4** – Distribution of duplicated population and sample by revenue classes (deciles).

### 3.1.3 Some adjustments

Before starting to test and define the distribution of the units in each cell, some operation needed to be done on the units of the sample themselves.

The products with a purchase price or with a selling price equal to zero were erased; this means a reduction of the sample from 42,245 to 38,164 units.

The profit margins and the ratios (variables already in the original survey's dataset) were computed again to check the consistency and the coherence of the data. Were some discrepancies was found, the new data was substituted to the previous one.

It was also decided to remove, for each cell, some outliers that would condition too much the distribution of the data. This means that firstly the units with a revenue equal to 1 or 0 were removed from the frame (as they were considered missing data and it was not possible to link them to one of the defined cells); it was then decided to remove also the units with a profit margin ratio lower than -50 (7 units) and higher that 100 (10 units), considering these last like outliers values.

All these phase brought to have a final number of 37,873 units to process.

### **3.1.4 Distributions of the units by cells**

Considering the NAICS code and the revenue classes, a total number of 250 cells are identified (25 NAICS codes multiplied by 10 decile classes). The distribution by cells of the sample selected for the Wholesale survey is shown in Appendix 3.2.

For some cells the initial number of available units was not considered enough to carry on the distribution analysis in the most appropriate way. For example, if the number of units is extremely reduced, it would be not possible to test the distribution in a given cell or the results of the test would have no significance. It was decided to further aggregate the revenue classes differently from a NAICS to the other one. 30 units or more is considered a number of units enough for each cell.

At the end of this phase, a code comprising the NAICS and the adjusted revenue's classes was given to each unit of the population.

Once an appropriate recoded distribution is made, starting from the NAICS and the establishment revenue data available from the *BR*, it is possible to generate a new simulated dataset based on the distribution observed in each cell, considering the sample data collected in the first year of the survey.

The following step is to study and identify the distribution of the studied variable (that is the profit margin ratio) for each single cell.

## **3.2 Testing the distribution by cells**

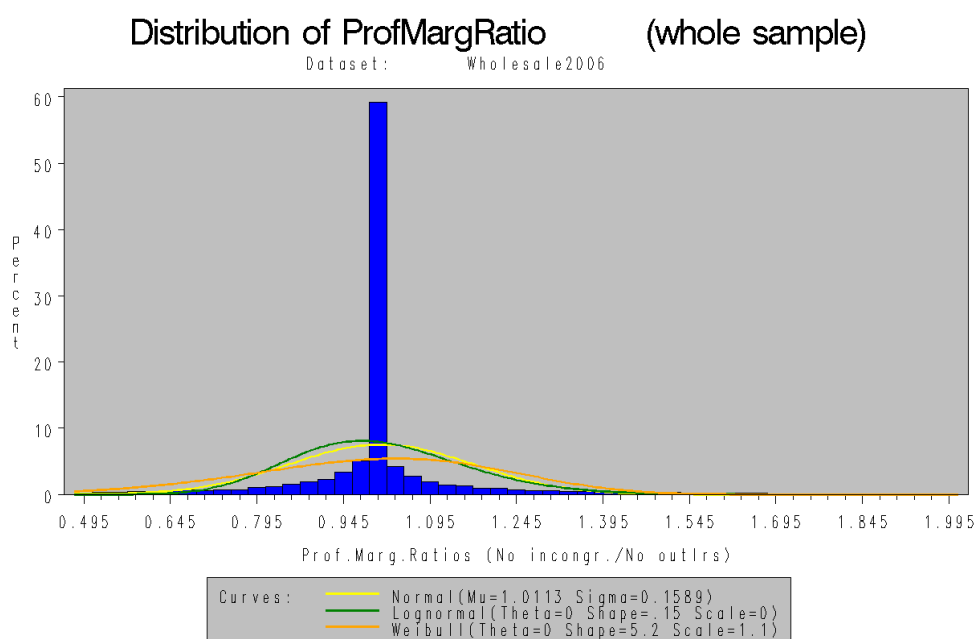
The goal of this phase of the work is the generation of new data that would be as close as possible to the real distribution observed in the sample selected for the first wave of the survey. To get this target, once the outliers are removed from the sample, it was decided to firstly generate data uniform distributed between the lowest and the highest observed value, for each cell. But this was considered a simple way to produce the new dataset, that would not bring to a simulated population really close to the characteristics of the sample selected.

Considering the Graph 3.1 it's clear that the general distribution of profit margin ratios in the whole sample observed is not always similar to the one observed in the two cells of the Graph 3.2 and Graph 3.3.

On the other hand, it was clear that the uniform distribution was not similar to any of the distributions observed on the sampled establishment, neither considering the stratification by NAICS, nor the stratification by revenue deciles.

Therefore the distribution is studied more deeply, that is by cell, to identify the theoretical distribution (and its parameters) more close to the observed one. So it was decided to test the normal, lognormal and Weibull distribution for each one of the cell<sup>46</sup>.

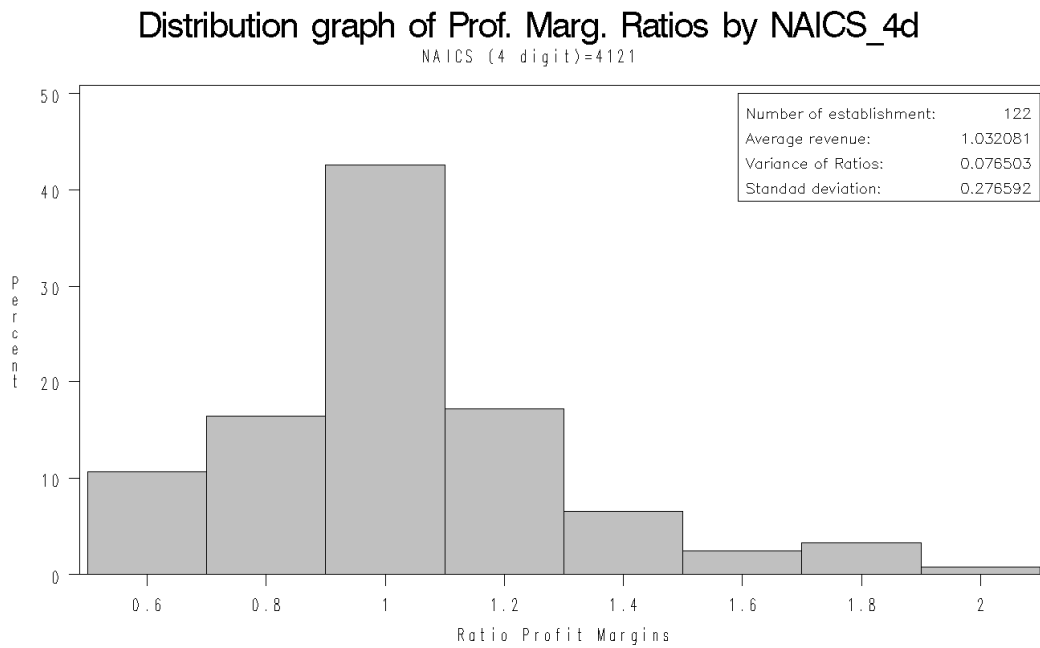
The main characteristics of these distributions are introduced in the following paragraphs<sup>47</sup>.



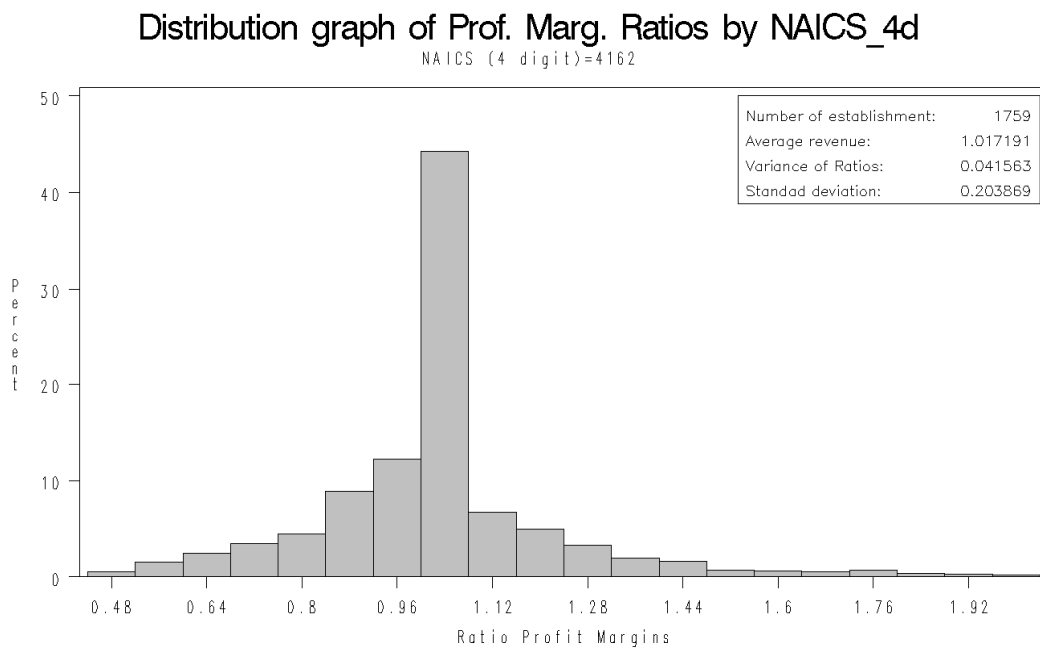
**Graph 3.1 – Profit margin ratios distribution (all the sample).**

<sup>46</sup> Where the number of collected data would make it possible.

<sup>47</sup> For further information about the characteristics of the distributions and about the distribution's tests, see also Murthy (1967), Cochran (1977), Stephens (1974), Chambers et al. (1983), Johnson et al. (1994, 1995).



**Graph 3.2** – Profit margin ratios distribution (NAICS: 4121).



**Graph 3.3** – Profit margin ratios distribution (NAICS: 4162).

### 3.2.1 Uniform distribution

The uniform distribution considers the units of the population uniformly distributed along a range comprised between the lowest and the highest observed value.

Seen the distribution graphs of all the sample data and the distributions by NAICS and by revenue, it was considered not useful to use the uniform distribution to generate data, because it was similar to no one of the distribution observed by cells.

### 3.2.2 Normal distribution

The Normal distribution is characterized by two parameters: the mean ( $\mu$ ) and the standard deviation ( $\sigma$ ).

The probability density function of the normal distribution is identified once the mean and the scale parameter (the standard deviation) are identified:

$$f(y) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp \left[ -\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right] \quad -\infty < y < \infty.$$

The cumulative distribution function is:

$$F(y) = \Phi \left( \frac{y - \mu}{\sigma} \right),$$

where  $\Phi$  is the cumulative distribution function of the standard normal variable, like the one shown in the following:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp \left( -\frac{u^2}{2} \right) du.$$

### 3.2.3 Lognormal distribution

The lognormal distribution is defined with the following three parameters: the threshold parameter ( $\theta$ ), the scale parameter ( $\zeta$ ) and the shape parameter ( $\sigma$ ).

The probability density function is:

$$f(y) = \frac{1}{y - \theta} \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp \left[ -\frac{1}{2} \left( \frac{\log(y - \theta) - \zeta}{\sigma} \right)^2 \right] \quad y > \theta.$$

The cumulative distribution function of the lognormal distribution is:

$$F(y) = \Phi \left( \frac{\log(y - \mu) - \zeta}{\sigma} \right) \quad y > \theta.$$

### 3.2.4 Weibull distribution

The parameters that define the Weibull distribution are: the threshold parameter ( $\theta$ ), the scale parameter ( $\sigma$ ,  $\sigma > 0$ ) and the shape parameter ( $c$ ,  $c > 0$ ).

The probability density function of a Weibull distribution is:

$$f(y) = \frac{c}{\sigma} \left( \frac{y - \theta}{\sigma} \right)^{c-1} \exp \left[ - \left( \frac{y - \theta}{\sigma} \right)^c \right] \quad y > \theta, c > 0.$$

The cumulative distribution is:

$$F(y) = 1 - \exp \left[ - \left( \frac{y - \theta}{\sigma} \right)^c \right] \quad y > \theta.$$

### 3.2.5 Tests to identify the distribution of each cell

To identify the distribution form of each cell, three different tests were used and applied: the Kolmogorov-Smirnov statistic ( $D$ ), the Anderson-Darling statistic ( $A^2$ ) and the Cramer-von Mises statistic ( $W^2$ ). All the used statistics are based on the *Empirical Distribution Function* and for this reason they are called *EDF (Empirical Distribution Function)* tests.

Given a group of  $n$  independent observations with the same normal distribution function  $F(x)$ , if we call these units as  $X_1, X_2, \dots, X_n$  and if we use the symbols  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  to indicate the ordered units<sup>48</sup>, the empirical distribution function,  $F_n(x)$ , is defined as a step function (with an height of each step equal to  $1/n$ ) like the following:

---

<sup>48</sup>  $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ .

$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ i/n & X_{(i)} \leq x < X_{(i+1)}, \quad i = 1, 2, \dots, n-1 \\ 1 & X_{(n)} \leq x \end{cases}$$

This means that  $F_n(x)$  is the proportion of the units with an observed value less than or equal to  $x$ .<sup>49</sup>

The *EDF* tests we used are based on the discrepancies between the empirical distribution function  $F_n(x)$  and the distribution function (also defined “parametric cumulative distribution function”)  $F(x)$ .

The power of this kind of test is bigger than the one of the chi-squared goodness-of-fit test; moreover it is invariant with respect to the histogram midpoints<sup>50</sup>.

The output of this kind of test comprises statistics of adaptation of the empirical distribution to the fixed distribution that one wants to test, but also parameters of the distributions tested themselves, already introduced in par. 3.2.2, 3.2.3 and 3.2.4. These could be used to build a population, referred to each cell, that has the same distribution of the tested distribution.

#### 3.2.5.1 TEST 1: the Kolmogorov-Smirnov statistic ( $D$ )

The Kolmogorov-Smirnov statistics ( $D$ ) is defined as the highest difference, for each  $x$ , between the empirical cumulated distribution function observed on the  $n$  units considered,  $F_n(x)$ , and the cumulated parametric distribution function,  $F(x)$ , that one wants to test.

$$D = \sup_x |F_n(x) - F(x)|.$$

#### 3.2.5.2 TEST 2: the Anderson-Darling statistic ( $A^2$ )

The Anderson-Darling statistic<sup>51</sup> is a test of the *EDF*-squared family: this means that the evaluated differences between the empirical distribution function and the cumulated distribution functions (seen above) are considered in their squared version. In fact the

---

<sup>49</sup> We remember that  $F(x)$  is the probability to have a unit with value less or equal to  $x$ .

<sup>50</sup> For more details about the *EDF* tests, see D’Agostino and Stephens (1986).

<sup>51</sup> Anderson and Darling (1954).

statistic  $A^2$  is computed as the integral on all the observed values of the squared differences between the empirical and the theoretical distribution.

$$A^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \psi(x) dF(x).$$

The differences are weighted with a function  $\psi(x)$ :

$$\psi(x) = [F(x)(1 - F(x))]^{-1}.$$

### 3.2.5.3 TEST 3: the Cramèr-von Mises statistic

The other test used to define the distribution of each cell was the Cramèr-von Mises statistic ( $W^2$ ), that is an *EDF*-statistics of the squared class too. The test, similar to the previous one, is based on the squared differences between the empirical distribution and the distribution that one wants to test. The differences are weighted by a function  $\psi(x)$  that, in this case, is considered equal to 1 ( $\psi(x) = 1$ ):

$$W^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 dF(x).$$

### 3.2.5.4 An example of distribution tests

In Graph 3.4 there is one of the outputs obtained for the distribution test (NAICS: 4181; revenue class # 4). The blue bars represent the real observed distribution of profit margin ratios in the considered cell.

The yellow line is the normal distribution estimated on the observed units of that cell: in the graph the parameters of the normal distribution ( $\mu = 1.0103$ ,  $\sigma = 0.0784$ ) are also shown. As seen above, the Kolmogorov-Smirnov statistic ( $D$ ) consider the highest difference (for each observed value  $x$ ) between the empirical distribution and the distribution that one wants to test.

In the same Graph 3.4 there are the parameters estimated by the software for the Lognormal ( $\theta = 0$ ,  $\zeta = 0.01$ ,  $\sigma = 0.08$ ) and the Weibull ( $\theta = 0$ ,  $c = 13$ ,  $\sigma = 1$ ) distribution. These parameters are useful to generate a distribution (of one of the three kind) basing on the characteristics observed on the real data.



Even in these last cases, to test the distribution the differences between empirical and theoretical distribution are used. But, this time, it is made considering all the range of observed values. So the differences used to test the distribution are computed along all the range and are summed up with an integral function.

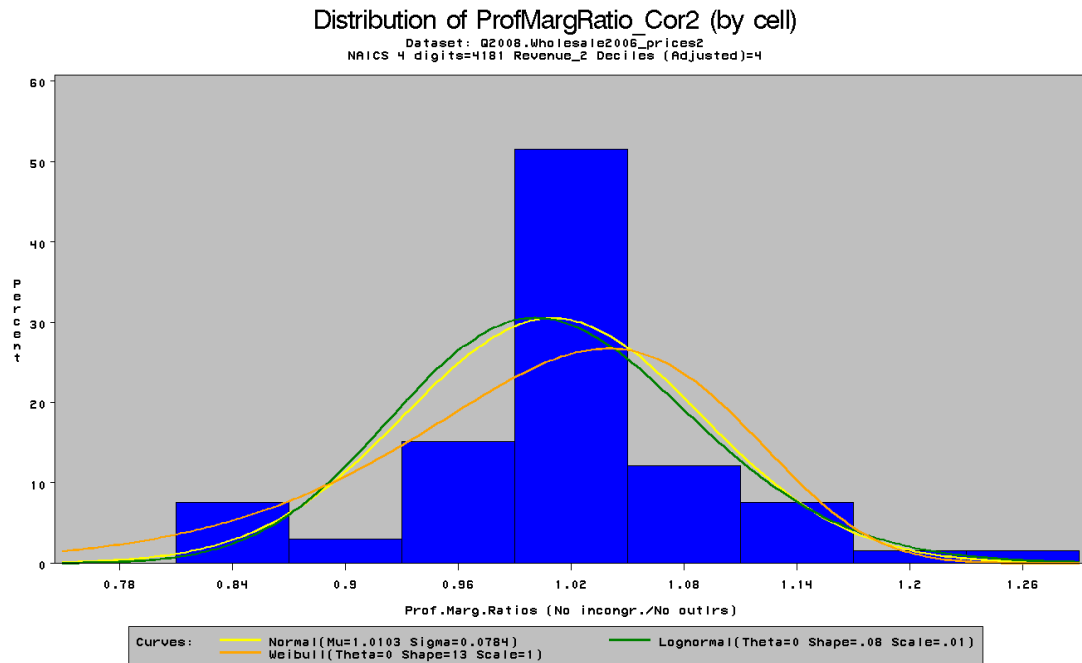
In Graph 3.5 there is the output of the Kolmogorov-Smirnov test obtained with the SAS/INSIGHT software<sup>52</sup>. In the first part of the graph, the cumulative distribution functions are represented; they are at the base of the tests computation<sup>53</sup>. In the second part of the graph (third, fourth and fifth columns) the estimated parameters for each one of the tested distribution (plus the exponential distribution) are available. The exponential distribution was excluded from the generation of data process, because, after the computation of the test on each cell, it was the most far from the observed data distribution. It's possible to see this in the first part of Graph 3.5, where the light blue line is very far from the empirical distribution (red line). More close to the empirical cumulated distribution function are the normal (pink), the lognormal (orange) and the Weibull (green) distributions.

In the last column of the second part of Graph 3.5 there is the statistical significance for the test. The value of the test is in the previous column: the better distribution (that is the closest distribution to the observed data) is the one who have the lowest value (the minimum distance between the empirical and the theoretical distribution tested).

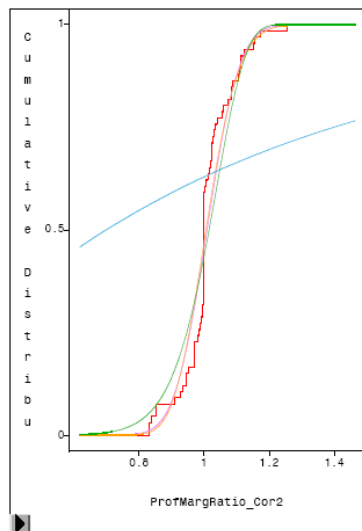
---

<sup>52</sup> For further details about the software used to test the distribution, see: SAS/INSIGHT User's Manual (SAS Institute Inc., 1999).

<sup>53</sup> For further details about the test, see Chambers et al. (1983).



**Graph 3.4** – Distribution tests (NAICS 4181, Revenue class: 4): graphs and estimated parameters.



Tests for Distribution						
Curve	Distribution	Mean/Theta	Sigma	Zeta/C	Kolmogorov D	Pr > D
<span style="color: blue;">—</span>	Normal	1.0103	0.0784	.	0.1498	<.01
<span style="color: red;">—</span>	Lognormal	0	0.0779	0.0072	0.1551	<.01
<span style="color: green;">—</span>	Exponential	0	1.0103	.	0.5604	<.01
<span style="color: magenta;">—</span>	Weibull	0	1.0466	12.5998	0.1813	<.01

**Graph 3.5** – Output tests (NAICS 4181, Revenue class: 4): cumulated distribution and Kolmogorov test.

### 3.2.6 Criteria to identify the distributions

Firstly the Anderson-Darling and the Cramer-von Mises statistics are considered to attribute, to each cell, the distribution more close to the real data. This choice was done because the first two tests are considered less approximate than the Kolmogorov-Smirnov statistic; in fact, they compute a sum of the differences between the theoretical and the empirical function along all the range of the observed values, and they do not only take into consideration the maximum difference.

It is not possible to apply a test to all the 250 cells: some cells are not represented by the sample selected, and for some others there are not enough data to test the distribution. So, working with frame data, it is only possible to find results about 240 cells, obtaining a percentage coverage of the frame population equal to 95.8% (in terms of number of cells) or 96.9% (in terms of number of units).

The classification's results obtained with the first two tests (the Anderson-Darling and the Cramer-von Mises test) are compared in the first part of Table 3.5. The cells with the same attributed distribution, using the two tests, are on the diagonal of the table (229 cells, that is 95.42% of the total number of studied cells).

	Cramer-von Mises				
<b>Anderson-Darling</b>	Normal	Lognormal	Weibull	<b>TOT</b>	<b>%</b>
Normal	41	6	1	<b>48</b>	<b>20.0</b>
Lognormal	2	173	0	<b>175</b>	<b>72.9</b>
Weibull	1	1	15	<b>17</b>	<b>7.1</b>
<b>TOT</b>	<b>44</b>	<b>180</b>	<b>16</b>	<b>240</b>	<b>100.0</b>
<b>%</b>	<b>18.3</b>	<b>75.0</b>	<b>6.7</b>	<b>100.0</b>	

**Table 3.5** – Distribution tests (all the cells – sample data): Anderson-Darling vs Cramer-von Mises.

For some cells (11, 4.58% of the total) the two first tests, used in Table 3.5, identify different distributions. This means that Anderson-Darling's test and Cramer-von Mises's statistic don't give the same results. For the undefined situations, the Kolmogorov-Smirnov statistics are considered together with the study of the empirical distribution of the units in each cell.

In Table 3.6 the results of the Kolmogorov-Smirnov test on all the 240 cells are shown.

Kolmogorov-Smirnov	Cells	%
Normal	68	28.3
Lognormal	172	71.7
Weibull	0	0.0
<b>TOT</b>	<b>240</b>	<b>100.0</b>

**Table 3.6** – Distribution tests (all the cells – sample data): Kolmogorov-Smirnov.

In Table 3.7 there are the final results of the distribution test for the 11 cells with incoherencies of classification considering the Anderson-Darling (And-Dar) and the Cramer-von Mises (Cra-vMis) tests. The final distribution was decided studying the Kolmogorov-Smirnov (Kol-Smir) statistics and the empirical distribution graphs in each cells of the list.

#	NAICS_4d	Revenue Decile	And-Dar	Cra-vMis	Kol-Smir	DEFINITIVE
1	4121	3	Normal	Weibull	Normal	Weibull
2	4133	6	Normal	Lognormal	Lognormal	Lognormal
3	4144	7	Normal	Lognormal	Normal	Normal
4	4151	3	Lognormal	Normal	Normal	Normal
5	4152	5	Normal	Lognormal	Lognormal	Lognormal
6	4153	5	Weibull	Lognormal	Lognormal	Weibull
7	4161	1	Normal	Lognormal	Normal	Normal
8	4161	6	Normal	Lognormal	Normal	Weibull
9	4181	4	Normal	Lognormal	Normal	Normal
10	4182	7	Lognormal	Normal	Normal	Normal
11	4183	1	Weibull	Normal	Normal	Weibull

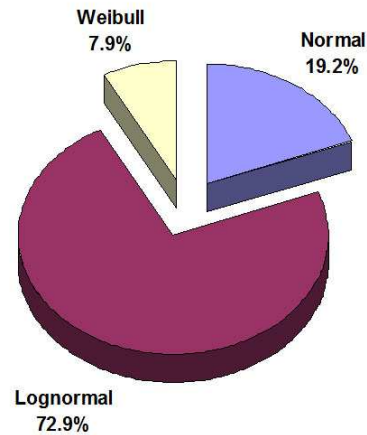
**Table 3.7** – Incoherencies of classification defined with Kolmogorov-Smirnov statistics and graphs.

After these two phases of the study of the distribution, the final classification is shown in Table 3.8. The main distributions are the lognormal (72.9% of the 240 cells) and the normal (19.2%). Less common is the Weibull distribution (7.9%).

#### Final distribution

	Cells	%
Normal	46	19.2
Lognormal	175	72.9
Weibull	19	7.9
<b>TOT</b>	<b>240</b>	<b>100.0</b>

**Table 3.8** – Final classification of the cells.



**Graph 3.6** – Final classification of the cells.

The attribution of the distribution that fits better to each cell, basing on the observed data is not the only result of this first phase of the project. The tests provided also the estimate of the parameters of each tested distribution in every cell. These parameters are useful to generate the simulated data.

### 3.3 Generation of simulated data

The following step of the work is the generation of a simulated population of profit margin ratios. The population generated is based, for each cell, both on the selected distribution and on the estimated parameters of that distribution (as explained in the previous paragraph 3.2).

The parameters for each of the tested distribution for each considered cell are obtained. After the test phase, the distribution parameters of all the tested distributions and of each sample-cell are merged with the original population frame. The used criterion is the belonging of the frame's unit to a specific cell. The parameters in this way could be used to generate three profit margin ratios' variables (one for each tested distribution), providing a simulated value for each unit of the frame.

The generation of the simulated population's profit margin ratios is explained more in detail in the following paragraph<sup>54</sup>.

Before starting with the generation process, it was decided to double the number of the units of the frame population. This was made to obtain an adequate number of cases in each cell and to make possible an appropriate selection of samples, representative of all the population cells. The fact to have a copy (with the same characteristics) of an original unit of the frame is not considered affecting in a bad way the results of the research, because of the random generation process that are used to generate the simulated profit margin ratios.

After the duplication process, we got a doubled frame population: from the original 77,025 units we obtained 154,050 units. For each unit of the population the parameters of the three tested distribution associated were estimated, and for each cell the most appropriate distribution was identified.

For each unit the profit margin ratios are generated in the way shown in the paragraphs 3.3.1, 3.3.2, 3.3.3 and 3.3.4.

Four different series of random number (called  $rn_0$ ,  $rn_1$ ,  $rn_2$  and  $rn_3$  in the following) are generated, one for each of the tested distributions (uniform, normal, lognormal, Weibull). A SAS procedure is used; this procedure starts with the choice of a seed: changing the initial seed is possible to obtain different series of random numbers with the same chosen distribution.

After the generation of three series of data with each of the tested distributions, another distribution, called "mixed", was generated<sup>55</sup>, considering the results of the distribution tests applied to the sample data. The "mixed" distribution, in fact, is made considering a different kind of distribution for each cell: the attributed distribution is chosen basing on the results of the best distribution's detection process seen in paragraph 3.2.6.

The "mixed" distribution is considered the distribution more close to the distribution of the real data, the ones observed on the sample. In this phase the uniform distribution was not considered.

