Problem Set 1, Solutions

Stats 506, Fall 2018

Due: Monday October 1, 5pm

The Rmarkdown file and all scripts used to create this document can be found on the Stats506_F18 (https://github.com/jbhender/Stats506_F18/) repository.

Question 1

In this question you will use command line tools to answer question about the 2015 Residential Energy Consumption Survey (RECS 2015) data set (https://www.eia.gov/consumption/residential/data/2015/index.php?view=microdata).

Part A [5 points; 2.5 each]

In part A, your solution to each question should be a Linux “one-liner”, i.e. a series of one or more commands connected by pipes “|”. Please provide both your solution and the result. Your solution must be written in text so that it can be copied and pasted if needed.

i. How many rows are there for region 3 in the RECS 2015 data set?

```
< recs2015_public_v3.csv cut -d, -f2 | grep -e "3" | wc -l
```

There are 2010 homes from region 3 in the data.

ii. Write a one-liner to create a compressed data set containing only the variables: DOEID, NWEIGHT, and BRRWT1-BRRWT96.

We can find the locations of these fields by looking in the codebook.

```
< recs2015_public_v3.csv cut -d, -f 1,475-571 > recs2015_weights.csv \
  && gzip recs2015_weights.csv
```

Part B [10 points; 5 each]

i. Write a Bash `for` loop to count and print the number of observations within each region.

Solution:

```
for r in `seq 4`
do
 echo -n "$r: "
 < recs2015_public_v3.csv cut -d, -f2 | grep -e $r | wc -l
done
```

Result:

```
1: 794
2: 1327
3: 2010
4: 1555
```

ii. Produce a file `region_division.txt` providing a sorted list showing unique combinations of values from `REGIONC` and `DIVISION`. Include the contents of that file in your solution. *Hint:* See `man uniq`.

Solution:

```
< recs2015_public_v3.csv head -1 | cut -d, -f2-3 > region_division.txt
< recs2015_public_v3.csv tail -n +2 | cut -d, -f2-3 | sort -t'"' -k4 -n | uniq >> region_divisi
on.txt
```

In this solution, the first line extracts the headers for the two columns of interest and writes them to `region_division.txt`. The second line removes the header row, extracts the columns of interest, sorts them and appends the unique combinations to our output file.

Result:

```
"REGIONC","DIVISION"
"1","1"
"1","2"
"2","3"
"2","4"
"3","5"
"3","6"
"3","7"
"4","8"
"4","9"
"4","10"
```

Question 2 [25 pts]

In this question, you will use **R** to answer questions about flights originating in New York City, NY (NYC) in 2013 and 2014. Data for 2013 can be found in the `nycflights2013` **R** package. Data through October 2014 is available here (https://raw.githubusercontent.com/wiki/arunsrinivasan/flights/NYCflights14/flights14.csv). Your answers should be submitted as nicely formatted tables produced using Rmarkdown.

Source code for solutions:

```
source('./ps1_q2.R')
```

a. Which airlines were responsible for at least 1% of the flights departing any of the three NYC airports between January 1 and October 31, 2013?

Solution:

Table 1. Airlines representing 1% of flights originating from the NYC area January - October, 2013.

Airline	Carrier Code	Flights	Percent
United Air Lines Inc.	UA	48,880	17.4%
JetBlue Airways	B6	45,605	16.2%
ExpressJet Airlines Inc.	EV	45,395	16.1%
Delta Air Lines Inc.	DL	40,168	14.3%
American Airlines Inc.	AA	27,447	9.8%

Airline	Carrier Code	Flights	Percent
Envoy Air	MQ	22,202	7.9%
US Airways Inc.	US	17,232	6.1%
Endeavor Air Inc.	9E	15,232	5.4%
Southwest Airlines Co.	WN	10,143	3.6%
Virgin America	VX	4,235	1.5%
AirTran Airways Corporation	FL	2,845	1.0%

b. Among the airlines from part “a”, compare the percent of annual flights in the first 10 months of 2013 and the first 10 months of 2014. Your table should include: the airline name (not carrier code), nicely formatted n’s (see `format()`), percents for each year with 95% CI, and change in percent with 95% CI. Which airlines showed the largest increase and decrease? Why do some airlines show an increase in the percent of flights but a decrease in the number of flights?

