Differential Privacy for the 2020 Census

David Van Riper

vanriper@umn.edu

Twin Cities Research Group Webinar
September 16, 2020

Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau

Fri Aug 17 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Protecting the Confidentiality of America's Statistics: Ensuring Confidentiality and Fitness-for-Use

Tue Sep 04 2018

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY

Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

Fri Feb 15 2019

WRITTEN BY: DR. RON JARMIN, DEPUTY DIRECTOR AND COO

Census Bureau Continues to Boost Data Safeguards

Tue Jul 30 2019

WRITTEN BY: RON JARMIN, DEPUTY DIRECTOR, US CENSUS BUREAU

SUBSCRIBE 

Disclosure Avoidance and the 2018 Census Test: Release of the Source Code

Thu Jun 06 2019

WRITTEN BY: DR. JOHN M. ABOWD, CHIEF SCIENTIST AND ASSOCIATE DIRECTOR FOR RESEARCH AND METHODOLOGY AND DR. SIMSON L. GARFINKEL, SENIOR COMPUTER SCIENTIST FOR CONFIDENTIALITY AND DATA ACCESS

2010 Demonstration Data Products

To help data users understand how differential privacy may or may not impact data products they are used to receiving, the Census Bureau created demonstration data products for review. This set of data products demonstrate the current computational capabilities of the 2020 Disclosure Avoidance System (DAS). The products include the 2010 Demonstration Public Law 94-171 (P.L. 94-171) Redistricting Data Summary File and the Demonstration Demographic and Housing Characteristics Summary File.

DAS Development Update 2020-05-27


Detailed Summary Metrics [<1.0 MB]


Data Metrics Overview [1.0 MB]


IPUMS NHGIS Privacy-Protected Census Demonstration Data


Download: 2010 Demonstration Privacy-Protected Microdata Files (PPMF)

Outline

- How is differential privacy implemented?
- What was the impact on 2010 data?
- What are the next steps?

HOW IS DIFFERENTIAL PRIVACY IMPLEMENTED?

“True” microdata

Sex	School
Male	Never
Male	Never
Male	Never
Male	Attending
Male	Attending
⋮	
Male	Attending
Male	Past
⋮	
Male	Past

x12

x33

Sex	School
Female	Never
⋮	
Female	Never
Female	Attending
⋮	
Female	Attending
Female	Past
⋮	
Female	Past

x4

x17

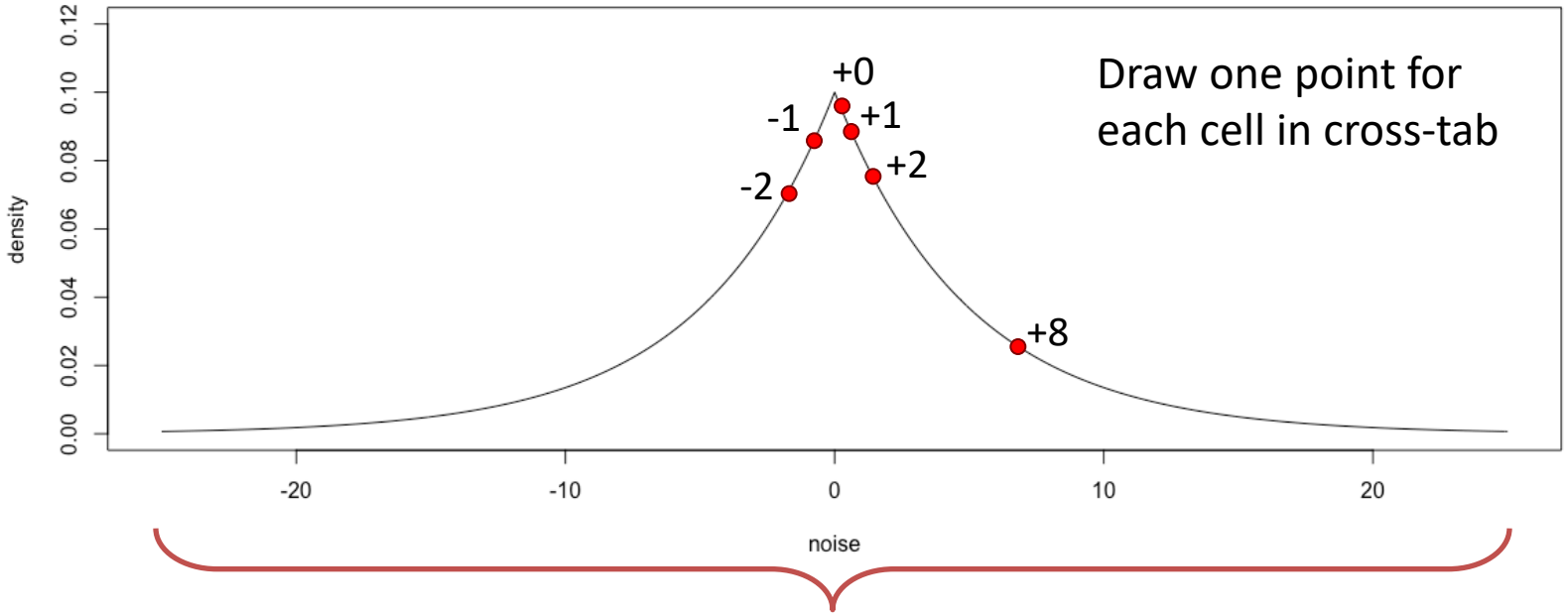
x31

Construct cross-tabs from “true” data

	School Attendance		
	Never	Attending	Past
Male	3	12	33
Female	4	17	31

Population = 100

Draw noise from Laplace distribution

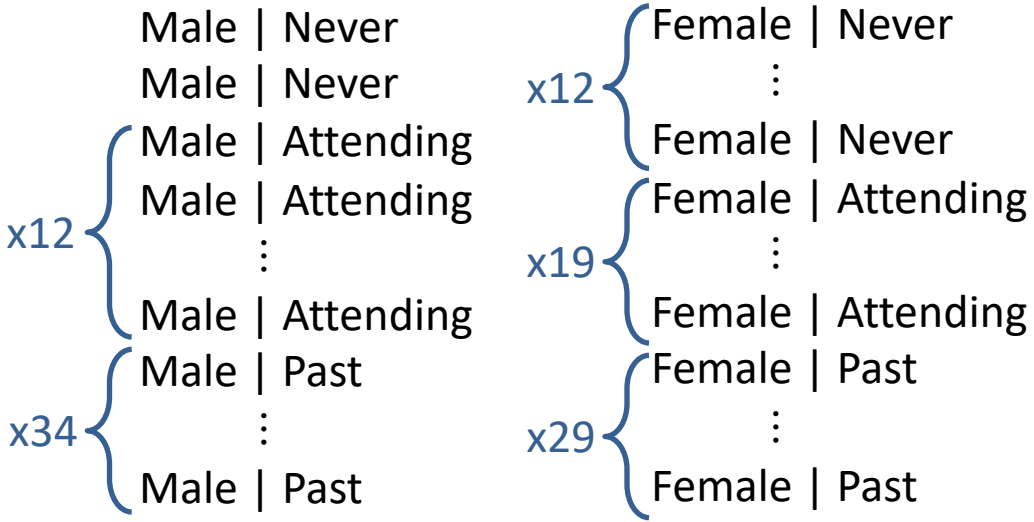


Add noise to cross-tab

	School Attendance		
	Never	Attending	Past
Male	3 – 1 = **2**	12 + 0 = **12**	33 + 1 = **34**
Female	4 + 8 = **12**	17 + 2 = **19**	31 – 2 = **29**

Sum = 108

Construct diff. private microdata



POLICY DECISIONS

Policy decisions

- Global privacy loss budget (**ε**)
- Fractional allocations
- Invariants and constraints

Policy decisions

- Global privacy loss budget (**ε**)
- Fractional allocations
- Invariants and constraints
- **Post-processing**

Global privacy loss budget

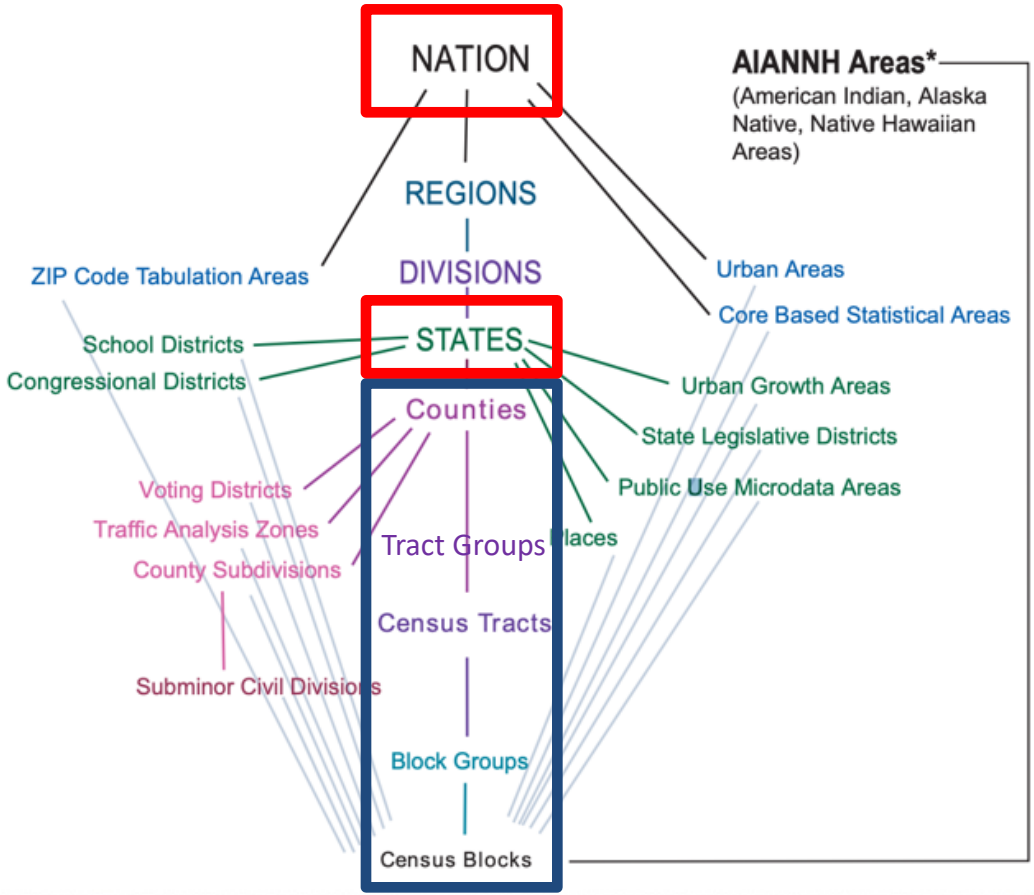
- Global privacy loss budget
  - $\varepsilon = 6.0$
- Person tables
  - $\varepsilon = 4.0$
- Housing tables
  - $\varepsilon = 2.0$

