MELDING DATA WITH SOCIAL JUSTICE IN UNDERGRADUATE STATISTICS AND DATA SCIENCE COURSES

Silas Bergen
Winona State University, Winona, MN, USA
sbergen@winona.edu

The world is increasingly filled with large, rich, and publicly-available data. These data are from a wide array of contexts including education, foreign policy, criminal justice, housing, and health care. Accordingly, undergraduate instructors of statistics and data science have an invaluable opportunity to engage students in social justice through the lens of quantitative analytics. In this paper I elaborate on ways I have incorporated these topics and data across a wide range of undergraduate statistics and data science courses. I also describe ways I seek to foster student reflection on the realities of social inequity, not just as data analysts but as world citizens. I conclude by discussing challenges I have faced and opportunities for future growth.

INTRODUCTION

We live in a world where the amount of large, rich, and publicly-available data grows ever greater. These data concern a wide array of topics ranging from education, foreign policy, criminal justice, housing, health care, and economic opportunity, serving as invaluable fodder both for training in quantitative proficiency and discussing social justice issues. I believe that we have a responsibility as educators to consider how to instill both analytic proficiency and social responsibility and consciousness in our students.

Lesser (2007) offers an excellent exposition on teaching statistics with social justice. He states that statistics could be called "the grammar of social justice," and discusses how a number of statistical concepts can be approached in the context of social justice. He also provides resources for those further interested. I resonate with Lesser on the need to engage students on this topic. Statistics especially offers a unique forum for social justice conversations. When trained to reason quantitatively with data, students can begin to fully understand large-scale social inequity in ways that were inaccessible to them before. Only then can informed and critically-sound conversations take place about solutions for inequity.

I introduce data topics relating to social justice in my courses with at least one of three goals in mind. Goal 1 is to promote growth in quantitative proficiency. Goal 2 is to use the results of quantitative analysis to raise awareness of social inequity. Goal 3 is to go beyond awareness, and to foster conversations about and reflections on social inequity. I would posit that Goal 1 is the responsibility of any instructor in statistics or data science. Of Goals 2 and 3, Goal 2 is easier to accomplish from an instructor's standpoint, as the instructor does little more than present or lead students to discover objective information gleaned from an analysis. An example might be showing students a graphic of home ownership across race, and asking if the percentages of those who own their own home differs across race. Goal 3 is most difficult. The instructor must not only consider which issues to raise, but how to foster dialogue and reflection in a constructive and meaningful way.

In this paper, I describe ways in which I have sought to accomplish all three goals with my undergraduate statistics and data science students. I conclude by describing the successes and challenges I have so far encountered and reflect on future directions.

INTRODUCTORY STATISTICS

The life-course framework

A major way I have sought to accomplish all three goals in my introductory statistics courses is with a final project, which comprises 15% of their overall grade. Prior to the final project throughout the entire semester, students learn traditional introductory statistics material such as data types and summaries, estimation and hypothesis testing for one or two parameters. For the final project, I use two large, publicly-available data sources. The first is microdata from the American Community Survey. This is a survey of millions of Americans conducted by the US Census Bureau, and is collected to help government officials, community leaders and businesses

In: J. Engel (Ed.), *Promoting understanding of statistics about society.*
Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), July 2016, Berlin, Germany. iase-web.org/Conference_Proceedings.php

understand the changes taking place in their communities. Data include information on employment status, income, gender, race, state of residence, travel time to work, housing situation, etc. The second data source is mother/infant data produced by the North Carolina State Center for Health Statistics and compiled by the North Carolina Vital Statistics Dataverse. This source contains remarkably detailed data on mother/infant dyads including access to prenatal care, race, infant birth weight and gestational age.

