County-level Discrepancies between SF-1 and DHC across the Rural-Urban Continuum

J. Tom Mueller

Research Assistant Professor

Department of Geography and Environmental Sustainability

University of Oklahoma

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Background

- Rural America is in the midst of a long pattern of population aging, chronic outmigration, and economic difficulties
- The decade from 2010 to 2020 was the first decade where there was net depopulation of rural counties
- Rural America is also prone to significant rates of poverty, hardship, and poor health outcomes

Background

- Although these trends are well-known, it has always been hard to document them
  - Data suppression has been an issue for rural demography for a long time
  - Many data sets are simply unusable for rural areas
    - Rural census tracts in the American Community Survey have massive margins of error
  - Other data sets simply don't have enough rural respondents to allow for rural insights
- The decennial census, particularly at the county level, has long-been the gold standard for rural demography

Background

- Rural demography generally happens at the county level
- Rural areas can be though of as being defined by low population density *and* a lack of connectivity to urban areas
  - The county is a valuable scale for determining connectivity
- Data is often only reliable at the county level
- Counties are political units, so policy implications can more easily be assessed
- Thus, we often treat nonmetropolitan counties as synonymous with rural
  - As defined by OMB where a metro county is any county with an urban population of greater than 50,000 or is connected to a core metro county by at least 25% of commuting

Goal and Methods

- My goal here is to display discrepancies between the traditional method of disclosure avoidance via the 2010 SF-1…
  - Data swapping, suppression, and top and bottom coding
- And the new differential privacy approach via TopDown
- I will be using the person-level DHC files from the March 16th release of demonstration data, retrieved via IPUMS-NHGIS
  - Privacy loss budget of 20.82

Goal and Methods

- I will be making comparisons along the rural-urban continuum via the nine-category Rural Urban Continuum Codes
  - 2013 vintage determined via the 2010 census
- Metro
  - 1 – 1 million or more
  - 2 – 250,000 to 1 million
  - 3 – Less than 250,000
- Nonmetro
  - 4 – Adjacent to metro, 20,000 or more
  - 5 – Not adjacent, 20,000 or more
  - 6 – Adjacent, 2,500 to 19,999
  - 7 – Not adjacent, 2,500 to 19,999
  - 8 – Adjacent, less than 2,500
  - 9 – Not Adjacent, less than 2500

Goal and Methods

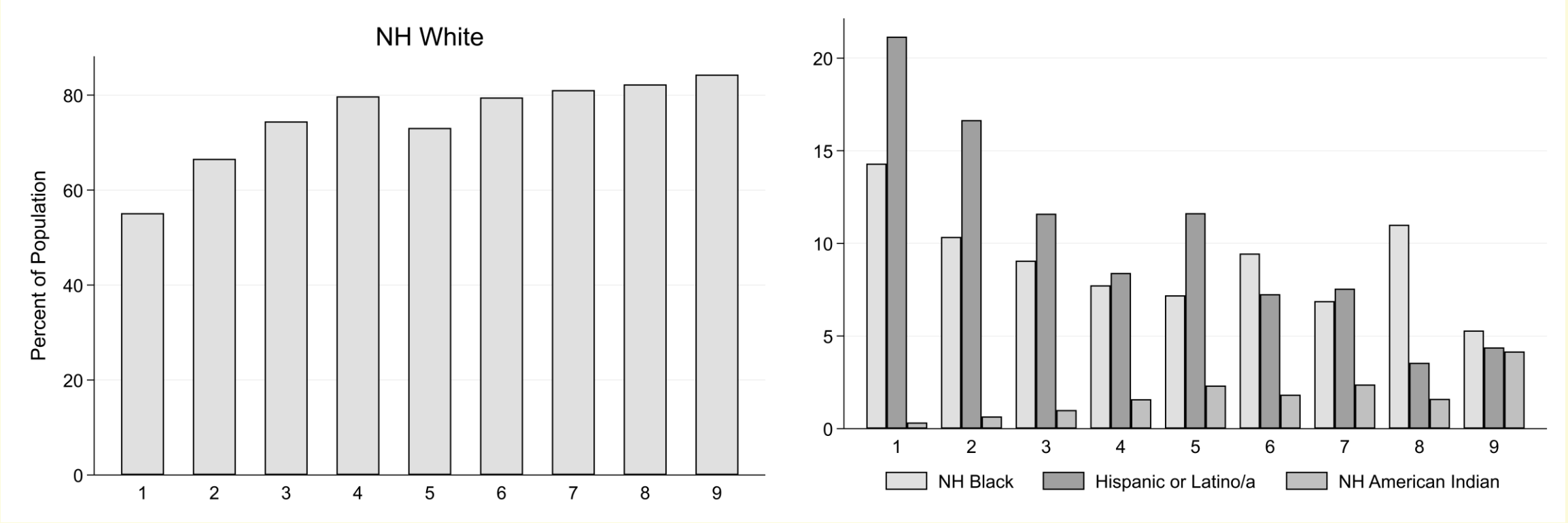
- I will first compare total population statistics, then median age statistics

