Data Analysis at Statistics Canada

(Message from the Chief Statistician, Ivan Fellegi, delivered at the First Award Ceremony for the Course on Principles of Data Analysis and Presentation, April 5 1993.)

Good morning!

I am seized by the importance of this occasion which is a historic moment in the annals of Statistics Canada. We are gathered in a room named after Simon Goldberg, a former Assistant Chief Statistician, whose intellectual legacy has influenced many of us. One of his characteristics was to have an analytic approach to everything he did and to encourage analytic thinking among those around him.

I would like to congratulate all of you on your courage - and I use this term advisedly - in undertaking this voyage of discovery on which you have embarked, and in successfully completing this course devoted to analysis.

I looked up the word "analysis" in the *Dictionnaire Robert* and this is what I found (translated): an intellectual exercise consisting of the decomposition of a work, or a text into its essential elements, so as to understand relationships and show the structure of the whole. The *Oxford English Dictionary* gives an even briefer definition: "to find or show the essence or structure of something." I find those definitions crystallize my own thoughts.

I want to give you all my reasons - nine of them, presented in no particular order - why I believe analysis is so important at Statistics Canada. **I am only talking about interpretative analysis**. I could add another fifteen reasons to emphasize the importance of methodological analysis.

1. First, there are certain types of analysis that can only be performed within the Bureau because of the need for access to microdata. If not by us, then these types of analysis will not be done at all. This includes all sorts of simulations using confidential microdata, as well as record linkages.

2. The second kind of analysis is the development of statistical products where the actual output is an analytical construct - not just a method. The best known example is the National Accounts. There is no such observable thing as a Gross Domestic Product - it is a concept that has been formulated. It is a very useful concept - an analytic concept – but it is not something that can be seen, touched, or consumed. Components of it can, but the Gross Domestic Product cannot.

Micro-simulation models are constructs based on analysis that other people can exploit for the production of information or for research. Micro-simulation models are developed at Statistics Canada to enable people outside the Bureau to assess the impact of alternative scenarios. For example:

- what would be the impact of a particular social policy on income distributions for different types of families?

- what would be the impact of such a policy on the fiscal situation of the federal or provincial governments?

The Consumer Price Index is another example of an analytical construct: the Consumer Price Index is not the change in the price of anything we can observe.

It is easy to say what these analytical constructs represent. Their creation, however, requires the kind of understanding that the dictionary was talking about: the decomposition of something into its elements and the understanding of the relationships between them in order to have a deeper understanding of the underlying essence or structure.

3. The third reason why analysis is important in Statistics Canada is because it is through analysis that we can assess the quality and consistency of the data we produce. We look to internal analysis to provide feedback to our own production process; to identify and resolve data problems before a product is released.

One can distinguish categories of quality assessment. Internal consistency checks are a type of analysis based on known or assumed relationships. Checking for consistency is analytic work designed to reveal unexpected deviations from this known or assumed relationship. Such deviations can highlight potential errors in the data. Alternatively, the data might not be in error, in which case the changing relationships represent important findings. So looking for consistency and for the patterns that one anticipates finding is a part of quality assurance. It is also part of the discovery of the meaning of the data. It is something that ought to be a feature of everything we do. It is the ultimate stage of quality assurance for both cross-sectional data and time series data.

A related type of quality assurance consists of comparing data from different sources, not just data from the same survey or process. There again, one looks simultaneously for quality assurance and the discovery of new insights. One also looks for consistency between data series that ought to be consistent, in order to find problems in their classification or treatment.

A very widely practiced kind of consistency test that we apply to our economic statistics occurs in their integration within the System of National Accounts. The relationships are not only formulated in terms of our prior expectations but are very often technical equations - identities which, to the extent that they are not satisfied, provide a signal that something is "wrong" with the component data.

"Wrong" can be within our tolerances or it can be outside our tolerances. If it is outside our tolerances, it is a signal that we ignore at our peril. The signal must be fed back; the component series must be carefully examined, and the source of divergence must be tracked down.

4. The discovery of data gaps is another analytical activity, at least if it is done with intelligence and a broad perspective. It would be easy to react to the feedback we get daily from clients that this, that, or the other piece of information that they would like to have is missing. While that is very important feedback, it is necessary to step back and get an understanding of the underlying phenomenon that we are trying to shed light on with our data, to ask the questions that users ought to have asked, and to

identify the gaps they ought to have identified. Perhaps they did not do so because they did not take that step back, they did not attempt that deeper understanding that is a prerequisite for a true identification of the underlying data gaps.

