Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

ESTIMATING UNDERCOUNTING OF VEHICLE-RELATED INJURY CASES IN NEW ZEALAND: A PROBABILISTIC DATA INTEGRATION AND CAPTURE-RECAPTURE APPROACH

Ricardo Enrico C. Namay II[1]

ABSTRACT

The Injury Statistics Project in Statistics New Zealand links three databases with the aim of producing a comprehensive view of injuries that happen within New Zealand. These databases are supplied by, the Accident Compensation Commission (ACC), the New Zealand Health Information Service (NZHIS) and Land Transport New Zealand (LTNZ).

Using probabilistic data linkage, a more comprehensive database of injuries is obtained. To assess the quality of the links produced, false positive rates and false negative rates are computed. These rates however do not give an indication of whether the databases used for linking have undercounted injuries (bias) nor do they provide error margins for the incidence generated.

Borrowing the idea of capture-recapture, a method for estimating undercounts and error margins for the estimated incidence of vehicular injuries is considered.

KEY WORDS: Capture-recapture; False Positive Rate; False Negative Rate.

1. INTRODUCTION

1.1 Data Integration Methodology

Data integration is a process of linking two or more data sources, two at a time, to identify records which belong to the same entity. Records from one database are compared against records from another in terms of selected comparison variables, called matching variables. One comparison round with respect to the matching variables is called a pass. When corresponding variables are compared and are deemed to agree, an agreement weight is assigned for that variable pair (some positive value). Otherwise, the pair receives a (negative) disagreement weight. The comparison is done on every corresponding pair of the selected matching variables. A composite weight is then calculated for the record pair being compared. This composite weight is merely the sum of all the agreement and/or disagreement weights of the corresponding matching variables of the record pair. Assuming the matching variables are not very highly correlated, a positive composite weight may be thought of to mean that the record pair belongs to the same entity (referred to as a link. When the record pair **truly** belongs to the same entity, we have a match and not a link). The more positive the composite weight is, the more likely the records belong to the same identity. Conversely, the more negative the composite weight is for a record pair, the more likely it is that the records making up the pair belong to different entities (a non-link).

In practice, not every positive composite weight for a record pair is taken to mean a link. A non-negative cut-off score to delineate links from non-links is set. Record pairs whose composite weights fall below this non-negative cut-off score are considered non-links. It is thus possible for record pairs with a positive composite weight to be

[1] Ricardo Enrico C. Namay II, Statistics New Zealand, Statistics House, Wellington, New Zealand, 6004 (ricardo.namay@stats.govt.nz)

regarded as a non-link if its composite weight falls below the cut-off score. In practice, partial agreement between variables (and thus partial agreement weight) is possible, should the analyst decide to allow such.

1.2 Capture-recapture Methodology

Capture-recapture was developed by biologists to estimate animal population in the wild. Among the earliest recorded use of the method is by Petersen in 1896 in estimating fish population, and Lincoln in the 1930s in estimating bird populations. Because of these pioneering works, it is not uncommon to see literature that refer to the population estimator as the Petersen-Lincoln index.

Capture-recapture aims to estimate the population size of a particular animal species in the wild. In this attempt, an initial sample of the species of interest is sampled (capture) and tagged. These tagged animals are then released in the wild. After some time, a second sample of the species is obtained (recapture) and the proportion of tagged vs. the untagged animals is used to come up with a population estimate.

Suppose the number of captured species in the first sample is n_A and the total number of recaptured species is n_B, with n_{AB} of these tagged. The natural estimator, $\hat{N}$, for the unknown population total N can be readily obtained from

$$\hat{N} = \frac{n_A n_B}{n_{AB}} \tag{1}$$

where $\frac{n_{AB}}{n_B}$ is the capture probability. Eberhardt (2003) notes that the relevant distribution is hypergeometric and provides an intuitive explanation to this.

As it is possible however that no tagged animal is in the recaptured sample, the estimator would have a zero denominator, a so-called "infinite bias".

