

2 Sources of Demographic Data

1. Populations: Open and Closed

We can think of a population size as a *process*. At any given time t a set of individuals satisfy the membership criterion of the population. In the case of a geographic area, for example, the criterion is “being in the area”. The population can increase via births and in-migration. It can decrease via deaths and out-migration.¹ Thus, births, deaths, and migration form the relevant *vital processes*.

Traditionally, the term *vital event* is used for births, deaths, marriages and divorces but not for migration (cf., Shryock and Siegel 1976, 20). Although this usage has an origin in civil registration, the distinction is not useful in statistical demography and we consider vital processes to include migration. Changes of marital status can be vital processes, if the population of interest has been defined in terms of marital status, but so can be such processes as getting a job or becoming unemployed, if the population is defined in terms of employment status.

In a limiting case we define a population as *closed* if it has no vital processes. A closed population is simply a set of individuals. (In demography it is common to call a population closed even if it experiences births and deaths. We take here a broader view.) In most demographic applications a population is open in some respects. For example, in a follow-up study of a fixed set of individuals, the population is closed with respect to births and in-migration, but it is open with respect to deaths. Annoyingly from the researcher’s point of view, such a population may, in practice, be open to out-migration and other forms of attrition or loss from follow-up, as well.

As discussed below, the distinction between closed and open populations is important in the design of the data collection for demographic studies. However, in most parts of this book we have the prototype of national population in mind. National populations are open to births, deaths, migration etc.

¹ A population can also change when its definition changes, e.g., when a country, state, or city annexes or de-annexes an area. Such changes do not involve vital processes, and analysis of past data on population change should make allowance for any significant boundary changes that occurred.

10 2. Sources of Demographic Data

At first thought nothing seems simpler than to define a population. National identity is so ingrained that a special effort is required to appreciate the conventional aspects of the membership criterion. Therefore, consider the following two examples.

Example 1.1. Who Counts in the U.S. Census? The United States Constitution (Article I, sec. 2) stipulates that “Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons.” Since nontaxed Indians were not included in these numbers, their coverage in historical censuses (that started in 1790) is dubious. Slaves were to be counted in a separate category in censuses prior to 1870. It seems that slaves were to be counted in full in the census and then their numbers reduced by two fifths for Federal apportionment – slaves did not figure into population counts for apportionment of state legislatures by southern states (cf., Shryock and Siegel 1976, 14–16; Savage 1982; Anderson and Fienberg 1999, 13). ◇

Example 1.2. Who Belongs to the Sami Population? In the mid-1990’s considerable controversy was caused in Northern Finland by the question of who belongs to the *Sami (Lapp)* population of Lapland. Some advocated a definition emphasizing the role of Sami language, others the length of family history in the area. Different cultures had mixed in Lapland over the centuries, so no clear-cut distinction between the families could be given. Fueling the controversy was the thought that the original people of the area may be treated preferentially in future legislation. In the Law on the Sami Cultural Self-Government from 1995 the following (freely translated) definition was given:

A person belongs to the Sami population, if he considers himself to be Lapp, provided that (1) he himself or at least one of his parents or grandparents has spoken Sami as his mother tongue; or (2) he is a descendant of a person who has been marked as mountain, forest, or fisher Lapp in the books of land or taxation; or (3) at least one of his parents has been marked or could have been marked as having the right to vote in the election of Sami representatives.

In addition, a map of the area within which this definition was to be applied, was published. ◇

These examples display many of the problems that one encounters in trying to define a membership criterion for a human population. Economic, cultural, and administrative considerations are typically involved. Even subjective factors (“... if he considers himself to be Lapp ...”) were involved in the very definition of the Sami population. How can or ought one define the “true size” of the Sami population at a given point in time? Not only is the definition subjective, but so is its measurement: a person’s self-identification may vary over time as well as how the question asking for self-identification is presented.

A similar issue arises forcefully in the definition and assignment of racial classifications. The American Anthropological Association concluded that “The concept of race is a social and cultural construction, with no basis in human biology – race can simply not be tested or proven scientifically.”² In the U.S., ever since the 1970 census a person’s race is based on self-identification. Since some people identify with more than one group, the United States began in the 2000 Census to allow for “multi-race” categories: 63 racial classifications with 6 categories³ for single-race only and 57 for combinations of races (U.S. Census Bureau 2000). Analysis of time series statistics for racial groups in the U.S. requires care for allowing for definition changes pre- and post-2000.

Below, we briefly discuss some aspects of the operational definition of national and sub-national populations and relate these to the coverage and classification errors that frequently occur. We next discuss censuses and population registers as sources of population data. We pay attention to historical aspects of the registration of the vital events, because analysis of past time series of statistics on vital events will help us understand the accuracy of forecasts. Similarly we introduce the concept of the Lexis diagram for insight into the complexities of using grouped data to estimate vital rates in open populations. After that we consider registers and cohort and case-control study designs as prototypes of data collection for specific demographic (or epidemiological) problems. We conclude the chapter by discussing the role of statistical sampling in population estimation. Sampling more generally will be discussed in Chapter 3.

2. *De Facto* and *De Jure* Populations

At any moment in time any specific geographic area has a *de facto* population, which consists of all individuals who are present in the area. This concept is unequivocal but may not always be highly relevant. Consider the following groups mentioned in the “Recommendations for the 1990 Censuses of Population and Housing in the ECE Region” (United Nations 1987, 9–10):

- (1) persons usually resident and present;
- (2) persons usually resident but absent;
- (3) persons temporarily present but usually resident elsewhere.

The *de facto* population comprises (1) and (3), but excludes (2). Often one is interested in the usually resident, or *de jure*, population consisting of (1) and (2). The distinction may seem simple until one considers the cases frequently encountered in practice:

² American Anthropological Association, Press Release/OMB 15, Sept. 8, 1997.

³ American Indian and Alaska Native, Asian, Black or African American, Native Hawaiian and Other Pacific Islander, Some Other Race, White.

12 2. Sources of Demographic Data

- (a) persons maintaining more than one residence;
- (b) students not living with parents;
- (c) persons living away from home during work week;
- (d) persons in military service;
- (e) military personnel who maintain a home elsewhere;
- (f) institutional populations such as hospitals, or prisons;
- (g) persons intending to return to a former home place;
- (h) persons who have arrived a short time ago who consider some other place as their home;
- (i) persons expected to return soon from elsewhere.

Categories (g)–(i) may consist of illegal aliens, nomads, vagrants, military, naval, or diplomatic personnel and their families. They may include merchant seamen, fishermen, transients in ships, trains, cars, or airplanes, refugees etc. For different purposes different choices can reasonably be made concerning which of these groups are included into the population. In many countries and many subnational areas these categories may be small and so their operational definitions may not matter in practice. Sometimes these groups do matter, however.

