12F0041
1995
no.3
c.3

STATISTICS CANADA STATISTIQUE CANADA
NOV 20 1995
LIBRARY BIBLIOTHÈQUE

Newsletter of the Data Analysis Resource Centre

Issue Number 3 Statistics Canada November 1995

In This Issue

1. Montreal Workshop
2. Course on Complex Survey Data Analysis
3. DARC and the Law
4. Ask Dr. DARC
5. UNIX Tips

Montreal Workshop

On July 9, 1995, DARC members David Binder and Pierre St-Martin gave a one-day workshop in Montreal on "Data Analysis from Complex Surveys." This workshop was sponsored by the Survey Methods Section of the Statistical Society of Canada, and was part of a series of workshops offered at the SSC Annual Meeting. The 16 very interested and enthusiastic participants came from Bureau de la statistique du Québec, University of Western Ontario, Concordia University, Montreal General Hospital, the National Center for Education Statistics in Washington, various universities in Canada and the United States and from Statistics Canada. David and Pierre plan to give a similar workshop in Ottawa soon, jointly sponsored by the Statistical Society of Ottawa and the Survey Methods Section of the SSC.

New Course on Statistical Methods of Complex Survey Data Analysis

Many social and economic survey data are often collected using complex sample designs that include multi-stage unequal selection probabilities, stratification and clustering. A well-known example is the Canadian Labour Force Survey from which many other social surveys draw their samples.

This departure from simple random sampling requires special methods for variance estimation for the construction of confidence intervals, hypothesis testing and model construction.

To illustrate our topic, consider the following example from Thomas (1991).

Consider the following 12 observations:

i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	6.0	2.0	4.5	1.0	3.0	1.5	5.0	0.5	4.0	5.5	3.5	2.5

We wish to test the hypothesis that the mean of the population from which these observations were drawn equals 2, i.e. we wish to test the null hypothesis $H_0: \mu_y = 2$ against the alternative $H_1: \mu_y \neq 2$.

The classical statistical approach, based on the assumption that the twelve observations are independent, yields the statistic $t_0 = 2.40$. With type I error probability of 5%, the null hypothesis is rejected since t_0 exceeds the critical value (from Student's t distribution) $t_{11;0.05} = 2.2$.

Suppose however, that these twelve observations had been selected by a simple random sample of n=3 clusters:

	i	y_i	Cluster Mean
	1	6.0	
Cluster	3	4.5	$\bar{y}_1 = 5.25$
1	7	5.0	
	10	5.5	
	9	4.0	
Cluster	11	3.5	$\bar{y}_2 = 3.25$
2	12	2.5	
	5	3.0	
	8	0.5	
Cluster	4	1.0	$\bar{y}_3 = 1.25$
3	6	1.5	
	2	2.0	

Clearly, the assumption that the twelve observations are independent, is not valid. Rather, it is the cluster means which should be considered as independent observations. Considering the same hypothesis as before, we get this time the value $t_0 = 1.08$ with the critical value $t_{2;0.05} = 4.3$. In this case, taking the sampling design into consideration, the null hypothesis cannot be rejected. The difference between the two results is explained by the fact that when the clustering effect and the intra-cluster correlation are ignored, we under-estimate the variance by a factor of 4.9. The inflation factor of the variance, introduced by the presence of clustering, is referred to by the term **design effect**. Taking the design effect into account is necessary to make many statistical analyses valid.

We have seen a very simple example of the dangers awaiting every user of complex survey data who wishes to apply classical statistical methods using software packages such as SAS and SPSS. To help statisticians and analysts carry out statistical analyses more appropriately, DARC has developed a new course dealing with many issues of the analysis of data from complex surveys. A pilot version of the course was offered in May 1995 in French, and is now available in both English and French. A detailed description of the course may be found in the Statistics Canada Course Catalogue under STC 4464.

Reference

Thomas, D.R.(1991), **Inference Using Complex Data from Surveys and Experiments**, Carleton University, Working paper series no. 91-4.

DARC and the Law

During the past couple of decades, statisticians have participated in the legal process with increasing frequency, both as expert witnesses and as consultants. This trend can likely be attributed to many factors, such as the increasing amount of information relevant to legal cases that needs statistical interpretation; the increased availability of computers which makes the access to and manipulation of statistical data easier and cheaper; and the increased ease lawyers have with statistical arguments. However, quoting from Statistics and the Law (ed. M. DeGroot, S. Fienberg, and J. Kadane, John Wiley and Sons, 1992),

> "There is a fundamental difference in outlook between lawyers and statisticians ...Statisticians are trained in the scientific method. It is their profession to collect and analyze data in such a way as to give them the deepest and most complete understanding possible of the processes they are studying...They are trained to recognize the uncertainties of their conclusions and to quantify these un certainties...Lawyers on the other hand are hired to present as strong a case as possible for their party, consistent with whatever facts and data are available to them, and ethical legal tenets."

