Simpson's Paradox: A Data Set and Discrimination Case Study Exercise

Stanley A. Taylor
Amy E. Mickel
California State University, Sacramento

Journal of Statistics Education Volume 22, Number 1 (2014),
www.amstat.org/publications/jse/v22n1/mickel.pdf

Key Words: Univariate analysis; Bivariate analysis; Specific variation; Outliers; Weighted average; Simpson's paradox.

Abstract

In this article, we present a data set and case study exercise that can be used by educators to teach a range of statistical concepts including Simpson's paradox. The data set and case study are based on a real-life scenario where there was a claim of discrimination based on ethnicity. The exercise highlights the importance of performing rigorous statistical analysis and how data interpretations can accurately inform or misguide decision makers.

1. Introduction

Statistics has played a key role in discrimination cases for decades. As the Supreme Court has stated, "our cases make it unmistakably clear that '[s]tatistical analyses have served and will continue to serve an important role' in cases in which the existence of discrimination is a disputed issue" (Int'l Bhd. of Teamsters v. United States 1973). When decision outcomes are heavily influenced by statistical evidence, it is imperative that data have been properly analyzed. Failure to perform a sufficient analysis can lead to misunderstandings and misguided decisions that can have far-reaching implications for a range of stakeholders.

One well-known arithmetic phenomenon is Simpson's paradox (Simpson, 1951) or the Yule–Simpson effect. This is a paradox when an association or comparison that holds for several groups reverses direction when the data are combined to form a single group (Moore, McCabe, and Craig 2012). An example of this phenomenon is when the University of California,

Berkeley was sued for bias against women who had applied for admission to graduate schools in 1973. Admission figures showed that men applying were more likely than women to be admitted, and the difference was so substantial that one would conclude that discrimination existed. However, when examining individual academic departments, it appeared that no department was significantly biased against women (Bickel, Hammel, and O'Connell 1975). In other words, there was no significant difference between the number of men and women admitted when looking at several groups (i.e., departments); however, this finding reversed (suggesting that more men were admitted than women) when the departmental groups were combined into one single group of all students admitted to UC Berkeley graduate schools.

While there are a number of other real-life examples of Simpson's paradox (see Guber 1999; Schneiter and Symanzik 2013), simple but convincing examples based on real data are limited (Appleton, French, and Vanderpump 1996). Despite research emphasizing the effectiveness of teaching statistical theory through application (e.g., open-ended data analyses and case studies) (e.g., Nolan and Speed 1999), there are even fewer data sets that students can use to experience this phenomenon first hand.

In this paper, we present a data set and case study exercise illustrating Simpson's paradox along with other statistical concepts. This exercise is based on a scenario that the lead author encountered on one of his many consulting engagements for the State of California. The situation involved an alleged case of discrimination privileging White non-Hispanics over Hispanics in the allocation of funds to over 250,000 developmentally-disabled California residents. Based on the initial analysis, it appeared that discrimination existed; however, a more in-depth analysis revealed that discrimination did not exist and that Simpson's-paradox had occurred.

This case study exercise is ideal for statistics courses for several reasons.

- The topic itself captures students' interest for claims of discrimination are prevalent in our society.
- Critical thinking is promoted; analysis, synthesis, and decision-making skills are used.
- The importance of identifying and analyzing all sources of specific variation (i.e., potential influential factors) in statistical analyses is highlighted.
- Students are introduced to the statistical concepts of weighted averages, outliers, univariate and bivariate analyses, and Simpson's paradox.

2. Case Study Exercise: Background and Learning Objectives

Most states in the USA provide services and support to individuals with developmental disabilities (e.g., intellectual disability, cerebral palsy, autism, etc.) and their families. The agency through which the State of California serves the developmentally-disabled population is the California Department of Developmental Services (DDS). Both authors have provided consulting services to this department, and one of the consulting engagements is the basis for this case study exercise.

One of the responsibilities of DDS is to allocate funds that support over 250,000 developmentally-disabled residents (referred to as "consumers"). A number of years ago, an allegation of discrimination was made and supported by a univariate analysis that examined average annual expenditures on consumers by ethnicity. The analysis revealed that the average annual expenditures on Hispanic consumers was approximately one-third (⅓) of the average expenditures on White non-Hispanic consumers. This finding was the catalyst for further investigation; subsequently, state legislators and department managers sought consulting services from a statistician (the lead author).

Understanding the concept of specific variation, the statistician looked for other potential sources of variation including age. A bivariate analysis examining ethnicity and age (divided into six age cohorts) revealed that ethnic discrimination did not exist. Moreover, in all but one of the age cohorts, the trend reversed where the average annual expenditures on White non-Hispanic consumers were less than the expenditures on Hispanic consumers—a classic example of Simpson's paradox!

Surprisingly, some of the members of the state legislative bodies and department staff still did not understand how this was possible, nor did they understand the related statistical concepts. This led to our desire to create a case study based on this real-life scenario with the following learning objectives:

(a) to increase students' knowledge of specific variation, outliers, univariate and bivariate analyses, weighted averages, and Simpson's paradox;
(b) to enhance student' analytical and critical thinking skills when making decisions based on statistical analyses; and
(c) to demonstrate the importance of performing rigorous statistical analysis and how decision outcomes are often profoundly impacted by interpretations of data.