Example 2.1. Accident Rates in Nordic Countries. A comparison of the rate of traffic accidents in the cities of Gothenburg, Helsinki, Oslo, and Stockholm from 1990–1994 (Nieminen 1996, 22) shows that Helsinki has had a lower rate of accidents involving passengers inside vehicles (about 1 passenger accident per 1,000 inhabitants in a year) than the other cities (1.5–2.5 per 1,000), but a higher rate of accidents involving pedestrians (about 0.5 per 1,000) than the other cities (0.35–0.5 per 1,000). There can be many causes for such differences, including possible variations in the completeness of the registration. However, a map of the locations of the accidents in Helsinki (Nieminen 1996, 13) shows that accidents concentrate near the central railway station, a major gateway for commuters to work. Although we cannot determine whether this explains the differences between the cities, it is clear that while the accidents are tabulated according to the place of occurrence, the denominator population is the *de jure* population. This is a mismatch. A proper denominator for the risk rate would be the *de facto* population because many accidents occur to individuals who commute to work. ◇

In the industrialized countries, the official population figures typically rely on some form of *de jure* definition (Shryock and Siegel 1976, 50). Once the definition of the population is agreed upon, it is important to consider the quality of demographic information. If the analysis of time trends is of interest, have the definitions remained the same over time? If comparisons between different areas are of interest, are the definitions the same in the different countries? Finally, if the definitions are comparable, are the counts and classifications accurate?

Example 2.2. Undercount in U.S. Censuses. Consider the population sizes reported by U.S. censuses of 1940–2000. The “net undercount” – true size minus census count – can be estimated by several methods (cf., Chapter 11). To appreciate the order of magnitude, consider the following estimates of the undercount (in %) by

race based on “demographic analysis” (Robinson et al. 1993, 1065, and Robinson, Adlakha, and West 2002, 26):

	Non-Black		Black	
year	male	female	male	female
1940	5.2	4.9	10.9	6.0
1950	3.8	3.7	9.7	5.4
1960	2.9	2.4	8.8	4.4
1970	2.7	1.7	9.1	4.0
1980	1.5	0.1	7.5	1.7
1990	1.6	0.6	8.1	3.1
2000	0.2	− 0.8	5.1	0.5

We see that Blacks have higher undercount rates than Non-Blacks, and males have higher undercount rates than females. Note that the rates show the *net* effect of both census misses and census duplications or other erroneous enumerations. By and large the net undercount rates declined from 1940 to 1980, and increased in 1990. It is possible that attempts to obtain a complete count may lead to increased erroneous enumerations, and the 2000 census appears to have overcounted non-black females. Demographic analysis also shows that net undercount varies markedly by age. For example, in 1990 Black males in ages 25–60 had the lowest probabilities of being enumerated in the census whereas non-blacks in ages 15–25 may even have been overcounted. Clearly, census numbers suffer from problems of comparability across sex, age, race or ethnic group, and time. ◇

Migration can also lead to surprising conceptual problems. In the case of international geographic migration most countries are unable to keep track of emigration, and many countries have difficulty in keeping track of (especially illegal) immigration. The United States, for example, does not have any statistics concerning emigration, and while it has annual statistics of legal immigration, only indirect estimates (e.g., Muller and Espenshade 1985, Espenshade 1997) are available for the much larger illegal immigration. In Europe, the quality of migration data varies considerably. The Nordic countries with well-functioning population registers have relatively good data on people moving in, because typically many aspects of daily life (health care, child care, opening of bank accounts, access to subsidized public transportation etc.) depend directly or indirectly on their being registered. It is somewhat harder to keep track of people moving out, unless the out-movers go to a country with a good register that agrees to supply information about new migrants received. The European countries that rely on censuses face problems similar to those of the United States. A practical problem in compiling statistics on migration is caused by the fact that the countries do not adhere to the same definition as to who is a (long term) *migrant* (Poulain 1993, 354). The U.N. has recommended that an intention of staying at least a year in a country (after an absence of at least a year) would be required to consider a person a migrant, but this is not followed by most European countries (Poulain 1993, 355; Eurostat 2004, 151–153). The use of different definitions of migrants implies that a person may be counted as belonging to the population of two countries at the same time,

14 2. Sources of Demographic Data

for example. Thus, even if the practices of census taking would agree between two countries, the definition of the population during intercensal years need not be the same across countries.

A further problem in published population statistics arises from possible misclassifications by age, race, marital status, place of residence etc. Although age is nowadays accurately known for inhabitants of most industrialized countries, a self-reported age may be in error. In non-industrialized countries age may have been less important, especially in the past. For example in the population of Philippines in 1960 showed remarkable *digit preference* (or *age heaping*) for multiples of 5 years. For example, the counts in ages 59, 60, and 61 were 72,206; 275,436; and 31,299, respectively (cf., Shryock and Siegel 1976, 116).^{4,5} Where feasible, such reporting problems may be mitigated by recording year and date of birth as well as age (to cross-check).

Although demographic methods typically are applied to human populations, demographic concepts have methodological value more broadly. Some notions that are basic for the study of human populations can be usefully extended to populations consisting of other types of elements. Populations of types of consumer goods (cars, refrigerators, . . .) or species of animals (rabbits, fish, insects, . . .) are obvious examples experiencing births, deaths and migration, and having a changing age structure. In addition, one can also study interesting populations consisting of human aggregates such as households and enterprises. Their definition often has an administrative, *de jure* basis, but for application one is typically interested in the *de facto* numbers.

Example 2.3. What Is a Household? Households can be defined in terms of house-keeping, or one or more persons live in a housing unit and provide themselves with food and possibly other necessities of life (cf., Van Imhoff and Keilman 1991, 10). Housing units often have not only *de jure* residents but *de facto* residents as well. Therefore, the composition of a household may only be revealed by special surveys. Note that no aspect of kinship is usually involved in the definition of a household even though many households are familial units also. In addition to births and deaths, households may also *split*. ◇

Example 2.4. Corporate Demography. In *enterprise or corporate demography* (cf., Ilmakunnas, Laaksonen, and Maliranta 1999; Carroll and Hannan 2000, 51) data often are available for individual *establishments*, such as factories, warehouses, restaurants, or stores. In some cases, data may exist for departments within establishments, such as different production lines in a factory. Enterprises, corporations

⁴ The age heaping was still present to a lesser extent in the 1990 census, where the numbers for the three ages were 275,560; 322,233; and 205,177, respectively (Hobbs 2004, 137).