- In all cases I will make comparisons across race and ethnicity
  - Non-Hispanic white
  - Non-Hispanic Black
  - Hispanic or Latino/a
  - Non-Hispanic American Indian

A Note on Race and Ethnicity

- Rural America is not monolithic
- Although many imagine it as being entirely white, this isn't true, hasn't ever been true, and is becoming less so every year
- Most importantly, rural non-white residents of the U.S. live in some of the most difficult and unequal conditions
  - High mortality rates, limited mobility, long-term neglect from local governments

A Note on Race and Ethnicity



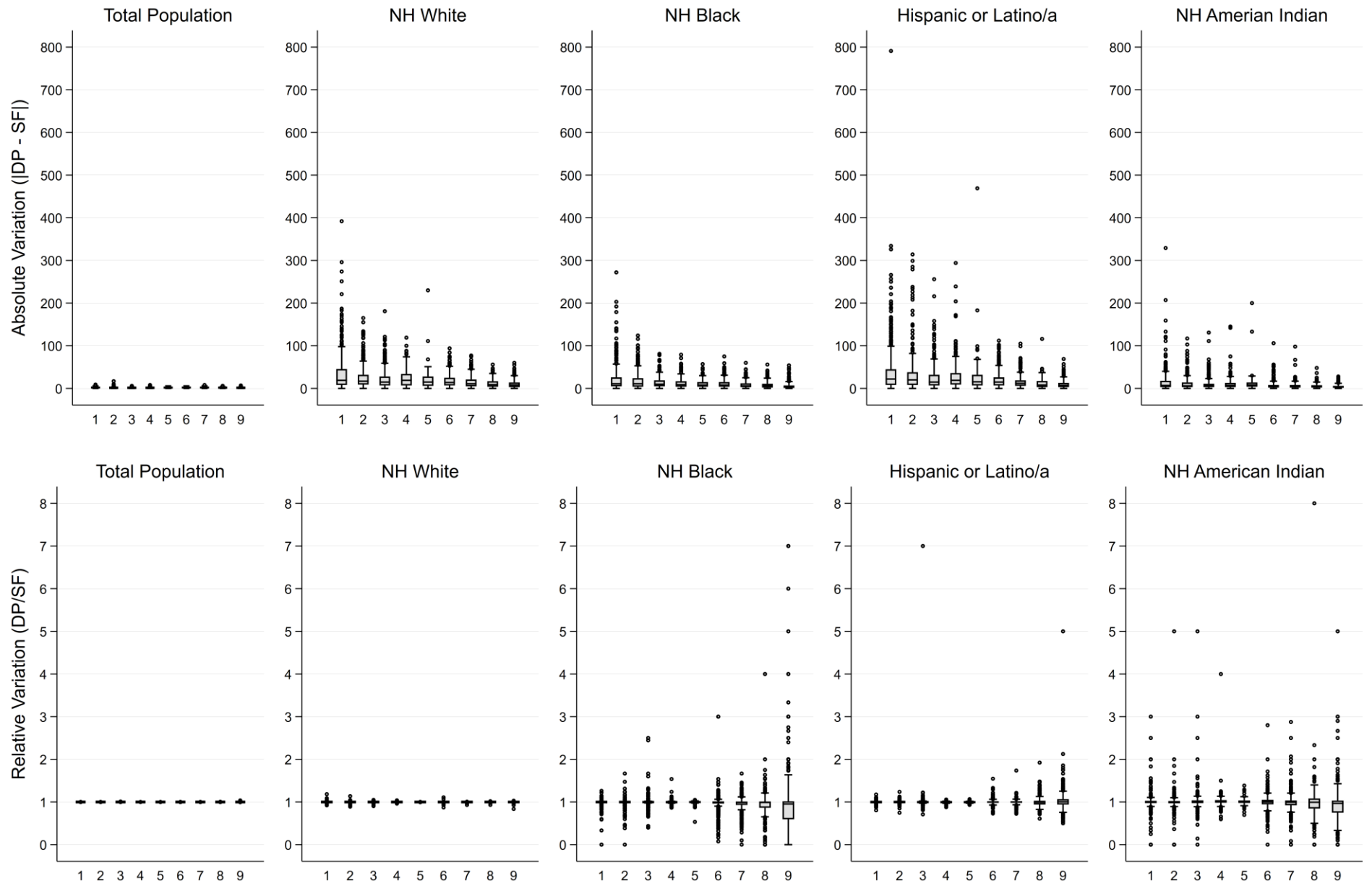
Percent of population in each group across 2013 RUCC Categories as reported by the 2010 SF-1.