Along with the quantitative aspects of the project, one of its primary objectives is to introduce students to the life course perspective (Lu and Halfon, 2003). The life course perspective is the idea that certain factors such as access to health care and nutritional foods, social support, and economic opportunity influence health over an entire life course. The presence of these factors serve to protect good health, while their absence serves as a risk factor for poor health. Current racial disparities in these factors have a "trickle down" effect in future generations. Racial inequity in health of and social support for mothers associates with the health of her infants, and hence racial disparities carry over to the health of future generations. On the project, students specifically investigate disparity across race in income, employment, prenatal care, and infant birth weight.

Goal 1 is accomplished by requiring students to synthesize all the concepts they have learned over the semester. The project comes at the end of the semester in which the analytic groundwork has been laid for students to conduct simple tests like the 2-sample t-test and chi-square tests of association. On the project students must answer open-ended questions about inequity without instructions as to which test to use. As a result of their analyses, students discover differences across race in employment, income, access to prenatal care, and low birth weight; hence Goal 2 is accomplished.

Prior to Spring Semester 2016, I sought to accomplish Goal 3 by asking a series of questions following the analytic portion of the project. These questions are designed to foster reflection on race and how it shapes the way we perceive the world. They are not required to answer these, but are offered extra credit if they do. Questions I have posed include the following:

1. *What are the racial stereotypes we hold in our heads? What are the things we think but don't say?*
2. *What would happen if we respectfully discussed our thoughts on and questions about different ethnic groups in a multiethnic setting? What would you ask? What would you say?*
3. *What are some ways you can think of to challenge your implicit stereotypes?*
4. *What are some steps you can take to better understand other cultures?*
5. *In what ways has your racial/ethnic background shaped your current values, habits, practices, and personal priorities?*

These questions following statistical analyses provide a way for me to engage the heart as well as the mind. My intent is that they realize an analysis doesn't just stop with the appropriate statistical conclusion, but should lead to societal changes. With these questions I hope to increase student awareness of racial-based stereotypes and help them think more deeply about how race influences the way they think about and interact with their classmates, co-workers, and beyond. I am often struck and encouraged by the thoughtfulness of student responses. Many of them echo a need for learning from and better understanding other cultures, often citing travel as a means to implement this. At this point I can encourage them to experience one of the many cultural engagement and education experiences hosted by the international students attending WSU, as travel is often prohibitively expensive for many students.

For the first time in Spring Semester 2016 thanks to reviews on this manuscript, I introduced a new component to this final project. This new component sought to tie in student's perspectives on racial inequality more explicitly with what they learned from the quantitative aspect of the project. This new component was inspired by an article by the Pew Research Center entitled *King's Dream Remains an Elusive Goal; Many Americans See Racial Disparities*. This article presents two sets of interesting findings. The first set includes the results of a survey designed to measure how Americans' views of racial equality differ across race. I was particularly interested in the results of two of these questions presented in the article. The first question was

"*How much progress toward Martin Luther King's dream of racial equality do you think the U.S. has made over the last 50 years?*" and the second was "*How much more needs to be done in order to achieve racial equality?"* The second set of findings presented in the article shows how racial gaps in metrics such as median household income, life expectancy, and educational attainment have narrowed, widened or stayed the same over time. I decided to incorporate these survey questions and the analyses of racial gaps in the final project in Spring Semester 2016. Thus, in addition to carrying out the formal statistical analyses using the ACS and North Carolina vital statistics data described previously to investigate racial inequality in the present time, I also had them investigate how racial gaps have varied over time using informal graphical summaries. For this part of the project, I used 2 data sets. The first was Integrated Public Microdata Series (IPUMS) from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) (Flood et al, 2015). These data contain historic trends on poverty rate (percent living below the poverty line, a metric that depends both on size and income of a household), median household income (in dollars), and high-school completion rate (in percentages). The second was data from the National Center for Health Statistics (2012) on neonatal mortality rate (number of deaths per 1000 live births); low birth weight rate (% of live births born low birth weight); teenage childbearing rate (% of live births born to women under 18 years of age); and life expectancy at birth (in years). Students were required to create graphs to investigate how racial gaps narrowed or widened over time. An example of where gaps have narrowed is in teenage childbearing rate, as shown in the top panel of Figure 1 below, whereas the racial gap in median income has widened over time as shown in the bottom panel.