In the following sections, we describe the data set and its variables. Next, a set of instructions on how to incorporate this exercise into course curriculum is provided. We conclude with a brief discussion on the value of this case study exercise.

3. Data Set

The data set presented is designed to represent a sample of 1,000 DDS consumers (which provides a 95% confidence interval with a margin of error of +-3.5% for this 250,000 consumer population).[1] The data set includes six variables (i.e., fields) which are: ID, age cohort/age (binned/unbinned), gender, expenditures, and ethnicity (see Appendix for data set).

"ID" is the unique identification code for each consumer. It is similar to a social security number and used for identification purposes.

"Age cohort/age" is a key variable in the case exercise. While age is a legal basis for discrimination in many situations, age is not an attribute that would be considered in a

[1] The data set originated 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. The provided data set is based on actual attributes of consumers.

discrimination claim for this particular population. The purpose of providing funds to those with developmental disabilities is to help them live like those without disabilities. As consumers get older, their financial needs increase as they move out of their parent's home, etc. Therefore, it is expected that expenditures for older consumers will be higher than for the younger consumers.

We have included both binned ("Age cohort") and unbinned ("Age") variables to represent a consumer's age. The binned age variable is represented in the data set as six age cohorts. Each consumer is assigned to an age cohort based on their years since birth. The six cohorts include: 0-5 years old, 6-12, 13-17, 18-21, 22-50, and 51+. The cohorts are established based on the amount of financial support typically required during a particular life phase.

The 0-5 cohort (preschool age) has the fewest needs and requires the least amount of funding. For the 6-12 cohort (elementary school age) and 13-17 (high school age), a number of needed services are provided by schools. The 18-21 cohort is typically in a transition phase as the consumers begin moving out from their parents' homes into community centers or living on their own. The majority of those in the 22-50 cohort no longer live with their parents but may still receive some support from their family. Those in the 51+ cohort have the most needs and require the most amount of funding because they are living on their own or in community centers and often have no living parents.

> *Teaching Note: We suggest instructors allocate some time discussing the age cohorts and how they were generated—based on theory. Instructors might want to discuss that age boundaries for the cohorts could differ as long as the theoretical basis is sound. The key teaching moment here is to emphasize the meaning of a cohort and the role theory plays in their creation.*

"Gender" is included in the data set as another variable to consider because it is an attribute on which many discrimination cases are based. In this exercise, gender does not play a significant role; there is no obvious difference in the distribution of funds between male and female consumers.

"Expenditures" variable represents the annual expenditures the State spends on each consumer in supporting these individuals and their families. It is important that students realize this is the amount each consumer receives from the State. Expenditures include services such as: respite for their families, psychological services, medical expenses, transportation, and costs related to housing such as rent (especially for adult consumers living outside their parent's home).

"Ethnicity" is the key demographic variable in the data set as it pertains to the case. Eight ethnic groups are represented in the data set. These groups reflect the demographic profile of the State of California.[2]

[2] www.dof.ca.gov/research/demographic

4. Incorporating the Data Set and Case Study Exercise

In this section, we describe how this case study exercise has been incorporated into our courses. A discussion on general instructions, analytical tools, and three phases of analyses are described. Several teaching notes are also included. These are designed to provide guidelines for instructors who are using the case for the first time. We encourage instructors to use the data set and case study exercise in ways that meet their learning objectives. In addition, we suggest those instructors teaching in non-traditional formats (e.g., on-line) to adapt the exercise accordingly.

4.1 General Instructions

In our courses, the students are first told that their primary task is to analyze the data set and determine whether or not discrimination exists by examining the expenditures (i.e., amount of money the State spends on the consumers). For this exercise, we explain that discrimination exists if the amount of expenditures for a typical person in a group of consumers that share a common attribute (e.g., gender, ethnicity, etc.) is significantly different when compared to a typical person in another group. For example, discrimination based on gender would occur if the expenditures for a typical female are less than the amount for a typical male. This usually leads to a discussion of what the terms "typical" and "significant" mean.

There are a few different statistics that can be used to measure the typical amount received including means and medians. We encourage instructors to lead a discussion about the differences between means and medians. Regarding the concept of significance, we instruct students to restrict their analysis to descriptive statistics rather than inferential statistics. In addition to keeping the case fairly simple, the purpose for this instruction is to promote a discussion about the differences between statistically significant and practically significant.

We require students to submit a report of their findings. We suggest the report be two-pages in length plus any tables, figures, and graphs that would help illustrate and support their claim as to whether or not discrimination existed in this case study exercise. We allow the students one week to complete the assignment. Prior to the submission of the final report, we allocate 45 minutes in two different class periods to discuss the case and data analyses.