A Note on Race and Ethnicity

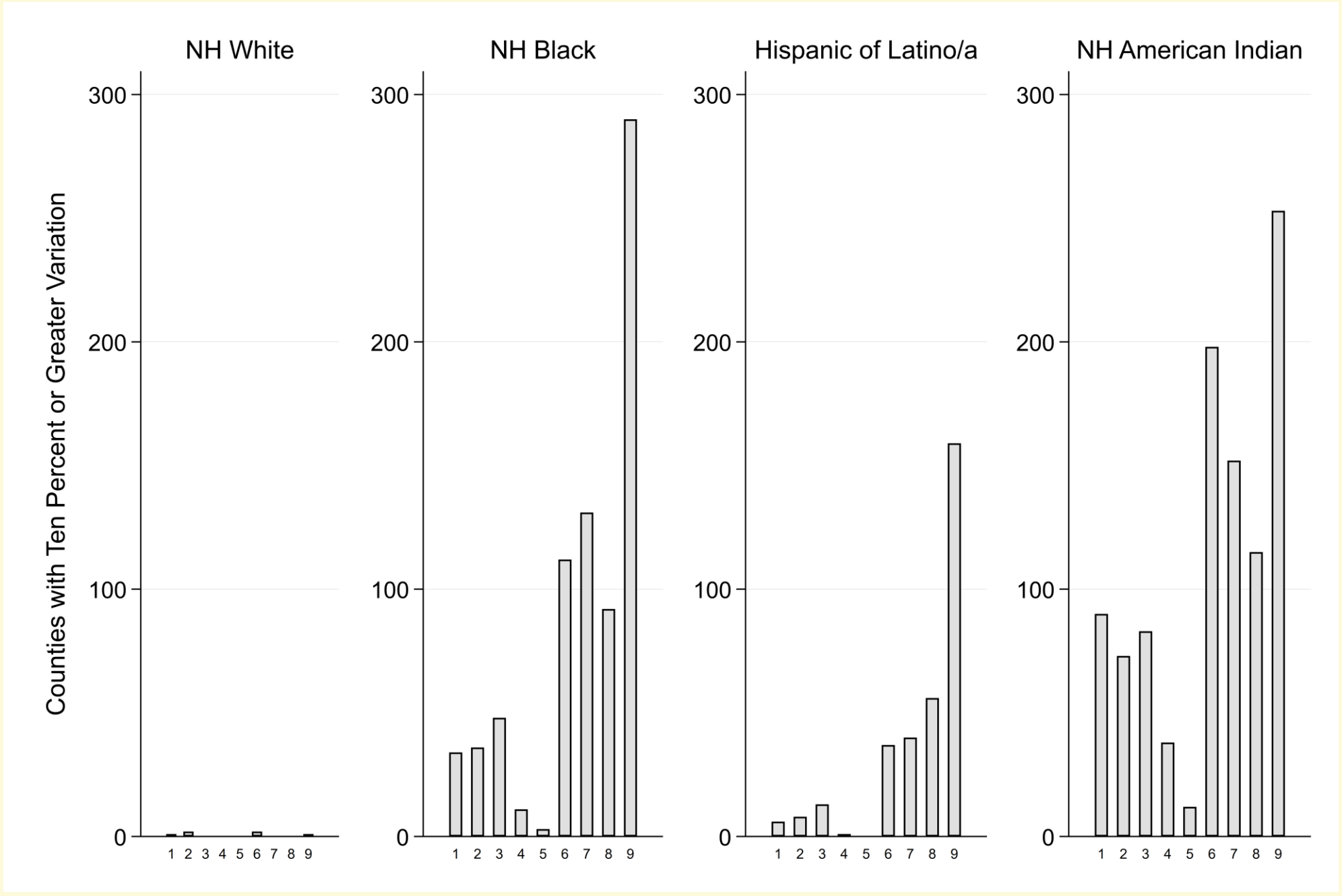
- Non-white populations are highly concentrated in specific areas throughout the United States

