Demography Final Project

January 3, 2019

In [5]:
```
#CHECKLIST
#1. Download Data
#2. Sort Data to get tables for Latinos and African Americans in the 4 time periods
#3. Get Regression Lines
#4. Make Scatterplots
#5. bootstrap the two regressions, resample, difference of the slopes ratio
#6. Confidence Intervals- ideally, seeing if one of the slopes is bigger than the othe
#7. p-values
#8. Draw conclusion
#how the immigration of Latino immigrants has affected the African American community

#I will analyze African Americans and Latinos in the labor force in both Suro's time p
#in different cities (percent of labor force that's hispanic on x axis, average wage o
#each dot represents a city, should look like negative slope).
#I will make scatterplots possibly with regression lines
#two slopes for 2000, 2017
#VARIABLE - MSA, composed of economically linked counties
#see if there has been a tangible effect and who has benefited from it.
#look at great migration lab
```

Background on Latino migration and its relation to African American labor in the United States, as told by Roberto Suro in his book, Stranger Among Us:

"The arrival of fresh Latino immigrants was only one part of all this churning, but because they were so new, so noticeable, and so numerous, they became the face of the change. One of the ways people saw the results was when whole job categories changed hands and newly arrived Latinos suddenly began doing work that had been performed by native-born workers before, usually African-Americans. It happened at car washes and on construction crews, in hotel banquet rooms and supermarkets. At first glance, it seemed easy to conclude that the new Latinos had 'taken' the jobs once held by blacks. That certainly seemed to be the case with an important job category that changed drastically in Los Angeles: the work of cleaning offices.

In 1980, many, if not most, of the janitors in big L.A. office buildings were African Americans, especially in the traditional downtown business district. They could count on wages and benefits worth more than twelve dollas an hour as part of secure, long-standing union contracts. Ten years later, most of the janitors in big buildings, especially in the newly developing areas such as Century City, were Latinos, usually recently arrived, often illegals. They worked for something close to the minimum wage with no contract, no vacation, no health insurance, no benefits, no overtime. The social contract that had provided thousands of black workers a decent wage was broken, and it was replaced wiht the sweatshop labor of immigrants" (Suro 213).

This idea that Suro explains is what inspired me to think about the influences of immigration on the labor and income of those currently living in the country. For my project, I decided to focus on these aspects he brings up in regards to Hispanic immigration and African American labor. My goal is to explore how the immigration of Latino Immigrants to the United States, in the labor force, affected African Americans living in the United States at the time, also in the labor force.

Null hypothesis: There is no difference between the effects of Latino immigration on African American labor and wages in 2000 and 2017. Any difference observed is due to random chance.

Alternative hypothesis: There is a difference between the effects of Latino immigration on African American labor and wages in 2000 and 2017. Any difference observed is not due to random chance.

Process: I will get the specific datasets for Latinos and African Americans as depicted above, use bootstrapping to get an array of the the regression line slopes, then make a 95% confidence interval of the difference between 2000 and 2017 in terms of these data to see if there has indeed been a difference over the years.

Test Statistic: the difference between the regression line slopes of the bootstrapped data with Latino immigrants in the labor force and African American average income in the labor force in 2000 and 2017.

There are a few aspects of this test statistic and the dataset I used that may affect the accuracy of the result. First of all, the use of both people who are employed and unemployed could make a difference as this might change the appearance of the effects on African Americans. The fact that I chose 2000 and 2017 may result in more insignificant differences than if I chose years that were farther apart or years in a different time period all together. My inclusion of both men and women may have made a difference as well; men might be more likely to be in the labor force and/or have higher incomes than women. Though I did specify age to be 20-60, a likely age at which people are in the labor force, perhaps age does make a difference in that whether the person is married, has children, needs to support a family, etc. Level of education of both Latino immigrants and African Americans may have caused an effect as well; perhaps those who are highly successful are outliers that are changing the reults. Additionally, the variable I used to select for Latinos is "HISPAN," or Hispanic origin. This means that the people in the data are not all necessarily first generation immigrants, or even completely Hispanic; the person just identifies as Hispanic. This may make a difference in the effects this variable actually has on African American labor.

