Exploratory Data Analysis: DDS Case Study

Chapter 1, Lab 2

OpenIntro Biostatistics

This lab presents the details of how to conduct the analysis discussed in Section 1.7.1 of *OpenIntro Biostatistics*. A reader interested in applied data analysis may benefit from working through this lab and reviewing the solutions instead of reading the section in the text.

Background information

In the United States, individuals with developmental disabilities typically receive services and support from state governments. The State of California allocates funds to developmentally-disabled residents through the California Department of Developmental Services (DDS); individuals receiving DDS funds are referred to as 'consumers'. The dataset `dds.discr` represents a sample of 1,000 DDS consumers (out of a total population of approximately 250,000), and includes information about age, gender, ethnicity, and the amount of financial support per consumer provided by the DDS. The dataset is available in the `oibiostat` package.

A team of researchers examined the mean annual expenditure on consumers by ethnicity, and found that the mean annual expenditures on Hispanic consumers was approximately one-third of the mean expenditures on White non-Hispanic consumers. As a result, an allegation of ethnic discrimination was brought against the California DDS.

Does this finding represent sufficient evidence of ethnic discrimination, or might there be more to the story? This lab provides a walkthrough to conducting an exploratory analysis that not only investigates the relationship between two variables of interest, but also considers whether other variables might be influencing that relationship.

Distributions of single variables

To begin understanding a dataset and developing a sense of context, start by examining the distributions of single variables.

1. Load the `dds.discr` dataset into *RStudio*. Descriptions of the variables are provided in the documentation file. Produce a table of the first five rows in the data matrix.
2. Using appropriate numerical and graphical summaries, examine the distributions of each of the variables in the dataset and answer the following questions.
    a) Describe the distribution of annual expenditures. For most consumers, is the amount of financial support provided by the DDS relatively high or low?

b) The variable `age` directly records a consumer's age; in the `age.cohort` variable, consumers are assigned to one of six age cohorts. Describe the distribution of age in this sample of consumers. Do consumers tend to be older or younger?

The cohorts are indicative of particular life phases. In the first three cohorts, consumers are still living with their parents as they move through preschool age, elementary/middle school age, and high school age. In the 18-21 cohort, consumers are transitioning from their parents' homes to living on their own or in supportive group homes. From ages 22-50, individuals are mostly no longer living with their parents but may still receive some support from family. In the 51+ cohort, consumers often have no living parents and typically require the most amount of support.

c) Is there an equal representation of ethnic groups in this sample of consumers?

d) Does gender appear to be balanced in this sample of consumers?

Relationships between two variables

After examining variables individually, explore how variables are related to each other. It is often useful to start by investigating the relationships between two variables, particularly between the primary response variable of interest and the exploratory variables. For this case study, the response variable is `expenditures`, the amount of funds the California DDS allocates annually to each consumer.

3. How do annual expenditures vary by age? Is there a large amount of variation in expenditures between age cohorts? Use the `age.cohort` variable.

4. Do annual expenditures seem to vary by gender?

5. How does the distribution of expenditures vary by ethnic group? Does there seem to be a difference in the amount of funding that a person receives, on average, between different ethnicities?

A closer look

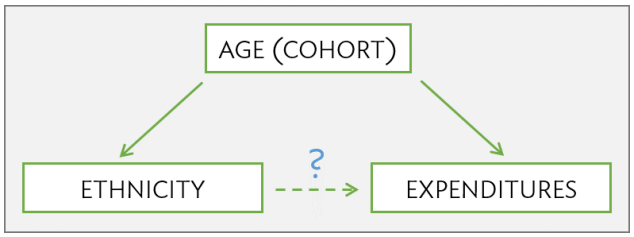
As shown in Question 2c), two of the ethnic groups, Hispanic and White non-Hispanic, comprise the majority of the data; some ethnic groups represent less than 10% of the observations. For ethnic groups with relatively small sample sizes, it is possible that the observed samples are not representative of the larger populations. The rest of this analysis will focus on comparing how `expenditures` varies between the two largest ethnic groups.

6. Compare the distribution of `expenditures` between Hispanic and White non-Hispanic consumers, graphically and numerically. Do Hispanic consumers, on average, seem to receive less financial support from the California DDS than a White non-Hispanic consumer?

7. Recall that expenditures is strongly associated with age—older individuals tend to receive more financial support. Is there also an association between age and ethnicity, for these two ethnic groups? Examine the distribution of age within each group and describe your findings.

When using data to investigate a question, it is important to explore not only how explanatory variables are related to the response variable(s), but also how explanatory variables influence each other.

Recall that a confounding variable is a variable that is associated with the response variable and the exploratory variable under consideration; confounding was initially introduced in the context of sunscreen use and the incidence of skin cancer, where sun exposure is a confounder. In this setting, age is a confounder for the relationship between `expenditures` and `ethnicity`. Just as it would be incorrect to claim that sunscreen causes skin cancer, it is essential here to recognize that there is more to the story than the apparent association between `expenditures` and `ethnicity`.



8. For a closer look at the relationship between age, ethnicity, and expenditures, compare how average `expenditures` differs by ethnicity within each age cohort. If age is indeed the primary source of the observed variation in expenditures, then there should be little difference in average `expenditures` between individuals in different ethnic groups but the same age cohort. Is this the case? Describe your findings.

9. Based on this exploratory analysis, does there seem to be evidence of ethnic discrimination in the amount of financial support provided by the California DDS? Summarize your findings in language accessible to a non-statistician.

Simpson's paradox

Identifying confounding variables is essential for understanding data. Confounders are often context-specific; for example, age is not necessarily a confounder for the relationship between ethnicity and expenditures in a different population. Additionally, it is rarely immediately obvious which variables in a dataset are confounders; looking for confounding variables is an integral part of exploring a dataset.

These data represent an extreme example of confounding known as **Simpson's paradox**, in which an association observed in several groups may disappear or reverse direction once the groups are combined. In other words, an association between two variables X and Y may disappear or

reverse direction once data are partitioned into subpopulations based on a third variable Z, the confounding variable.

Mean `expenditures` is higher for Hispanics than White non-Hispanics in all age cohorts except one. Yet, once all the data are aggregated, the average expenditures for White non-Hispanics is over twice as large as the average for Hispanics. This paradox can be explored from a mathematical perspective by using weighted averages, where the average expenditure for each cohort is weighted by the proportion of the population in that cohort.

10. Calculate the overall weighted average expenditures for Hispanics and for White non-Hispanics, using the proportions of individuals in each age cohort (Question 7) and the average expenditures for each Cohort (Question 8). How does the weighting lead to overall average expenditures for White non-Hispanics to be higher than for Hispanics?