DARC became involved in a legal dispute - and thus got first hand knowledge of the use of statistics in the courts in Canada - when Georgia Roberts and Carleton University's Prof. J.N.K. Rao were asked to be expert witnesses in a case

involving Canada's refugee determination system. A new inland refugee determination system came into effect in Canada in January 1989. By this system, any refugee claimant who is not allowed to remain in Canada on humanitarian and compassionate grounds must proceed to a two-stage inquiry process. If a negative decision is reached in this process, a claimant's only recourse is to apply for permission to appeal to the Federal Court of Appeal. This permission is granted or rejected by a single judge of that court. A legal dispute arose when a refugee claimant who had been refused permission to appeal filed a case against the Department of Employment and Immigration claiming that the permission granting stage of the refugee determination system was unjust. The major piece of evidence entered by the claimant was a statistical study conducted by two political scientists, comparing the rates of granting permission observed from 12 different Court of Appeal judges in 1990. DARC and Prof. Rao were asked by the Department of Justice (the lawyers for E and I) to become involved in the case through the provision of critiques of the original statistical study and of subsequent statistical work carried out for the side of the refugee claimant. Among the points of criticism made, two were about aspects of data analysis that are emphasized in courses on statistical methods but are frequently overlooked in practice:

(1) The randomization assumptions underlying the statistical methods used did not hold for the data collected. Therefore, inferences made from the analysis do not necessarily follow.

(2) When using a model fitted to data to draw conclusions, the model is assumed to be a good description of the data. It is therefore essential to do a thorough study of the goodness of fit of the model chosen. This was not done.

Therefore, it was argued that, while the statistical studies would be acceptable as social science research, where others would have opportunity to confirm or refute the claims of the studies, they were not of sufficient quality to be used as evidence.

The impact of the statistical work done by both sides in the resolution of the case is yet to be seen - since the case has been moved from provincial to federal court, and could wait years to be heard.

ASK DR. DARC

This is the second in a series of articles in this newsletter. In this series we answer readers' questions or other commonly asked questions on data analysis. If you have a question, please send it to the DARC Newsletter Editor, M. Feder, Household Survey Methods Division, RHC 16-L.

DARC's Newsletter was read even in Australia... and has prompted questions from Prof. Danny Pfeffermann (on Sabbatical at the Australian Bureau of Statistics from the Hebrew University in Jerusalem).

> I read Dr. DARC's answer to the question about "*target population*" or "*population of interest*" and I think that he is wrong there. What he refers to is the distinction between the finite population and the superpopulation which is something different. The target population is the *finite* population for which you like to make your inference. The sample may actually represent another population. For example, in case of nonresponse, the target population is the finite population from which the sample is taken but the sample only represents the respondents. You may take a sample from a big population but then be interested in a subpopulation.

DR. DARC ANSWERS: I thank Prof. Pfeffermann for his comments to the Dr. DARC article in the last issue. I would encourage more people to submit comments and questions.

Prof. Pfeffermann criticizes my use of the term *target population* to refer to a conceptual infinite population. It is true that in the usual sampling literature, the target population is the finite population about which we wish to make inferences; for example, the confidence intervals are with respect to some finite population quantity such as the finite population total. The literature, however, does not indicate that the target population must be finite. Särndal, Swensson and Wretman (p.13) define target population as "the set of elements about which information is wanted and parameter estimates are required". They do make the point that frame imperfections can make the target population different than the *frame population*. Prof. Pfeffermann also points out that non-response can also make the sample non-representative.

Satin and Shastry (p.7) define the target popula-

tion as the "population for which information is required". Cochran uses a similar definition. It is true that all these authors are addressing the problem of sampling from a finite population, but for data analytic requirements, the parameter of interest could be the infinite population quantity, such as a regression coefficient. In this case the target population should be the infinite one.

Prof. Pfeffermann also raises the question of what is a superpopulation. Normally superpopulations are introduced to model the values of a finite population as coming from a stochastic process. Such superpopulations are often (but not always) used for model-based inference. I feel that for some analytical inferences, it is this superpopulation which becomes the target population. Perhaps the easiest way to avoid this semantic controversy is to use *population of interest* instead.