Fractional allocations

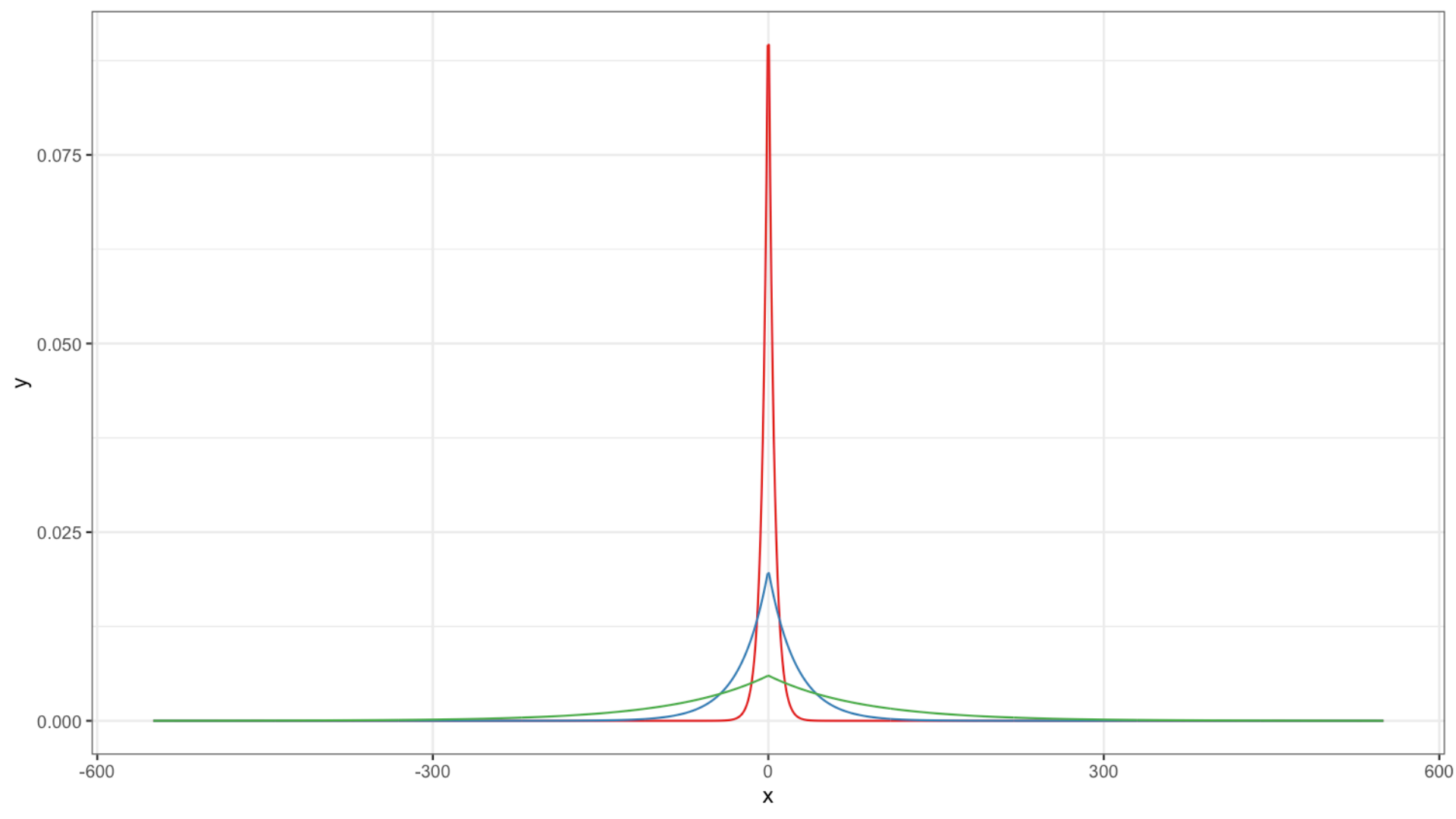
- Geographic levels
- Queries

20% each

12% each



Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Household – Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5


0.075
0.050
0.025
0.000
y
-600
-300
0
300
600
x

Invariants and Constraints

- Invariants are counts not subject to noise injection

2010 Decennial Invariants	2010 Demonstration Data Invariants
Total population (block)	Total population (state)
Total housing units (block)	Total housing units (block)
Group quarters count (block)	Group quarters count (block)
Group quarters type count (block)	Group quarters type count (block)
Occupancy status (block)	
Voting age population (block)	

2010 Decennial Invariants	2010 Demonstration Data Invariants
Total population (block)	Total population (state)
Total housing units (block)	Total housing units (block)
Group quarters count (block)	Group quarters count (block)
Group quarters type count (block)	Group quarters type count (block)
Occupancy status (block)	
Voting age population (block)	

Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints

Invariants and Constraints

- Invariants are counts not subject to noise injection
- Constraints
  - Non-negativity
  - Consistency

Post-processing

- Non-negative least squares + constraints = positive bias for small counts and negative bias for large counts

ANALYZING DIFFERENTIALLY PRIVATE 2010 CENSUS DATA

Data

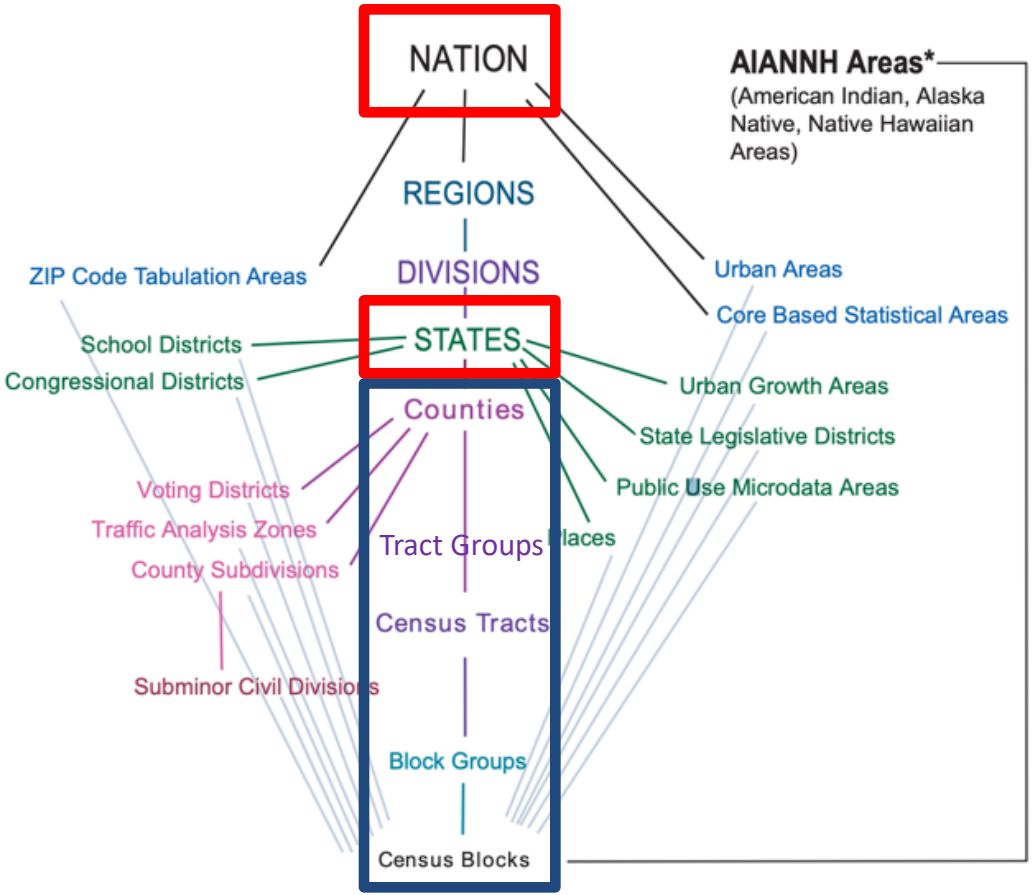
- 2010 Summary File 1
- Vintage 1 (October 2019)
- Vintage 2 (June 2020)