4.2 Analytical Tools

There are a number of statistical software packages and analytical tools that can be used for this exercise. We require our students to analyze this data set using pivot tables[3] which is a built-in feature of Microsoft Excel and to focus on comparing means. A pivot table is a dynamic table that allows students to interpret data in different ways without having to enter a formula. Pivot table reports are particularly useful in narrowing down larger data sets or analyzing relationships between data points. There are a number of good online tutorials about pivot tables. In our opinion, one of the best on-line tutorials used by our students can be found on YouTube (www.youtube.com/watch?v=Vx-Fuw46VbY).

[3] Our motivation for exposing students to pivot tables is in response to feedback from two Fortune 500 companies in the high tech industry that frequently hire our graduates. They explained that they use pivot tables regularly and would like to have future employees prepared to use and appreciate this analytical tool.

We focus on the mean, versus the median, for the following reasons: (a) it is a statistic with which students are familiar and refer to as an average and (b) pivot table functionality is limited to means (i.e., medians are not offered as an option). For those instructors that prefer to focus on medians, they will want to consider using other analytical tools.

For instructors who do not teach Excel but want to focus on means, the same analysis may be accomplished by using other software and analytical tools. Examples of software and relevant commands include: (a) Minitab [Start, Basic Stat, Display Descriptive Statistics], (b) SPSS [Analyze, Compare Means, Means], SAS [Proc Means], and R [Aggregate ( )].

In subsequent sections, we describe the analysis process which we divide into three different phases. It is important to note that our classes are conducted in computer labs on campus. We encourage instructors to adapt these instructions accordingly based on their delivery format.

4.3 Phase 1 of the Analysis

After we discuss the general instructions in class (see 4.1), we then ask students to open the data file in a data analysis software program (e.g., Excel) to answer some basic questions. We pose the following questions: "What is the average of expenditures for: (a) all males, (b) all Hispanics, (c) all 22-50 year olds, (d) all male, White non-Hispanics, and (e) all Asian, 22-50 year olds?" We do this to ensure that students are comfortable with the analytical tools and with the overall exercise.

> *Teaching Note: In our classes, we verbally pose these questions and allow a short amount of time for responses between each question. We do this to illustrate that a pivot table is an efficient way to summarize data from different perspectives. We encourage instructors using other analytical tools to consider doing a similar exercise. This process can be adapted to a non-classroom setting as well.*

Students are then assigned the homework of preparing data tables to support their findings as to whether or not discrimination exists. In the following class period, we have students present the tables they have generated. Typically, there is a general consensus that discrimination exists. In the next paragraphs, we present the four most common tables students generate, along with graphical displays of the data.

4.3.1 First Typical Table: Ethnicity and Average of Expenditures

Students often present the data highlighting the differences in average expenditures by ethnicity (see Table 1 and Figure 1). Students typically focus on the extreme values, especially the discrepancy between the high values of the Native Hawaiian and American Indian consumers and the low values of the Multi-Race and Other groups. Post discussion, students are typically convinced that ethnic discrimination exists.

Ethnicity of Consumers	Average of Expenditures ($)
American Indian	$ 36,438
Asian	$ 18,392
Black	$ 20,885
Hispanic	$ 11,066
Multi Race	$ 4,457
Native Hawaiian	$ 42,782
Other	$ 3,317
White non-Hispanic	$ 24,698
All Consumers	**$ 18,066**

Table 1. Average Expenditures by Ethnicity

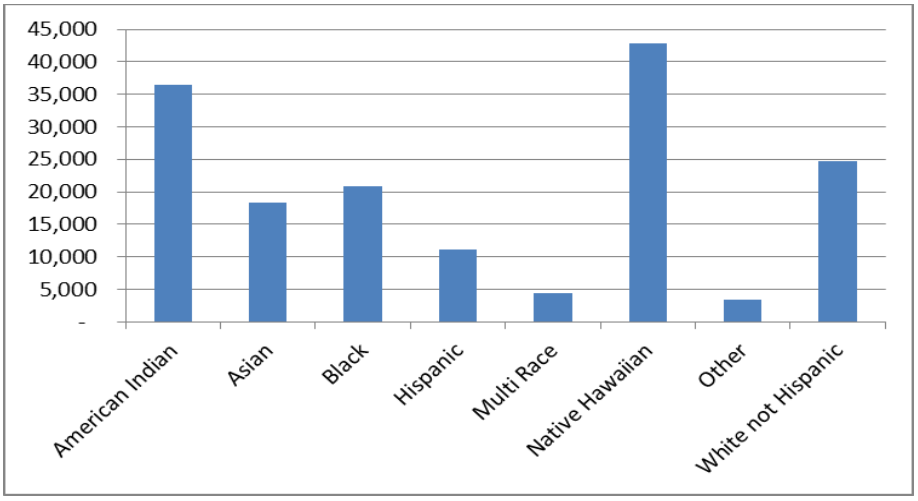

Figure 1. Average Expenditures by Ethnicity

4.3.2 Second Typical Table: Gender and Average of Expenditures

The students often present the data comparing average expenditures by gender (see Table 2 and Figure 2). With regards to possible gender discrimination, the students usually conclude that there is no evidence of discrimination.