> QUESTION: When I look at all the standard references on sampling, the discussions on optimizing or improving the sample design focus on means and totals, but no mention is made of ratios, regression coefficients or other more complex non-linear measures. How should I design a survey for estimating these non-linear quantities?

DR. DARC ANSWERS: This question is an important one and often not given proper treatment, even though the answer is quite straightforward. First of all, we must generally assume that the sample size is sufficiently large that the asymptotic versions of the mean square error can be derived. The mean square error (or, equivalently, the variance for unbiased estimators) is the most commonly used measure for assessing the performance of an estimator in sampling theory. This is certainly an excellent measure for large-scale surveys where the central limit theorem implies that the estimator is approximately normally distributed. We often develop a sampling strategy that minimizes the mean square error of the estimator of a population mean or total for a particular variable, or a set of variables.

For the most common case where the non-linear statistic is a differentiable function of the estimators of population means of one or more variables, the mean square error can be derived easily and explicitly. The mean square error can also be shown to be equivalent to the mean square error of an estimated mean based on a linearized variable. Therefore to answer your question, we can apply all the methods used for getting an appropriate design for estimating population means, but using instead this linearized variable.

Let me give a couple of examples. First, suppose that we are interested in estimating $\ln(\bar{Y})$ where $\bar{Y}$ is the mean of the variable Y, and that we estimate this mean by $\hat{\bar{Y}}$, using sample values, $y_1, y_2, \ldots, y_n$. It can be shown that $\ln\left(\hat{\bar{Y}}\right)$ has an asymptotic variance that is equivalent to that of an estimated mean based on sample values $z_1, z_2, \ldots, z_n$, where $z_i = y_i / \bar{Y}$. In this case optimizing on the y's to estimate $\bar{Y}$ is equivalent to optimizing on the z's to estimate $\ln(\bar{Y})$.

A more complex example is estimating a ratio. Suppose we estimate a ratio by $\hat{R} = \hat{Y}/\hat{X}$. The asymptotic variance is now equivalent to that of an estimated mean, using sample values $z_1, z_2, \ldots, z_n$, where $z_i = (y_i - Rx_i)/\bar{X}$ and $R = Y/X$. Therefore, we would choose a design to minimize the asymptotic variance of a population total based on the z_i's.

The general methods for deriving these linearized variables is beyond the scope of this column. If you need help in these matters, please contact any member of DARC.

UNIX Tips

The graphical interface to UNIX is X and most of what we do in UNIX is accomplished via this interface i.e. via the windows and fonts provided by X. The windows and fonts are automatically scaled to the screen size and resolution of the display (monitor) we use. As a result, the view on screen may not be what is needed or preferred. Following is a way of choosing a font.

Font availability depends on the X server being used because the server provides access to its own set of fonts and/or to a remote font server. The X server you use depends on your means of access to the UNIX workstations. Since DARC is standardizing on access via PCs using "Xoftware for Windows" from AGE, this procedure is for Xoftware but is very similar to what would be

required with any similar package.

1. Login to darc1 (darc2 requires knowing Sun's path for X commands)
2. Enter the command: `xfontsel`
At the top right you will see the count of the number of fonts available
3. Left click and hold on: `fndry`
4. Slide down the list to make a choice then let go of the button
5. Repeat for each of the mnemonics
NOTE: You will see the sample change as you make choices. Be careful. Most windows are x-terms which require monospaced fonts. If results are peculiar, you have chosen a proportional font.

The mnemonic `spc` indicates:

c	char cell/monospaced
m	monospaced
p	proportional
any #	scaled font (only in X11 R6)

6. When your choice is made, click on "select".
Now edit your `.xsession` file
(`.xsession-darc1` and
`.xsession-darc2` in DARC)
7. Insert into the xterm (or other window) command: `-fn -fb`
8. Place your cursor after `-fn` and click the middle button to paste the font.
You should use the same font for both `-fn` and `-fb` except that for `-fb` choose bold for: `wght`.
You may create an alias for each font you select or wish to use and place it in:
`fontpath/fonts.alias.`
If you take care to choose the same font across platforms, you will get a consistent look.
darc1 and dc2s run a font server. To access these fonts too, add them to your font path:
For just the current session:
`xset +fp tcp/darc1:7000/`
Permanently:
1. Be sure you have no other job running because restart is automatic
2. Click on the icon of the active session to obtain the main menu
3. Select `options | fonts`
4. Add to the font path as follows:
`,tcp/darc1:7000/`
5. Repeat the above, if desired, for each font server.
Use a different port for each tcp - 7001, 7002, etc.
Confirm the entry with: `xset -q`

Near the bottom of the output you will see:
Font Path: then the path

For help with changing fonts for SAS, call Reid Gregson at 951-1388.

