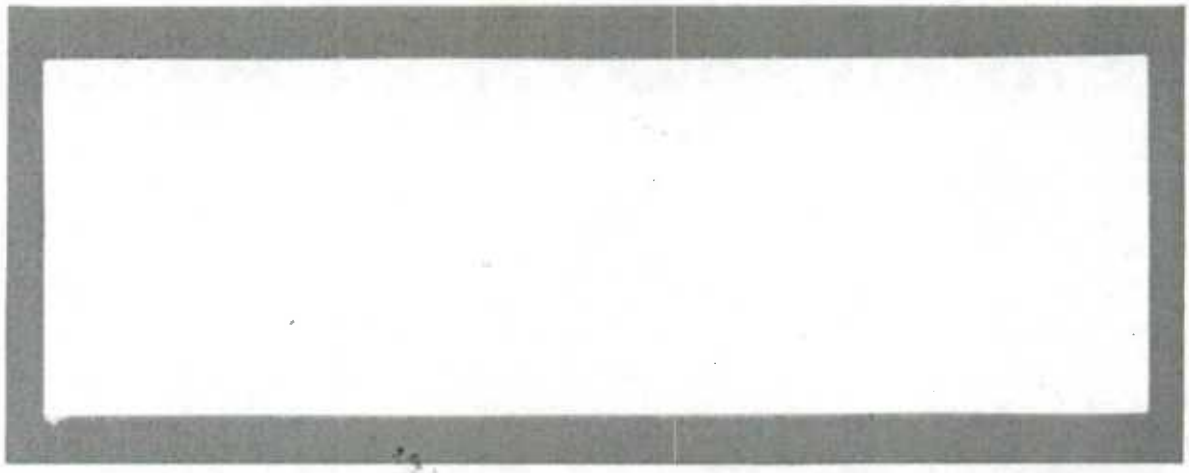
Statistics Canada | Statistique Canada



Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER

METHODOLOGY BRANCH

WEIGHTING AND ESTIMATION METHODOLOGY OF THE CANADIAN LABOUR FORCE SURVEY

HSMD-98-002E

Brian Kennedy

Survey and Analysis Methods Development

Household Survey Methods Division

Statistics Canada

February 1998

WEIGHTING AND ESTIMATION METHODOLOGY OF THE CANADIAN LABOUR FORCE SURVEY

Brian Kennedy[1]

Household Survey Methods Division
Methodology Branch, Statistics Canada

ABSTRACT

The Canadian Labour Force Survey (LFS) is a monthly survey of 52,300 households. It is used to derive the estimates necessary to monitor labour market conditions in the ten provinces of Canada. The purpose of this paper is to describe the estimation methodology used to derive these estimates. In doing so, the intent is to provide a detailed account of LFS weighting procedures so as to serve as a reference document, and to provide the rationale behind the particular methods employed. Included is a description of how the design weights are computed, how nonresponding households are treated, and how the generalized regression estimator is used to exploit demographic population estimates. A section devoted to the computation of sampling errors is provided, as is an appendix that describes the changes made to the estimation methodology between the present (1995) and previous (1985) sample designs.

---

[1] Brian Kennedy, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

LA MÉTHODE D'ESTIMATION ET DE PONDÉRATION DE L'ENQUÊTE SUR LA POPULATION ACTIVE

Brian Kennedy[2]

Division des méthodes d'enquêtes des ménages
Statistique Canada

RÉSUMÉ

L'enquête sur la population active (EPA) est une enquête mensuelle menée auprès de 52,300 ménages. Elle est utilisée pour obtenir les estimations nécessaires à l'évaluation de l'état du marché du travail dans les dix provinces du Canada. Le but de ce document est de décrire la méthode d'estimation utilisée pour produire ces estimations. Pour ce faire, l'intention est de fournir un compte-rendu détaillé des procédures de pondération de l'EPA qui pourra servir de document de référence, et de motiver le choix des méthodes utilisées. Sont inclus dans le document, une description de la façon dont les poids du plan sont calculés, la façon dont les ménages non-répondants sont traités, et comment l'estimateur généralisé par régression est utilisé pour exploiter les estimations démographiques de la population. Une section consacrée au calcul des erreurs échantionnales est fournie sous forme d'annexe. Les changements apportés à la méthode d'estimation entre les plan d'échantillonnage présent (1995) et précédent (1985) y sont décrits.

---

2 Brian Kennedy, Division des méthodes d'enquêtes des ménages, Statistique Canada, Ottawa, (Ontario), Canada, K1A 0T6.

WEIGHTING AND ESTIMATION METHODOLOGY OF THE CANADIAN LABOUR FORCE SURVEY

Brian Kennedy
Household Survey Methods Division

1. INTRODUCTION

The mandate of the Canadian Labour Force Survey (LFS) is to monitor labour market conditions in the ten provinces of Canada. The standard monthly **characteristics of interest** are, the total number of persons employed, unemployed and not participating in the labour force. Several functions of these totals, namely the unemployment rate, the participation rate and the employment rate, are also of interest. Detailed definitions of these and other labour force characteristics can be found in **GUIDE TO LABOUR FORCE SURVEY DATA** (Statistics Canada catalogue 71-528). Information about these characteristics are required nationally and provincially for various age/sex, industry and occupational groupings, as well as for sub-provincial regions.

In order to obtain the desired information, a sample of 52,300 households is selected each month. The survey excludes inmates of institutions, members of the armed forces and persons living on native reserves. An LFS questionnaire is completed for all persons 15 years of age and over living in the selected household. Estimates are obtained from the sample data using knowledge of the sample design and by employing estimation techniques from the theory of survey sampling. As will be described in this paper, each person in the sample receives a survey weight. This weight represents the respondent's contribution to the total population and is used for deriving estimates for all characteristics of interest. This weight is derived as the product of three factors. A design weight, which incorporates design information; a nonresponse adjustment, which compensates for nonresponding households; and a final weight that calibrates the sample to known population counts.

Once the estimates are derived, it is necessary to judge their reliability. Because the LFS is a **probability sample** it is possible to compute the sampling error associated with each estimate. The sampling error can be used to make probability based statements, or inference, concerning survey estimates.

The purpose of this paper is to describe the methodology used by the LFS to derive estimates and to provide the rationale behind the particular methods employed. This is followed by a description of the method of estimating the sampling error. Finally, an appendix is provided that describes the changes made to the estimation methodology between the present and previous sample designs.