On the other hand, I decided to use these slopes and the confidence interval because I felt that it would be the clearest way with which I could see if there is a difference between 2000 and 2017, what that difference might be, and how it may or may not have affected African Americans.

```
In [6]: #Imports
        import pandas as pd
        import numpy as np
        import scipy as sp
        import matplotlib.pyplot as plt
        plt.style.use('fivethirtyeight')

        %matplotlib inline
        from datascience import Table
        from datascience.predicates import are
        from datascience.util import *
```

```
In [7]: #tables
        general_tbl = Table.read_table("DemogFinalProject.csv")
```

```
general_tbl = general_tbl.drop("BPL", "BPLD", "YRIMMIG")
general_tbl
```

Out[7]:

YEAR	DATANUM	SERIAL	CBSERIAL	HHWT	MET2013	GQ	PERNUM	PERWT	AGE
2000	1	17	nan	1147.01	39300	1	1	1446.23	42
2000	1	67	nan	1296.62	39300	1	1	1446.23	25
2000	1	117	nan	1047.27	39300	1	1	1296.62	41
2000	1	166	nan	748.05	39300	1	1	398.96	41
2000	1	216	nan	1047.27	39300	1	1	748.05	20
2000	1	266	nan	947.53	39300	1	5	1047.27	31
2000	1	565	nan	349.09	39300	4	1	349.09	58
2000	1	665	nan	947.53	39300	1	2	897.66	57
2000	1	915	nan	1296.62	39300	1	1	1944.93	26
2000	1	964	nan	648.31	39300	1	1	897.66	43

... (250387 rows omitted)

Before even getting to this table, in Ipums itself I selected for age: (the AGE variable) 20-60, people whom I knew were more likely to be participants in the labor force and earning members. Additionally, I also selected for people who are in the labor force (the EMPSTAT variable) since including those not in the labor force would throw off the data as we are looking at average salary for African Americans.

YEAR: 2000 and 2017. These were used since they're during a time period that's close to current day, where Latino immigration was definitely present, and also far enough apart that we can see if there is a long term difference in years.

MET2013: This represents larger, identifiable metropolitan areas. These would be the most effective places to look for differences over longer periods of time, as they are more likely to be frequented by immigrants and can easily be grouped to map out the differences.

HISPAN: This is to select those only with Hispanic origin for the dataset.

HISPAND: This is a more specific version of the general HISPAN which I did not end up using.

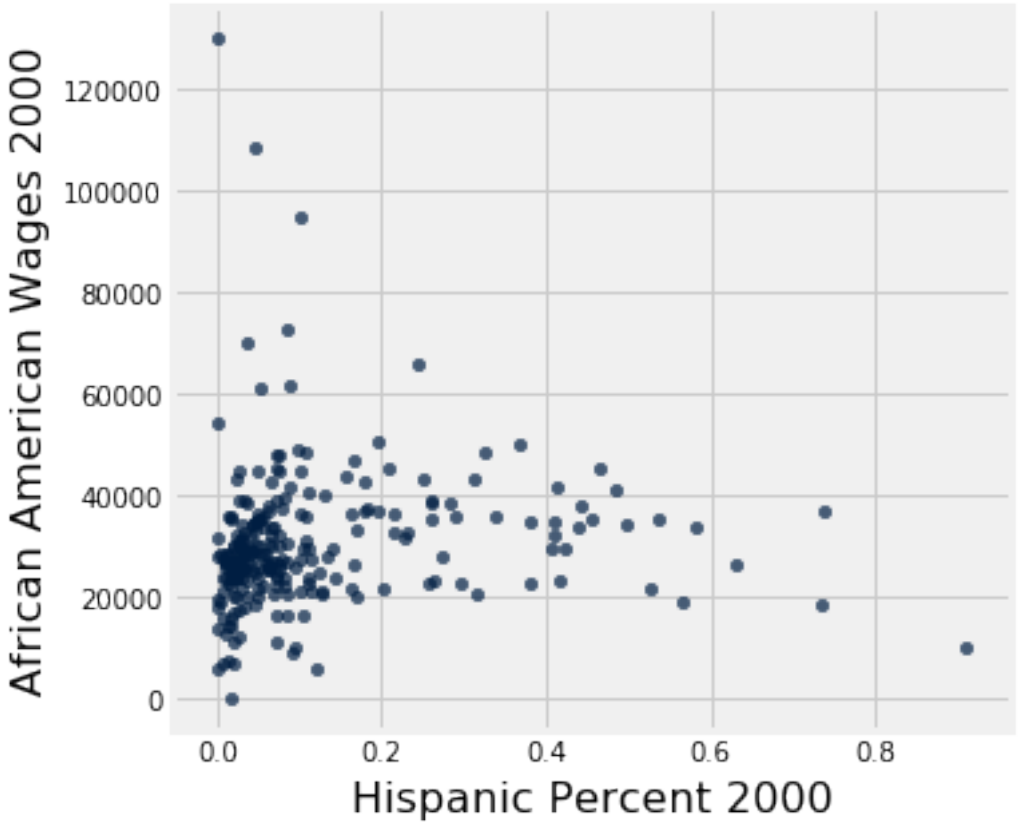
RACLBK: This is to select those only with African American origin for the dataset.