Solution:

Table 2. Top airlines serving NYC in 2013 and 2014. The table shows all airlines representing at least 1% of flights originating from NYC during January - October, 2013. Columns show total flights and percent of all flights for that period and the corresponding period from 2014. The final column shows the absolute change in each airline’s share of all flights.

Airline	2013, n	2013, %	2014, n	2014, %	Change in flight share, %
United Air Lines Inc.	48,880	17.4% (17.0, 17.7)	46,267	18.3% (17.9, 18.6)	0.9% (0.4, 1.4)
JetBlue Airways	45,605	16.2% (15.9, 16.5)	44,479	17.6% (17.2, 17.9)	1.4% (0.9, 1.8)
ExpressJet Airlines Inc.	45,395	16.1% (15.8, 16.5)	39,819	15.7% (15.4, 16.1)	-0.4% (-0.9, 0.1)
Delta Air Lines Inc.	40,168	14.3% (13.9, 14.6)	41,683	16.5% (16.1, 16.8)	2.2% (1.7, 2.7)
American Airlines Inc.	27,447	9.8% (9.4, 10.1)	26,302	10.4% (10.0, 10.8)	0.6% (0.1, 1.1)
Envoy Air	22,202	7.9% (7.5, 8.2)	18,559	7.3% (7.0, 7.7)	-0.6% (-1.1, -0.0)
US Airways Inc.	17,232	6.1% (5.8, 6.5)	16,750	6.6% (6.2, 7.0)	0.5% (-0.0, 1.0)
Endeavor Air Inc.	15,232	5.4% (5.1, 5.8)	-	-	-5.4% (-5.8, -5.1)
Southwest Airlines Co.	10,143	3.6% (3.2, 4.0)	11,902	4.7% (4.3, 5.1)	1.1% (0.6, 1.6)
Virgin America	4,235	1.5% (1.1, 1.9)	4,797	1.9% (1.5, 2.3)	0.4% (-0.1, 0.9)
AirTran Airways Corporation	2,845	1.0% (0.6, 1.4)	1,251	0.5% (0.1, 0.9)	-0.5% (-1.1, 0.0)

Endeavor air has no 2014 flights originating from NYC in our data and consquently shows the largest decrease in flight share: from 5.4% in 2013 to 0% in 2014. Delta's flight share increased 2.2% from 14.3% to 16.5%, significantly larger than the next largest increase of 1.4% by Jet Blue ($z = 2.34$; $p = 0.02$).

Some airlines, such as United, increased their flight share despite a decrease in the number of flights because there were fewer flights overall in the first ten months of 2014.

c. Among the three NYC airports, produce a table showing the percent of flights each airline is responsible for. Limit the table to the airlines identified in part a and include confidence intervals for your estimates. Which airline is the largest carrier at each airport?

Solution:

Table 3. Airline share of flights originating from each of three NYC area airports. Percentages represent all flights form January 2013 - October 2014.

Airline	EWR share	JFK share	LGA share
United Air Lines Inc.	39.5% (39.1, 39.8)	4.4% (4.0, 4.8)	7.5% (7.1, 8.0)
JetBlue Airways	5.8% (5.4, 6.2)	39.6% (39.2, 39.9)	5.7% (5.3, 6.1)
ExpressJet Airlines Inc.	34.7% (34.4, 35.1)	1.3% (0.8, 1.7)	10.2% (9.8, 10.6)
Delta Air Lines Inc.	4.1% (3.7, 4.5)	20.5% (20.1, 20.9)	22.1% (21.7, 22.5)
American Airlines Inc.	2.9% (2.5, 3.4)	13.3% (12.9, 13.8)	14.4% (14.0, 14.8)
Envoy Air	1.2% (0.7, 1.6)	6.6% (6.1, 7.0)	15.8% (15.4, 16.2)
US Airways Inc.	3.8% (3.4, 4.2)	2.9% (2.5, 3.4)	12.6% (12.1, 13.0)
Endeavor Air Inc.	0.6% (0.2, 1.0)	7.6% (7.2, 8.0)	1.3% (0.9, 1.8)
Southwest Airlines Co.	5.3% (4.9, 5.7)	-	7.0% (6.5, 7.4)
Virgin America	1.5% (1.1, 2.0)	3.5% (3.1, 3.9)	0.0% (-0.4, 0.5)
AirTran Airways Corporation	-	-	2.4% (1.9, 2.8)

United has the largest share of flights (40%) originating from EWR, Jet Blue (40%) from JFK, and Delta (22%) from LGA.