- Even in the most rural counties (RUCC=9), there are still counties with…
  - 72% NH Black
  - 92% Hispanic or Latino/a
  - 94% NH American Indian

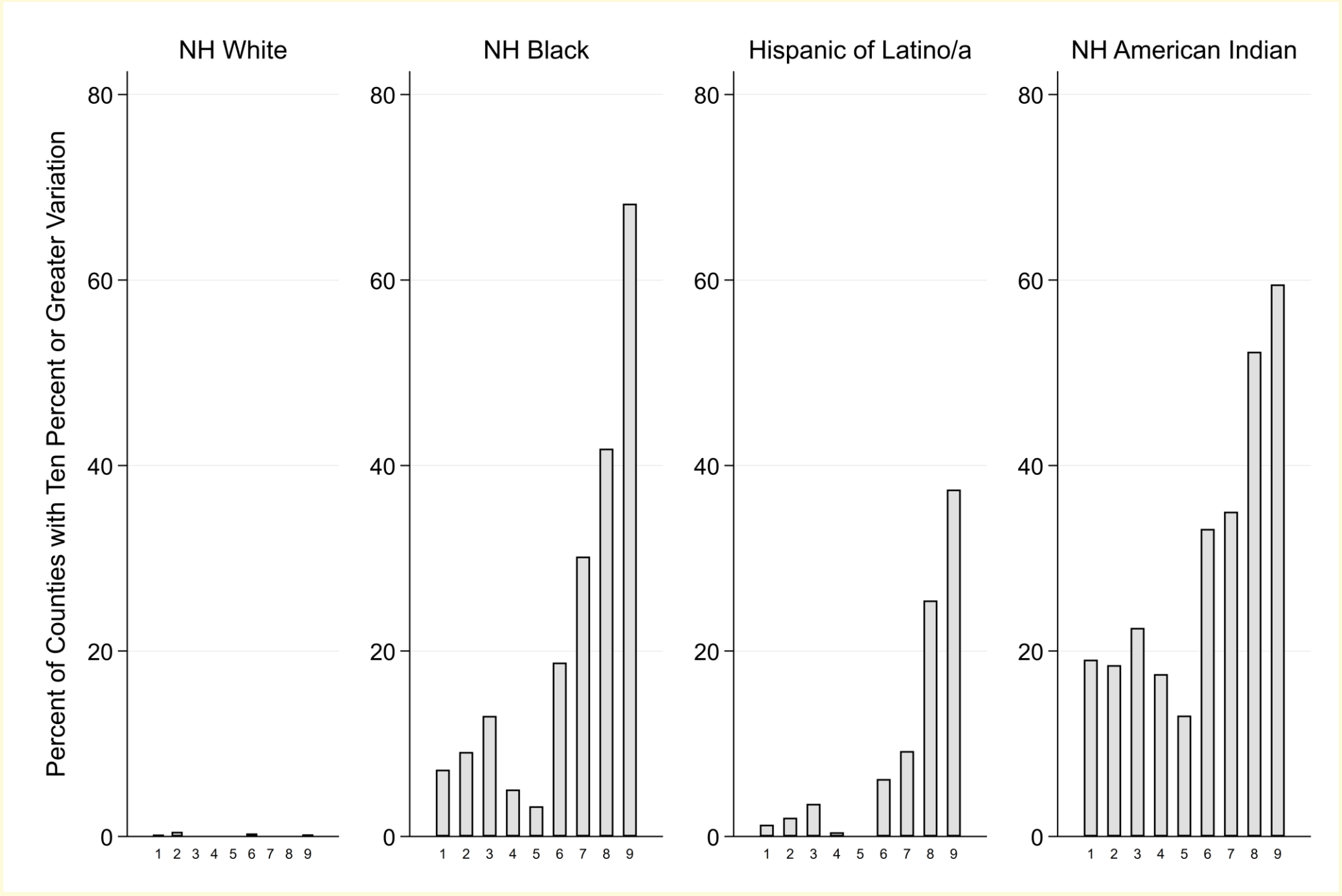
- Thus, the discrepancies I am about to show would impact many people and communities across rural America, many of whom are in difficult situations already