Gender	Average of Expenditures ($)
Female	$ 18,130
Male	$ 18,001
All Consumers	**$ 18,066**

Table 2. Average Expenditures by Gender

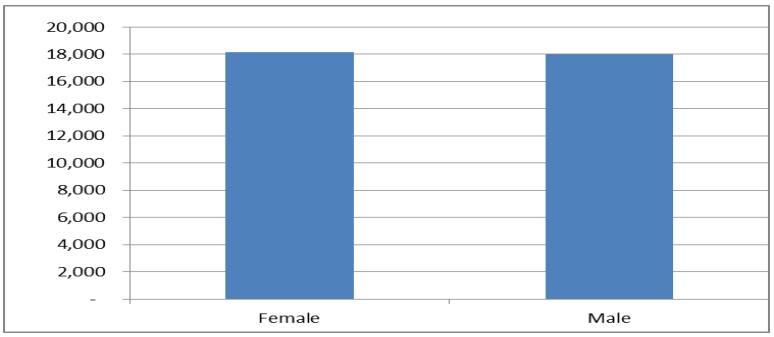

Figure 2. Average Expenditures by Gender

4.3.3 Third Typical Table: Age Cohort and Average of Expenditures

The students also present the data comparing average expenditures by age cohort (see Table 3 and Figure 3). With regards to possible age discrimination, there is typically a fair amount of discussion about these findings. We remind students that the needs for consumers increase as they become older which results in higher expenditures.

Age Cohort	**Average of Expenditures ($)**
0 – 5	$ 1,415
6 – 12	$ 2,227
13 – 17	$ 3,923
18 – 21	$ 9,889
22 – 50	$ 40,209
51 +	$ 53,522
All Consumers	**$ 18,066**

Table 3. Average Expenditures by Age

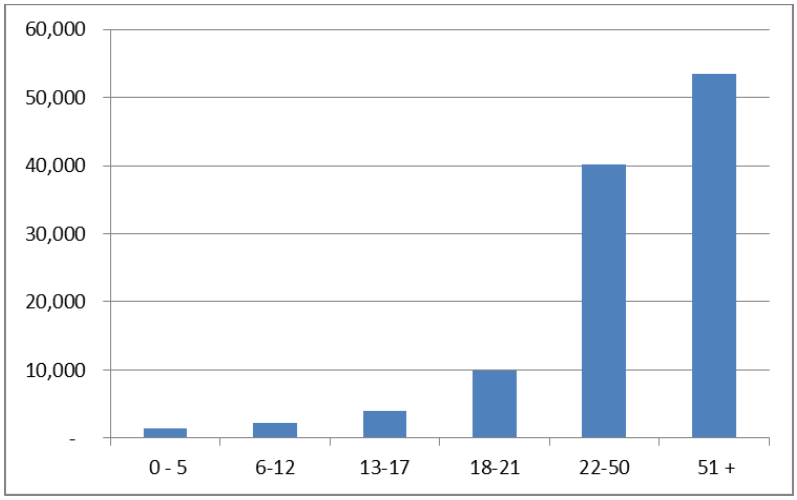

Figure 3. Average Expenditures by Age

4.3.4 Fourth Typical Table: Percentages of Ethnic Groups and Expenditures

The students sometimes present data (see Table 4 and Figure 4) that compare the percentages of: (a) the sum of expenditures and (b) the number of consumers in that ethnic sub-population. Students explain that if discrimination did exist then the percentage profiles should be very different. During this discussion, we introduce the idea of comparing means and/or medians which sets the stage for the next phase of analysis.

Ethnicity	Sum of Expenditures ($)	% of Expenditures	# of Consumers	% of Consumers
American Indian	$ 145,753	1%	4	0%
Asian	$ 2,372,616	13%	129	13%
Black	$ 1,232,191	7%	59	6%
Hispanic	$ 4,160,654	23%	376	38%
Multi Race	$ 115,875	1%	26	3%
Native Hawaiian	$ 128,347	1%	3	0%
Other	$ 6,633	0%	2	0%
White non-Hispanic	$ 9,903,717	55%	401	40%
Total	**$ 18,065,786**	**100%**	**1,000**	**100%**