If you wish to be on our mailing list, please send an e-mail request to me or call.
Georgia Roberts (951-1471)

DARC's Manager	Mike Hidiroglou (951-4767)
Newsletter Editor	Moshe Feder (951-0795)

For general information on DARC and its services, contact Mike Hidiroglou (951-4767) or Georgia Roberts (951-1471).

DARC Members

Mike Hidiroglou, BSMD (951-4767)
Georgia Roberts, BSMD (951-1471)
Moshe Feder, HSMD (951-0795)
Milorad Kovačević, HSMD (951-9891)
Harold Mantel, HSMD (951-4150)
Zdenek Patak, BSMD (951-9434)
Boxin Tang, BSMD (951-0498)
Paul Wong, BSMD (951-0049)
Wesley Yung, HSMD (951-4699)
Advisor: David Binder, BSMD (951-0980)

c 3

STATISTICS CANADA STATISTIQUE CANADA
NOV 20 1995
LIBRARY BIBLIOTHÈQUE

Bulletin du

Centre de ressources en analyse de données

Numéro 3 **Statistique Canada** **Novembre 1995**

Dans ce numéro

1. L'Atelier de Montréal
2. Le cours traitant des méthodes statistiques pour l'analyse des données provenant d'enquêtes à plan complexe
3. CRAD et la Loi
4. Demande à DR CRAD
5. Conseils pour UNIX

L'Atelier de Montréal

Le 9 juillet, 1995, deux des membres du CRAD David Binder et Pierre St - Martin ont présenté un atelier à Montréal sur "Analyse des données pour les sondages complexes". Cet atelier était parrainé par la Section des méthodes de sondage de la Société statistique du Canada (SSC), et faisait partie d'une série d'ateliers offerts à l'assemblée annuelle de la SSC. Seize participants très enthousiastes et intéressés provenant du Bureau de la statistique du Québec, de l'Université de Western Ontario, de l'Université de Concordia, de l'Hôpital général de Montréal, du Centre national pour les statistiques de l'éducation à Washington, et de diverses universités à travers le Canada et les Etats-Unis et de Statistique Canada ont participé à cet atelier. David et Pierre ont l'intention de présenter bientôt un atelier similaire à Ottawa, qui sera conjointement parrainé par la section de la Société statistique d'Ottawa et la Section des méthodes de sondage de la SSC.

Nouveau cours traitant des méthodes statistiques pour l'analyse de données provenant d'enquêtes à plan complexe

Pour des raisons théoriques et pratiques, les données de nombreuses enquêtes sociales ou économiques sont souvent obtenues à partir de plans d 'échantillonnage complexes incluant des processus de sélection à plusieurs degrés avec probabilités variables, des méthodes de stratification ou encore l'utilisation de grappes. À titre d'illustration, on n'a qu'à penser à la complexité du plan d 'échantillonnage de l'Enquête sur la population active dont dépendent un certain nombre d'autres enquêtes sociales qui y puisent leurs échantillons.

Il ne s'agit plus alors d'échantillonnage aléatoire simple et ces enquêtes complexes requièrent des méthodes spéciales pour l'estimation de la variance ainsi que pour l'inférence statistique telles que la construction d'intervalles de confiance, les tests d'hypothèses ou la construction de modèles. Pour bien illustrer notre propos, considérons

10102038 25
STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA

l'exemple suivant tiré de l'article de Thomas (1991).

On dispose des 12 observations suivantes:

i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	6,0	2,0	4,5	1,0	3,0	1,5	5,0	0,5	4,0	5,5	3,5	2,5

On désire vérifier l'hypothèse selon laquelle la moyenne de la population dont on a tiré ces observations est égale à deux. En fait, on désire confronter l'hypothèse $H_0 : \mu_y = 2$ contre l'alternative $H_1 : \mu_y \neq 2$.

