In a perfect world, data would always be complete, accurate, current, pertinent and unambiguous. In the real world, data is generally flawed on some or all of these dimensions. Data assessment in practice has tended to focus on completeness and accuracy, and that is the focus of these notes. Currency, pertinence and clarity deserve more attention than they receive, perhaps, but their assessment requires very different methods.

Assessment is sometimes thought of as a preliminary to analysis proper. This is a useful distinction is some circumstances, but in general the assessment of error and the drawing of substantive conclusions are two sides of the same coin. This is suggested by the symbolic equation "Data = Reality + Error", in which "Reality" represents conclusions drawn from the data that are valid despite error and "Error" represents spurious conclusions suggested by the data as a result of error. Since all conclusions fall into one or the other of these two categories, conclusions about error are at the same time conclusions about reality, and conversely.

Data may be defined as systematic information about the members of some statistical aggregate. *Systematic* means that the same information is available for every entity, with exceptions only for missing values and inapplicable cases (e.g., age at first marriage for a never married woman). *Statistical aggregate* refers to a collection of entities (e.g., persons, births, deaths, households) defined by explicitly stated rules for inclusion. Data consist concretely of (1) a collection of *records*, one for each entity in the statistical aggregate, each record containing information about the entity it represents, and (2) one or more texts describing the statistical aggregate and the content of the records. The records and associated documentation are often referred to as a *data set*. *Statistics* are indicators, usually but not necessarily numerical, derived from one or more data sets.

The term *data* in common usage may refer either to data or to statistics. Census data, for example, may refer either to census information on individual persons and households or to the tabular data contained in census publications, such as total population size and the distribution of population by age and sex.

Direct Assessment

There are two general approaches to the assessment of data, direct and indirect. Direct assessment consists of evaluating the coverage and content of a data set. *Coverage* refers to the faithfulness of the correspondence between the records that constitute the data set and the statistical aggregate the data set represents. Data sets may omit records for some entities that should be represented and include records that should not be included. Improper inclusions occur when a data set includes more than one record for the same entity, includes records for entities not in the statistical aggregate, or includes fictitious records. *Content* refers to the completeness and accuracy of the information contained on the records in the data set.

Direct assessment requires a record-matching study, in which two data sets are compared. The records in each data set are divided into two groups: matched records, which represent entities represented by records in the other data set, and unmatched records, the remainder. Numbers of matched and unmatched records provide a basis for assessing coverage. Comparison of corresponding values on matched records provides a basis for assessing content.

The value of information on coverage provided by a record-matching study is limited by *response correlation bias*, which exists whenever the inclusion of an entity in one data set is not independent of its inclusion in the other data set. In the extreme case of perfect correlation, both data sets would consist entirely of records representing the same entities, and matching would provide no information on coverage. Strict independence is unattainable in practice, but a modicum of independence is necessary for a record-matching study to yield useful information on coverage.

In a 1974 publication, Eli S. Marks, William Seltzer, and Karol J. Krotki provided a useful general discussion of record-matching studies. John G. C. Blacker wrote a 1977 article containing a critical assessment of record-matching studies in demography.

Record-matching studies for population censuses require a *post-enumeration survey*, a survey taken after the census for the purpose of evaluating its quality. Matching studies for civil registration data may involve special surveys or draw on other sources of data, such as newspaper reports of births and deaths.

Record-matching studies may be used to assess content error in population surveys. Coverage is less important for survey data than for population census or civil registration data because information is obtained only for a sample. Assessment of coverage for survey data generally focuses on the percentage of households or persons in the sample for whom it was possible to obtain data and on any selection biases that might arise from the exclusions.

Indirect Assessment

Direct assessment of data sets is expensive, both because a second data set is required for comparison and because matching is often a complex and difficult process. The results of direct assessment are, moreover, limited by response correlation bias and by the tendency of data sets collected at the same or nearly the same time to have similar content error. The indirect approach, by which data sets are assessed by analyzing the accuracy of statistics derived from them, is generally far less expensive and will often give results as good as or better than direct assessment.

Assessment of a statistic is concerned with its accuracy, that is, with how close it is to the true value it represents. The principal means for assessing the accuracy of a statistic is comparison with other statistics. Comparison may take many forms. In some cases it may rely on general knowledge rather than on specific comparison statistics. Sex ratios at birth in national populations, for example, tend to be about 105 male births per 100 female births. Should survey data indicate a much higher value, it might be concluded that the

completeness of reporting of female births was deficient. Such conclusions must always take due account of context, however. A sex ratio at birth of 130, for example, might indicate sex-selective abortion rather than defective data.

Direct comparisons with other statistics generally provide the strongest conclusions about data quality. Consider for example Figure 1, which shows retrospective estimates of total fertility rates for Pakistan from four successive data sets. Taken in isolation, each retrospective series of estimates shows a sharp decline in fertility followed by a rise at the end of the series. Comparison of the four series, however, shows that these declines are spurious, for none of the declines indicated by the first three data sets is confirmed by any of the following data sets. The four series taken together suggest not only that was there no fertility decline but that fertility rose slightly between 1960 and the late 1970s.