Table 4. Percentages of Expenditures and Consumers

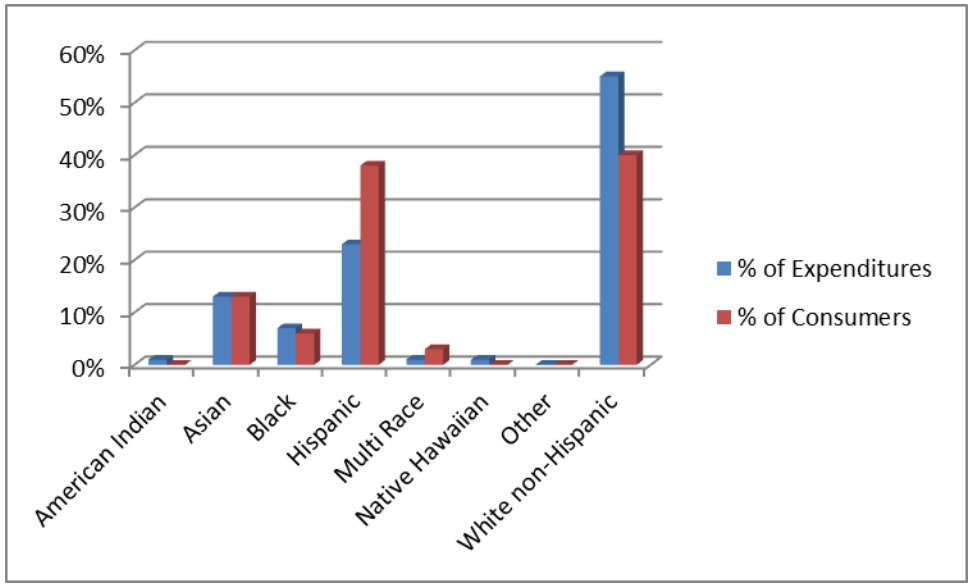

Figure 4. Percentages of Expenditures and Consumers

4.4 Phase 2 of the Analysis

In this phase, we ask students to consider Table 5 and Figure 5. In this discussion, we focus on the definition and impact of outliers when doing data analysis. After the outlier discussion, students conclude that ethnic groups with small sample sizes should not be considered. At this point, we encourage students to focus on the two largest ethnic groups: White non-Hispanic and Hispanic. (Table 5.1 is the "collapsed" version of Table 5.) We do stress that by focusing the analysis on the two largest groups does not imply that the other ethnic groups are unimportant.

Ethnicity	**Average of Expenditures ($)**	**% of Consumers**
American Indian	$ 36,438	0%
Asian	$ 18,392	13%
Black	$ 20,885	6%
Hispanic	$ 11,066	38%
Multi Race	$ 4,457	3%
Native Hawaiian	$ 42,782	0%
Other	$ 3,317	0%
White non-Hispanic	$ 24,698	40%
All Consumers	**$ 18,066**	**100%**

Table 5. Average Expenditures and % of Consumers by Ethnicity

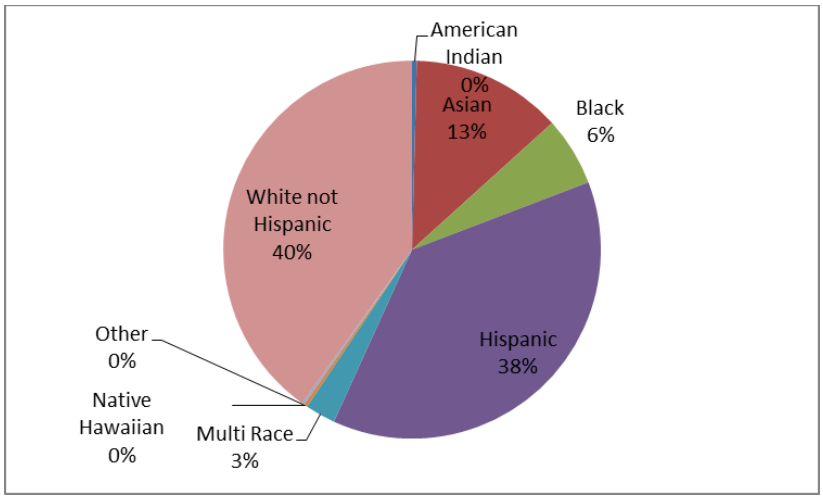

Figure 5. % of Consumers by Ethnicity

Ethnicity	Average of Expenditures ($)	% of Consumers
Hispanic	$ 11,066	38%
White non-Hispanic	$ 24,698	40%

Table 5.1 Average Expenditures and # of Consumers by Ethnicity

After the students examine Table 5.1, there is general consensus among the students that there is a significant difference in the average amount of expenditures between the White non-Hispanic and Hispanic groups. We then ask them to generate justifiable reasons as to why there might be differences in the averages and to determine if discrimination truly exists.

> *Teaching Note: At this point, we remind the students that this case study is based on a real-life scenario and that we are focusing on practically significant differences (as opposed to statistically significant differences).*

Similar to the responses that those working for the State of California generated when asked the same question during our consulting engagement, students came up with the following reasons: (a) Hispanics have more family support, and therefore, are less likely to seek government-funded assistance, and (b) Hispanics are less informed about how to seek assistance. Both of these reasons are difficult to model and could lend support for allegation of discrimination. Next we instruct students to conduct a bivariate analysis of the data by including age cohorts (in addition to their ethnicity). The results of this analysis for the Hispanic and White non-Hispanic sub-populations are presented in Table 6 and Figure 6.

