SSC Annual Meeting, June 1996
Proceedings of the Survey Methods Section

VARIANCE ESTIMATION AND CONFIDENTIALITY: THEY ARE RELATED!

J.E. Mayda, C. Mohl and J.-L. Tambay[1]

ABSTRACT

Statistics Canada is conducting the National Population Health Survey (NPHS), a comprehensive longitudinal household survey covering a variety of aspects related to health. The initial wave of the survey in 1994-95 provided a panel of respondents who will be followed-up every two years for up to twenty years. The products available from the first wave included basic tabulations of important variables by age and sex, a custom data request service for user-defined tabulations and a public use microdata file.

Statistics Canada operates under the Statistics Act, which contains confidentiality provisions that prohibit the disclosure of information related to an identifiable individual person, business or organization. Public use microdata files may be made available provided that the data cannot be associated with a particular respondent. Several measures were taken with the NPHS microdata file to ensure the confidentiality of respondents, including the removal of design information such as stratum and cluster identifiers from the file.

To provide an indication of the quality of the NPHS data, approximate sampling variability tables were included with the microdata file documentation. However, these tables are very general and do not address sophisticated users' needs. An exact variance program that calculates variances for specific user requests is available, but must be executed by Statistics Canada personnel due to the absence of design information on the public use microdata file. A research project has been undertaken to determine how design strata can be collapsed so that users may be provided with enough information to calculate a reasonable variance estimate while continuing to preserve the confidentiality of the data. This paper will describe how the collapsing of strata was accomplished and present the results of a study comparing the variances calculated using the detailed design information to those estimated using the aggregated design information.

KEY WORDS: Public use microdata files; Collapsing; Replicates.

RÉSUMÉ

Statistique Canada poursuit son Enquête nationale sur la santé de la population (ENSP), une grande enquête longitudinale des ménages couvrant une variété d'aspects reliés à la santé. La phase initiale de l'enquête a produit en 1994 un panel de répondants qui feront l'objet d'un suivi tous les deux ans jusqu'à une période maximale de vingt ans. Les produits disponibles de cette première phase comportaient entre autres des tableaux croisés élémentaires de variables importantes par âge et par sexe, un service de données sur mesure pour des tableaux définis par l'usager et un fichier de micro-données à grande diffusion.

Statistique Canada est régit par la Loi sur la statistique, qui contient des dispositions sur la confidentialité qui interdisent la publication d'information reliée à une personne individuelle, commerce ou organisme identifiable. Les fichiers de micro-données à grande diffusion peuvent être rendus disponibles, à condition que les données qu'ils contiennent ne pourrons être associées à un répondant particulier. Plusieurs mesures ont été prises avec le fichier de micro-données de l'ENSP pour garantir la confidentialité des répondants, y inclus l'élimination de l'information du plan échantillonnal, telle que les identificateurs de strates ou de grappes, du fichier.

Pour donner une idée de la qualité des données de l'ENSP, des tableaux de variabilités de l'échantillonnage approximatifs sont inclus dans la documentation du fichier de micro-données. Cependant ces tableaux sont très généraux et ne répondent pas aux besoins de l'utilisateur averti. Un logiciel de variance exacte qui calcule des variances pour des demandes spécifiques d'un usager est disponible, mais il doit être exécuté par le personnel de Statistique Canada dû à l'absence d'informations du plan échantillonnal dans le fichier de micro-données. Un projet de recherche a été entamé pour déterminer comment les strates du plan peuvent être groupées de telle façon qu'on puisse fournir aux usagers assez d'informations pour calculer une estimation raisonnable de la variance, tout en préservant la confidentialité des données.

---

[1] Jackey E. Mayda, Christopher Mohl and Jean-Louis Tambay, Statistics Canada, 16th floor, R.H. Coats Building, Ottawa, Ontario, K1A 0T6.

Cet article explique comment ce regroupement des strates a été accompli et présente les résultats d'une étude qui compare les variances calculées en utilisant l'information détaillé du plan aux estimations obtenues en utilisant les informations groupées du plan.

MOTS CLÉS: Fichiers microdonnées à grande diffusion; regroupement; réplicats.