Comparisons

- Comparing data from vintage 1 and 2 with data from Summary File 1
- Summary File 1 essentially serves as our "ground truth"
  - Acknowledging that prior disclosure avoidance techniques introduced error into SF1

20% each

12% each



Vintage 1

Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Relation to HH/Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5

Vintage 2

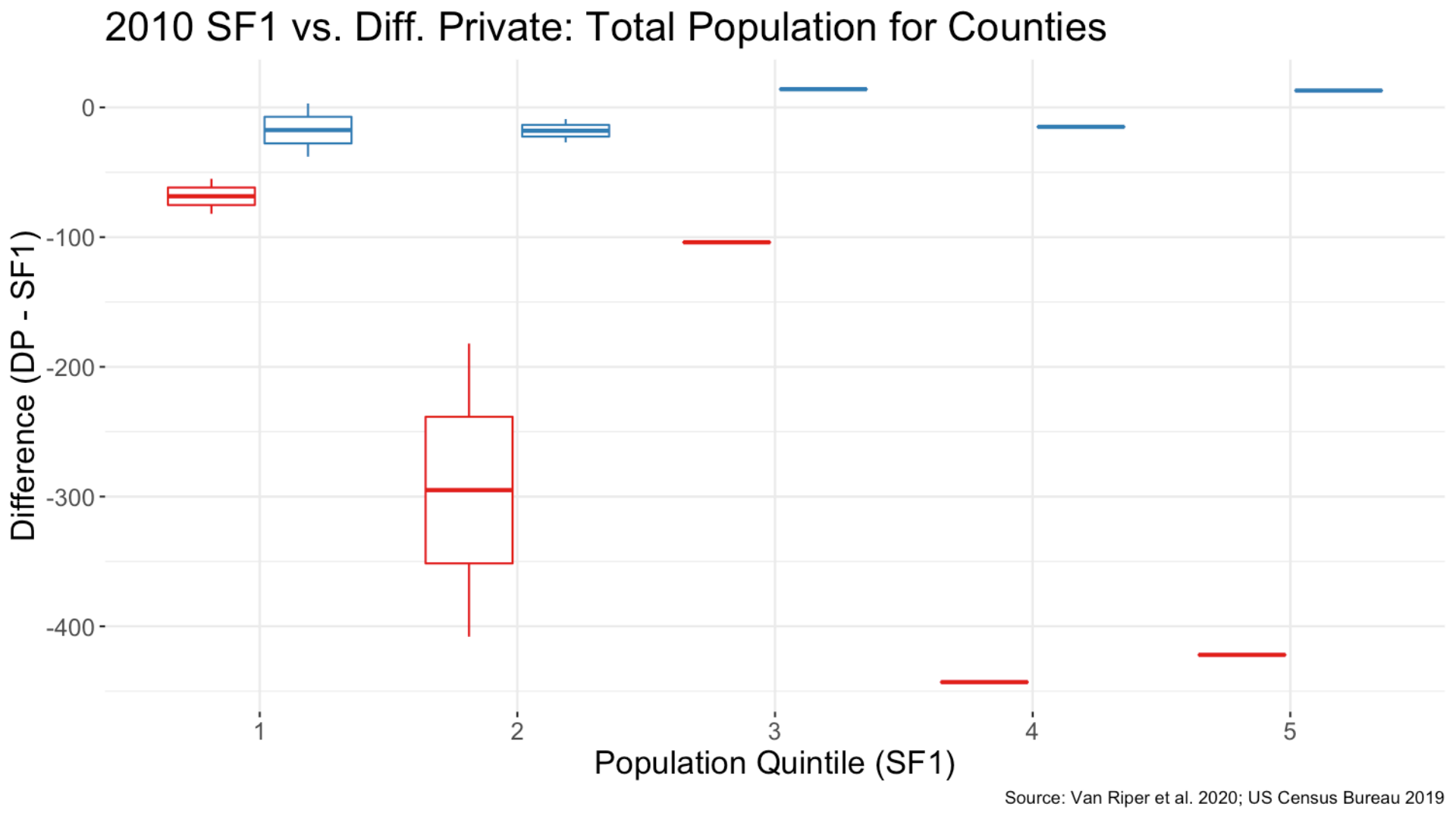
Query	Allocation (%)
Total population	30
Voting age * Hispanic * Race	29
Age * Sex * Hispanic * Race	25
Relation to HH/Group quarters	15
Detailed	1