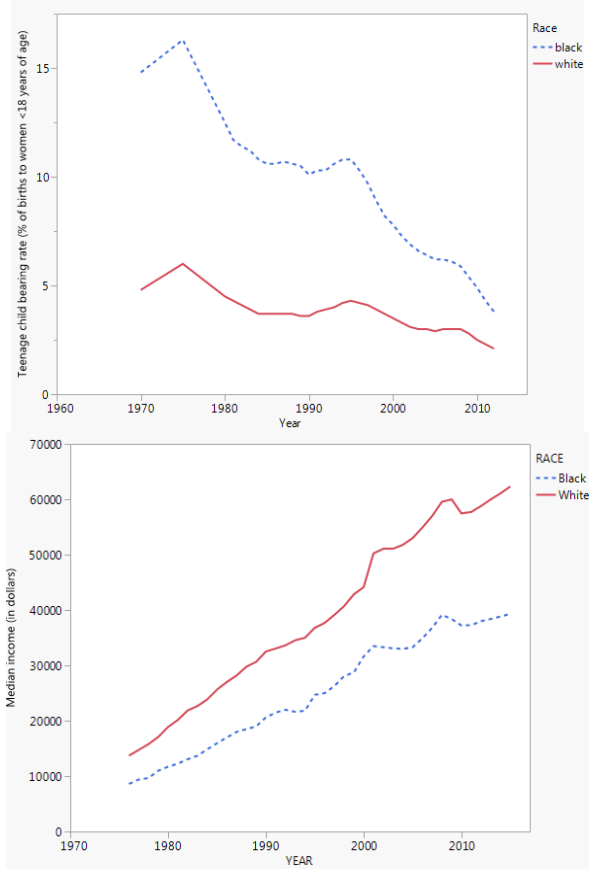

Figure 1: Racial gaps in teenage childbearing rate and median income over time

In order to more rigorously measure changes in student perspectives on racial inequality due to the project and to measure the project's overall effectiveness, I designed pre-project and post-project surveys using the Qualtrics software (Qualtrics Labs, 2016). The pre-project survey asked students the same two questions from the Pew article described above. Students had to select one of the following responses to each question: "A lot," "Some," "A little," or "None at

all." These were the same options given to participants in the Pew survey. The post-project survey contained these same two questions asked again, along with two additional questions not asked in the pre-project survey: *"I am more informed about the nature of racial inequality having completed this project"* (Choose: Strongly agree; Agree; Disagree; or Strongly Disagree) and *"The nature of racial inequality seems more complex after completing the project"* (Choose: Strongly agree; Agree; Disagree; or Strongly Disagree). Both the pre- and the post-project surveys were optional, and extra credit opportunities were offered to those who participated. The results of the 37 students who responded to both the pre- and the post-surveys are shown in Tables 1 and 2 below. 44 student responses to the two questions that were only on the post-survey are shown in Tables 3 and 4.

	Post-survey responses			
Pre-survey responses	None at all	A little	Some	A lot
None at all	0	0	0	0
A little	0	0	1	0
Some	0	5	14	1
A lot	0	1	7	9

Table 1: Counts of pre- and post-project student responses to the question: *How much progress toward Martin Luther King's dream of racial equality do you think the U.S. has made over the last 50 years?*

	Post-survey responses			
Pre-survey responses	None at all	A little	Some	A lot
None at all	0	0	0	1
A little	0	0	4	0
Some	0	2	8	8
A lot	0	1	1	13

Table 2: Counts of pre- and post-project student responses to the question: *How much more needs to be done in order to achieve racial equality?*

Post-survey responses			
Strongly disagree	Disagree	Agree	Strongly agree
2 (4.5%)	3 (6.8%)	28 (63.6%)	11 (25.0%)

