

Title

data management — Introduction to data-management commands

Description

This manual, called [D], documents Stata's data-management features.

Data management for statistical applications refers not only to classical data management—sorting, merging, appending, and the like—but also to data reorganization because the statistical routines you will use assume that the data are organized in a certain way. For example, statistical commands that analyze longitudinal data, such as `xtreg`, generally require that the data be in long rather than wide form, meaning that repeated values are recorded not as extra variables, but as extra observations.

Here are the basics everyone should know:

[D] use	Use Stata dataset
[D] save	Save datasets
[D] describe	Describe data in memory or in file
[D] inspect	Display simple summary of data's attributes
[D] codebook	Produce a codebook describing the contents of data
[D] data types	Quick reference for data types
[D] missing values	Quick reference for missing values
[D] list	List values of variables
[D] edit	Edit and list data using Data Editor
[D] rename	Rename variable
[D] format	Specify variable display format
[D] label	Manipulate labels

You will need to create and drop variables, and here is how:

[D] generate	Create or change contents of variable
[D] functions	Functions
[D] egen	Extensions to generate
[D] drop	Eliminate variables or observations

For inputting or importing data, see

[D] use	Use Stata dataset
[D] sysuse	Use shipped dataset
[D] webuse	Use dataset from web
[D] input	Enter data from keyboard
[D] insheet	Read ASCII (text) data created by a spreadsheet
[D] infile	Overview of reading data into Stata
[D] infile (fixed format)	Read ASCII (text) data in fixed format with a dictionary
[D] infile (free format)	Read unformatted ASCII (text) data
[D] infix (fixed format)	Read ASCII (text) data in fixed format
[D] hexdump	Display hexadecimal report on file
[D] odbc	Load, write, or view data from ODBC sources
[D] xmlsave	Save and use datasets in XML format
[D] fdasave	Save and use datasets in FDA (SAS XPORT) format
[D] icd9	ICD-9-CM diagnostic and procedure codes

and for exporting data, see

[D] outfile	Write ASCII-format dataset
[D] outsheet	Write spreadsheet-style dataset
[D] fdasave	Save and use datasets in FDA (SAS XPORT) format
[D] odbc	Load, write, or view data from ODBC sources

The ordering of variables and observations (sort order) can be important; see

[D] order	Reorder variables in dataset
[D] sort	Sort data
[D] gsort	Ascending and descending sort

To reorganize or combine data, see

[D] merge	Merge datasets
[D] append	Append datasets
[D] reshape	Convert data from wide to long form and vice versa
[D] collapse	Make dataset of summary statistics
[D] fillin	Rectangularize dataset
[D] expand	Duplicate observations
[D] expandcl	Duplicate clustered observations
[D] stack	Stack data
[D] joinby	Form all pairwise combinations within groups
[D] xpose	Interchange observations and variables
[D] cross	Form every pairwise combination of two datasets

In the above list, we particularly want to direct your attention to [D] **reshape**, a useful command beginners often overlook.

For random sampling, see

[D] sample	Draw random sample
[D] drawnorm	Draw sample from multivariate normal distribution

For file manipulation, see

[D] type	Display contents of a file
[D] erase	Erase a disk file
[D] copy	Copy file from disk or URL
[D] cd	Change directory
[D] dir	Display filenames
[D] mkdir	Create directory
[D] rmdir	Remove directory
[D] cf	Compare two datasets
[D] filefilter	Convert ASCII text or binary patterns in a file
[D] checksum	Calculate checksum of file

The entries above are important. The rest are useful when you need them:

[D] type	Display contents of a file
[D] notes	Place notes in data
[D] label language	Labels for variables and values in multiple languages
[D] labelbook	Label utilities
[D] encode	Encode string into numeric and vice versa
[D] recode	Recode categorical variable
[D] impute	Fill in missing values
[D] ipolate	Linearly interpolate (extrapolate) values
[D] destring	Convert string variables to numeric variables and vice versa
[D] mvencode	Change missing values to numeric values and vice versa
[D] pctile	Create variable containing percentiles
[D] range	Generate numerical range
[D] by	Repeat Stata command on subsets of the data
[D] statsby	Collect statistics for a command across a by list
[D] compress	Compress data in memory
[D] recast	Change storage type of variable

[D] assert	Verify truth of claim
[D] clonevar	Clone existing variable
[D] compare	Compare two variables
[D] contract	Make dataset of frequencies and percentages
[D] corr2data	Create dataset with specified correlation structure
[D] count	Count observations satisfying specified conditions
[D] duplicates	Report, tag, or drop duplicate observations
[D] isid	Check for unique identifiers
[D] lookfor	Search for string in variable names and labels
[D] memory	Memory size considerations
[D] obs	Increase the number of observations in a dataset
[D] separate	Create separate variables
[D] shell	Temporarily invoke operating system
[D] split	Split string variables into parts

There are some real jewels in the above, such as [D] **notes**, [D] **compress**, and [D] **assert**, which you will find particularly useful.

Also See

Complementary: [D] **intro**

Background: [R] **intro**