Question 3 [45 pts; 15 pts each]

In this question, you will use **R** to answer questions about the RECS 2015 data. You should read the section on computing standard errors available here (https://www.eia.gov/consumption/residential/data/2015/pdf/microdata.pdf). For each question, produce a nicely formatted table and graph to support you answer. In your tables and graphs please provide standard errors for all point estimates.

a. What percent of homes have stucco construction as the *major outside wall material* within each division? Which division has the highest proportion? Which the lowest?

Solution:

Table 3. Airline share of flights originating from each of three NYC area airports. Percentages represent all flights form January 2013 - October 2014.

Census Division	% Stucco Homes (95% CI)
Mountain South	64.2% (55.4, 73.0)
Pacific	44.6% (41.3, 47.9)
Mountain North	16.6% (10.2, 23.0)
South Atlantic	10.6% (7.8, 13.4)
West North Central	4.9% (0.9, 8.8)
West South Central	3.0% (1.6, 4.3)
Middle Atlantic	2.1% (0.6, 3.5)
New England	1.2% (0.0, 2.8)
East North Central	0.7% (0.1, 1.2)
East South Central	0.4% (0.0, 1.2)

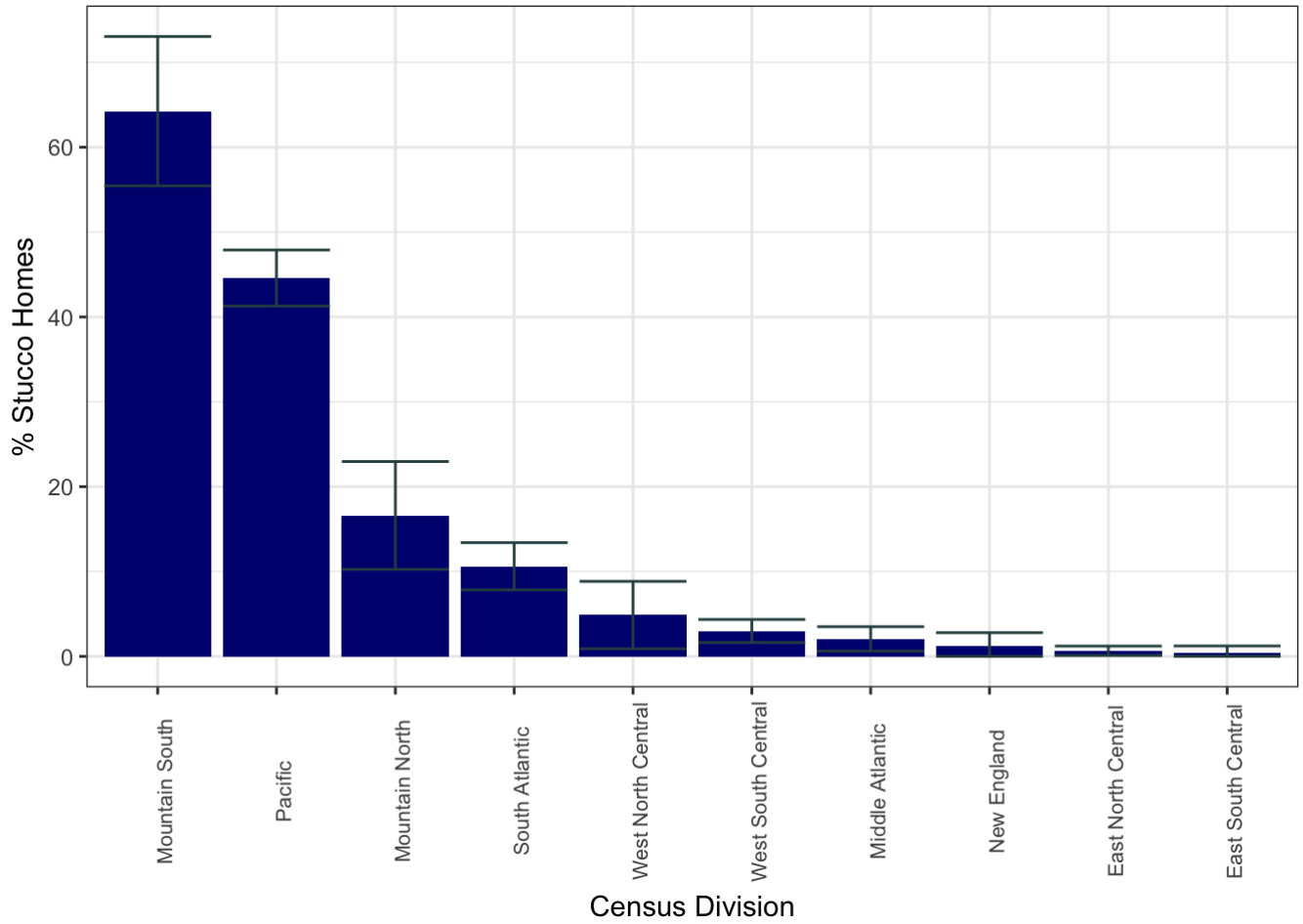


Figure 1. Estimated percent of homes within each census division with major wall type of stucco.

Solution:

The Mountain South division has 64% (95% CI: 55-73%) of homes built with stucco. It is the only division in which more than half the homes are built with this material; it and the Pacific division are the only divisions with more than 25% stucco homes. The East South Central and East North Central divisions both have less than 1% estimated stucco homes but other divisions are comparably low within sampling uncertainty.

b. What is the average total electricity usage in kilowatt hours in each division? Answer the same question stratified by urban and rural status.

Solution:

Table 4. Average annual electricity utilization by Census Division in kwh/home.

Census Division	Average Electricity Usage, kwh/home (95% CI)
East South Central	14,536, (13,320 - 15,752)
West South Central	14,324, (13,495 - 15,153)
South Atlantic	13,447, (12,904 - 13,989)
West North Central	10,524, (9,635 - 11,413)
Mountain South	10,442, (7,950 - 12,934)
East North Central	9,129, (8,730 - 9,528)
Middle Atlantic	8,465, (8,071 - 8,860)
Mountain North	8,384, (7,121 - 9,648)
Pacific	8,100, (7,750 - 8,450)
New England	7,515, (6,472 - 8,557)