Table 3: Counts (percents) of post-project student responses to the question: *I am more informed about the nature of racial inequality having completed this project*

Post-survey responses			
Strongly disagree	Disagree	Agree	Strongly agree
2 (4.5%)	5 (11.4%)	20 (45.4%)	18 (40.9%)

Table 4: Counts (percents) of post-project student responses to the question: *The nature of racial inequality seems more complex after completing the project*

The overall takeaways from Tables 1 and 2 appear to be that after completing the project, students tended to grow more pessimistic about U.S. progress towards racial equality and the amount that needs to be done in order to achieve racial equality. Tables 3 and 4 indicate that students tended agree or strongly agree that they learned a lot about racial inequality and its complexities from the final project.

Race and out-of-school suspensions

Another activity I assign in my introductory statistics courses was inspired by a National Public Radio (NPR) show *This American Life*. The episode was titled "Is This Working?" and centers on discipline in public schools. One of the stories from the episode focuses on racial disparity in preschool suspensions. The story describes the "School-to-prison pipeline", the idea that disciplining students with out-of-school suspension can push them into the juvenile and criminal justice systems. Hence, racial disparity in suspension can create racial disparity in educational opportunity as the suspended students are sent home, as well as exacerbating the disparity in racial demographics of the criminal justice system.

The activity I give my students uses data from the Office for Civil Rights overseen by the U.S. Department of Education. The Civil Rights Data Collection makes publicly available a wide range of education access and equity data across the country's public schools. The project I give to students uses data from Winona Senior High School in 2013. Figure 2 summarizes the data.

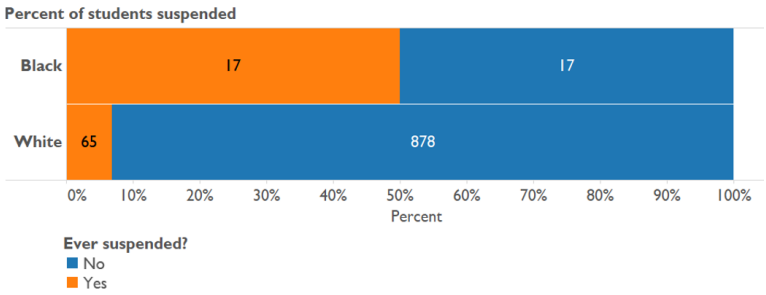

Figure 2. Out-of-school suspensions by race.

I have my students listen to the *This American Life* episode, then show them the data. This project provides a great example of Fisher's exact test. I ask them, "If I walked into the school and chose 82 students at random (i.e., without bias) and suspended them, would it seem strange to randomly select 17 black students, given there are only 34 black students overall compared to 943 white students?" Often their intuition agrees that this seems very unlikely, and they can use the p-value from Fisher's Exact Test to officially measure that likelihood. Admittedly, valid arguments could be made that using any formal statistical test is unnecessary here, since the data are on the entire population. However, as an introductory concept I find this to be a helpful feature of the data, as it is more conducive to motivating the fixed racial and suspension rates that Fisher's exact test requires than if it came from a random sample, where in practice marginal totals would likely differ from sample to sample. The project concludes with 2 questions:

1. *Write 3-4 sentences summarizing your findings, as if you were reporting them to a WSHS administrator or at a parent-teacher meeting. What "action steps" might you recommend to school administrators and teachers as a result of your findings?*
2. *What do you think about using suspension as a method of disciplining bad behavior?*

I find this project successfully engages students on many different levels. First, they are exposed to excellent reporting on racial inequity in the form of a popular radio show. Second, they see evidence of racial disparity in a high school just a couple miles from WSU. Third, they are required to reflect on their findings as if they were a consultant to school administrators. Since many of my students are mathematics education majors, they especially are able to appreciate the responsibility teachers and school officials have in addressing racial disparity in classroom discipline.