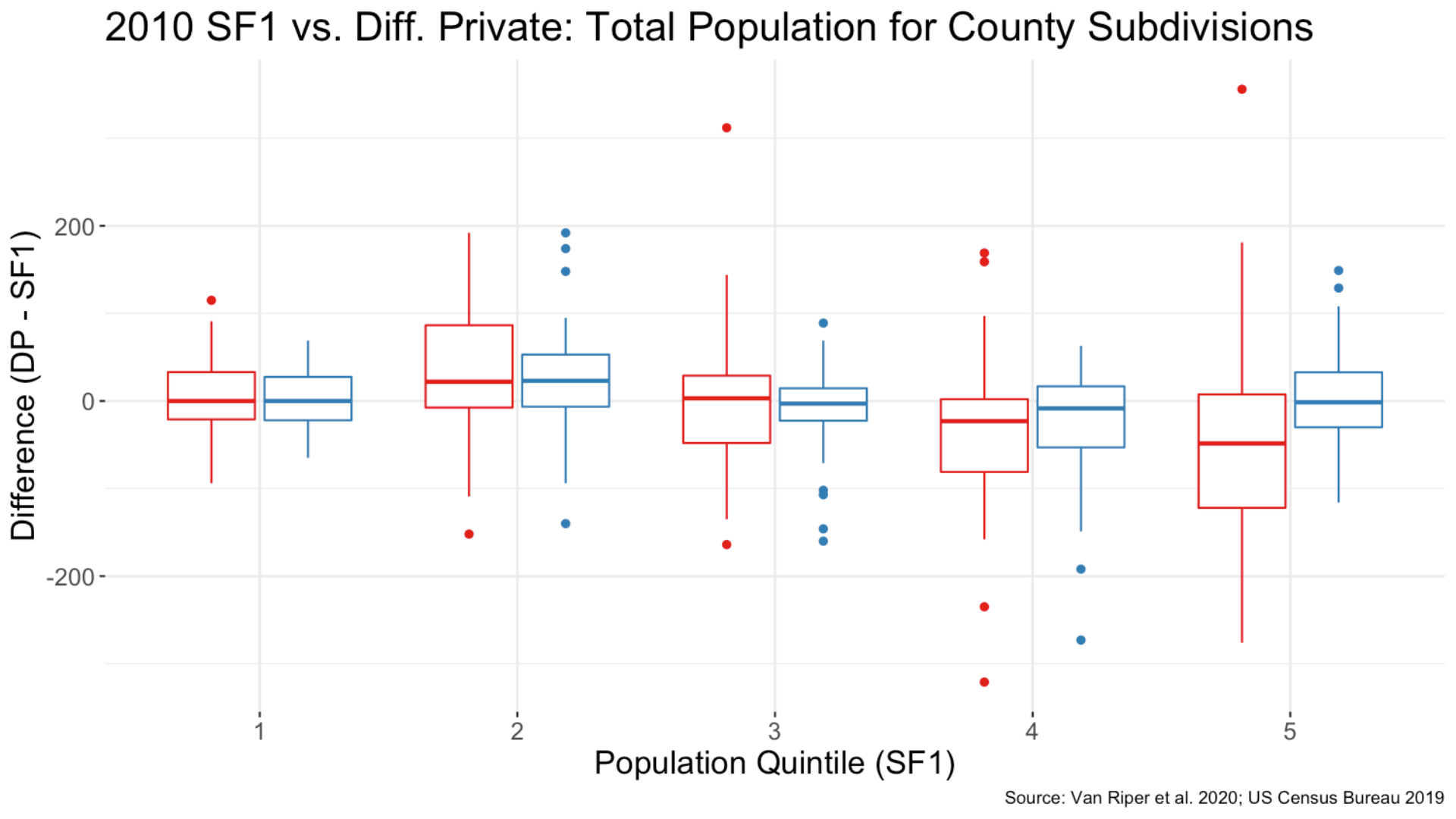
Vintage 1

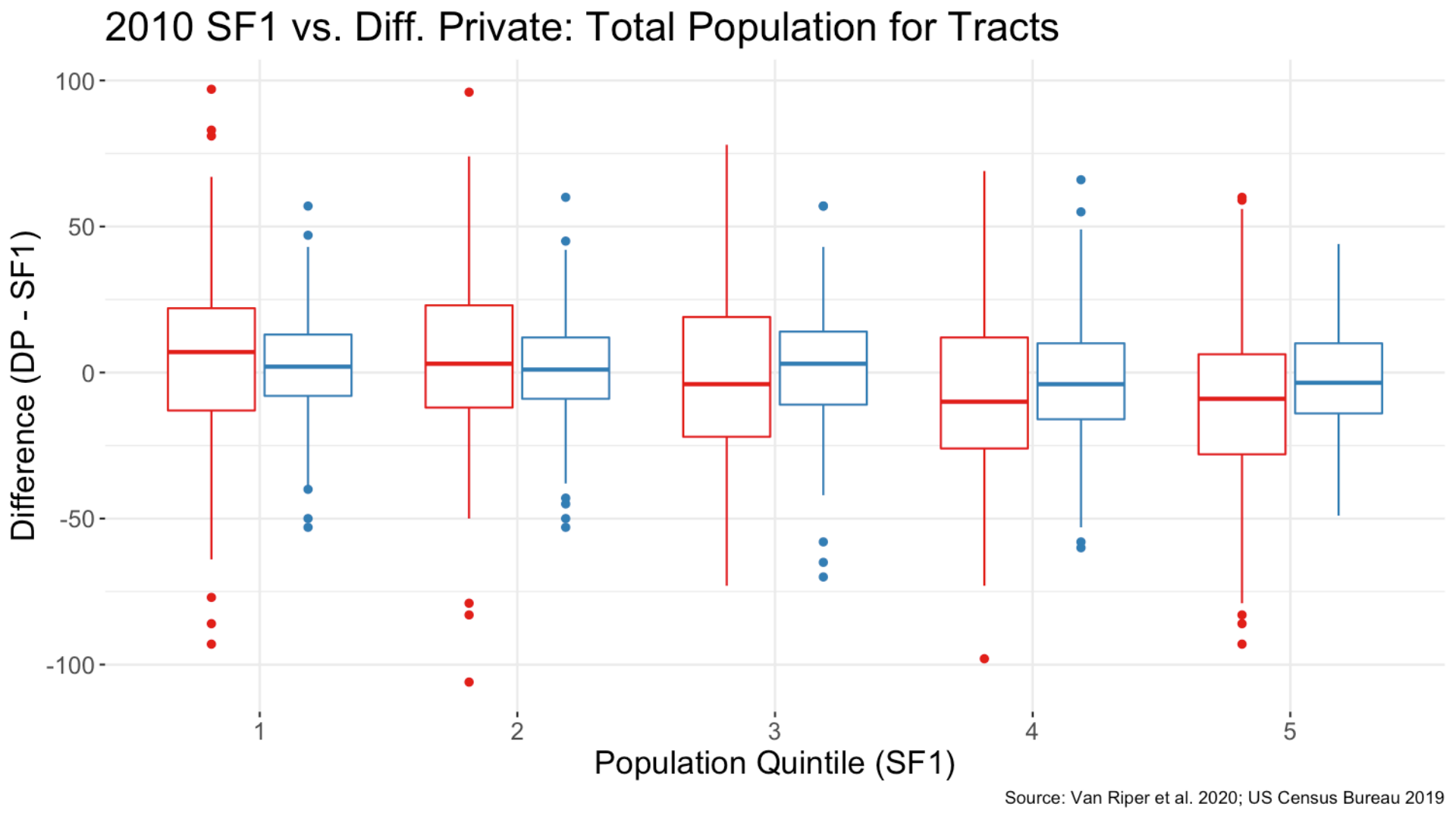
Query	Allocation (%)
Voting age * Hispanic * Race * Citizen	50
Relation to HH/Group quarters	20
Detailed	10
Sex * Age (single year of age)	5
Sex * Age (4-year age bins)	5
Sex * Age (16-year age bins)	5
Sex * Age (64-year age bins)	5

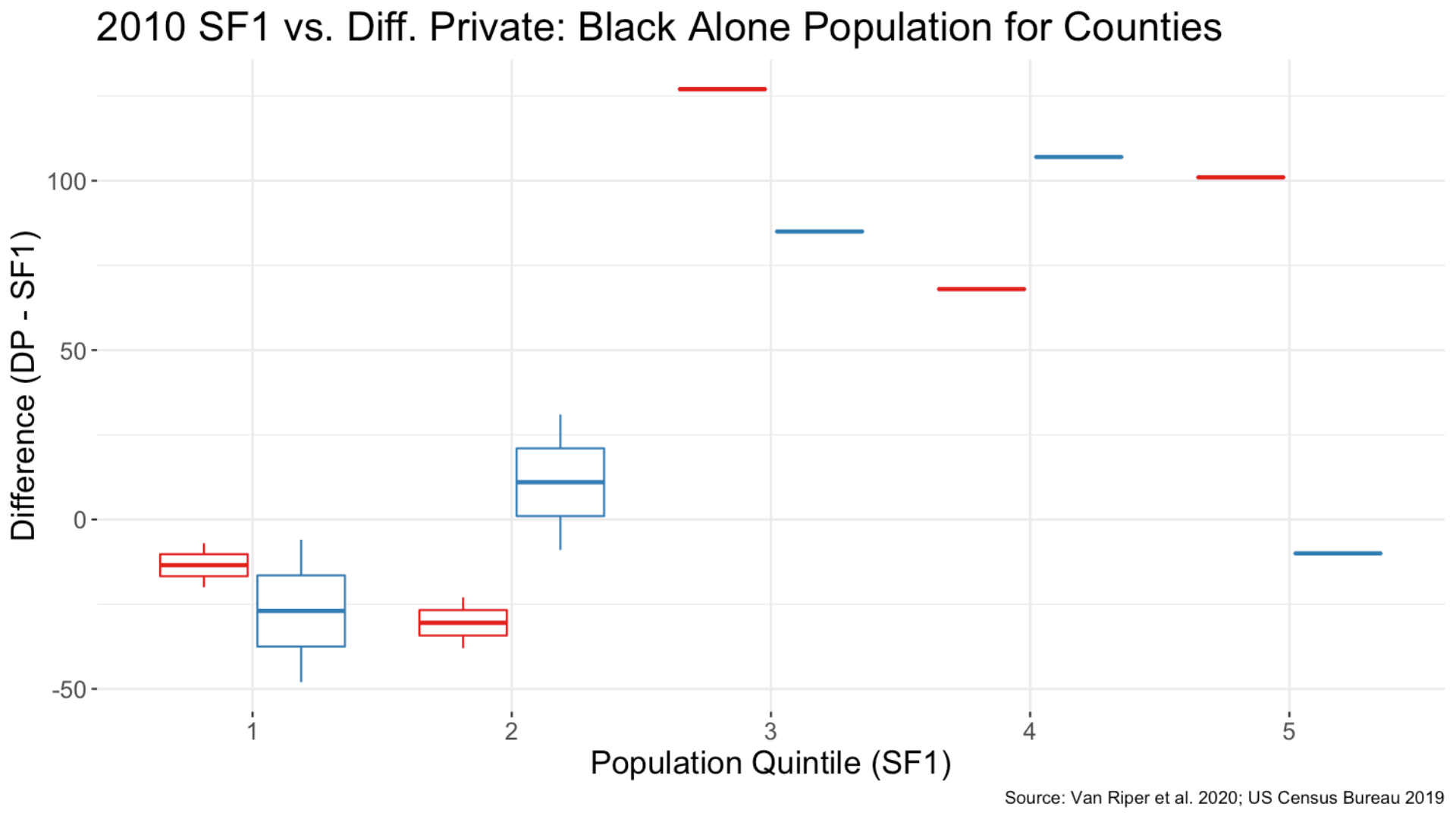
Vintage 2

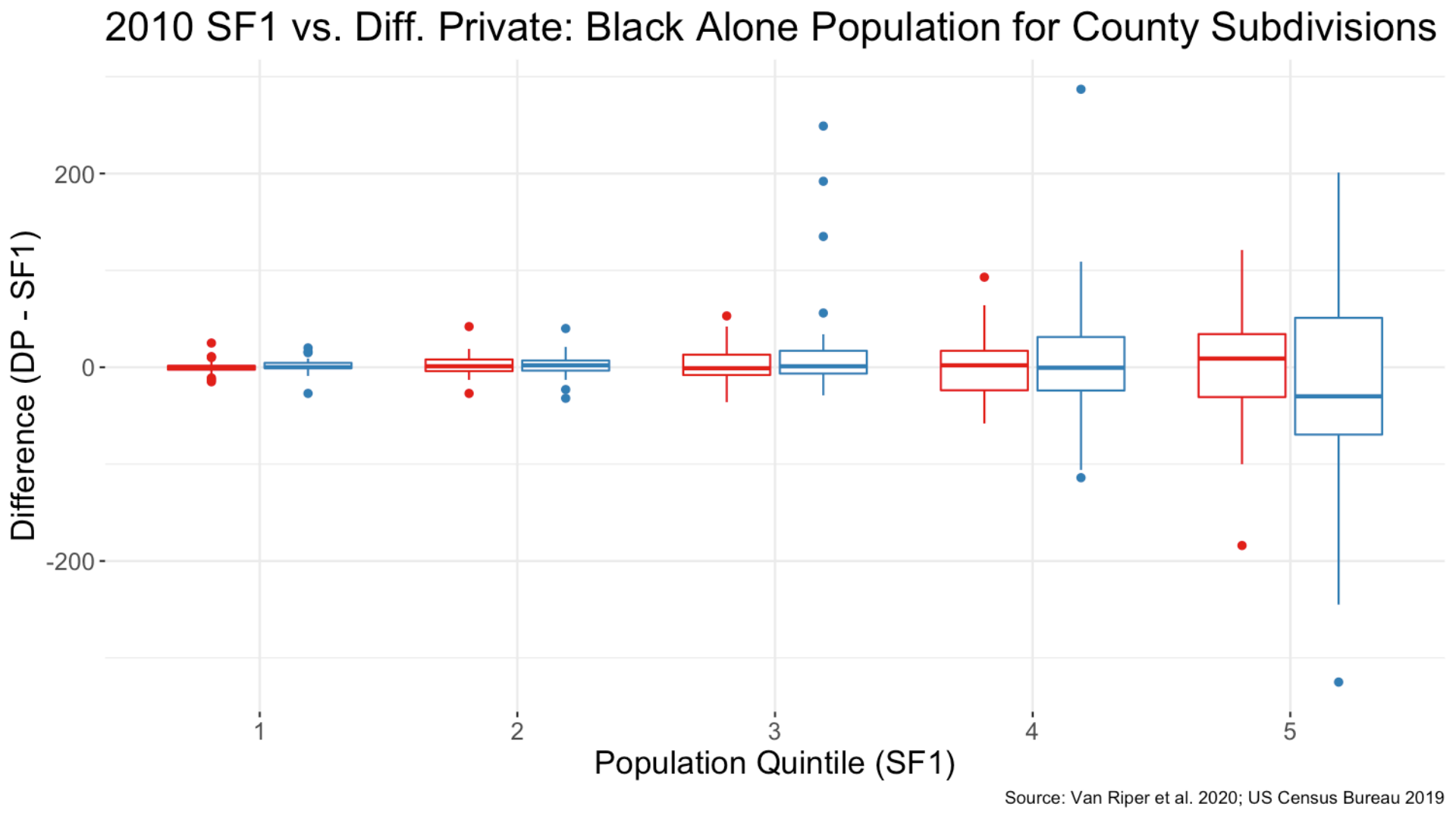
Query	Allocation (%)
Total population	30
Voting age * Hispanic * Race	29
Age * Sex * Hispanic * Race	25
Relation to HH/Group quarters	15
Detailed	1