---

<sup>54</sup> See Eandt (1961) and Cohen (1951) for further details.

<sup>55</sup> See also paragraph 3.3.5.

In the following paragraphs  $i$  indicates the  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell ( $i = 1, 2, \dots, n_j$ , where  $n_j$  is the number of unit in the  $j^{\text{th}}$  cell;  $j = 1, 2, \dots, p$ , where  $p = 240$  is the total number of the tested cells).

### 3.3.1 Uniform distribution

To generate the estimated ratio of each  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell ( $\hat{r}_{ij}$ ), a random number  $rn_{0,ij}$  ( $i = 1, 2, \dots, n_j; j = 1, 2, \dots, p$ ) is used. The random number is generated in a way that would have range between 0 and 1 and would have a uniform distribution.

To get this target the SAS' *CALL RANUNI* routine was used<sup>56</sup>.

This random number was generated in a new dataset made of a number of units enough to cover the number of units of the frame's population. After verifying the presence of a random distribution of these generated numbers, this dataset was merged with the population frame using the SAS' simple one-to-one *MERGE* statement<sup>57</sup>.

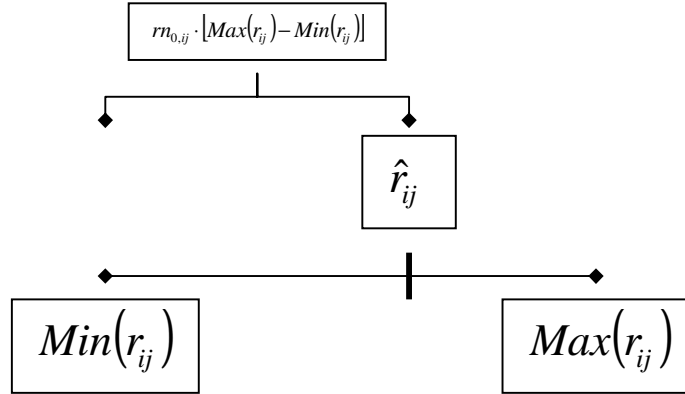
The new simulated ratio  $\hat{r}_{ij}$  is generated using the minimum  $[Min(r_{ij})]$  and the maximum  $[Max(r_{ij})]$  observed ratios, in the  $j^{\text{th}}$  cell (for  $i = 1, 2, \dots, k$ ). The used generation process, starting from random number and the minimum and maximum ratios, is the following:

$$\hat{r}_{ij} = Min(r_{ij}) + \{rn_{0,ij} \cdot [Max(r_{ij}) - Min(r_{ij})]\}.$$

---

<sup>56</sup> See Fisherman and Moore (1982) and SAS User's Manual (SAS Institute Inc., 2004) for details.

<sup>57</sup> See SAS User's Manual (SAS Institute Inc., 2004). The used merging procedure matches the cases of two different datasets line by line: the first unit of the first database is merged with the first unit of the second database; the second unit of the first database is merged with the second unit of the second database, and so on (for this reason the number of units of the first and of the second database should be the same). This one-to-one merging process is possible even if the order of the units of one of the datasets would be changed, seen that the random numbers that we want to match with the frame's units are randomly distributed.



**Figure 3.1** – Generation of a random number with a uniform distribution.

The simulated ratios uniform distributed were not used for the following definition of the “mixed” distribution.

### 3.3.2 Normal

The ratios are generated starting from a random number  $rn_{1,ij}$  ( $0 \leq rn_{1,ij} \leq 1$ )<sup>58</sup> that has a standardized normal distribution with mean  $\mu^{rn_1} = 0$  and mean square error  $\sigma^{rn_1} = 1$ .

The general formula for the standardization of a non-standardized variable  $x_{ij}$  with a normal distribution is (we are referring to a generic  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell):

$$\frac{x_{ij} - \mu_j}{\sigma_j} = z_{ij}, \quad i = 1, 2, \dots, n_j; j = 1, 2, \dots, p$$

Given the  $\mu_j$  and the  $\sigma_j$  of a normally distributed variable and the random number  $z_{ij}$  with a standardized normal distribution, we can use the inverted formula to find a number with a normal distribution ( $x_{ij}$ ):

$$x_{ij} = (z_{ij} \cdot \sigma_j) + \mu_j.$$

---

<sup>58</sup>  $i = 1, 2, \dots, n_j; j = 1, 2, \dots, p$ .



In our case we want to generate the profit margin ratios for the  $j^{\text{th}}$  cell with a normal distribution. We have the parameters of the  $j^{\text{th}}$  cell ( $\mu_j$  and  $\sigma_j$ ) and we have a random number with a standardized normal distribution ( $rn_{1,ij}$ ) with the same characteristics of  $z_{ij}$ .

So, using the same inverted formula, we can generate the estimated ratio  $\hat{r}_{ij}^N$  (where  $N$  indicates a normally distributed ratios) for the unit  $i$  ( $i = 1, 2, \dots, n_j$ ) of the fixed  $j^{\text{th}}$  cell:

$$\hat{r}_{ij}^N = (rn_{1,ij} \cdot \sigma_j) + \mu_j. \quad 3.1$$

As seen from the previous formula 3.1, the generation process depends from the estimated (for each cell) parameters of the normal distribution; this means that, in the inverted formula, we can use the estimated parameters we've obtained testing the normal distribution in each cell.

Using formula 3.1, to generate a normally distributed population with the same observed distribution of the considered cell, the variables containing generated random number  $rn_{1,ij}$  and the estimated parameters for each cell (mean and mean square error) are used.

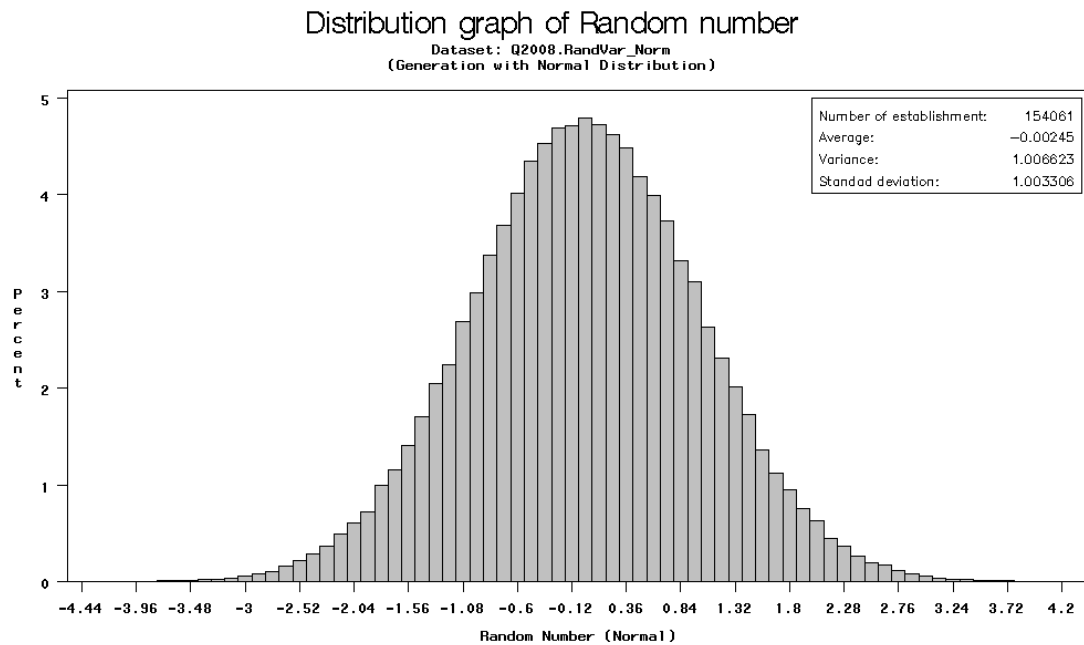
As a result, we obtain another new variable with ratios characterized by the same normal distribution observed on each cell of the sample (that is: the ratios have a distribution, for each cell, characterized by parameters equal to the ones observed in the corresponding cell of the sample).

In the Graph 3.7 the distribution of the 154,061 generated random numbers  $rn_{1,ij}$  with standardized normal distribution is shown (these numbers were generated using the SAS' *CALL RANNOR* routine<sup>59</sup>): the mean is close to 0 and the variance is close to 1, as expected.

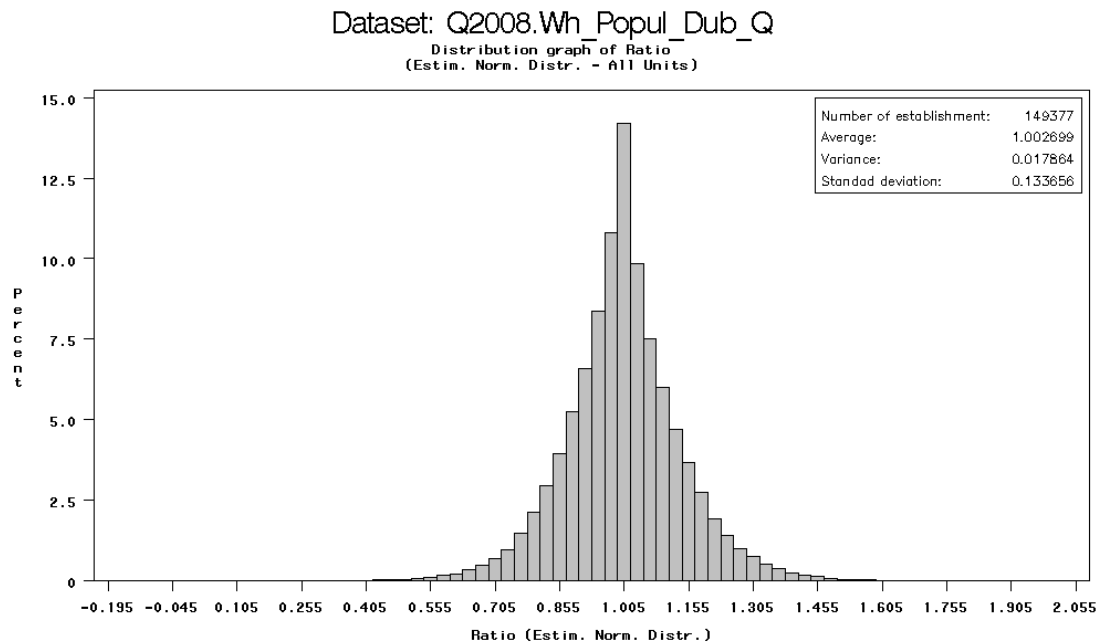
The Graph 3.8 presents the distribution of the profit margin ratios generated with a normal distribution obtained starting from the random numbers  $rn_{1,ij}$  and from the parameters observed in each sample cell.

---

<sup>59</sup> See SAS User's Manual (SAS Institute Inc., 2004) for details. Similarly to the SAS' *Call RANUNI* routine, the *Call RANNOR* routine allows the choice of a seed to start the generation process.



**Graph 3.7** – Distribution of random number  $m_{i,j}$  generated with standardized normal distribution.



**Graph 3.8** – Distribution of profit margin ratios generated with normal distribution.

In this case the profit margin ratios' average (on 149,377 units) is equal to 1.002699. We remind that the unweighted average of the sample was 1.0113171 (33,721 units), a little bit higher, probably for the influence of a proportional bigger presence of the biggest units of the frame<sup>60</sup>.

### 3.3.3 Lognormal

If the variable  $X_2$  is distributed like a normal variable with mean equal to  $\mu$  and mean square error equal to  $\sigma$ , the variable  $Y_2 = \log(X_2)$  has a lognormal distribution:

$$Y_2 \approx \log N.$$

The mean of the  $Y_2$  variable is equal to:

$$\mu_{Y_2} = e^{\left(\mu + \frac{\sigma^2}{2}\right)}$$

and the mean square error is equal to:

$$\sigma_{Y_2} = e^{\sigma^2 - 1} \cdot e^{2\mu + \sigma^2}.$$

To get a generated population of ratios with the same lognormal distribution observed in each cell, it is firstly generated a random number  $rn_{2,ij}$  ( $0 \leq rn_{2,ij} \leq 1$ ) that has a standardized normal distribution with mean  $\mu_j^{rn_2} = 0$  and mean square error  $\sigma_j^{rn_2} = 1$ . The process of generation of the random number is similar to the one used in the previous paragraph 3.3.2 (it's based on the SAS' *Call RANNOR* routine).

This random number has to be transformed to obtain values with a normal distribution (not standardized) similar to the one observed in each cell.

The process is based on the inverted standardization formula:

$$x_{ij}^N = (rn_{2,ij} \cdot \sigma_j) + \mu_j,$$

where  $\mu_j$  and the  $\sigma_j$  are the observed parameters for the  $j$ -th cell ( $j = 1, 2, \dots, p$ ), the random number  $rn_{2,ij}$  has a standardized normal distribution.

---

<sup>60</sup> The sample was selected with a *PPS* (Probability Proportional to Size) sampling method: biggest units have a bigger probability to be included in the sample.

The estimated value  $x_{ij}^N$  for the unit  $i$  ( $i = 1, 2, \dots, k$ ) of the fixed cell is normal distributed.

This value must be transformed to get a lognormal distributed series of values in the following way:

$$x_{ij}^{LogN} = \mu_j^{LogN} + x_{ij}^N \cdot \sigma_j^{LogN},$$

where:

$$\mu_j^{LogN} = \ln(\mu_j) - \frac{1}{2} \ln \left[ 1 + \left( \frac{\sigma_j}{\mu_j} \right)^2 \right]$$

and

$$\sigma_j^{LogN} = \sqrt{\ln \left[ 1 + \left( \frac{\sigma_j}{\mu_j} \right)^2 \right]}.$$

To obtain log-normally distributed simulated ratios, the last step is the following transformation:

$$\hat{r}_{ij}^{LogN} = e^{x_{ij}^{LogN}}.$$

The Graph 3.9 shows the distribution of the profit margin ratios generated with a lognormal distribution (so, the generation process starts from the random numbers  $rn_{2,ij}$  and from the parameters obtained in each sample cell testing the lognormal distribution).

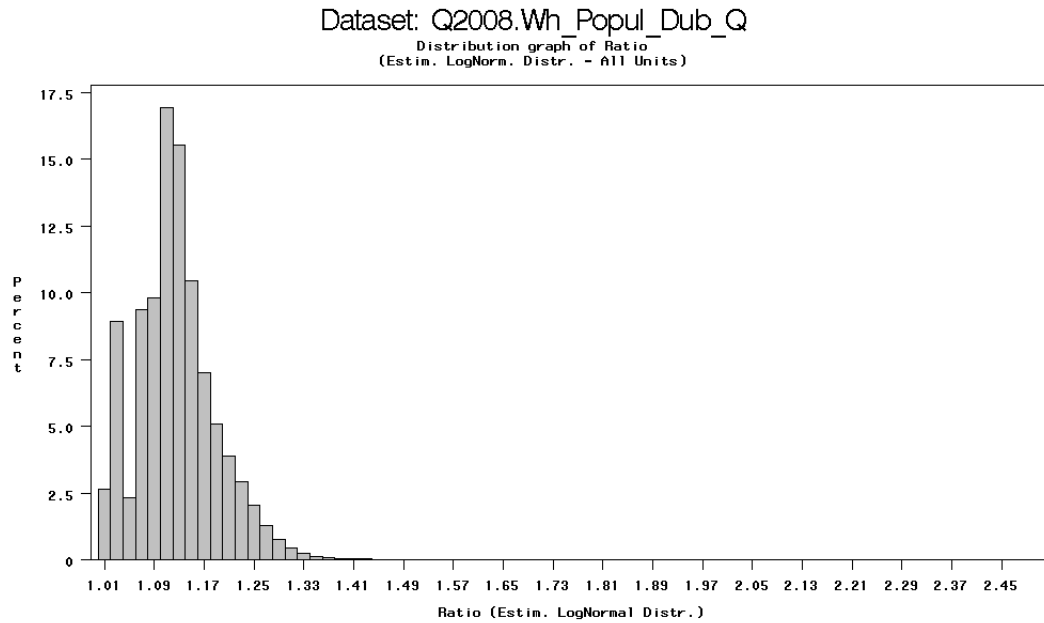
### 3.3.4 Weibull

To generate profit margin ratios that would have a Weibull distribution as close as possible to the one observed on the sample, it is necessary, firstly, to generate a series of random numbers  $rn_{3,ij}$  with uniform distribution<sup>61</sup>.

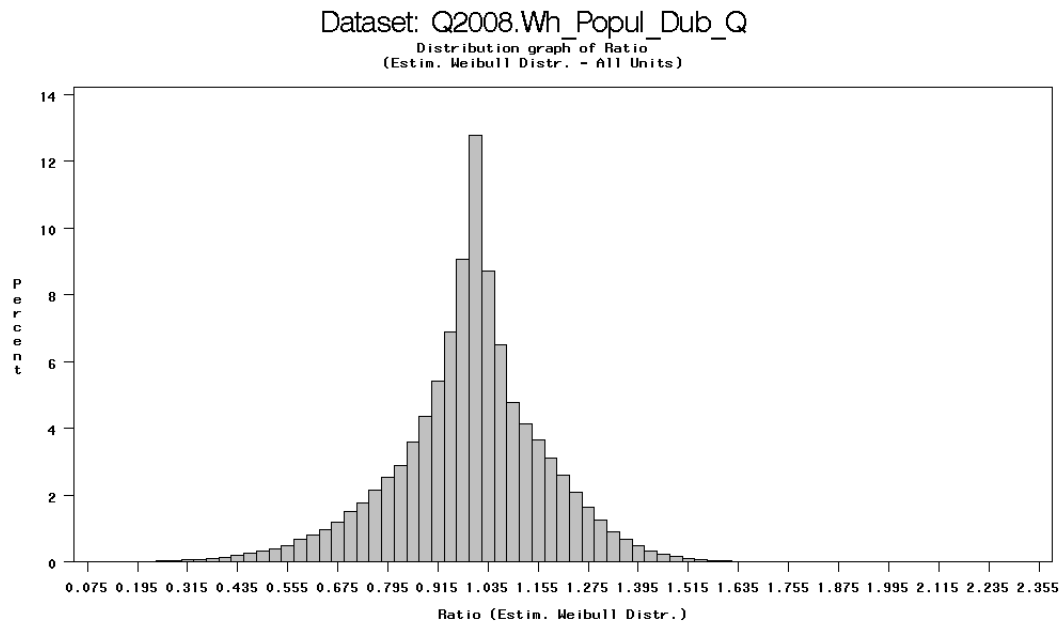
Then, using the Weibull parameters estimated in each sample cell with the distribution test<sup>62</sup>, the random numbers are converted into profit margin ratios that have a Weibull distribution. The following way is used:

---

<sup>61</sup> The same *Call RUNUNI* SAS routine seen in par. 3.3.1 is used. See SAS User's Manual (SAS Institute Inc., 2004) for further details.



**Graph 3.9** – Distribution of profit margin ratios generated with lognormal distribution.



**Graph 3.10** – Distribution of profit margin ratios generated with Weibull distribution.

<sup>62</sup>  $\sigma_j$  ( $\sigma_j > 0$  for each  $j$ ) is the estimated scale parameter of the  $j^{\text{th}}$  cell;  $c_j$  ( $c_j > 0$  for each  $j$ ) is the estimated shape parameter of the  $j^{\text{th}}$  cell.

$$\hat{r}_{ij}^W = \sigma_j \cdot \left[ -\log(1 - m_{3,ij}) \right]^{1/c}.$$

The Graph 3.10 shows the distribution of the profit margin ratios generated with a Weibull distribution is shown.

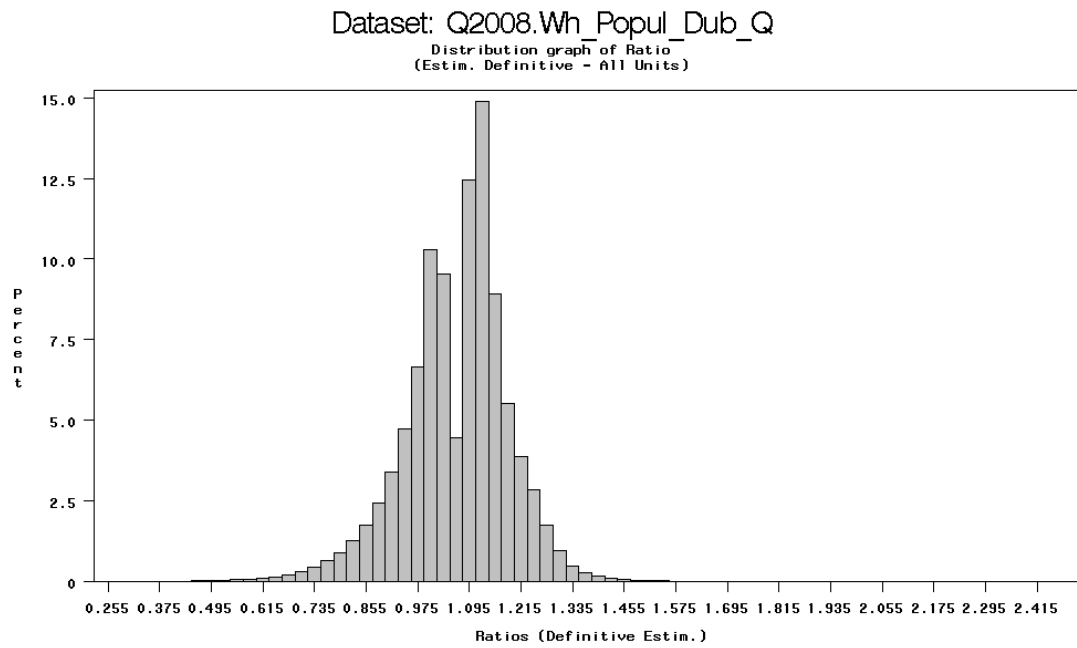
### 3.3.5 Mixed distribution

To generate the data of the new simulated population frame it is decided to do not attribute the same distribution at every cell (for example, the prevalent distribution in terms of number of cells or the one supposed to be more close to the real distribution of the profit margin ratios). It is instead attributed a different distribution to each one of the cells, considering the results of the tests used to outline the cells distribution, that were presented in par. 3.2.5 and 3.2.6. The new kind of global distribution of all the simulated ratios is called “mixed” distribution, because is the final results of the mixing of the three distributions presented previously (normal, lognormal or Weibull)<sup>63</sup>.

The final “mixed” distribution is shown in the Graph 3.11; the obtained distribution underlines the main influence of the normal and of the lognormal distribution, that gives the global effect of a bimodal distribution. The two peaks show the influence of both the lognormal (on the right) and the normal (on the left) distribution. The influence of the Weibull distribution is not relevant, seen the small number of cell with this kind of distribution (only 7.9% of the total).

---

<sup>63</sup> The proportion of the cells attributed to each of the three distributions is presented at the end of the par. 3.2.6.



**Graph 3.11** – Distribution of profit margin ratios – “mixed” distribution.





## **4 Relative efficiency of the methods of generation**

After the identification of the distribution that reproduces in the best way the observed sample, we want to test the relative efficiency of the other methods of generation against the “mixed” method. This means that we want to understand the loss of precision that we would obtain if we would use a one-way generation process, that is a process based on one of the tested distributions only: the normal, the lognormal or the Weibull.

The main goal of this phase is to understand if the simulated frame could be built in a more efficient way (that is considering a single distribution) rather than using a “mixed” selection process of distribution by cells.

In the following paragraph 4.1 we evaluate the bias of the single-way generation processes in comparison with the one based on the “mixed” distribution; moreover we evaluate the relative precision of the use of stratification (4.1.2.1) and we compare the results obtained with simple random sampling and with the systematic sampling (4.1.2.2).

Another interesting topic that will be discussed is the convergence to the parameter that we want to study in correspondence to different numbers of selected samples: the selection of a higher number of samples brings to more precise estimates? This topic is discussed in paragraph 4.1.2.6.

In the paragraph 4.2 we want to evaluate the bias of the estimates obtained with different kind of probability proportional to size sample selection methods.

### **4.1 Bias of the parameters**

The aim of this paragraph is to study the bias of the parameters. The bias is computed with reference to the values of the “mixed” distributed frame population on one hand and of the

wholesale observed sample on the other hand, comparing these to the values obtained selecting samples in different ways and using the different simulated populations (generated in chapter 3).

The objective of this first phase of the analysis is firstly to identify the more efficient sampling selection method: does stratified or not-stratified (see par. 4.1.2.1), simple random sampling or systematic sampling (par. 4.1.2.2) bring to more precise results? We also want to understand what is the gain (in terms of precision) that we obtain generating a population using a “mixed” distribution, rather than generating a population with an unique distribution (normal, lognormal, Weibull) for every cell. Staying in the same path, we will try to identify the method that brings to results closer to the “mixed” generated data (considered the best method of generation; see paragraph 4.1.2.3) or to the observed data (par. 4.1.2.4). The different kinds of units composing the frame and the selected wholesale sample can furthermore give some suggestions about the behaviour of the variable we are studying (the profit margin ratios) in the two groups (4.1.2.5).

Another goal of this first phase of the analysis is to understand what is the gain, in terms of precision, that we obtain using a smaller rather than a bigger sample size (4.1.2.6).

#### **4.1.1 Methodology**

To get the first results, a series of samples is selected from the frame population. Each one of them has a number of units equal to 1,000 ( $n = 1,000$ ).

With a bootstrap process a number of samples equal to 1, 2, 5, 10, 50, 200, 500, 1,000 are consequentially selected.

Each sample was selected using all of the following four sample selection methods: simple random sampling (*SRS*), stratified simple random sampling, systematic sampling, stratified systematic sampling.

1. *SRS (Simple Random Sampling)*: the  $n = 1,000$  units of the sample are selected without repetition. The probability of selection  $p$  of each unit is equal to the one of the others:

$$p = \frac{n}{N},$$

where  $n$  is the number of units of the sample (sample size) and  $N$  is the number of units in the population frame.

2. *Stratified SRS*: in each 4 digits NAICS group, 40 units are selected without repetition from the population frame. Each unit of a NAICS group has the same probability of selection of the other ones of the same group. This means that:

$$p_j = \frac{n_j}{N_j}$$

where  $j$  indicate the  $j^{\text{th}}$  group ( $j = 1, 2, \dots, 25$ ).

3. *Systematic sampling*: the units are firstly ordered by revenue classes, and then 1,000 units are selected with sampling interval (also called interval size) equal to  $k$ :

$$k = \frac{N}{n}.$$

The starting point of selection is randomly selected. The probability of selection for each unit is equal to:

$$p = \frac{1}{k} = \frac{n}{N}.$$

4. *Stratified systematic sampling*: the sample selection is stratified by 4 digits NAICS. From each NAICS group, 40 units are selected using a systematic method similar to the one used on all the population frame seen at point 3.; the starting point, in each NAICS, is randomly selected and the sampling interval is, for each  $j^{\text{th}}$  group ( $j = 1, 2, \dots, 25$ ):

$$k_j = \frac{N_j}{n_j}.$$

The probability of selection of each unit in the  $j^{\text{th}}$  group is equal to:

$$p_j = \frac{1}{k_j} = \frac{n_j}{N_j}.$$

On each selected sample the mean and the variance of the profit margin ratio variable are computed; on the average values of all the samples is then computed an un-weighted mean. Some results of this first phase of the analysis are shown in the following paragraph 4.1.2.

#### 4.1.2 Some results

The means of the samples' values, one for each kind of simulated population (generated with a normal, lognormal, Weibull and "mixed" distribution) is compared with the value computed on the entire simulated frame population, generated using the "mixed" distribution. This last average value will be defined, in the following, as the *frame reference parameter*.

The percentage differences from the frame reference parameter, are shown in Table 4.1. The second and the fourth columns of the table (called "Frame r. p.", that is "Frame reference parameter") show the percentage differences from the frame reference value, while the third and the fifth columns show the percentage differences from the reference parameter computed on the observed wholesale sample used to generate the population. This last parameter will be called *sample reference parameter*.

	ALL THE FRAME % average differences from...		ALL THE FRAME - Strata % average differences from...	
	Frame r. p.	Sample r. p.	Frame r. p.	Sample r. p.
<b>Simple Random Sampling</b>				
Mixed	-0.17	5.19	0.72	6.13
Normal	-6.00	-0.95	-5.48	-0.41
Lognormal	5.50	11.16	5.85	11.54
Weibull	-7.07	-2.07	-6.60	-1.58
<b>Systematic Sampling</b>				
Mixed	-0.12	5.25	0.75	6.16
Normal	-6.13	-1.08	-5.52	-0.44
Lognormal	5.56	11.23	5.87	11.56
Weibull	-7.19	-2.20	-6.47	-1.44

**Table 4.1** – Bias of the sampling averages from the reference parameters (frame and sample reference parameter).

Detailed results by sampling method are shown in Table 4.2, Table 4.3, Table 4.4, Table 4.5 and Table 4.6.