Let me give you examples: there is currently a great deal of controversy about health costs and rationing or pseudo-rationing of the health system. It is very easy to identify all sorts of immediately visible "data gaps" - why don't we have detailed information on all the elements of the health system that would help decision makers in their immediate task of deciding which hospital or what kind of hospital bed to close in preference to another? But deciding which hospitals to close might not be the real issue. The real issue is to optimize the allocation of health expenditures so as to maximize population health. So the real issue is to understand what are the determinants of health and to understand them so that health policy can be articulated intelligently in full understanding of its impact on the ultimate objective - on which there is no misunderstanding and no ambiguity - the improvement of the health of the Canadian population. It is an extremely difficult and ambitious task to understand in detail the determinants of public health. But we fail in our duty if we do not try to do so because otherwise we would simply respond to superficial demands that our clients articulate, and not to the fundamental requirements of society which is what we are here to try to do.

The same applies in every other field. I could have mentioned education as readily. We are not here to provide data for their own sake - about the number of students that enrol in educational institutions, or the number of teachers, or the subjects that are taught, or the graduation rate, and so on. To the extent that those are important determinants of the state of education in Canada, we should provide them. But a prerequisite is to understand, firstly, what we mean by education, and secondly, what are its determinants. Most know what we mean by education. But what are the determinants of good education, and how do we measure them? This calls for analytic work of a very high order because it involves looking for fundamental relationships and structures - the essential characteristics that the dictionaries have identified as being at the core of analytic activity.

We need product champions in this agency once we discover data gaps. Very often we are steered to act in response to the discovery of important data gaps by some persistent champion. I encourage all of you who will end up working as analysts in this agency to think about that role. A product champion is often unappreciated in the short run, particularly as it often involves becoming a nuisance, but in the long run it is a very productive role. It is due to product champions in the agency that we developed the System of National Accounts. Simon Goldberg was one of its product champions. Income distributions, which we now take for granted, would not have been developed - or at least they would have been developed much later - without a product champion in the person of Jenny Podoluk. More recently, the range of data that we produce on aging is the result of the product championship of Leroy Stone.

All these people are analysts. They are product champions because their desire to understand and analyze certain phenomena runs up against the data gaps and compels them to act as internal agitators to fill these gaps. We need those beneficial troublemakers in the agency.

5. A fifth reason why I am convinced of the importance of analysis is because analytic techniques provide the means for the extrapolation and interpolation of data that cannot be measured directly. A number of components of GDP are determined through extrapolation and interpolation. For example, several components of production are obtained by using employment data together with the relationship between employment and production.

6. A sixth reason for the importance of analysis is provided by the important contacts it engenders between Statistics Canada and policy departments. If Statistics Canada is a successful agency, it is in large part because it has the support of policy departments. There is a need for close links between analysts within policy departments and analysts at Statistics Canada - analysts who understand their concerns and interests and their approach to problems. Although the Statistics Canada analyst must not be a policy analyst, the analysis nevertheless is the common link.

7. The same applies to the seventh category - contacts with the academic community. Sociologists, economists, analysts in quantitative methods in academia are analysts first and foremost. If we want to have good contacts and serve them well we need people capable of maintaining strong relationships with them and understanding their needs. It is like a club. If you want academics to treat you as a member, you must become an analyst - you have to publish and speak at conferences, and so on.

8. The eighth point, which sounds controversial, is that we need analysis to develop our leadership in subject matter areas. The kind of people we ought to assign to head up our subject matter divisions are people who are able to understand, to discover the essence of things - i.e. people with a demonstrated aptitude for analytic thinking. Now that does not mean that everyone heading a subject matter division should be a world famous subject matter analyst. I am not talking about simply analysis, I am talking about a way of thinking: the ability to relate, to think analytically, to discover relationships. In management, we also need analysis in order to make the right decisions: whom do we need to train, what kind of human resource development needs do we have? Whether we talk about setting priorities, what data gaps exist, what redundancies exist, what information is less important than it used to be, it involves analysis. Where does our money go, what are the relationships between the different error components on the one hand, and where do we put our money on the other hand? Can we optimize our deployment of resources to lead to more reliable data at lower cost? That is analysis. So I do not think it exaggerated to say that analytic thinking is fundamental to the development of our leadership. I regard myself as an analyst even though I have not published too many subject matter articles. It is the kind of thinking that I emphasize, that is best developed through doing analytic work at some point in one's career.

9. The ninth and final reason why I regard analysis as so fundamental for the Bureau is because it contributes to the popularisation and highlighting of our findings.