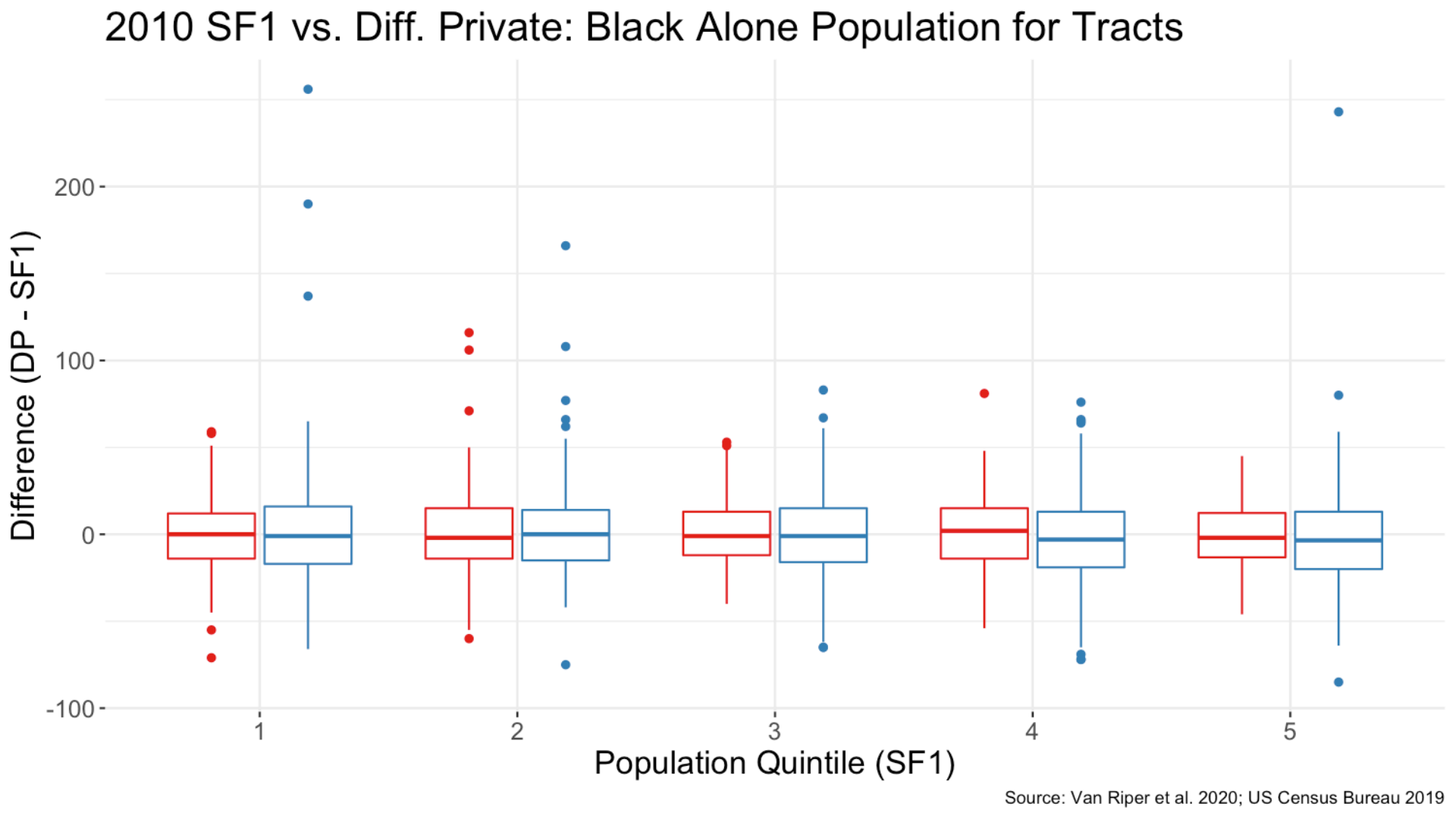

2010 SF1 vs. Diff. Private: Total Population for Counties
0
-100
-200
-300
-400
Difference (DP - SF1)
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019


2010 SF1 vs. Diff. Private: Total Population for County Subdivisions
Difference (DP - SF1)
200
0
-200
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019

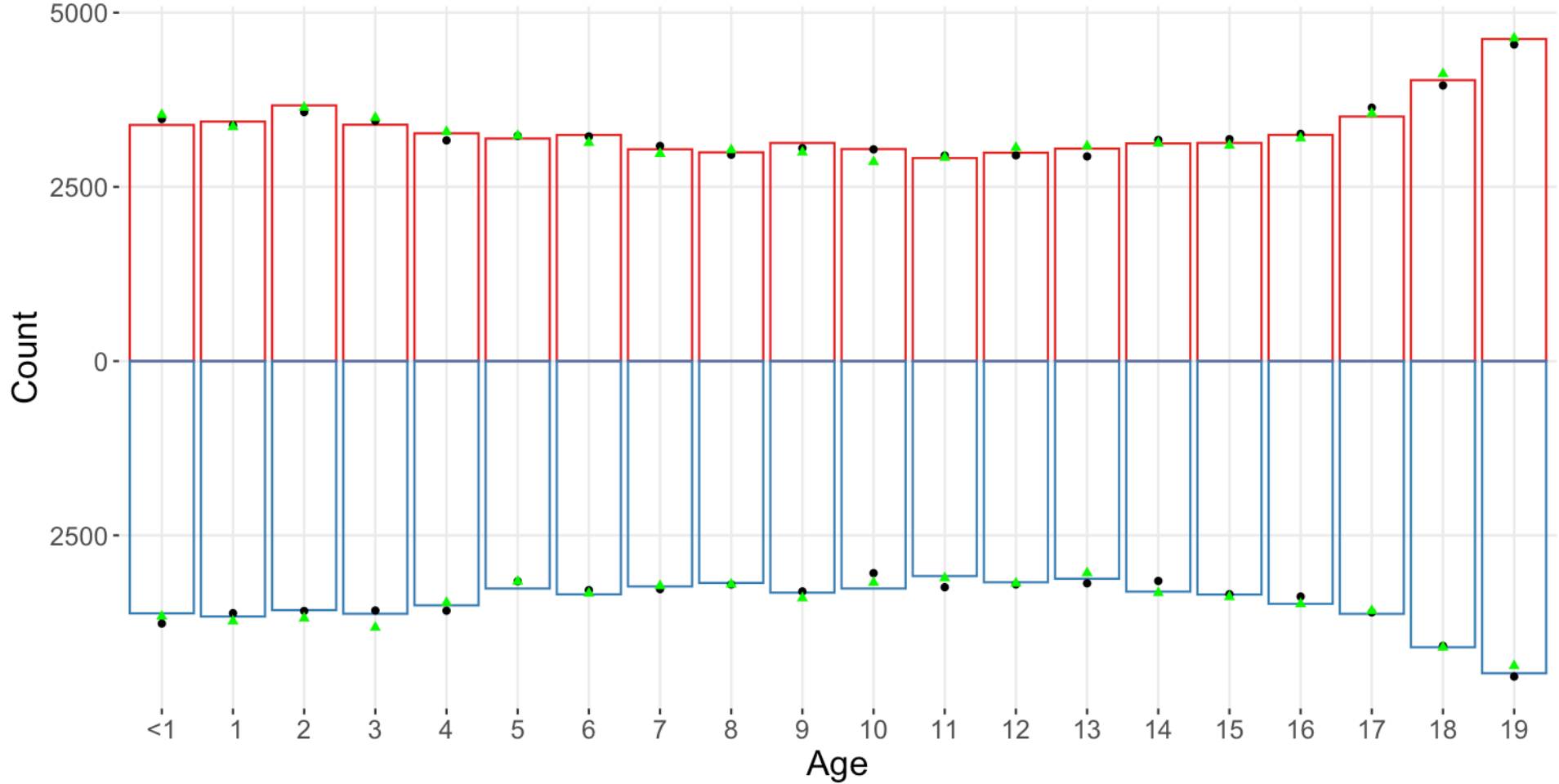

2010 SF1 vs. Diff. Private: Total Population for Tracts
Difference (DP - SF1)
100
50
0
-50
-100
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019