#### 4.1.2.1 Precision of not-stratified vs stratified selection methods

The first reading key of the Table 4.1 shows the differences in the sample selection methods' efficiency using a not-stratified rather than a stratified method of selection, summarized in Table 4.2. For this comparison the mixed distribution (considered to be the one closest to the real distribution) is used.

Sampling methods			
SRS	SRS – Stratified	Systematic	Systematic - Stratified
-0.17	+0.72	-0.12	+0.75

**Table 4.2** – Percentage bias from the frame reference parameter of the simple random sampling and of the systematic methods (not-stratified and stratified version).

Considering the average bias of the samples (selected from the “mixed” population) from the frame reference parameter, the use of stratification brings to worse results for the simple random sampling and for the systematic sampling too. In fact, there is an over-estimation, respectively, of 0.72% and of 0.75%, while, if we don't use the stratification, there is a lower bias: an under-estimation of the reference parameter of, respectively, -0.17% and -0.12%.

Coming back to the detailed results shown in Table 4.1, also considering sample selected from lognormally distributed population, the bias is higher (of about 0.3/0.35 percentage points) if we use the stratification. The results are not the same for the normal and the Weibull distributed populations: the use of the stratification brings to less distorted results (about 0.5/0.6 percentage point; see Table 4.1).

#### 4.1.2.2 Simple random sampling vs systematic sampling

A second key to read Table 4.1 is related to the differences between the results obtained with the simple random sampling rather than the results obtained with the systematic sampling selection method. The Table 4.3 shows that there are no relevant differences

between the results of the two selection methods. If, for example, we consider the percentage differences from the frame reference parameter, we obtain an average bias of -1.94% when we use a SRS method of selection; the bias is similar (-1.97%), when we use a systematic sampling method.

The same thing happens for the bias of the stratified samples' parameters: the difference is 0.04% only.

We find similar results taking a look to the (higher) bias of the samples' averages from the sample reference parameter.

The global average bias of samples selected with a SRS rather than a Systematic method are exactly the same (-2.62%).

This means that the use of one rather than the other selection method doesn't give any advantage in term of bias from the reference parameter.

	% Differences from...				
	Frame ref. parameter		Sample ref. parameter		
Sampling method:	Not stratified	Stratified	Not stratified	Stratified	
SRS	-1.94	-1.38	-3.31	-3.85	-2.62
Systematic S.	-1.97	-1.34	-3.27	-3.88	-2.62

**Table 4.3** – Average bias of the samples' averages from the reference parameters.

The first conclusion is that the kind of process used to generate the population is the factor that influences the bias of the computed parameter most. The use of the stratification can only marginally change the results in terms of bias, and the use of a systematic rather than of simple random sampling doesn't bring to a relevant advantages in terms of precision.

#### 4.1.2.3 *Relative bias of the uniformly distributed simulated populations*

The third key to read the Table 4.1 is comparing, for each kind of sampling selection method, the bias of the average values of the samples selected from one specific simulated population (normal, lognormal, Weibull), considering the reference parameters, versus the bias obtained with the selection from the "mixed" population. This comparison was decided, because the samples selected from a simulated population generated with a mixed

distribution are supposed to provide the closest means to the parameter of reference, as confirmed from the data in the Table 4.1.

The results of the Table 4.1 show that using a simulated population generated using one distribution only (normal, lognormal or Weibull) brings to a global bias from 5.48 to 7.19 percentage points.

The results, in terms of differences from the bias of the samples selected from the “mixed” population and for each kind of simulated distribution, are shown in the Table 4.4.

Simulated population's distributions	Sampling selection methods (differences in % points)				MEDIA
	SRS	SRS – Stratified	Systematic	Systematic – Stratified	
NORMAL	-5.83	-6.20	-6.01	-6.27	-6.08
LOGNORMAL	5.67	5.13	5.68	5.12	5.40
WEIBULL	-6.90	-7.32	-7.07	-7.21	-7.12

**Table 4.4** – Bias from the bias of the “mixed” parameters (computed on samples selected from the “mixed” population).

The simulated population that brings to less distorted results is the lognormal population: +5.40 percentage points (+5.67 if we use simple random sampling, +5.68 with the systematic sampling). If we use a stratified selection method, we obtain less overestimated results (respectively: +5.13/+5.12 percentage points).

The use of the normal rather than of the Weibull distribution brings to an additional under estimation of, respectively, -6.08 and -7.12 (in terms of percentage points differences from the parameter estimated with the mixed distribution). In particular, the bias, if we use the simulated population generated with the normal distribution only, gives an additive underestimation of about -5.83 (for SRS selection method) or -6.01 (with systematic method) percentage points, in comparison with the parameter estimated with the selection from a mixed distributed population. The bias is a little bit higher if we use the stratified methods: we get an additive bias of -0.37 with the SRS-stratified and of -0.20 with the systematic-stratified.

Also the samples selected from a Weibull simulated population make the underestimation stronger: about 7 percentage points for the sample selected without any stratification (SRS:

-6.90; Systematic: -7.07), and about 7.3 points for the sample selected using the stratification by cells (SRS-stratified: -7.32; systematic-stratified: -7.21).

All these results are useful to suggest to the researcher which one is the more efficient method to generate a simulated population, considering the longer process that is necessary to generate a mixed population rather than a population based on one distribution only.

#### 4.1.2.4 *Bias from the sample reference parameter*

Table 4.5 shows the gained precision of a certain sampling selection method versus the precision of the simulated population. This time the reference parameter is computed on the selected wholesale sample and not on the frame reference parameter.

	% differences from sample reference parameter				
	SRS		Systematic Sampling		AVERAGE
Simulated population	Not stratified	Stratified	Not stratified	Stratified	
Mixed	5.19	6.13	5.25	6.16	5.68
Normal	-0.95	-0.41	-1.08	-0.44	-0.72
Lognormal	11.16	11.54	11.23	11.56	11.37
Weibull	-2.07	-1.58	-2.20	-1.44	-1.82

**Table 4.5** – Percentage differences of the sampling averages from the sample reference parameter (wholesale selected sample).

Looking to the last column of Table 4.5 (the average bias for each simulated population) we notice that the two generated populations more close to the reference parameter computed on the sample's data are the Weibull (-1.82%) and, most of all, the normal distributed population (-0.72%): they both present an underestimation of the parameter.

If, for the Weibull distributed population, with a not-stratified sampling selection method the underestimation is more than 2 percentage points (-2.07% for the SRS, -2.20% for the systematic sampling), using a stratification selection method we get parameters more close to the sampling data: in fact we have more precise estimates of about 0.49 percentage points (for SRS) and of about 0.76 (for systematic sampling). The same happens for the normally distributed generated population: we get a higher precision of 0.54 (for SRS)



rather than 0.64 (for systematic sampling) percentage points using the stratification in the sampling selection methods.

On the other hand, the population generated with a lognormal distribution overestimates the sample reference parameter of 11.37 percentage points. The bias are quite similar considering all the kind of selection methods, but this time, using a stratified method of selection, we get a bias a little bit higher (0.38 percentage points for the SRS, 0.33 for the systematic sampling).

The higher bias of the parameters computed on the lognormally distributed population can be attributed to the specific conformation of the lognormal distribution used to generate the data. To have a confirmation of the asymmetry on the right side of the lognormal distribution in comparison with the normal distribution, see Graph 3.11; in this last graph, the right peak is the one of the lognormal generated ratios. The graph shows clearly the high influence of the lognormally distributed cells. In fact, the weight of the lognormal distribution in the definition of the “mixed” distributed population<sup>64</sup> brings to obtain overestimated average parameters selecting samples from the “mixed” distributed population too (an average overestimation of 5.68%). Furthermore, for the “mixed” distributed population the bias is higher if we use a stratified selecting method (+0.97% for SRS, +0.91% for the systematic sampling).

#### 4.1.2.5 *The biggest units effect*

It's also interesting to notice the different degree of bias that is obtained comparing the average parameters computed selecting samples from the “mixed” distributed population and the sample reference parameter (on one hand) and the frame reference parameter (on the other hand). These differences are shown in Table 4.6. The average bias of the samples' parameter from the frame reference parameter is lower (+0.29%) than the bias from the wholesale sample's reference parameter (+5.68%). The difference can be attributed to the different kind of units that we can find in the wholesale sample rather than in the frame population.

---

<sup>64</sup>The 72.9% of the cells has a lognormal distribution: see par. 3.2.6.

	% differences from reference parameter				
	SRS		Systematic Sampling		
Simulated population: MIXED	Not stratified	Stratified	Not stratified	Stratified	AVERAGE
Sample reference parameter	+5.19	+6.13	+5.25	+6.16	+5.68
Frame reference parameter	-0.17	+0.72	-0.12	+0.75	+0.29

**Table 4.6** – Percentage differences of the sampling averages from the sample and frame reference parameters.

In fact, the wholesale sample is selected from the frame population using a PPS (Probability Proportional to Size) sampling selection method<sup>65</sup>; this gives a bigger probability of selection to the biggest units of the frame (these are the units with a higher revenue). Seen that the proportional weight of this kind of units is bigger in the sample, we can attribute the higher overestimation of the reference parameter to the biggest units' effect. This means that the biggest units have probably a higher average level of profit margin ratios, so their margins should vary more than happens in a frame where the smaller units are proportionally more numerous.

#### 4.1.2.6 Precision's improvements with more samples

In the previous paragraph 4.1.1 we told that, to compute the shown results, we selected consequentially, using a bootstrap process, a number of samples equal to: 1, 2, 5, 10, 50, 200, 500 and 1,000 samples (every sample is made by 1,000 units). An average parameter for each of these groups of samples was computed (as the mean of all the samples' profit margin ratios). This was also done to evaluate the convergence to the reference parameters (the one of the frame population or the one computed on the selected wholesale sample) selecting a bigger number of samples.

In this paragraph we want to study the gain in terms of precision of the estimated parameters that we obtain using 1,000 samples rather than 1 sample only, selected with a certain method (Table 4.7 and Table 4.8).

The Table 4.7 shows the convergence to the average value of the reference parameter, computed on the entire simulated population (second and fourth columns) or on the

<sup>65</sup> For the selection of the sample, the cut-off Poisson selection method is used.

wholesale selected sample's preliminary data (third and fifth columns). The improvements are computed as differences and are expressed in percentage points.

Sampling Selection Methods / Population	No stratification		Stratified samples	
	Improvements in terms of % points differences from...		Improvements in terms of % points differences from...	
	Frame par.	Sample par.	Frame par.	Sample par.
<b>Simple Random Sampling</b>				
Mixed	0.55	0.58	-0.23	-0.24
Normal	0.03	0.03	-0.36	-0.38
Lognormal	0.17	0.18	0.07	0.07
Weibull	0.05	0.05	0.28	0.29
<b>AVERAGE (SRS)</b>	<b>0.20</b>	<b>0.21</b>	<b>-0.06</b>	<b>-0.07</b>
<b>Systematic Sampling</b>				
Mixed	0.32	0.34	-0.38	-0.40
Normal	0.50	0.52	-0.05	-0.05
Lognormal	0.04	0.04	0.15	0.15
Weibull	0.59	0.62	-0.43	-0.45
<b>AVERAGE (Systematic)</b>	<b>0.36</b>	<b>0.38</b>	<b>-0.18</b>	<b>-0.19</b>
<b>AVERAGE (GLOBAL)</b>	<b>0.28</b>	<b>0.30</b>	<b>-0.12</b>	<b>-0.13</b>

**Table 4.7** – Improvements (in percentage points) of the precision of the parameter obtained selecting 1,000 rather than 1 samples to compute the samples' mean.

This means, for example, that using a simple random sampling on a “mixed” population we obtain a higher precision (of 0.55 percentage points, compared to the frame reference parameter) computing the average value on 1,000 samples rather than on 1 sample only. The precision's improvement is of 0.58 points, if we are referring to the sample's reference parameter.

On the other side, if we select 1,000 samples, rather than 1 sample only, from a “mixed” population, we don't obtain any convergence: the bias is higher of 0.23 percentage point (-0.23), if we consider the frame's reference parameter, and of 0.24 percentage points (-0.24), if we consider the sample's reference parameter.

Taking a look to the average of the four simulated populations by sampling selection method (Table 4.7), we can notice that we get an improvement of about 0.3 percentage points in the precision of the parameter if we don't use the stratification; furthermore the

precision is improved of other about 0.15 percentage points if we use the systematic sampling rather than the SRS. At this regard, the stratified sampling selection methods have an opposite behaviour. In these cases we loose, in terms of precision, about 0.6 percentage points with the SRS, and about 0.18 percentage points with the stratified systematic sampling.

To test if the kind of sampled population influences the improvement that is possible to obtain going from 1 to 1,000 selected samples, we can observe the data shown in Table 4.8.

% differences from reference parameter	
<b>Population</b>	<b>Improvement</b>
Mixed	0.55
Normal	0.03
Lognormal	0.17
Weibull	0.05
<b>AVERAGE (SRS)</b>	<b>0.20</b>

**Table 4.8** – Improvements (in percentage points) of the precision of the average obtained using 1,000 rather than 1 selected samples by simulated population.

The maximum improvement is obtained with the “mixed” population (0.55 percentage points). Selecting 1,000 samples rather than 1 sample only from the normal and the Weibull population seems to bring to a very small improvement (respectively, 0.03 and 0.05 percentage points).

Generally speaking, in no one of the examined cases there seems to be a significant improvement in the precision (or convergence) of the estimated parameter if we select 1,000 samples, rather than 1 sample. This brings to the conclusion that selecting a higher number of samples, usually doesn’t allow to obtain better results. In some cases (using stratified sampling methods for the “mixed” and normal population, for example) we even obtain a little bit higher bias.

These data contribute to prove the reliability of the used sampling selection methods (SRS and systematic): in fact, seen that the convergence to the real value of the parameter is so

slow and small, we can conclude that one sample only is already representative of the target population and selecting a much more bigger number of samples doesn't affect in a remarkably way the precision of the results: the improvement goes from 0.03 to 0.62 max percentage points (if we don't use any stratification) and, in terms of lost precision, from 0.38 max to 0.05 min points (with stratified sampling selection methods)<sup>66</sup>.

## **4.2 Bias of the different PPS selection methods**

The aim of this second phase of the analysis is to understand which is the probability proportional to size sample selection method that would bring to better estimates, using the simulated populations generated in chapter 3.

The methodology is the same introduced in par. 4.1 (see par. 4.1.1): a bootstrap process was applied, with the selection of a number of samples equal to 1, 2, 5, 10, 50, 200, 500, 1,000 (each sample is made by 1,000 units) and the computation, for each of these samples' groups, of an average value of the simulated profit margin ratios.

To select these samples the different probability proportional to size sampling selecting methods are used (they will be briefly introduced in the par. 4.2.1). This was done because the original sample (the surveyed wholesale sample) was selected using the cut-off Poisson sampling, that is a probability proportional to size sampling method.

If the main aim of the previous paragraph 4.1 was the test of the generated population and the gain in terms of convergence using numerous group of samples, this paragraph has the objective to understand if any other method is considerable better than the one actually used to select the sample for the survey.

---

<sup>66</sup> See Table 4.7 for details.

#### 4.2.1 PPS sampling selecting methods

The Probability Proportional to Size (PPS) sampling methods used to select the samples of the groups seen above are: PPS, PPS-stratified, PPS systematic, PPS systematic-stratified, PPS sequential random sampling, SPS (Sequential Poisson Sampling). They'll be briefly introduced in the following paragraphs (4.2.1.1, 4.2.1.2, 4.2.1.3, 4.2.1.4, 4.2.1.5, 4.2.1.6).

##### 4.2.1.1 PPS (*Probability Proportional to Size*)

The sampling units are selected without repetition and with probability proportional to a size variable, that we identify with the revenue. This means that the units characterized by a bigger size have a bigger probability of being selected.

At the first step of the sample selection (as happens for the selection of the wholesale sample), the units with relative size bigger than  $1/n$  are classified as *TA* (*Take All*) units. The relative size of  $i^{\text{th}}$  unit is the ratio of the size of that unit (its revenue,  $rev_i$ ) on the total size of the units of the population (the total revenue of the frame population,  $\sum_i (rev_i)$ )<sup>67</sup>:

$$relative\ size(i) = \frac{rev_i}{\sum_i (rev_i)}.$$

After the first group of TA units is identified, the relative size of the remaining units are computed again to eventually identify further groups of TAs; this process is applied until no more TA units are found.

After the identification of all the TA units, a sample of a fixed size is selected between the remaining units, called *TS* (*Take Some*) units. If, for example, the sample size is  $k$  and the number of TA units is  $g$ , the number of units to select from the group of TS is  $m$  (where  $m = k - g$ ). The selection of  $m$  units in the TS group is done with probability proportional to size. This means that the bigger units (the ones whose have a bigger revenue) have a bigger probability of selection.

---

<sup>67</sup> For further details about the selection of TA units, see Cochran (1977) and Madow (1949).

This algorithm of selection was introduced by Hanurav (1967) and Vijayan (1968)<sup>68</sup>.

#### 4.2.1.2 PPS/Stratified

The sampling selection method is similar to the one explained in the par. 4.2.1.1. The only difference is that the units are firstly stratified by 4-digits NAICS code. This means that the iterative identification of the TA and the TS units is done inside the single cell (identified by NAICS). In this way, the relative size of the  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell is computed in the following way:

$$relative\ size(i_j) = \frac{rev_{ij}}{\sum_i (rev_{ij})},$$

where:  $rev_{ij}$  is the revenue of the  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell and  $\sum_i (rev_{ij})$  is the sum of all the revenues of the units belonging to the  $j^{\text{th}}$  cell.

The  $i^{\text{th}}$  unit of the  $j^{\text{th}}$  cell is considered a TA unit, if its relative size is bigger than  $(1/n_j)$ , where  $n_j$  is the sample size of the  $j^{\text{th}}$  cell.

For this sampling selection method, each of the 25 cells has a constant sample size equal to 40 units (1,000 units of the global sample divided by 25 cells).

#### 4.2.1.3 PPS – Systematic

The probability proportional to size sampling method selects randomly the ordered units<sup>69</sup> with a fixed sampling interval. The probability of selection is proportional to the size of each unit (that is to its revenue). This method works in a way similar to the systematic sampling, but, depending on the size, each unit can be selected more than once. The units selected more than once are classified as TA units. So, also in this case the selection of the sample is divided into two different phases: the first one is finalized to the recursive selection of the TA units. The second phase selects with a PPS systematic method, between

---

<sup>68</sup> For further details, see also Fox (1989), Golmant (1990), Watts (1991) and the SAS' User Manual (SAS Institute, Inc., 2004).

<sup>69</sup> The units are ordered by the revenue variable, using the *CONTROL* option of the *SURVEYSELECT* SAS procedure.

the TS units' group, a number of unit equal to the final sample size, minus the number of TA units already identified, that are included by default in the sample.

#### 4.2.1.4 PPS – Systematic/Stratified

This method works similarly to the previous one (par. 4.2.1.3), but the identification of the TA units (the ones identified more than once) and the PPS systematic selection of the remaining units within the TS units is done considering a stratification by cells (that is by 4-digits NAICS). The sample size for each cell is 40 units.

#### 4.2.1.5 PPS – Sequential random sampling

Ordering the units by revenue classes, we used a probabilistic version of the Chromy selection method (1980)<sup>70</sup>. This method selects randomly the first unit with probability proportional to size; this selection is done within all the units of the population or within the strata's units, if we are using a stratified version of the method. Starting from this point (inside the whole population or inside a certain strata), all the remaining ordered units are numbered. The population (or each stratum) is treated as it was a loop, so that the selection, once the units are finished, can start again from the beginning of the same group (the population or the stratum). The selection is with probability proportional to size, sequential and with minimum replacement of the sampling units<sup>71</sup>. Similarly to the previous presented methods (4.2.1.3 and 4.2.1.4), one unit can be selected (within the entire population or within a certain strata) more than once (this means that the corresponding number of hits could be bigger than 1). The units with a number of hits bigger than 1 are considered TA units, while the other units are considered TS.<sup>72</sup>

---

<sup>70</sup> See Chromy (1978).

<sup>71</sup> "Selection with minimum replacement means that the actual number of hits for a unit can equal the integer part of the expected number of hits for that unit, or the next largest integer. This can be compared to selection without replacement, where each unit can be selected only once, so the number of hits can equal 0 or one. The other alternative is selection with replacement, where there is no restriction on the number of hits for each unit, so the number of hits can equal 0, 1, ...,  $n_h$ , where  $n_h$  is the  $h^{\text{th}}$  stratum sample size" – Source: SAS' User Manual (SAS Institute Inc., 2004).

<sup>72</sup> For further details about Chromy selection method, see Chromy (1979), Williams and Chromy (1980) and the SAS' User Manual (SAS Institute Inc., 2004).



#### 4.2.1.6 SPS (Sequential Poisson Sampling)

The Sequential Poisson Sampling method (SPS) belongs to the probability proportional to size methods' family: the bigger units<sup>73</sup> have a bigger probability of being selected.

The sequential Poisson sampling is a version of the Poisson sampling method<sup>74</sup>, and the main difference (that is also one of its advantages) is that the sample size is not a random value.

We will now explain briefly how the SPS method works.

Using an auxiliary dimensional variable (the revenue) the probabilities of inclusion can be computed, for each unit  $i$  ( $i = 1, 2, \dots, N$ ), in the following way:

$$\pi_i = n \cdot \frac{x_i}{X},$$

where  $\pi_i$  is the probability of inclusion,  $n$  is the sample size,  $x_i$  is the size variable for the  $i^{\text{th}}$  unit and  $X$  is the sum of the size variable among all the  $N$  units of the population, that is:

$$X = \sum_{i=1}^N x_i.$$

Usually some units have a probability of selection bigger than 1 ( $\pi_i > 1$ ). These units are considered as TA. The TAs are the units considered more representative of the population and they're inserted in the sample by default.

After the first selection of TA units, the probabilities of selection are computed again among the remaining units to check if other TA units are found. This identification is recursively done until no more TAs are selected in the population.

When no more TA units are selected, we get a definitive group of TA and another group made of TS units.

The sample's selection process goes on with the generation of a random number for each  $i^{\text{th}}$  unit:

$$RN_i.$$

This number has uniform distribution in the interval 0-1 and is used to compute a transformed value ( $\zeta_i$ ), for each  $i^{\text{th}}$  unit:

---

<sup>73</sup> Also in this case the size variable is considered the revenue.

<sup>74</sup> For more details about the Poisson method, see Särndal et al. (1992).

$$\xi_i = \frac{RN_i}{\pi_i}.$$

The  $i^{\text{th}}$  unit is inserted in the sample if the following condition is verified:

$$\xi_i \leq n.$$

This means that the  $n$  units with smaller  $\xi_i$  values are inserted in the sample.<sup>75</sup>

The sequential Poisson sampling selection method is considered very important in our analysis, because is the one used to select the units for the wholesale survey. So the results of the application of this method of selection will be compared to the ones obtained with the other probability proportional to size methods to outline the relative efficiency of the different selection methods.

#### 4.2.2 Results

In the Table 4.9 there are the percentage differences of the average computed on the groups of samples (selected from the different simulated populations) from the reference parameters. In the second and in the fourth column there are the percentage differences of the averages from the frame reference parameter (computed on the whole frame population), while in the third and fifth column the same differences are computed from the sample reference parameter (obtained from the wholesale observed sample's data). The results are split by sampling selection method (all methods are probability proportional to size). In the second and in the third column the results are referred to selection methods applied without stratification, while in the last two columns the sampling selection was made using a stratification by 4 digits NAICS.

If we consider the differences of the applied methods, we can notice that, at an average level, the stratification of the units of the frame population brings to better results only when we use the PPS-Systematic sampling method: 1.51/6.97% of average bias with the stratification versus 1.55/7.01% obtained without using strata. In the PPS and in the PPS-

---

<sup>75</sup> For further details about the sequential Poisson sampling method, see Särndal et al. (1992) and Ohlsson (1990 and 1998).

	No stratification		Stratification	
	% average differences from...		% average differences from...	
	Frame par.	Sample par.	Frame par.	Sample par.
<b>PPS (Probability Proportional to Size)</b>				
Mixed	6.91	12.65	7.13	12.89
Normal	-5.15	-0.06	-5.03	0.07
Lognormal	10.10	16.02	10.22	16.14
Weibull	-5.77	-0.71	-5.55	-0.48
<b>AVERAGE</b>	<b>1.52</b>	<b>6.98</b>	<b>1.69</b>	<b>7.16</b>
<b>PPS – Systematic</b>				
Mixed	7.02	12.77	6.84	12.58
Normal	-5.00	0.10	-5.06	0.04
Lognormal	10.07	15.99	10.02	15.93
Weibull	-5.88	-0.83	-5.75	-0.69
<b>AVERAGE</b>	<b>1.55</b>	<b>7.01</b>	<b>1.51</b>	<b>6.97</b>
<b>PPS – Sequential</b>				
Mixed	6.90	12.64	6.90	12.64
Normal	-5.04	0.06	-4.91	0.20
Lognormal	10.01	15.92	10.06	15.98
Weibull	-5.86	-0.80	-5.74	-0.67
<b>AVERAGE</b>	<b>1.50</b>	<b>6.96</b>	<b>1.58</b>	<b>7.04</b>
<b>SPS (Sequential Poisson Sampling)</b>				
Mixed	7.61	13.39		
Normal	-4.73	0.38		
Lognormal	10.75	16.70		
Weibull	-6.05	-1.00		
<b>AVERAGE</b>	<b>1.89</b>	<b>7.37</b>		

**Table 4.9** – Average percentage differences from the frame and the sample reference parameters by sampling methods and by used population.

	% average differences from frame's parameter		% average differences from sample's parameter		Improvements (stratific. – no stratific.)	
	No stratification	Stratification	No stratification	Stratification	Frame ref. parameter	Sample ref. parameter
Mixed	6.94	6.95	12.70	12.69	0.01	0.01
Normal	-5.06	-5.00	0.10	0.04	0.06	0.07
Lognormal	10.06	10.10	16.02	15.98	0.04	0.04
Weibull	-5.84	-5.68	-0.61	-0.78	0.16	0.17
<b>AVERAGE</b>	<b>1.53</b>	<b>1.59</b>	<b>7.05</b>	<b>6.98</b>	<b>0.07</b>	<b>0.07</b>

**Table 4.10** – Average percentage differences from the frame and the sample reference parameters by used population (all probability proportional to size methods).

Sequential method the average bias is bigger if we use strata: respectively 1.69/7.16% versus 1.52/6.98% for PPS and 1.58/7.04% versus 1.50/6.96% for PPS-Sequential. In all the cases the precision we obtained using or not the stratification is quite similar: the differences, in percentage terms, are small. This conclusion is confirmed from the following Table 4.10, where the averages differences by kind of population are shown. In the last two columns of the table we can see the percentage improvements we can obtain using the stratification in the different selection methods by kind of population. The average improvement is 0.07%: the maximum one is obtainable with the Weibull population (0.16/0.17%), the minimum with “mixed” population (0.01/0.01%).

Coming back to Table 4.9, other considerations can be done considering the subdivision of the table, to evaluate the relative degree of precision of the different sampling selection methods. At an average level, comparing the averages with the sample and the frame reference parameters, we can notice an overestimation of about 1.5% (1.52 for PPS, 1.55 for PPS-Sys., 1.50 for PPS-Seq., 1.51 for the stratified PPS-Sys., 1.58 for the stratified PPS-Seq.). The over estimation is bigger for the stratified PPS (1.69%), but the biggest one is obtained with the SPS method (1.89%).

This could make us believe that the SPS sampling method is the one that has less precision. Nevertheless, the main objective of the survey is to study the biggest units of the frame population, the ones whose have the biggest influence on the evolutions of the prices (and of the profit margin) in a certain market. We already told about the conclusion that the biggest units are probably characterized by biggest profit margin ratios<sup>76</sup>. The 1.89% average overestimation (the biggest one) of the frame’s reference parameter means that, thanks to the SPS method, we can select the part of the target population more interesting for our research, that is the group of biggest units.