1. INTRODUCTION

The National Population Health Survey (NPHS) is a longitudinal household survey being conducted by Statistics Canada. Data from the first wave of the survey were collected at four points in time during 1994-95. This first wave of the survey provided a panel of respondents who will be followed-up every two years for up to 20 years. The panel respondents were chosen by randomly selecting one person per surveyed household.

The sample design of the NPHS (Tambay and Catlin, 1995), used in all the provinces except Quebec, is based upon the sampling methodology of the redesigned Labour Force Survey (LFS). The LFS design is a stratified multi-stage sample of dwellings selected within clusters. The LFS usually selects six clusters in its regular strata, exceptions being some rural strata, remote strata and apartment strata. Under the LFS survey design, a cluster contains on average about 6 sample dwellings. The NPHS strata are groupings of LFS strata, with a subset or all of the clusters selected from an LFS stratum. In Quebec, the NPHS sample is selected from dwellings that participated in the *Enquête Sociale et de Santé* (ESS), a health survey conducted by *Santé Québec* in 1992-93. The design of that survey was a 2-stage stratified cluster sample similar to that of the LFS.

It is important that data providers supply measures of data quality such as sample variances for complex survey designs like the NPHS. For such multi-stage, clustered designs, simple formulas for exact variance calculations are not available and more sophisticated methods such as jackknifing or balanced repeated replication (BRR) must be used. To compute such variances, specific design information must be made available to the data users. For the NPHS, this is not a trivial problem due to the fact that for confidentiality reasons, the stratum and cluster information cannot appear on public use data files.

The paper is organized as follows. Section 2 describes the problem in more detail. Alternative methods that were discussed and the preferred solution to the problem are covered in sections 3 and 4. Section 5 describes the general rules used to create new strata for an empirical study, and section 6 presents the results of the study. The conclusions and areas for future work are provided in section 7.

2. DESCRIPTION OF THE PROBLEM

Two public use microdata files (PUMF) released in the fall of 1995 contain detailed information from the initial wave of the survey. The first file contains general demographic and health information for all members of the sampled households (approximately 50,000 individuals). The second file contains specific health information for the panel respondents only (approximately 17,500 individuals).

Under the Statistics Act (1970), public use microdata files may be released if the data cannot be associated with a particular respondent. The Microdata Release Committee at Statistics Canada reviews all PUMF before they are released to ensure that individuals' confidentiality is not being jeopardized. Measures such as top and bottom coding, grouping of response categories and removal of some variables that were deemed to reveal too much information were applied to the NPHS files before they were released. In addition, the design information such as stratum and cluster identifiers were not made available on the PUMF due to the extremely detailed level of geography they represented. Providing cluster information on PUMF could also allow the users to form households and thus significantly raise the possibility of identifying individuals. However, removing such information restricts the users' ability to calculate variance estimates.

In order to provide some measure of the quality of the NPHS data, approximate sampling variability tables were provided with the PUMF documentation (Statistics Canada, 1995). These tables are very general and are only useful to estimate variances of totals and rates for certain domains. Many users of the NPHS data apply logistic and linear regression techniques to the data and the tables are not appropriate in these situations. An exact variance program has been written which employs a jackknife variance estimator (Wolter, 1985). However, this program requires the detailed design information at the replicate level, which in most cases is the cluster for NPHS. Since this information is not on the PUMF, and there are no other data items on the PUMF that could be used to create replicates, users cannot calculate exact variance estimates. A research project was undertaken to try to find solutions to this problem so that the confidentiality of individuals could be preserved while allowing users to calculate reliable variance estimates.

3. ALTERNATIVES

Several options were discussed to examine how to balance the need for valid variance estimates while not giving away too much information on individuals. The first solution was to renumber the strata and cluster identifiers so that they are not meaningful. Although this is easy to do, it does not solve the problem since the detailed cluster information would still be available on the file. A user could still identify persons within the same cluster and/or household. This method would not pass the microdata release standards.