Selon l'approche statistique classique fondée sur l'indépendance des 12 observations, on obtient une valeur de la statistique $t_0 = 2{,}40$. En acceptant un risque d'erreur de première espèce de 5%, on doit rejeter l'hypothèse nulle puisque la valeur critique lue dans une table de Student est égale à $t_{11;0,05} = 2{,}2$.

Supposons maintenant qu'en réalité, ces douze observations auraient été obtenues par l'entremise d'un plan d'échantillonnage aléatoire simple de n=3 grappes selon la représentation suivante:

i	yi		moyenne de grappe
	1	6,0	
Grappe	3	4,5	$\bar{y}_1$ = 5,25
1	7	5,0	
	10	5,5	
	9	4,0	
Grappe	11	3,5	$\bar{y}_2$ = 3,25
2	12	2,5	
	5	3,0	
	8	0,5	
Grappe	4	1,0	$\bar{y}_3$ = 1,25
3	6	1,5	
	2	2,0	

L'hypothèse classique d'indépendance des 12 observations n'est évidemment plus valide. Ce sont plutôt les moyennes des grappes qui doivent être considérées comme des observations indépendantes. En reprenant le test d'hypothèses précédent, on obtient cette fois-ci une valeur $t_0 = 1{,}08$ et une valeur critique $t_{2;0,05} = 4{,}3$. Dans ce cas, lorsque l'on tient compte du plan d'échantillonnage, l'hypothèse nulle ne peut être rejetée. La différence entre les résultats s'explique par le fait que lorsqu'on ne tient pas compte de l'effet des grappes et de la corrélation intra-grappe des observations qui en résulte, on sous-estime la variance d'un facteur égal à 4,9. Le facteur de gonflement de la variance introduit par la présence de grappes est ce que l'on appelle habituellement **effet de plan**. Il va de soi que la prise en compte de l'effet de plan s'avère essentielle à toute analyse statistique valide.

Voilà un exemple très simple des dangers qui guettent tout utilisateur de données d'enquêtes ayant des plans d'échantillonnage complexes, désireux de mettre en application les méthodes statistiques classiques par l'entremise de logiciels tels que SAS, SPSS ou autres. Dans le but d'aider les statisticiens et les analystes à effectuer des analyses statistiques plus appropriées, le CRAD a élaboré un nouveau cours de formation traitant des nombreuses questions liées à l'analyse des données provenant d'enquêtes à plans complexes. Le cours a été offert en version pilote au début du mois de mai et est maintenant disponible en anglais et en français. Vous trouverez une description détaillée du cours dans le Répertoire de cours de Statistique Canada, sous la cote STC4464.

Référence

Thomas, D.R. (1991), **Inference Using Complex Data from Surveys and Experiments**, Carleton University, Working paper series no. 91-4.

Le CRAD et la loi

Au cours des dernières décennies, les statisticiens ont de plus en plus participé au processus légal, à titre de témoins experts et de conseillers. Cette tendance peut vraisemblablement être attribuée à plusieurs facteurs, telle que la croissance de l'information pertinente aux causes légales qui requièrent une interprétation statistique ; une plus grande disponibilité d'ordinateurs qui rend plus facile et moins chère l'accès et la manipulation de données statistiques; de plus en plus d'avocats qui se sentent de plus en plus à l'aise avec les arguments statistiques. Cependant, citant Statistics and the law (ed. M. DeGroot, S. Fienberg, et J. Kadane, John Wiley et Son, 1992),

> "Il y a une différence fondamentale entre la vision des avocats et celle des statisticiens. .. Les statisticiens sont formés dans la méthode scientifique. Il est de leur profession de recueillir et analyser des données afin de donner une compréhension plus profonde et complète aux processus qu'ils étudient. .. Ils sont formés afin de reconnaître les incertitudes de leurs conclusions et à les évaluer..... Les avocats, par contre, sont embauchés afin de présenter la cause de leur client aussi bien que possible, en tenant compte des faits et des données qui leur sont disponibles, et des principes légaux éthiques."

Le CRAD s'est impliqué dans une dispute légale, et s'est donc familiarisé pour la première fois avec l'utilisation des statistiques dans les cours du Canada - lorsqu'on a demandé à Georgia Roberts et au professeur J.N.K. Rao de l'Université de Carleton d'être témoins experts dans un cas impliquant le système de détermination des réfugiés au Canada.