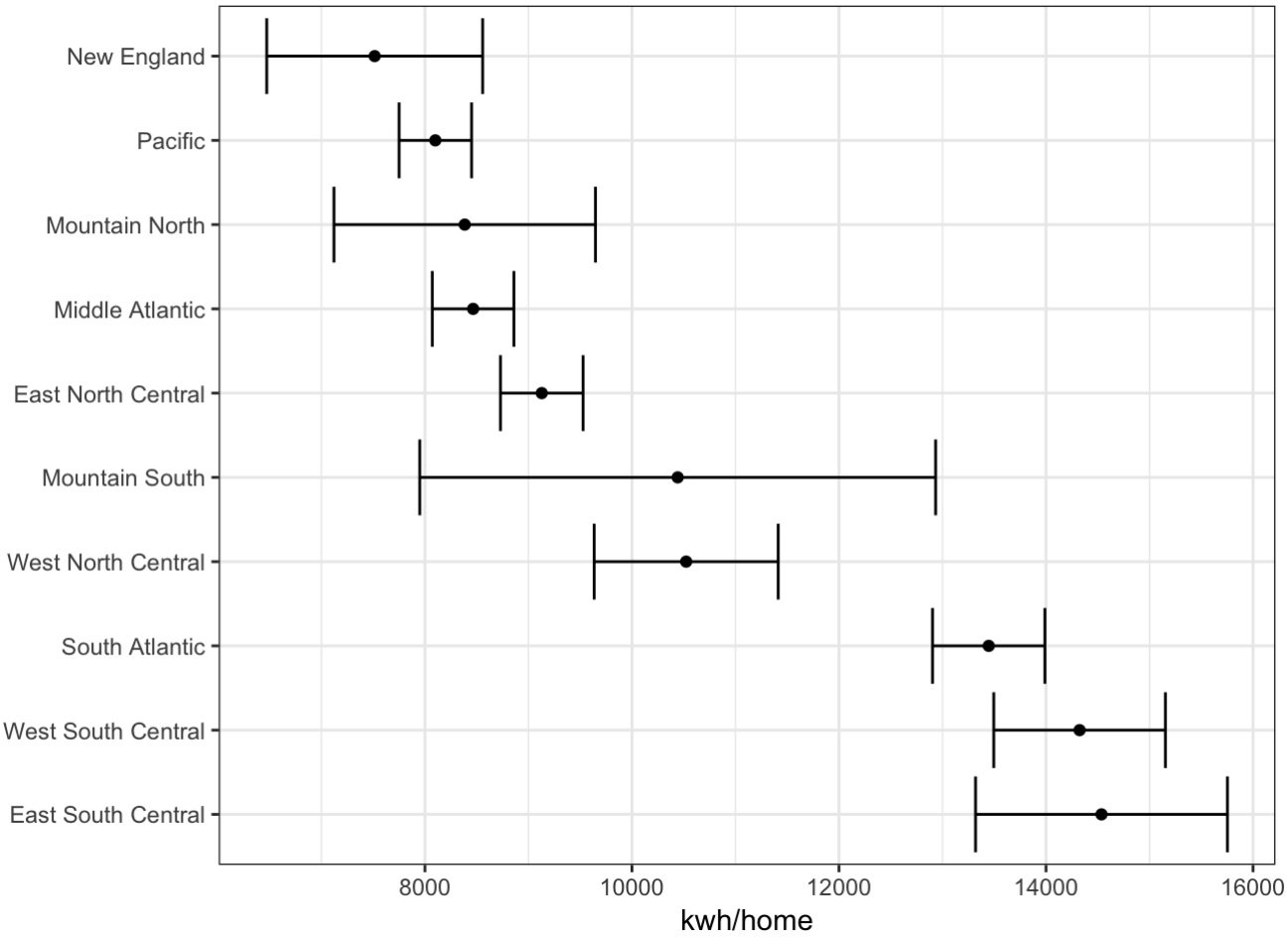