⁵ Similar phenomena occur in other statistics. For example, Breslow and Day (1987, 163) presented data on smoking from the so-called British Doctors' study (cf., Example 5.1 below). Smoking status was classified into classes 0, -4, 5-9, . . . , 30-34, 35-40 cigarettes/day. An estimate of the average number of cigarettes is also given for each class. The averages are quite close to lower limits of the classes suggesting that the respondents have had a clear digit preference of multiples of five.

and other economic organizations with a legally defined (*de jure*) status may consist of several establishments. Finally, conglomerates consisting of legally separate corporations may form a unit of analysis. Data on enterprises are usually collected for some administrative purpose such as taxation or occupational health. Enterprises with low level of economic activity may be inadequately surveyed or even completely omitted by the legal definitions in use. Therefore, the size of the enterprise population may be underestimated in official statistics at the same time that total employee population statistic is relatively accurate. In addition to births, deaths, and splits, enterprises may also *merge*. ◇

3. Censuses and Population Registers

In statistics it has become customary to contrast censuses and samples. A *census* is a study comprising the whole population of interest, whereas a sample involves only a part. A population census refers more specifically to a complete count of the population of an area at a given time. Censuses may be combined with samples in various ways. Some data (e.g., age) may be collected for 100% of the population and other data (e.g., income) collected from, say, every 100th unit. A census can be *de facto* or *de jure* based and typically collects such basic information as age, sex, and, perhaps on a sample basis, marital status, literacy, educational attainment, occupation, industry, place of usual residence, place of birth (cf., Shryock and Siegel 1976, 32; United Nations 1987, 5–7). Most countries of the world (including the United States, England, France, China, and India) rely on censuses as the basic source of population data. In practice, censuses are carried out via mail questionnaires and door-to-door interviewing. Since population counts are often used to apportion political power, for military conscription, or for taxation, a census may not always be an innocuous operation.

Example 3.1. Nigerian Censuses. Prior to the 1991 census the population of Nigeria was estimated to be 95.7 million in 1985 by the United Nations, 110 million in 1988 by the World Bank, and 112 million in 1987 by the Nigerian government. Estimates for the year 1991 were in the range 112–123 million (*Population Today*, June 1992, No. 6). The history of the Nigerian censuses goes back to the 1860's but apparently the quality of the results, including that of the previous census, in 1973, has been less than satisfactory. Presumably, the ethnic diversity of the country has played a part in this. With this background it was quite a shock that the 1991 census count was 88.5 million, or more than 20% less than the estimates. Evidently, any attempt at a statistical analysis of the population of African countries must somehow account for the uncertainty of the census results. ◇

In countries using censuses a separate system has been in place for the estimation of births, deaths, marriages, migration etc. For example, in the United States death registration became fairly complete in Massachusetts around 1865 (Shryock and Siegel 1976, 21). In the year 1900 a “death registration area” was established comprising the District of Columbia and ten states. A “birth registration area” was

16 2. Sources of Demographic Data

established in 1915 with the same area included. Complete geographic coverage was achieved in 1933 although only 90% registration was required for the admission of a state into the area (Shryock and Siegel 1976, 274). We see that even in the industrialized world one cannot expect long time-series of known statistical quality, on vital events.

In contrast to the statistics usage, in demography censuses typically are contrasted with population registers. Registers provide continuous information about all members of the (typically *de jure*) population. The Nordic countries, Japan, and Russia are examples of countries with population registers. Although nowadays population registers are maintained as computerized databases in many countries, they have a long history. Finland and Sweden have continuous, register based population statistics from the year 1749 onwards. The registers were kept by the church based on an ecclesiastic law of 1686. Each parish would keep track of the vital processes of births, deaths, marriages, and changes of parish. Initially, these registers developed out of books that were maintained since the 1500's for the follow-up of parishioners' progress in the knowledge of reading, writing, and the Bible (Nieminen and Markelin 1974). The establishment of the population statistics around 1750 seems to have occurred in part because estimates compiled by the Royal Academy in Stockholm showed that the true population was only about 2 million instead of the generally believed figure of 3 million (Teräsvirta 1987, 3), a situation not unlike the one that occurred much later in Nigeria!

The reliability of the Finnish vital statistics has been studied using parish level data by Pitkänen (1977), for example. He has shown that many infant deaths were omitted from the registers during the 18th century, because unbaptized children were recorded as stillborn, and baptized infants who died young were deliberately omitted. Pitkänen (1986) also shows that a curious increase in the mortality of the middle-aged and older men during the first decades of the 20th century may have been an artifact caused by migration to the United States. Apparently a fairly large number of deaths that occurred overseas were recorded in the parish registers, even though the persons themselves had been marked as emigrated. The mis-match of the numerator and denominator (as in Example 2.1) could have caused an artificial increase of a few percent in the estimated mortality (Pitkänen and Laakso 1999).

Countries with population registers do conduct censuses every five or ten years to provide occupational and educational details that are not included in the population register itself. The situation varies between countries but for example in Finland this involves the linking of computerized databases rather than door-to-door activities (Harala and Tammilehto-Luode 1999).

4. Lexis Diagram and Classification of Events

A formal aspect of the recording of the vital events is their classification by age and time. Much the same way as with defining populations, initially nothing seems simpler. However, since it is customary to compile statistics on vital events by discrete time, rather annoying complications arise. To appreciate the problem, we

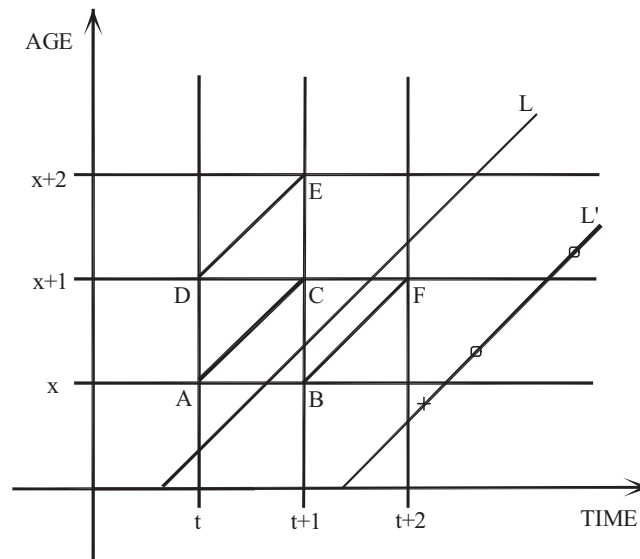


FIGURE 1. Lexis Diagram.

introduce the concept of a Lexis diagram, one of the most useful technical devices of demography.⁶

We let horizontal axis refer to time t and vertical axis to age x in Figure 1. For each person in a well-defined population we may draw a *life line* that starts at a time and age when the person enters the population and ends at the time and age when the person exits the population. Typically the entry would occur at birth and the exit at death, but entries or exits due to other vital processes (e.g., migration) may occur at other ages. The line L of Figure 1 is an example of a life line.