2. DESIGN WEIGHT

An important concept in estimation is that of the **design weight**. In any sample survey a target population is defined. The target population is the subset of the population that the characteristics of interest refer to. It coincides with the population being sampled. In any given sample, some members of the target population are selected and others are not. The selected members can be thought of as representing the non-selected members. In a probability sample, a member has a known probability of being selected. If that selection probability is one in fifty, then the member represents 50 persons in total. One could make 50 copies of the survey responses and by repeating this procedure for every member in the sample, create a pseudo population. This pseudo population could be used for deriving the required estimates, since if the sample is representative of the population, then tabulations carried out on the pseudo population will be very

close to what would have been obtained had the true population been used. In practise, the records are not duplicated but rather are assigned a weight. The design weight is the number of times the record would have been replicated.

For the LFS, the following observations are important in determining the estimation procedure:

1) The survey uses a stratified, multi-stage design, with sampling conducted by probability proportional to size (PPS) at all stages except the final stage which uses systematic sampling. That is, given the list of households in the penultimate units (or clusters), every K^{th} household is selected. How K is determined is described later.
2) Since ultimate sampling units are households, the design weights in the LFS refer to households. As mentioned earlier, information is collected on every member of the household. Every person in the household is given their household weight in order to eventually derive estimates referring to persons.
3) The LFS is a repeated survey. Once the survey is designed, the same design is used month after month until a new design is introduced. Historically, the survey has been re-designed every 10 years. It is expected that growth in the population will occur over life of the design and appropriate adjustments are made at the sampling and estimation stages.
4) At the time of the design of the survey, information from the most recent census is used. In this case, the **design counts** are from the 1991 census.

Given the sample allocation, the survey design itself determines an initial set of design weights. These weights could be used as long as the design and allocation remain unchanged. Because the penultimate units experience growth over time and the systematic sampling rate is fixed, this would lead to an ever increasing sample size (and ever increasing costs). It would also lead to extreme variations in interviewer assignment sizes both within the same assignments over time and between different assignments. To avoid this sampling methods are employed that control the sample size. The methods, sample stabilization and cluster sub-sampling (described below), change a household's probability of inclusion in the sample. It is necessary to adjust the initial design weights to compensate for these methods. The adjustment factors are called the stabilization weight and the cluster subweight.

Sample stabilization and cluster subsampling involve dropping households in order to address problems with sample size growth. Stabilization accommodates the slow growth over time that is the result of the increasing size of the population, which left unchecked would lead to an undesirable increase in the sample size. Cluster subsampling accommodates isolated growth in relatively small areas that could present interviewers with work load problems.

In this paper the **initial design weights** are referred to as the **basic weights**. The term **design weight** is reserved to mean the inverse inclusion probability that applies at the time of sampling. Usually the terms design weight and basic weight are used interchangeably. Here we only make the distinction to emphasize this difference.

To summarize, the **design weight** for a particular household is equal to the household's inverse probability of inclusion in the sample. In the Labour Force Survey, it is computed as the product of three factors. These are referred to as the **basic weight**, the **stabilization weight** and the **cluster subweight**.

2.1. BASIC WEIGHT

When designing the survey, strata were formed by grouping together geographic units. The geography used may have been either census enumeration areas (EA's) or census sub-divisions (CSD's) depending on the area being stratified. Details of the stratification can be found in Gambino (1996). From each stratum, the number of households to be selected is determined and fixed. For stratum **h** we will call this $\mathbf{n_h}$. We also know the number of households in the

stratum at the time of design of the survey. Denote this by N_h. The inverse stratum sampling rate is given as:

$$R_h = \frac{N_h}{n_h}$$

In the following, the specific methods of probability proportional to size (PPS) sampling used varies across different types of areas as described in Gambino (1996). Here, if we accept that when the method of random groups is used, the random groups are essentially stratum with one first stage unit (FSU) selected, we need only know that it is PPS sampling. Also, note that sampling rates are necessarily integers. Here we ignore this sufficing it to say that integers are obtained by taking the integer greater than (or less than) the decimal number.

Because the LFS uses multi-stage sampling, it is necessary to determine the number of units to be selected at each stage of sampling. Consider the case of two stage sampling. The expected sample take based on design counts from each FSU is fixed. This is called the density factor and for FSU **j** in stratum **h**, will be denoted by n^*_{hj} The number of FSU's to select, n_{1h} is given by n_h/n^*_{hj}. If N_{hj} is the number of households in FSU j in stratum **h**, the sampling rate for the FSU is N_{hj}/n^*_{hj}. This is denoted by R_{hj}. (R_{hj} corresponds to the value **K** described in point 1 in the introduction to section 2.)

In some cases n_{1h} is fixed and n^*_{hj} is determined as n_h/n_{1h}. In either case, the size of the stratum is set to obtain desirable sample sizes.

We can now determine a household's inclusion probability as the product of selection probabilities at each stage. We use R_{hj} as the size measure for PPS sampling, for the j^{th} FSU in stratum **h**. The first stage inclusion probability for FSU **j** is :

$$\pi_{1hj} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}} R_{hj}$$

The conditional inclusion probability of selecting household **k** given FSU **j** is selected is, by definition:

$$\pi_{k/j} = \frac{n^*_{hj}}{N_{hj}} = \frac{1}{R_{hj}}$$

The inclusion probability of household **k** in stratum **h** then is:

$$\pi_{hk} = \pi_{1hj}\,\pi_{k/j} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}} R_{hj} \frac{1}{R_{hj}} = \frac{n_{1h}}{\sum_{j\in h} R_{hj}}$$

Note that,

$$\sum_{j\in h} R_{hj} = \sum_{j\in h} \frac{N_{hj}}{n^*_{hj}} = \frac{n_{1h}}{n_h} \sum_{j\in h} N_{hj} = n_{1h} R_h$$

The inclusion probability equals the original stratum sampling rate, $1/R_h$. In general, sample designs with equal basic weights within each stratum are called **self weighting designs within each stratum**. The LFS is self-weighting within each stratum (with respect to the basic weight) and the basic weight is R_h.

2.2 CLUSTER SUBWEIGHT

As described earlier, the LFS follows a multistage design. The penultimate units, or clusters, are sampled at a fixed rate determined on the basis of the 1991 census counts, in order to yield between 6 and 10 dwellings. In urban areas in particular, new development often takes place and the number of dwellings in a cluster can grow substantially. When this occurs, given the fixed sampling rate, an interviewer's assignment size can become overburdening. This can affect the quality of the interviewer's work in addition to his/her ability to complete the assignment. In such cases, a request to subsample the cluster may come from the responsible Regional Office. When growth in a cluster is less than 200%, the request is rejected. In cases where the request is accepted, one of three methods is employed to address it. All of these methods involve randomly dropping sampled households from the problem cluster, with the result that a households probability of inclusion is altered. Instead of continually re-computing the basic weight, it is easier to compute a weight adjustment and apply it to the original basic weight. This adjustment factor is called the cluster subweight. The three methods used to drop the households and determine the cluster subweight are:

Method I: Subclustering

When growth exceeds 300% and street patterns are well enough defined to delineate clusters, the problem cluster is re-delineated into several smaller clusters. A sample of the smaller clusters is taken, say $\mathbf{n}_{2hj}$. The smaller clusters are sampled in a manner that will reduce the overall take. Let $\mathbf{R}_{hj}$ equal the sampling rate of the original cluster. The size of the new clusters, $\mathbf{N}_{hji}$, and the expected sample take, $\mathbf{n}_{hji}$, gives the sampling rate for the new clusters, $\mathbf{R}_{hji}$. We now ask the question, at what rate would we have to sample the original cluster in order to obtain the total sample size obtained from the new subclusters? This is given by:

$$\mathbf{R}^{*}_{hj} = \sum_{i \in j} \frac{\mathbf{R}_{hji}}{\mathbf{n}_{2hj}}$$

The cluster subweight is given as

$$\mathbf{K} = \frac{\mathbf{R}^{*}_{hj}}{\mathbf{R}_{hj}}$$

The original basic weights given to the households which end up being selected are multiplied by this factor to reflect their actual selection probability.

Method II: Self- Representing Cluster

When the characteristics of the growth dwellings are distinct from the remainder of the stratum or, the size of the cluster is at least 20% of the size of a stratum, the cluster is first re-classified as a stratum. Call this new stratum **(hj)**. New clusters are formed within this new stratum and a sample is drawn. The sample from the cluster represents the cluster itself, rather than only being a contribution to the larger original stratum. If the design count of the new stratum is $\mathbf{N}^{N}_{(hj)}$ and the expected sample take is $\mathbf{n}^{N}_{(hj)}$, the stratum sampling rate is given by,

$$\mathbf{R}^{N}_{(hj)} = \frac{\mathbf{N}^{N}_{(hj)}}{\mathbf{n}^{N}_{(hj)}}$$

Using the same logic as applied in section 2.1, we know that $\mathbf{R}^{N}_{(hj)}$ is the basic weight to be assigned to households selected from this new stratum. As households selected from this new stratum are assigned the weight from the original

stratum, R_h, the appropriate factor is,

$$K = \frac{R^N_{(hj)}}{R_h}$$

It is also necessary to apply an adjustment to all the sampled households in the remainder of the original stratum. Consider a stratum from which six clusters are selected. The six clusters are used to represent the full stratum's population. After removing the growth cluster from the stratum, the weights for households in the five remaining clusters must be adjusted so that they represent the remainder of the stratum.

Let $N^R_h = N_h - N^N_{(hj)}$ be the design count of the remainder of the stratum and if n_{hj} is the original expected take from the cluster that has been removed from the stratum, let $n^R_h = n_h - n_{hj}$ be the expected sample take from the remainder of the stratum. The new inverse sampling rate for the stratum is

$$R^R_h = \frac{N^R_h}{n^R_h}$$

This leads to the cluster subweight,

$$K = \frac{R^R_h}{R_h}$$

Method III: Cluster Subsampling

When a cluster is to be subsampled , and neither methods I or II apply, this, the simplest and most common case of subsampling, is used. First the cluster is sampled based on the original design sampling rates. This yields a set of sampled households. A second selection is made from the sampled households. The households that remain after the second selection are interviewed while the remainder are dropped from the sample. If the cluster was originally sampled at a rate of R_{hj} , and subsampling leads to a sampling rate of R^*_{hj} then the cluster subweight is

$$K = \frac{R^*_{hj}}{R_{hj}}$$

For example, if every second selected household is chosen to remain in the sample then the new sampling rate for the cluster is twice the old sampling rate. The above ratio would equal two. For households in this cluster, the basic weight will be multiplied by two in order to compensate for the discarded households. Due to outlier problems encountered by special surveys that use the LFS frame, the maximum value the cluster subweight can be when using method III is 3.

2.3 STABILIZATION WEIGHT

The final stage of sampling is conducted using systematic sampling at a fixed rate. As the sampling rate is employed consistently over time, growth in the population, (and hence number of households) will lead to an ever increasing sample size and escalating survey costs. In order to control costs, sample stabilization is carried out. Sample stabilization is the random dropping of dwellings from the sample in order to maintain the sample at its desired level. By randomly dropping dwellings, a household's inclusion probability is changed. For example, suppose we define a stabilization area, **a** , in which households have a probability of inclusion of 1 in 200 at the time of design. If the area has a desired take of 300 dwellings and sampling using the probability assigned yielded 350 dwellings, then 50 dwellings must be dropped. After dropping the 50 dwellings, the inclusion probability is no longer 1 in 200, but rather

3 in 700 (ie. 1/200 times 300/350). As with cluster subsampling, it is simpler to adjust the basic weights where necessary rather than to continually re-compute them. The basic weight is retained as 200 but is multiplied by the factor 350/300 in order to yield the desired weight. The adjustment factor is called the stabilization weight.

It is first necessary to define **stabilization areas**. For the present design, a stabilization area is defined as all dwellings belonging to the same Employment Insurance Economic Region (EIER) and the same rotation group. For each stabilization area, **a** , a base sample size is determined. This is the desired sample based on the sample allocation. The base sample size for area **a** is denoted, b_a. If sampling took place without stabilization a number of dwellings would be selected. Call this n_a. If n_a exceeds b_a it is necessary to drop $n_a - b_a$ dwellings. This is done systematically at random. Once this is done we must adjust the basic weight.

The LFS follows the rule that if a cluster has been subsampled using method III of section 2.2, then the cluster should be excluded from stabilization. No dwellings from that cluster can be dropped nor is the stabilization weight applied. It is felt the area should not lose more households than is already the case. Denote the total number of dwellings in stabilization area **a** , excluded from stabilization in this manner by c_a.

There are two other cases when a household in a stabilization area does not receive the stabilization weight. On occasion, a group of households that were originally believed to be one household are encountered. These households, called multiples, are all included in the sample. As they did not have an opportunity to be excluded via stabilization, they do not receive a stabilization weight. Also, over the lifetime of a cluster new dwellings are built and added to the cluster list of households. Again, since they were not eligible to be dropped, no stabilization weight is applied.

Once the dwellings have been dropped, the stabilization area is partitioned into sub-areas. A stabilization sub-area is the collection of strata within the stabilization area which have a common inverse sampling rate, R_h. The stabilization weights are calculated separately for each sub-area. In the notation we ignore this subtle point.

The stabilization weight to apply to households in area **a** is,

$$F_a = \frac{n_a - c_a}{b_a - c_a} .$$

To conclude this section we repeat, the **design weight**, or inverse inclusion probability, is given as the product of the **basic weight**, the **stabilization weight** and the **cluster subweight**.