Looking to the more detailed data of Table 4.9, we notice that the precision of the sampling methods varies a lot basing on the kind of simulated population used for the selection of the sample. In fact, the selection of samples from a simulated population generated with normal distribution brings to an average bias of about -5% (from -4.91 min, excluding the SPS’ -

---

<sup>76</sup> See paragraph 4.1.2.5.

4.73, to -5.15 max). A little bit higher (for all the sampling selection methods) is the underestimation of the samples selected from a simulated Weibull population (from -5.55 to -5.88, the SPS' -6.05 excluded). Extracting samples from the lognormal generated population, we obtain an average overestimation of about 10% (10.75% for the SPS). The mixed distribution is conditioned from the heavy presence of cells with a lognormal distribution, and for this reason there is an average overestimation of about 6.9/7% (7.61% for the SPS).

All this can bring to the first general conclusion: the kind of probability proportional to size sample selection method chosen to select the sample doesn't affect much the precision of the parameter object of study. A similar situation it's shown by the results of sample selection methods applied using or not the stratification of sampling units. The stratification doesn't affect considerably the precision of the estimates. The factor that seems to affect the precision of the results most is the kind of procedure used to generate the simulated population.

Generally speaking (Table 4.9) the smaller percentage bias is obtained extracting samples from the normal or from the Weibull population (from -5% to -6%), while using a "mixed" population the bias is positive (overestimation of the parameter of about 6/7%); the maximum bias is the one obtained with lognormally distributed population (more than 10%).

The results seen above, in terms of bias, are different if we consider the differences from the wholesale sample's reference parameter shown in the same Table 4.9 (third and fifth column). The precision is very high for samples selected from the normally distributed generated population: the maximum bias is the one obtained with the SPS method (+0.38%) and the maximum precision is obtained with the stratified PPS-Systematic method (overestimation of +0.04% only). Extracting samples from the Weibull distributed population makes the bias grow up a little bit, with a minimum underestimation of -0.48% (obtained with the PPS stratified method) and a maximum of -1% (SPS). Higher bias is obtained considering the "mixed" distribution, that is between +12.58% and +12.89%, with

a peak corresponding to the SPS selection (+ 13.39%). These highly distorted averages are extremely conditioned by the high bias obtained from the simulated population generated with a lognormal distribution (an overestimation of about 16% of the wholesale sample's parameter).

This high bias of the lognormal simulated population can be seen as the cumulative effect of the asymmetrical distribution we are considering<sup>77</sup> and of the proportional to size selection methods that we used<sup>78</sup>.

The wholesale sample used to generate the simulated populations was selected with a probability proportional to size selection method. Seen this, the samples selected with the PPS methods should be closer to the sample's than to the frame's reference parameter. But this doesn't happen, apparently, if we take a look to the averages of the four populations for each method. The reason is that, in the comparison with the frame reference parameter, the underestimation (about -5%, -6%) obtained from the normal and Weibull distribution is useful to compensate the overestimation obtained with the lognormal (about +10%) and with "mixed" (about +7%) populations. This doesn't happen if the comparison is made with the sample reference parameter: the normal and Weibull populations have a small bias (between -1% and +0.38%) that is not enough to compensate the highly distorted results of the "mixed" and lognormal populations (respectively, about 12.6% and 16%).

The previous results by sampling methods and by reference parameters are summed up in the Table 4.11.

This table shows the average differences by kind of population; this, as we already told, is the factor that influences most the bias of the results. In the table there are the averages of all the differences (by PPS sampling methods and with or without stratification), considering both the differences from the sample's and from the frame's reference parameters.

---

<sup>77</sup> The lognormal distribution (the prevalent one and the most important in the definition of the "mixed" distribution) is asymmetric on the right, in comparison with the normal distribution (symmetrical).

<sup>78</sup> We already told that the biggest units (the ones whose have a bigger probability of selection) are characterized by bigger profit margins ratios.

<b>Distribution of simulated population</b>	<b>% average differences from reference parameters</b>
Mixed	9.82
Normal	-2.48
Lognormal	13.04
Weibull	-3.23
<b>AVERAGE</b>	<b>4.29</b>

**Table 4.11** – Average percentage differences from the reference parameters by population (all probability proportional to size methods, with and without stratification, compared to both the reference parameters).

### 4.3 Some preliminary conclusion about the process of generation

In this paragraph we will discuss schematically some evidence underlined by the results shown in the previous paragraphs.

About the simulated population process, the main part of the efficiency (in term of discrepancy between the average level of the ratios and the reference parameters) is due to the process used to generate the population (that is the kind of distribution attributed to the cell of the simulated population).

This means that, if we generate a population with the normal or with the Weibull distribution for each cell, we usually get an underestimation of the level of the profit margin ratios. On the contrary, if we use a population of simulated profit margin ratios generated as they have a lognormal distribution in all the cells, we obtain an overestimation of the level of the variable. The heavy presence of the lognormal distribution in the generation process used for the “mixed” distribution, also affects these results bringing to have an overestimation of the level of profit margin ratios.

The overestimation obtained with the “mixed” and with the lognormal populations is due to the asymmetry on the right of the lognormal distribution; but it is also marked when we use a probability proportional to size sampling selection method.

The results show also which is the generated population that gives estimates more close to the reference parameter we are interested: if we consider the wholesale sample’s reference

parameter, the best precision is obtainable selecting samples from the normal distributed simulated population (this was supposed from the way the population was generated). Using another kind of population, uniformly distributed across the different cells, brings to bias higher of almost 5%. Nevertheless, if we are interested to have results more close to the frame's reference parameter, the results suggest the selection of samples from the "mixed" simulated population.

The first part of the analysis also shows that the effect of selecting bigger units (that is: the use of a probability proportional to size method of selection) brings to higher average level of the profit margin ratios.

Other conclusions are more referable to the way we can use to improve the quality of the results (that is the precision of the computed averages). Selecting 1,000 rather than small quantity of samples (e.g. 1, 2, 5, 10 samples) from the simulated population doesn't bring to have a significantly better precision in the estimates.

Moreover, if we refer to the random sampling methods (SRS and Systematic sampling) the stratification of the units by NAICS doesn't brings to significantly better results. And the Systematic sampling gives a precision quite close to the one obtained with the SRS.

If we consider the probability proportional to size sampling selection methods, the Sequential Poisson Sampling it the one that brings to a more distorted estimates, comparing it to the other methods of the same kind. This overestimation is particularly high considering the bias of the estimates from the wholesale sample's reference parameter. This is probably caused by the absence (and the proportionally less weight) of the smaller units of the frame. In fact, the bigger units seem to be characterized, at a general level, by higher profit margin ratios.

The other PPS selection methods (PPS, PPS-Systematic, PPS-Sequential) have a similar bias from the value of the reference parameters, but the highest precision is the one registered using the PPS-Sequential, without using the stratification by NAICS.



## **5 From profit margin ratios to price indexes**

The aim of this chapter is to explain the computation of the price indexes (at a various level of aggregation) from the original data (the observed prices). Usually the process used by most of the statistical agencies starts computing an elementary version of the price indexes (using the prices' data); afterwards, the process computes weighted averages using, as weights, a size variable (for example the establishment revenue). This brings to more aggregate level of indexes for groups of units classified according to some criteria: trade group, sector, national level.

The process will be synthetically explained step by step from the micro elementary index level (par. 5.1) to a more aggregate level (par. 5.2); the economic and the sampling weights (par. 5.2.1 and 5.2.2) will be also introduced.

### **5.1 Micro elementary indexes**

The first step to obtain the higher level price index is starting from the computation of a micro elementary index, that, in our case, is an index computed for each establishment of the selected sample. An establishment is a first stadium unit and for each establishment the prices of three different representative products (secondary stadium units) were collected.

The first phase to compute the elementary micro-index is the computation of a synthesis of the observed prices by establishment. By computing a simple geometric unweighted mean of the primary stadium units (the three observed prices of the selected products) we obtain an index called *Jevon's index*:

$${}_b^i I_t^J = \left( \frac{\prod_{g=1}^k {}^i p_t^g}{\prod_{g=1}^k {}^i p_b^g} \right)^{\frac{1}{k}} = \left( \prod_{g=1}^k \frac{{}^i p_t^g}{{}^i p_b^g} \right)^{\frac{1}{k}} = \left( \prod_{g=1}^k {}_b^i P_t^g \right)^{\frac{1}{k}}; \quad \text{Jevon's Index} \quad 5.1$$

In the formula 5.1:

- $g$  is the  $g^{\text{th}}$  product of a certain establishment  $i$  ( $g = 1, 2, 3$ ),
- $t$  is the actual time,
- ${}^i p_t^g$  is the price of the product  $g$  of the establishment  $i$  observed at time  $t$ ,
- $b$  is the base time of reference,
- ${}^i p_b^g$  is the price of the product  $g$  of the establishment  $i$  observed at time  $b$ ,
- $k = 3$  is the number of selected products for each establishment.

From the formula 5.1 it's easy to understand that we can compute (obtaining the same results) the unweighted geometric mean of the ratios of the observed prices in the time  $t$  and  $b$  or the unweighted geometric mean of the simple index computed with the prices of a single product  $g$  ( ${}_b P_t^g$ ):

$${}_b P_t^g = \frac{p_t^g}{p_b^g}.$$

The Jevon's index is introduced as a method to compute elementary price indexes by Producer Price Index Manual (International Monetary Fund, 2004).

Other ways to compute elementary price indexes are, for example, the Carli's index and the Dutot's index.

Carli's index is computed as the unweighted average of the products' simple index of a fixed establishment  $i$ :

$${}_b^i I_t^C = \frac{1}{k} \sum_{g=1}^k \left( \frac{{}^i p_t^g}{{}^i p_b^g} \right); \quad ^{79}$$

---

<sup>79</sup> The definitions of all the terms used in the formula are the same of formula 5.1 explained above (the Jevon's index). Basing on an economic approach, a discussion about the choice of Carli's index rather than of

Dutot's index is computed as the ratio of the arithmetic means of the prices in the two considered times:

$${}^i I_t^D = \frac{\frac{1}{k} \sum_{g=1}^k {}^i p_t^g}{\frac{1}{k} \sum_{g=1}^k {}^i p_b^g} .^{80}$$

The Jevon's index is one of the best index, from an axiomatic viewpoint; it is better, for example, than the Carli's and the Dutot's indexes, because it has some properties that these last indexes don't have (for details, see par. B.2.1 of Producer Price Index Manual, p. 218-219<sup>81</sup>). Furthermore the Jevon's index is supposed to give good estimates of the ideal index when the revenue's shares of the considered establishments remain constant from time to time (Producer Price Index Manual, par. B.2.2, p. 219-221<sup>82</sup>).

Nevertheless, usually the choice of the best index should be done considering the kind of products used to compute the elementary aggregates.

For further information about the Jevons index and its properties and about the computation of other elementary micro-index, see also: the Boskin Report (Advisory Commission to Study the Consumer Price Index, 1996), the Product Price Index Manual (International Monetary Fund, 2004) and Patak and Rais (2005).

The second phase to compute a higher-level price index is the aggregation of the elementary indexes (like, for examples, indexes computed with the Jevon's formula), to obtain a monthly index by NAICS and, subsequently, to compute the whole business sector's index (the wholesalers for services) at a national level.

---

Jevon's index in different contexts could be found in the Producer Price Index Manual (International Monetary Fund, 2004, p. 219-221).

<sup>80</sup> The definitions of all the terms used in the formula are the same of the Jevon's index formula (5.1) explained above.

<sup>81</sup> International Monetary Fund (2004).

<sup>82</sup> International Monetary Fund (2004).

## 5.2 Aggregate index

An aggregate index, also called “higher-level index” is usually built starting from an aggregation of elementary indexes (or a “lower-level index”) using some kind of weighting variable (usually the revenue). In fact, this is the definition of an aggregate index of the “Producer Price Index Manual” (par. C2, p. 230): “*The higher-level indexes are calculated simply as weighted arithmetic averages of the elementary price indexes*”<sup>83</sup>.

Once again, the system of weights is considered the revenue of a specific establishment (or of a specific group of establishment). It’s clear that these kind of relative weights change from time to time. So, sometimes it’s necessary to update the weighting system. Some countries’ statistical agencies update the weights’ system every year (to estimate as close as possible the product’s evolutions and the market’s shares), and some other don’t change them for many years, but the “Producer Price Index Manual”<sup>84</sup> suggest to update the weights every five years only.

The computation of the aggregate *SPPI* (*Services Producer Price Index*) is based on the Laspeyres’ methodology, using two different kinds of weights to aggregate the elementary indexes: the economic weights and the sampling weights.

### 5.2.1 Economic weights

If we use  $i$  to indicate the  $i^{\text{th}}$  statistical unit ( $i = 1, 2, \dots, n$ ), and  ${}_t z_{ih}$  to indicate the economic weights of the  $i^{\text{th}}$  unit of the  $h^{\text{th}}$  stratum at the time  $t$ , the formula used to compute the economic weights for the unit  $i$  at the time  $t$  is:

$${}_t z_{ih} = \frac{{}_{t-1} x_{ih}}{{}_{t-1} X_h},$$

where:

---

<sup>83</sup> International Monetary Fund (2004).

<sup>84</sup> International Monetary Fund (2004).

- ${}_{t-1}x_{ih}$  is the annual revenue of the  $i^{\text{th}}$  establishment of the stratum  $h$  in the previous year ( $t-1$ );
- ${}_{t-1}X_h$  is the total revenue generated by all the surveyed establishments of the  $h$  stratum in the previous year ( $t-1$ ).<sup>85</sup>

This is a generally valid formula: the annual revenue of the establishment (the revenue variable “ $x$ ” in the formula) is the system of weights often used by statistical agencies to evaluate the weight of the statistical units computing the Service Producer Price Index (SPPI). Usually this kind of weight could be referred to the market’s shares of the current year ( $t$ ). However, sometimes, for some SPPI components, the economic weights of the present year are not available; in these cases (as in the case shown from the formula) the economic weights referred to a previous time ( $t-1$ ) are used, assuming them as correlated to economic weights of the current time  $t$ .

Using the economic weights for the computation of the higher-level index means that the bigger units have weights proportional to their size (where the size is indicated by the revenue of the unit).

### 5.2.2 Sampling weights

The sampling weights, computed for each establishment, are the inverse of the probability of selection ( $\pi$ ):

$$w_{ih} = \frac{1}{\pi_{ih}} = \frac{{}_{t-1}X_h}{n_h \cdot {}_{t-1}x_{ih}}.$$

In the previous formula:

- $w_{ih}$  is the sampling weight of the establishment  $i$  of stratum  $h$ ,
- $n_h$  is the number of unit of the  $h^{\text{th}}$  stratum,
- $\pi_{ih}$  is the probability of inclusion of the unit  $i$  of the stratum  $h$ ;

---

<sup>85</sup> In our case, for the 2006 wholesale survey, the 2005 revenue data are used.

- ${}_{t-1}X_h$  is the total revenue obtained in the previous year,  $t-1$  (in our case we are considering the 2005's data), by all the establishment of the  $h^{\text{th}}$  stratum.
- ${}_{t-1}x_{ih}$  is the annual revenue of the establishment  $i$  of the  $h^{\text{th}}$  stratum in the previous year  $t-1$  (2005);

In the wholesale survey the probability of inclusion is given by the Poisson's sampling selection method and gives bigger weight to the bigger units of the sample.

The effect of the sampling weights compensates the economic weight's effect; this last one would give a proportionally bigger importance to the smaller units.

### 5.2.3 Computation of aggregate indexes

Once the Jevon's index  $({}_b^i I_t^J)$  or any other kind of elementary index is computed, using the economic weights  $(w_{ih})$ <sup>86</sup> and the sampling weights referred to a certain year  $t$   $({}_t z_{ih})$ <sup>87</sup>, we can obtain the aggregate index for the stratum  $h$ <sup>88</sup> as shown in the following formula 5.2:

$${}_b \hat{I}_t^h = \frac{\sum_{i=1}^{n_h} w_{ih} \cdot {}_t z_{ih} \cdot {}_b^i I_t^J}{\sum_{i=1}^{n_h} w_{ih} \cdot {}_t z_{ih}}. \quad 5.2$$

This weighted average is a micro-index for the  $h^{\text{th}}$  stratum.

In the formula:

- $n_h$  is the number of units included in stratum  $h$ ,
- $w_{ih}$  is the sampling weight of the establishment  $i$  of stratum  $h$ ,
- ${}_t z_{ih}$  is the economic weight of the  $i^{\text{th}}$  unit at the time  $t$  (computed using a size variable, that is the revenue, referred to the previous year),
- ${}_b^i I_t^J$  is the elementary index of the establishment  $i$  (introduced in par. 5.1).

<sup>86</sup> The economic weights were introduced in par. 5.2.1.

<sup>87</sup> The sampling weights were introduced (using the revenue of the previous year,  $x_{t-1}$ ) in par. 5.2.2.

<sup>88</sup> A stratum is usually referred to a specific NAICS.

If we write in a more extended way the product of the weights seen in the previous formula of the micro index, we obtain:

$$w_{ih} \cdot z_{ih} = \frac{_{t-1}X_h}{n_h \cdot _{t-1}x_{ih}} \cdot \frac{_{t-1}x_{ih}}{_{t-1}X_h} = \frac{1}{n_h}.$$

So we can conclude that usually the economic weights and the sampling weights simplify, if they both are function of the same variable (the revenue, in this case). Nevertheless, this simplification is not always possible. In our research, for example, there is a further adjustment of the sampling weights due to the missing values, so the simplification is not feasible.

When there are no adjustments and it's possible to simplify the weights in the way shown above, the index of the  $h^{\text{th}}$  stratum is a simple (or unweighted) arithmetic average of the elementary indexes  ${}_b^i P_t$ . Already mentioned in par. 5.1, this index, is called Carli's index  $\left({}_b^C \hat{I}_t^h\right)$ .

$${}_b^C \hat{I}_t^h = \frac{\sum_{i=1}^{n_h} {}_b^i P_t}{\sum_{i=1}^{n_h} 1} = \frac{\sum_{i=1}^{n_h} {}_b^i P_t}{n_h} \quad \text{Carli's index}^{89} \quad 5.3$$

In the Carli's formula:

- ${}_b^i P_t$  is the prices' elementary index of the establishment  $i$  in the time  $t$  (it compares the prices in time  $t$  to prices in the base time  $b$ ); this index can be, for example, an elementary Jevon's index  $\left({}_b^i I_t^J\right)$ .
- $n_h$  is the number of the establishments in the stratum  $h$ .

So, if we substitute the general elementary index's formula, we obtain:

---

<sup>89</sup> The index is introduced as one of the possible formula reliable to compute elementary price indexes by the Producer Price Index Manual (International Monetary Fund, 2004).

$${}_b^C \hat{I}_t^h = \frac{\sum_{i=1}^{n_h} {}_b^i P_t}{n_h} = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} \frac{{}_b^i P_t}{{}_b^i P_b},$$

where:

- ${}_b^i P_t$  is the price of a general product of the establishment  $i$  in the time  $t$ ;
- ${}_b^i P_b$  is the price of a general product of the establishment  $i$  in the base time  $b$ ;
- $n_h$  is the number of the establishments in the stratum  $h$ ;

Seen this, for the computation of Carli's index we don't need to use economic weights.

#### 5.2.3.1 More aggregate indexes

The general indexes of all the considered strata  $h$  seen above  $({}_b \hat{I}_t^h)^{90}$  can be aggregated at a more general level.

If we use the economic weights of the strata, we can compute, for example, the general index of the trade group ( $tg$ ) or of the economic sector.

The economic weight for the  $h^{\text{th}}$  stratum is:

$$z_h = \frac{\sum_{i=1}^{n_h} x_{ih}}{\sum_h X_h} = \frac{X_h}{X_{tg}},$$

where:

- $X_h = \sum_{i=1}^{n_h} x_{ih}$  is the sum of the revenues in the  $h^{\text{th}}$  stratum<sup>91</sup>;
- $n_h$  is the number of units in the  $h^{\text{th}}$  strata;
- $\sum_h X_h = X_{tg}$  is the total revenue of all the considered strata.

The general index of the trade group ( $tg$ )<sup>92</sup> is computed using the following formula:

---

<sup>90</sup>  $h = 1, 2, \dots, 25$  (we remember that we are using 25 NAICS codes in our research).

<sup>91</sup>  $i = 1, 2, \dots, 25$ .



$${}_b\hat{I}_t^{tg} = \frac{\sum_h z_h {}_bI_t^h}{\sum_h z_h},$$

where:

- $tg$  is the considered trade group,
- $z_h$  is the economic weight of the  $h^{\text{th}}$  stratum of the  $tg$  trade group,
- ${}_b\hat{I}_t^h$  is the aggregate micro index for stratum  $h$  seen in previous formula 5.2,
- $h = 1, 2, \dots, m$  (where  $m$  is the number of the strata in the considered trade group  $tg$ ).

Using a similar way of aggregation for enterprises' groups, we can obtain a national prices' index. If  $X$  is the national revenue,  $X_{tg}$  is the total revenue of a certain trade group ( $tg$ ) and  $z_{tg}$  is the economic weight of the trade group  $tg$ , the national index can be computed in the following way:

$${}_b\hat{I}_t = \frac{\sum z_{tg} {}_bI_t^{tg}}{\sum z_{tg}},$$

where:

$$z_{tg} = \frac{X_{tg}}{X}.$$

---

<sup>92</sup> An analogous formula can be used for a specific sector or for other kind of groups of units.



## **6 Further researches: data issues and imputation methods application**

This chapter studies some issues related to the data quality and some methods commonly used to improve the quality of collected data.

In the first paragraph (6.1) the main data issues and the research's plans will be shortly presented; in the paragraph 6.2 the classification and the different kind of imputation methodologies will be introduced together with some criteria to measure their effects on the data (par. 6.3).

### **6.1 Data quality issues**

We already mentioned (par. 2.4 and chapter 5) that the collected preliminary data had some quality challenges. The two main aspects are the presence of missing data and of outlier values.

Both the issues are extremely important. And this is why the preliminary collected data are subsequently being reviewed by Price Division. On one hand, the presence of missing data can condition the quality of the estimates causing a nonresponse bias; this challenge is very important, especially when the missing responses are concentrated in a specific category of respondents of the sample/target population. On the other hand, also the presence of outliers can bring to strong biased estimates, because the outlier values have a strong impact on the synthesis parameters.

The missing data and the outlier values were excluded from the data processing shown in the previous chapters, where we generated and tested a simulated population. Nevertheless,

an in-depth study of these units is extremely important to improve the quality of the estimates.

Imputation methods can be useful to study the issues under investigation. We can test the different kind of imputation methods to find out the most appropriate for the data we are working on (which is the best imputation method for profit margin ratios' data?).

Once the most appropriate method of imputation is chosen, we can solve the thinning of the sample caused by missing data and we can get more reliable estimates.

Moreover, we can use imputation methods to impute values whose would substitute the observed outlier values. In this way we can understand what's the impact of the outliers on the estimated index and we can also get more precise estimates of the parameter we are interested to.

But imputation methods could also be useful to understand and measure the impact of the biggest units on the estimates<sup>93</sup> (what's the impact of the biggest units on the level of the index and on its variance?). We can gain this objective, for example, selecting the biggest units and imputing the observed values, as they were missing and without considering their size, to evaluate the impact on the estimates.

Our study of the outliers, of the missing data and of the impact of the biggest units on the estimates is a simulation study because it is based on the study of the simulated population we generated. The results of this study are supposed to be of a general interest and applicable to various other contexts.

Our study is planned as an experimental analysis. The following factors are experimentally changed: the criteria to identify the outliers, the imputation methods, the units' size (the bigger rather than the smaller units). Their impact on the final estimates is then evaluated.

The study's strategy can be summarized in the following points:

- Random selection of some missing units/outliers (especially between the biggest units);

---

<sup>93</sup> We remember that the biggest units are considered the ones with probability of selection equal or bigger than one ( $\pi_{ih} \geq 1$ , where  $\pi$  is the probability of inclusion in the sample of the  $i^{\text{th}}$  establishment of the  $h^{\text{th}}$  group of units;  $i = 1, 2, \dots, n_h$ ;  $h = 1, 2, \dots, k$ ).

- Selection of samples with the bootstrap methods to evaluate the convergence of the variance and of the estimated index;
- Computation of index and its variance using different kind of imputation methods for all the units selected as missing and/or outlier;
- Comparison with the value of the index computed on the whole simulated frame population;
- Identification of the more efficient imputation methods (evaluation considering the convergence to the real values of the computed index);
- Evaluation of efficacy of all the tested methods also in term of variance;
- Experimental analysis: impact of the biggest units (in term of size, that is considering the revenue) on the index value and on the variance.

The research plan show the fundamental relevance of the imputations methods, as tools to improve the data quality: they will be introduced in the next paragraph (6.2).

## **6.2 The imputation methods**

When a survey is conducted, there are often problems related to the participation and to the partial compilation to the survey itself. These issues are related to the incompleteness of data files that usually brings to biased results in the final estimates. The non-response bias due to missing data, in fact, is one of the main issues we have to manage with carrying out surveys of different kind. It's usually less important in the administrative data, but becomes more and more important in other kind of surveys<sup>94</sup>.

In the following paragraph the main typologies of missing data will be presented.

---

<sup>94</sup> Mueller et al. (1995).

### 6.2.1 Missing data

The missing data are usually classified into two main categories corresponding to two different kinds of nonresponse.

The first one is when we have *nonresponse units*: these are units who refuse to participate to the survey, units that are not available or contactable, moved or dead people and so on. The presence of units who refuse to participate to a survey is growing in the last years (Nordholt, 1998). The presence of non-response unit is even more relevant if we consider the probability design of the surveys; in fact, if the probability sample reduces the mean square error, the problem of low participation rates becomes very important. Nevertheless, the main problem related to the nonresponse units is that they are usually selective; this “means that the answers of the non-respondents differ from the answers of the respondents”<sup>95</sup>. This problem is also called “selectivity” of respondents: the respondents (that is the complete cases) could be not representative of the entire population, but only of a part of it. This is also called a problem of representativeness of the sample. Consequently the nonresponse bias grows up as much as the population of respondents is different from the nonrespondents’ one for one or many characteristics. The presence of non-response units or of units not participants to a survey is an important issue that usually cannot be ignored.

The second category of nonresponse is called *item nonresponse* (or *partial nonresponse*): one interviewed unit refuses or can’t answer to one or more of the items proposed in the survey, but for some other items we get an answer. The item nonresponse is caused by different kind of reasons: the interviewed doesn’t know which is the right answer, he/she doesn’t have any opinion about the topic, he/she doesn’t want to answer to that specific item or to a specific part of the questionnaire, he/she cannot decide the answer to chose, the interview is stopped before the end, the answer is not valid or incoherent with other answers given in the questionnaire, and so on. The main problem related to the item non

---

<sup>95</sup> Nordholt (1998).

response is also that usually this kind of missing data are selective. For this reason, we could get results not representative of the target population.

The problem of missing data is also important in our specific research's field. The missing data about prices of a certain establishment (or, worse, of a certain group of establishments) relative to a certain time not only interrupt the historical series of the data, but can also affect the reliability of the estimates, bringing to biased results. This problem is even more serious if the missing data come from the biggest units. In fact, not only they are more important (because they are supposed to have more influence on the considered market), but they also are not so numerous, in the frame population, as the smaller units (so the substitution of a nonresponse unit in the selected sample could be not so easy, or even impossible).

In conclusion, generally speaking, if in a dataset we have one of the two kinds of nonresponse (or both of them), we have to manage with considerable problems due to these missing data.<sup>96</sup>

Three ways to manage with these problems are presented in Little (1988):

1. The first way is to leave the missing data in the dataset giving them a special code. The disadvantage of this approach is that some statistical software doesn't process units that have one or more missing data (that is units with an incomplete line of information). This means a discard of uncompleted cases, and sometimes they could be numerous. This method, based on the elaboration of the complete response cases only, is also called "available case method". It is based on the assumption, usually not realistic, that the non-respondents are a non-selective subset of the set of all the units; this means that the respondents are always considered representative of the target population. If we don't want to lose too many units, another kind of suggested approach is to use the *Maximum Likelihood (ML)* introduced in Little (1982) and Little and Rubin (1987) to model the incomplete data.
2. Weighting complete cases, when we have nonresponse units, is another possible approach to solve the problem of nonresponse bias. In fact, if there are many

---

<sup>96</sup> A review of the problem of missing data can be also found in Madow, Nisselson, Olkin and Rubin (1983).

nonresponse units, we might have, as told before, problems connected with the selectivity of respondents.<sup>97</sup> To solve this problem is possible to weight the collected data using some additional information about the entire population; this is usually derivable from auxiliary variables, appropriately chosen basing on the topic of the survey. In the following analysis the nonrespondent units are not considered, and the respondents units are re-weighted to cover also the weight that nonrespondents should have in the sample.

