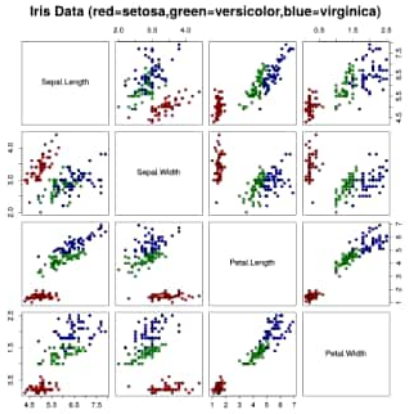
Data set

A **data set** (or **dataset**) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set.

Data sets can also consist of a collection of documents or files.[2]



Various plots of the multivariate dataset Iris flower data set introduced by Ronald Fisher (1936).[1]

In the open data discipline, data set is the unit to measure the information released in a public open data repository. The European data.europa.eu portal aggregates more than a million data sets.[3] Some other issues (real-time data

sources,[4] non-relational data sets, etc.) increases the difficulty to reach a consensus about it.[4]

Properties

Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis.[5]

The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data

(i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values are normally all of the same kind. However, there may also be *missing values*, which must be indicated in some way.

In statistics, data sets usually come from actual observations obtained by sampling a statistical population, and each row corresponds to the observations on one element of that population. Data sets may further be generated by algorithms for the

purpose of testing certain kinds of software. Some modern statistical analysis software such as SPSS still present their data in the classical data set fashion. If data is missing or suspicious an imputation method may be used to complete a data set.[6]

Classic data sets

Several classic data sets have been used extensively in the statistical literature:

- Iris flower data set – Multivariate data set introduced by Ronald Fisher (1936).[1]

- MNIST database – Images of handwritten digits commonly used to test classification, clustering, and image processing algorithms
- *Categorical data analysis* – Data sets used in the book, *An Introduction to Categorical Data Analysis*.

- Anscombe's quartet – Small data set illustrating the importance of graphing the data to avoid statistical fallacies

See also

- Data
- Data blending

- Data (computing)
- Data samples
- Data store
- Interoperability
- Data collection system
- List of datasets for machine-learning research