3. TREATMENT OF NONRESPONSE AND DERIVATION OF THE SUBWEIGHT

As with all surveys, the LFS experiences nonresponse. Nonresponse is classified as one of two types:

1. **Item nonresponse**

Those households for which only some information is missing. This could mean some, but not all items are missing for one or more household members, or all information is missing for some but not all household members.

2. **Whole unit nonresponse**

Those households for which there is no information available for any members of the household.

Item nonresponse is treated entirely by **imputation.** For a particular item nonrespondent, a donor record is found from among the respondents. The donor's responses to the corresponding missing information is used. Typically, a suitable donor is a person who has similar geographic and demographic characteristics and for those items for which responses are available, similar response patterns. The details of the imputation can be found in Lorenz (1995).

In the case of whole unit household nonresponse, if a nonresponding household had responded in the previous month, then the previous month's responses are "carried forward". This method is employed only if there was a response in the previous month (ie. carried forward data is not carried forward again).

Finally, all remaining whole unit nonresponse is treated by the method of weight adjustment. The principle of weight adjustment is that the responding households can be used to represent both responding and nonresponding households. The design weight is multiplied by this adjustment factor and the result is called the **subweight**.

In order to carry out this weight adjustment, the sample is first partitioned into weight adjustment classes or **nonresponse areas**. The nonresponse areas are defined in such a way as to improve the chances that the respondents will have characteristics similar to those of the nonrespondents. Presently in the LFS, the nonresponse area is defined as all households that belong to the same Employment Insurance Economic Region (EIER), the same type of area (see below) and have been in the sample for the same number of months. EIER's are sub-provincial regions defined by Human Resources Development Canada in order to administer the Employment Insurance Program. Type of area refers to the type of frame the sample is drawn from (see Gambino, 1996). The classifications are as follows:

CMA Apartment design.
CMA Regular design
Non-CMA Computer Assisted Districting Program (CADP) design
Urban EA design
Urban Cluster design
Urban 3 stage design
Rural EA design
Rural 3 stage design
Remote area design.

Tenure in sample is included in the definition of a nonresponse area because it is known that both the magnitude and patterns (refusals vs. non-contacts etc.) of nonresponse differ depending on how long a household has been in the survey. In the context of nonresponse adjustment, this is discussed in Kennedy et. al. (1994) . One feature of the new design is the formation of high income strata. Because of their unique characteristics, high income strata are treated as nonresponse areas on their own. Note that the nonresponse areas do not overlap and together cover the entire target population.

Within each nonresponse area, a **nonresponse adjustment factor** is computed. The adjustment factor for a nonresponse area is given as the ratio of sampled households, weighted using the design weight, to represent the number of households in the area, to responding households weighted to estimate the number of households in the area that would respond. If we denote by **n** the number of sampled households in nonresponse area **b** , and **r** the number of responding households, then the nonresponse factor is given as:

$$F_b = \frac{\sum_{k=1}^{n} \pi_k^{-1}}{\sum_{k=1}^{r} \pi_k^{-1}}$$

where π_k^{-1} is the design weight assigned to the household as described in section 2.

It is believed a value greater than two for the above weight is undesirable so when this occurs the nonresponse area is collapsed with another nonresponse area chosen so that when the pooled weight is computed, it will be less than two. The nonresponse area to collapse with, must come from the same province, the same type of frame and the same EIER (collapsing across rotation groups).

This weighting factor is applied to all responding households in the area. The **subweight** is defined as the product of the **design weight** and the **nonresponse factor**.

4. FINAL WEIGHT

In principle the subweight defined above could be used to produce estimates of the characteristics desired. However, from estimation theory, it is known that if auxiliary information about the target population is available, and this information is correlated with the characteristics of interest, then it can be used to produce more efficient point estimates. This is somewhat intuitive. Consider a sample that by chance consists of 50% women and 50% men. If the true distribution of males and females in the population is 51% women and 49% men, then the sample under represents females. Many labour force characteristics depend on gender. For example, a higher proportion of men are employed. Adjusting the weight so that the true proportion of each gender group was represented would lead to a better estimate. The adjustment factor computed to exploit auxiliary information is called the **g-weight**. The product of the **subweight** and the **g-weight** is called the **final weight**.

To obtain the g-weight the LFS uses a form of the general regression estimator (GREG) based on the final weighting methodology proposed by Lemaitre and Dufour (1987). For each province, post censal estimates of population, projected to the current time period, are used as auxiliary information. Specifically, the estimates used are population totals for 30 age/sex groups within each province, Economic Regions, and CMA/CA's. Also, for any particular rotation group, the sample of persons 15 years of age and older within the rotation group, is benchmarked to one sixth of the total province level population 15 years and older. The regression estimator uses all the auxiliary information simultaneously. The population counts are produced each month by Labour and Household Surveys Analysis Division.

As mentioned earlier, the LFS weight is a household weight. The GREG estimator computes a final weight at the household level, derived in such a manner that the sum of the final weights in a particular age/sex grouping, or in a particular sub-provincial region agree identically with the population estimates used as auxiliary information. Also, the estimates of employed, unemployed and not in the labour force will sum to the population totals used as auxiliary information. Because the weight is the same for all persons in the same household, family level estimates and person level estimates are also consistent. This was not true in the methodology employed before the regression estimator was used.

To conclude, the following are some advantages to using the final weight step:

- consistency of estimates with demographic estimates of population.
- an adjustment for coverage error.
- a common weight for all members of the same household.
- reduction in sampling error of estimates.

5. ALGEBRAIC DESCRIPTION OF WEIGHTING A RECORD

The following is an algebraic description of weighting. It is necessary to begin by introducing notation. The LFS sample design consists of a nested hierarchy of geography. Consider the following:

Let:
p = 1 , ... , 10 denote the province.
u = 1 , ... , U denote the EIER u, within province p.
f = 1 , ... , F denote the type of frame within EIER u.
h = 1 , ... , H denote stratum h within frame f.
r = 1 , ... , 6 denote the rotation group within stratum h.
j = 1, ... , J denote cluster j of rotation group r.
k = 1, ... , K denote household k in cluster j .
i = 1 , ... , c_k denote individual i within household k.

With this notation a household is identified with the subscript **pufhrjk** . A subscript containing periods or missing subscripts indicates a reference to a level of accumulation. For example, **pu..r** refers to all households in province p, EIER u, and rotation group r , collecting households over the missing subscripts.

In some cases, design strata cut across EIER boundaries. For the most part this has occurred because HRDC redelineated its EIER's after the redesign of the LFS. Special estimation techniques are used to produce estimates for these regions. Specifically, a description of the method used, the sample size dependent estimator, can be found in Drew et. al. (1982). For now we note that the above geography is not quite perfectly nested. This will present no problem for the standard estimation methods discussed in this section.

