Debate Fact Checking: If you are a Data Analyst, It's Just What You Do

Dale Lehman, Center for Business Analytics

Watching the recent Democratic presidential debate from Milwaukee, I was struck by a disturbing statistic cited by both candidates. The relative incarceration rate of African Americans to Whites in Wisconsin has been twice the national average – and the national average of 5.6 is already disturbingly high. I applaud the attention focused on this issue, but data analysts are never satisfied with unscrutinized data. So, I decided to look into this statistic, and that set off a chain of inquiries that exemplifies what data analysts do and what keeps them awake at night.

FIRST: CHECK THE DATA

It was correct. According to a 2005 report from the Bureau of Justice Statistics, the incarceration rate for African Americans was 10.6 times as high as for Whites in Wisconsin compared with a 5.6 national average rate. More recent data from the University of Wisconsin at Milwaukee found that 12.8% of Black adult males residing in Wisconsin were incarcerated compared with 6.7% nationally.[1] Actually, it is not that simple (it never is). The Census counts the "residence" of prisoners as their place of incarceration rather than the home that they lived in when they became incarcerated. So, the use of Census data may be skewed towards the location of prisons and may not be representative of where the incarcerated population was living when they first became incarcerated. I will not consider this issue further, but I will note that this issue has been examined by the Census Bureau and they explain in detail the difficulties of trying to establish a permanent "place of residence" for each prisoner.[2]

Given the Census Bureau's definition of residence, it also means that prisoners in Federal facilities should not be counted in any analysis of state variations in incarceration rates by race. A number of states have no federal prisons, so presumably federal crimes committed in those states result in imprisonment in other (probably geographically close by) states. Federal prison populations would then distort any analysis of variations across states. Also, "local" jails are provided by the state prisons in Alaska, Connecticut, Delaware, Hawaii, Rhode Island, and Vermont, so these states will show no local jail prisoners but proportionally higher incarceration rates in state prisons.

I found it refreshing that presidential candidates would focus on the incarceration rates of African Americans – this is surely a national tragedy that needs to be addressed as a priority. To understand and tackle this problem, however, we need to go beyond the superficial comments offered in the presidential debate. We need to explain two separate, but perhaps related, phenomena: why are African Americans incarcerated at so much higher rates than Whites, and why is their relative rate in Wisconsin so much higher than in most of the country? The candidates raised some issues that might be relevant, such as lack of economic opportunity and the large incarceration rates for drug-related crimes. There was also some reference to the need for police to better represent the demographics of the areas

[1] This is from a 2013 report based on 2010 Census Data, https://www4.uwm.edu/eti/2013/BlackImprisonment.pdf.
[2] http://www.census.gov/newsroom/releases/pdf/2006-02-21_tabulating_prisoners.pdf.

they serve and the need to hold police accountable for their actions. But are these the relevant issues? Are they somehow "worse" in Wisconsin than elsewhere?

The reason this matters is that any proposed remedy must be based on an understanding of the cause. Without accurate diagnosis of the problem, how can we hope to find cures? If income inequality is the cause, then we should observe more inequality in Wisconsin than in other states. If biased policing is the cause, then we should find that Wisconsin police officers behave "worse" than police in other states. If enforcement of drug laws is the cause, then data should show that Wisconsin enforces drug laws on African Americans more than other states? To the extent that the answers to these questions is that Wisconsin is no worse than elsewhere, then these "causes" would fail to explain what we see in the data.

SECOND: FIND MORE DATA

So, this set me on a search for data I could use to investigate the relationships between incarceration rates, crime rates, economic opportunity, population characteristics, etc. Finding such data at the state level, broken down by race, proved to be fraught with difficulty. That fact is critical.

WITHOUT GOOD DATA, MANY PHENOMENA WILL NOT BE EXPLORED, AND WILL REMAIN POORLY UNDERSTOOD.

While there is much crime data, good data on the interaction of crime, race, geography, and economics are hard to find. I was finally able to combine state and local incarceration data from the 2010 Census with population data from the Census, economic data from Remapping Debate sponsored by the Anti-Discrimination Center in New York[3], segregation data from the University of Michigan[4], marijuana arrest data from the ACLU, and crime rates from the Bureau of Justice Statistics. Once assembled, I could investigate relationships among these variables.

My goal was to see if I could find an explanation for the variation in the relative incarceration rates of African Americans to Whites across the states. Note that in this analysis I am not trying to explain why incarceration rates for African Americans are higher than for Whites (which they are) but to explain why they are so much higher in some states than in others. Both questions are important and the second should offer some clues to answering the first.

