The *DEMOGRAPHY-STATISTICS-INFORMATION TECHNOLOGY* Letter
FROM: Griffith Feeney feeney@gfeeney.com
Letter No. 3, 31 October 2013

Statistics are produced from “data”, but what exactly *is* data? Dictionary definitions tend to be too broad to be useful. Specialist definitions tend to be tied to particular software applications or IT concepts.

This letter presents concepts and definitions I find specific enough to be useful, but general enough to be applicable across many computer software applications and IT environments.

Here are my suggestions for what more or less everyone who works with data and statistics should know, to understand clearly what they are doing, and to communicate effectively with others involved in the same work.

Data

Data may be defined as systematically organized information about the entities comprising some statistical aggregate. The entities may be persons, households, dwelling units, births, cities, automobiles, or just about anything else. All that is required is that individual entities be clearly defined and identifiable. Most examples below are for persons, but the concepts and definitions are general.

“Statistical aggregate” means a collection of entities defined in a way that clearly identifies a meaningful aggregate. “Clearly” means simply that there is no ambiguity over which entities are members and which are not. Lacking clear definition, data collection is impossible because we don’t know which entities to collect data for.

“Meaningful” depends on context and resists formal definition. ”The first 300 people listed in the Boston telephone book” is not a meaningful aggregate most contexts. “Persons physically present in the Republic of South Africa at midnight 9/10 October 2011” is a meaningful aggregate in many contexts.

Records, variables, and values

“Systematically organized” might mean all sorts of things, but in practice it tends to mean four fairly specific things.

First, for each entity in the aggregate we have a ***record*** containing information about the entity. The information takes the form of ***values*** of ***variables*** representing characteristics of the entity. For persons, for example, “sex” is a variable whose values are “male” and “female”.

Values, codes, and codebooks

Second, information on records is ***encoded***, “1” signifying “male”, for example, and “2” female. “1” and “2” are ***codes***. “Male” and “Female” are ***values*** represented by these codes. Values are meaningful. Codes are semantically arbitrary.

Codes and values may be identical or nearly identical. If age is recorded to two digits, for example, codes “00” through “98” may refer to ages 0 through 98 years and code “99” to “99 years old and older”.

The correspondence between codes and values for a variable is established in a **codebook**. A number not assigned a value in the codebook for a variable is—if the codebook is accurate—an ***invalid code*** for this variable.

Record layout

A third usual aspect of data structure is that the code representing the value of a particular variable occupies the same position on every record. The code for sex might be in position 28, the code for marital status in position 37. A list showing the position of each variable on a record is the ***record layout***

Data as a matrix

Data organized in this manner may be thought of as a matrix. The cells in each row give (codes representing) values of variables defined for the person represented by the record. The cells in each column give (codes representing) values of the variable represented by the column for each person for whom this variable is defined.

Domain of a variable

The ***domain*** of a variable is the set of persons for whom this variable is defined. The domain of age at first marriage, for example, is ever married persons.

The domain of age at first marriage is a matter of logic, but some domains reflect decisions based on convenience and convention. Literacy, for example, tends to be asked only of children age 5 years old or older—even though some three year old children can read and write.

Undefined values, domain errors and missing values

Consider a cell in a data set matrix giving the value of the variable represented by the column for the person represented by the row.

If the person is *not* in the domain of the variable, we have an ***undefined value***. The cell should be empty. If it contains a (code representing a) value, we have a ***domain error***.

If the person *is* in the domain of this variable, the cell should contain a value. If it does not, we have a ***missing value***.

A cell that is not a missing value cell or an undefined value cell is a ***valid value*** cell. Ideally, every cell in a data set matrix will be either a valid value cell or an empty undefined value cell. There will be no domain errors or missing values.

Filter variables, filtered variables and skip instructions

Population censuses and surveys typically include some questions that are asked of some persons and not of others. Information on place of birth, for example, is typically obtained by an initial question asking whether a person was (a) born in the province in which they are enumerated, (b) another province, or (c) in different country.

Province of birth is then asked of persons born in another province, and country of birth is asked of persons born in another country. For persons born in the province in which they are enumerated, the enumerator is instructed to skip the province and country of birth questions.

More generally, variables whose values for a particular person determine whether some subsequent question should be asked of this person are ***filter variables***. The questions that solicit values of these variables are ***filter questions***.

Subsequent questions that are or are not asked depending the value of these variables are **filtered questions**. The corresponding variables are ***filtered variables***.

Filtered questions are necessarily preceded by ***skip instructions*** that instruct the enumerator either to proceed to the next question or to skip to some subsequent question, depending on the response to the filter question. Skip instructions provide operational definitions of the domains of the various variables whose values are solicited by the questionnaire.

DEMOGRAPHY, STATISTICS, and INFORMATION TECHNOLOGY is an every-other-month email letter from Griffith Feeney, Ph.D., an international consultant based in Scarsdale, New York, USA. For more information see griffithfeeneyconsulting.com.

This letter and past letters are available at demographer.com. To SUBSCRIBE, send email to feeney@gfeeney.com with “subscribe” in the subject field. To UNSUBSCRIBE, send email to the same address with “unsubscribe” in the subject field.