UPPER-LEVEL STATISTICS

One of the upper-level statistics course I have taught recently is an introduction to statistical modeling course. This course introduces simple and multiple linear regression, one- and two-way ANOVAs, and logistic regression. Taylor and Mickel (2014) provided an excellent case study of Simpson's Paradox that blends statistics with social justice. I have adapted their paper

into a final project for this course. The data (which Taylor and Mickel make publicly available) come from the California Department of Developmental Services (DDS), and provide age, ethnicity, and annual expenditures on developmentally disabled Californian residents (referred to as "consumers"). By performing a two-sample t-test, it is clear that the DDS's average expenditure on Hispanic consumers is significantly lower than on white non-Hispanic consumers. However, using a variety of visualization as well as modeling techniques such as linear regression and analysis of variance which they learn in the course, students find that expenditures are higher for older consumers, and that the white consumers tend to be older. Thus, students should realize that the expenditure/ethnicity association is confounded by age. Fitting a model that controls for age, there is no longer any association of expenditure with ethnicity. The final question on the project is as follows:

> *Write a paragraph clearly summarizing your "big picture" findings, and stating whether you believe these data provide evidence of discrimination. If you think there is discrimination, describe the nature of the discrimination. If you do not, clearly explain why not.*

This question requires them to synthesize all their models to write a conclusion. My intent with this project is to build in my students a healthy skepticism of sensational headlines, and to consider other factors that might be in the background of any apparent bivariate association. Students often grasped the need to consider multiple variables when looking at an association. Sample student responses to the question asked above are provided below:

"The allegation was that the DDS was discriminating against Hispanic consumers, since average annual expenditure on Hispanic consumers was less than spending on White non-Hispanic consumers. …[T]his allegation is not correct…. [A]nnual average expenditure increases as age increase, and there are more Whites in the data set that are older than Hispanics…. [W]e investigated how annual average expenditures differed across Ethnicity group, when comparing people of the same Age Cohort. We learned from this analysis that Hispanics actually have a slightly higher, yet insignificant, average annual expenditure than Whites!"

"When purely looking at the expenditures between the two ethnic groups, it is easy to see how someone might mistakenly assume that there is a significant difference between the expenditures of these two groups. However, this difference is not due to the difference in ethnicity, but instead due to the difference in ages of the people that make up these two groups…. [T]he average age for Whites in this data set are higher than the average age for Hispanics…. [A]s Age increases, so does annual expenditures. When looking at the two ethnic groups, while controlling for age, there is no evidence of a difference between Whites and Hispanics."

DATA VISUALIZATION

An exciting and unique arena for considering social justice topics is in data visualization. This is a new course which was offered at WSU for the first time in Fall 2015 as part of our new Data Science major. I had the privilege of teaching it the first time it was offered. One of the projects I assigned was used individual income data from the American Community Survey. Students were required to use the Tableau software to visualize gender disparity in median income across the U.S. One of the visualizations could be a map, while the other was required to be something other than a map. Figure 3 shows a map of median income disparity, along with a different take on visualizing the same data. The District of Columbia has the greatest gender disparity, but this is obscured in the map by its small size. The slope graph at right visualizes this much more clearly, where we can see median income for females on the left of the graph and median income for males on the right of the graph. In the slope graph, we see that of all states, the District of Columbia exhibits the greatest disparity between median male and female income. We were unable to determine this based on the map shown at left, due to the District of Columbia's small size.