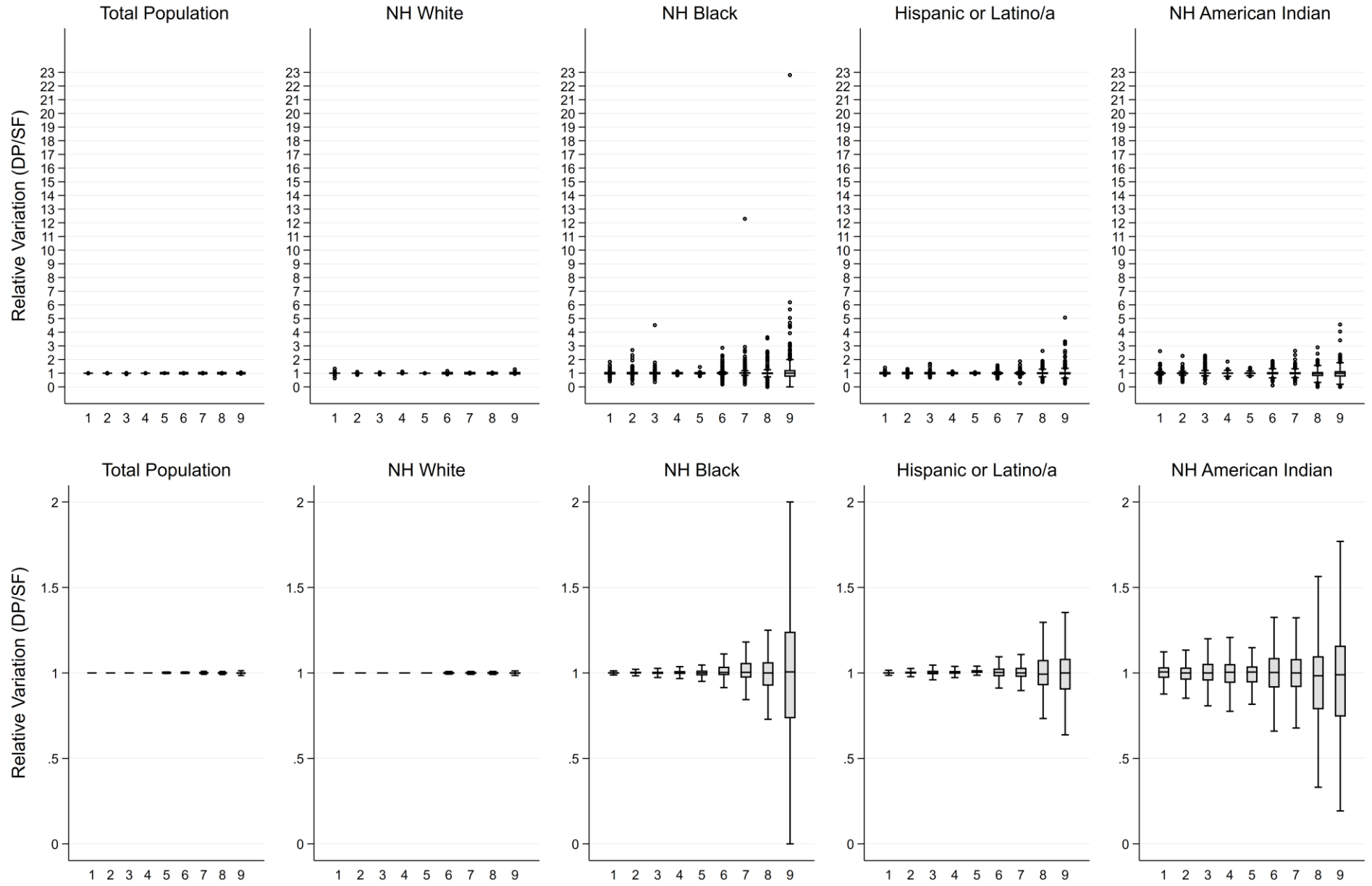
Absolute and relative differences in county-level population counts between 2010 DHC demonstration and SF-1 data across 2013 RUCC Categories



