Creating a Data Set from a Data Table

Paul R. Yarnold, Ph.D.

Optimal Data Analysis, LLC

This note addresses transforming a data table into a data set.

Tabular presentation of data is ubiquitous in scientific, business, sports, journalistic, and organizational reports. For a detailed discussion of how to prepare data for analysis by ODA (or any statistical) software, please see the article written by Fred Bryant and Patrick Harrison.[1]

Figure 1 is an example of a tabular cross-tabulation of two variables: Gender with two categories, and Reaction to Treatment with three categories (hypothetical treatment and data are presented for exposition).

Figure 1: Two-Variable Table

Gender	Patient's Reaction		
	Happy	Relaxed	Optimistic
Male	324	198	556
Female	429	217	499

First, name the two variables: GENDER (rows) and REACTION (columns) will suffice.

Second, code each of the variables. For GENDER, the codes are 1 for male, and 0 for female (arbitrary integers may be used, the user assigns and defines the codes). For REACTION, the codes are 1 for happy, 2 for relaxed, and 3 for optimistic (again, arbitrary codes).[1-4]

Third, transform the coded cross-classification table into a data set that ODA software can read and analyze.[1-3] The data for each of the (324 + 198 + 556) 1,078 males in the sample, and for the (429 + 217 + 499) 1,145 females in

the sample, must appear on separate lines in a text (also called ASCII) file. The total data set will thus consist of 1,078 + 1,145 = 2,223 space-delimited lines of data.

The first entry on each line is the code for GENDER, followed by a blank space, and then the second entry on each line is the code for REACTION. If there were more variables, the data would be coded and entered in the same manner as, and following, REACTION on each observation's line in the data set.[1-4]

Figure 2 presents space-delimited data for one observation from each cell in the Table in Figure 1. There will be 324 separate lines coded as "1 1" (Male, Happy); 198 separate lines coded as "1 2" (Male, Relaxed); 556 separate lines coded as "1 3" (Male, Optimistic); 429 separate lines coded as "0 1" (Female, Happy); 217 separate lines coded as "0 2" (Female, Relaxed); and 499 separate lines coded as "0 3" (Female, Optimistic).

Figure 2: Sample Two-Variable Data File Entries for Gender and Reaction

1	1
1	2
1	3
0	1
0	2
0	3

Whereas transforming a data table into a data set is not excessively laborious for applications consisting of relatively few observations, often a program is needed to help construct (write) the data set. The ODA laboratory uses software such as SAS[TM], SPSS[TM], STATA[TM] and EXCEL[TM], for example, to help transform data tables into data sets.[1-3]

References

[1]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software. *Optimal Data Analysis*, *2*, 2-6. URL: http://optimalprediction.com/files/pdf/V2A1.pdf

[2]Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, DC: APA Books, 2005.

[3]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[4]Ebert TA, Yarnold PR (2017). Running ODA™ software on Windows™: Assigning missing values using Excel™, and using longer file names. *Optimal Data Analysis*, *6*, 3-4. URL: http://optimalprediction.com/files/pdf/V6A1.pdf

Author's Notes

No IRB review is needed for this programming note, and no conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
      6348 N. Milwaukee Ave., Suite 163
      Chicago, IL 60646