To begin, note that the **basic weight** is the same for every household in the same stratum. At the time of design of the survey, the inverse selection probabilities are the same for all households in the same stratum. The basic weight can be denoted as:

$$w_{pufh}$$

The next two weighting factors, the **cluster subweight** and the **stabilization weight** adjust the basic weight to account for various adjustments to the sample yields as described in section 2. The method of computing the cluster subweight depends on the method of subsampling employed.

Method I: Area Subsampling

In this case, the cluster is redelineated into smaller clusters. A sample of clusters is then selected and sampled to obtain some fixed total yield. If the sampling rate from the original cluster was $R_{pufh.j}$, and if the sampling rate at which the

original cluster had to be sampled at in order to obtain the new total sample yield is $\mathbf{R}^{*}_{\mathbf{pufh.j}}$, then the cluster subweight is:

$$\mathbf{c}_{\mathbf{pufh.j}} = \frac{\mathbf{R}^{*}_{\mathbf{pufh.j}}}{\mathbf{R}_{\mathbf{pufh.j}}}$$

Method II: Self Representing Cluster

In this case, the problem cluster is removed from the stratum and forms a new stratum. The new stratum, **h'** say, is delineated into clusters and sampled. If the new strata were sampled at the same rate as the old strata, no adjustment weight would be required. However this would likely yield a very small take. Typically it is sampled at a lower rate. If the sampling rate of the original stratum is $\mathbf{R}_{\mathbf{pufh}}$, and the sample rate of the new stratum is $\mathbf{R}^{*}_{\mathbf{pufh'}}$, then the cluster subweight to be assigned to households in the new stratum only is:

$$\mathbf{c}_{\mathbf{pufh'}} = \frac{\mathbf{R}^{*}_{\mathbf{pufh'}}}{\mathbf{R}_{\mathbf{pufh}}}$$

It is also necessary to adjust the remainder of the clusters in the old strata. This is to compensate for losing a cluster. Recall the sampling rate of the original strata is $\mathbf{R}_{\mathbf{pufh}}$. Denote by $\mathbf{R}^{\mathbf{R}}_{\mathbf{pufh}}$ the rate at which the remainder of the original stratum would be sampled at to get the expected design sample take from the remaining clusters. This leads to the following factor which is applied to all households in the remainder of the stratum:

$$\mathbf{c}_{\mathbf{pufh}} = \frac{\mathbf{R}^{\mathbf{R}}_{\mathbf{pufh}}}{\mathbf{R}_{\mathbf{pufh}}}$$

Method III: Cluster Subsampling

In this the simplest case the selected households are subsampled and only the subsampled households interviewed. If $\mathbf{R}_{\mathbf{pufh.j}}$ is the original sampling rate for the cluster and $\mathbf{R}^{*}_{\mathbf{pufh.j}}$ is the cluster sampling rate required to achieve the appropriate level of subsampling, then the cluster subweight is :

$$\mathbf{c}_{\mathbf{pufh.j}} = \frac{\mathbf{R}^{*}_{\mathbf{pufh.j}}}{\mathbf{R}_{\mathbf{pufh.j}}}$$

As outlined in section 2.3, stabilization weights are computed within stabilization areas. In the present design, a stabilization area is defined as the collection of all strata belonging to the same EIER. This area is then divided into common rotation groups. Within each stabilization area, a base sample size is determined. This is the number of household the area should sample, based on the sample allocation. This number is denoted as $\mathbf{b}_{\mathbf{pu..r}}$. When sampling takes place, a realized number of households is encountered, say $\mathbf{n}_{\mathbf{pu..r}}$. If $\mathbf{n}_{\mathbf{pu..r}} > \mathbf{b}_{\mathbf{pu..r}}$ then the area is being over sampled and the excess households are dropped at random, using systematic sampling. As clusters that were subsampled by Method III of cluster subsampling above are not eligible for stabilization, they are excluded when computing the stabilization weight. Denote the total of these dwellings in a stabilization are as $\mathbf{c}_{\mathbf{pu..r}}$.

When an area is subject to stabilization the following weight is applied to households in this area:

$$\mathbf{s}_{\mathbf{pu..r}} = \frac{\mathbf{n}_{\mathbf{pu..r}} - \mathbf{c}_{\mathbf{pu..r}}}{\mathbf{b}_{\mathbf{pu..r}} - \mathbf{c}_{\mathbf{pu..r}}}$$

Note that **some** households in a stabilization area do not receive the stabilization weight. These households are defined in section 2.3. Essentially, they are households which were not eligible to be dropped via stabilization.

We can now compute the **design weight** for each household as follows:

$$\pi^{-1}_{pufhrjk} = w_{pufh} \times c_{pufh.j} \times s_{pu..r}$$

The design weight is the inverse inclusion probability for the given household. When referring to the design weight in the following, the cumbersome subscripting will be dropped. That is,

$$\pi^{-1}_{k} = \pi^{-1}_{pufhrjk}$$

The next adjustment, described in section 3, is the nonresponse adjustment. Nonresponse areas are defined and an adjustment weight applied to compensate for complete household nonresponse. Presently the LFS defines nonresponse areas as all sampled households belonging to the same EIER, the same type of frame and the same rotation group. The adjustment factor is computed as the weighted ratio of sampled households to respondent households. That is,

$$f_{puf.r} = \frac{\sum_{k \in s} \pi^{-1}_{k}}{\sum_{k \in r} \pi^{-1}_{k}}$$

where summation over **s** indicates summation over all households in the nonresponse area and summation over **r** is over all responding households in the area. All households in the same nonresponse area receive the same nonresponse adjustment factor.

The **subweight** is given as the product of the design weight and the nonresponse adjustment:

$$a_k = f_{puf..r} \times \pi^{-1}_{k}$$

Note that all members of the same household receive the same value of the subweight.

As mentioned earlier, we could use the subweight to estimate the desired characteristics. Given a characteristic **Y**, employment say, we are interested in the total number of persons employed in the population. This can be denoted as:

$$t_y = \sum_{U} y_i$$

where summation over U indicates summation over all **persons** in the in scope population (the subscript **i** above refers to persons) and y_i has a value of one if an individual **i** is employed and zero otherwise.

The survey estimate based on the subweights defined above would give:

$$\hat{t}_{ya} = \sum_{s} y_i a_i$$

where summation over **s** indicates summation over sampled persons only.