Age Cohort	Hispanic (avg. of expenditures)	White non-Hispanic (avg. of expenditures)
0 – 5	$ 1,393	$1,367
6-12	$ 2,312	$2,052
13-17	$ 3,955	$3,904
18-21	$ 9,960	$10,133
22-50	$ 40,924	$40,188
51 +	$ 55,585	$52,670
All Consumers	**$11,066**	**$24,698**

Table 6. Average Expenditures by Ethnicity and Age Cohort

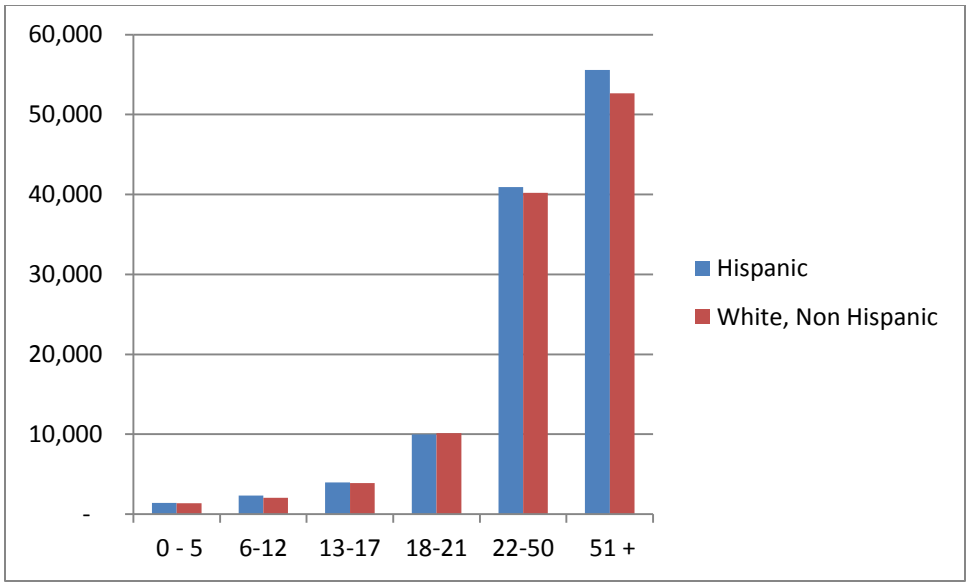

Figure 6. Average Expenditures by Ethnicity and Age Cohort

When asked to interpret these findings, most students still focus on the difference in the average of expenditures for all consumers in the ethnic groups. When asked about the age cohorts, they are perplexed. Before explaining the paradox, we redirect the students to the original question of whether or not discrimination exists. In other words, we ask: "Is the typical Hispanic receiving fewer funds (i.e., expenditures) than the typical White non-Hispanic?"

We point out that if a Hispanic consumer was to file for discrimination based upon ethnicity, s/he would more than likely be asked his/her age. Since the typical amount of expenditures for Hispanics (in all but one age cohort) is higher than the typical amount of expenditures for White non-Hispanics in the respective age cohort, the discrimination claim would be refuted.

> *Teaching Note: We have found that an actual example helps most students see this more clearly. If a Hispanic consumer was to claim discrimination because s/he is Hispanic (vs. White non-Hispanic), s/he might do so based on the overall average of expenditures for all consumers in their group ($11,066 vs. $24,698). However, if the Hispanic consumer states that s/he is 25 years old, the average of expenditures for this age cohort is slightly higher than the White non-Hispanic in the same age cohort ($40,924 vs. $40,188).*

At this point, most students are confused. There exists a paradox (Simpson's paradox!) which they do not understand. Rather than provide the answer, we instruct the students to answer the question of: "Why is the overall average for all consumers significantly different indicating ethnic discrimination of Hispanics, yet in all but one age cohort (18-21) the average of expenditures for Hispanic consumers are greater than those of the White non-Hispanic population?"

4.5 Phase 3 of the Analysis

In order to answer the previous question, an analysis similar to the one presented in Table 7 is conducted. The main difference between Table 6 and Table 7 is that the number of consumers—rather than the average of expenditures—are presented.

Age Cohort	**Hispanic** (# of consumers)	**White non-Hispanic** (# of consumers)
0 – 5	44	20
6-12	91	46
13-17	103	67
18-21	78	69
22-50	43	133
51 +	17	66
All Consumers	**376**	**401**

Table 7. # of Consumers by Ethnicity and Age Cohort

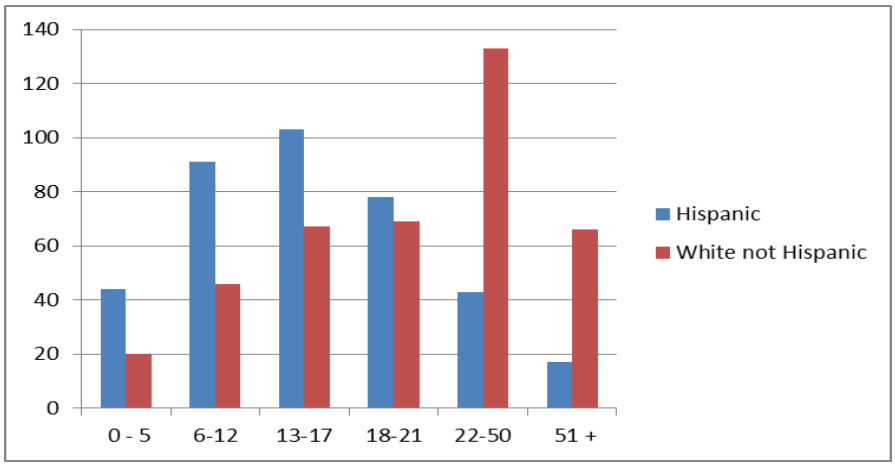

Figure 7. # of Consumers by Ethnicity and Age Cohort

Most students realize that there are more Hispanics in the youngest four age cohorts, while the White non-Hispanics have more consumers in the oldest two age cohorts. Since the two populations are close in overall counts (376 vs. 401), students are generally able to use this information along with fact that consumers expenditures increase as they age (see Table-Figure 3) to see the paradox.