3. The third method is the data imputation. Imputation methods usually are not used to handle the nonresponse units (like weighting method), but the item nonresponse units<sup>98</sup>. This means that not all the answers of a respondent (or into a record) are missing. The imputation has the objective to adjust the results of a survey for the presence of this kind of missing data. These methods, generally speaking, usually replace a missing value in the dataset using other information contained in the incomplete questionnaire (that is answers given by the respondent through other items). The main principle of imputation is to substitute each missing data item with at least one possible response. In this way we can get a complete dataset available for further elaborations<sup>99</sup>.

The “*main reason for weighting or imputation in large surveys is to produce a more representative rectangular file for analysis*” without missing values<sup>100</sup>, even if “*the imputation is not as widely accepted as the technique of weighting*”<sup>101</sup>.

Nordholt (1998) underlined that is always important to have the choice between non-imputed and imputed variables: he proposes to create two different datasets: “*one without imputations and one including the imputations, or one without imputations and one with the imputations only*” (in this way the imputations could be easily added to the rest of the

---

<sup>97</sup> The problems of selectivity and of representativeness are also discussed in Nordholt (1998) and Little (1988).

<sup>98</sup> That is the second kind of problem seen in the classification presented above.

<sup>99</sup> For more details about imputation methods, see Montaquila and Ponikowsky (1993, 1995).

<sup>100</sup> Little (1988).

<sup>101</sup> Nordholt (1998).



data). Therefore, it is understandable how is important to have, in both the proposed cases, a system to flag the imputed data in the dataset.

The main advantages of imputation methods are underlined by Kalton and Kasprzyk (1982). First of all, the survey estimates' biases caused by missing data are reduced; then is simpler both to work and to present data, because the work is done as all the data were available and as the dataset was complete and no algorithm to estimate population's parameters is necessary; thirdly, the results obtained from different analyses are consistent. If advantages are clear, it has to be taken into consideration that the application of this kind of technique of adjustment generates also some risks. For example, a general method could not fit well (or in the same way) to every context and to every kind of analysis. Nevertheless, this doesn't mean that an adjustment shouldn't be done; this only means that the adjustments should be carefully developed and the way to adjust data should be carefully chosen, considering the context of the research and the purposes of the researchers. To better develop the adjustment procedures, *"the data producer should also communicate the operating characteristics of the adjustment procedure to the user so that its limitations are clear. Moreover, [as already told] imputations should be flagged so that users have the option of developing their own adjustment"*<sup>102</sup>.

Another defect of imputation methods is that we get an overestimation of the precision of the estimates, when we treat imputed data as observed data.

There is also the risk to compare two surveys that had different results: one of them with imputed data and the other one without imputed data. Some of the differences that could be found in the results of the two surveys can be referred both to real inference and to imputation effect. The risk could be solved by comparing imputed (or non-imputed) datasets only (Nordholt, 1998).

---

<sup>102</sup> Little (1988).

## 6.2.2 Principles of imputation methods

As already told, generally speaking imputation methods are based on the concept that, if a nonresponse item is found, we can obtain some information about that topic from a set of other items provided into the questionnaire (or, at least, another single item). An appropriate selection of these variables (called “auxiliary variables”) can be helpful to assign a value to the missing response (Kalton and Kasprzyk, 1982); to select this set of variables, one can usually use, for example, regression methods or log-linear models. But the imputation is much more useful if the researcher starts working on that from the early-stage, that is from the writing down phase of the questionnaire. In fact, if a large amount of nonresponses for an item is expected, in the questionnaire a set of auxiliary variables useful to predict the missing values could be used. For example, if a survey about workers of a certain area is in the developing stage and if the questionnaire should be provided with a question regarding the monthly revenue, in the case a high number of non respondents is expected for that item, we can insert in the questionnaire two (or more) other items usable as auxiliary variables. This is what usually happens when we have question about the salary. If we provide the questionnaire with two questions regarding the weekly average hours of work ( $z$ ) and the hour salary ( $k$ ) and if the data about the monthly revenue ( $y$ ) of some worker (the  $i^{\text{th}}$  worker, for example) is missing, we could impute the missing data in the following way:

$$\hat{y}_i = z_i \cdot k_i.$$

So, once an appropriate set or auxiliary variables is chosen, an imputed value for the variable  $y$  (and relative to an item non respondent) could be seen as a function of the chosen auxiliary variables ( $z_1, z_2, \dots, z_p$ ) and of an estimated residual ( $e$ ) (Santos, 1981a and 1981b). This means that, if  $m$  is one nonresponse unit considering the item  $y$ , the value of this unit ( $\hat{y}_i$ ) could be estimated through the following mostly linear function:

$$\hat{y}_i = f(z_1, z_2, \dots, z_p) + e.$$

This function is often estimated from the data available from other items (item 1, 2, ...,  $p$ ) while  $e$  is a random effect.

It is possible to choose between different imputation methods, and every method fits better to one rather than to another context or research. Nevertheless, to better understand which kind of method is more appropriate to a specific context, Little (1988) underlines the desirable properties of any imputation methods. They will be explained in the following.

1. “*Imputations should be based on the predictive distribution of the missing values, given the observed values for a case [...] to preserve the distribution of the variables*” considered in the study. Through a modeling of the observed data, we can obtain imputation of a certain quality level only if the model underlies in a good way the data. For example, using average cannot be helpful (the estimates provided are the same, but the associated standard error is worse), nor can be helpful to add residuals or random errors to the mean imputation (because it usually distorts the measures of the spread).
2. In the first step of the study “*all observed items for a case*”, eventually selected with accuracy, “*should be taken into account in developing imputations*” because we must “*make the best use of the available information on incomplete case*”. This is because some measures of association could be biased due to the absence of relevant variables (Kalton and Kasprzyk, 1982). The imputation conditioned on observed items has two main advantages: reduces the nonresponse bias and reduces the variance (Little, 1986). The main kind of conditioning are:

- a. Conditioning limited to a *single item*. For example, let us suppose that  $\bar{x}$  is an observed value and  $\bar{y}$  is missing; to make the imputation we take into consideration  $\bar{R}$ , the average ratio of the variable  $Y$  on  $X$  within the group of respondents:

$$\bar{R} = \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{y_i}{x_i}$$

where  $n_r$  is the number of respondents and  $i$  ( $i = 1, 2, \dots, n_r$ ) is a response unit for the two variables.

In this way the estimated  $y_i$  (for a missing unit  $i$ ), called  $\hat{y}_i$ , would be computed as follows:

$$\hat{y}_i = \bar{R} \cdot x_i.$$

The results of imputation are as better as we use more than two recorded items.<sup>103</sup>

- b. Conditioning using *adjustment cells*. We can use observed items to build adjustment cells. The data about the nonrespondents for some items could be imputed, in each cell, using the information available about the respondents in the same cell. If Welniak and Coder (1980) suggested to use, imputing in a certain cell, the value from an individual donor, Lillard et al. (1986) preferred as potential predictors a larger number of observed items. In this way the imputation would assume an implicit model including the main effects and the interactions between the considered variables.
  - c. *Regression method* is preferred by Little (1988): he considers more important the direct effect within the variables rather than the interaction between them. This approach works well especially for longitudinal imputations, based on the use of information coming from other waves of the survey (in this way the number of predictors could be larger).
3. Imputation “*should take into account contextual knowledge about the variables being imputed*”. This brings advantages in the choice of adjustment cells or when they are too small and criteria to decide how to collapse them are needed. For this reason the imputation model should include all variables that could be considered predictive for a certain item; moreover, subject-matter specialists should be involved in the imputation process, as happens in the work of Greenberg and Surdy (1984), even if input from specialists could be expansive. To limit the impact of this input, Little and Smith (1987) propose a more automatic and empirically based imputation schemes.

---

<sup>103</sup> For further details see Little and Smith (1983 and 1987) and Kusch and Clark (1979).

4. Imputation methods “*should avoid excessive extrapolation beyond the range of the data, unless objective evidence is available to substantiate these models*”. This could happen, for example, using regression imputation or certain nonrandom nonresponse models that sometimes could fit well to the data for the only reason that nonresponse mechanisms are often nonrandom. Lillard et al. (1982) and Greenlees et al. (1982) implemented an adjustment for nonrandom nonresponse based on the stochastic censoring models<sup>104</sup>; these are highly sensitive to distributional assumptions and to the choice of predictors.
5. “*Imputations should be drawn from the predictive distribution in principle 1., not means, to preserve the distribution of the variables in the filled-in data set*”. This means that some noise has to be added to the predicted means, if we are not interested in means and totals. Hot-deck methods (like the one used in Colledge et al., 1978) tend to preserve the variables’ distributions, while methods based on regression (Dixon, 1983) or ratios (Kusch and Clark, 1979) need to be modified to preserve the distributions of variables, for example adding a residual from a matched respondent.
6. “*A method should be provided for computing sampling error of [appropriate] estimates*”.

The principles 1. and 3. are particularly connected with the modeling approach. Nevertheless, the Bureau of the Census practice prefers a design-based descriptive analysis (Chapman et al., 1986; Hanson, 1978).

### 6.2.3 Classification of imputation methods

Kalton and Kasprzyk (1982) classified the imputation methods considering two aspects: the use of auxiliary variables (some methods use these kind of variables, some others no) and the value assigned to the residuals.

---

<sup>104</sup> See Amemiya (1984), Hausman and Wise (1977), Heckman (1976).

If we consider the first aspect, we can have a further sub-classification:

- *Categorical* (or *continuous recoded as categorical*) auxiliary variables: they are useful to classify respondents in a certain number of classes.
- *Continuous variable*: in this case the categorical variables are converted into dummy variables.

If we consider the other aspect, that is the randomization of the imputation process, we have the following classification, based on three different categories (Nordholt, 1998): deductive, stochastic and deterministic imputation methods.

- *Deductive imputation*: the imputed value is given from known information available about the considered unit. The deductive imputation method is discussed in par. 6.2.4.1.
- *Stochastic imputation*: it is used when the unknown value cannot be deduced from the value of other variables or from the value observed on the same variable in the past. In the stochastic imputation we have to manage with a prediction and with a residual,  $e$ . This is randomly generated and assigned to a non-response item unit to get the imputed value<sup>105</sup>. Usually this random residual is assigned to a prediction made with deterministic imputation (see more forward). This random component is useful “to preserve the variances and covariances of the imputed data” (Nordholt, 1998).

The main problem with this method is that the imputed values need to be feasible; nevertheless, adding a random residual to the deterministic component, we could get an unfeasible value even if the imputed value is feasible.

A different version of this method is based on the imputation of a random component of a value observed on a different unit, randomly selected (*random hot-deck* or *overall method*): this could solve the last problem seen previously. The following imputation methods can be classified as stochastic: random imputation

---

<sup>105</sup> Usually the random residual should be normally distributed with variance  $\sigma^2$  and average equal to zero. This variance “is equal to the residual variance of the regression of the variable that has to be imputed on some explanatory variables” (Nordholt, 1998).

overall (see par. 6.2.4.6), random imputation within classes (6.2.4.7), hot deck - random selection (6.2.4.8), flexible matching imputation (6.2.4.12).

- *Deterministic imputation*: the residual  $e$  introduced for the stochastic imputation methods is set to zero. The following ones are some of the methods considered deterministic: mean imputation overall (par. 6.2.4.2), running mean (6.2.4.3), mean imputation within classes (see par. 6.2.4.4), sequential hot deck (6.2.4.10<sup>106</sup>), cold deck (6.2.4.13).

A disadvantage of all these methods, caused from the imputation of the best prediction at the record level, is that the variance is underestimated; in fact the deterministic imputations methods make the distribution too peaked. For this reason sometimes a worse imputation method is preferred, if it's useful to get a less distorted distribution (Nordholdt, 1998<sup>107</sup>).

The one proposed above is a general classification only: some imputation methods are not easily classifiable. For example, usually hot-deck imputation could be considered either deterministic or stochastic method; it is considered stochastic if the units are randomly ordered or if a unit is randomly chosen from a group; it is considered deterministic if the group of units is not randomly sorted and/or if a record is not randomly selected from the group.

After the main principles of the imputation methods and after their classification, in the following paragraph, the main imputation methodologies are described.

---

<sup>106</sup> A particular version of the sequential hot deck can be considered a stochastic imputation method (see par. 6.2.4.10).

<sup>107</sup> See also Kalton (1983a).

## 6.2.4 The different kind of imputation methods

The knowledge about the survey that has to be implemented should suggest which kind of imputation methods to use. In the following we will give a brief introduction about the main imputation methods.

### 6.2.4.1 Deductive imputation

In the deductive imputation overall method a missing value for the variable  $Y$  can be imputed as an exact function of the values observed on the same unit about one or more than one (for example,  $k$ ) other auxiliary variables ( $z_{ij}$ ):

$$\hat{y}_i = f(z_{ij}) \quad j = 1, 2, \dots, k$$

where  $i$  indicates the  $i^{\text{th}}$  unit with a missing value for the variable  $Y$ , the hat indicates an imputed value,  $z_{ij}$  is the variable used for the imputation ( $j = 1, 2, \dots, k$ ) of the unit  $i$ .

This happens, for example, when we can compute the age of a  $i^{\text{th}}$  respondent ( $\hat{y}_i$ ) if we have the date of birth ( $z_{i1}$ )<sup>108</sup>.

Another similar situation is a panel survey, where we can know or deduce data if we have the same kind of data referred to the previous survey's wave. For example we can impute the missing age of a product obtained in the last survey, if we have the age of the same product twelve months before this last survey.

Sometimes deductive imputation is considered part of the editing process; methodologically speaking it is not considered very interesting (Nordholt, 1998).

### 6.2.4.2 Mean imputation overall (MO)

The mean imputation method, also called *MO* (*Mean imputation Overall*<sup>109</sup> or *Mean of Observations*<sup>110</sup>), is based on the substitution of a  $Y$  variable's missing value with the mean of the overall group of respondents for the same variable ( $\bar{y}_r$ ).

---

<sup>108</sup>  $z_{i1}$  is the only variable used to impute the date of birth; so in this case  $j = 1$ .

<sup>109</sup> Karlton and Kasprzyk (1982).

<sup>110</sup> Lawrence (1980).



$$\hat{y}_i = \bar{y}_r \quad i = 1, 2, \dots, p,$$

where  $\hat{y}_i$  is the value imputed for the  $i^{\text{th}}$  unit and  $p$  is the number of unit with a missing value for the variable  $Y$ .

This method could be seen as a degeneration of the deterministic process of imputation in which there are no auxiliary variables.

#### 6.2.4.3 Running Mean (RM)

To apply this kind of method, we need to have a sequence of a variable's observed values referred to different times. To impute a missing value of the considered variable, the mean of all the previous observed values in the sequence is used.

In this method the variable to be imputed at previous times can be considered as the auxiliary variable.<sup>111</sup>

#### 6.2.4.4 Mean imputation within Classes (MC)

In this method, also called *Cell Mean* or *Group Means*<sup>112</sup>, one or more variables are used to share the sample into  $H$  different classes or strata.

Fixed a time  $t$ , for each class  $h$  ( $h = 1, 2, \dots, H$ ), to all the nonrespondent units the same value of the variable  $Y$  is assigned; this value is the mean computed on all the respondents of the considered class.

$$\hat{y}_{t,hi} = \bar{y}_{t,rh} = \sum_i \frac{y_{t,rih}}{n_{t,rh}}$$

where:

- $\hat{y}_{t,hi}$  is the imputed value of the variable  $Y$  for the  $i^{\text{th}}$  unit ( $i = 1, 2, \dots, k$ ) of the  $h^{\text{th}}$  class ( $h = 1, 2, \dots, H$ ) in the time  $t$ ,
- $\bar{y}_{t,rh}$  is the average of the same variable  $Y$  computed on the respondent units of the same class  $h$  (time  $t$ ),
- $y_{t,rih}$  is the value observed on the respondent ( $r$ ) unit  $i$  of the group  $h$  in the time  $t$ ,

<sup>111</sup> For more details, see Lawrence (1980).

<sup>112</sup> This method is also called *MN* or *Mean Imputation method* (West, Butani and Witt, 1988).

- $n_{t,rh}$  is the number of respondent units in the group  $h$  at the time  $t$ .

This method, furthermore, is very useful if the response rate is high, because it will not alter an estimate of the stratum mean. If the response rate is low, the distribution of the sample could be skewed toward the mean from this kind of imputation.<sup>113</sup>

A different version of this method of imputation suggests to use the median of each strata, instead of the average (West, Kratzke and Robertson, 1993).

As suggested by Kalton and Kasprzyk (1982), the sample could be divided into different classes using, for example, two categorized variables. These allow to obtain, from their cross tabulation, cells useful for the classification. It is also possible to use some classification variables for a part of the sample and some others for the other part.

With the mean imputation within classes, *“if all the cells in the cross-tabulation are used, the linear function can be expressed as a model with the main effects and all levels of interaction for the auxiliary variables”*<sup>114</sup> using a model like the following:

$$\hat{y}_i = b_{ro} + \sum b_{rh} z_{hi}$$

where:

- $\hat{y}_i$  is the imputed value for the unit  $i^{\text{th}}$  ( $i = 1, 2, \dots, k$ );
- $b_{ro}$  is the main effect;
- $b_{rh}$  is the effect of the class  $h^{\text{th}}$  ( $h = 1, 2, \dots, H$ );
- $z_{hi}$  is a dummy variable with value 1 if the  $i^{\text{th}}$  respondent is in the class  $h$ , 0 otherwise;
- the stochastic residual,  $e_i$ , is equal to zero (deterministic model).

This method allows to attenuate the cavariances (Kalton and Kasprzyk, 1986).

The same criteria to define the classes could be used for the random within-cell hot-deck and for the nearest neighbor within-cell hot-deck method (we'll talk about these in the following).

---

<sup>113</sup> West, Butani, Witt and Adkins (1989).

<sup>114</sup> Kalton and Kasprzyk (1982).

#### 6.2.4.5 Mean of Log variable (MNL)

The Mean of Log variable method, similar to the previous Mean imputation within classes (6.2.4.4), uses the log-transformation of the variable we want to impute<sup>115</sup>.

$$\hat{y}_{t,hi} = \sum_i \frac{\log(y_{t,rhi})}{n_{t,rh}},$$

where  $\log(y_{t,rhi})$  is the log-transformation of the value of the variable  $Y$  for the unit  $i$  of group  $h$  observed in time  $t$  and  $n_{t,rh}$  is the number of respondent units in the same class at time  $t$ .

#### 6.2.4.6 Random imputation Overall (RO)

With the Random imputation Overall (RO) method, if we obtain a missing value of the  $Y$  variable for the  $i^{\text{th}}$  unit, we can impute this value randomly choosing between the same variable's remain observed values.

This is a stochastic method, and it is a degenerate version of a linear function without auxiliary variables:

$$\hat{y}_i = \bar{y}_r + e_i \quad i = 1, 2, \dots, p,$$

where  $\hat{y}_i$  is the imputed value,  $p$  is the number of unit with a missing value for the variable  $Y$ ,  $\bar{y}_r$  is the average of the observed values,  $e_i$  is a stochastic residual.

If we use this kind of imputation, in terms of mean and variance, it is like to ignore the missing data (Lawrence, 1980).

#### 6.2.4.7 Random imputation within Classes (RC)

This method works in the same way of the *RO* method (par. 6.2.4.6). In fact it is based on the random selection of an observed value of the variable that has to be imputed. The difference is that this is applied within the imputation classes; these classes are defined by one or more auxiliary variables conveniently chosen. The imputation classes are usually defined using an auxiliary variable that has known values for all the units.

---

<sup>115</sup> West, Butani and Witt (1988).

To each non-respondent of one considered class a value randomly chosen within the observed values of the same class is assigned. The selection is made randomly and without replacement. The unit randomly selected is called *donor*, while the unit that has a missing data is called *recipient*<sup>116</sup>.

The RC is a stochastic method; in fact, for each class  $h$  ( $h = 1, 2, \dots, H$ ), we have a stochastic residual ( $e_{hi}$ ). So, the imputed value,  $\hat{y}_{hi}$ , can be seen as:

$$\hat{y}_{hi} = \bar{y}_{rh} + e_{hi} \quad i = 1, 2, \dots, p,$$

where:

- $\hat{y}_{hi}$  is the imputed value of the variable  $Y$  for the  $i^{\text{th}}$  unit ( $i = 1, 2, \dots, k$ ) of the  $h^{\text{th}}$  class ( $h = 1, 2, \dots, H$ ),
- $\bar{y}_{rh}$  is the average of the same variable  $Y$  computed on the respondent units of the same class  $h$ ,
- $e_{hi}$  is the stochastic residual of a  $i^{\text{th}}$  respondent randomly selected within the class  $h$ .

From the previous formula we can deduce that:

$$e_{hi} = y_{rhk} - \bar{y}_{rh} \quad i = 1, 2, \dots, p$$

where  $y_{rhk}$  is the real mean of the  $k$  observed values of the variable  $Y$  in the  $h^{\text{th}}$  class. And so:

$$\hat{y}_{hi} = y_{rhk} \quad i = 1, 2, \dots, p.^{117}$$

An alternative formulation for the imputed value for the  $i^{\text{th}}$  unit of the  $h^{\text{th}}$  class ( $\hat{y}_{hi}$ ) is:

$$\hat{y}_{hi} = b_{ro} + \sum b_{rh} z_{hi} + e_{hi} \quad i = 1, 2, \dots, p; h = 1, 2, \dots, H,$$

where:

- $b_{ro}$  is the main effect;
- $b_{rh}$  is the effect of the class  $h^{\text{th}}$  ( $h = 1, 2, \dots, H$ );
- $z_{hi}$  is a dummy variable with value 1 if the  $i^{\text{th}}$  respondent is in the class  $h$ , 0 otherwise.

---

<sup>116</sup> Montaquila and Ponikowski (1995).

<sup>117</sup> Kalton and Kasprzyk (1982).

- $e_{hi}$  is the stochastic residual of the  $i^{\text{th}}$  respondent of the class  $h$ .

One advantage of the random imputation within classes method is that it retains the respondents' distribution of the variable within the considered cell.

#### 6.2.4.8 *Hot deck - random selection*

The most common hot-deck procedure is the one used by the Bureau of the Census in the Current Population Survey (CPS).<sup>118</sup>

This kind of imputation requires the specification of imputation classes.

For each class one starting point has to be chosen: if  $Y$  is the variable that we need to impute, this point is one of the  $Y$ 's observed values within the same class. As different alternative choices, one could choose a representative value of the class (like the average or the median), a representative value registered in the same class in the previous run of the survey, and so on.

The process goes sequentially from a unit to the other of the same cell. It gives results similar to a SRS (Simple Random Sampling) with replacement selection of donors within the considered class.

Once the initial selection of one unit is made:

- If the selected record has a valid observed value about the variable  $Y$ , this value is stored for the considered imputation class.
- If the selected record has a missing value for the  $Y$  variable, this value is substituted with the value previously stored for the same class.

This imputation method gives better results if the units of the considered cell are not randomly ordered, but ordered using an auxiliary variable; this is done in order to obtain a positive autocorrelation (Kalton and Kasprzyk, 1982).

---

<sup>118</sup> Brooks and Bailer (1978).

When, within the same class, a record with a missing data is followed by one or more records that are missing too, to all these records the same value is assigned (the value observed on the last not-missing unit).

From this last situation we can understand one of the main drawbacks of the use of the hot-deck random selection method, because it could bring to a significant loss of precision of the estimates of the survey.

The same disadvantage occurs when the random within class method is applied, but in this last case the problem could be reduced sampling donors without replacement.

#### 6.2.4.9 Distance function matching – Hot deck-Nearest neighbour

The distance function matching method (also known as hot deck-nearest neighbour) uses an auxiliary variable and a predefined method to measure the distance between a nonrespondent unit and a respondent one.

All the units in the sample are previously ordered by an auxiliary variable conveniently chosen; after this, using the same auxiliary variable, the distance between each nonrespondent and the nearest respondents is computed. Considering the variable we want to impute, to the nonrespondent is assigned the same value observed on the nearest respondent. This last unit is called *donor*.

For example, we can consider, in a certain observed sample, the nonrespondent unit's value  $y_t$  (that is the observed value of the chosen auxiliary variable  $Y$  in the time  $t$  on that unit). Once we have ordered the units of the sample itself by the  $Y$  auxiliary variable, we can compute the distance between this non respondent ( $y_t$ ), the previous unit ( $y_t^{(-1)}$ ) and the following unit ( $y_t^{(+1)}$ ). The simplest way to compute this distance between two units, in terms of a certain auxiliary variable  $Y$ , is using the absolute differences between the values of the same variable  $Y$  observed on those units. The first step is the computation of the distances, as follows:

- $y_t^{(-1)} - y_t$  and
- $y_t^{(+1)} - y_t$ .

The distances are then converted into absolute values. The value to be imputed is chosen basing on the following conditions:

- If  $|y_t^{(-1)} - y_t| \leq |y_t^{(+1)} - y_t|$ , the value observed on the unit  $y_t^{(-1)}$  is assigned to the nonresponse unit  $y_t$ .
- In the opposite case, when  $|y_t^{(-1)} - y_t| \geq |y_t^{(+1)} - y_t|$ , we assign the value  $y_t^{(+1)}$  to the missing value.

There are different choices about the distance function that can be used to evaluate the distance between two units: for example one can use the absolute difference between the values of the two variables referred to the two units (as seen before); the Mahalanobis distance<sup>119</sup> is another option, but other kind of evaluation methods are also available.

Another variation of this method (based on the computation of the distance using an auxiliary variable) is to assigns to a nonresponse unit an average of two or more values observed on the units that are more close to the missing unit. Ford (1976), for example, used two respondents close to the non-respondent. The disadvantage of these procedures is that the distribution could get distorted (Ford, 1976).

The nearest neighbour method is based on the concept of “covariates”<sup>120</sup>. The basic idea of this method is that two units that are close (or identical) considering a variable (*covariate*) should be similar, or almost equal, also considering another variable (the one with a missing value to impute). This is the criteria used to match a non respondent unit with its donor. If this principle is respected, the donor’s value can be used to impute the missing value of the nonresponse unit.

To decide if two units are matched, one can consider a single variable (called “covariate”) or a set of variables (*covariates*) that define the criteria of decision. For example, if we consider one auxiliary variable only, the nearest respondent is the one that has the minimum absolute difference, considering the auxiliary variable used or one of its function. But what can happen is that there could be more than one donor for the unit we want to

---

<sup>119</sup> For more details about Mahalanobis distance, see Vacek and Ashikaga (1980).

<sup>120</sup> The concept of covariates is also known as “matching key”: two “records match if they have the same values on the covariates” (both the concepts were used in Nordholt, 1998).

impute. This could also happen if we are using a set of variables (covariates): a unit could matches with several other donors. In both these last cases, it would be possible to add one or more adjunctive variables to the group of covariates to reduce the number of matching records. The distance function is also useful to reduce the multiple use of donors (Kalton and Kaspritz, 1982): if we find more than one donor for the units we want to impute, we can change the used distance function to refine the research of donors.

In the same situation, if there are two or more values observed on units whose are at the same distance from the unit we need to impute, another solution could be imputing one of them randomly (Lawrence, 1980).

Seen how this method works, it is clear that the choice of the covariates is fundamental. The method works with quantitative variable most, but also qualitative variables could be included in the distance function; it is also possible to consider these qualitative auxiliary variables to form imputation classes. In fact the hot deck imputation method based on the nearest neighbor could be used also within each class. These classes can be defined using different kind of variable (for example: size and period of time  $t^{121}$ ). The definition of the cells is much more appropriate as much the auxiliary variable (or variables) are highly correlated with the characteristic used to choose the donor. Imputation cells are usually defined using the same criteria seen for the random within-cell hot-deck method.

We said that it's possible to work with a group of covariates. A good solution to use more than one auxiliary variable is to transform them with ranks. These ranks represent the importance of one variable in the distance function.