2010 SF1 vs. Diff. Private: Black Alone Population for Counties
Difference (DP - SF1)
100
50
0
-50
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019

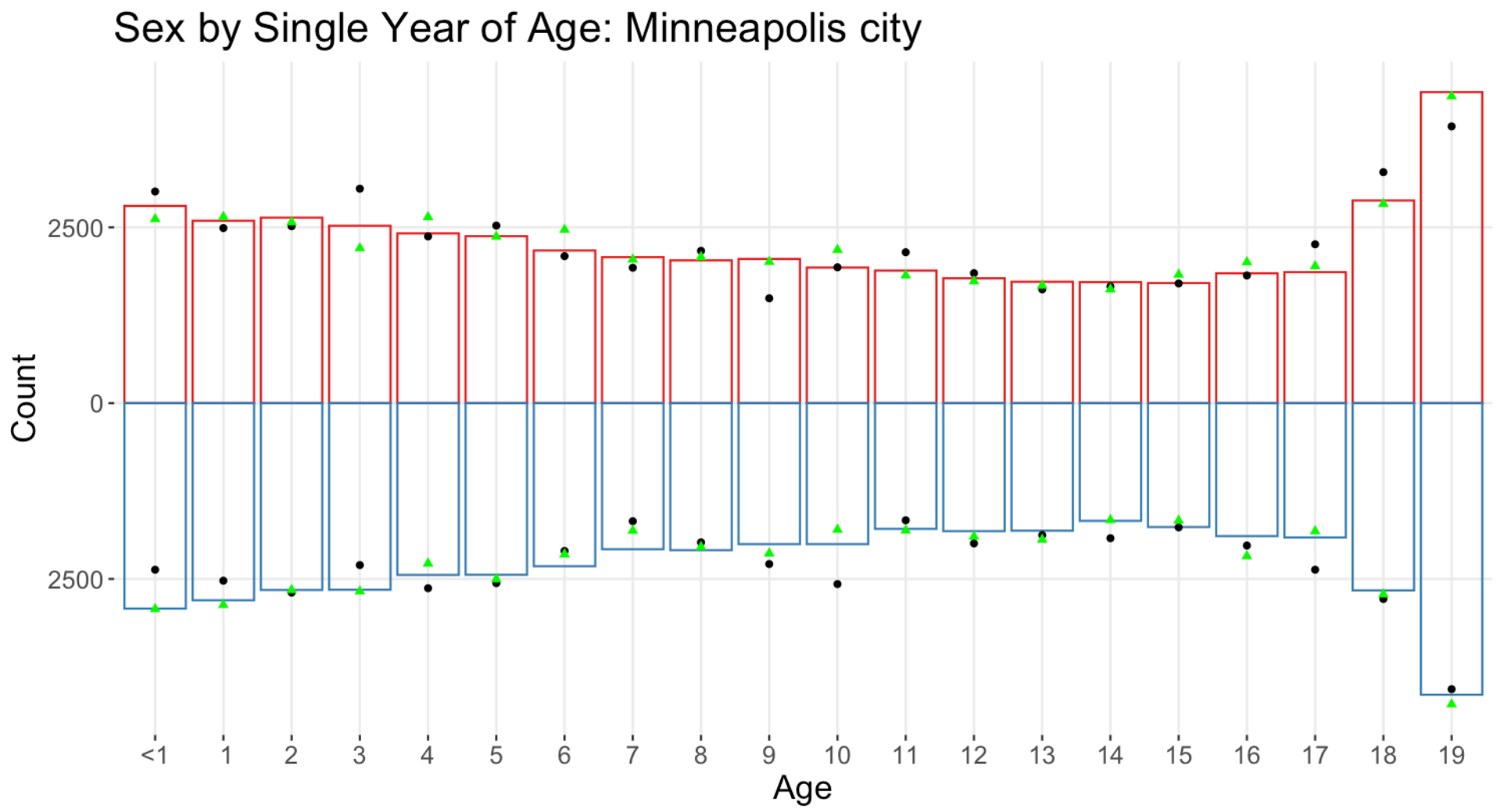

2010 SF1 vs. Diff. Private: Black Alone Population for County Subdivisions
200
0
-200
Difference (DP - SF1)
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019


2010 SF1 vs. Diff. Private: Black Alone Population for Tracts
Difference (DP - SF1)
200
100
0
-100
1
2
3
4
5
Population Quintile (SF1)
Source: Van Riper et al. 2020; US Census Bureau 2019

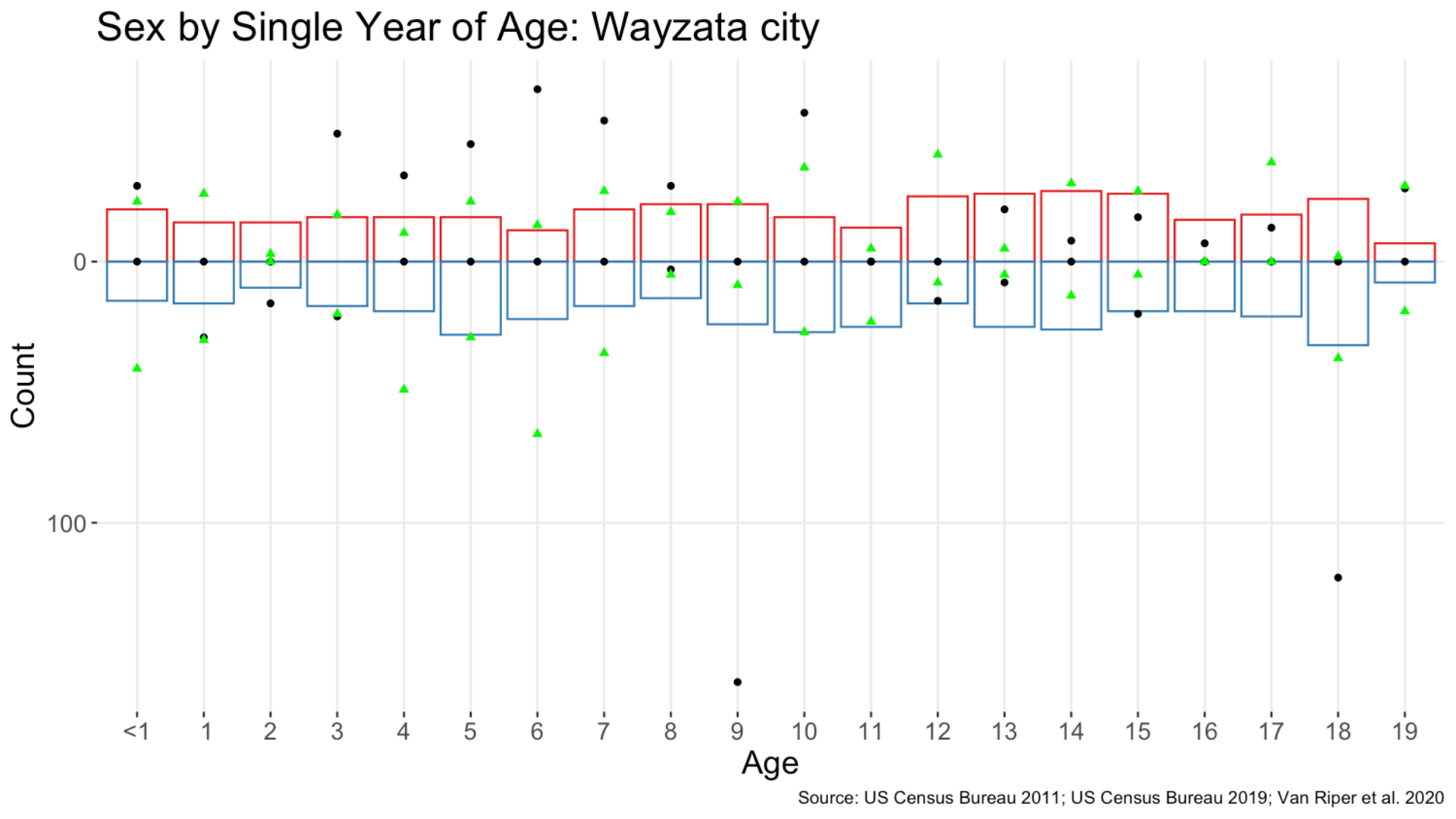
Sex by Single Year of Age: Ramsey County