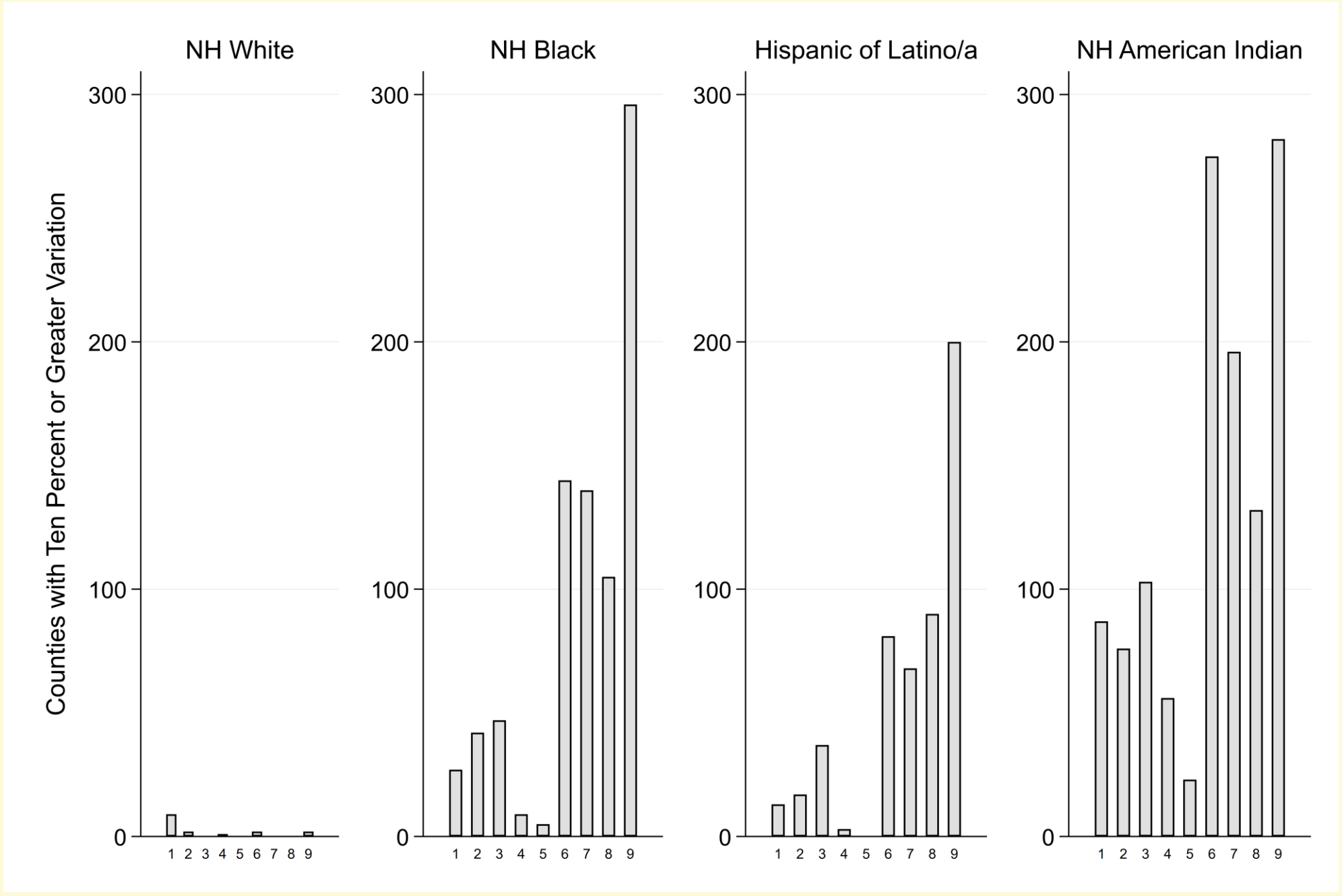
Number of counties in each 2013 RUCC group with a discrepancy between population counts in 2010 DHC and SF-1 greater than 10%