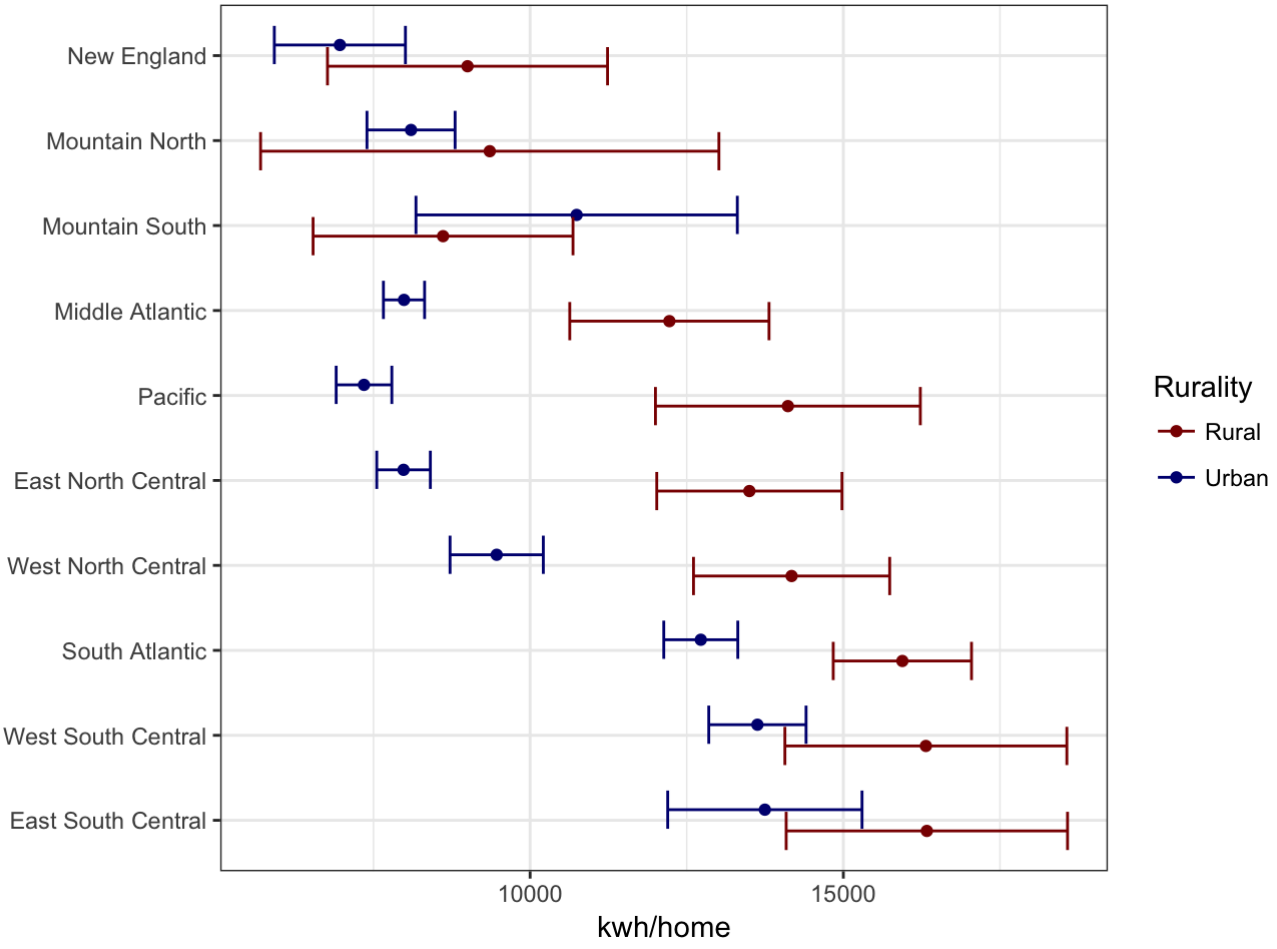