In a recent interview with W5, one of the questions I was asked was: how can you keep on top of 35 million pages of printed material that this agency produces? (I did not know that we produced 35 million pages or where they got this piece of information but I presume it is valid.) One of the messages that I tried to convey is

that I do no try to keep on top of 35 million printed pages. That is not how I see my role. Nor is it the role of Statistics Canada. Rather, I try to keep on top of the main findings - the information that we provide to society about itself. I try to keep on top of our failings and our successes and in particular, steer us away from failure.

The popularisation and highlighting of our information is a very important function. The image - and in many cases, unfortunately, the fact also - is that too often we produce numbers and we do not take the trouble to understand what the data show about a particular aspect of society. How can society set priorities intelligently if Canadians do not understand what the data show? A large part of the Canadian population is not used to analyzing numbers. We are not serving them if that is all we produce. We have to call data to their attention. We cannot expect people to sift through 35 million pages per year in order to find the occasional golden nugget.

If we do not do it, who will? And if it is not done, aren't we wasting not only the $270 million of our budget, but more importantly the enormous opportunity that is given to us to educate and provide feedback to our country about itself? This is why we have such a wonderful job, such an enormously satisfying job, so long as we do it property - not simply by producing 35 million pages of raw material (terribly important raw material nevertheless).

Intelligent popularisation and highlighting of our findings is doubly important. First of all, because it is the most effective way of reaching our clients - most of them indirectly through the media - to tell Canadians about their country, as reflected in our statistical activities. Secondly, because it is a way of popularising Statistics Canada. Every time the papers say: "Statistics Canada reports that...", we are legitimising our activity through a subliminal message to Canadians that what Statistics Canada is doing is important. And that encourages collaboration with Statistics Canada on the part of individual Canadians when an interviewer knocks at the door or when a questionnaire arrives in the mail. In the end, this is important to our political masters, and is translated into support through the political process. So highlighting intelligently and understandably the findings is utterly, totally, basically fundamental. If I could discover a few more adjectives, I would add them to the previous sentence. I do not know how I can emphasize it more. It is <u>the</u> most important message I am attempting to convey.

<u>Constraints to analysis</u>

There are some necessary constraints on analytic activity.

i. It is essential that our analytical activities be focussed, objective, and scientific. That is why we have peer reviews. Let's not undertake shabby analysis because we won't get away with it.

ii. Secondly, political objectivity and non-partisanship are more difficult to achieve in the area of analysis than in data production. And in order to maintain political neutrality we have to be very careful that we highlight the relationships and not the causalities. Causalities are seldom discovered by analysis. Causalities are something we attribute to relationships if we think we understand the underlying mechanism. But what we observe in data are

the relationships only, not causalities.

We have to be very careful about causalities. It is best to avoid them. But sometimes they are unavoidable, in which case it is essential that we identify all possible causalities and not single one out of many. So this is my second truly important message: distinguish between relationships and causalities.

iii. The third constraint relates to rigorous scientific objectivity. It involves the need to describe all the assumptions, methods and limitations, in the interest of reproducibility. This means that somebody else doing the analysis, knowing all the methods we used and their limitations, would come to the same conclusions.

iv. A fourth constraint is not to be prescriptive. This is somewhat similar to the question of objectivity - yet different. Being prescriptive implies giving advice on how to proceed in developing policy. This is something for policy analysts - it is not up to us.

v. The fifth constraint is to avoid forecasting. We are in the business of identifying relationships. Forecasting is extrapolation, and very often it is not even based on a mechanical model, but is judgemental or partially judgemental. Projections are O.K - they are not forecasting. Projections are produced within a model, whose limitations are identified; they are reproducible. With projections, you can modify the model and arrive in exactly the same way at an alternative projection. Indeed, it is part of our policy that when we produce projections, we produce several of them corresponding to different scenarios. This means that, under stated assumptions, the model implies the following outcomes. That is analysis. Forecasting is what we think will happen. That is not appropriate.

Concluding remarks

Unambiguously, I regard it as terribly important to establish the right prerequisites for an analytic environment in this agency. This is not to say that everybody ought to be an analyst. But there should be a community within the agency that is sympathetic and understanding of analysis.

It involves taking analytic capability into account when we do our hiring. It involves taking analytical capability into account in our promotional competitions. It involves providing opportunities for analytic work as part of the day-to-day work, or as part of the sabbatical program that we have established. It involves providing exposure to the best practices through our focal points of analytic activity, such as the Analytical Services Branch, and the analytical divisions and sections that exist in various branches. And it involves studying and emulating best external practices through visiting fellows that we bring in. It is reflected through recognition that we provide, such as travel and, of course, training. That is what fostering analysis involves. So our objective is not to have an agency of analysts, but an agency that cultivates an analytic environment and analytic thinking. I know I can count on you in helping us to achieve that objective.