THIRD: EXPLORE THE DATA

How much variation is there across the states in the relative incarceration rates? Using 2010 data, the national average is 6.8 (i.e., the average incarceration rate for African American adult males is 6.8 times as high as for adult White males), ranging from a low of 3 in New Mexico to a high of 16 in Vermont. Figure 1 shows the varying ratio across the states:

[3] http://www.remappingdebate.org/map-data-tool/racial-disparities-median-household-income-remain-enormous-most-states

[4] http://www.psc.isr.umich.edu/dis/census/segregation.html

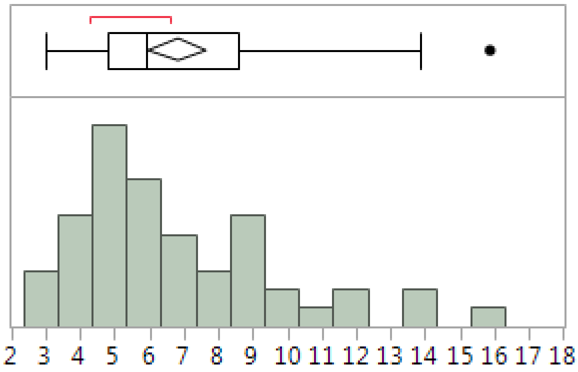
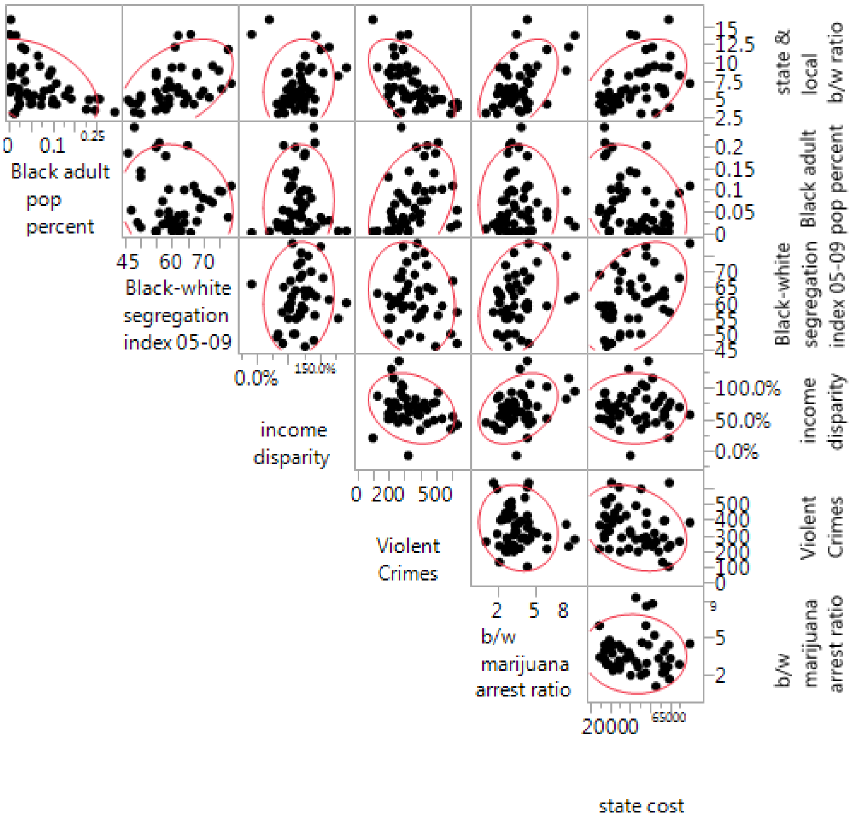


Figure 1: Ratio of African American adult male incarceration rate to White adult male rate

Wisconsin is at a relatively high 11.8, while Iowa stands at 13.7 and Minnesota at 12.1. This certainly suggests that something is going on in the upper Midwest that is not prevalent elsewhere – but what is it? Is it related to economics, demographics, geography, policing? A proper data analysis would examine the relationship with each of these factors and then return to the entire data to see if the patterns are robust or idiosyncratic – in other words, is it Signal or Noise?[5]

Firgure 2 gives a visual image of how the ratio relates to a number of variables:



[5] This is the subject of the best selling book by Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*, Penguin Books, 2015.

Each point is a different state, and the red ellipses encircle 90% of the data in each diagram giving an idea of the strength (or lack thereof) of the relationship. Glancing across the top row, there appears to be inverse relationships between the incarceration ratio and (i) the African American percent of the population in each state, and (ii) the violent crime rate in each state. Positive relationships are apparent between the incarceration ratio and (i) the segregation index,[6] (ii) the ratio of marijuana arrests for African Americans compared with Whites, and (iii) the cost per prisoner of incarceration in each state. Little relationship is evident between income disparity (measured by the ratio of median White household income in each state to median African American household income) and the incarceration ratio.