The complications referred to above arise from the following. Suppose we are interested in describing the mortality of the population in age x during year t . We have three options. (1) We may take those who were in age $\in [x, x + 1)$ at exact time t , and observe their mortality experience during year t . The life lines of these individuals touch or cross the line AD and the deaths among them occur in the parallelogram $ACED$. The problem is that these individuals have their $(x + 1)^{\text{st}}$ birthday during the year, so the deaths occur to both x and $x + 1$ year-olds. (2) We may take those whose x^{th} birthday occurs during year t . Their life lines cross the line AB and their deaths occur in the parallelogram $ABFC$. The obvious problem is that the deaths occur in part during year $t + 1$. (3) We may consider those who are present in the population in age x during any part of the year t . Their life lines cross either AB or AD , and the deaths are recorded in the rectangle $ABCD$. One problem

⁶ Wilhelm Lexis (1837–1914) was a German statistician and economist who was among the first users of the diagram in Lexis (1875). Others (e.g., Gustav Zeuner, Karl Becker) had used similar graphics in the 1870's also.

in this approach is that it mixes deaths from two *birth cohorts*: life lines crossing AD belong to those born during calendar year $t - x - 1$; life lines crossing AB belong to those born during year $t - x$. Also, unlike the other two approaches, it is less directly applicable to forecasting because forecasts are typically formulated in terms of cohorts.

Many countries routinely compile their vital statistics based on the rectangles. They give rise to *period measures* (i.e., measures relating to a particular observation period such as a calendar year) of life expectancy, for example. Since such calculations combine data concerning different cohorts (mortality experience of the $x + 1$ years olds is recorded from the rectangle above DC, for example), one often thinks of them as referring to *synthetic cohorts*, whose experience corresponds to those alive during any part of the year t .

A more refined analysis is feasible if continuous-time data are available. Consider the lifeline L' of Figure 1. Suppose it refers to a woman, whose marriage is marked by '+', whose first and second children were born at mark 'o'. The analysis of the "waiting times" between the marks is called *event history analysis*. Statistical techniques for such analyses will be discussed in Chapters 4 and 5.

In general, the follow-up of cohorts requires that events are classified by the year of occurrence, age, and birth year. With the *triple classification of vital events*, the events of interest can be divided into the *triangles* of Figure 1, so any of the above approaches could be implemented. In modern computerized registration systems triple classification poses no particular problems. However, one should note that in all countries of the world demographic statistics have earlier been based on separate tabulations that have been extracted from the primary source materials by hand. In many countries they still are. In non-automated tabulations the requirement of triple classification is an additional burden. Consequently, one cannot expect long time-series based on triple classification in any country in the world.

There is an even more fundamental problem in some demographic and related statistics. Above, we have taken for granted that the events are classified by the *year of occurrence*. However, sometimes events are tabulated by the *year of reporting*. This seemingly illogical practice may sometimes be followed because it is desired to published statistics in a timely fashion. One can argue that if the number of missed reports during year t equals the number of those reports that actually relate to events from earlier years, but come in during t , then no error occurs. This argument is misleading, however, since much of the interest in official statistics is in changes of trends, and the trends will be distorted if tabulations are made by the year of reporting.

The timeliness requirement does produce a problem for all statistics, even those based on the most modern computer systems. For example, it is typical that information about deaths occurring abroad come into the registration system months, or years, after the event. For this reason, statistical agencies establish rules as to how long they wait for reports of the events. Statistics compiled in this manner may sometimes have to be revised, if the missing events are numerically important. The historical Finnish parish registers discussed above are a case in point.

One should also note that there are events of demographic interest for which the time of occurrence is not easily observable. For example, the onset time of many cancers, or that of HIV infection, is not directly observable, and the presence of a disease may only become known when the disease has progressed sufficiently. In other cases, such as noise-induced hearing loss, the impairment may progress gradually, and no clear-cut definition is feasible. In such cases the reporting of the events may depend crucially on the severity of the symptoms and the efficiency of medical screening. In these cases there may not exist any estimates of actual onset times, and tabulation by year of reporting is the only practical possibility. Nevertheless, we caution that the statistics thus obtained may misrepresent actual trends.

5. Register Data and Epidemiologic Studies

5.1. *Event Histories from Registers*

Much of demography deals with data classified by age group, time period etc. With modern computing power, the analysis of data sets consisting of individual level data has become feasible. Computerized population registers contain life histories of all individuals in a population (cf., Harala and Tammilehto-Luode 1999). These have been supplemented by information from other registers, or from censuses, to analyze mortality, for example (Valkonen and Martelin 1999). Census data are entered into databases, and historical parish records have been available in computerized form (e.g., the Umeå Demographic Database at <http://www.ddbumu.se>, or the Scanian Demographic Database at <http://ddss.nu/Ldd/fortext.htm>, both in Sweden). Social security systems or insurance companies often have highly detailed work histories that are continually updated.

In addition to the administrative data sources mentioned above, computerized data bases have been created for specific research tasks. For example, cancer incidence data are available in many countries from specific cancer registries (e.g., Teppo and Hakulinen 1999). Some countries, such as Finland, maintain a large number of other special purpose databases on births, congenital malformations, occupational diseases, causes of death, abortions, sterilizations, implants, visual impairments, intellectual disabilities, diabetes, infectious diseases etc. (Gissler 1999)

The strength of the continuously operating administrative and special purpose registers is their ability, in principle, to provide information on trends. However, their usefulness may be limited by narrow data content and their information may be biased for specific research uses because they cover only certain groups of persons.

5.2. *Cohort and Case-Control Studies*

Complementing census or register based information, we have increasingly available databases from large epidemiological studies and from social surveys. These

databases have the advantage that they have been created with specific research hypotheses in mind, so, in general, they can be expected to provide superior data sources for certain kinds of causal research.

In Section 4, we used “cohort” to refer to those born in a given year. More generally, a *cohort* consists of those individuals that have experienced a given event at the same time. Strictly speaking, one can then think of a cohort as a closed population. In practice, the term is often used in a way that allows for the possibility that a cohort is depleted by deaths. Or, a cohort can be open with respect to deaths.

In addition to birth cohorts, those entering college during a given semester form a cohort, women who have given birth on the same day form a cohort, etc. In response to the increased public interest in effects of environment and individuals’ behavior on health, governments have funded increasingly many follow-up studies to try to unravel the causal chains involved. As a result, there is an increasing number of high quality data sets containing individual-level information on cohorts.

An alternative, *case-control* (or *case-referent*) study design in epidemiology tries to assess relative risk by comparing those who have fallen ill (“cases”) to those who could have fallen ill, but have not (“controls” or “referents”). Case-control data typically are collected from an open population by sampling, so its study design is quite different from that of a cohort study.⁷

Both designs are much used in epidemiology, and they are both well-suited to demographic studies. We briefly introduce their basic logic and point out some possible pitfalls. For a more detailed discussion, Breslow and Day (1980, 1987), Kleinbaum, Kupper and Morgenstern (1982), Woodward (1999) or dos Santos Silva (1999) may be consulted.

5.3. *Advantages and Disadvantages*

A cohort study is based on the idea that one follows a cohort over time, records the exposures or the occurrence of other potential causal agents, and estimates the extent to which the subsequent illnesses among the members of the cohort vary by exposure history. Since specific illnesses typically are rare and may have a long latency time, cohort studies can be both costly and time consuming.