INCTOT: This is the total income for the individual in a year. It is useful to find the average income for African Americans and the two years specified.

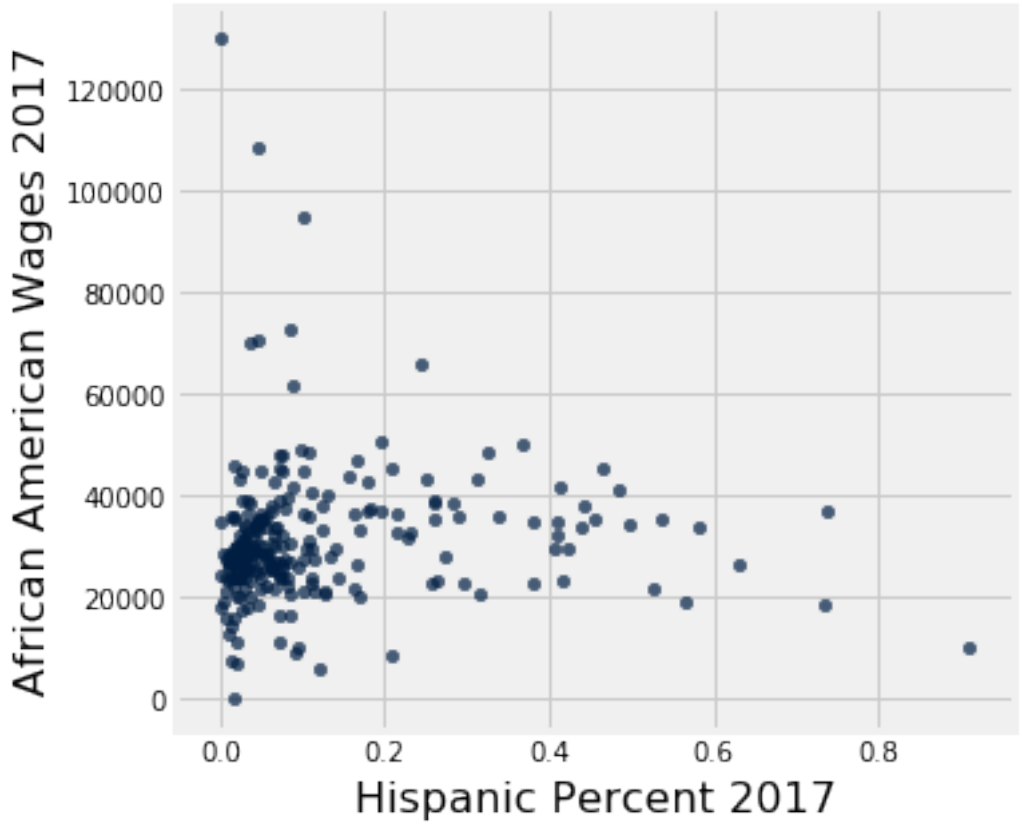
In [8]:
```
grouped2000 = general_tbl.where("YEAR", are.equal_to(2000)).group("MET2013")
grouped2000 = grouped2000.where("MET2013", are.not_equal_to(0))
num_hispanic2000 = make_array()
af_wage2000 = make_array()
for i in np.arange(grouped2000.num_rows):
    tbl_specific = general_tbl.where("MET2013", are.equal_to(grouped2000.column("MET20
    num_hispanic2000 = np.append(num_hispanic2000, (tbl_specific.where("HISPAN", are.n
    af_tbl = tbl_specific.where("RACBLK", are.equal_to(2))
    af_wage2000 = np.append(af_wage2000, np.average(af_tbl.column("INCTOT")))
```

```
/srv/app/venv/lib/python3.6/site-packages/numpy/lib/function_base.py:1128: RuntimeWarning: Mean
  avg = a.mean(axis)
/srv/app/venv/lib/python3.6/site-packages/numpy/core/_methods.py:80: RuntimeWarning: invalid va
  ret = ret.dtype.type(ret / rcount)
```

```
In [9]: grouped2017 = general_tbl.where("YEAR", are.equal_to(2017)).group("MET2013")
        grouped2017 = grouped2017.where("MET2013", are.not_equal_to(0))
        num_hispanic2017 = make_array()
        af_wage2017 = make_array()
        for i in np.arange(grouped2017.num_rows):
            tbl_specific = general_tbl.where("MET2013", are.equal_to(grouped2017.column("MET20
            num_hispanic2017 = np.append(num_hispanic2017, (tbl_specific.where("HISPAN", are.no