FOURTH: IT'S MORE COMPLICATED THAN THAT

While these two dimensional pictures can be revealing, they can also be spurious. An apparent relationship (or lack thereof) may disappear (or appear) when a third or fourth variable is taken into account. The most common technique to accomplish this is multiple regression analysis, where the simultaneous effects of all the variables are taken into account. Complicating this analysis is a lack of complete and consistent data across all states. Should we use the total incarceration rate for African Americans and Whites in Federal, State, and local facilities ("jails" generally refers to local facilities), or just in state prisons, or only local jails? Remember that some states do not have Federal prisons at all – and also some states do not have local jails.

States also vary in the way they categorize criminal offenses and their sentencing – and state laws change over time. For example, in 2015, several states (e.g., Connecticut, Maine, North Dakota) reclassified some crimes from felonies to misdemeanors. Generally, local prisons house people convicted of misdemeanors while state and Federal prisons house those with felony convictions. Many states limit local incarceration to those with sentences of 1 year or less, although some states use 2 years. Given these issues with the Federal and local data, my primary measure is the incarceration rate in state and local facilities together, but I looked at all options in case state-level anomalies distort the relationships.

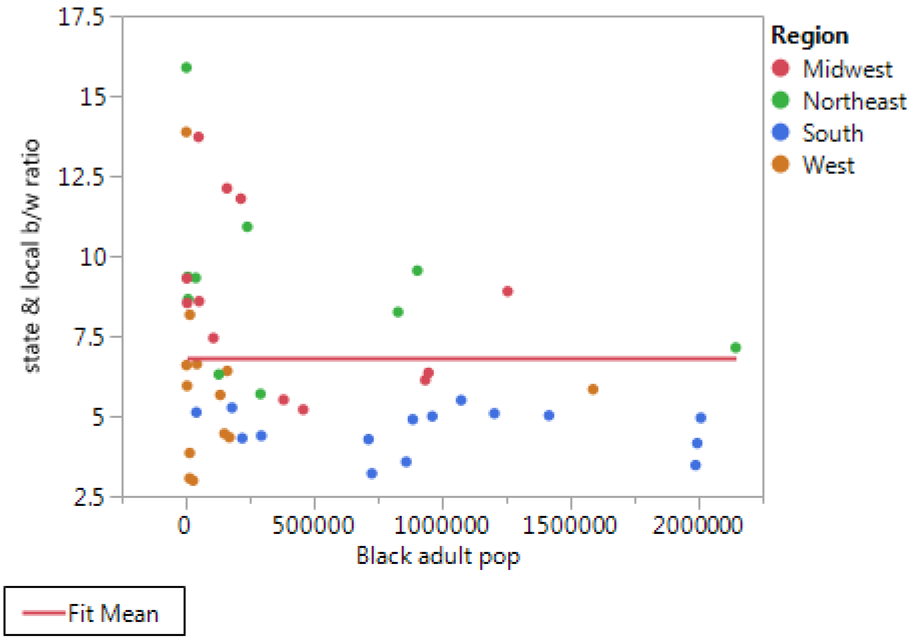
A consistent pattern does emerge: the African American percent of state population and the violent crime rate are significantly and inversely related to the incarceration ratio in all of the models. The ratio of marijuana arrests for African Americans to Whites and the per-prisoner cost of incarceration are significantly and positively related to the incarceration ratio. The segregation index is also positive and statistically significant in most of the models (not for the 2005 data, but it is for all the 2010 models). Income disparity is not statistically significant in any of the models except for the incarceration ratio in local jails (where more income disparity is associated with a higher ratio).

[6] The segregation index was developed by University of Michigan researchers and measures the percent of African American households that would need to move in order to have the geographic distribution of African Americans look similar to that of White households in each state.

FIFTH: BE CAREFUL ABOUT STORIES!

One must be careful not to interpret these associations as cause and effect. For example, consider the strong inverse relationship between the violent crime rate and the ratio. This could be due to relatively high incarceration rates for African Americans causing them to move to other states and resulting in a lower violent crime rate (requiring the assumption that African Americans are relatively more involved in violent crimes than Whites). Or it could mean that where violent crime rates are relatively high, police and courts are focused on preventing and solving such crimes and so spend relatively less energy prosecuting non-violent crimes, such as drug offenses. Indeed, considerable research provides evidence of the latter story.[7]

One must also be careful not to find a causative story in every pattern. This is a point amply made by Daniel Kahneman.[8] It is easier for the ratio of incarceration rates to vary more widely in states with lower African American populations than in those with greater populations – this is simply a function of the sample size and not due to characteristics of those states. Figure 3 shows the ratio of incarceration rates in relation to the sizes of the African American population in each state:



There is much more variability in the ratio for states with small African American populations than with large ones. Still, there are strong regional patterns that are asking for an explanation – the ratio is lowest in the Southern and Western states and highest in the Northeast and Midwest.