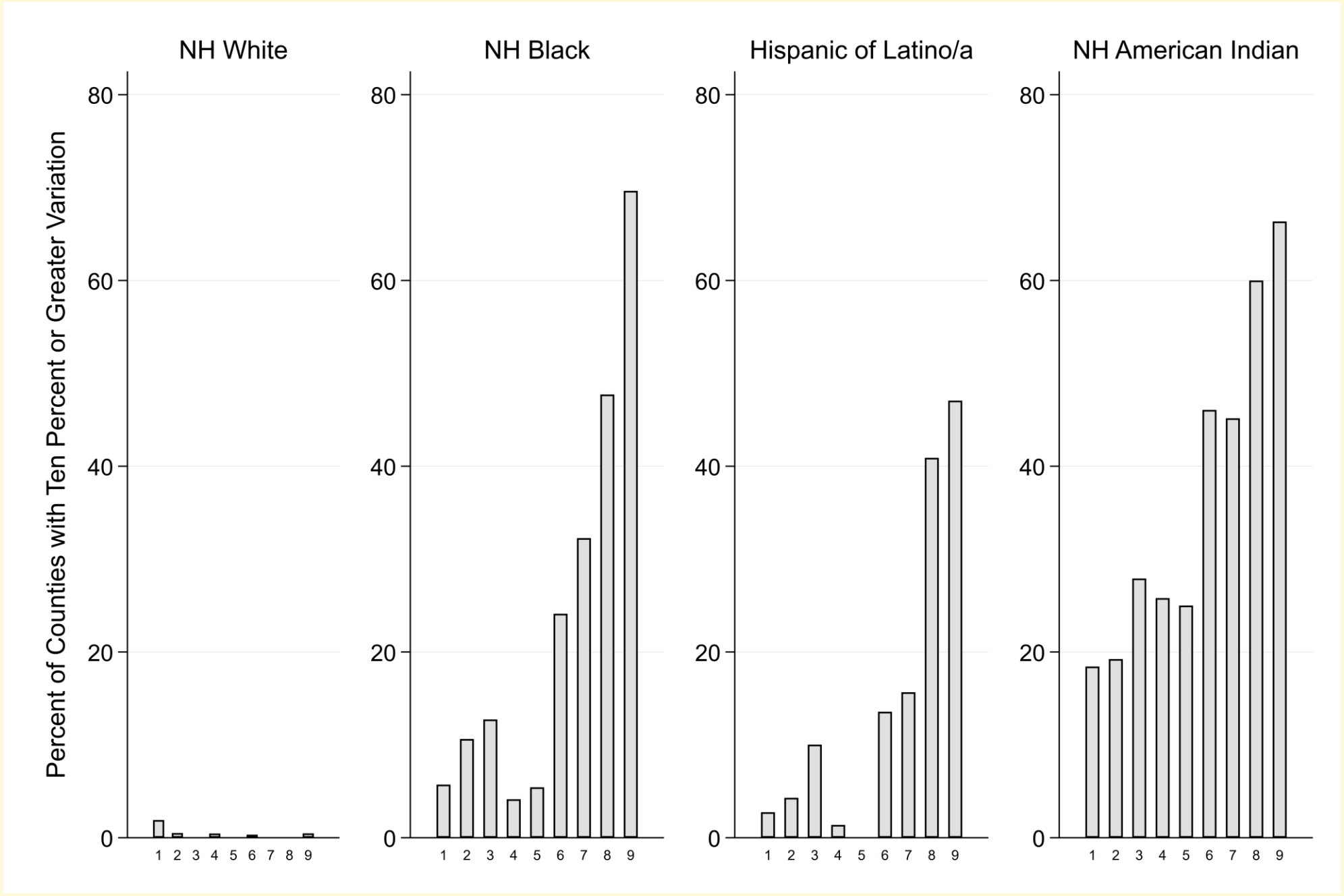
Number of counties in each 2013 RUCC group with a discrepancy between population counts in 2010 DHC and SF-1 greater than 10%