For example, one way to compute the distance between two units using the ranks is the following: if  $R_{hi}$  is the rank of one nonrespondent (the unit  $i$ ) and  $R_{hk}$  is the rank of the potential donor  $k$  (both of them referred to the same variable  $h$ ), one kind of distance function could be:

$$D(i, k) = \text{Sup}_h w_h |R_{hi} - R_{hk}|.^{122}$$

---

<sup>121</sup> West, Butani, Witt (1988).

<sup>122</sup> For more details, see Sande (1979a).



There are two more complex variations about the hot deck-nearest neighbour method: the first one, NNI, is based on the linear interpolation of the ordered list based on the auxiliary variable; the second one, NNIR, is similar to the first but provides, for the border case, a ratio adjustment. For more details, see West, Kratzke and Robertson (1993).

A particular case is when we are working with panel data. In these situations, the imputation strategy is similar to the cold deck one, but the donor is not chosen from a previous wave of the survey, but from the same wave.

#### *6.2.4.10 Sequential hot deck*

Another variation of the hot-deck method is suggested by Nordholt (1998). It is the sequential hot-deck method. To use this method the units of the sample should be ordered in a special way; if not, the following random hot-deck within classes (see par. 6.2.4.11) is preferred.

Once the units of the sample are ordered basing on some auxiliary variable, the donor for a non respondent unit is considered the last matching record before the record with the missing value, taking into consideration the variable that we want to impute.

This kind of imputation is deterministic, but could be turned into a stochastic one in two different way: adding a random residual to the imputed value or randomly ordering the units of the sample.

#### *6.2.4.11 Random hot deck within classes*

Another variation of the hot-deck imputation method, more common of the sequential hot-deck, is made with a random selection that considers groups of units. These classes can be defined, for example, using a size variable.

Let us consider a class with some values to impute, considering a certain variable. If all the units in the class have the same value of the variable we want to impute, that value is the one that will be imputed (*exact matching*).

If we want to impute the missing values of a variable  $Y$ , and if the values of the same variable  $Y$  observed in the same class are different, the imputed value for a nonrespondent unit  $i$  of a class  $j$  at the time  $t$  ( $\hat{y}_{t,ij}$ ) is<sup>123</sup>:

$$\hat{y}_{t,ij} = y_{t,ij}^*,$$

where  $y_{t,ij}^*$  is the value of the same variable  $Y$  observed on a randomly selected respondent in the time  $t$  and in the class  $j$  (*random matching*). Selection is made independently within strata and with replacement.

This method, if we have one class only, corresponds to the overall random hot-deck method (paragraph 6.2.4.8).

The random hot deck in general (and, more appropriately, in its “within classes” version) could be also used to impute more than one variable simultaneously. One variable is considered the main one and, once the donor is selected, the same donor provides values for the imputation of the missing values of all the other variables we want to impute too (*record matching* or *common donor rule*). This method obviously works well when the variables we want to impute (the main one and the secondary ones) are strongly correlated; it is usually used when there are people with several missing values related to these variables. Moreover this method avoids inconsistency due to imputations.

A limitation of its application is the fact that only the units whose have valid values for all the variables in the group can be considered as donors.

Further discussions about the hot-deck methods could be found in: Allen (1990), Bailar, Bailey and Corby (1978), Bailar and Bailar (1978, 1979), Ford (1980), Kalton (1983a, 1983b), Kalton and Kasprzyk (1986), Oh and Scheuren (1980), Oh, Sheuren and Nisselson (1980), Sande (1979a, 1979b).

---

<sup>123</sup> West, Butani, Witt (1988).

#### 6.2.4.12 *Flexible Matching Imputation*

With the flexible matching imputation method, used from 1976 for the CPS March Income Supplement, respondents and nonrespondents are sorted into a large number of imputation classes, using auxiliary variables about size.

This method works hierarchically: if a nonrespondent is not matched with a respondent in an initial classification, classes are collapsed and the process goes on at a lower level. This procedure, in comparison with the hot-deck random selection, brings to obtain closer matches for nonrespondents units. It is also useful to avoid the multiple use of respondents. Further discussions about the flexible matching imputation method could be found in Coder (1978) and Welniak and Coder (1980).

#### 6.2.4.13 *Cold deck*

The cold deck method (Nordholdt, 1998), a deterministic imputation methods, is based on the use of the information coming from another dataset to impute the missing values of a certain dataset. It is usually used with panel survey, where the same unit could be interviewed for more than one time: the value(s) observed, on the same unit, in a previous wave of the survey can be imputed to the missing value of the current survey. The main difference from the methods that will be introduced in the following paragraphs (6.2.4.14, 6.2.4.15, 6.2.4.16) is that the data are available into different datasets.

#### 6.2.4.14 *Establishment trend times the last observed value (UILT)*

The establishment trend times last observed value method<sup>124</sup> is specifically applied to establishments, object of study of administrative surveys, or to other survey where we have historical data for the same units.

If, for example, data are monthly collected, the missing value of a certain establishment  $i$  in time  $t$  ( $y_{t,i}$ ) is imputed projecting the same over-the-month change of the variable  $Y$  observed one year earlier, that is:

$$\hat{y}_{t,i} = \frac{y_{t-12,i}}{y_{l-12,i}} \cdot e_{l,i},$$

---

<sup>124</sup> The method is also defined as UILT in Mueller et al. (1995).

where:

- $\hat{y}_{t,i}$  is the imputed value for the establishment  $i$  at time  $t$ ,
- $y_{t-12,i}$  is the value observed on the same establishment 12 months ago,
- $y_{l,i}$  is the last available data (supposed at time  $l$ ) for the establishment  $i$ ,
- $y_{l-12,i}$  is the value observed 12 months before the last available data on the same unit.

The main assumption of this method is that the over the month change of the last year is considered similar to the over the month change observed one year before: this means that the last movements are considered the extension of the previous<sup>125</sup> observed ones.

#### 6.2.4.15 Sample trend times the last observed value (UIST)

The sample trend times the last observed value method<sup>126</sup> was originally applied to establishments of monthly administrative surveys; it is a partially different version of the method introduced in par. 6.2.4.14, because it considers the over-the-month change from the previous time referred to all the respondent establishments. This means that the missing value of a certain establishment  $i$  in time  $t$  ( $y_{t,i}$ ) is imputed projecting the same over-the-month change of the variable observed on the establishments that participated at the survey:

$$\hat{y}_{t,i} = \frac{s_t}{s_l} \cdot y_{l,i},$$

where:

- $\hat{y}_{t,i}$  is the imputed value for the establishment  $i$  at time  $t$  (variable  $Y$ ),
- $s_t$  is the sum of the observed values of the variable we want to impute at time  $t$ ,
- $s_l$  is the analogue sum on the same respondent units considered at a previous point, but at time  $l$ ,

---

<sup>125</sup> As “previous” we intend the movements observed 12 months before, because in this way we exclude the possible influence of the seasonal evolution of the variable that we might have considering a change between a month and a previous one.

<sup>126</sup> This method is also known as UIST (Mueller et al., 1995).

- $y_{l,i}$  is the last available month value observed before time  $t$  on the same unit.<sup>127</sup>

The main assumption of this method is that the over the month change doesn't vary so much across the time: the trend of the data is considered almost constant. This method doesn't work so well if there is a strong seasonal behavior in the studied data (in this case, the method introduced in par. 6.2.4.14 is preferred).

#### 6.2.4.16 *Last observed value for the establishment*

The last observed value method for the establishment<sup>128</sup> is also called Previous Observations (PO) or Carryover Method of imputation (CO). The data of a certain non respondent unit  $i$  in time  $t$  ( $y_{t,i}$ ) is imputed using the data of the previous time, as follows<sup>129</sup>:

$$\hat{y}_{t,i} = y_{t-1,i},$$

where  $\hat{y}_{t,i}$  is the imputed value for the unit  $i$  in time  $t$  and  $y_{t-1,i}$  is the value observed at a previous time  $t-1$ , or at the last time we have an available data.

The general principle is that, once we have a sequence of data, for each missing value the immediately preceding observed value in the sequence is imputed. This principle is based on the assumption that the observed data have not a strong trend (no big changing are expected from a time to the following one). If there is no a preceding observed value, then the missing value is imputed in other ways (see Lawrence, 1980).

#### 6.2.4.17 *Mean and median Ratio variation (MeanR and MedR)*

The Mean Ratio method (MeanR) was introduced by West, Kratzke and Robertson (1993) and is used mainly with the survey with a high response rate<sup>130</sup>. This method applies the imputation to predefined groups of units: for example a classification made by NAICS code can be useful. This means that we need at least two variables for the units we are

---

<sup>127</sup> Usually  $l$  is considered equal to  $t-1$ . But if there is no available value at time  $t-1$ , we have to choose the same value at time  $t-2$ , and so on.

<sup>128</sup> This method is also known as UILO (Mueller et al., 1995).

<sup>129</sup> West, Butani and Witt (1988).

<sup>130</sup> For survey with a low response rate, this method has the effect to skew the distribution toward the mean (West et al., 1993).

considering: the main variable (the one we need to impute on certain units) and a secondary variable (usually a size variable) or a variable correlated with the main variable (for example, for data about the employees' annual revenue, a wage variable).

If, for example, we are studying the total employment (variable called  $E$ ) of different groups of establishment (defined by NAICS), we can compute, for each one of these groups, the mean of the employment, that is the mean of the main variable:  $\bar{E}_{j,t}$ .

$$\bar{E}_{j,t} = \sum_{i=1}^{n_j} \frac{E_{ij,t}}{n_j}$$

where:

- $\bar{E}_{j,t}$  is the mean referred to the  $j^{\text{th}}$  group at time  $t$ ,
- $E_{ij,t}$  is the employment (the main variable's value) of the unit  $i$  of the group  $j$  observed at time  $t$ ,
- $n_j$  is the number of establishments included in the class  $j^{\text{th}}$ .

One auxiliary variable we can choose is the establishment's wages ( $W_{ij,t}$ ). Similarly to what we did with the variable  $E$ , it is possible to compute the mean of the wages in the time  $t$  for the units of the  $j^{\text{th}}$  group ( $\bar{W}_{j,t}$ ). This is the average by NAICS group of the secondary variable.

The two parameters ( $\bar{E}_{j,t}, \bar{W}_{j,t}$ ) are computed over the whole group of units whose have both the data about the main variable and the secondary variable for the time  $t$ .

If we have a unit  $i$  of the  $j^{\text{th}}$  group that has, for time  $t$ , the wage's value (secondary variable) but not the data of the main variable, the Mean Ratio method (MeanR) imputes the missing value  $E_{ij,t}$  in the following way:

$$\hat{E}_{ij,t} = \frac{\bar{E}_{j,t}}{\bar{W}_{j,t}} \cdot W_{ij,t}$$

where  $W_{ij,t}$  is the observed value of the wages for the unit  $i$  of group  $j$ .

In a similar way it's possible to use the median instead of the mean: in this case, the process is called MedR (or Median Ratio method).

#### 6.2.4.18 Regression models

One kind of regression model used to impute missing data was introduced by Afifi and Elaskoff; it is based on the least square criteria.<sup>131</sup>

West proposed the proportional regression models<sup>132</sup>. Using this criteria, the value of the  $Y_{t,i}$  variable for a missing unit  $i$  at time  $t$  is proportional to the unit's previous time values ( $Y_{t-1,i}$ ), given the following vector of the  $Y$  values for the previous time  $t-1$ :

$$\underline{Y}_{t-1} = [Y_{t-1,1}, Y_{t-1,2}, \dots, Y_{t-1,k}].$$

This means that, if we consider a constant  $\beta$  depending on the time  $t$ :

$$E(Y_{t,i} | Y_{t-1,i} = y_{t-1,i}) = \beta y_{t-1,i}. \quad i = 1, 2, \dots, k$$

The regression model can be written in the following way:

$$Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i} \quad i = 1, 2, \dots, k$$

where:

- $E(\varepsilon_{t,i}) = 0$  and
- $E(\varepsilon_{t,i}, \varepsilon_{t,j}) = \begin{cases} \nu_{t,i} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$
- $\nu_{t,i}$  is the conditional variance of  $Y_{t,i}$  that usually depends on  $Y_{t-1,i}$ .

The regression model, that works well, for example, for employment data<sup>133</sup>, can be fitted in each of the considered strata considering the units that have the data about both the main and the secondary variable.

<sup>131</sup> For more details, see Afifi and Elaskoff (1969).

<sup>132</sup> West (1982, 1983), West et al. (1989).

<sup>133</sup> West et al. (1993).

#### 6.2.4.19 Predicted Regression imputation (PR)

The Predicted Regression imputation method (PR) is considered a generalization of the group mean<sup>134</sup>; it is deterministic and it uses auxiliary variables (quantitative or qualitative converted to dummy variables) to predict the  $Y$  missing value.

Only the units without item nonresponse are used. On these units the regression parameters are estimated considering the chosen explanatory variables. Once the regression parameters are estimated, they can be used to predict the values of the missing data. These values are then imputed on the missing data of the variable object of study.

The simplest case of regression model is the one with a single auxiliary variable ( $z_i$ ) and an intercept equal to zero (Ford, Kleweno and Tortora, 1980):

$$\hat{y}_i = b_r z_i.$$

But usually the most common regression equation is similar to the following one:

$$\hat{y}_{ij} = b_{ro} + \sum b_{rj} z_{ij} \quad i = 1, 2, \dots, p,$$

where:

- $b_{ro}$  is the fixed effect (that is the intercept),
- $b_{rj}$  is the coefficient estimated for the group  $j$ ,
- $z_{ji}$  is the auxiliary variable for the  $i^{\text{th}}$  unit of the group  $j$ ,
- $e_i = 0$  is the stochastic residual.

Sometimes an interaction term (or more than one) can be included in the regression's equation.

This method is more accurate than the mean imputation, but one of its problem is that is usable only if the missing values are coming from a continuous variable; moreover sometimes an invalid value could be imputed.

---

<sup>134</sup> Nordholt (1998).



#### 6.2.4.20 *Random Regression imputation (RR)*

The random regression imputation (RR) is a stochastic version of the predicted regression method (PR), because it usually adds a residual term ( $e_i$ ) to the predicted values of the PR equation.

As told in Kalton and Kasprzyk (1982), the residual can be:

- Homoschedastic and with a normal distribution (mean equal to zero, variance equal to residual variance from the regression);
- If residuals come from the same unspecified distribution, a random selection from respondents' residuals can be made;
- It's better to select residuals from respondents with similar values regarding the auxiliary variables.

For more details about the RR method, see Schieber (1978), Herzog and Lancaster (1980) and Herzog (1980).

### 6.3 Evaluation criteria

In the previous paragraph the different kinds of imputation methods were presented. Some of them are not applicable to our data: for example we've lost the time linkage of the data, in the generation phase of the simulated population; for this reason we cannot use imputation methods based on the regression or other methods based on the evaluation of the month over month change or on the projection of the observed trend of the values.

Anyway, for the methods we want to apply to our data some evaluation criteria are needed to understand which one fits better to the data we are studying. The main criteria generally available will be discussed in the following paragraphs.

### 6.3.1 Kalton and Kasprzyk methods

Kalton and Kasprzyk (1982) reviewed the effects of six imputation methods<sup>135</sup> on the estimates of means, distributions, variances, covariances, regression and correlation coefficients. They first noted that the stochastic methods give different kind of results depending from the way the residuals ( $e_i$ ) are obtained.

Then they fixed some assumptions:

- respondents are assumed to always respond (“over conceptually repeated applications of the survey”) and non respondents never do this;<sup>136</sup>
- the missing responses are assumed to be randomly distributed (*Missing At Random* – *MAR* assumption);<sup>137</sup>
- the last assumption is that, due to the fact that the sample is large, the sample population factor may be ignored.

The review made by Kalton and Kasprzyk (1982) considers mainly when some standard estimator and their variances must be computed if there are some imputed values.

#### 6.3.1.1 Sample mean

If in a sample there are some imputed observations, the overall mean could be divided into two sub-means: the first one comes from the real values ( $\bar{y}_r$ ), the second one from the imputed values ( $\hat{\bar{y}}$ ). To get the overall mean ( $\bar{y}$ ) is possible to put together the two means with the following formula (Kalton and Kasprzyk, 1982):

$$\bar{y} = \bar{r}\bar{y}_r + \bar{m}\hat{\bar{y}} = \frac{\sum_{k=1}^r y_{rk} + \sum_{i=1}^m \hat{y}_i}{n},$$

---

<sup>135</sup> The six methods considered by category are the following. Deterministic methods: Mean Overall (MO), Mean within classes (MC), Predicted Regression (PR). Stochastic methods: Random Overall (RO), Random within classes (RC), Random Regression (RR).

<sup>136</sup> This is a simplification discussed more deeply by Platek, Singh and Trembaly (1978) and by Platek and Gray (1978, 1979) that introduced a more complex probability response model.

<sup>137</sup> This is considered an unrealistic assumption.

where:

- $\bar{r} = \frac{r}{n}$  is the proportion of actual responses and
- $\bar{m} = \frac{m}{n}$  is the proportion of the imputed responses.

All the six methods considered by Kalton and Kasprzyk (1982) give approximately unbiased estimators with no significant differences between stochastic and deterministic methods<sup>138</sup>.

The bias of the mean and random within classed methods for a variable  $Y$  is computed as follows:

$$B(\bar{y}_{MC}) = B(\bar{y}_{RC}) = \sum_h M_h \frac{\bar{y}_{rh} - \hat{\bar{y}}_h}{N} = B,$$

where:

- $h$  is the imputation class,
- $M_h$  is the number of non respondent,
- $\bar{y}_{rh}$  is the mean of the  $Y$  variable for the respondents of class  $h$ ,
- $\hat{\bar{y}}_h$  is the analogue mean for non respondents,
- $N$  is the size of the population.

If we use the overall imputation methods (MO and RO), the general bias is:

$$B(\bar{y}_{MO}) = B(\bar{y}_{RO}) = \left[ \sum W_h \frac{(\hat{\bar{y}}_h - \bar{y}_r) \cdot (\bar{R}_h - \bar{R})}{\bar{R}} \right] + B = A + B,$$

where:

---

<sup>138</sup> The demonstration is available in Kalton and Kasprzyk (1982, p. 25). It is demonstrated that, if we indicate with the notation  $B(\bar{y}_z)$  the bias of the imputation method  $z$ :

- $B(\bar{y}_{MO}) = B(\bar{y}_{RO})$ ,
- $B(\bar{y}_{MC}) = B(\bar{y}_{RC})$ ,
- $B(\bar{y}_{PR}) = B(\bar{y}_{RR})$ .

- $W_h$  is the proportion of the population in class  $h$ ,
- $\bar{R}_h$  is the response rate in class  $h$ ,
- $\bar{y}_r$  is the overall mean of respondents,
- $\bar{R}$  is the overall response rate.

Thomsen (1973) and Kalton (1981) found the following conclusions:

- When A and B have the same sign, the absolute bias produced by class methods is less of the bias produced by the overall methods of an amount equal to  $|A|$ .
- When A and B have different sign, imputation class methods produce smaller absolute bias only if  $|A| > 2|B|$ .

Kalton and Kasprzyk (1982) also considered the effect of the imputation method (the ones that use auxiliary variables excluded) on the variance of  $\bar{y}$  and noted a loss of precision in  $\bar{y}_{RO}$  from using the stochastic imputation method. According to the results obtained from Kalton and Kish (1981), they underlined the reduction in the imputation variance, due to the exclusion of the multiple employ of donors that could be obtained using SRS rather than unrestricted sampling.

Another way to reduce the imputation's variance is using a proportionate stratifying sampling (by  $Y$  variable) or systematic sampling with respondents ordered by the same variable.

One can reduce the variance also using a larger sample of donors or multiple imputations<sup>139</sup>.

#### 6.3.1.2 *Distribution and variance*

As we saw, the deterministic imputation methods avoid the introduction of imputation variance, but they distort the distributions attenuating the variance. On the other side,

---

<sup>139</sup> In Kalton and Kasprzyk (1982, p. 26) a way to handle with multiple imputations is suggested.

stochastic imputation methods usually give approximately unbiased variances and estimates of distributions, if the missing units are MAR (Missing At Random)<sup>140</sup>.

### 6.3.1.3 Covariance

Kalton and Kasprzyk<sup>141</sup> demonstrate that biases of the covariance between the variable  $Y$  and an additional variable  $X$  ( $s_{xy}$ ) under the stochastic methods and corresponding methods are the same:

- $B(s_{xy_{MO}}) = B(s_{xy_{RO}})$ ;
- $B(s_{xy_{MC}}) = B(s_{xy_{RC}})$ ;
- $B(s_{xy_{PR}}) = B(s_{xy_{RR}})$ .

The conclusion is that  $s_{xy}$  computed with imputed values is, in all the cases, subject to substantial bias, even if we have a MAR model.

But when one uses imputation classes and regression methods, the  $s_{xy}$  estimate is unbiased only if partial covariance  $s_{xy.z}$  is zero (where  $z$  is an auxiliary variable).

If  $x = z$ ,  $s_{xy.z}$  is zero, so  $Z$  should be used as an auxiliary variable in imputing the missing values of  $Y$ , when the covariance between  $X$  and  $Y$  is important.

If we apply imputation considering a simple regression of  $Y$  on  $X$  (when no one  $Y$  and  $X$  value is missing), we get an attenuation of the estimated covariance also of the regression coefficient and if there are more than one independent variable, their relative importance could be biased. Kalton and Kasprzyk suggest to use this  $X$  variable in the imputation scheme to attenuate the covariance. The same method could be used to attenuate the general effect we obtain with imputation on the correlation between the two chosen variable ( $X$  and  $Y$ ); this last can be considered as “*a combination of its effects on the covariance and the standard deviations of the two variables*”<sup>142</sup>. The correlation is usually

---

<sup>140</sup> For a demonstration, see Kalton and Kasprzyk (1982, p. 26).

<sup>141</sup> Kalton and Kasprzyk (1982, p. 26).

<sup>142</sup> Kalton and Kasprzyk (1982).

“overestimated with deterministic imputation methods which employ auxiliary information even when the missing data are MAR”<sup>143</sup>.

For the specific cases where both  $X$  and  $Y$  have missing data, see Kalton and Kasprzyk (1982, p. 27).

#### 6.3.1.4 Standard error estimation

The main risk of using imputation, underlined by Kalton and Kasprzyk (1982, p. 27), is that usually the standard error is computed on a dataset as it was done by all real observed data, but in this way we forget that the variance arises due to the imputation variance. If a dataset has some imputed data, Rubin (1978, 1979) suggests to handle with the problem of the standard error estimation applying the same imputation method several time and computing means on the values obtained at each iteration of the process. The more is the number of iteration, the more the precision of the estimated variance increases.

This aspect suggest to Kalton and Kasprzyk (1982) a feasible way to test the results obtainable with the different imputation methods: if “*the use of multiple imputations reduces the imputation variance [...], multiple imputations may be generated from different imputation procedures, making different assumptions about the nonrespondents. Comparisons of the survey estimates then indicate the sensitivity of the results to the imputation procedures employed*”. This way will be useful, in our further research, to understand which kind of imputation method is more efficient to be applied to data about profit margin ratios.

#### 6.3.2 Other criteria of evaluation

To evaluate the different kind of imputation methods, West, Kratzke and Robertson (1993) propose to consider the error in the imputed value for the unit  $k$  at a certain time  $t$ :

$$\varepsilon_{k,t} = \hat{E}_{k,t} - E_{k,t},$$

where:

---

<sup>143</sup> Kalton and Kasprzyk (1982).

- $\hat{E}_{k,t}$  is the predicted value of the variable we want to impute,
- $E_{k,t}$  is the observed data for the unit  $k$  in the time  $t$ .

It is then possible to summarize this error, within the different strata considered in the survey, in two main ways:

- Percent Relative Error (RE):

$$RE = 100 \cdot \frac{\sum_{k \in \text{stratum}} \mathcal{E}_{k,t}}{\sum_{k \in \text{stratum}} E_{k,t}}$$

- Percent Relative Absolute Error (RAE):

$$RAE = 100 \cdot \frac{\sum_{k \in \text{stratum}} |\mathcal{E}_{k,t}|}{\sum_{k \in \text{stratum}} E_{k,t}}$$

It's also possible to evaluate the mean errors using the presented measures across strata. In this way RE could be considered a macro level statistic, while RAE is a micro level statistics that shows the effect of imputation on each unit.

### 6.3.3 The West, Butani, Witt and Adkins evaluation criteria

West, Butani, Witt and Adkins (1989) used two different kinds of evaluation: the first is based on the computation of the mean unit error, the second one on the mean unit absolute error. For both of the criteria is possible to compute the corresponding relative error. The criteria will be presented in the following.

#### 6.3.3.1 Mean Unit Error

We can define  $\hat{E}_{t,i,m}$  as the error in the prediction for the month  $t$  and the unit  $i$  obtained with a certain imputation procedure  $m$ ; this means that:

$$E_{t,i,m} = (\hat{y}_{t,i,m} - y_{t,i}),$$

where  $\hat{y}_{t,i,m}$  is the predicted data about the month  $t$  and the unit  $i$  obtained with the used imputation procedure  $m$  and  $y_{t,i}$  is the data recorded about the variable object of study for the unit  $i$  and the month  $t$ .

Given these definitions, the Mean Unit Error ( $ME_m$ ) can be computed as:

$$ME_m = \frac{\sum_{\substack{size \\ class}} \sum_t \sum_i E_{t,i,m}}{\sum_{\substack{size \\ class}} \sum_t N_{t,m}^p}$$

where  $N_{t,m}^p$  is the number of respondent units whose have also a value for the previous month,  $t-1$ , and whose are in the domain of imputation procedure  $m$ .

In the study of West, Butani, Witt and Adkins “ $ME_m$  represents a macro level statistic that indicates the effect that the imputation procedure has on total employment”<sup>144</sup> or on the total amount of the considered variable.

### 6.3.3.2 Mean Unit Absolute Error

If we define  $AE_{t,i,m}$  as the absolute error in the prediction for the month  $t$  and the unit  $i$  obtained with the imputation procedure  $m$ :

$$AE_{t,i,m} = |\hat{y}_{t,i,m} - y_{t,i}|,$$

and if we use the same definitions seen for the Mean Unit Absolute Error<sup>145</sup>, the Mean Unit Error ( $MAE_m$ ) can be computed as:

---

<sup>144</sup> West, Butani, Witt and Adkins (1989).

<sup>145</sup>  $\hat{y}_{t,i,m}$  is the predicted data about the month  $t$  and the unit  $i$  (using the imputation procedure  $m$ );  $y_{t,i}$  is the data recorded about the variable object of study,  $Y$ , for the establishment  $i$  and the month  $t$ ;  $N_{t,m}^p$  is the number of respondent units whose have a value for the month ( $t-1$ ) and whose are in the domain of imputation procedure  $m$ .



$$MAE_m = \frac{\sum_{\substack{size \\ class}} \sum_t \sum_i AE_{t,i,m}}{\sum_{\substack{size \\ class}} \sum_t N_{t,m}^p}.$$

In this second criteria,  $MAE_m$  is considered a micro level statistic because it measures the effect that the imputation procedure has on the considered units<sup>146</sup>.

#### 6.3.3.3 Relative Error

Given the previous definitions, the relative error can be calculated as<sup>147</sup>:

$$RE_m = \frac{\sum_{\substack{size \\ class}} \sum_t \sum_i E_{t,i,m}}{\sum_{\substack{size \\ class}} \sum_t \sum_i y_{t,i}} \cdot 100,$$

where:

- $E_{t,i,m} = (\hat{y}_{t,i,m} - y_{t,i})$  is the error in prediction,
- $y_{t,i}$  is the data of unit  $i$  in time  $t$  ;
- the unit  $i$  is included in the set of nonrespondents of procedure  $m$ .

This error measures a macro level statistics that is useful to quantify the effect of imputation on the total amount of the variable.

#### 6.3.3.4 Relative Absolute Error

Given the previous definitions and analogously to what seen in the previous point, we can compute the relative absolute error<sup>148</sup>:

$$RAE_m = \frac{\sum_{\substack{size \\ class}} \sum_t \sum_i AE_{t,i,m}}{\sum_{\substack{size \\ class}} \sum_t \sum_i y_{t,i}} \cdot 100$$

---

<sup>146</sup> West, Butani, Witt and Adkins (1989).

<sup>147</sup> West, Butani, and Witt (1988).