Figure 2. Estimated average annual electricity usage in khw/home for each of 10 census divisions.

Table 5. Average electricity utilization in kwh per home for urban and rural areas witihin each census division.

Census Division	Rural, kwh/home (95% CI)	Urban, kwh/home (95% CI)
East South Central	16,333, (14,088 - 18,578)	13,747, (12,197 - 15,298)

Census Division	Rural, kwh/home (95% CI)	Urban, kwh/home (95% CI)
West South Central	16,317, (14,067 - 18,567)	13,629, (12,852 - 14,405)
South Atlantic	15,942, (14,839 - 17,045)	12,725, (12,134 - 13,316)
West North Central	14,174, (12,608 - 15,740)	9,467, (8,722 - 10,211)
East North Central	13,500, (12,022 - 14,978)	7,980, (7,552 - 8,408)
Pacific	14,115, (12,001 - 16,229)	7,349, (6,905 - 7,793)
Middle Atlantic	12,223, (10,633 - 13,814)	7,987, (7,659 - 8,316)
Mountain South	8,610, (6,536 - 10,685)	10,743, (8,178 - 13,308)
Mountain North	9,356, (5,698 - 13,014)	8,099, (7,396 - 8,803)
New England	9,001, (6,766 - 11,236)	6,964, (5,918 - 8,010)



c. Which division has the largest disparity between urban and rural areas in terms of the proportion of homes with internet access?

Solution:

Table 6. Urban and rural disparity in internet access for the ten US Census Division in 2015.

Census Division	Urban Internet Access, % (95% CI)	Rural Internet Access, % (95% CI)	Difference, % (95% CI)
Mountain South	85.3 (81.3, 89.2)	66.7 (58.3, 75.2)	18.5% (7.2, 29.8)

Census Division	Urban Internet Access, % (95% CI)	Rural Internet Access, % (95% CI)	Difference, % (95% CI)
East South Central	78.4 (70.5, 86.2)	69.0 (63.5, 74.6)	9.3% (-1.4, 20.1)
West North Central	88.0 (84.6, 91.4)	80.3 (71.5, 89.2)	7.7% (-2.5, 17.8)
Mountain North	87.4 (82.0, 92.9)	81.9 (73.8, 90.0)	5.5% (-6.2, 17.2)
West South Central	81.6 (76.4, 86.8)	76.5 (72.1, 80.9)	5.1% (-2.3, 12.5)
Pacific	88.7 (86.2, 91.2)	85.3 (77.4, 93.1)	3.4% (-4.5, 11.4)
South Atlantic	85.3 (82.6, 88.0)	82.0 (76.3, 87.8)	3.3% (-3.5, 10.1)
New England	87.6 (82.5, 92.6)	85.8 (82.4, 89.2)	1.8% (-2.5, 6.0)
East North Central	86.3 (83.8, 88.7)	86.2 (81.6, 90.8)	0.0% (-5.3, 5.4)
Middle Atlantic	89.3 (83.9, 94.8)	91.3 (85.3, 97.3)	-1.9% (-9.1, 5.2)

In the Mountain South division there is an 18.5% disparity between Urban and Rural internet access. This is approximately twice as large as the next largest estimated disparity and the only estimate whose confidence interval does not include zero.