Example 5.1. British Doctors’ Study. In the famous *British Doctors’ Study* (Doll and Peto 1976) the primary objective was to study the lung cancer risk caused by smoking. In October 1951, all men and women in the British Medical Register who were believed to be resident in the U.K. were sent a questionnaire. The first analyses related to the men only. A total of 34,440 men (or 69% of the men alive at the time) gave their name, address, age, and sufficient information about their smoking habits to be included in the study. Follow-up started in November 1, 1951, and

⁷ Increasingly, case-control studies are conducted within cohorts, i.e., both cases and controls are restricted to members of a predefined cohort. The cohort is followed and controls are selected over time as cases appear. These hybrid designs are called *nested case-control*, *case-cohort*, or *case-base designs* (cf., Prentice, Self and Mason 1986; Flanders, Dersimonian and Rhodes 1990).

continued until October 31, 1971. Repeat questionnaires were sent in 1957, 1966, and 1972 to collect current information on smoking. The numbers of respondents (as proportion of those alive in parenthesis) were 31,318 (98.4%), 26,163 (96.4%), and 23,299 (97.9%), respectively. A total of 10,072 deaths were observed during the follow-up, with 441 caused by lung cancer. In addition, much information was obtained concerning other cancers, cardio-vascular diseases and other diseases. Among the results, one may note that the age-standardized death rate (Section 3.3 of Chapter 5) due to lung cancer was 0.1 per 1,000 person years among the non-smokers and 1.4 among the cigarette smokers – the relative risk of the smokers is about 14-fold. Among the latter, the risk increased from 0.78 for those smoking 1–14 cigarettes/day, to 1.27 for those smoking 15–24 cigarettes/day, to 2.51 for those smoking over 25 cigarettes/day. The evidence on increasing dose-response was clear. ♦

A case-control study is based on the idea that if we find a group of people with a specific illness, and select a group of those who could have the illness (i.e., are at risk) but do not have the illness, then any differences in the earlier exposures of the two groups may be causally related to the illness. The difficulty in carrying out the study centers on the investigator's ability to find controls that can be validly compared to the cases (Feinstein 1985). No exact rules are available, but if one can identify the population out of which the cases arose, then a random sample of the same population are eligible for being controls. (For a lively debate on the matter, see the 1985 contributions of O. Miettinen, J. Schlesselman, A. Feinstein and O. Axelsson in *Journal of Chronic Disease* 38, 543–558.)

Example 5.2. Doll and Hill Study. Prior to the British Doctors' Study, Doll and Hill (1950) had used the case-control design to investigate the role of smoking and atmospheric pollution as risk factors for lung cancer. The study was planned in 1947. Twenty London hospitals were asked to notify the investigators of all carcinomas of the lung, stomach, colon, or rectum. The latter three cancers were investigated to provide a possible contrast to lung cancer. Although complete notification was not achieved, the authors believe that omissions could not bias the inquiry by being a select group, since the hospitals did not know the detailed hypotheses being studied. Between April 1948 and October 1949 a total of 2,370 cancers were reported. It had been decided beforehand that patients 75 years of age and older would not be admitted, so 150 cases were excluded from the study. In 80 cases the cancer diagnosis was found to be erroneous, so 2,140 patients were left. Of these, 408 could not be interviewed due to early discharge (189), being too ill (116), death (67), deafness (24), being unable to speak English clearly (11). One case was excluded due to "wholly unreliable" replies. Thus, 1,732 cancer cases remained. Of these, 709 were lung cancer cases. Despite the exclusions, the authors claimed that the cases were "a representative sample of the lung-carcinoma patients attending selected London hospitals". As controls for the lung cancer cases, the investigators chose 709 patients at the same hospitals who had come there for some other illness. For each case, the control had to be of the same sex, within the same 5-year age-group, and have come to the same hospital at about the same time.

22 2. Sources of Demographic Data

In other words, the controls were individually *matched* to the cases. Somewhat more of the cases turned out to live outside London than of the controls, but again the authors believe that this can hardly influence the results. As one indication of the excess risk they mention that the odds of never smoking were 2:647 among the male lung carcinoma patients, whereas the odds were 27:622 among the male controls. Alternatively, one could say that the odds of cancer were 2:27 among the non-smokers and 647:622 among the smokers. (I.e., there were 29 non-smokers in the data set with 2 lung cancers, and 1,269 smokers with 647 lung cancers.) The resulting odds ratio for cancer is $647:622/2:27 = 14$ indicating a similar relative risk as the one later found in the British Doctors' Study. (This analysis does not allow for the matching that was used in the study, however, and the analysis would now be done in a different way, see Example 7.5 of Chapter 5). ◇

Examples 5.1 and 5.2 suggest the following, simplified characterization of the merits of the two approaches. The cohort study is often relatively *slow and costly*, especially if the illness is rare and the latency time of the illness is long, but the *results are more trustworthy*. The case-control study typically is *quicker and less expensive* but it may be *less reliable* if the choice of controls is biased in some way. We will come back to this issue in Section 2.3 of Chapter 5. Moreover, when cohort studies are carried out prospectively, the exposures and illnesses both occur after the study has been initiated.⁸ In contrast, often a case-control study is retrospective, so that information on exposures must be obtained from remaining records, or it must be remembered by the subjects or by other people who have known them.⁹ Therefore, the exposure information is typically weaker, and possibly biased, and imperfect controls may also cause bias.

However, the potential gains in efficiency are often seen to outweigh the risk of bias, and the case-control design has become a standard tool of epidemiologic investigation. With this background it is surprising that in demography, most investigations with causal goals have cohort designs.

A very large number of demographic studies are cross-sectional, so they follow neither paradigm. Since the time element is missing from those designs, they often lack credibility for causal inferences.

5.4. Confounding

A defining feature of experimental research is that the researcher can manipulate and control the causal factors of interest. For example, in a study of drug efficiency, groups with precise dosage are formed and subjects are randomized into them. In many epidemiologic studies, such as those discussed in Examples 5.1 and 5.2, ethical considerations prohibit manipulation of exposures. Similarly, in most demographic studies (e.g., when investigating the determinants of fertility) the

⁸ Logically, a retrospective cohort study is also a possibility. In this case one defines a historical cohort and collects information on it from existing records.

⁹ In nested case-control studies data collection is usually prospective.

researcher has no choice but to observe what happens, and to try to make comparisons in as valid a manner as possible. We call such studies *observational*. The validity of an observational study with causal aims can sometimes be compromised by unobserved interdependencies of the variables being studied.

Two variables are said to be *confounded* in a study if their separate effects cannot be distinguished from each other (Moses 1986, 9–10).¹⁰ If one variable has negligible effects then the possible confounding may not be important (cf., Bailey 1982). There are also a multitude of other ways in which a comparative study may fail. Yet, possible confounding is often a major concern.