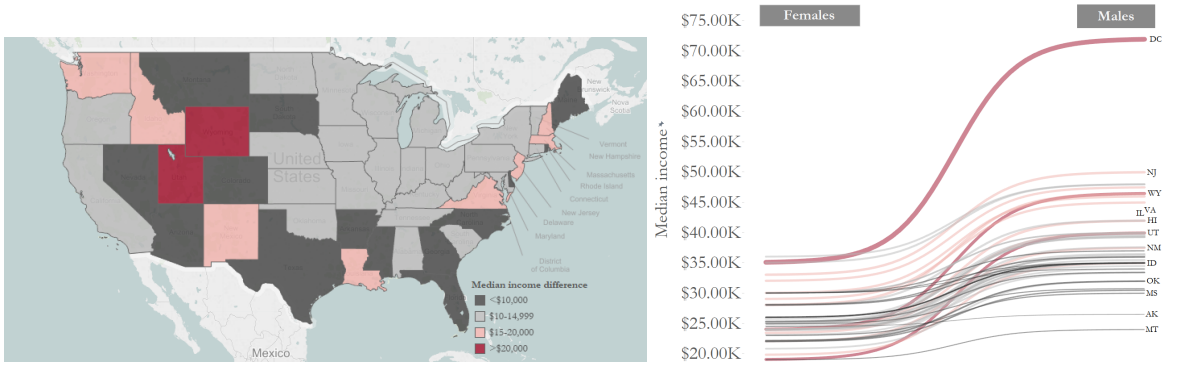

Figure 3: Two methods of visualizing income inequity across gender

Another project involved visualizing racial inequity in health care access using data from the Behavioral Risk Surveillance System (BRFSS). Specifically, interest lay in investigating disparity in ability to see a doctor because of cost and access to a personal health care provider. The first challenge was to think of a metric with which to define inequity; the second challenge was to visualize the metric. One of the visualizations is shown in Figure 4. On the left is the "disparity score" measuring inequity in ability to see a doctor because of cost across race and on the right is the disparity score of the state measring racial inequity in access to a personal health care provider. The disparity score is simply a standardized χ^2 statistic measuring heterogeneity in the percent unable to afford doctor's care across race (on the left) and percent without access to a health care provider (on the right).

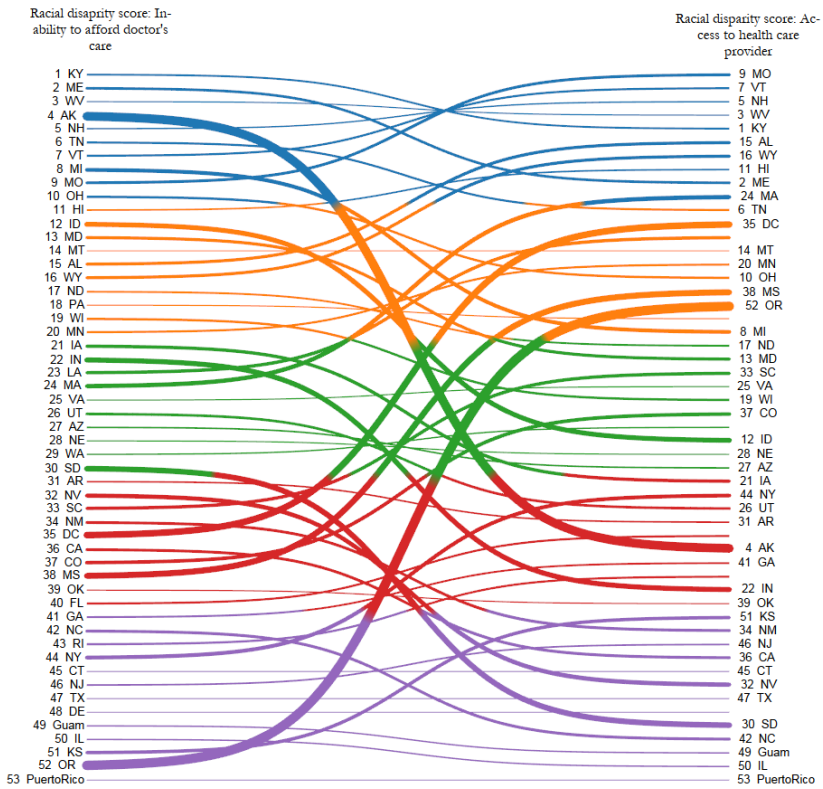

Figure 4: Examing racial disparity in health care access across state

The data visualization class used public data to create exciting visualizations addressing social topics. One particular student in the class ran with these ideas, and for his final project created an impressive visualization investigating the success of the United Nations' Millennium Development Goals in Southeast Asia. His visualization can be viewed at http://tabsoft.co/1UmY2UQ.

