Paper ST04

Research Across Multiple Systems: Probabilistic Population Estimation (PPE)

Diane Haynes, Rebecca Larsen, Shabnam Mehra, University of South Florida, Tampa, FL

Abstract

Today, social service administrators are examining client service utilization using cross system analysis, because often a client's needs require accessing government-funded services from multiple organizations. One technical problem that arises is that organizations do not share common unique identifiers from which to link one individual's information together (i.e., system #1 uses Social Security Number (SSN) and system #2 uses Personal Identification Number (PIN)). Different methods have been employed to deal with the issue of working with information across data sets when there is no common unique identifier. Probabilistic Population Estimation (PPE), Caseload Segregation/Integration Ratio (C/SIR), and Probabilistic Population Matching (PPM) are methods used in our shop. This paper discusses the use of SAS® to perform the PPE & C/SIR methods of cross system analysis. These methods accurately identify the number of individuals who cross multiple systems without using a unique ID, while keeping the identity of an individual confidential. PPE is a statistical procedure for deriving unduplicated counts of the number of people represented in data sets that do not include unique person identifiers and the number of people shared by data sets that do not share common personal identifiers (Banks & Pandiani, 2001).

Introduction

Cross system analysis is being used more and more today as local communities find it beneficial to understand the complete picture of how services, that are funded by local, state, and federal dollars, are being accessed and by whom. Individuals interact with multiple agencies in order to have various needs met. Understanding a more complete picture of service utilization requires information from multiple agencies, or systems, to be accessed and combined when conducting analyses. Thus far, each agency has developed its information system in isolation from other agency's systems. One problem that often appears when attempting to share and integrate information from multiple systems is that the unique identifier (ID) that distinguishes an individual is not common across all systems. For example, in one system the unique ID maybe the Social Security Number (SSN) and in the other system the unique ID maybe a Personal Identification Number (PIN). It is not possible to link one individual's information from one system to the other using the unique ID. Methods for integrating information across systems when the unique ID is not shared between agencies include Probabilistic Population Estimation (PPE), Caseload Segregation/Integration Ratio (C/SIR), and Probabilistic Population Matching (PPM).

This paper discusses the PPE and C/SIR methods, which has been coded using SAS® and used to conduct analyses across systems. PPE is a statistical procedure for deriving unduplicated counts of the number of people represented in data sets that do not include unique person identifiers and the number of people shared by data sets that do not share common personal identifiers (Banks & Pandiani, 2001). C/SIR is a ratio rating of 0 to 100 of the amount of overlap

of individuals between multiple files. The formula is as follows:

C/SIR = [(Duplicated Count/Unduplicated Count) – 1]
/ [(Duplicated Count/Largest Unduplicated Count) – 1]

This methodology provides valid and reliable research, while it also protects the personal privacy of individuals (Pandiani et al., 1998).

Methods

The SAS® code (attachment A) accomplishes the following:

- Computes the actual number of individuals in the file (using the unique ID)
- Computes the frequency distribution of the number of DOB and gender combinations in the file
- Computes the expected number of individuals needed to fill the number of DOB and gender combinations found in the file being used and computes the estimated number of individuals in the file
- Computes the lower and upper bounds for the 95% confidence intervals and the zscore difference between the actual and estimated number of individuals
- Repeats the first four steps above for the other file
- Combines both files and repeats the first four steps
- Computes the overlap of individuals between the two files
- Computes the Caseload Segregated/Integrated Ratio (C/SIR)
- Creates a report

The two examples below will examine the overlap and/or impact of dealing with mental health and substance abuse within the local criminal justice and the EMS systems

Example 1

For the purposes of simplicity, the data from only two agencies or systems will be used at one time. The first system contains service data of individuals receiving mental health services (MH/SA). The second system contains arrest information from a county criminal justice system (CJIS). The goal of the analysis, in this example, is to understand the impact of mental health and substance abuse illness by looking at the amount of overlap of persons with a mental health and/or substance abuse and the arrests in the CJIS system.

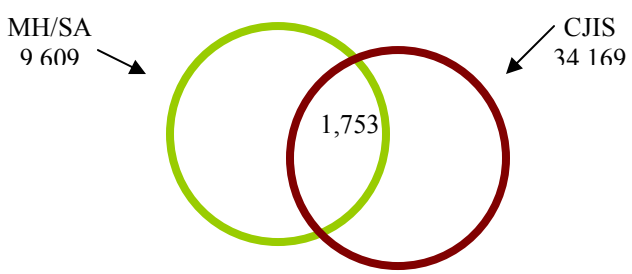
MH/SA uses SSN and CJIS uses a unique person number (UPN) that they created. Therefore, we are not able to link the data from each system for an individual directly using the unique ID. So, we turn to the PPE process, which requires only the date of birth (DOB) and gender for each person in each of the systems. There are 9,609 individuals identified in the MH/SA system during a 12-month period. Their SSN, DOB and gender were preformatted and written to a file. There were 34,169 individuals who had been arrested during the same period of time, and their UPN, DOB, and gender were preformatted and written to a file.