            af_tbl = tbl_specific.where("RACBLK", are.equal_to(2))
            af_wage2017 = np.append(af_wage2017, np.average(af_tbl.column("INCTOT")))

In [10]: tbl_2000 = Table().with_columns("Hispanic Percent 2000", num_hispanic2000, "African Am
                                         af_wage2000)
         tbl_2017 = Table().with_columns("Hispanic Percent 2017", num_hispanic2017, "African Am
                                         af_wage2017)

In [11]: #regression equations
         def std_u(arr):
             return (arr - np.mean(arr))/np.std(arr)

         def find_r(tbl, col_x, col_y):
             return np.mean(std_u(tbl.column(col_x))*std_u(tbl.column(col_y)))

         def slope(tbl, col_x, col_y):
             r = find_r(tbl, col_x, col_y)
             return r*np.std(tbl.column(col_y))/np.std(tbl.column(col_x))

         def intercept(tbl, col_x, col_y):
             return np.mean(tbl.column(col_y)) - slope(tbl, col_x, col_y)*np.mean(tbl.column(co

In [12]: #scatterplot2000
         line_2000 = (slope(tbl_2000, "Hispanic Percent 2000", "African American Wages 2000") *
         tbl_2000.scatter("Hispanic Percent 2000", "African American Wages 2000", fit_line = Tr
```




```
In [13]: #scatterplot2017
         line_2017 = (slope(tbl_2017, 0, 1) * tbl_2017.column(0)) + intercept(tbl_2017, 0, 1)
         tbl_2017.scatter("Hispanic Percent 2017", "African American Wages 2017", fit_line = Tr
```