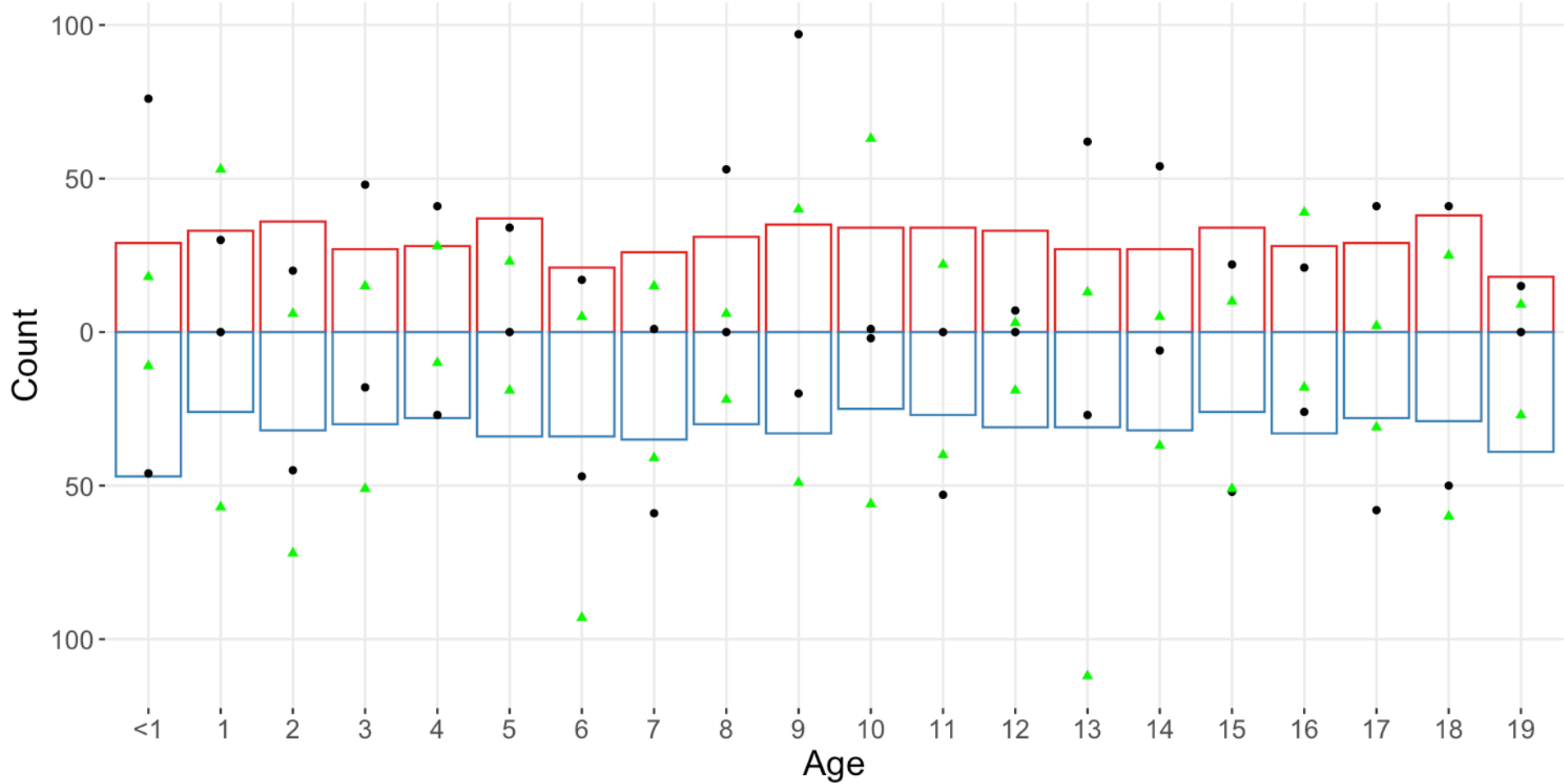
Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020



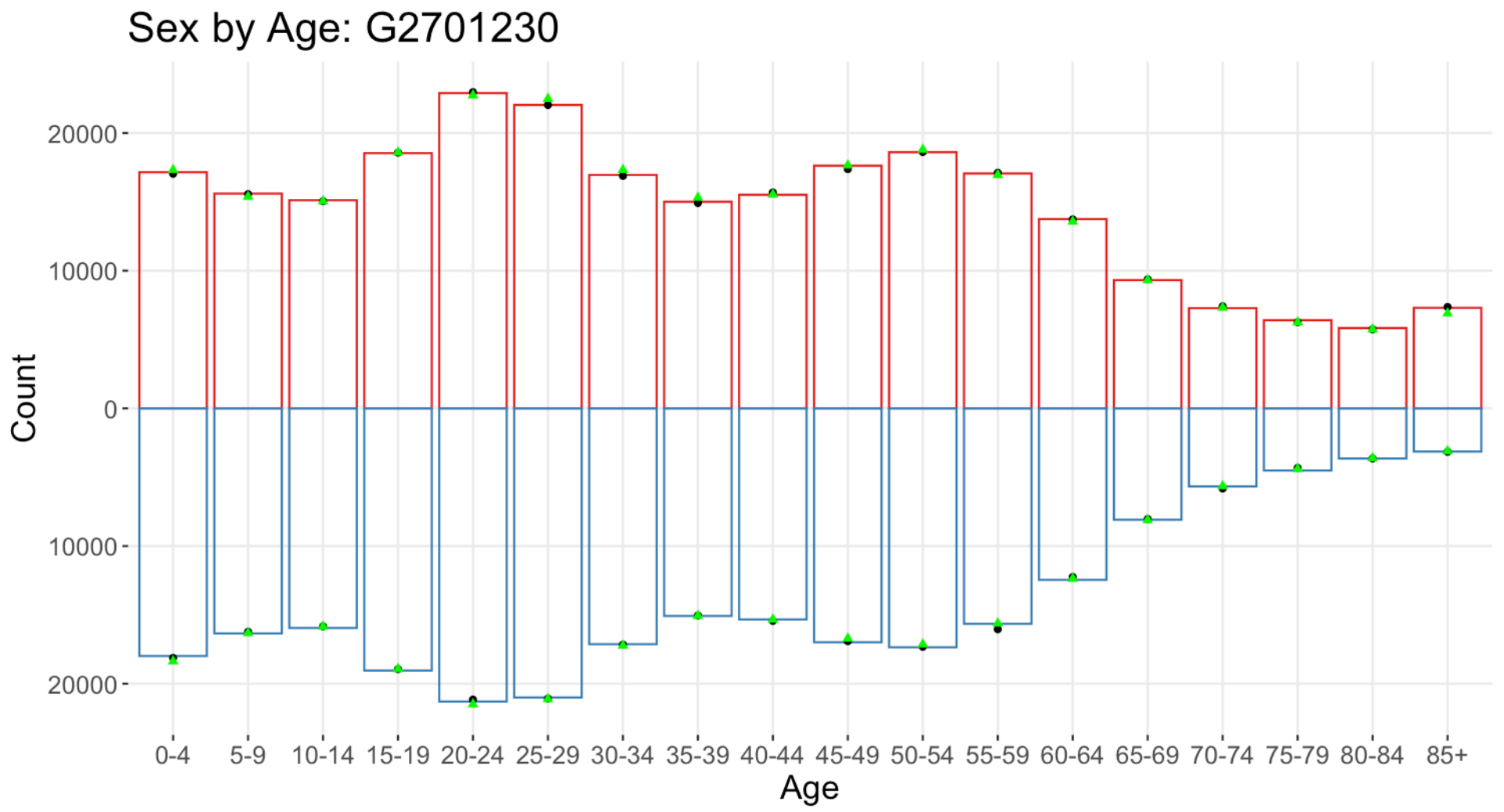
Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020


Sex by Single Year of Age: Wayzata city
Count
0
100
<1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Age
Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

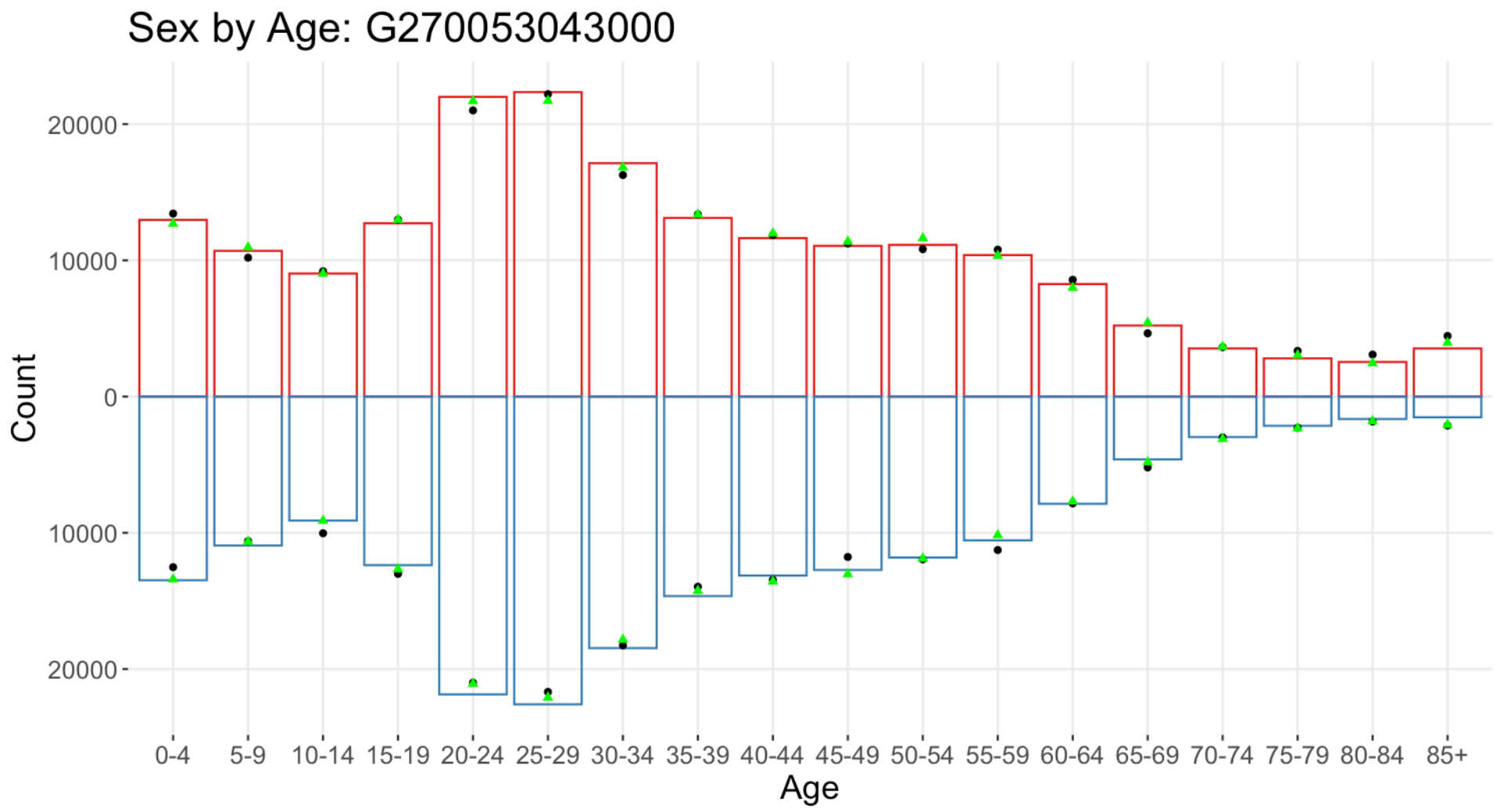
Sex by Single Year of Age: Census Tract 307.03