Eberhardt cites Chapman's (1951,1954) proposed adjustment to get around this problem by introducing some correction terms and the resulting Chapman estimator (also called Petersen-Lincoln-Chapman) is

$$\hat{N} = \frac{(n_A + 1)(n_B + 1)}{n_{AB} + 1} - 1 \tag{2}$$

with variance given by

$$Var(\hat{N}) = \frac{(n_A + 1)(n_B + 1)(n_A - n_{AB})(n_B - n_{AB})}{(n_{AB} + 1)^2 (n_{AB} + 2)}. \tag{3}$$

Further, if the unknown population size N is big ($N > 1000$) and the recapture probability $\frac{n_{AB}}{n_B}$ is greater than 0.05, then the normal distribution can be used to approximate the hypergeometric distribution. This result is particularly convenient for constructing a confidence interval around an estimate. Thus for a large N and recapture probability greater than 0.05, the 95% confidence interval for the estimate $\hat{N}$ is given by

$$CI = \hat{N} \pm 1{,}96 * \sqrt{Var(\hat{N})}. \tag{4}$$

2. DATA SOURCES

2.1 The Accident Compensation Corporation (ACC)

The Accident Compensation Corporation (ACC) is a Crown entity that administers New Zealand's accident compensation scheme. This scheme provides accident insurance for New Zealand citizens, residents and temporary visitors to New Zealand.

The reporting unit on the ACC source file is a claim. However, during the course of developing the data integration passes for this project, it became apparent that there were pairs of duplicate claims in the ACC data, that is, two claims lodged for the same injury. The purpose of the integration was not only to count the total number of injuries, but also to attach information from each administrative source to an injury, thus generating a more complete picture. The duplicates affect both aims as the number of injuries would be slightly over counted. Furthermore, information about the same injury could well be attached to two duplicate claims. For example, the costs of an injury would be split over a pair of duplicates. This issue was addressed and a methodology developed to identify and remove any duplicates from the data.

Any injury claims that are declined by ACC are included in the database as injuries, because they meet the definition of injury as outlined in *'Injury Statistics Project Pilot: Definitions of Injury'*.[10]

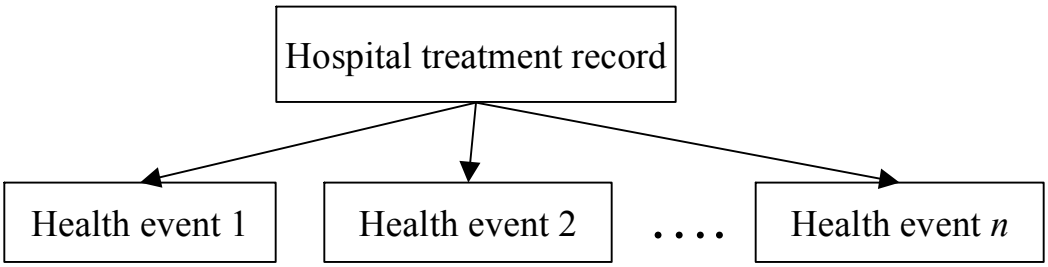
2.2 The New Zealand Health Information Service (NZHIS)

The New Zealand Health Information Service (NZHIS) is a group within the Ministry of Health responsible for the collection and dissemination of health-related data.

Hospitals must supply a record to NZHIS every time a patient who has received publicly-funded healthcare is discharged from hospital. These discharge records are the reporting units on the NZHIS source file and are called 'health events'.

Several health events may be required to treat a particular injury. Prior to carrying out record linkage between NZHIS and ACC data, a process has been developed that identifies health events that belong to the same injury, groups them together and creates a single record for that injury. The term used to refer to a group of health events relating to the same injury is a 'hospital treatment record'.

Figure 1. Relationship of hospital treatment records with health events



Although a great deal of work has been put into developing grouping rules that give as accurate results as possible, the quality of some of the data on the individual health events means that there may be errors in the resulting hospital treatment records. For simplicity, for the purposes of this paper, it is assumed that there are no errors in the groupings of health events.

2.3 Land Transport New Zealand (LTNZ)

Formerly known as the Land Transport Safety Authority (LTSA) at the start of this project, LTNZ is an agency under the Ministry of Transport that collects vehicular accident data and registers these in the crash analysis system (CAS).

At the scene of a vehicular accident, a police officer fills out the crash report form. The names of the injured drivers and/or passengers are indicated in the form. Sometimes, in serious accidents, the ambulance may have left the scene with the people involved in the accident before the police officer arrives or other people involved in the accident may have dispersed so the reports are completed using information from bystanders. In cases where injured people have left the scene before the officer arrives, the officer is meant to follow up on this to gather the additional information needed to complete the report.

LTNZ keeps unique person and crash identifier fields that allow them to update their records when new information regarding the accident becomes available.

3. THE DATA INTEGRATION SEQUENCE AND THE CAPTURE-RECAPTURE METHODOLOGY

ACC and NZHIS data were first linked using various combinations of available common fields. Unique identifiers (e.g. claim numbers, claim form numbers) and personal identifiers (e.g. name, age) have been used to block and link the records from these databases. After the ACC-NZHIS linkage process, false positive and false negative rates were estimated based on business rules formulated to define false positives and false negatives.