The next option that was considered was to create pseudo-strata and pseudo-clusters by regrouping individuals in the sample. The idea was to keep the same number of strata per province and a similar number of pseudo-clusters per pseudo-stratum, but create the pseudo-strata and clusters so that persons in the same household would not necessarily be in the same pseudo-stratum/cluster. This could be accomplished by forming pseudo-strata using groups of similar strata and forming pseudo-clusters from units in different clusters in the same set of strata. The advantages to this method are that the user would not be able to identify the original clusters with certainty, and that the pseudo-strata and clusters would mimic the original design. The disadvantage of this method is that detailed cluster identifiers are still on the files, which could lead a user to a conclusion about a particular household, even if it was a false one. There is also a concern about how to assign the pseudo-strata and clusters so that the resulting variance estimates would be accurate and unbiased.

4. THE PREFERRED SOLUTION

The next option that was investigated was based upon a method proposed by Rust (1986) in the context of reducing the time required for jackknifing. This method involves collapsing the design strata to form "super-strata" and then collapsing the original replicates together within the collapsed strata. This is done in such a way that the new replicates contain old replicates from the original strata. In addition, all of the old replicates become part of collapsed "super-replicates". Rust suggests that with relatively few collapsed replicates, variance estimates are close to those based upon the detailed design. The advantages of this method are many for NPHS. The NPHS design favours collapsing because the NPHS strata are already groupings of LFS strata. Also, the super-strata and super-replicates created using this method would contain a larger group of individuals than the original strata and clusters (replicates). Another advantage is that the original strata and replicates can be collapsed over geographic regions to form the super-strata and replicates, which creates a larger area that the user would have to identify. The main advantages of this method are that the original strata and clusters would not be present on the PUMF and that, under certain conditions, collapsing can allow unbiased variance estimates. The disadvantages of this method are that the exact variance corresponding to the original design will not be generated, and that collapsing reduces the number of degrees of freedom and in turn the precision of the variance estimate. It was felt that these disadvantages were surmountable and the effects could be measured by an empirical study.

Other issues must also be kept in mind with the implementation of this method. There is a fine balance between collapsing the original design strata and replicates enough so that the confidentiality of individuals is preserved, and yet producing valid variance estimates. The other issue facing the NPHS as a longitudinal survey is that the collapsing strategy should be effective for future waves, that is, how it is done and the number of dwellings in the resulting collapsed replicates should still bring useful results in future waves (after attrition, migrations and other events are observed).

5. COLLAPSING METHODOLOGY

The creation of original NPHS strata and replicates from the LFS and ESS designs was carefully conceived by keeping a number of rules and constraints in mind. Therefore, it was important that there be some reasonable procedures for creating super-strata and super-replicates rather than just a random assignment. Since Rust's paper does not give any specific rules, several guidelines were set out prior to the collapsing in order to ensure that the process took place in a logical manner. There were occasions, however, that some of the guidelines could not be followed exactly. Also, the modifications were different in Quebec compared to the other provinces due to the differing designs.

5.1 General Guidelines

1) Preserve replicate structure - Different NPHS strata had differing numbers of replicates created within them. When collapsing into super-strata, an attempt was made to combine only those NPHS strata with the same number of replicates. There would then be no need to renumber or re-create replicates within the super-strata, i.e., the first replicate from each of the original NPHS strata would be the first replicate within the super-

stratum, etc. In certain cases this was not possible because there were not enough NPHS strata with a given number of replicates to warrant a separate super-stratum. In these cases, the number of replicates was collapsed within the NPHS stratum and assigned to an existing super-stratum.

2) Preserve the urbanization of the stratification - Strata that belonged to the same 'population density' category were collapsed together when possible. This means that NPHS strata within cities or large towns were joined together rather than being combined with rural strata. The LFS and ESS stratification followed similar rules.

3) Preserve the geographic characteristics of the stratification - An attempt was made to collapse neighbouring NPHS strata together in the less populated areas as these strata tend to have similar characteristics.

5.2 Constraints

1) Sample Size and Confidentiality - The reason for creating super-replicates was to build groupings large enough so that more information on clustering could be disclosed without sacrificing the individual's identity. The super-replicates had to be large enough to preserve this confidentiality not only after the first wave, but as well in the future waves as the panel size diminishes because of attrition and migration. An attempt was made to set a threshold value based upon the number of clusters or dwellings originally selected from the survey. However, because of the differing designs in each of the provinces, it was not possible to give a single minimum sample size. Instead, a rule stating that there should be a minimum of about sixty households in a super-replicate after wave 1 was implemented.