We then explain Simpson's paradox and discuss how the paradox is relevant to this case exercise: expenditure average for Hispanic consumers are higher in all but one of the age

cohorts, but the trend reverses when the groups are combined resulting in a lower expenditure average for all Hispanic consumers when compared to all White non-Hispanics. Table 8 also helps illustrate this paradox by showing the percentages of consumers in each age cohort. This leads into a discussion of weighted averages.

Age Cohort	Hispanic (%)	White non-Hispanic (%)
0 – 5	12%	5%
6-12	24%	11%
13-17	27%	17%
18-21	21%	17%
22-50	11%	33%
51 +	5%	16%
All Consumers	**100%**	**100%**

Table 8. Bivariate Table: Percentages by Ethnicity and Age Cohort

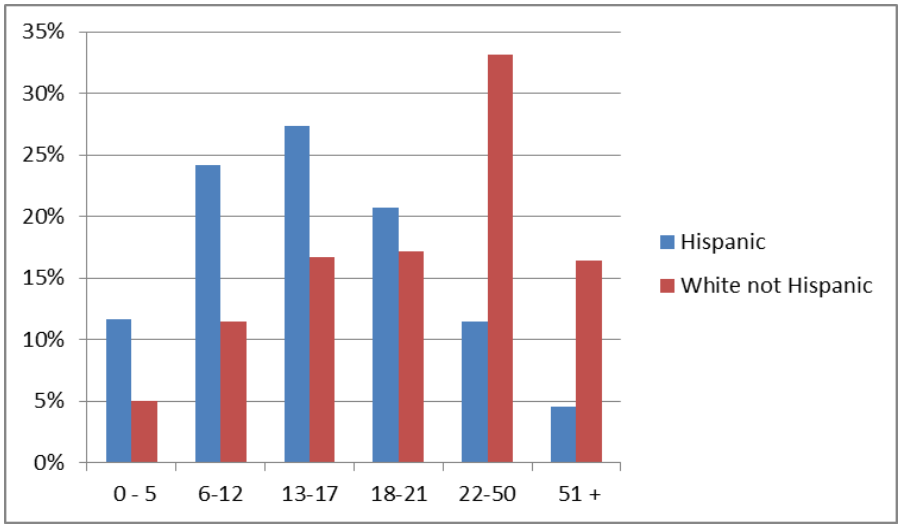

Figure 8. Percentages by Ethnicity and Age Cohort

The students are then provided the formula for a weighted average.

$$\bar{X}_k = \sum_{i=1}^{6} w_{k,i}\bar{X}_{k,i}$$

Where

$\bar{X}_k$ represents the mean of the k[th] ethnic group,

$w_{k,i}$ represents the percentage of the k[th] ethnic group in the i[th] age group, and

$\bar{X}_{k,i}$ represents the mean for the k[th] ethnic group in the i[th] age group.

We discuss the paradox and how the weights for the Hispanic population are higher for the youngest four age cohorts and lower for the oldest two age cohorts when compared to the White non-Hispanic population. In other words, the overall Hispanic consumer population is a relatively younger when compared to the White non-Hispanic consumer population. Since the expenditures for younger consumers is lower, the overall average of expenditures for Hispanics (vs White non-Hispanics) is less.

> *Teaching Note: To reinforce the mathematics involved in this scenario, we have the students use the weights (percentages) in Table 8 , along with the corresponding age cohort expenditure averages in Table 6 to calculate the overall expenditure means.[4] We first instruct them to let K represent the Hispanic population and then to let K represent the White non-Hispanic population.*

One additional way to visually illustrate the distribution of ages for each ethnic group is through Figure 9 where the vertical axis represents the population frequency per ethnic group and the horizontal axis represents age by year.[5] Students are able to see that the White non-Hispanic consumers are overall an older population than the Hispanic consumers.

[4] There will be some rounding; as a result, the estimated overall means may not be exactly the same as those shown in Table 6.

[5] Please note that we truncated the data set to include only those up to 50 years old. We did this to highlight the described phenomenon and make the figure easier to read (scaling).

