

How can I transpose data using PROC TRANSPOSE in SAS?

Authored by
stats writer

June 24, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I transpose data using PROC TRANSPOSE in SAS?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=150362>

PROC TRANSPOSE is a SAS procedure that allows for the reorganization of data by transposing columns into rows or vice versa. This procedure is useful when there is a need to switch the orientation of data for easier analysis or reporting. Using the TRANSPOSE procedure, users can specify the input and output datasets, as well as the variables to be transposed. This process is particularly helpful in situations where data needs to be rearranged to fit a specific analysis or presentation format.

What is a transpose?

Ideally, datasets are structured so that each row corresponds to one unique subject or object, and each column corresponds to a single variable. However, data can be recorded or collected in many different arrangements, depending on what is convenient or cost-effective for the data collector. Furthermore, we may need our data to be arranged in a particular format in order to use a specific analysis or procedure. This is where *transposing* or *reshaping* a dataset comes in to play.

Simple transposes

The simplest possible case of transposing switches the rows and columns of a matrix or dataset.

For example, consider the following 2x3 matrix (2 rows, 3 columns):

1 2 3

4 5 6

Transposing this matrix would turn it into a 3x2 matrix (3 rows, 2 columns):

1 4

2 5

3 6

Notice that the first row of matrix 1 becomes the first column of matrix 2. The matrix is essentially turned around a diagonal axis.

Wide datasets

A "wide" dataset contains exactly one row per subject, and uses a unique ID (called a "key") to identify each subject. Any measurements about that subject are recorded in columns. If a given measurement was taken under different conditions -- for example, in a within-subjects experiment where a subject completes a questionnaire more than once under different conditions -- there will be separate columns for each instance of the measurement.

When to use "wide" data

Statistical software packages typically require data to be in "wide" format for procedures like:

Paired t-tests Repeated measures ANOVA Correlation

Long datasets

A "long" dataset contains more than one row per subject, and uses a unique ID to identify each subject. Panel (or longitudinal) data is often recorded in this "long" format.

When to use "long" form data

Statistical software packages typically require data to be in "long" format for procedures like:

Time series Mixed and multilevel model

Visualizing a set of data in "long" and "wide" forms

Consider a clinic where patients come in for 1-month and 3-month follow-up visits after some procedure. As patients come into the clinic, each visit is recorded in the clinic's records. That is, each row of the "appointments" dataset corresponds to *visit*. A single visit record might contain information about the patient's name, the type of visit, and the weight of the patient during that visit. In this situation, the patient identifier and the type of visit are both "key" variables that uniquely identify each record; while the patient identifier uniquely identifies a given subject. This arrangement would be considered "long format", since there are multiple rows associated with each subject.

Unique Patient ID	Time	Weight
Patient1	1-month follow-up visit	Weight
Patient1	3-month follow-up visit	Weight
Patient2	1-month follow-up visit	Weight
Patient2	3-month follow-up visit	Weight
.	.	.
.	.	.
.	.	.
Patient n	1-month follow-up visit	Weight
Patient n	3-month follow-up visit	Weight

Now suppose you want to create a scatterplot of how the patients' weights changed between their 1-month and 3-month follow-up visits, or compute the correlation between these measurements. To do this, you might want to transpose the data so that each patient has one line of data that includes both weight values (i.e., a wide dataset), like below:

Unique Patient ID	Weight at Time=1-month visit	Weight at Time=3-month visit
Patient1	Patient1's Weight at 1-month visit	Patient1's Weight at 3-month visit
Patient2	Patient2's Weight at 1-month visit	Patient1's Weight at 3-month visit
.	.	.
.	.	.
.	.	.
Patient n	Patient n 's Weight at 1-month visit	Patient n 's Weight at 3-month visit

The TRANSPOSE Procedure

In SAS, PROC TRANSPOSE can perform simple transposes, as well as wide-to-long and long-to-wide restructuring of datasets. The general format is:

```
PROC TRANSPOSE DATA=Dataset-name OUT=New-dataset-name;
BY variable(s);
ID variable;
VAR variable(s);
RUN;
```

In the SAS code above:

The PROC TRANSPOSE statement tells SAS to execute the transpose procedure on an existing dataset called Dataset-name.

The OUT keyword says that the transposed dataset should be created as a new dataset called New-dataset-name. The BY statement is used to determine the row structure of the transposed dataset. You can include more than one variable in the BY statement. Your data must be sorted on your BY variables before running PROC TRANSPOSE.

For long-to-wide transposes, the BY variable(s) should uniquely identify each row. For wide-to-long transposes, the BY variable(s) determine the row structure of the long data; that is, it determines the repetition of the rows. The ID statement assigns names to the transposed value columns that match the values in the variable listed in the ID statement.

For long-to-wide transposes, the ID variable(s) determine the structure of the columns in the

transposed dataset. There will be one column for each unique value of the ID variable (or if multiple ID variables are present, one column for each unique combination of values). For wide-to-long transposes, you typically do not need an ID variable. However, if you do supply an ID variable, it will determine the column structure. The VAR statement is where you actually tell SAS what variables you want transposed. These are the values that will appear in the cells of the transposed variables.

For long-to-wide datasets, there is usually one variable in the VAR statement. For wide-to-long datasets, there are usually multiple variables in the VAR statement. The resulting dataset will have one row for each variable identified in the VAR statement.

Other options available in the PROC TRANSPOSE statement that can be found in the [SAS Help Guide](#).