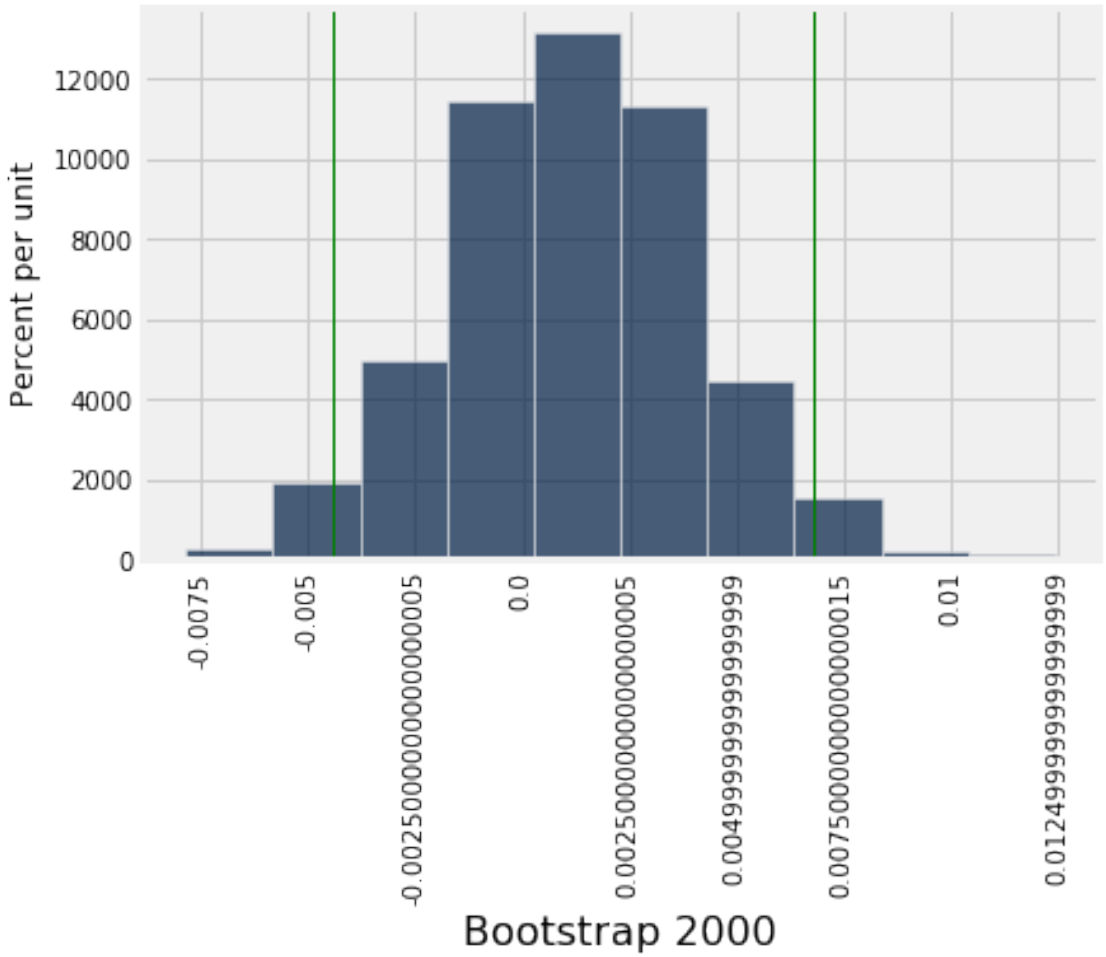



```
In [14]: #bootstrap2000
         bootstrap2000 = make_array()
         for i in np.arange(1000):
             boot_samp00 = grouped2000.sample()
             resampled_slope = slope(boot_samp00, "MET2013", "count")
             bootstrap2000 = np.append(bootstrap2000, resampled_slope)

In [15]: #bootstrap2017
         bootstrap2017 = make_array()
         for i in np.arange(1000):
             boot_samp17 = grouped2017.sample()
             resampled_slope = slope(boot_samp17, "MET2013", "count")
             bootstrap2017 = np.append(bootstrap2017, resampled_slope)

In [16]: #bootstrap table
         bootstrap_tbl = Table().with_columns("Bootstrap 2000", bootstrap2000, "Bootstrap 2017"
                                              "Differences", bootstrap2017-bootstrap2000)
         bootstrap_tbl

Out[16]: Bootstrap 2000 | Bootstrap 2017 | Differences
         0.00415987     | -0.00181996    | -0.00597983
```

```
0.000351561    | -0.00200555    | -0.00235711
0.00371415     | 0.00210236     | -0.00161179
0.00439542     | -0.00530635    | -0.00970177
0.000431207    | -0.00231045    | -0.00274166
-0.00212087    | -0.00254947    | -0.000428591
-0.000127023   | 0.00952129     | 0.00964831
0.00195974     | 0.00183005     | -0.00012969
-0.00284928    | 0.00698844     | 0.00983772
0.00144561     | 0.00287548     | 0.00142987
... (990 rows omitted)
```

