Guidelines for data submissions of high complexity

This document describes data formatting guidelines for high complexity cases (HCC) which standard data format types (see Data for Statistics) cannot handle. Submission of high complexity cases which this document describes require custom documentation by the submitting firm. *Any questions regarding selection of data format type for a specific submission should be directed to CVB Statistics. Contact CVB Statistics directly.*

When to use this guideline

In general, it may be useful to adopt the methods described when one or more of the following cases apply:

1. Use of the General data format type requires repeated entry of a single observation.
   - EX: Litter and multiple variables associated with source are reported. All records associated with a single source share a set of common variables such as vendor name and vendor city.
2. Use of the General data format type requires excessive use of "NA" due to experimental design.
   - EX: Clinical signs are observed at different frequencies, e.g. some are recorded daily and others weekly.
3. Observations are recorded for multiple unit types
   - EX: Clinical signs are recorded for sows and piglets

When NOT to use this guideline

Data submitted using an existing Data Format types may be prioritized for analysis. Only use HCC guidelines in cases for which existing Data Formats do not apply.

Vocabulary

- **Data Format Type**: One of General, ELISA, Bioassay, Titration, Checkerboard, or Software export as specified in Data for Statistics
- **Field**: A single column. See "variable"
- **Foreign Key**: A variable in a table where the value corresponds to a primary key of another table. Foreign keys are observations.
- **Observation**: One field of a single variable, not including primary keys.
- **Primary key**: An identifier variable unique to the table. Every table must have a primary key. No two rows within the table may have the same value for the primary key. A primary key is not an observation.
- **Record**: A row in the table. Also known as a *tuple*.

- **Variable**: A column in a table. Every variable has attributes of name, role, type and value. Depending on the role, options for type may be limited. Depending on the type, additional attributes may be required. The role and type of a variable may impose constraints on valid values.

What to submit

1. A CSV file for the variables table.

2. Separate CSV files for each custom table (minimum two per submission)

   – NOTE: If the submission contains two or fewer custom tables, it may be well-suited for the General data format type. Please contact CVB Statistics for help with a specific submission

File nomenclature

The set of data files generated using this guideline must have the same prefix, e.g. *dataset1*. A submission may have more than one set of data files and this prefix allows for grouping of related files. For the CSV files listed above, the suffix is as follows:

- **Variables**: variables

- **Custom**: user-defined string of maximum length 8 alphanumeric values

Offset the prefix from the suffix using underscore("_") as a separator. For example, files grouped with prefix *dataset1* could be named:

- dataset1_variables.csv

- dataset1_A.csv

- dataset1_B.csv

- ...

Variables data file contents

The variables data captures metadata for unique fields from the custom tables and allows CVB Statistics to understand relationships between tables. Each unique field is listed as a single row on the variables data table, along with the following attributes:

- **Table**: Name of custom table which the variable is the primary key (NA if variable is never a primary key)

- **Data Role**: Role the data plays in statistical analysis.

  – *identifier*: Unique identification given to each animal unit.

  – *grouping*: grouping factors are categories important in study design such as treatments or clusters. All efficacy studies have treatment groups, such as vaccine or placebo. Many have clusters, such as litter or pen. Note that testing device may be treated as a grouping variable if it is used for multiple tests (either of multiple test types or for testing of multiple animal units).

  - *observation*: direct observation or measurements. Examples would include clinical signs, testing tech, one-time use devices)
  - *timing*: Date or time. May be an actual date or time (Data type: date), or may be a count, such as number of days in relation to an event such as challenge.
  - *derived*: Value calculated or derived from other variables. The other variables must also be provided.

- **Data Type**: Primitive data type expected, e.g. *dichotomous*, *nominal*, *continuous*, *date*, *ranked*, *nominal*, *ordinal*, *censored*
- **Units**: For some data types, a unit of measure is required
- **Level Code(s)**: Binary and ordinal data types require level definitions. See General Format Instructions for specifics.
- **Comment**: Additional comments

Note that every custom table must have exactly one variable that is the primary key and each variable may be listed primary key a maximum of once. Some variables (e.g. body temperature) may not be a primary key on any custom table.

Additional discussion of the Data Role, Data Type, Units and Level Codes may be found in the General Format Instructions

Custom data files

While each submission is unique, we advise firms to structure tables to meet normality associated with third normal form. An overview of all five normal forms may be found here. A brief introduction to third normal form may be found here. Additional discussion and examples of normal forms may be found in literature specific to relational database theory.

Further reading

Codd, E.F. "Further Normalization of the Data Base Relational Model." (Presented at Courant Computer Science Symposia Series 6, "Data Base Systems," New York City, May 24th–25th, 1971.) IBM Research Report RJ909 (August 31st, 1971). Republished in Randall J. Rustin (ed.), Data Base Systems: Courant Computer Science Symposia Series 6. Prentice-Hall, 1972.

Hernandez, M.J. "Database Design for Mere Mortals: A Hands-on Guide to Relational Database Design", Ann Arbor: Addison-Wesley. 2013

Kent, William, "A Simple Guide to Five Normal Forms in Relational Database Theory", Communications of the ACM 26(2), Feb. 1983, 120-125. Also IBM Technical Report TR03.159, Aug. 1981.