Confounding may be present in an observational study when those subjects who receive a treatment differ systematically from those who do not. For example, when the large-scale randomized (and double-blind placebo-controlled) Salk vaccine trials were conducted, an observational study was also done to compare (i) polio incidence rates for second-grade students who were vaccinated and whose parents gave permission for vaccination with (ii) the rates for first-grade and third-grade students in the same schools. Comparison with a randomized controlled experiment showed the risk of contracting polio was confounded with parental permission – higher income children more readily received permission but had lower immunity from the disease.

Confounding may also be present even in a randomized controlled experiment when subjects leave the study or otherwise do not follow protocol for reasons related to the assignment of the treatment. For example, subjects assigned a placebo or a treatment may perceive it as inferior and leave the study to pursue other treatment.

For an illustration, consider the artificial data of Figure 2. The aim of the study is to understand what might explain variations in Y . Two groups are involved: there are 24 individuals marked with a ‘+’ and 36 individuals marked with a ‘o’, and there is one continuous explanatory variable X . Define $G = 1$, for the individuals of type ‘+’, and let $G = 0$ otherwise. The data are well described by the estimated regression equation

$$Y_i = 1.47 + 6.65G_i + 0.915X_i + e_i, \quad (5.1)$$

where the estimated residuals e_i , $i = 1, \dots, 50$, have the variance 2.19². The coefficient of G has a t -value = 10.27 and the coefficient of X has a t -value = 5.90. With P -values < 0.001, both effects appear highly significantly different from 0.

Suppose now that an investigator has no knowledge of the two types of individuals, and fits a simple linear regression with X alone as an explanatory variable. The estimated equation is

$$Y_i = 9.94 + 0.192X_i + e_i, \quad (5.2)$$

¹⁰ This is a rather general characterization. In particular, it does not include specific assumptions concerning the causal roles of the variables. For a review of the many complexities that arise when the concept is operationalized in an epidemiologic context, see Geng, Guo and Fung (2002).

24 2. Sources of Demographic Data

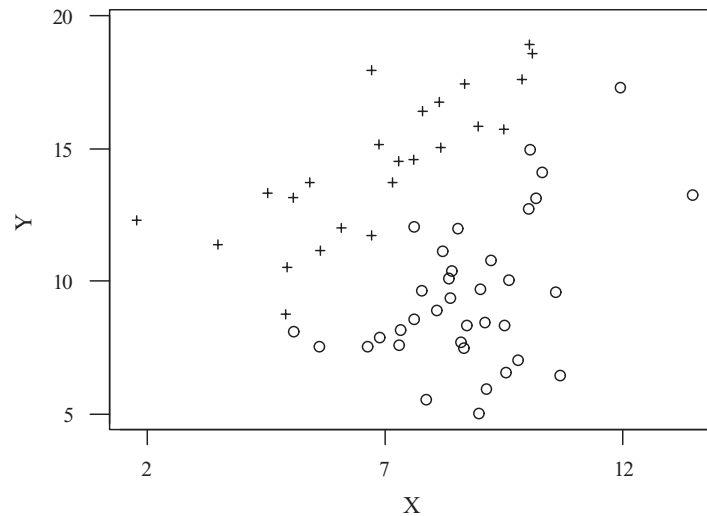


FIGURE 2. Example of Confounding.

where the residual variance is 3.67^2 . The coefficient of X has a t -value = 0.83 and a P -value = 0.41, suggesting that X has no influence on Y . The estimated effect of X is tangled up with the unmeasured group indicator, and the conclusion of the study is incorrect.

Note that had the researcher restricted his or her study to those of type '+' only, and regressed Y on X , the estimated slope would have been 0.83 with a P -value of 0.003, so the correct conclusion would have been reached. The same is true if only those of type 'o' had been studied (resulting in the estimated slope = 0.99, and P -value < 0.001). This suggests that restricting the scope of the study by controlling a variable is one way to avoid confounding.

On the other hand, suppose the investigator was interested in comparing the two groups, and did not measure X . Using a two-sample t -test, he or she would have found that a 95% confidence interval for the mean of those of type '+' minus the mean of those of type 'o' is (3.47, 6.37). The conclusion that those with a '+' have a higher mean would have been correct, but the difference would have been underestimated by approximately a half due to the confounding of G and X .

Both cohort and case-control designs often give rise to contingency tables whose analysis can be invalid, if confounding is present. In complements we indicate some classical procedures for handling suspected confounding via stratified analysis. In Chapter 5 we show how regression techniques can be used to do the same.

6. Sampling in Censuses and Dual System Estimation

If it were not for the need of geographic detail (for municipalities, city neighborhoods or blocks, etc.), sample surveys would probably have replaced censuses a long time ago. Samples would be less expensive to carry out and they reduce

the burden of respondents because only a fraction is included. More extensive information can be collected by well-trained personnel in a sample survey than in a census that has to rely on temporary work force. In addition, being based on deliberate randomization, the precision of statistical sampling can be assessed based on the sample itself (Chapter 3), whereas errors in a census cannot be evaluated based on the census only. These advantages have been used to complement census information in various ways.¹¹

Sampling has been used in the U.S. decennial censuses since 1940 to collect part of the information. The so-called long form requesting detailed data on income and other characteristics is given to approximately 10% of the respondents, the fraction being larger in smaller areas and smaller in larger areas. Major savings in response burden are achieved by this without unduly compromising data quality.

Sampling has also been used in the United States to evaluate the accuracy of the decennial censuses. The “demographic analysis” estimates of Example 2.2 are essentially based on consistency checks between the current census, earlier censuses, and the recorded vital events. A problem in such estimates is that they rely on the assumption that such other pieces of earlier information are trustworthy, an uncertain proposition at best, and they depend on consistency in definitions (e.g., racial classification) among the various data sources.

A direct statistical evaluation of the census can be made by redoing the census on a sample basis in different parts of the country. Suppose the unknown population of an area is N , with n_1 individuals counted in the census. Suppose the second census count is n_2 , and one can verify that m individuals were counted in both censuses. A more refined analysis will be given in Section 5 of Chapter 5, but let us condition here on n_1 and n_2 . Assume that the two counts are independent, and that individuals are equally likely to be counted during either occasion. The probability of counting m individuals in both censuses is equal to the number of ways of choosing m from the n_1 in the first census times the number of ways of choosing $n_2 - m$ from the $N - n_1$ not counted in the first census, divided by the number of ways of choosing n_2 from N . The resulting probability of observing m can be written as $P(m; n_1, N - n_1, n_2)$, when we first define

$$P(x; \alpha, \beta, \gamma) \equiv \binom{\alpha}{x} \binom{\beta}{\gamma - x} / \binom{\alpha + \beta}{\gamma}. \quad (6.1)$$

Here $\max\{0, \gamma - \beta\} \leq x \leq \min\{\alpha, \gamma\}$ and $P(x; \alpha, \beta, \gamma) = 0$ otherwise (Exercise 8). This probability distribution is called the *hypergeometric distribution* (DeGroot 1987, 247–250) and we can use it to calculate the probability of observing m when we know N (and n_1 and n_2). In the census context, we observe values of n_1 , n_2 , and m but we do not know N . One way to formulate a guess

¹¹ The existence of censuses is very important for many sample surveys, because the census can provide a frame or list from which a probability sample can be drawn. The census can also provide information adjusting a sample or calibrating estimates based on the sample to agree with observations on the whole population. We will not pursue these aspects, however.