<sup>148</sup> West, Butani, and Witt (1988).

where the unit  $i$  is included in the set of nonrespondents when the the  $m^{\text{th}}$  imputation method is used.  $RAE_m$  is a statistics at a micro-level that gives the affect on the units' data.

## 7 Conclusions

This research project started from the necessity to study the price and its evolutions along the time as a marketing lever to plan the business' strategies. In fact, the knowledge of the price and of its movements is useful to develop a firm's evaluation and planning in terms of comparison with the competitors in a certain market, to enact ad-hoc or more general promotional strategies, and, at least, to define the final price (and, indirectly, the product's or service's image) for the customer. The knowledge of the price movements is also useful to plan the enterprise's oncoming strategic decisions in both a short and a long term perspectives. For all this reason, to have a better view of the state of art of a specific market and, more in general, of a specific sector or economy, it is clear how can be important to have a methodology to measure the prices' movements across time.

In this thesis a solution (that is the price indexes' computation) to measure the variation of prices across time is proposed.

Nevertheless the Statistics Canada's wholesale price survey (implemented in the field of the more general SPPI project) and the elaboration of the first wave's preliminary survey data, brought to face with some data's issues (the presence of missing data and of outlier values, in particular).

For this reason, before the computation of the index using the preliminary collected data, it was considered useful to test some methodological aspects.

The best way to make these tests was considered the generation of a simulated population that would be as close as possible to the observed population (in terms of distribution of the studied variable, that is the profit margin ratio). Moreover, seen the final data are supposed to be released in 2009 and considering the quality improvement process actually on running, no computation could be done on the first preliminary whole collected data (taking

also into consideration the confidentiality challenges); also for this reason there was the necessity to generate and work with a simulated population.

The simulated frame population is also useful because it allows to improve the quality of the survey path and the computation process from the earliest stage, and because it can provide results useable, more in general, also for other kinds of survey.

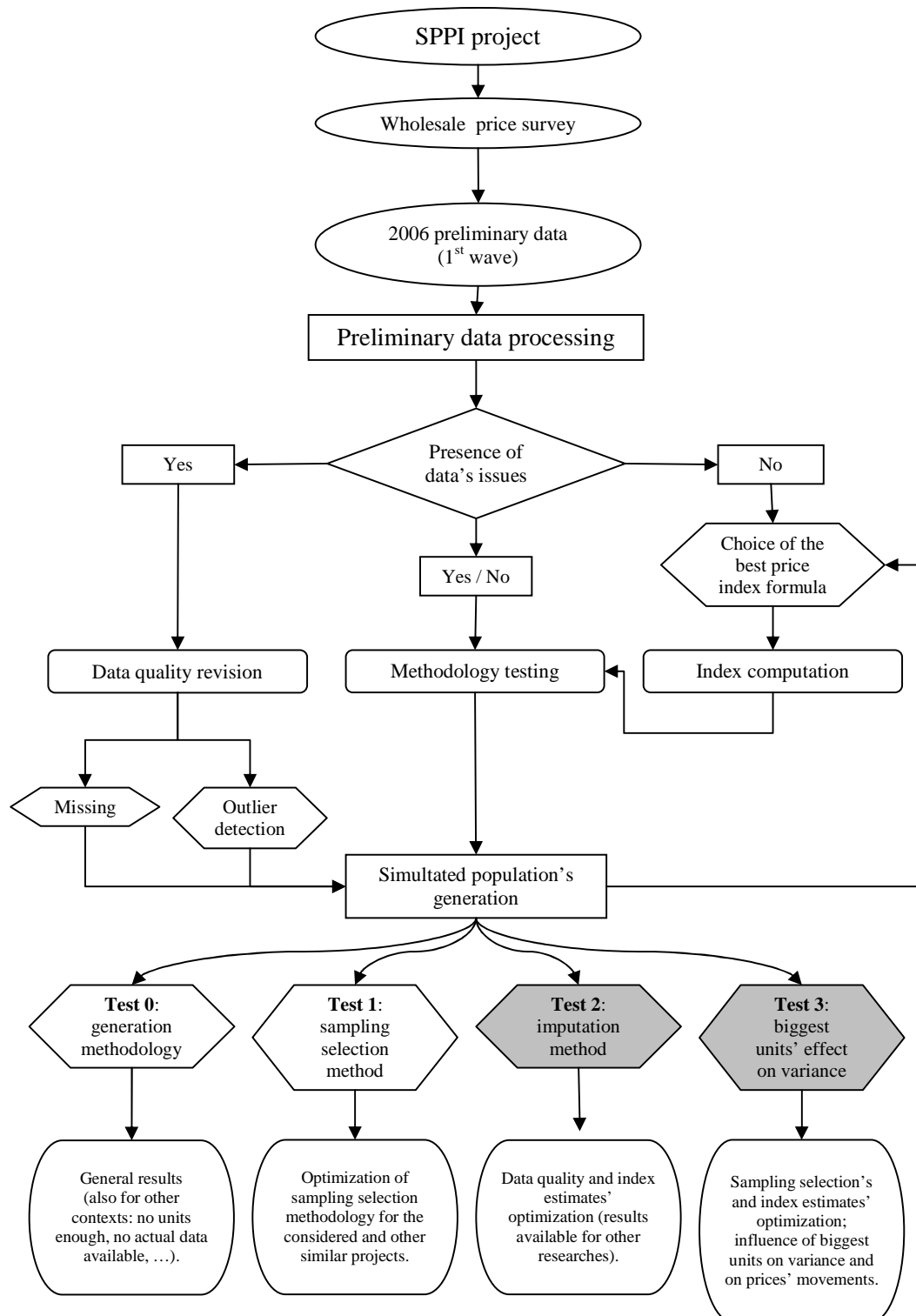
Firstly, the simulated population makes it possible to test the generation methodology itself: it is more efficient to generate a population with a single distribution? Or it's better to simulate a "mixed" distribution, based on the detailed results of statistical tests applied to detailed cells? How is possible to choose the most appropriate available variables to make a stratification? Other useful conclusions of this phase regard the gain in terms of precision that we can obtain using one rather than the other methodology of generation.

All these results can be used for other contexts or projects, where in the studied sample there are no units enough to go on with the elaboration process or where no official data can be used.

Secondly the simulated population can be used to evaluate the comparative efficiency of the sampling selection methodology: this could be useful to understand which is the best sampling selection method (probability proportional to size, simple random sampling, stratified sampling, ...) for a specific field to get a value as close as possible to the real value of an index. What is the gain we can obtain, in term of precision of the index, selecting the population with the various selection methods? Is it useful to select a high number of samples, to get better estimates?

The last two interesting aspect will be studied in further researches. The imputation methodology, introduced in this thesis, will be applied to the simulated data to understand which is the best method to face the problem of missing data and of outlier values. Moreover, applying the different imputation methods it would be possible to evaluate the impact of the biggest units (considered their high influence in the market) not only on the variance of the index (that is a measure of the index' quality), but also on the level of the index itself. This also allows the evaluation of the impact of the prices' policies of the biggest units on the price movements and on the marketing strategies of a certain market (and, in this case in particular, of the wholesalers).


The scheme shown in the following Figure 7.1 can be useful to understand the general structure of the project: the cells underlined in grey will be object of further researches.



**Figure 7.1** – Scheme of the project.

## **APPENDICES**

# APPENDIX # 2.1 - Wholesale price survey questionnaire<sup>149</sup>



Statistics Canada - Prices Division

**WHOLESALE PRICE REPORT**

**Purpose of this survey**

This survey is being conducted to collect prices of representative product and service transactions. The prices you report are essential to the production of indexes measuring the movement of prices for important industries in the Canadian economy as well as for international comparability of productivity, inflation and trade. The resulting indexes are used in developing estimates for real wholesaling output and valuation of imports. In order to enhance the information you provide in this survey, Statistics Canada plans to combine the responses relating to your organization with the information you previously provided on this survey.

**Confidentiality**

Statistics Canada is prohibited by law from publishing any statistics which would divulge information relating to your business without your prior written consent. **The data reported on your questionnaire will be treated in strict confidence, used for statistical purposes and published in aggregate form only.** The confidentiality provisions of the Statistics Act are not affected by the Access to Information Act or by any other legislation.

**CONFIDENTIAL when completed.**

Collected under the authority of the Statistics Act, Revised Statutes of Canada, 1985, Chapter S19. Completion of this questionnaire is a legal requirement under this Act.

**Si vous préférez recevoir ce questionnaire en français veuillez composer le (613) 951-6916.**

**Your Participation is important**

Your participation is vital to ensuring that the information collected in this survey is accurate and comprehensive.

**Fax or Other Electronic Transmission Disclosure**

Statistics Canada advises you that there could be a risk of disclosure during the facsimile or other electronic transmission. However, upon receipt, Statistics Canada will provide the guaranteed level of protection afforded to all information collected under the authority of the Statistics Act.

**Return Procedures.... Need Help?**

We ask that you complete and return this questionnaire within 30 days of receipt. If you require assistance in completing this questionnaire or expect delays in returning the survey please contact:

Statistics Canada - Prices Division  
Tel: 1-888-881-3686  
Fax: 1-888-883-7999  
E-mail: [Kim.Lacroix@statcan.ca](mailto:Kim.Lacroix@statcan.ca)

0001 Legal Name

0002 Business Name

0003 Contact Name

0004 Address


0005 Province

0006 Postal Code/Zip Code

0007 City

0008 Country

STCPR1-420-754/06 2007-02-12

 Statistics Canada

0009

0010

Please make any necessary address changes below

Canada

<sup>149</sup> Source: [www.statcan.gc.ca](http://www.statcan.gc.ca).



**Wholesale Activities for this Business**

**Definition**

A wholesale service is defined as the:

- buying and / or selling of goods on your own account (taking title to goods), or
- engaging in the buying and / or selling, on a commission or fee basis, the goods owned by others

This service may also include secondary activities incidental to the sale of goods including:

- breaking of bulk
- in-store or co-op promotions
- inventory management
- marketing services
- product training
- shipping
- warehousing

**1. Is this business unit primarily a wholesaler (merchant, agent broker, drop shipper, distributor)?**

C0100 1 ☐ Yes

C0101 3 ☐ No, if no, please provide a brief description of your main activity.

C0101

**2. Please provide pricing information for three (3) products identified in the previous quarter.**

C0200 **Product 1:**

C0201 **Product 2:**

C0202 **Product 3:**

**3. Which wholesale activities does this business perform for the product(s) listed above?**

C0301 ☐ breaking of bulk

C0302 ☐ in-store or co-op promotions

C0303 ☐ inventory management

C0304 ☐ marketing services

C0305 ☐ product training

C0306 ☐ shipping

C0307 ☐ warehousing

C0308 ☐ other activity (please specify) C0308TXT

#### 4. Price Information for Product 1

C0400H_1	Data reported for	for:
C0401H_1	Imported:	
C0441H_1	Wholesale average purchase price:	C0441TH_1 Unit of Measure:
C0441AH_1	Wholesale average selling price:	C0441ATH_1 Unit of Measure:

##### Step 1. Is this product currently imported?

C0400\_1 1 O Yes C0401\_1 Country: \_\_\_\_\_  
3 O No

##### Step 2. Please report the average purchase price and selling price per unit for

Month	Average Purchase Price \$CDN Exclude: Only GST and HST	Average Selling Price \$CDN Exclude: GST, HST, PST, TVO and freight	Main reason for any price change (if applicable)
	C0402_1 \$	C0402A_1 \$	C0403_1 1 O Change in supplier 2 O Change in product 3 O Inflation 4 O Exchange rate 9 O Other (specify) C0403TXT_1
	C0402T_1 Unit of Measure	C0402AT_1 Unit of Measure	C0408TXT_1
	C0421_1 \$	C0421A_1 \$	C0427_1 1 O Change in service 2 O Change in customer 3 O Inflation 4 O Exchange rate 9 O Other (specify) C0427TXT_1
	C0421T_1 Unit of Measure	C0421AT_1 Unit of Measure	C0427TXT_1
	C0441_1 \$	C0441A_1	C0447_1 1 O Change in service 2 O Change in customer 3 O Inflation 4 O Exchange rate 9 O Other (specify) C0447TXT_1
	C0441T_1 Unit of Measure	C0441AT_1 Unit of Measure	C0447TXT_1

5. Price Information for Product 2

Data reported for _____ for _____	
CO201H_2 CO401H_2 Imported:	CO441TH_2 Unit of Measure: _____
Wholesale average purchase price: _____	
CO441AH_2 Wholesale average selling price:	CO441AT_2 Unit of Measure: _____

Step 1. Is this product currently imported?

CO400\_2 1 ☐ Yes ☐ No ☐ Country: \_\_\_\_\_

Step 2. Please report the average purchase price and selling price per unit for

Month	Average Purchase Price \$CDN Exclude: Only GST and HST	Average Selling Price \$CDN Exclude: GST, HST, PST, TVQ and freight	Main reason for any price change (if applicable)
	CO402_2 \$	CO402A_2 \$	CO408_2 1 <input type="radio"/> Change in service 2 <input type="radio"/> Change in customer 3 <input type="radio"/> Inflation 4 <input type="radio"/> Exchange rate 9 <input type="radio"/> Other (Specify) _____ CO408TXT_2
	CO402T_2 Unit of Measure	CO402AT_2 Unit of Measure	CO408TXT_2
	CO421_2 \$	CO421A_2 \$	CO427_2 1 <input type="radio"/> Change in service 2 <input type="radio"/> Change in customer 3 <input type="radio"/> Inflation 4 <input type="radio"/> Exchange rate 9 <input type="radio"/> Other (Specify) _____ CO427TXT_2
	CO421T_2 Unit of Measure	CO421AT_2 Unit of Measure	CO427TXT_2
	CO441_2 \$	CO441A_2 \$	CO447_2 1 <input type="radio"/> Change in service 2 <input type="radio"/> Change in customer 3 <input type="radio"/> Inflation 4 <input type="radio"/> Exchange rate 9 <input type="radio"/> Other (Specify) _____ CO447TXT_2
	CO441T_2 Unit of Measure	CO441AT_2 Unit of Measure	CO447TXT_2

### 6. Price Information for Product 3

C0402H_3	Data reported for	for
C0401H_3	Imported:	
C0441H_3	Wholesale average purchase price:	C0441TH_3 Unit of Measure:
C0441AH_3	Wholesale average selling price:	C0441ATH_3 Unit of Measure:

#### Step 1. Is this product currently imported?

C0400\_3 ☐ Yes ☐ No Country: \_\_\_\_\_

#### Step 2. Please report the average purchase price and selling price per unit for

Month	Average Purchase Price \$CDN Exclude: Only GST and HST	Average Selling Price \$CDN Exclude: GST, HST, PST, TVQ and freight	Main reason for any price change (if applicable)
C0402_3 \$	C0402A_3 \$	C0403_3 1 O Change in supplier 2 O Change in product 3 O Inflation 4 O Exchange rate 9 O Other (specify)	C0408_3 1 O Change in service 2 O Change in customer 3 O Inflation 4 O Exchange rate 9 O Other (specify)
C0402T_3 Unit of Measure	C0402AT_3 Unit of Measure	C0403T_3 Unit of Measure	C0408T_3 Unit of Measure
C0401_3 \$	C0401A_3 \$	C0402_3 1 O Change in supplier 2 O Change in product 3 O Inflation 4 O Exchange rate 9 O Other (specify)	C0407_3 1 O Change in service 2 O Change in customer 3 O Inflation 4 O Exchange rate 9 O Other (specify)
C0401T_3 Unit of Measure	C0401AT_3 Unit of Measure	C0402T_3 Unit of Measure	C0407T_3 Unit of Measure
C0401_3 \$	C0401A_3 \$	C0402_3 1 O Change in supplier 2 O Change in product 3 O Inflation 4 O Exchange rate 9 O Other (specify)	C0407_3 1 O Change in service 2 O Change in customer 3 O Inflation 4 O Exchange rate 9 O Other (specify)
C0401T_3 Unit of Measure	C0401AT_3 Unit of Measure	C0402T_3 Unit of Measure	C0407T_3 Unit of Measure

<b>Comments</b>			
We welcome any suggestions that you may have for improving our Wholesale Price Report.			
C9920			
C9913			
C9914			
C9915			
C9916			
C9917			
C9918			
<b>Certification</b> (I certify that the information contained herein is complete and correct to the best of my knowledge.)			
Signature of authorized person		C0016 Date Completed	
<b>Name of person to contact for further information (please print)</b>			
C0015 First Name		C0054 Last Name	
C0014 Title			
C0017 Telephone Number	C0027 Ext.	C0016 Fax No.	C0018 E-mail address
C9909			
<b>Time to complete questionnaire</b>		<b>Minutes</b>	
How long did you spend collecting and reporting the information needed to complete this questionnaire?			
<b>Pre-filled Questionnaire</b>			
In order to facilitate the completion of next quarter's questionnaire, we can provide you with a copy of the information you provided this quarter. Do you authorize us to send a pre-filled questionnaire containing the information you provided this quarter?			
C0003			
Please check		<input type="checkbox"/> YES Please send a pre-filled questionnaire <input type="checkbox"/> NO Please send a blank questionnaire	
Date:		Signature:	
Please make a copy of this completed questionnaire for your records.			
Thank you for completing this questionnaire.			

## APPENDIX # 3.1 - NAICS (*North American Industry Classification Methods*)

Version: 2007<sup>150</sup>

The superscript symbols used to signify comparability are:

- CAN Canadian industry only
- MEX Canadian and Mexican industries are comparable
- US Canadian and United States industries are comparable
- [Blank] [No superscript symbol] Canadian, Mexican and United States industries are comparable.

### 41 Wholesale Trade

#### 411 Farm Product Wholesaler-Distributors <sup>CAN</sup>

##### 4111 Farm Product Wholesaler-Distributors <sup>CAN</sup>

- 41111 Live Animal Wholesaler-Distributors <sup>CAN</sup>
- 411110 Live Animal Wholesaler-Distributors <sup>CAN</sup>
- 41112 Oilseed and Grain Wholesaler-Distributors <sup>CAN</sup>
- 411120 Oilseed and Grain Wholesaler-Distributors <sup>CAN</sup>
- 41113 Nursery Stock and Plant Wholesaler-Distributors <sup>CAN</sup>
- 411130 Nursery Stock and Plant Wholesaler-Distributors <sup>CAN</sup>
- 41119 Other Farm Product Wholesaler-Distributors <sup>CAN</sup>
- 411190 Other Farm Product Wholesaler-Distributors <sup>CAN</sup>

#### 412 Petroleum Product Wholesaler-Distributors <sup>CAN</sup>

##### 4121 Petroleum Product Wholesaler-Distributors <sup>CAN</sup>

- 41211 Petroleum Product Wholesaler-Distributors <sup>CAN</sup>
- 412110 Petroleum Product Wholesaler-Distributors <sup>CAN</sup>

#### 413 Food, Beverage and Tobacco Wholesaler-Distributors <sup>CAN</sup>

##### 4131 Food Wholesaler-Distributors <sup>CAN</sup>

- 41311 General-Line Food Wholesaler-Distributors <sup>CAN</sup>
- 413110 General-Line Food Wholesaler-Distributors <sup>CAN</sup>
- 41312 Dairy and Milk Products Wholesaler-Distributors <sup>CAN</sup>
- 413120 Dairy and Milk Products Wholesaler-Distributors <sup>CAN</sup>
- 41313 Poultry and Egg Wholesaler-Distributors <sup>CAN</sup>

---

<sup>150</sup> Source: Statistics Canada's web site ([www.statcan.gc.ca](http://www.statcan.gc.ca)). For further information, see also Barzyk (2008).

413130	Poultry and Egg Wholesaler-Distributors	CAN
41314	Fish and Seafood Product Wholesaler-Distributors	CAN
413140	Fish and Seafood Product Wholesaler-Distributors	CAN
41315	Fresh Fruit and Vegetable Wholesaler-Distributors	CAN
413150	Fresh Fruit and Vegetable Wholesaler-Distributors	CAN
41316	Red Meat and Meat Product Wholesaler-Distributors	CAN
413160	Red Meat and Meat Product Wholesaler-Distributors	CAN
41319	Other Specialty-Line Food Wholesaler-Distributors	CAN
413190	Other Specialty-Line Food Wholesaler-Distributors	CAN
<b>4132</b>	<b>Beverage Wholesaler-Distributors</b>	CAN
41321	Non-Alcoholic Beverage Wholesaler-Distributors	CAN
413210	Non-Alcoholic Beverage Wholesaler-Distributors	CAN
41322	Alcoholic Beverage Wholesaler-Distributors	CAN
413220	Alcoholic Beverage Wholesaler-Distributors	CAN
<b>4133</b>	<b>Cigarette and Tobacco Product Wholesaler-Distributors</b>	CAN
41331	Cigarette and Tobacco Product Wholesaler-Distributors	CAN
413310	Cigarette and Tobacco Product Wholesaler-Distributors	CAN
<b>414</b>	<b>Personal and Household Goods Wholesaler-Distributors</b>	CAN
<b>4141</b>	<b>Textile, Clothing and Footwear Wholesaler-Distributors</b>	CAN
41411	Clothing and Clothing Accessories Wholesaler-Distributors	CAN
414110	Clothing and Clothing Accessories Wholesaler-Distributors	CAN
41412	Footwear Wholesaler-Distributors	CAN
414120	Footwear Wholesaler-Distributors	CAN
41413	Piece Goods, Notions and Other Dry Goods Wholesaler-Distributors	CAN
414130	Piece Goods, Notions and Other Dry Goods Wholesaler-Distributors	CAN
<b>4142</b>	<b>Home Entertainment Equipment and Household Appliance Wholesaler-Distributors</b>	CAN
41421	Home Entertainment Equipment Wholesaler-Distributors	CAN
414210	Home Entertainment Equipment Wholesaler-Distributors	CAN
41422	Household Appliance Wholesaler-Distributors	CAN
414220	Household Appliance Wholesaler-Distributors	CAN
<b>4143</b>	<b>Home Furnishings Wholesaler-Distributors</b>	CAN
41431	China, Glassware, Crockery and Pottery Wholesaler-Distributors	CAN
414310	China, Glassware, Crockery and Pottery Wholesaler-Distributors	CAN
41432	Floor Covering Wholesaler-Distributors	CAN
414320	Floor Covering Wholesaler-Distributors	CAN
41433	Linen, Drapery and Other Textile Furnishings Wholesaler-Distributors	CAN
414330	Linen, Drapery and Other Textile Furnishings Wholesaler-Distributors	CAN
41439	Other Home Furnishings Wholesaler-Distributors	CAN
414390	Other Home Furnishings Wholesaler-Distributors	CAN
<b>4144</b>	<b>Personal Goods Wholesaler-Distributors</b>	CAN
41441	Jewellery and Watch Wholesaler-Distributors	CAN

414410	Jewellery and Watch Wholesaler-Distributors	CAN
41442	Book, Periodical and Newspaper Wholesaler-Distributors	CAN
414420	Book, Periodical and Newspaper Wholesaler-Distributors	CAN
41443	Photographic Equipment and Supplies Wholesaler-Distributors	CAN
414430	Photographic Equipment and Supplies Wholesaler-Distributors	CAN
41444	Sound Recording Wholesalers	CAN
414440	Sound Recording Wholesalers	CAN
41445	Video Cassette Wholesalers	CAN
414450	Video Cassette Wholesalers	CAN
41446	Toy and Hobby Goods Wholesaler-Distributors	CAN
414460	Toy and Hobby Goods Wholesaler-Distributors	CAN
41447	Amusement and Sporting Goods Wholesaler-Distributors	CAN
414470	Amusement and Sporting Goods Wholesaler-Distributors	CAN
<b>4145</b>	<b>Pharmaceuticals, Toiletries, Cosmetics and Sundries Wholesaler-Distributors</b>	
		CAN
41451	Pharmaceuticals and Pharmacy Supplies Wholesaler-Distributors	CAN
414510	Pharmaceuticals and Pharmacy Supplies Wholesaler-Distributors	CAN
41452	Toiletries, Cosmetics and Sundries Wholesaler-Distributors	CAN
414520	Toiletries, Cosmetics and Sundries Wholesaler-Distributors	CAN
<b>415</b>	<b>Motor Vehicle and Parts Wholesaler-Distributors</b>	CAN
<b>4151</b>	<b>Motor Vehicle Wholesaler-Distributors</b>	CAN
41511	New and Used Automobile and Light-Duty Truck Wholesaler-Distributors	CAN
415110	New and Used Automobile and Light-Duty Truck Wholesaler-Distributors	CAN
41512	Truck, Truck Tractor and Bus Wholesaler-Distributors	CAN
415120	Truck, Truck Tractor and Bus Wholesaler-Distributors	CAN
41519	Recreational and Other Motor Vehicles Wholesaler-Distributors	CAN
415190	Recreational and Other Motor Vehicles Wholesaler-Distributors	CAN
<b>4152</b>	<b>New Motor Vehicle Parts and Accessories Wholesaler-Distributors</b>	CAN
41521	Tire Wholesaler-Distributors	CAN
415210	Tire Wholesaler-Distributors	CAN
41529	Other New Motor Vehicle Parts and Accessories Wholesaler-Distributors	CAN
415290	Other New Motor Vehicle Parts and Accessories Wholesaler-Distributors	CAN
<b>4153</b>	<b>Used Motor Vehicle Parts and Accessories Wholesaler-Distributors</b>	CAN
41531	Used Motor Vehicle Parts and Accessories Wholesaler-Distributors	CAN
415310	Used Motor Vehicle Parts and Accessories Wholesaler-Distributors	CAN
<b>416</b>	<b>Building Material and Supplies Wholesaler-Distributors</b>	CAN
<b>4161</b>	<b>Electrical, Plumbing, Heating and Air-Conditioning Equipment and Supplies Wholesaler-Distributors</b>	CAN
41611	Electrical Wiring and Construction Supplies Wholesaler-Distributors	CAN
416110	Electrical Wiring and Construction Supplies Wholesaler-Distributors	CAN
41612	Plumbing, Heating and Air-Conditioning Equipment and Supplies Wholesaler-Distributors	CAN



416120	Plumbing, Heating and Air-Conditioning Equipment and Supplies Wholesaler-Distributors	CAN
<b>4162</b>	<b>Metal Service Centres</b>	CAN
41621	Metal Service Centres	CAN
416210	Metal Service Centres	CAN
<b>4163</b>	<b>Lumber, Millwork, Hardware and Other Building Supplies Wholesaler-Distributors</b>	CAN
41631	General-Line Building Supplies Wholesaler-Distributors	CAN
416310	General-Line Building Supplies Wholesaler-Distributors	CAN
41632	Lumber, Plywood and Millwork Wholesaler-Distributors	CAN
416320	Lumber, Plywood and Millwork Wholesaler-Distributors	CAN
41633	Hardware Wholesaler-Distributors	CAN
416330	Hardware Wholesaler-Distributors	CAN
41634	Paint, Glass and Wallpaper Wholesaler-Distributors	CAN
416340	Paint, Glass and Wallpaper Wholesaler-Distributors	CAN
41639	Other Specialty-Line Building Supplies Wholesaler-Distributors	CAN
416390	Other Specialty-Line Building Supplies Wholesaler-Distributors	CAN
<b>417</b>	<b>Machinery, Equipment and Supplies Wholesaler-Distributors</b>	CAN
<b>4171</b>	<b>Farm, Lawn and Garden Machinery and Equipment Wholesaler-Distributors</b>	CAN
41711	Farm, Lawn and Garden Machinery and Equipment Wholesaler-Distributors	CAN
417110	Farm, Lawn and Garden Machinery and Equipment Wholesaler-Distributors	CAN
<b>4172</b>	<b>Construction, Forestry, Mining, and Industrial Machinery, Equipment and Supplies Wholesaler-Distributors</b>	CAN
41721	Construction and Forestry Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
417210	Construction and Forestry Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
41722	Mining and Oil and Gas Well Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
417220	Mining and Oil and Gas Well Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
41723	Industrial Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
417230	Industrial Machinery, Equipment and Supplies Wholesaler-Distributors	CAN
<b>4173</b>	<b>Computer and Communications Equipment and Supplies Wholesaler-Distributors</b>	CAN
41731	Computer, Computer Peripheral and Pre-Packaged Software Wholesaler-Distributors	CAN
417310	Computer, Computer Peripheral and Pre-Packaged Software Wholesaler-Distributors	CAN
41732	Electronic Components, Navigational and Communications Equipment and Supplies Wholesaler-Distributors	CAN

417320 Electronic Components, Navigational and Communications Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

**4179 Other Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>**

41791 Office and Store Machinery and Equipment Wholesaler-Distributors <sup>CAN</sup>

417910 Office and Store Machinery and Equipment Wholesaler-Distributors <sup>CAN</sup>

41792 Service Establishment Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

417920 Service Establishment Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

41793 Professional Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

417930 Professional Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

41799 All Other Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

417990 All Other Machinery, Equipment and Supplies Wholesaler-Distributors <sup>CAN</sup>

**418 Miscellaneous Wholesaler-Distributors <sup>CAN</sup>**

**4181 Recyclable Material Wholesaler-Distributors <sup>CAN</sup>**

41811 Recyclable Metal Wholesaler-Distributors <sup>CAN</sup>

418110 Recyclable Metal Wholesaler-Distributors <sup>CAN</sup>

41812 Recyclable Paper and Paperboard Wholesaler-Distributors <sup>CAN</sup>

418120 Recyclable Paper and Paperboard Wholesaler-Distributors <sup>CAN</sup>

41819 Other Recyclable Material Wholesaler-Distributors <sup>CAN</sup>

418190 Other Recyclable Material Wholesaler-Distributors <sup>CAN</sup>

**4182 Paper, Paper Product and Disposable Plastic Product Wholesaler-Distributors <sup>CAN</sup>**

41821 Stationery and Office Supplies Wholesaler-Distributors <sup>CAN</sup>

418210 Stationery and Office Supplies Wholesaler-Distributors <sup>CAN</sup>

41822 Other Paper and Disposable Plastic Product Wholesaler-Distributors <sup>CAN</sup>

418220 Other Paper and Disposable Plastic Product Wholesaler-Distributors <sup>CAN</sup>

**4183 Agricultural Supplies Wholesaler-Distributors <sup>CAN</sup>**

41831 Agricultural Feed Wholesaler-Distributors <sup>CAN</sup>

418310 Agricultural Feed Wholesaler-Distributors <sup>CAN</sup>

41832 Seed Wholesaler-Distributors <sup>CAN</sup>

418320 Seed Wholesaler-Distributors <sup>CAN</sup>

41839 Agricultural Chemical and Other Farm Supplies Wholesaler-Distributors <sup>CAN</sup>

418390 Agricultural Chemical and Other Farm Supplies Wholesaler-Distributors <sup>CAN</sup>

**4184 Chemical (except Agricultural) and Allied Product Wholesaler-Distributors <sup>CAN</sup>**

41841 Chemical (except Agricultural) and Allied Product Wholesaler-Distributors <sup>CAN</sup>

418410 Chemical (except Agricultural) and Allied Product Wholesaler-Distributors <sup>CAN</sup>

**4189 Other Miscellaneous Wholesaler-Distributors <sup>CAN</sup>**

41891 Log and Wood Chip Wholesaler-Distributors <sup>CAN</sup>

418910 Log and Wood Chip Wholesaler-Distributors <sup>CAN</sup>

41892 Mineral, Ore and Precious Metal Wholesaler-Distributors <sup>CAN</sup>

418920 Mineral, Ore and Precious Metal Wholesaler-Distributors <sup>CAN</sup>

41893 Second-Hand Goods (except Machinery and Automotive) Wholesaler-Distributors

	<small>CAN</small>	
418930	Second-Hand Goods (except Machinery and Automotive) Wholesaler-Distributors	
	<small>CAN</small>	
41899	All Other Wholesaler-Distributors	<small>CAN</small>
418990	All Other Wholesaler-Distributors	<small>CAN</small>
<b>419</b>	<b>Wholesale Electronic Markets, and Agents and Brokers</b>	<small>US</small>
<b>4191</b>	<b>Wholesale Electronic Markets, and Agents and Brokers</b>	<small>US</small>
41911	Business-to-Business Electronic Markets	<small>US</small>
419110	Business-to-Business Electronic Markets	<small>US</small>
41912	Wholesale Trade Agents and Brokers	<small>US</small>
419120	Wholesale Trade Agents and Brokers	<small>US</small>

## APPENDIX # 3.2 - Distribution of the sample by cells

Distribution of the sample (37,873 units) by cells. The cells are defined using the NAICS code (4 digits version) and revenue classes (defined by deciles, for each cell).