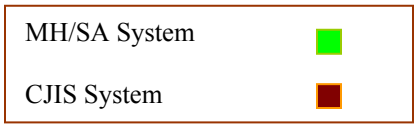
Results of Example 1

The PPE process is executed on each system's data to obtain the PPE for that system. The estimate is compared to the actual count of unique individuals in that system to verify that the PPE is within a 95% confidence interval of the actual count. The reasoning of this is discussed in more detail further on in the paper.

Next the PPE process is executed using a file concatenating both systems data. This gives the estimate of the number of unique individuals in the combined file to be 42,025. That means there are an estimated 1,753 individuals (18%) served by a publicly funded mental health and/or substance abuse agency that are also arrested by local law enforcement during the same 12-month period.

Overlap of populations between MH/SA & CJIS - C/SIR rating of 13.9





The average cost to the county when an individual is arrested is $ 714.00 ($238.00 per day and the average length of stay is 3 days). The minimum estimate of the cost to the county last year would be $1,251,642.

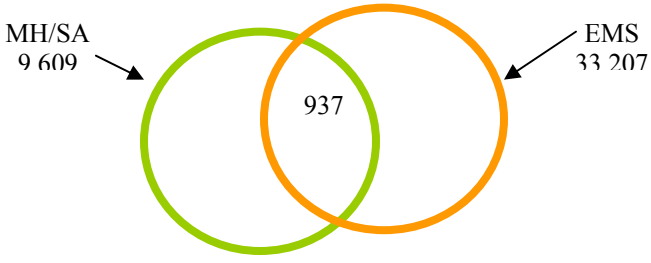
Example 2

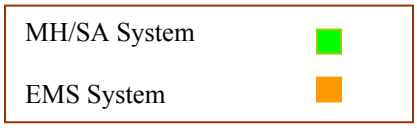
This example examines the overlap in another system, EMS, which is also impacted by mental health and substance abuse. The same MH/SA system is used, where there are 9,609 individuals identified. In the EMS system there are 33,474 individuals where EMS went out and rendered aid and actually transported.

Results of Example 2

The same process is run as in example 1. PPE estimated the number of individuals in the combined systems to be 41,879 individuals. That means there are 937 individuals served by a publicly funded mental health and/or substance abuse agency that are also interacting using EMS services, which is approximately 10%.

Overlap of populations between MH/SA & EMS - C/SIR rating of 6.7





The average cost to the county when an individual is served by EMS varies dependent upon the type of medical needs, the range is from 440.90 to 733.90. This does not include the additional mileage rate. The estimated of the cost to the county last year would be from $413,123 to $687,664.

Issues of PPE and C/SIR

There are three issues when using PPE and C/SIR that need to be considered. The first is the concern of data validity. This is where the 95% confidence interval test is used. The second issue is the difference in the sizes of the files being used, which is dealt with a 1:20 ratio rule. Finally, the third issue, that you can fill up all the days of birth, which could require a modification to the PPE process.

95% Confidence Interval

The first issue of data validity can be checked. It is a check to verify whether the systems unique ID count falls within a 95% confidence level of the PPE estimates. If it does not it could alert the potential of a data validity problem. If the confidence level is lower than 95%, then further analysis needs to be done to verify if the lack of confidence is due to the poor data quality of DOB & gender, or the inability to identify one individual in the system using the system's unique ID.

If the concern is the latter then the PPE may actually be a more accurate count of the number of individuals in the system. This could happen if the unique ID is used to identify multiple individuals or if one individual may be assigned multiple unique system IDs. Examples of these cases would be 1) if the mother's SSN is also used for the child's SSN because the application for child's SSN has not been filed yet; or 2) if the SSN is not unknown and the administrative policy is to use a psuedo SSN, but the individual was already in the system using the actual SSN. Thus one individual is in the system twice, under two different SSNs. In both examples, the SSN

would actually undercount or over-count the number of individuals in the system. If this is the case then PPE could still be used with this file.

