Basic Data Quality Tools

In this chapter, we describe a number of basic data editing techniques. We begin by defining a data element. We then discuss a number of different types of deterministic tests applied to these data elements. We conclude with a brief discussion of both probabilistic tests and exploratory data analysis techniques.

Many data systems employ data editing techniques. For simple data systems, the editing procedures may consist of a few basic checks and tests. As such systems increase in both size and complexity, however, the number of such tests usually increases as well. Large complex databases, containing many different types of data elements, might require several hundred interrelated tests.

When dealing with large databases, the number of tests often grows rapidly. When a large number of tests is involved, they need to be organized into a dictionary. The effort compiling such a dictionary can be minimized if the data editing is organized from the beginning in an orderly fashion. Even in large systems this is often not done. We consider the organizational problem in more detail in Chapter 7, our chapter on editing and imputation. The interested reader could also refer to Chapter 2 of Naus [1975].

5.1. Data Elements

The first step in the construction of an editing/imputation system is the creation of a list of the data elements contained within the data system. A data element or a data field is defined as an aspect of an individual or object that can take on varying values among individuals. Every piece of information in the data system can be viewed as a measurement on a data element.

In most data systems, there are several distinct types of data elements. Some data elements can only assume a discrete set of values (i.e., a countable number of values); other data elements can be viewed as having an underlying continuous set of possible values taken from one or more continuous intervals of values. Sometimes, the values of the data elements are numeric — that is, quantities or magnitudes; other times, these values are qualities. When organizing the edits, it is useful to distinguish between qualitative data elements and quantitative data elements.

The different types of data elements can be illustrated by a simple example. In Table 5.1, we consider a few data elements on a small number of case records in a database of mortgages.
Table 5.1. Extract of database of FNMA/FHLMC\(^1\) conforming mortgages

<table>
<thead>
<tr>
<th>Name of primary borrower</th>
<th>Year of mortgage origination</th>
<th>Month of mortgage origination</th>
<th>Gender of primary borrower</th>
<th>Term of mortgage (in months)</th>
<th>Annual contract interest rate (in %)</th>
<th>Original mortgage amount (in $)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane Smith</td>
<td>1999</td>
<td>08</td>
<td>F</td>
<td>360</td>
<td>6.5</td>
<td>250,000.00</td>
</tr>
<tr>
<td>Robert Brown</td>
<td>2001</td>
<td>06</td>
<td>M</td>
<td>360</td>
<td>6.25</td>
<td>275,000.00</td>
</tr>
<tr>
<td>Harry Turner</td>
<td>2000</td>
<td>03</td>
<td>M</td>
<td>240</td>
<td>5.875</td>
<td>230,000.00</td>
</tr>
<tr>
<td>Mary Sunshine</td>
<td>2003</td>
<td>02</td>
<td>F</td>
<td>300</td>
<td>6.25</td>
<td>300,000.00</td>
</tr>
<tr>
<td>Susan Stone</td>
<td>2002</td>
<td>11</td>
<td>F</td>
<td>360</td>
<td>6</td>
<td>295,000.00</td>
</tr>
</tbody>
</table>

Some data elements can only take on a finite number of (admissible) values or codes. For example, the data element on the “gender of the primary borrower” can only be one of two values – “M” or “F”. Some data elements such as month of mortgage origination have a natural set of values – the month is coded to be a two-digit number between 01 and 12. The quantitative data element “annual contract interest rate” can, in theory, assume positive values in a continuous range. However, in practice, it rarely exceeds five decimal places – for example, a typical value in calendar year 2005 was 0.05875 or 5.875%, whereas in calendar year 1981 a more typical value was 0.15875 or 15.875%. The quantitative data element “original mortgage amount” can only take on a finite number of values because, by statute, there is a maximum amount of a conforming mortgage and the amount must be in dollars and cents. During 2006, the maximum amount of a conforming single-family loan on a home located in one of the 48 contiguous States was $417,000.

5.2. Requirements Document

Much of the idea of data quality begins with the creation of a conceptual framework that answers the following questions:

- Why should the database be established?
- How are the data to be collected?
- How do analysts plan to use the data?
- What database-design issues may affect these data uses?

As a possible first step, the potential users may simply make a list of requirements for the data and associated system. This list is created to help ensure that the system designers will meet the needs of potential users. The final list may evolve into a requirements document that provides an overview of the system design.

---

\(^1\) FNMA = Federal National Mortgage Association and FHLMC = Federal Home Loan Mortgage corporation.