CONCLUSION

The current data climate provides a unique opportunity to engage in social justice coversations through the quantitative lens of data analytics. I have presented ways in which I have sought to accomplish this across a variety of different undergraduate classes. In my introduction, I outlined 3 goals one should keep in mind when introducing social justice concepts with data. I also explained why Goal 3 is the hardest to accomplish. Of the projects I have described in this paper, the only ones where I feel I have successfully accomplished all 3 goals are my introductory statistics courses. However, there is still room for growth even in how I approach Goal 3 for these courses. With all of my projects and reflection questions, I do not engage in face-to-face conversations about social justice with my students. This is partly to allow them space to reflect honestly, but also reflects some discomfort and insecurity about my ability as a white male to succesfully lead face-to-face classroom dialogues about disparity across race and gender. Another challenge is to know when and how to discuss the *whys* of racial inequity. All of my projects and exercises are mainly intented to raise awareness of what racial inequity is from a data perspective, but I do little (if any) follow-up to help students understand the root causes of inequity or ways it can be alleviated. This is partly due to time constraints, and also due to not feeling well-equipped to lead such discussions. One of the reasons I am eager to attend the IASE roundtable is to hear others' experiences and ideas to improve in this respect.

Another area of potential future growth is in my upper level statistics and data visualization courses. The projects I described in my upper level statistics and data visualization courses accomplish Goals 1 and 2, but it is not clear that they accomplish Goal 3. I am again eager to get feedback on how to successfully accomplish Goal 3 in these courses.

Yet a third area of potential growth is more subtle, and perhaps most difficult of all. Is it possible to measure whether one is successful in engaging students about social inequity, and in bringing about real change in the way they interact with others who are different from them? I am still unsure how to answer this question for myself, and not even sure it can be answered. However, I eagerly look forward to thoughtful roundtable discussions on this question as an avenue for potential growth in this area.

REFERENCES

Centers for Disease Control and Prevention (CDC). (2014). Behavioral Risk Factor Surveillance System Survey Data. *U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.*

Flood, S; King, M; Ruggles, S; and Warren, JR. (2015). *Integrated Public Use Microdata Series, Current Population Survey: Version 4.0***.** [Machine-readable database]. Minneapolis: University of Minnesota.

Lesser, L. (2007). Critical Values and Transforming Data: Teaching Statistics with Social Justice. *Journal of Statistics Education*, *Volume 15*(1), 1-21.

Lu, MC and Halfon, N. (2003). Racial and ethnic disparities in birth outcomes: a life-course perspective. *Maternal and Child Health Journal*, *Volume 7*(1), 13-30.

National Center for Health Statistics. (2012). Health, United States: With Special Feature on Emergency Care. Hyattsville, MD. Accessed via http://www.cdc.gov/nchs/data/hus/hus12.pdf

National Public Radio (2008). Is this working? *This American Life*. WBEZ. Radio.

Pew Research Center. (2013). King's Dream Remains an Elusive Goal; Many Americans See Racial Disparities. Web. http://www.pewsocialtrends.org/2013/08/22/kings-dream-remains-an-elusive-goal-many-americans-see-racial-disparities/.

Taylor, SA and Mickel, AE. (2014). Simpson's Paradox: A Data Set and Discrimination Case Study Exercise. *Journal of Statistics Education, Volume 22*(1), 1-18.

United States Department of Education. (2016). Civil Rights Data Collection. *Office for Civil Rights.* Web. http://ocrdata.ed.gov/.

United States Census Bureau. (2016). 2008 – 2010 American Community Survey. *U.S. Census Bureau's American Community Survey Office.* Web. http://ftp2.census.gov/.

University of North Carolina. (2016). North Carolina Vital Statistics. *The Odum Institute Dataverse.*