If the concern is the quality of the DOB, gender, or both variables, then the PPE process should not be used on the file, until the data is cleaned up. This may happen if the DOB is unknown and an administrative policy is in place to use a default administrative date of '01/01/1999'. In this case the PPE process would undercount the number of individuals in a file because multiple individuals would be identified as one individual. PPE calculates how often it would be expected that two individuals with the same gender would share the same birthday in the general population.

1:20 Ratio

The proportion difference between the files cannot be larger than 1:20. Meaning if the smallest file had 6,417 individuals in it, then the other file being used with it in the PPE process must not have over 128,340 individuals. If the proportion does not meet this requirement then PPE cannot be used on those two files.

Fill up all the dates of birth/gender cells

If you deal with huge data files there could be the potential of the number of individuals being large enough to have at least one individual (Gender & DOB) fill every DOB Year possibility or cell. PPE cannot be used in this case, unless you make changes to either the code or the files. The file(s) could be made smaller by only sub-selecting only individuals you are interested in by some characteristic (i.e. age) or another unique and static characteristic could be added to the gender and DOB in the PPE code to create a larger number of cells for the larger file to fill (i.e. race or mother's maiden name).

Conclusion

PPE and C/SIR are two useful tools with which to conduct cross system analysis, especially during a time when pressures from government and other funding sources are increasing their demand for accountability across multiple systems. These two statistical methods were created by Steve Banks & John Pandiani and for more information about these methods and how they are being used, check the following web site: www.thebristolobservatory.com/. The statistical methods are independent non-identically distributed geometric distributions and are based on two assumptions: 1) That DOBs are uniformly distributed, meaning it is just as likely to be born on one day as any other day of the year. 2) DOBs are independent of one another. The formula that estimates the expected number of individuals is determined by (Banks, 2001).

$$P_j(d) = \sum_{I=1}^{d} \frac{365}{365 - I}$$

For more information about our shop, PSRDC, check the following web site: psrdc.fmhi.usf.edu.

References

Banks, Steven M. & Pandiani, J (2001). Probabilistic population estimation of the size and overlap of data sets based on the date of birth. Statistics In Medicine 20, pp. 1421-1421.

Pandiani, J., Banks, S., & Schacht, L. (1998) Personal privacy versus public accountability: A technological solutions to an ethical dilemma. The Journal of Behavioral Health Services & Research, 25, pp 456-463

Trademark Information

Contact Information

Diane Haynes
Phone: 813-974-9244
e-mail: Haynes@fmhi.usf.edu
Becky Larsen
Phone: 813-974-7206
e-mail: rlarsen@fmhi.usf.edu
Shabnam Mehra
Phone: 813-974-9315
e-mail: smehra@fmhi..usf.edu

Acknowledgements

Steve M. Banks
John A. Pandiani

I would like to express special thanks to Paul Stiles, who guided me toward PPE, & Martha Lenderman for her encouragement and feedback. And all the authors express their appreciation to a unique community effort The Pinellas County MH/SA Data collaborative for making data available to this project.