Un nouveau système interne pour déterminer du statut de réfugié à été mis en vigueur en janvier 1989 au Canada. Selon ce système, tout individu demandant le statut de réfugié, et auquel il n'est pas permis de rester au Canada sur des bases humanitaires et compatissantes, doit suivre une procédure d'enquête à deux étapes. Si cette procédure aboutit à une décision négative, alors le seul recours du demandeur est de demander la permission de faire un appel à la Cour d'appel fédérale du Canada. Cette permission est accordée ou rejetée par un seul juge de cette cour. Une dispute légale est survenue lorsqu'on a refusé à un demandeur du statut de réfugié la permission de faire appel, et il a conséquemment entamer une poursuite contre le ministère de l'Emploi et de l'Immigration (ELTI) en disant que la procédure reliée à l'étape accordant la permission d'appeler était injuste. La principale preuve présentée par le prétendant était une étude statistique faite par deux chercheurs en sciences politiques, comparant les taux de permissions accordées pour 12 juges différents de la Cour d'appel en 1990. Le ministère de la Justice (les avocats pour E et I) a demandé au CRAD et au professeur Rao de s'impliquer dans cette cause afin d'élaborer une critique de l'étude statistique originale et du travail statistique subséquent exécuté par la partie du demandeur du statut de réfugié.

Parmi les critiques avancées, deux points souvent mentionnés lors des cours sur les méthodes statistiques, mais fréquemment oubliés dans la pratique, étaient reliés à l'analyse des données:

> (1) Les suppositions de randomisation sous-jacentes aux méthodes statistiques utilisées ne tenaient pas pour les données recueillies. Donc, les inférences subséquentes de l'analyse n'étaient pas nécessairement vraies.
> (2) Lorsqu'on ajuste un modèle à des données afin d'arriver à des conclusions, on suppose que le modèle décrit bien les données. Il est donc essentiel de faire une étude complète de la validité du modèle choisi. Ceci n'avait pas été fait.
> Donc, il a été prétendu que, quoique l'étude statistique soit acceptable à des fins de recherche pour les sciences sociales, où d'autres auraient l'occasion

de confirmer ou de réfuter les revendications de l'étude, elle n'était pas de qualité suffisante pour être utilisée en cour.

Il reste à voir l'impact du travail statistique puisque ce cas a été porté de la cour provinciale à la cour fédérale, et on pourrait attendre des années avant que la couse soit entendue.

DEMANDER à DR CRAD

Ceci est la deuxième discussion dans une série d'articles dans ce bulletin. Dans cette série, nous répondons aux questions des lecteurs ou autres questions normalement posées sur l'analyse des données. Si vous avez une question, envoyer la au Rédacteur du bulletin de CRAD, M. Feder, gestionnaire de DMEM, RHC 16-L.

Le Bulletin du CRAD est même lu en Australie. .. et a incité des questions du professeur Danny Pfeffermann (en année sabbatique au Bureau Australien de statistiques de l'université Hebrew à Jérusalem).

J'ai lu la réponse du DR CRAD au sujet de la question *"population cible"* ou *"population d'intérêt"* et je pense qu'il a tort.
Le point auquel il fait allusion est la distinction entre la population finie et la superpopulation, qui sont deux idées différentes. La population cible est la population finie pour laquelle on voudrait faire une inférence. En réalité l'échantillon peut représenter une autre population. Par exemple, en cas de non- réponse, la population cible est la population finie où on pige l'échantillon, mais cet échantillon ne représente que les unités répondantes. Vous pouvez tirer un échantillon d'une population de grande taille, mais être intéressé à un sous-population.

DR CRAD RÉPOND: Je remercie le professeur Pfeffermann pour ses commentaires à DR CRAD dans l'article du numéro précédent. J'encourage les lecteurs à soumettre plus de commentaires et questions.

Le professeur Pfeffermann critique mon emploi du terme population cible pour référer à une population infinie conceptuelle. Il est vrai que dans la littérature traditionnelle sur l'échantillonnage, la population cible est la population finie pour laquelle on veut faire des inférences; par exemple, les intervalles de confiance sont associés à un certain paramètre de la population finie telle que le total de la population. La littérature, cependant, n'indique pas que la population cible doit être finie. Särndal, Swensson et Wretman (p.13) définissent la population cible comme" l'ensemble des éléments pour lesquels on désire de l'information et des estimations de paramètres ". Ils font aussi la remarque que des imperfections peuvent rendre la base de sondage différente de la population cible. Le professeur Pfeffermann indique aussi que la non-réponse peut aussi rendre l'échantillon non représentatif. Satin et Shastry (p.7) définissent la population cible comme la" population pour laquelle l'information est nécessaire ". Cochran emploie une définition similaire. Il est vrai que tous ces auteurs examinent le problème d'échantillonner une population finie. Mais pour des fins analytiques, le paramètre d'intérêt pourrait être celui de la population infinie, telle qu'un coefficient de régression. Dans ce cas la population cible devrait être la superpopulation. Le professeur Pfeffermann soulève aussi la question de ce qu'est une "superpopulation". D'habitude les "superpopulations" sont introduites afin de modéliser les valeurs de la population finie comme provenant d'un processus stochastique. De telles superpopulations sont souvent (mais pas toujours) employées pour les inférences basées sur un modèle. Il me semble que pour certaines inférences analytiques, c'est la superpopulation qui devient la population cible. Peut-être qu'une façon plus facile d'éviter cette controverse