2) Maximize the degrees of freedom - All of this collapsing means that there will be a much smaller number of replicates in the collapsed design. This in turn implies that the number of degrees of freedom will be very small at the provincial level. (The degrees of freedom are calculated as the number of replicates minus the number of strata.) The collapsing should reduce the degrees of freedom as little as possible.

6. EMPIRICAL STUDY

The empirical study used some NPHS data from the first wave to determine the impact of creating super-strata and super-replicates on the resulting variances and coefficients of variation (CVs). Variances were calculated using the jackknife method. The rules described in section 5 were used to generate super-strata and super-replicates in Quebec and the Atlantic provinces. Table 1 shows the original number of NPHS strata, replicates and degrees of freedom as well as the new values after the collapsing. Less collapsing takes place in Quebec than in the other provinces due to the larger size of the NPHS replicates in Quebec. For practical reasons, we concentrated on comparing the variances and CVs themselves. The effect of collapsing on the *variance* of the variance and CV estimates (proper collapsing should not yield biased variances) should also be evaluated.

Estimates were created for a number of totals, ratios and regression coefficients using both the original and collapsed designs. Some results are shown in Tables 2 and 3.

6.1 Variance estimates for totals and ratios

Totals were created at the provincial level for four variables: total number of people with activity restrictions, with food allergies, with other allergies and number of daily smokers. Provincial ratios for the average number of cigarettes smoked per day by daily smokers and an allergy ratio of people with food allergies compared to the total with other allergies were computed. These values were also calculated by sex within provinces (see Table 2). In the Atlantic provinces, about one-half of the provincial estimates of CVs increased when the replicates were collapsed while the other half remained the same or decreased slightly. Whether the CV goes up or down after collapsing does not appear to be affected by either the province or variable in question. In Quebec, the super-replicated results had a slightly higher variance than the true variance in most cases. Usually the differences in CVs were very small and in no instance would the releasability of the results be affected.

For the estimates by sex, the number of CVs that increase or decrease under collapsing is similar to that observed for the provincial estimates. Notice that in most cases the collapsed CV is higher at the sex level compared to the provincial level. The differences in CVs are more variable at the sex level due in part to the smaller sample size.

6.2 Variance estimates for linear regression models

Table 3 compares CVs for linear regression coefficients under collapsed and uncollapsed designs. The linear models that are used in the study predict the value of a person's health status (a continuous variable between zero and one) based upon four variables - restriction of activities, age group (4 levels), type of

drinker and household income (5 levels). In most of these cases, the collapsed replicates give smaller CVs for the regression coefficients than the original ones. Large differences in CVs are seen when the original CVs are very high. However, when the original CVs are moderate or low, the collapsed CVs are close in value. The most stable values are seen in Quebec. In this province, unlike in the case of totals and rates, there is not a tendency for the collapsed CVs to be larger than the originals. Note that collapsing severely reduces the number of parameters that the models can use. Newfoundland, with six degrees of freedom, has thus attained the maximum allowable number of parameters to estimate (5). This leaves only one degree of freedom for the error term, a situation that will not generally be allowed by researchers.

7. CONCLUSIONS AND FUTURE WORK

Under collapsing, no systematic effect on the estimated variances for totals and rates was noted. The average absolute difference between the uncollapsed and collapsed estimated CVs was approximately 1.3% across the provinces studied. In fact, most of the differences in the CVs were less than 2%. The largest differences occurred in the Atlantic provinces. This was due in part to the collapsing, which created a low number of degrees of freedom in each province. For the provinces and variables that were examined in the study, if the CVs using the collapsed strata were available to users, it would not change the users' ability to release any of the results. This is based upon a release criterion of a CV of less than 33% for estimates from survey samples.

In the case of linear regression coefficients, the estimated variances under collapsing tend to be smaller than the design variances. This may be simply due to the choice of the model for the study. If this approach is to be used in practice, more studies will need to be done using linear and logistic regression models.