As the LFS is a household survey, all persons in the same household receive the same subweight. It is useful to note that we could re-write the above formulae as

$$t_y = \sum_{k=1}^{N} \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{N} y_k$$

and,

$$\hat{t}_{ya} = \sum_{k=1}^{n} a_k \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{n} y_k \, a_k$$

where c_k is the number of persons in household **k** , **N** is the number of households in the population and **n** is the number of households in the sample. The y_k are household totals of the characteristic of interest. In this case it represents the total number of person employed in the household. **This notation is somewhat subtle and it is important to remember the subscript k will always refer to household totals and the subscript i to an individuals value.**

As mentioned in section 4, the LFS has access to post-censal population estimates that are derived independent of the sample. These are used as auxiliary information to derive a final set of weights. In order to exploit the auxiliary information we use the regression estimator. The version used for the LFS can be found in Lemaitre and Dufour (1987). The reader may wish to consult chapter 6 of Sarndal et al. (1992) for a thorough explanation of the regression estimator.

To this point, we have focussed on a weighting approach to obtaining estimates from the sample. We will get back on that line, but the following helps explain the utility of the estimation procedure used.

To begin consider the following notation:

- y_i **is the value of the characteristic of interest for individual i.**
- y_k **is the household total of the characteristic of interest for household k.**
- **Q is the number of auxiliary variables used in estimation. Each auxiliary variable will be denoted by q = 1,..., Q.**
- x_{qi} **is the value of the q^{th} indicator variable for individual i. The indicator variable assumes a value of one if individual i belongs to the j^{th} auxiliary category and zero otherwise.**
- x_{qk} **is the total of the values of the q^{th} indicator for all persons in household k.**
- x_k **is a Q x 1 vector whose q^{th} entry is the corresponding household total x_{qk} .**
- c_k **is the size of the k^{th} household. .**
- $\hat{t}_{ya}$ **is the subweight based estimate described above.**
- t_{xq} **is the known population total for the q^{th} auxiliary variable.**
- $\hat{t}_{x_q a}$ **is the subweight based estimate for the q^{th} auxiliary variable.**

That is,

$$\hat{t}_{x_q a} = \sum_{s} x_{qi} \, a_i$$

The **regression estimator** used can be written as,

$$\hat{t}_{yr} = \hat{t}_{ya} + \sum_{q=1}^{Q} \hat{B}_q \, (t_{x_q} - \hat{t}_{x_q a})$$

The $\hat{B}_q$ will be defined below. From the above formula, we can see that the regression estimator can be viewed as the subweighted estimator plus an adjustment term. If the sample based estimate is close to the known total for x_q , then

the adjustment term will be close to zero. If they are different, the adjustment term will be large. In some sense we are suggesting that if the sample is doing a good job of estimating the auxiliary variables, it is probably doing a good job of estimating the unknown characteristic of interest. Likewise, if the sample is doing a poor job of estimating the auxiliary variables, it is probably doing a bad job of estimating the characteristic of interest. This seems likely if the characteristics are correlated with the auxiliary variables.

To define the $\hat{\mathbf{B}}_q$ matrix notation is used.

$$\hat{\mathbf{B}} = (\hat{\mathbf{B}}_1,...,\hat{\mathbf{B}}_Q) = \left(\sum_{k=1}^{n} \frac{\mathbf{x}_k \mathbf{x}_k^t \mathbf{a}_k}{\mathbf{c}_k} \right)^{-1} \sum_{k=1}^{n} \frac{\mathbf{x}_k \mathbf{y}_k \mathbf{a}_k}{\mathbf{c}_k}$$

where,

$$\left(\sum_{k=1}^{n} \frac{\mathbf{x}_k \mathbf{x}_k^t \mathbf{a}_k}{\mathbf{c}_k} \right)^{-1}$$

is a **qxq** matrix. This matrix is the inverted weighted sum of squared cross products matrix seen as usual in regression estimation, and,

$$\sum_{k=1}^{n} \frac{\mathbf{x}_k \mathbf{y}_k \mathbf{a}_k}{\mathbf{c}_k}$$

is a **p x 1** vector.

As promised, we return to the weighting approach to estimation. By re-arrangement of algebra, the above estimator can be rewritten as:

$$\hat{\mathbf{t}}_{yr} = \sum_{s} \mathbf{y}_k \mathbf{a}_k \mathbf{g}_k$$

where

$$\mathbf{g}_k = 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{xa})^t \left(\sum_{s} \frac{\mathbf{x}_k \mathbf{x}_k^t \mathbf{a}_k}{\mathbf{c}_k} \right)^{-1} \frac{\mathbf{x}_k}{\mathbf{c}_k}$$

The g_k or **g-weights** are the factors applied to the **subweights** to obtain the **final weights**.

Comments:

The final weighting step is carried separately for each province. National totals are obtained by summing the provincial totals.

The fact that the final weights do not depend on the characteristic y means that the same weight is used for tabulating all characteristics of interest.

One curiosity in the method of estimation is the division by the household size ($\mathbf{c}_k$) through out. This is the result of the original formulation of the regression estimator for the LFS as described in Lemaitre and Dufour (1987). In their approach, weighting was carried out at the person level with the design matrix containing household averages rather than person level information. This was done in order to ensure each household member received the same weight. The above formulation at the household level is equivalent to their method. That is, they yield identical weights.

Algebraically, we could compute the g-weights as follows:

$$g_i = 1 + (t_x - \hat{t}_{xa})^t (\sum_s z_i z_i^t a_i)^{-1} z_i$$

Remember the subscript **i** refers to persons. The a_i is the subweight assigned to the i^{th} person in the sample. The z_i contain, for every person, the average of the values of the indicator variables for each person in the same household. That is,

$$z_i = \frac{1}{c_k} \sum_i^{c_k} y_i$$

Every person **i** in household **k** receives the same value of the indicator variable **z**, namely the household average. The issue of household versus person level weighting is discussed in Wu et. al. (1997).

6. VARIANCE ESTIMATION

6.1 OVERVIEW

To this point it is has been shown how point estimates are derived. For a characteristic of interest **y** we use,

$$\hat{t}_{y_r} = \sum_{k \in s} y_k a_k g_k$$

as an estimate for,

$$t_y = \sum_U y_k$$

In this discussion, we think of t_y as the measure that would be obtained if the entire population was surveyed.

The question arises, how good a proxy is the estimate of t_y for the census value? We know the estimate is unlikely to equal t_y exactly, but how close can we expect it to be?

Conceptually we can address this problem by considering how variable estimates from different samples might be. Suppose we know t_y, and we know the value of the estimates that would be obtained from each possible sample that could be drawn under the sample design. Denote the estimate obtained from sample **s, s = 1, ... , S** by:

$$\hat{t}_{y_r}^s$$

Consider the expected value of the squared differences between the sample estimates and the census value over all possible samples:

$$E(\hat{t}_{y_r}^s - t_y)^2 = \sum_{s=1}^{S} p(s) (\hat{t}_{y_r}^s - t_y)^2$$

where **p(s)** is the probability of drawing sample **s** .
The above measure is by definition the **variance** of the sample estimate. The square root of this measure would give us an idea of how much an estimate from the sample might differ from the census value.