[7] See, "Wisconsin's Mass Incarceration of African American Males: Workforce Challenges for 2013," https://www4.uwm.edu/eti/2013/BlackImprisonment.pdf for an example.

[8] Daniel Kahneman, *Thinking Fast and Slow*, Farrar, Straus, and Giroux, 2013. Kahneman's example compared kidney cancer rates, where rates are highest in rural, Republican states, but also lowest in the same states. The difference is due to smaller sample sizes, not to the facts that the states are rural or Republican.

SIXTH: GOOD DATA ANALYSIS REQUIRES GOOD SUBJECT MATTER EXPERTISE

At this point in my analysis I felt that I had identified some factors that are clearly related to the incarceration ratios across the states. When more African Americans are arrested for marijuana offenses relative to Whites, the state exhibits relatively higher incarceration rates for African Americans. So, the police and court systems are part of the story (but cause and effect are still unclear). States with lower African American populations that are more segregated are associated with larger incarceration ratios. Where violent crime rates are higher, the incarceration ratio of African Americans relative to Whites is lower. But more questions kept emerging.

I was somewhat surprised that the income disparity was not statistically significant. Perhaps the ratio of median incomes is not the relevant factor. Might relative unemployment rates be more associated with relative incarceration rates? Perhaps it is youth unemployment that matters? What is the best way to measure relative economic opportunities for African Americans and Whites?

I made several inquiries to the Bureau of Justice Statistics (BJS) to better understand the data. They confirmed that 2005 was the most recent study they had published about the racial incarceration rates across the states (this is itself politically relevant: it reflects priorities, and if we don't have good data on something, it is easier to deny existence of the problem). I was also told of further complicating factors in any analysis. States vary considerably in the way they report the race of prisoners – some states use self-reporting (by the prisoner), others use visual recording at intake, while others use whatever was reported on the RAP sheet. The BJS considers the state race data to be unreliable.

Some states (e.g., New York, New Jersey, Michigan, Kansas) made substantial cuts in their prison populations over the decade 1999-2009. What policies did they change and how does this affect the data on incarceration rates? To really understand this data I need a better understanding of our legal and penal system – it is not just a matter of analysis, it requires subject matter expertise.

These are the kinds of questions that keep a data analyst awake at night. There is no end to the set of variables that might be relevant, nor do questions about data integrity and relevance ever cease.

SEVENTH: DON'T REJECT ANALYSIS

Some people become skeptical of any data analysis as a result – I am reminded of Nobel Prize winner Ronald Coase's quote that "If you torture the data long enough, it will confess." But I think that conclusion is unfounded. An effective counterargument has been offered by the famous statistical educator Frederick Mosteller, "it is easy to lie with statistics, but easier without them."[9]

Indeed, data analysis can be misused and done poorly. But it can also be done well and honestly. Part of being a data analyst is knowing when you have been thorough enough, knowing when you are cherry picking results to tell the story you want to tell (when the data do not really support it), and knowing

[9] Quoted in Gelman and Loken, 2013, "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time," unpublished manuscript.

how to *not* overstate your findings. Understanding proceeds incrementally and slowly – and this is at odds with our ever-faster world of tweets and Instagram posts. Data analysts need to have a moral compass. And so, too, do politicians. Finding solutions to the horrifying incarceration rates will require a deeper understanding than we've seen from the Democrats (who acknowledged the problem but avoided digging deep enough to know why Wisconsin was so extreme or exactly how extreme it was or wasn't) or the Republicans (who have not yet raised the issue at all).

After all this analysis, the most puzzling feature to me is the strong regional pattern showing much higher incarceration ratios in the Midwest and Northeast and relatively low ratios in the South and West. Indeed, the ratio is very high in North and South Dakota, Minnesota, Iowa, Wisconsin, Illinois, and Iowa. Perhaps policing practices are different in these states, but I would find that surprising. It might be related to historical migration and settlement patterns. Or it might be that incarceration is only part of a larger picture – criminal offenses can result in prison sentences, fines, home detention, other institutionalized settings, or combinations of these. Does the disparate treatment of African Americans and Whites really differ that much across the states or would a more holistic view find that states are really similar, but rely on different enforcement mechanisms?

I suppose there is more to keep me awake at night.