sémantique serait d'utiliser plutôt l'expression "population d'intérêt".

QUESTION: Lorsque je lis les références habituelles sur la théorie de l'échantillonnage, les discussions sur l'optimisation ou l'amélioration du plan de sondage se rapportent à des moyennes ou à des totaux, mais on ne discute jamais de proportions, de coefficients de régression ou d'autres mesures non linéaires plus complexes. Comment devrais-je concevoir un plan de sondage afin d'estimer ces quantités non linéaires ?

DR CRAD RÉPOND: Cette question est importante et souvent elle n'est pas bien traitée, bien que la réponse soit assez facile. Tout d'abord, nous devons généralement supposer que la taille de l'échantillon est suffisamment grande afin que des versions asymptotiques de l'erreur quadratique moyenne puissent être dérivées. L'erreur quadratique moyenne (ou, d'une façon équivalente, la variance des estimateurs sans biais) est la mesure la plus utilisée afin d'évaluer la performance d'un estimateur dans la théorie des sondages. C'est certainement une mesure excellente pour des sondages à grandes échelles où le théorème central limite implique que l'estimateur est approximativement distribué selon une loi normale. Nous développons souvent un plan d'échantillonnage en adoptant une stratégie qui minimise l'erreur quadratique moyenne de l'estimateur de la moyenne ou du total de la population pour une variable bien spécifique, ou un ensemble de variables.

Pour le cas plus commun où la statistique non linéaire est une fonction différenciable des estimateurs de moyennes de la population, l'erreur quadratique moyenne peut être dérivée aisément et de façon explicite. On peut montrer qu'elle est équivalente à l'erreur quadratique moyenne de l'estimateur de la moyenne d'une variable linéarisée. Donc, pour répondre à votre question, nous pouvons appliquer toutes les méthodes employées afin d'obtenir un plan approprié pour estimer les moyennes de la population, mais en utilisant plutôt une linéarisation de cette variable.

Je vous présente quelques exemples. Le premier, suppose que nous sommes intéressés à estimer $\ln(\bar{Y})$ où $\bar{Y}$ est la moyenne de la variable Y, et que nous estimons cette moyenne par $\hat{\bar{Y}}$, en utilisant les valeurs de l'échantillon $y_1, y_2, \ldots, y_n$. Il peut être démontré que $\ln\left(\hat{\bar{Y}}\right)$ a une variance asymptotique qui est équivalente à celle à un estimateur de la moyenne se basant sur des valeurs observées $z_1, z_2, \ldots, z_n$, ou $z_i = y_i / \bar{Y}$. Dans ce cas, optimiser sur le y afin d'estimer $\bar{Y}$ est équivalent à optimiser sur les z afin d'estimer $\ln(\bar{Y})$.

Un exemple plus complexe estime une proportion. Supposons que nous désirions estimer une proportion par $\hat{R} = \hat{Y}/\hat{X}$. La variance asymptotique est maintenant équivalente à celle de l'estimateur d'une moyenne qui utilise des valeurs observées $z_1, z_2, \ldots, z_n$, où $z_i = (y_i - Rx_i)/\bar{X}$ et $R = Y/X$. Nous choisirons un plan de sondage qui minimise la variance asymptotique du total basée sur les z_i.

Les méthodes générales pour dériver ces variables linéarisées dépassent les objectifs de ce bulletin. Si vous désirez plus d'information à ce sujet contactez n'importe quel membre du CRAD.