The set of captured entities consists of vehicle-related ACC and/or NZHIS injury records. This means that the records of interest for the first sample (capture) could be either a linked ACC-NZHIS record, an unlinked ACC record or an unlinked NZHIS record as long as such records pertain to a vehicular injury. The ACC and NZHIS records have a vehicle injury indicator field and this readily allows one to filter the records pertinent to the study.

Since the false positive and false negative rates for the linked ACC and NZHIS records have been estimated regardless of the vehicle injury indicator value, that is whether or not the accident is vehicle-related, the estimation of undercounts in the LTNZ database in this paper assumes homogeneity of these rates between vehicle and non-vehicle injury sub-samples. The count, *M*, of vehicle-related injuries in the first sample (capture) will thus be the sum:

$$n_A = ACC_{uv} + (1 - fn_{ACC-NZHIS}) * NZHIS_{uv} + (1 + fp_{ACC-NZHIS}) * (ACC \leftrightarrow NZHIS)_{lv} \quad (5)$$

where

ACC_{uv} = the number of unlinked ACC vehicle-related injury records

$NZHIS_{uv}$ = the number of unlinked NZHIS vehicle-related injury records

$(ACC \leftrightarrow NZHIS)_{lv}$ = the number of linked ACC and NZHIS vehicle-related injury records

$fp_{ACC-NZHIS}$ = false positive rate for ACC-NZHIS linking
= percentage of *linked* NZHIS health treatment records incorrectly linked to an ACC record

$fn_{ACC-NZHIS}$ = false negative rate from ACC-NZHIS linking
= percentage of *unlinked* NZHIS health treatment records which should have linked to an ACC record

After the ACC and NZHIS linking, the LTNZ records were linked to the integrated ACC and NZHIS records. As before, false negatives and false positives rates were estimated. The size of LTNZ records therefore that link with vehicle-related ACC and/or NZHIS records would comprise the set of tagged recaptured species of size n_{AB}, adjusted to reflect the false positive and false negative rates for linking in the set LTNZ with the set ACC/NZHIS. If for simplicity, a constant false positive rate for the different linked LTNZ record types is assumed, then,

$$n_{AB} = [1 + fp_{ACC/NZHIS-LTNZ}] * [(ACC \leftrightarrow LTNZ) + (NZHIS \leftrightarrow LTNZ) + (ACC \leftrightarrow NZHIS \leftrightarrow LTNZ)] \quad (6)$$

where

$(ACC \leftrightarrow LTNZ)$ = the number of linked ACC and LTNZ records

$(NZHIS \leftrightarrow LTNZ)$ = the number of linked NZHIS and LTNZ records

$(ACC \leftrightarrow NZHIS \leftrightarrow LTNZ)$ = the number of linked ACC, NZHIS and LTNZ records

$fp_{ACC/NZHIS-LTNZ}$ = false positive rate from linking ACC/NZHIS with LTNZ
= percentage of *linked* LTNZ records incorrectly linked to a record in the ACC/NZHIS file

The count of vehicle-related injuries in the second sample (recapture), will thus be number of elements in the LTNZ file and is given by

$$n_B = n_{AB} + [1 - fn_{ACC/NZHIS-LTNZ}] * LTNZ_u \, . \quad (7)$$

If the false positive and/or false negative rates are suspected to be non-homogeneous, stratification with respect to the various record types may improve the estimates.

4. CONCLUSION

What is shown herein is a simple framework for improving estimates of populations of interest using the idea of capture-recapture and probabilistic record linkage. Moreover, through capture-recapture, a simple way of constructing confidence intervals around estimates coming from probabilistically-linked data has been presented.

REFERENCES

Chapman, D.G. (1951), "Some properties of the hypergeometric distribution with application to zoological censuses", *Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp. 131-160.

Chapman, D.G. (1954), "The estimation of biological populations", *Annals of Mathematical Statistics*; 25, pp. 1–15.

Eberhardt, L.A. (2003), *Course in Quantitative Ecology*, National Marine Mammal Laboratory, Alaska Fisheries Science Center.

Gill, G.V., Ismail A.A. and Beeching, N.J. (2001), "The Use of Capture-recapture Techniques in Determining the Prevalence of Type 2 Diabetes", *Q J Med*; 94, pp. 341-346.

Gill, L. (2001), *Methods for Automatic Record Matching and Linking and their Use in National Statistics*, London: Office for National Statistics.

Nanan, D.J. and White, F. (1997). "Capture-recapture: Reconnaissance of a Demographic Technique in Epidemiology", *Chronic Diseases in Canada*, 18, pp. 144-148.

Rohatgi, V.K. (1976) *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.