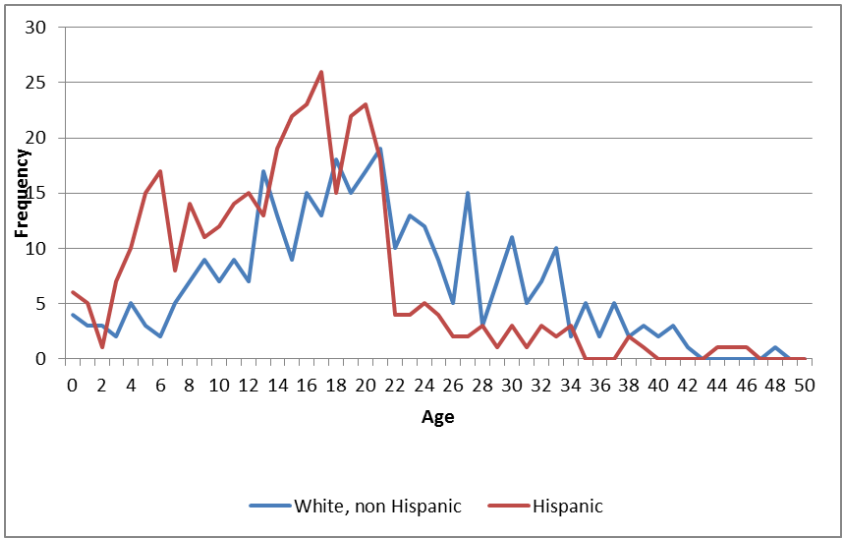

Figure 9. Population Profiles: Ethnicity and Age (unbinned)

The final discussion of this case study emphasises the importance of conducting rigorous statistical analysis by considering all possible variables that may be contributing to findings. We discuss how the outcome of important decisions (such as discrimination claims) are often heavily influenced by statistics and how an incomplete analysis may lead to poor decision making. We encourage students to think about how the initial interpretation of data could have led to misguided decisions and how such decisions can have far-reaching implications for a range of stakeholders.

5. Closing Remarks

In addition to the case study exercise presented in the article, we encourage instructors to use this rich data set in other ways to meet their learning objectives. For example, if the data set was treated as representing an entire population, the data could be used to illustrate simple random sampling vs. stratified sampling inferences. The data could also be used to teach other exploratory data analysis techniques such as dot plots or statistical concepts such as regression analysis.

Beyond the data set, the case may also set the stage for future exercises. For example, we have required students to compare ethnic profiles of different geographic areas by analyzing data provided by the California Department of Finance at www.dof.ca.gov/research/demographic. Moreover, the case presented could lead into future discussions about discrimination. Instructors may want to incorporate exercises such as the one Miao (2010) describes involving the use and interpretation of statistics in a legal case involving claims of discrimination.

This article makes several contributions. First, a data set based on a real-life scenario is presented and can be used by educators throughout the country. In addition, we provide guidelines to help instructors incorporate the data set and case study exercise into their courses.

Most importantly, the proposed learning objectives are achieved by engaging students in an exercise that focuses on an issue that is pertinent to today's society—discrimination. More specifically, students gain a greater understanding of statistical concepts related to specific variation, outliers, univariate and bivariate analyses, weighted averages, and Simpson's paradox. They hone their analytical and critical thinking skills. Lastly, students gain an appreciation for the importance of performing rigorous statistical analysis and how decision outcomes are often profoundly impacted by interpretations of data.

Appendix

Expenditure Data

The "Expenditure Data" includes 1000 observations with 6 variables in the data file.

This data set is available as a comma-separated value Excel file and can be accessed at: http://www.amstat.org/publications/jse/v22n1/mickel/paradox_data.csv

A documentation file for the data set is available as a .pdf file and can be accessed at: http://www.amstat.org/publications/jse/v22n1/mickel/paradox_documentation.docx

---

References

Appleton, D.R., French, J.M., and Vanderpump, M.P.J. (1996), "Ignoring a Covariate: An Example of Simpson's Paradox.'' *The American Statistician, 50(4),* 340-341.

Bickel, P.J., Hammel, E.A., and O'Connell, J.W. (1975), "Sex Bias in Graduate Admissions: Data from Berkeley.'' *Science, 187(4175),* 398-404.

Guber, D. (1999), "Getting What You Pay For: The Debate Over Equity in Public School Expenditures" *Journal of Statistics Education* [online], 7(2). Available at http://www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm

Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 339 (1973).

Miao, W. (2010), "Did the Results of Promotion Exams Have a Disparate Impact on Minorities? - Using Statistical Evidence in *Ricci v. DeStefano.*" *Journal of Statistics Education* [online], 18(3). Available at www.amstat.org/publications/jse/v18n3/miao.pdf

Moore, D.S., McCabe, G.P., and Craig, B.A. (2012), *Introduction to the Practice of Statistics (Seventh Edition).* W. H. Freeman and Company, New York, NY.

Nolan, D. and Speed, T. P. (1999), "Teaching Statistics Theory through Applications." *The American Statistician, 53(4),* 370-375.

Schneiter, K. and Symanzik, J. (2013), "An Applet for the Investigation of Simpson's Paradox." *Journal of Statistics Education* [online] 21(1). Available at www.amstat.org/publications/jse/v21n1/schneiter.pdf

Simpson, E.H. (1951), "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society (Series B), 13*, 238–241.

Stanley A. Taylor
California State University, Sacramento
College of Business Administration
6000 J Street
Sacramento, CA 95819-6088
(916) 278-5439
sataylor@csus.edu

Amy E. Mickel
California State University, Sacramento
College of Business Administration
6000 J Street
Sacramento, CA 95819-6088
(916) 278-7195
mickela@csus.edu