Conseils pour UNIX

L'interface graphique avec UNIX est X et la plupart de ce qui est fait dans UNIX est fait avec cette interface i.e. par les fenêtres et polices (fonts) inclus dans X. L'échelle des fenêtres et des polices sont automatiquement ajustées en tenant compte de la résolution et

de la taille de l'écran (moniteur) dont nous nous servons. Donc, la vision sur l'écran n'est pas nécessairement celle qui est requise ou préférée. Voici une façon de choisir les polices. La disponibilité des polices dépend du serveur X étant utilisé parce que le serveur fournit l'accès à son propre ensemble de polices et ou à un serveur éloigné de polices. Le serveur X dont vous vous servez dépend de votre façon d'accéder aux stations de travail UNIX. Puisque le CRAD standardise l'accès par micro-ordinateur en utilisant "Xoftware for Windows" d'AGE, cette procédure est pour Xoftware mais est très semblable à celle qui serait requise pour tout logiciel similaire.

1. Etablissez la communication avec DARC1 (DARC2 requiert que l'on commaisse le chemin d'accès du SUN pour les commandes X)
2. Entrez la commande: `xfontsel`
En haut et à droite vous verrez le nombre de polices disponibles
3. Cliquez à gauche sur les caractéristique, et maintenir : `fndry`
4. Descendez en bas de la liste pour faire un choix, et relâchez le bouton
5. Répétez pour chacune des caractéristiques

NOTER: Vous verrez l'échantillon changer au fur et à mesure que vous faites vos choix. Faites attention. La plupart des fenêtres sont x-terms et ont besoin de polices espacées uniformément. Si les résultats sont bizarres, vous avez choisi une police proportionnelle.

La caractéristique `spc` indique:
`c` cellule en caractère / espacée uniformément
`m` espacée uniformément
`p` proportionnelle
tous # polices proportionnelles (seulement dans X11 R6)
6. Quand votre choix est fait, cliquez sur "sélect". Maintenant mettez à jour votre fichier`#.xsession`
(`.xsession-DARC1` et `.xsession-DARC2` dans CRAD)
7. Insérez dans la commande xterm (ou autre fenêtre): `- fn - fb`
8. Placez votre curseur après `- fn` et cliquez sur le bouton du milieu afin de coller la police.

Vous devriez employer la même police pour les deux `- fn` et `- fb` sauf pour `- fb`, choisissez "bold" pour : `wght`.

Vous pouvez créer un alias pour chaque police que vous sélectionnez ou désirez utiliser et le placer dans :
`fontpath / fonts.alias.`

Si vous choisissez la même police pour diverses plates-formes, vous obtiendrez des fenêtres consistantes. darc1 et dc2s utilisent un serveur de polices. Afin d'accéder à ces polices, ajoutez à votre chemin d'accès de polices : Pour la séance présente: `xset +fp tcp / DARC1:7000 /`
Définitivement:
1. Assurez-vous qu'aucune autre tâche n'est en cours parce que le redémanage est automatique
2. Cliquez sur l'icône de la séance active pour
obtenir le menu principal
3. Choisissez: `options | fonts`
4. Ajoutez au chemin d'accès de vos polices comme suit:
`,tcp/darc1:7000/`
5. Répétez les étapes précédentes si désiré, pour chaque police du serveur.
Employez un port différent pour chaque tcp - 7001, 7002, etc.
Confirmez l'entrée avec: `xset - q`
Près de la fin des résultats en sortie vous verrez:
Font Path: suivit du chemin d'accès

Si vous avez besoin d'aide pour changer les polices pour SAS, appelez Reid Gregson à 951-1388.

Si vous désirez faire partie de notre liste d'envoi, veuillez nous faire parvenir une demande par l'entremise du courrier électronique ou communiquer avec moi: Georgia Roberts, DMEE (951-1471)

Gestionnaire du CRAD	Mike Hidiroglou, DMEE (951-4767)
Éditeur du Bulletin	Moshe Feder (951-0795)

Pour obtenir de plus amples informations au sujet du CRAD et de ses services, communiquer avec Mike Hidiroglou (951-4767) ou Georgia Roberts (951-1471).

Membres du CRAD

Mike Hidiroglou, DMEE (951-4767)
Georgia Roberts, DMEE (951-1471)
Moshe Feder, DMEM (951-0795)
Milorad Kovačević, DMEM (951-9891)
Harold Mantel, DMEM (951-4150)
Zdenek Patak, DMEE (951-9434)
Boxin Tang, DMEE (951-0498)
Paul Wong, DMEE (951-0049)
Wesley Yung, DMEM (951-4699)
Conseilleur-expert: David Binder, DMEE (951-0980)