Computing the variance based on the formula above requires knowing the census value and drawing all possible samples, but this information is not available. Instead the variance must be estimated from the sample. (Note that this is an estimate and is itself subject to error). Because of the complexity of the sample design and estimation technique, it is not possible to explicitly derive a formula that can be used to estimate the variances. Many methods of deriving variance estimates are available for such cases (see Wolter (1985)). The LFS uses the Jackknife method and is outlined in section 6.3.

At this point it is worth noting the factors that contribute to the variability of a particular estimate.

1. First, in the whole population there is some underlying variability in the characteristic of interest. This will contribute to the variance of an estimate of that characteristic. If one characteristic is more variable than another, then all other things being equal, the corresponding estimate will be more variable.

2. The sample size is an important factor in the variance of an estimate. The larger the sample size, the smaller the variance. However, the sample size will reach a state of diminishing returns. At some point large increases in the sample size will lead to only small reductions in the variance.

3. The sample design will impact on variance. Optimal stratification and PPS sampling will reduce the sampling variance when they are used properly. The necessity of a multistage design leads to a clustered sample which inflates the variance.

4. The auxiliary information used in estimation reduces the variance of those characteristics that are correlated with them.

Variances of estimates of month to month change and of both three month and annual averages are also of interest. The computational method is included in section 6.3. We conclude by noting a fifth factor that will contribute to the variances of estimates of change and averages.

5. The positive correlation between monthly estimates reduces the variance of the difference of the estimates. This correlation is enhanced by the overlapping sample. Also the predictors used in regression will contribute. Variances of averages may be slightly increased by this positive correlation.

6.2 A NOTE ON SAMPLING ERRORS

We briefly described the concept of the sampling variance in the previous section. In practice the measure we are interested in is the **sampling error**. The sampling error of an estimate is the square root of the variance of the estimate.

The sampling error is used to make inference. For example, we can make statements to the effect that the census value for a corresponding estimate has roughly a 68% chance of being within plus or minus one sampling error of the estimate and a 95% chance of being within plus or minus two sampling errors of the estimate. The point estimate itself is the mode, or the point with the highest probability of equalling the census value.

For monthly and average estimates we often present the errors as the Coefficient of Variation (CV) of the estimate. This is simply the standard error expressed as a percent of the estimate.

$$CV(\hat{t}_{y_r}) = 100\frac{SE(\hat{t}_{y_r})}{\hat{t}_{y_r}}$$

The estimate of the CV is obtained by replacing the standard error by the estimated standard error.

6.3 JACKKNIFE ALGORITHM

The variance estimator implemented for the LFS is the jackknife. In the general case a description and justification of the Jackknife is found in Chapter 4 of Wolter (1985). Here we only outline the Jackknife as it is applied in the LFS. The first step in the jackknife method is to create **replicate samples** from the LFS data. Within each design strata a first stage sampling unit is selected. This FSU is deleted from the sample and the subweights in the remainder of the stratum are adjusted to compensate for the loss. We then recompute the final estimates based on the replicated sample. By repeating this procedure for every FSU in the sample, we obtain estimates for every replicate sample. In the LFS the number of replicated samples is equal to the number of FSU's. When this procedure is done in an appropriate manner, the variability amongst the replicate sample estimates can be used as an estimate of the variability of the sample estimate.

In keeping with the operational terminology of the LFS, we will refer to the FSU being deleted from the sample in order to obtain a replicate sample as a replicate.

To obtain a variance using the jackknife procedure we proceed as follows, pointing out that the notation is refreshed from the previous sections.

For every characteristic within a specific province,

(i) Remove all the households from a specific replicate. Replicates will be denoted by $\mathbf{a} = 1, \ldots, J_h$. That is, the $\mathbf{h}^{th}$ stratum contains J_h replicates each one denoted by **a**.

The total number of replicates in the sample is :

$$J = \sum_{h=1}^{H} J_h$$

where **H** is the total number of stratum in the sample.

(ii) In the given stratum, for all the households in the remaining $J_h - 1$ replicates, an adjustment to the subweights is made. This is done to compensate for dropping the households. The adjustment weight is given as,

$$a_k^a = \frac{J_h}{(J_h - 1)} a_k$$

(iii) Using the remaining sample, with the adjusted subweights, we recompute the final weights to obtain a new estimate of the desired characteristic. The new estimate can be denoted as:

$$\hat{t}_{yr(ha)}$$

The notation **(ha)** indicates that the $\mathbf{a}^{th}$ replicate from the $\mathbf{h}^{th}$ stratum was deleted in order to obtain the new estimate.

This procedure is repeated for every replicate that has been defined in the sample. This will lead to a total of **J** different estimates of the desired characteristic. The key to the jackknife technique is that the variability amongst these estimates provides an estimate of the variability of our overall estimate.

The actual formula for the variance of the estimate of a total is:

$$\hat{V}(\hat{t}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{t}_{yr(ha)} - \hat{t}_{yr})^2$$

It is useful to note that interest often centres on the ratio of two totals. For example, the unemployment rate is the ratio of total unemployed to total labour force expressed as a percentage. In the general case a ratio **100(y/z)%** will use the variance formula:

$$\hat{V}\left(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}} \right) = (100)^2 \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} \left(\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}} - \frac{\hat{t}_{yr}}{\hat{t}_{zr}} \right)^2$$

Variances of the estimates of month to month change and for averages over a number of months require linking the jackknife estimates over time. Consider the difference estimate:

$$\hat{D}_{yr} = \hat{t}^2_{yr} - \hat{t}^1_{yr}$$

and the corresponding jackknife estimates:

$$\hat{D}_{yr(ha)} = \hat{t}^2_{yr(ha)} - \hat{t}^1_{yr(ha)}$$

where the superscripts refer to consecutive months. The estimate of variance is given by :

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{D}_{yr(ha)} - \hat{D}_{yr})^2$$

Variance of averages are obtained in a similar way. Consider the average over **n** months,

$$\hat{A}_{yr} = \sum_{i=1}^{n} \frac{\hat{t}^i_{yr}}{n}$$

and the jackknife estimates:

$$\hat{A}_{yr(ha)} = \sum_{i=1}^{n} \frac{\hat{t}^i_{yr(ha)}}{n}$$

The estimate of variance is given by :

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{A}_{yr(ha)} - \hat{A}_{yr})^2$$

The above method provides estimates of the variance of characteristics within each province. Variances for national totals are obtained as the sum of all the corresponding provincial variances.

ACKNOWLEDGEMENT

Many individuals have reviewed this document. The author is grateful to J. D. Drew and J. Lindeyer whose extensive comments on an earlier version of this paper led to a major revision and resulted in this more thorough account. I am also grateful to J. Gambino, A.C. Singh, T. Merkouris, D. Sunter, J.M. Levesque, P. Lorenz for their comments and suggestions. Finally, a previous description of LFS weighting methodology under the old sample design appears in chapter 8 of Singh et. al (1990). That description was most useful in writing this document.