(or estimate) of N is to choose the value that makes the observed data as likely as possible. We view (6.1) as a function of N (a likelihood function) and choose the value of N that maximizes (6.1) (cf., Feller, 1968, 45–46). The maximizing N is the maximum likelihood estimator. Here, the MLE is essentially $\hat{N} = n_1 n_2 / m$ (Exercises 9, 10).

Example 6.1. Underreporting of Occupational Diseases. The Finnish Register of Occupational Diseases obtains its information from two sources. A suspected case of occupational disease must be reported by the examining physician to authorities (first capture). The case must also be reported to the insurance institution responsible for compensation (second capture). The following data were obtained in 1980: $n_1 = 3,769$, $n_2 = 3,053$, and $m = 1,591$. The total number of cases reported was $M = 3,769 + 3,053 - 1,591 = 5,231$. In this case $\hat{N} = 3,769 \times 3,053 / 1,591 = 7,232$, or the ratio between the estimated cases to the reported cases would appear to be $c \equiv \hat{N} / M = 1.38$. However, it was suspected that the likelihood of reporting would depend on the diagnosis. The main diagnostic groups were (a) noise-induced hearing loss with $M = 1,856$ and $c = 1.20$, (b) diseases caused by repetitive or monotonous work with $M = 1,448$ and $c = 2.47$, (c) skin diseases with $M = 1,171$ and $c = 1.23$, (d) other diseases with $M = 756$ and $c = 1.34$. Adding the disease specific estimates leads to an overall estimate of 8,258 cases in 1980. The fact that diseases in category (b) are poorly reported is understandable, because the connection between working conditions and the disease is particularly hard to establish for them. \diamond

Some populations are especially hard to estimate, because their membership criterion involves illegal activities. Drug use is an example in which users are expected to be reluctant to reveal their user status (cf., Turner, Lessler and Gfroerer 1992). Yet, a drug user may end up being registered in several administrative registers. This provides a basis for population estimation.

Example 6.2. Numbers of Drug Users. In Finland, information about heavy drug use is available through several registers. The most important ones are the Hospital Discharge Register and the Criminal Report Register. In 2001 there were $n_1 = 446$ reports from the former, $n_2 = 825$ reports from the latter, and $m = 53$ reports from both registers, for heavy drug use in the Helsinki Region (Helsinki, Espoo, Vantaa, Kauniainen). This yields the estimate $\hat{N} = 446 \times 825 / 53 = 6,942$. We will come back to this in Example 3.7 of Chapter 5. \diamond

A form of this *capture-recapture* method was used by Sir Francis Bacon in the study of wildlife populations around 1650 (Cormack 1968). Laplace applied it to human populations in the 1780's. The method has been reinvented many times, whence the names “Petersen's method” or “Lincoln index” in ecology. Its modern use in demography is usually accredited to Chandra Sekar and Deming (1949). In demography it is often called *dual systems estimation (DSE)* (Marks, Seltzer, and Krotki 1974).

Simple as $\hat{N} = n_1 n_2 / m$ may seem, in practice the application of dual systems estimation to the study of the census is complicated by several factors. First, the

population may be heterogeneous with respect to the probability of being captured. If the heterogeneity is observable, it can be modeled by stratification (Chandra Sekar and Deming 1949) as we did in Example 6.1 or by logistic regression (Huggins 1989, Alho 1990b). Second, error in n_1 , n_2 , and m may arise from data errors (names, addresses etc.) that should be corrected. Third, actual human populations are typically open, so the *de facto* population of an area may not be the same during the two counts (cf., Alho et al. 1993). Nevertheless, the dual systems approach provides a practical way to analyze the coverage of a census (cf., Mulry and Spencer 1993; Kostanich 2003a,b; U.S. Census Bureau 2004). A more detailed discussion of population heterogeneity will be taken up in Section 5 of Chapter 5, and Chapter 10 presents an overview of the whole problem of census evaluation using dual systems techniques in the U.S. context.

Exercises and Complements (*)

1. Consider (a) your own country, (b) the city you live in. Which is bigger, the *de jure* or the *de facto* population?
2. Digit preference has been quantified in demography using statistics that are based on comparing the size of the enumerated population to the population one would expect to see in the absence of digit preference. Define V_x = enumerated population in age x . *Whipple's index* (for digit preference of ages 25, 30, ..., 60) is defined as,

$$\sum_{y=1}^8 V_{20+5y} / \frac{1}{5} \sum_{x=23}^{62} V_x.$$

This is of the observed/expected form if in reality all V_x 's are equal. Give some more general conditions, under which this index still works. (Hint: Consider 5-year intervals [23, 27], [28, 32], ... and assume that V_x is (a) linear in each interval, (b) an odd function around the center of the interval: $V_{25-x} - V_{25} = -(V_{25+x} - V_{25})$ for $x = 1, 2$, etc.) For more information about quantifying digit preference, see Shryock and Siegel (1976, 116–118).

3. Consider Example 5.2, where an odds ratio for disease (among smokers and non-smokers has been calculated as 647:622/2:27. (a) Show that the odds ratio for smoking (among those diseased and non-diseased) has the same value. Therefore, the value of the odds-ratio does not depend on whether the data come from a case-control, or a cohort study. (b) Given that the data come from a case-control study, can one say that the risk of cancer is 2/29 for the non-smokers and 647/1269 among the smokers?
- *4. Suppose the results of either a cohort or a case-control study are presented as a 2×2 table,

	Ill	Not	Total
Exposed	a	b	n_1
Not	c	d	n_2
Total	m_1	m_2	N

28 2. Sources of Demographic Data

Here $N = n_1 + n_2 = m_1 + m_2$ is the total number of subjects. The odds ratio is estimated as $OR = ad/bc$ under both study designs. Condition on all the margins n_1, n_2, m_1, m_2 . Then, any one element of the matrix defines the others. Denote the upper left hand corner of the matrix by A and its value in a particular experiment by a . Under the null hypothesis that the true odds ratio is $= 1$, the probability of having a exposed who are ill is $P(a; n_1, n_2, m_1)$ as defined in (6.1). Thus, $E[A] = m_1 n_1 / N$ and $\text{Var}(A) = m_1 (n_1 / N) (n_2 / N) (N - m_1) / (N - 1)$ (e.g., DeGroot 1987, 247–250). As discussed by Feller (1968, 194) the variable $X = (A - E[A]) / \text{Var}(A)^{1/2} \sim N(0, 1)$ asymptotically, so $X^2 \sim \chi^2$ distribution with one degree of freedom. Thus, the null hypothesis is rejected at risk level α , if $X^2 \geq k_{1-\alpha}$, where $k_{1-\alpha}$ is the $1 - \alpha$ fractile of the χ^2 distribution. Show that the observed value of the test statistic can be written as

$$X^2 = (N - 1) \frac{(ad - bc)^2}{n_1 n_2 m_1 m_2}.$$

- *5. Continuation. When one wants to control for the values of a potentially confounding third variable with values, say, $k = 1, \dots, K$, then we have K independent strata with