This is a promising approach that has the added benefit of reducing the jackknifing time. It is felt that this method does preserve confidentiality without affecting the variance estimate negatively. This approach could also be useful for other surveys with complex sample designs that release PUMF.

There are a couple of cautions that would be provided to users if they were to apply this method. First, the results seem to be slightly unstable for the rarer characteristics (such as food allergies), so users would have to be careful interpreting the estimates for those kinds of analyses. In addition, the differences between the collapsed and design variances for subsets of the data (such as by sex) are more variable due to smaller sample size. It would be suggested that the collapsed estimates be calculated at the regional level (Atlantic provinces, Quebec, Ontario, Western provinces) rather than by province. For provincial or sub-provincial estimates, the design information should be used as opposed to the collapsed strata, since the degrees of freedom have been reduced and the sample sizes are smaller. This would involve Statistics Canada personnel running the exact jackknife variance program.

The next step in this project is to present the approach to Statistics Canada's Microdata Release Committee to see if it complies with their standards. If this approach is approved, the collapsing methodology will be applied to the other provinces. If this approach is not feasible, an alternative is the Remote Access Project at Statistics Canada. Under this scenario, users would be able to have indirect access to an "enhanced" public use microdata file that contains the design information plus other confidential variables. Computer programs written by the user to compute estimates from this file would be transmitted electronically to Statistics Canada via the internet and applied to the data by Statistics Canada personnel. Once the programs have been run and results verified to ensure that no confidential information is disclosed, the user would receive the output.

REFERENCES

Rust, K. (1986). "Efficient replicated variance estimation", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 81-87.

Statistics Act, 1970 - 71 - 72, c. 15, s. 1.

Statistics Canada (1995). *National Population Health Survey (1994-1995)*, Public Use Microdata File documentation, Statistics Canada Catalogue No. 82F0001XDB.

Tambay, J.-L., and Catlin, G. (1995). "Sample design of the National Population Health Survey", *Health Reports*, Statistics Canada Catalogue No. 82-003, Volume 7 (1), 33-42.

Wolter, K. (1985). *Introduction to Variance Estimation,* New York: Springer Verlag, 153-195.

Table 1
Number of strata and replicates before and after collapsing

Province	Original Design			Collapsed Design		
	# NPHS	# NPHS	# degrees	# super-	# super-	# degrees
Newfoundland	48	186	138	6	12	6
PEI	29	200	171	4	12	8
Nova Scotia	75	196	121	5	12	7
New Brunswick	65	193	128	4	10	6
Quebec	37	123	86	15	41	26
Total	**254**	**898**	**644**	**34**	**87**	**53**

