NAME: Expenditure data for developmentally-disabled California residents
TYPE: Sample
SIZE: 1000 observations, 6 variables
ARTICLE TITLE: Simpson's paradox: A data set and discrimination case study exercise

DESCRIPTIVE ABSTRACT:

The State of California Department of Developmental Services (DDS) is responsible for allocating funds that support over 250,000 developmentally-disabled residents (referred to as "consumers"). The data set represents a sample of 1,000 of these consumers. Biographical characteristics and expenditure data (i.e., the dollar amount the State spends on each consumer in supporting these individuals and their families) are included in the data set for each consumer.

SOURCE:

The data set originates from DDS's "Client Master File." In order to remain in compliance with California State Legislation, the data have been altered to protect the rights and privacy of specific individual consumers.

VARIABLE DESCRIPTIONS:

The data reside in a csv file and are tab-delimited. A header line contains the name of the variables. There are no missing values.
Id: 5-digit, unique identification code for each consumer (similar to a social security number and used for identification purposes)
Age Cohort: Binned age variable represented as six age cohorts (0-5, 6-12, 13-17, 18-21, 22-50, and 51+)
Age: Unbinned age variable
Gender: Male or Female
Expenditures: Dollar amount of annual expenditures spent on each consumer
Ethnicity: Eight ethnic groups (American Indian, Asian, Black, Hispanic, Multi-race, Native Hawaiian, Other, and White non-Hispanic)

STORY BEHIND THE DATA:

These data are from a dataset used in an alleged case of discrimination privileging White non-Hispanics over Hispanics in the allocation of funds. Based on the initial analysis, it would appear that discrimination existed; however, a more in-depth analysis revealed that discrimination did not exist and that Simpson's-paradox phenomenon had occurred.

PEDAGOGICAL NOTES:

This data set can be used to teach a range of statistical concepts including Simpson's paradox. The importance of considering all variables in an analysis by conducting a bivariate (instead of just a univariate analysis) is highlighted in this data set.

SUBMITTED BY:

Name: Stanley Taylor and Amy Mickel
Affiliation: California State University, Sacramento
Address: College of Business Administration, CSUS, Sacramento, CA 95819-6088, USA
Email: sataylor@csus.edu, mickela@csus.edu

Documentation sources:
AMSTAT: http://www.amstat.org/publications/jse/v22n1/mickel/paradox_documentation.docx
STATLIT: www.StatLit.org/pdf/2014-Taylor-Mickel-Paradox-Documentation.pdf

Data sources:
AMSTAT: http://www.amstat.org/publications/jse/v22n1/mickel/paradox_data.csv
STATLIT: www.StatLit.org/XLS/2014-Taylor-Mickel-Paradox-Data.xlsx