Figure 1. Total fertility rates for Pakistan, estimated from four data sources

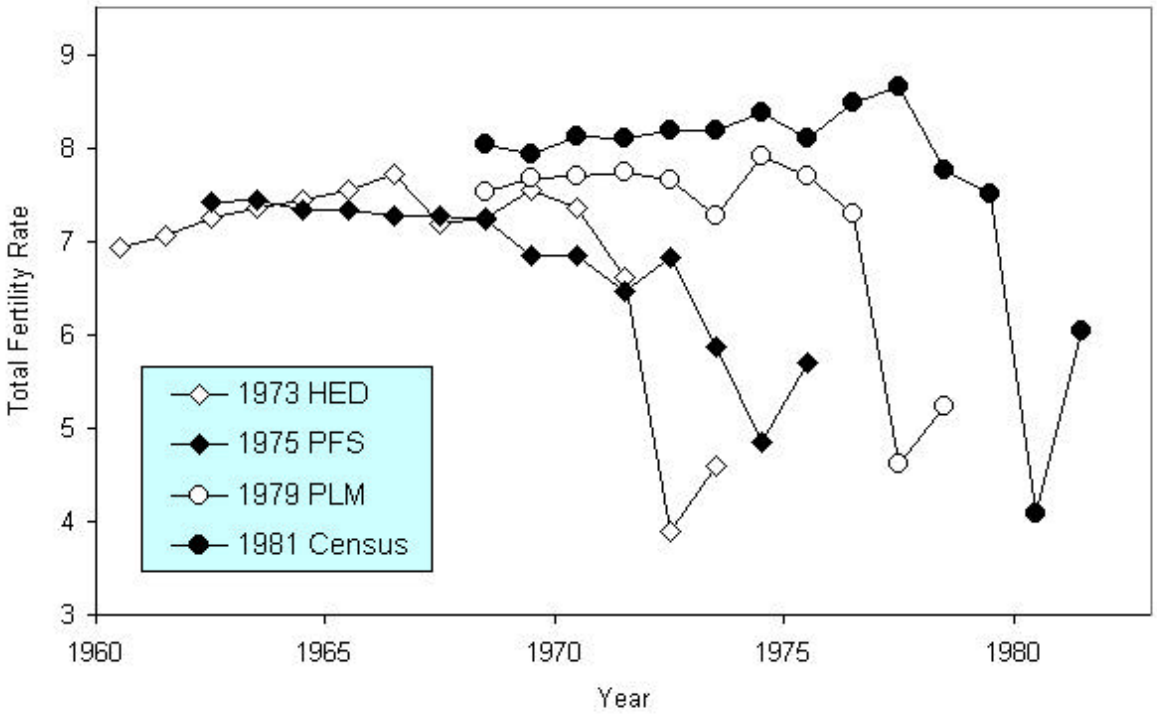

Figure 1

Source note: Redrawn from Retherford, Mirza, Irfan, and Alam 1987. (Estimates for more than five years prior to each data-collection operation are smoothed.)

This example illustrates how error patterns in statistics derived from data sets can be used to draw conclusions about data quality. In each of the four data sets there is a tendency for too few births to be reported during the second and third years prior to the year of data collection. Direct assessment is unlikely to reveal these errors because they tend to affect

all data collection operations. The errors are revealed by the comparison of retrospective series of estimates derived from data sets collected in different years. Because many demographic estimation procedures provide such retrospective series of estimates, such comparisons are often possible.

Statistics derived from higher quality data sets will generally be more accurate than statistics derived from lower quality data sets, but there is no simple, general relation between the quality of a data set and the accuracy of statistics derived from it. A population census with perfect coverage would yield a perfectly accurate total population, for example, but a census that omitted some persons who should have been included and included the same number of persons who should have been omitted would also yield a perfectly accurate total. The latter scenario is unlikely, as omissions nearly always exceed improper inclusions, but the example illustrates the indeterminacy of the relation between data quality and the accuracy of statistics.

Bibliography

Blacker, John G. C. 1977. "Dual Record Demographic Surveys: A Re-assessment." *Population Studies* **31**: 585–597.

Marks, Eli S., William Seltzer, and Karol J. Krotki. 1974. *Population Growth Estimation: A Handbook of Vital Statistics Estimation.* New York: Population Council.

Retherford, Robert D., G. Mujtaba Mirza, Mohammad Irfan, and Iqbal Alam. 1987. "Fertility Trends in Pakistan: The Decline That Wasn't." *Asian and Pacific Population Forum* **1**(2): 1 and 3–10.

Griffith Feeney
Scarsdale, New York
December 26, 2002