Table 2
Comparison of CVs for Totals and Ratios

			Both Sexes			Males			Females		
		Estimate (rounded)	CV Uncollapsed	CV Collapsed	Difference Uncoll-Coll	CV Uncollapsed	CV Collapsed	Difference Uncoll-Coll	CV Uncollapsed	CV Collapsed	Difference Uncoll-Coll
Newfoundland	restricted	73200	8.39	8.67	-0.28	10.70	10.19	0.51	12.47	11.17	1.30
	food allergies	22700	18.84	16.95	1.89	31.37	22.19	9.18	21.18	22.68	-1.50
	other allergies	56200	9.98	9.52	0.46	17.74	14.29	3.45	14.06	12.30	1.76
	daily smoker	124200	6.84	7.84	-1.00	8.91	10.32	-1.41	9.61	7.46	2.15
	avg cigarettes	17.09	3.81	3.13	0.68	5.05	4.83	0.22	4.55	3.34	1.21
	allergy ratio	0.40	17.57	14.78	2.79	30.59	22.91	7.68	12.39	20.64	-8.25
Prince Edward Island	restricted	24100	6.34	6.30	0.04	9.32	7.68	1.64	9.12	5.90	3.22
	food allergies	5100	15.70	15.20	0.50	29.07	20.46	8.61	18.91	20.93	-2.02
	other allergies	16100	9.96	9.60	0.36	19.48	22.50	-3.02	11.38	12.66	-1.28
	daily smoker	29700	6.30	7.17	-0.87	7.23	7.46	-0.23	9.52	11.20	-1.68
	avg cigarettes	20.10	3.26	2.04	1.22	4.38	3.85	0.53	5.10	4.24	0.86
	allergy ratio	0.32	16.15	14.21	1.94	29.29	26.57	2.72	19.64	20.20	-0.56
Nova Scotia	restricted	210800	6.74	8.18	-1.44	9.54	8.07	1.47	8.50	12.70	-4.20
	food allergies	45200	18.08	22.21	-4.13	27.93	19.89	8.04	21.46	26.52	-5.06
	other allergies	122400	8.10	9.67	-1.57	13.51	8.41	5.10	10.00	12.66	-2.66
	daily smoker	208200	6.17	5.23	0.94	8.59	9.51	-0.92	8.78	6.40	2.38
	avg cigarettes	18.72	3.17	2.28	0.89	4.30	3.77	0.53	4.65	3.95	0.70
	allergy ratio	0.37	15.50	15.68	-0.18	25.59	14.99	10.60	17.61	18.96	-1.35
New Brunswick	restricted	122700	6.61	4.24	2.37	9.38	7.06	2.32	9.24	7.98	1.26
	food allergies	47800	11.65	9.78	1.87	21.10	13.85	7.25	14.07	11.15	2.92
	other allergies	109000	7.36	12.45	-5.09	10.98	7.73	3.25	9.21	16.46	-7.25
	daily smoker	164100	6.54	5.59	0.95	9.99	10.84	-0.85	8.11	5.15	2.96
	avg cigarettes	18.97	3.08	3.92	-0.84	3.91	4.04	-0.13	3.84	4.29	-0.45
	allergy ratio	0.44	10.09	6.83	3.26	20.03	12.37	7.66	14.84	14.11	0.73
Quebec	restricted	1013900	6.41	6.32	0.09	8.18	8.61	-0.43	7.00	6.32	0.68
	food allergies	156300	15.54	16.44	-0.90	20.08	21.19	-1.11	21.47	23.36	-1.89
	other allergies	862200	5.70	6.50	-0.80	9	9.79	-0.79	7.42	6.00	1.42
	daily smoker	1752000	3.66	4.34	-0.68	5.16	5.38	-0.22	5.52	6.37	-0.85
	avg cigarettes	19.90	2.04	2.43	-0.39	2.67	2.85	-0.18	2.75	3.58	-0.83
	allergy ratio	0.18	15.63	16.01	-0.38	19.56	19.67	-0.11	21.58	23.29	-1.71

Table 3

Comparison of CVs for Linear Regression Parameters

		Estimate	CV Uncollapsed	CV Collapsed	Difference Uncoll-Coll
Newfoundland	intercept	0.631	6.70	6.31	0.39
	restricted	0.140	12.90	11.30	1.60
	age group	-0.010	40.93	36.93	4.00
	type of drinker	0.011	84.91	66.97	17.94
	household income	0.006	25.94	19.27	6.67
Prince Edward	intercept	0.674	4.89	4.83	0.06
Island	restricted	0.139	11.43	10.05	1.38
	age group	-0.022	16.85	14.26	2.59
	type of drinker	-0.008	123.85	93.17	30.68
	household income	0.006	32.63	36.37	-3.74
Nova Scotia	intercept	0.617	6.23	5.96	0.27
	restricted	0.156	9.91	7.36	2.55
	age group	-0.017	29.31	27.93	1.38
	type of drinker	-0.005	231.01	167.44	63.57
	household income	0.005	50.11	37.14	12.97
New Brunswick	intercept	0.596	5.84	8.07	-2.23
	restricted	0.153	9.06	10.61	-1.55
	age group	-0.017	22.65	19.38	3.27
	type of drinker	0.020	50.17	47.17	3.00
	household income	0.006	26.17	46.21	-20.04
Quebec	intercept	0.684	3.28	3.25	0.03
	restricted	0.129	8.13	7.34	0.79
	age group	-0.023	11.20	12.72	-1.52
	type of drinker	0.020	30.08	30.53	-0.45
	household income	0.003	32.69	23.81	8.88