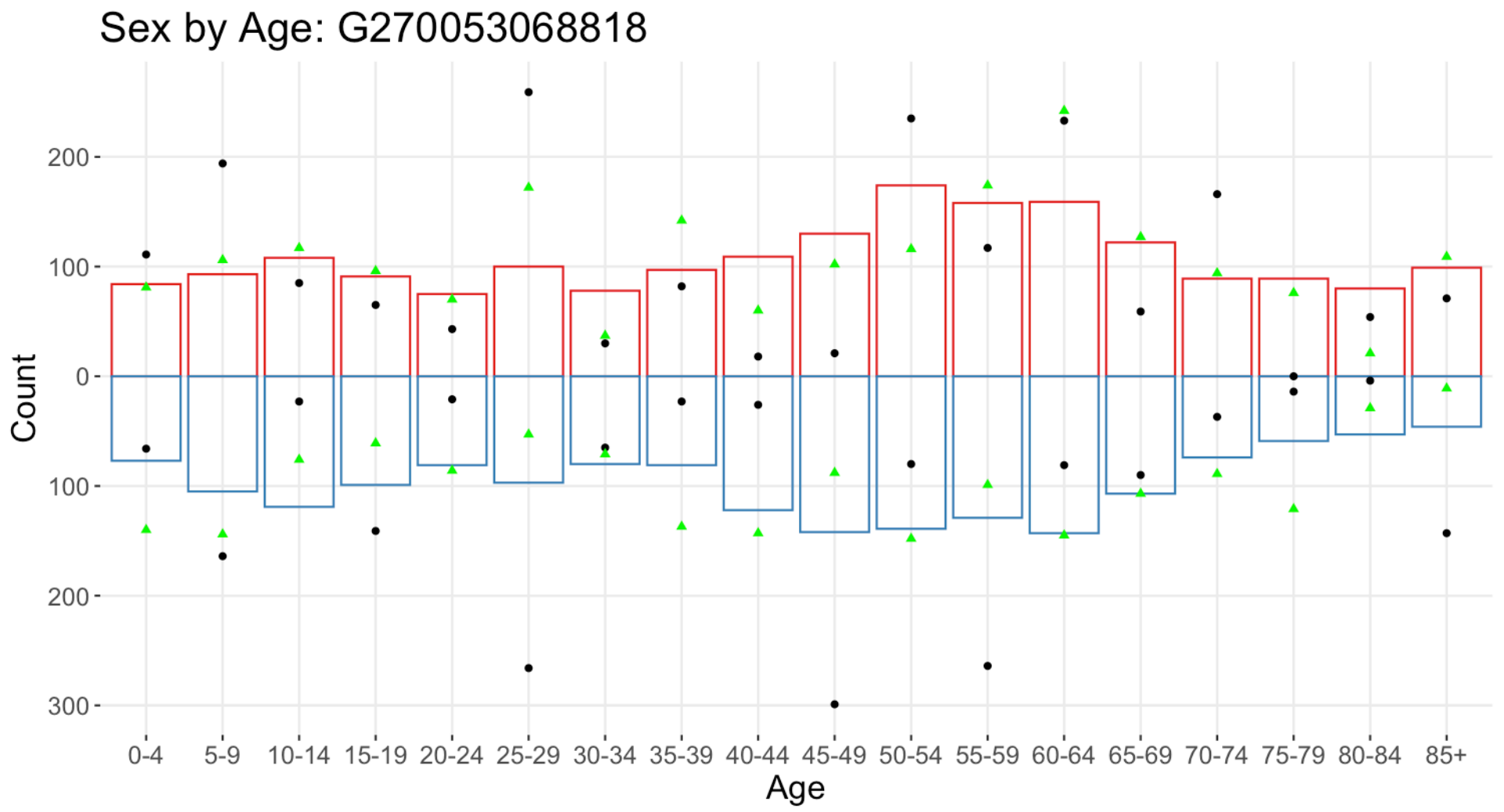


Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

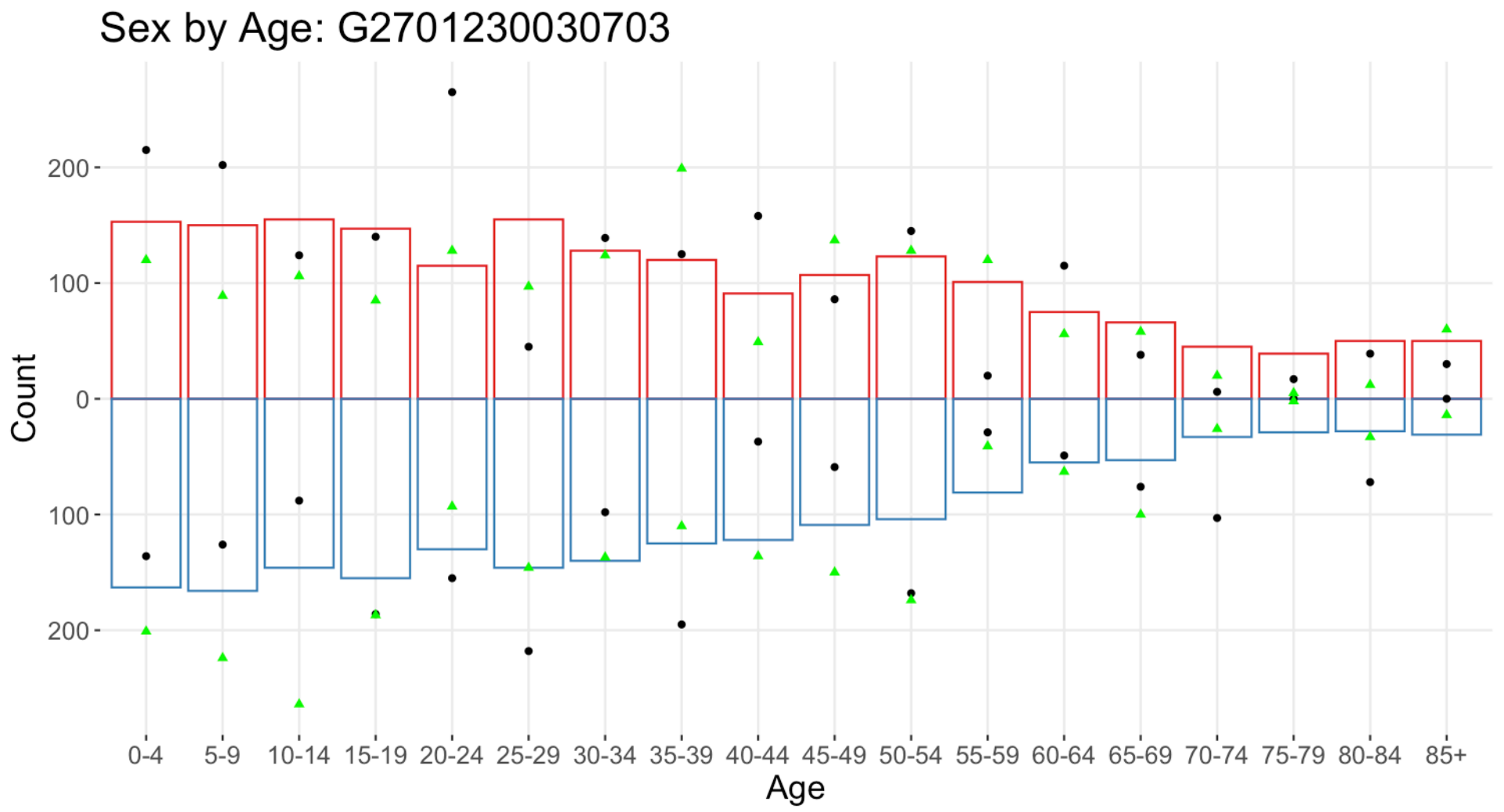


Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020


Sex by Age: G270053043000
20000
10000
0
10000
20000
Count
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
Age
Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020



Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020


Sex by Age: G2701230030703
Count
200
100
0
100
200
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
Age
Source: US Census Bureau 2011; US Census Bureau 2019; Van Riper et al. 2020

What's next?

- Census will release another demo. dataset
  - This week?
  - Only supports redistricting tables

What's next?

- Census will release another demo. dataset
  - This week?
  - Only supports redistricting tables
- Invariants will be set next week

What's next?

- Census will release another demo. dataset
  - This week?
  - Only supports redistricting tables
- Invariants will be set next week
- Final privacy loss budget and allocation set in early January