REFERENCES

Demography Division (1987). Population Estimation Methods, Canada. Statistics Canada, Catalogue 91-528E.

Deville, J.C. and Sarndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association,* 87, 418, Theory and Methods.

Drew J.D., Singh M.P., and Choudhry G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology* 8, 17-47.

Fuller W. A. (1975). Regression analysis for sample surveys. *Sankhya*, C37, 117132.

Gambino, J. (1996). The 1995 Labour Force Survey Sample Design, draft internal document, Household Surveys Methods Division, Statistics Canada.

Kennedy, B. (1994). Labour Force Survey Weighting and Estimation System. System Specifications. Internal Document, Household Surveys Division, Statistics Canada.

Kennedy B., Drew J. D., and Lorenz P. (1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Presented at the 5th International Workshop on Household Survey Nonresponse. Ottawa, Canada.

Labour And Household Surveys Analysis Division (1997). Guide to Labour Force Survey Data. Statistics Canada, Catalogue 71-528.

Lemaitre, G.E. (1988). Integrated Person/Family Weighting Functional Description. Internal Document, Social Survey Methods Division, Statistics Canada.

Lemaitre, G.E. and Dufour J. (1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13, 199-297.

Lorenz, P. (1995). Labour Force Survey Head Office Imputation System Specifications. Version 3. Internal Document, Household Surveys Division, Statistics Canada.

Sarndal. C-E (1980) On π - inverse weighting versus linear unbiased weighting in probability sampling. *Biometrika*, 67, 639 - 650.

Sarndal, C.E., Swensson, B and Wretman J., (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Singh A.C. (1993). On Weight Adjustment in Survey Sampling. Discussion paper for the 18th meeting of the Advisory Committee on Statistical Methods, Statistics Canada, Ottawa Oct.25-26.

Singh A.C. and Mohl C. (1996). Understanding Calibration Estimators in Survey Sampling. *Survey Methodology*, 22, 107-115.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda F. (1990). Methodology of the Canadian Labour Force Survey. Statistics Canada, Catalogue 71-526.

Wolter K. M., (1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.

Wu, S., Kennedy, B. , and Singh, A.C. (1997). Household-level versus person-level regression weight calibration for household surveys. Proceedings of the Survey Methods Section, Statistical Society of Canada, to appear.

APPENDIX

CHANGES FROM THE PREVIOUS METHODOLOGY

Several changes to LFS weighting methodology were introduced with the implementation of the new sample design. Most notable were the elimination of the rural-urban factor, a change in definition of a nonresponse area, a change in the definition of the stabilization area, and the elimination of the replication of special area records. Each of these are outlined below as are the methods used to accommodate weighting during the phase-in of the new sample design.

1. RURAL-URBAN FACTOR

In the old LFS design, some strata in non-self representing areas consisted of both rural and urban parts. This made it possible to have an over or under representation of the rural or urban population. A factor was used to adjust the subweight so that the **proportion** of rural and urban population in any Economic Region was the same as it was at the time of the 1981 census. In the present design, the stratification is explicit and hence the sample is representative. Therefore the factor is no longer necessary due to this design enhancement.

2. NONRESPONSE ADJUSTMENT

For the majority of whole unit nonresponse, the LFS uses a weight adjustment. Applying such an adjustment requires making the assumption that nonrespondents can be represented by their responding counterparts in the so called nonresponse areas. The old method of defining a nonresponse area was as a stratum in self representing and special areas and as the rural or urban part of a primary sampling unit in the non-self representing areas.

Nonresponse patterns for the survey have been observed over time as part of the ongoing quality monitoring program. It has long been noted that response patterns differ among different rotation groups. Tenure in sample tends to affect the magnitude of nonresponse. The proportions of the different types of nonrespondents (non-contact, temporarily absent and refusal) also differs with the tenure in sample. Therefore it seemed reasonable to include tenure in survey in the definition of nonresponse area. Simply adding tenure in the old definition would have led to areas that had too small a sample take for adjustment. The definition was changed to be all households in the same Employment Insurance Economic Region, the same frame and with the same tenure in sample.

3. STABILIZATION AREAS

As with nonresponse areas, a change in the definition of a stabilization area was implemented. Previously, such areas were defined as all strata within a province with the same basic weight. Presently a stabilization area is defined as all households in the same Employment Insurance Economic Region and with the same tenure in sample, for the purpose of dropping dwellings. The weight adjustment is then computed within the stabilization area pooling all stratum with a common sampling rate. This change reflected the added importance placed on EIER's in the sample redesign.

4. SPECIAL AREA REPLICATION

The old LFS design had three frames that were small in terms of population. Namely, the institutions frame, the remote area frame and the Quebec remote urban frame. Combined, these frames covered about two percent of the population. The cost of interviewing in these areas was substantially more than in other areas and the sample yields were typically quite small. This latter point led to small interviewer assignments covering large areas. Because of the operational

difficulties of small assignments and the fact that such small populations were involved, it was decided to ignore the subprovincial regions when sampling these areas. For example not every Economic Region had a representative sample of the special area households (though each province did). In order to correct for this during estimation, all the special area records were replicated across Economic Regions that had population of corresponding type. The province level weights were then proportioned to represent the subprovincial region. For example if an Economic Region contained 10 percent of the provinces remote population, the sample records for remote areas were replicated into that region and their basic weight multiplied by 0.1 .

In the new design populations in institutions are not sampled from a special frame. They are no longer treated as a special case. The remaining remote area frame is sampled in the same manner as in the old design. Replication of the records is no longer required as the impact on estimates is minimal.

5. WEIGHTING DURING THE SAMPLE PHASE IN

The new LFS sample design was introduced by replacing the old sample, one rotation group a month over six months. Whenever households selected via the old design were due to rotate out of the sample, they were replaced with households from the new design. This process began in October of 1994 , with the new sample being fully implemented in March of 1995. Because of changes to the LFS numbering system, new sub-provincial geography, and modifications to the weighting methodology, some special considerations in weighting were required. For example:

- No stabilization of the new sample was carried out during the phase in period.
- It was decided to stop using the rural/urban factor in October 1994. As mentioned elsewhere the new sample did not require this factor. Applying it through the phase-in period to the old sample only would have lead to very unstable weighting factors. As old sample rotated out the imbalance in rural/urban populations for the old sample would have been exaggerated as the new sample contribution to these populations could not be accounted for.
- Nonresponse was adjusted using the old method for the old sample and the new method for the new sample.
- Special area replication continued to be carried out for the old sample only.
- Once the subweights were computed the two samples were combined for the final weighting step.