Produced by the Policy and Services Research Data Center, Louis de la Parte Florida Mental Health Institute, University of South Florida -- Contact: Diane Haynes - 813-974-9244 Created for SESUG 2002 conference

```
/**********ATTACHMENT A
******************
PROGRAM NAME: PPE.SAS
AUTHOR:      DIANE HAYNES
DATE CREATED: 6/01/00
PROJECT NAME: PINELLAS DATA
COLLABORATIVE
PROJECT DESC: THE PINELLAS DATA
COLLABORATIVE PROJECT IS A COUNTY-
LEVEL EFFORT TO SHARE DATA ACROSS
MULTIPLE SYSTEMS FOR THE PURPOSE OF
ANALYZING THE COMBINED DATA. THE
INTENT OF THE COLLABORATIVE IS TO
COORDINATE THE DELIVERY OF SERVICES
IN A MORE SYSTEMATIC,
EFFICIENT AND EFFECTIVE MANNER AND TO
ASSIST IN HEALTH POLICY DECISION-
MAKING.

PROGRAM DESC: THIS PROGRAM TAKES
AGENCY FILES THAT CONTAIN DISTINCT
DATE OF BIRTH AND GENDER FOR EACH
INDIVIDUAL WHO HAS RECEIVED SERVICES
AND CALCULATES THE PROBABILISTIC
POPULATION ESTIMATION (PPE) FOR EACH
AGENCY. (NOTE THAT THE PPE FOR EACH
AGENCY FILE NEEDS TO HAVE MET
A 95% CONFIDENCE LEVEL TO CALCULATE
THE CASELOAD OVERLAP (CSIR) BETWEEN
THE SYSTEMS ACCURATELY.)

THEN THE FILES WILL BE CONCATENATED
AND THE PPE WILL BE RUN ON THE
COMBINED FILES. THESE PPEs WILL BE USED
TO CALCULATE THE NUMBER OF
OVERLAPPING INDIVIDUALS BETWEEN
FILES AND THE C/SIR.

FINALLY A REPORT WILL BE PRINTED WITH
THE CASELOAD CROSSOVER BETWEEN
EACH FILE.

INTRUCTIONS:Prep work: each of the files should
be in the following format:
ID          $9.     System / SSN / Rec. nbr
DOB          8.     mmddyyyy, Date of Birth
FILE        $5.     File ID
SEX         $1.     Gender (1-male, 2-female)
YMDSEX      $9.     Concatenation of
                    DOB Year,month,day,&
                    gender
************************************/
%macro ppe(file);

Proc sql;    /* creates a record by year */
  create table yrsex as   /* and gender */
  select yrsex,
    sum(head_ct) as no_ind format = 5.0,
    freq(yrsex) as unq_dob format =5.0
    from (select ymdsex, freq(id) as
    head_ct format = 5.0,
    substr(ymdsex,1,4) || substr(ymdsex,9,1)
    as yrsex format = $5.
      from &file. group by ymdsex)
    group by yrsex;
quit;

data yrsex
  (drop = leap rleap leapyear year I);
  set yrsex;
  year = substr(yrsex,1,4); /* Test for*/
  leap = year / 4;          /* leap year*/
  rleap = int(year / 4);
  leapyear = leap - rleap;
   if leapyear > 0  then leapyr = "N";
   else leapyr = "Y";

  if leapyr = "N" and unq_dob >= 366
       then  put "ERROR - FILLED EVERY DOB
                  CELL, yrsex = "  yrsex;
   else  if leapyr = "Y" and unq_dob >= 367  then
       put "ERROR - FILLED EVERY DOB CELL,
                yrsex = " yrsex;
   else  do;
    estp = 0;
    varp = 0;
   do I = 1  to unq_dob;
    if leapyr = "N"  then  do;
     estp = estp + (365 / (365 - I));
     varp = varp + ((365 * I ) /  ((365 - I)**2));
     end;
   else
     if leapyr = "Y"  then  do;
      estp = estp + (366 / (366 - I));
      varp = varp + ((366 * I) /  ((366 - I)**2));
           end;
   end;
  ci95 = (varp**.5)*1.96;
  l_ci95 = estp - ci95;
  u_ci95 = estp + ci95;
  zdif = ((no_ind - estp) / varp**.5);
  format estp 8.2 varp 8.6 ci95 6.2 l_ci95
       6.2 u_ci95 6.2 zdif 5.2;
  end;
run;

proc sql;
  create table tot as
  select sum(no_ind) as tot_ind,
          sum(estp) as tot_ppe,
     sum(varp) as tot_var,
     calculated tot_ppe + (1.93 *
   (calculated tot_var **.5))
    as h_nbr,
     calculated tot_ppe - (1.93 *
              (calculated tot_var **.5))
              as l_nbr,
     case when ((calculated tot_ind >=
              Calculated h_nbr) and
              (Calculated  tot_ind <=
```

```
                Calculated l_nbr)) then "Y"
      else "N" end as ok, "&file" as file
from yrsex;
quit;

data totals;
  set totals tot;
run;
proc sql;  /* verify numbers look correct*/
  select * from totals;
quit;

 %mend ppe;


/***** create table here ****/

options mprint mlogic;

 proc sql;
  create table totals ( type = data,
           file char(15),
                    h_nbr num ,
                    l_nbr num ,
                    ok char(1),
                    tot_ind num ,
                    tot_ppe num ,
                    tot_var num );
quit;



 %ppe(ba.ba_youth);
 %ppe(ems_all);

data ba_ems;
  set ba.ba_youth ems_all;
run;

 %ppe(ba_ems);

/* Done once at the end - calculate the C/SIR Rating
on concatenated file */

data csir ;
  set totals end=eof;
  length dup_cnt largest_undup_cnt
       undup_cnt 8. csir_t $8;
  retain dup_cnt largest_undup_cnt
       undup_cnt 0;
   if eof then do;
           undup_cnt = tot_ppe;
     csir =(((dup_cnt / undup_cnt) - 1 )/
           ((dup_cnt/largest_undup_cnt)-1))*100;
           csir_t = 'csir is ';

    end;
  else do;
    dup_cnt = dup_cnt + tot_ppe;
    if tot_ppe > largest_undup_cnt then
       largest_undup_cnt = tot_ppe;
         end;
run;

proc sql;
  select * from totals;
  select * from csir;
quit;
```