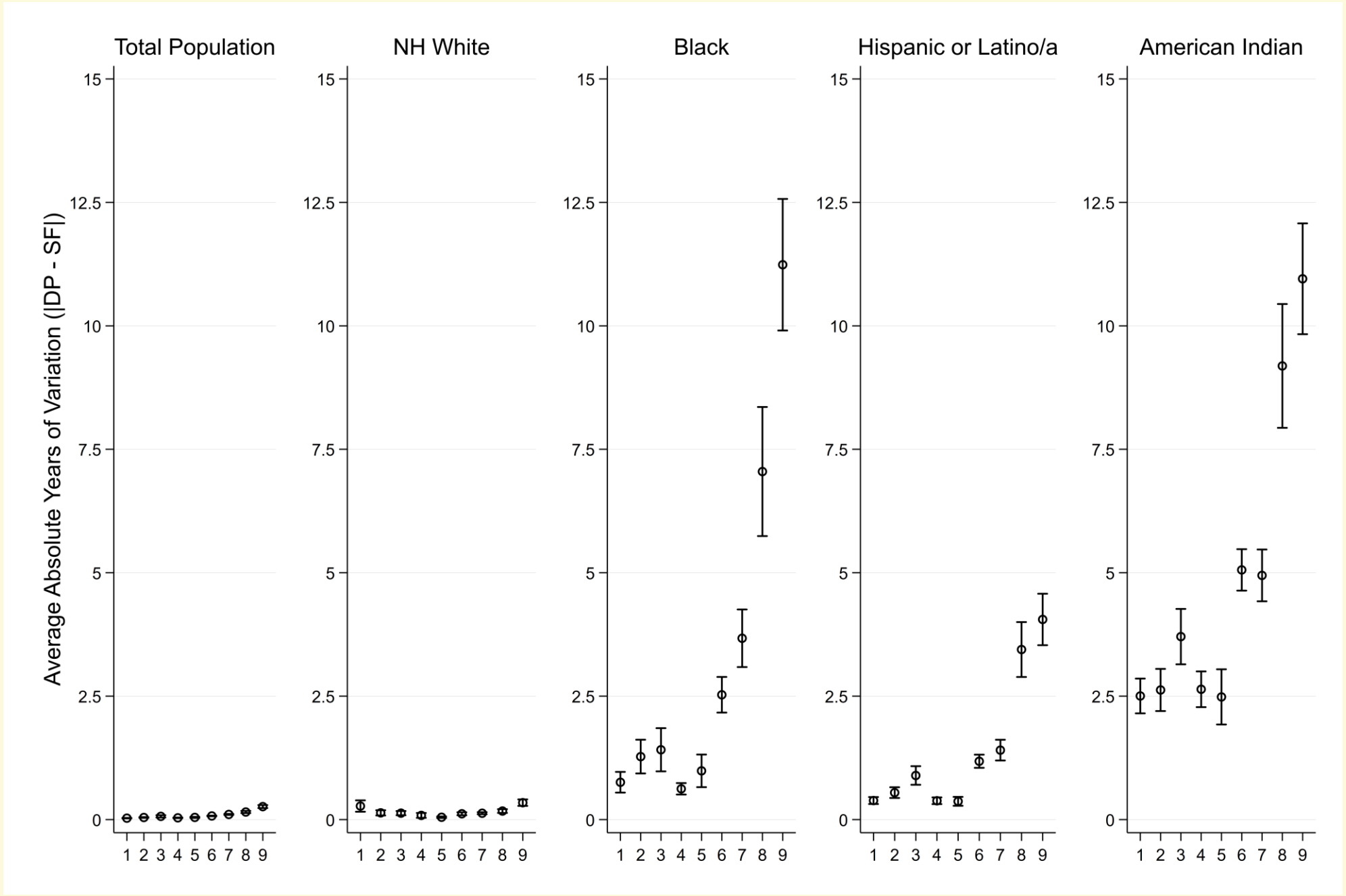
Absolute and relative differences in county-level median age between 2010 DHC demonstration and SF-1 data across 2013 RUCC Categories



Number of counties in each 2013 RUCC group with a discrepancy between median age statistics in 2010 DHC and SF-1 greater than 10%



Percent of counties in each 2013 RUCC group with a discrepancy between median age statistics in 2010 DHC and SF-1 greater than 10%



Average county-level discrepancy between median age between 2010 DHC and SF-1 in years by 2013 RUCC

Conclusion

- To be clear, most of the ratio averages remain quite close to 1.0
  - But this isn't really how demographers and other users use this data.
- State demographers rely on these estimates to know specific things about specific rural counties
- Further, population counts are used by agencies and policies to determine aid and other resources
  - These discrepancies will *certainly* impact our upcoming poverty rates
- The discrepancies I illustrate will cause rate estimates to become very inaccurate for many counties

Conclusion

- The issues I have shown will exist for all small-n population groups
  - However, in the US that basically assures that white and urban populations get to have accurate data and non-white and rural populations do not
- The picture is likely even worse for any intersectional subgroups
- I am concerned about the ability of the academic and practitioner communities to do their jobs if this method is used for 2020 data

Conclusion

- It is not clear to the rural demography community how differential privacy makes sense when accuracy for small groups is important
  - This is the difference between big data and large-n data
  - Prior methods seemingly did a much better job of preserving headcount totals
- Small groups and places have just as much of a right to accuracy and representation as the large groups and places
- I know many have concerns about reidentification, but if our 2020 data is released with these discrepancies, I am not sure I will be able to use it or recommend others use it.
  - Sadly, we don't have an alternative source of data for rural areas in the United States

Thank you

Email: tom.mueller@ou.edu

Website: JTomMueller.com