The units with a missing profit margin ratio are in the last column (missing data).

NAICS 4 digits	Rev. classes	1	2	3	4	5	6	7	8	9	10	Units	Missing	TOT UNITS
4111	% Row	7.4	11.1	16.0	10.4	13.2	8.8	11.6	9.5	5.3	6.7	100.0	26.7	100.0
4121	% Row	8.2	13.1	9.0	13.9	9.0	9.0	9.8	12.3	0.8	14.8	100.0	35.8	100.0
4131	% Row	9.3	9.9	9.5	9.9	10.5	9.5	10.4	10.3	10.5	10.2	100.0	15.5	100.0
4132	% Row	7.4	7.4	14.8	7.4	9.4	9.4	11.4	10.8	7.2	14.8	100.0	10.3	100.0
4133	% Row	2.5	2.5	2.5	25.3	2.5	5.1	12.7	5.1	13.9	27.8	100.0	15.1	100.0
4141	% Row	9.8	9.5	9.7	9.5	10.1	10.5	9.0	11.3	9.8	10.9	100.0	13.8	100.0
4142	% Row	7.9	9.9	11.1	11.0	10.7	9.3	9.9	9.0	10.5	10.7	100.0	10.0	100.0
4143	% Row	9.1	10.5	9.9	9.6	9.0	9.8	10.3	10.3	10.8	10.7	100.0	9.8	100.0
4144	% Row	9.9	9.6	9.5	9.5	9.5	10.2	10.5	10.1	10.3	10.9	100.0	8.6	100.0
4145	% Row	9.6	10.4	10.0	9.1	10.6	9.5	10.2	10.2	9.7	10.7	100.0	3.0	100.0
4151	% Row	8.1	9.6	10.4	9.1	9.6	10.4	9.4	10.9	11.3	11.3	100.0	9.2	100.0
4152	% Row	10.1	9.8	9.9	10.6	9.6	9.0	9.0	11.0	10.3	10.7	100.0	8.0	100.0
4153	% Row	5.7	9.7	10.1	10.4	10.4	8.2	8.2	11.3	9.7	16.4	100.0	9.4	100.0
4161	% Row	9.4	9.5	10.8	10.6	8.6	10.6	9.6	10.2	9.3	11.4	100.0	7.8	100.0
4162	% Row	8.5	11.3	9.9	9.1	9.3	10.8	9.0	11.3	10.9	9.9	100.0	12.4	100.0

*Continues on the next page*

Continues from the previous page

NAICS 4 digits	Rev classes	1	2	3	4	5	6	7	8	9	10	Units	Missing	TOT UNITS
4163	% Row	8.7	10.0	9.8	10.9	10.4	10.9	9.7	9.3	10.0	10.3	100.0	10.1	100.0
4171	% Row	10.1	8.5	8.7	11.1	7.9	10.2	11.2	10.6	7.9	13.8	100.0	11.4	100.0
4172	% Row	9.3	10.3	10.1	10.2	10.1	9.4	9.1	10.7	10.2	10.5	100.0	10.5	100.0
4173	% Row	8.7	10.3	10.4	10.6	10.3	10.9	10.0	9.7	9.6	9.7	100.0	11.8	100.0
4179	% Row	9.2	9.6	9.8	9.9	10.0	10.6	9.8	10.0	10.1	11.0	100.0	7.9	100.0
4181	% Row	8.0	9.5	3.7	12.3	10.1	8.2	12.9	12.7	9.7	12.9	100.0	19.6	100.0
4182	% Row	3.8	7.7	7.7	16.2	4.6	6.9	17.7	7.7	8.5	19.2	100.0	15.6	100.0
4183	% Row	9.1	10.1	10.6	12.5	10.3	11.8	11.5	6.5	7.9	9.8	100.0	30.6	100.0
4184	% Row	4.2	11.5	11.5	10.1	14.3	9.8	2.1	10.1	6.3	20.2	100.0	18.5	100.0
4189	% Row	7.9	12.0	9.5	11.1	8.7	10.3	8.4	9.9	9.3	13.0	100.0	18.6	100.0
TOTAL	Units	3046	3364	3362	3424	3319	3387	3309	3462	3333	3715	33721	4152	37873



## REFERENCES

- Advisory Commission to Study the Consumer Price Index (1996), *Toward a More Accurate Measure of the Cost of Living* – *The Boskin Commission Report*.
- Afifi A. A., Elaskoff R. M. (1969), *Missing Observation in Multivariate Statistics III: Large Sample Analysis of Simple Linear Regression*, in *Journal of the American Statistical Association*, vol. 64, 337-358.
- Allen J.D. (1990), *An overview of imputation procedures*, Staff Report SMB-90-06, US Department of Agriculture, Washington, D.C..
- Amemiya T. (1984), *Tobit Models: A Survey*, in “*J. Econometrics*”, 24 (1984), 3-61.
- Anderson T. W., Darling D. A. (1954), “*A Test of Goodness-of-Fit*” in *Journal of the American Statistical Association*, 49, 765-769.
- Bailar III J.C., Bailar B.A. (1978), *Comparison of two procedures for imputing missing survey values*, in “*Proc. Sect. Survey Research Methodology - American Statistical Association*”, 462-467, American Statistical Association.
- Bailar B.A., Bailar III J.C. (1979), *Comparison of the biases of the “hot-deck” imputation procedure with an “equal-weights” imputation procedure*, in “*Symposium on Incomplete Data: Preliminary Proceedings (Panel on Incomplete Data of the Committee on National Statistics/National Research Council)*”, 422-447, U.S. Department of Health, Education and Welfare, Washington, D.C..
- Bailar B.A., Bailey L., Corby C.A. (1978), *A comparison of some adjustment and weighting procedures for survey data*, in “*Survey Sampling and Measurement*”, Namboodiri, N.K. ed., 175-198, Academic Press, New York.

- Barzyk F. (2008), *SPPI for Wholesale Services in Canada*, 23<sup>rd</sup> Vooburg Group Meeting, September 22<sup>nd</sup>-26<sup>th</sup> 2008, Aguascalientes, Mexico (paper available at the following link: <http://www4.statcan.ca/english/vooburg/Documents/2008%20Aguascalientes/Papers/2008%20-%2019.pdf>; presentation available at the following link: [http://www.inegi.gob.mx/vooburg/docs/WholesaleSPPI\(CANADA\).ppt](http://www.inegi.gob.mx/vooburg/docs/WholesaleSPPI(CANADA).ppt)).
- Bérard H., Pursey S., Rancourt E. (2005), *Re-thinking Statistics Canada's Business Register*, Statistics Canada (paper available at the following link: [http://www.fcsn.gov/05papers/Rancourt\\_Berard\\_Pursey\\_IVB.pdf](http://www.fcsn.gov/05papers/Rancourt_Berard_Pursey_IVB.pdf)).
- Brooks C.A., Bailar B.A. (1978), *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3, U.S. Department of Commerce, U.S., Government Printing Office, Washington, D.C..
- Busacca B. (1994 ), *Le risorse di fiducia dell'impresa*, Utet, Torino.
- Busacca B., Costabile M., Pasini P. (1993), *Decidere il prezzo*, Etas, Milan.
- Castonguay E., Monty A. (2000), *Recent Developments in the Statistics Canada Business Register*, Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association, 61-66.
- Carù A., Cugini A. (2000 ), *Valore per il cliente e controllo dei costi*, Egea, Milan.
- Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A. (1983), *Graphical Methods for Data Analysis*, Wadsworth International Group, Belmont, Calif..
- Chapman D.W., Bailey L., Kasprzyk D. (1986), *Nonresponse adjustment procedures at the U.S. Bureau of the Census*, in "Survey Methodology", 12, 161-179.
- Chromy J.R. (1978), "Sequential Sample Selection Methods", in *Proceedings of the American Statistical Association*, Survey Research Methods Section, 401-406.
- Cochran W.G. (1977), *Sampling Techniques*, John Wiley & Sons, New York.
- Coder J. (1978), *Income data collection and processing from the March Income Supplement to the Current Population Survey*, in "The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing", February 23-24 1978,



- D. Kasprzyk ed., Chapter II, U.S. Department of Health, Education and Welfare, Washington, D.C..
- Cohen A. C. (1951), "*Estimating Parameters of Logarithmic-Normal Distributions by Maximum Likelihood*", in *Journal of the American Statistical Association*, 46, 206-212.
- Colledge M. J. (1987), *The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada*, Proceedings of the Third Annual Research Conference, Bureau of the Census, 550-576.
- Colledge M.J., Johnson J.H., Pare R., Sande, I.G. (1978), *Large scale imputation of survey data*, Proc. Section Survey Research Meth., American Statistical Association, 1978, 431-436.
- Costabile M. (1992), *Prezzo e consumatore. Il ruolo del prezzo nel processo d'acquisto*, Egea, Milan.
- Costabile M. (1996), *Misurare il valore per il cliente. Aspetti metodologici e implicazioni per la gestione dei processi di scambio*, Utet, Turin.
- Costabile M. (2001), *Il capitale relazionale. La gestione delle relazioni e della customer loyalty*, McGraw-Hill, Milan.
- Costabile M. (2003), *Le ricerche per la definizione e il controllo del prezzo dei prodotti*, in Molteni L., Troilo G. (2003), *Ricerche di Marketing*, McGraw-Hill, Milan.
- Consumer Price Index reference paper, The* (1995) Statistics Canada, Price Division – Minister of Industry, Ottawa, Canada.
- Cuthill I. (1990), *The Statistics Canada Business Register*, Internal document, Informatics Branch, Statistics Canada. Revised: 1997.
- D'Agostino R.B., Stephens M. (1986), *Goodness-of-Fit Techniques*, Marcel Dekker, Inc., New York.
- Dixon W.J. (1983), *BMDP Statistical Software* (3<sup>rd</sup> ed.), Berkeley, CA, University of California Press.

- Dolan R.J., Simon H. (1996), *Power Pricing. How Managing Price Transforms the Bottom Line*, The Free Press, New York.
- Elandt R. C. (1961), "*The Folded Normal Distribution: Two Methods of Estimating Parameters from Moments*", in *Technometrics*, 3, 551-562.
- Fishman G.S., Moore L.R. (1982), "A Statistical Evaluation of Multiplicative Congruential Generators with Modulus ( $2^{31} - 1$ )", in *Journal of the American Statistical Association*, 77, 1 29-136.
- Ford B. (1976), "*Missing data procedures: a comparative study*", in *Proc. Sect. Soc. Stat.*, Amer. Stat. Assoc., 1976, 324-329.
- Ford B. (1980), "*An overview of hot deck procedures*", Draft paper for *Panel on Incomplete Data*, Committee on National Statistics, National Academy of Sciences.
- Ford B. L., Kleweno D. G., Tortora L. D. (1980), *The effects of the procedures which impute for missing items: a simulation study using an agricultural survey*, *Proc. Sect. Survey Res. Meth.*, American Statistical Association, 1980, 251-256.
- Fox D.R. (1989), "*Computer Selection of Size-Biased Samples*", in *The American Statistician*, 43(3), 168-171.
- Gagné P. (2004), *Projet de Refonte du Registre des Entreprises*, Internal document, Statistics Canada.
- Golmant J. (1990), "*Correction: Computer Selection of Size-Biased Samples*", in *The American Statistician*, 44(2), 194.
- Greenberg B.G., Surdy R. (1984), "*A Flexible and Interactive Edit and Imputation System for Ratio Edits*", in *SRD report RR-84/18*, U.S. Bureau of the Census, Washington, D.C..
- Greenlees W.S., Reece J.S., Zieschang K.D. (1982), *Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed*, in "*Journal of the American Statistical Association*", Vol. 77, 251-261.
- Hanson R.H. (1978), *The current population survey, design and methodology*, Technical Paper 40, Bureau of the Census, Washington, DC 20233.

- Hanurav T.V. (1967), “*Optimum Utilization of Auxiliary Information: Sampling of Two Units from a Stratum*”, in *Journal of the Royal Statistical Society, Series B*, 29, 374-391.
- Hausman J.A., Wise D.A. (1977), *Social Experimentation, Truncated Distributions, and Efficient Estimation*, in “*Econometrica*”, 45, 919-938.
- Heckman J. (1976), *The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models*, in “*Annals of Economic and Social Measurement*”, 5, 475–492.
- Herzog T. N. (1980), *Multiple Imputation of Individual Social Security amounts*, Part II, Proc. Sect. Survey Res. Meth., American Statistical Association, 1980, 404-407.
- Herzog T. N., Lancaster C (1980), *Multiple imputation of individual Social Security amounts*, Part I, Proc. Sect. Survey Res. Meth., American Statistical Association, 1980, 398-403.
- Hidioglou M.A. (1986), *The construction of a self-representing stratum of large units in survey design*, in “*The American Statistician*”, 40, 27-31.
- International Monetary Fund (2004), *Produce Price Index Manual – Theory and Practice*, International Monetary Fund, Washington, D.C. (<http://www.imf.org>).
- Johnson N. L., Kotz S., Balakrishnan N. (1994), *Continuous Univariate Distributions - 1*, Second Edition, John Wiley & Sons Inc., New York.
- Johnson N. L., Kotz S., Balakrishnan N. (1995), *Continuous Univariate Distributions - 2*, Second Edition, John Wiley & Sons Inc., New York.
- Kalton G. (1981), “*Compensating for Missing Survey Data*”. Income Survey Development Program, Department of Health and Human Services report. Survey Research Center, University of Michigan.
- Kalton G. (1983a), *Compensating for missing survey data*, in “*Research Report Series*”, Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor, Michigan.

- Kalton G. (1983b), *Introduction to Survey Sampling*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035, Sage Publications Inc., Beverly Hills, CA and London.
- Kalton G., Kasprzyk D. (1982), *Imputing for Missing Survey Responses*, in “*Proceedings of the Section on Survey Research Methods*”, American Statistical Association, 1982, 22-31.
- Kalton G., Kasprzyk D. (1986), *The treatment of missing survey data*, in “*Survey Methodology*”, 12, 1-16.
- Kalton G., Kish L. (1981), *Two efficient random imputation procedures*, Proc. Sect. Survey Res. Meth., American Statistical Association, 1981, 146-151.
- Kish L. (1965), *Survey Sampling*, John Wiley & Sons Inc., New York.
- Kish L. (1987), *Statistical Design for Research*, John Wiley & Sons Inc., New York.
- Krishnamurthi L. (2001), *Pricing Strategies and Tactics*, in “*Kellogg on Marketing*” edited by Iacobucci D., Wiley & Sons, New York, 270-301.
- Kush G.L., Clark D.F. (1979), *Annual Survey of Manufactures General Statistics Edit*, in “*Proceedings of the Business and Economic Statistics Section*”, American Statistical Association, 183-187.
- Laspeyres E. (1884), *Hamburger Warenpreise 1851-1860 und die californisch-australischer Soldentdeckung seit 1848, Ein Beitrag zur Lehre von der Geldentwertung*, in “*Jahrbücher für Nationalökonomie*”.
- Lavallée P., Hidirolou M.A. (1988), *On the stratification of skewed populations*, Survey Methodology, Vol. 14, No. 1, 33-43.
- Lawrence, R.E. (1980) Variance of the Estimated Mean for Several Imputation Procedures, U.S. Bureau of the Census.
- Lillard L., Smith J.P., Welch F. (1982), *What do we really know about wages: the importance of nonreporting and census imputation*, technical report, the Rand Corporation, 1700 Main Street, Santa Monica, California.

- Lillard L., Smith J.P., Welch F. (1986), *What do we really know about wages? The importance of nonreporting and census imputation*, in “*Journal of Political Economy*”, 94, 489-506.
- Little R.J.A. (1982), *Models for nonresponse in sample survey*, in “*Journal of American Statistical Association*”, 77, 237-250.
- Little, R.J.A. (1986), *Survey Nonresponse Adjustments for Estimates of Means*, in “*International Statistical Review*”, 54, 139-157.
- Little R.J.A. (1988), *Missing-Data Adjustments in Large Survey*, in “*Journal of Business & Economic Statistics*”, July Vol. 6, No. 3, 287-296.
- Little R.J.A., Rubin D.B. (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons Inc., New York.
- Little R.J.A., Smith P.J. (1983), *Multivariate Edit and Imputation for Economic Data. Contributed paper to the Survey, Methods Section of the 143rd Annual Meeting of the American Statistical Association.*
- Little R.J.A., Smith P.J. (1987), *Editing and imputation for quantitative data*, in “*Journal of the American Statistical Association*”, 82, 58-69.
- Madow W.G. (1949), “*On the Theory of Systematic Sampling, II*”, in *Annals of Mathematical Statistics*, 20, 333-354.
- Madow W.G., Nisselson H., Olkin I., Rubin D.B. (1983), *Incomplete data in sample survey*, vol. 1, 2, 3, Academic Press, New York.
- Monroe K.B. (2002), *Pricing. Making Profitable Decision*, McGraw-Hill, New York (first edition: 1979).
- Montaquila J.M., Ponikowski C.H. (1993), *Comparison of methods for imputing missing response in an establishment survey*. Proceedings of the Survey Research Methods Section of the American Statistical Association, 446-451.

- Montaquila J.M., Ponikowski C.H. (1995), *An evaluation of alternative imputation methods*. Proceedings of the Survey Research Methods Section of the American Statistical Association, 281-286.
- Molteni L., Troilo G. (2003), *Ricerche di Marketing*, McGraw-Hill, Milan.
- Mueller K., Stamas G., Butani S. (1995), *Nonresponse adjustment in certainty strata for an establishment survey*, in *Proceedings of the Section on Survey*, American Statistical Association.
- Murthy M.N. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, India.
- Nagle T.T., Holden R.K. (1995), *The Strategy and Tactics of Pricing. A guide to profitable decision*, Prentice Hall, Englewood Cliffs.
- Nordholt E. S. (1998), *Imputation: Methods, Simulation, Experiments and Practical Examples*, in "International Statistical Review", Vol. 66, 2, 157-180, International Statistical Institute, Mexico.
- Oh H.L., Scheuren F. (1980), *Estimating the variance impact on missing CPS income data*, Proceedings of the Survey Research Methods Section, American Statistical Association, 1980, 408-415.
- Oh H.L., Scheuren F., Nisselson H. (1980), *Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data*, Proceedings of the Survey Research Methods Section, American Statistical Association, 1980, 416-420.
- Ohlsson E. (1990), "Sequential Poisson Sampling from a Business Register and Its Applications to the Swedish Consumer Price Index", in *R&D Report*, n. 6, Statistics Sweden, Stockholm.
- Ohlsson E. (1998), "Sequential Poisson Sampling", in *Journal of Official Statistics*, 14, n. 2, pp. 149-162.

- Patak Z., Lothian J. (2007), *Enhancing the Quality of Price Indexes*, Third International Conference on Establishment Surveys - Survey Methods for Businesses, Farms, and Institutions (ICES-III), June 18–21, 2007, Montréal, Québec, Canada.
- Patak Z., Rais S. (2005), *Survey methodology for new business services price indexes*, Third International Conference on Establishment Surveys - Survey Methods for Businesses, Farms, and Institutions (ICES-III), June 18-21, 2007, Montréal, Québec, Canada.
- Platek R., Gray G.B. (1978), *Nonresponse and Imputation*, In “*Survey Methodology*”, 6, 93-132.
- Platek R., Gray G.B. (1979), *Methodology and application of adjustment for nonresponse*, ISI, 42 Session, Manila.
- Platek R., Singh M.P., Tremblay V. (1978) *Adjustment for Nonresponse*, in “*Survey Sampling and Measurement*”, p. 157-74, ed. N. K. Namboodiri, Academic Press, New York.
- Predetti A. (2006), *I Numeri Indici – Teoria e pratica dei confronti temporali e spaziali*, 11° ed., Giuffrè, Milan.
- QAF (*Quality Assessment Framework*). Statistic Canada website: [www.statcan.gc.ca](http://www.statcan.gc.ca).
- Romani S. (2000), *L’analisi del comportamento del cliente per la determinazione del prezzo di vendita di beni e servizi*, FrancoAngeli, Milan.
- Rubin D.B. (1978), *Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse* in “*Proc. Sect Survey Research Meth.*”, American Statistical Association, 20-34.
- Rubin D.B. (1979), *Illustrating the use of multiple imputations to handle nonresponse in sample surveys* in “*Bull. Int. Statist. Inst.*”.
- Sande I.G. (1979a), *A personal view of hot deck imputation procedures*, in “*Survey Methodology*”, 5, 238-258

- Sande I.G. (1979b), *Hot deck imputation procedures*. Symposium on Incomplete Data: Preliminary Proceedings (Panel on Incomplete Data of the Committee on National Statistics/National Research Council), 484-498, U.S. Department of Health, Education and Welfare, Washington, D.C.
- Särndal C.E., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Santos R.L. (1981a), *Effect of Imputation on Complex Statistics*, Survey Research Center, University of Michigan, Ann Arbor.
- Santos R.L. (1981b), *Effect of Imputation on Regression Coefficients*, Proc. Sect. Survey Res. Meth., American Statistical Association, 140-145.
- SAS Institute Inc. (1999), *SAS/INSIGHT User's Guide*, version 8, SAS Institute Inc., Cary, NC.
- SAS Institute Inc. (2004), *SAS OnlineDoc® 9.1.3.*, SAS Institute Inc, Cary, NC.
- Schieber S.J. (1978), *A comparison of three alternative techniques for allocating unreported Social Security Income on the Survey of the Low-Income Aged and Disabled*, in Proc. Sect. Survey Res. Meth., American Statistical Association, 1978, 212-218.
- Simon H. (1989), *Price Management*, North Holland, Amsterdam.
- Statistics Canada's web site: [www.statcan.gc.ca](http://www.statcan.gc.ca).
- Stephens M. A. (1974), *"EDF Statistics for Goodness of Fit and Some Comparisons"*, in *Journal of the American Statistical Association*, 69, 730-737.
- Thomsen I. (1973), *A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data*, Statistik Tidskrift, 16, 191-196.
- Vacek P.M., Ashikaga T. (1980), *An examination of the Nearest Neighbour Rule for imputing missing values*, in *"Proceedings of the Statistical Computing Section"*, American Statistical Association, 326-331.



- Valdani E. (1989), *Pricing. Tattiche e strategie per definire con successo il prezzo di vendita*, Etas, Milano.
- Valdani E. (1995), *Marketing Strategico*, Etas, Milano.
- Vijayan K. (1968), “An Exact Sampling Scheme: Generalization of a Method of Hanurav”, in *Journal of the Royal Statistical Society, Series B*, 30, 556-566.
- Watts D.L. (1991), “Correction: Computer Selection of Size-Biased Samples”, in *The American Statistician*, 45(2), 172.
- Welniak E.J., Coder J.F. (1980), *A measure of the bias in the March CPS earnings imputation system*, in “*Proc. Sect. Survey Res. Meth.*”, American Statistical Association, 1980, 421-425.
- West S.A. (1982), *Linear Models for Monthly All Employment Data*, Bureau of Labor Statistics Report.
- West S. A. (1983), *A Comparison of Different Ratio and Regression Type Estimators for the Total of a Finite Population*, Proceedings of the Section in Survey Research Methods, American Statistical Association, 388-393.
- West S., Butani S., Witt M. (1988), *Alternative Imputation Methods for Wage Data*, Bureau of Labor Statistics, Washington DC, 254-259.
- West S., Butani S., Witt M., Adkins C. (1989), *Alternative Imputation Methods for Employment Data*, in “*Proceedings of the Survey Research Methods Section*”, Washington D.C., American Statistical Association, 227-232.
- West S.A., Kratzke D.T., Robertson K.W. (1993), *Alternative Imputation Procedures for Item Non-response from New Establishments in the Universe*, in “*Proceedings of the Survey Research Methods Section*”, American Statistical Association, 1993.
- Williams R.L., Chromy J.R. (1980), “SAS Sample Selection Macros”, in *Proceedings of the Fifth Annual SAS Users Group International Conference*, 5, 392-396.

My grateful thanks go to:

***Silvia***

*(who made this possible),*

***Sylvie***

*(who made this great),*

***Zdenek***

*(who made this interesting and enjoyable)...*

*... and to **all the friends and colleagues I met in***

*“the True North strong and free” Country*

*where I spent some of the most memorable and fantastic months of my life.*

*“...from far and wide,*

*o Canada, we stand on guard for thee.”*

*Daniele*