	Ill	Not	Total
Exposed	a_k	b_k	n_{1k}
Not	c_k	d_k	n_{2k}
Total	m_{1k}	m_{2k}	N_k

Denote the true odds ratio in stratum k by θ_k . Consider the situation in which $\theta_k \equiv \theta$ for $k = 1, \dots, K$. Now test $H_0: \theta = 1$ against $H_A: \theta \neq 1$. The famous *Cochran-Mantel-Haenszel statistic* for this hypothesis is

$$X^2 = \left\{ \sum_{k=1}^K (A_k - E[A_k]) \right\}^2 / \sum_{k=1}^K \text{Var}(A_k),$$

where the expectation and variance are calculated as in Complement 4 for each table $k = 1, \dots, K$ (Cochran 1954, Mantel and Haenszel 1959). The remarkable fact is that asymptotically $X^2 \sim \chi^2$ distribution with *one degree of freedom* even if the strata are very small (e.g., $N_k = 2$), as long as K is large. (For large strata the result is obvious.) Show that the observed value of the test statistic can be written as

$$X^2 = \left\{ \sum_{k=1}^K \left(\frac{a_k d_k - b_k c_k}{N_k} \right) \right\}^2 / \sum_{k=1}^K \left(\frac{n_1 n_2 m_1 m_2}{N_k^2 (N_k - 1)} \right).$$

- *6. Continuation. In the setting of Complement 5, the so-called *Mantel-Haenszel estimator of the common odds ratio* is defined (Mantel and Haenszel 1959) as

$$\hat{\theta} = \sum_{k=1}^K \left(\frac{a_k d_k}{N_k} \right) / \sum_{k=1}^K \left(\frac{b_k c_k}{N_k} \right).$$

Show that if $b_k c_k > 0$ for all $k = 1, \dots, K$, then we can write

$$\hat{\theta} = \sum_{k=1}^K w_k \hat{\theta}_k,$$

where $\hat{\theta}_k = a_k d_k / (b_k c_k)$, and $w_k = (b_k c_k / N_k) / \sum_j b_j c_j / N_j$.

*7 Continuation. In a matched case-control study in which one case is matched with one control, each pair forms a stratum $k = 1, \dots, K$ because the matching criteria may correspond to possible confounders. The results of such a study are often represented as 2×2 table as follows:

		Control	
		Exposed	Not
Case	Exposed	a	b
	Not	c	d

This table is a sum of the K stratum specific tables of the type considered in Complement 5. In this case $N_k = 2$ for all $k = 1, \dots, K$ because there is one case and one control in each stratum. There are $N = 2K$ individuals in all. There are four types of tables: (i) a tables with both the case and the control exposed, (ii) b tables with the case exposed but the control is not, (iii) c tables with the case not exposed but the control is, (iv) d tables with neither the case nor the control exposed. In case (i), for example, the table is of the form

	Ill	Not	Total
Exposed	1	1	2
Not	0	0	0
Total	1	1	2

- (a) Verify that in cases (i) and (iv) we have $a_k d_k = b_k c_k = 0$, in case (ii) $a_k d_k = 1$, $b_k c_k = 0$, and in case (iii) $a_k d_k = 0$, $b_k c_k = 1$. (b) Show that the Cochran-Mantel-Haenszel test statistic is then of the form $X^2 = (b - c)^2 / (b + c)$. This is also known as the *McNemar test statistic*. (c) Show that the Mantel-Haenszel estimator of the common odds ratio is $\hat{\theta} = b/c$. Thus, in both statistics only the “discordant pairs” matter.
8. Consider the capture-recapture case in which n_1 is the number of first captures, n_2 recaptures, and m is the number caught both times. (The traditional notation used in capture-recapture literature does not follow the usual conventions of statistics; note that these symbols have here a meaning different from the one in the previous examples!) Show that the labeling of the censuses as first or second in Section 6 does not matter, so that $P(m; n_1, N - n_1, n_2) = P(m; n_2, N - n_2, n_1)$, as defined in (6.1).
9. Show that by equating m to its expected value (that is given in Complement 4) one obtains the classical estimator, $\hat{N} = n_1 n_2 / m$.
10. Show that the MLE based on (6.1) is essentially the same as \hat{N} defined above. (Hint: show that $P(m; n_1, N - n_1, n_2) / P(m; n_1, N - 1 - n_1, n_2) = (N - n_1)(N - n_2) / (N - n_1 - n_2 + m)$, which is increasing when $n_1 n_2 / m > N$, so

30 2. Sources of Demographic Data

that (6.1) is increasing for $N < n_1 n_2 / m$ and decreasing for $N > n_1 n_2 / m$. Conclude that the exact MLE is $= \lfloor n_1 n_2 / m \rfloor$, where $\lfloor x \rfloor$ is the largest integer $\leq x$.)

11. To estimate $\text{Var}(\hat{N})$ under the hypergeometric model in which n_1 and n_2 are fixed, note first that $E[m] = n_1 n_2 / N$ and $\text{Var}(m) = n_2(n_1/N)((N - n_1)/N)(N - n_2)/(N - 1)$. Since \hat{N} is a nonlinear function of m we linearize the statistic at $E[m]$ using a Taylor series, or $\hat{N} \approx n_1 n_2 / E[m] - (n_1 n_2 / E[m]^2)(m - E[m])$. This yields the approximate variance, $\text{Var}(\hat{N}) \approx (n_1 n_2 / E[m]^2)^2 \text{Var}(m)$. Assume that N is large enough so that $N - 1$ can be replaced by N in $\text{Var}(m)$. Show that by plugging in the estimator \hat{N} the approximate variance of the capture-recapture estimator can be estimated by $n_1 n_2 u_1 u_2 / m^3$, where $u_j = n_j - m$, $j = 1, 2$. This is an example of the so-called delta method that will be discussed in more detail in Section 7.2 of Chapter 3.

<http://www.springer.com/978-0-387-23530-1>

Statistical Demography and Forecasting

Alho, J.; Spencer, B.

2005, XXVIII, 412 p. 34 illus., Hardcover

ISBN: 978-0-387-23530-1