

PROBABILISTIC POPULATION ESTIMATION

Probabilistic Population Estimation is a statistical procedure that determines the number of people (with known confidence intervals) who are represented in data sets that do not contain unique person identifiers. Probabilistic Population Estimation uses information on the distribution of birth dates in a data set to determine the number of people represented in the data set. The number of people necessary to produce the number of birthdays observed in a single birth year cohort, for instance, would be calculated using the following formula:

$$P_j(l_j) = \sum_{i=1}^{l_j} \frac{365}{365-i}$$

where “ P_j ” is the number of people and “ l_j ” is the number of birth dates observed. Similar logic is used to determine the number of people who appear in more than one data set. The table below provides illustrative results of Probabilistic Population Estimation for populations of specified size.

Population Estimates for Specified Numbers of Birth Dates Within a Year

Birth Dates	Number of People	Birth Dates	Number of People
1	1.003 ± .103	180	249 ± 20
10	10.15 ± .776	250	423 ± 38
20	20.6 ± 1.54	300	632 ± 64
50	54. ± 4	330	860 ± 101
100	117. ± 9	360	1603 ± 325

POPULATION OVERLAP

In order to probabilistically determine the number of people shared across data sets that do not include a common person identifier, the sizes of three populations are determined and the results are compared. The number of people in each of the original data sets are the first two populations. The number of people in a data set that is formed by combining the two original data sets is the third data set.

The number of people who are shared by the two data sets is the difference between the sum of the numbers of people represented in the two original data sets and the number of people represented in the combined data set. This occurs because the sum of the number of people represented in the two original data sets includes a double count of every person who is represented in both data sets. The number of people represented in the combined data set does not include this duplication. The difference between these two numbers is the size of the duplication between the two original data sets, the size of the caseload overlap. In terms of mathematical set theory, the intersection of two sets is the difference between the sum of the sizes of the two sets ($A+B$) and the union of the two sets ($A \cup B$):

$$(A \cap B) = (A + B) - (A \cup B).$$