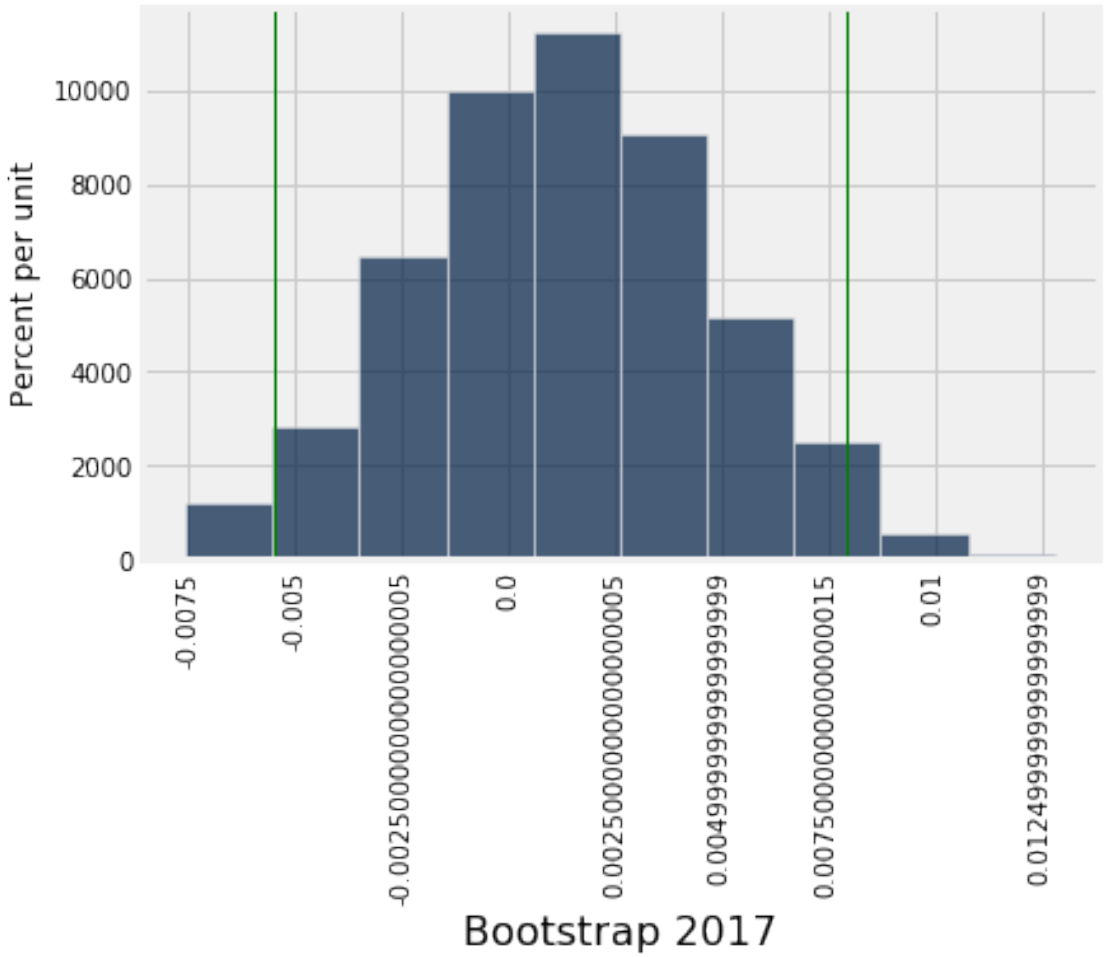
In [17]:
```
#bootstrap hist 2000
bootstrap_tbl.hist("Bootstrap 2000", unit="")

p05_2000=Table().with_column('dind',bootstrap_tbl.column("Bootstrap 2000")).percentile
p95_2000=Table().with_column('dind',bootstrap_tbl.column("Bootstrap 2000")).percentile
plt.axvline(x=p05_2000,color='green',linewidth=1)
plt.axvline(x=p95_2000,color='green',linewidth=1)

left_2000 = percentile(2.5, bootstrap_tbl.column("Bootstrap 2000"))
right_2000 = percentile(97.5, bootstrap_tbl.column("Bootstrap 2000"))
print("The 95% confidence interval for the difference between the two data sets is:",
     left_2000 , "and", right_2000)
```

```
The 95% confidence interval for the difference between the two data sets is: -0.00445039185522
```



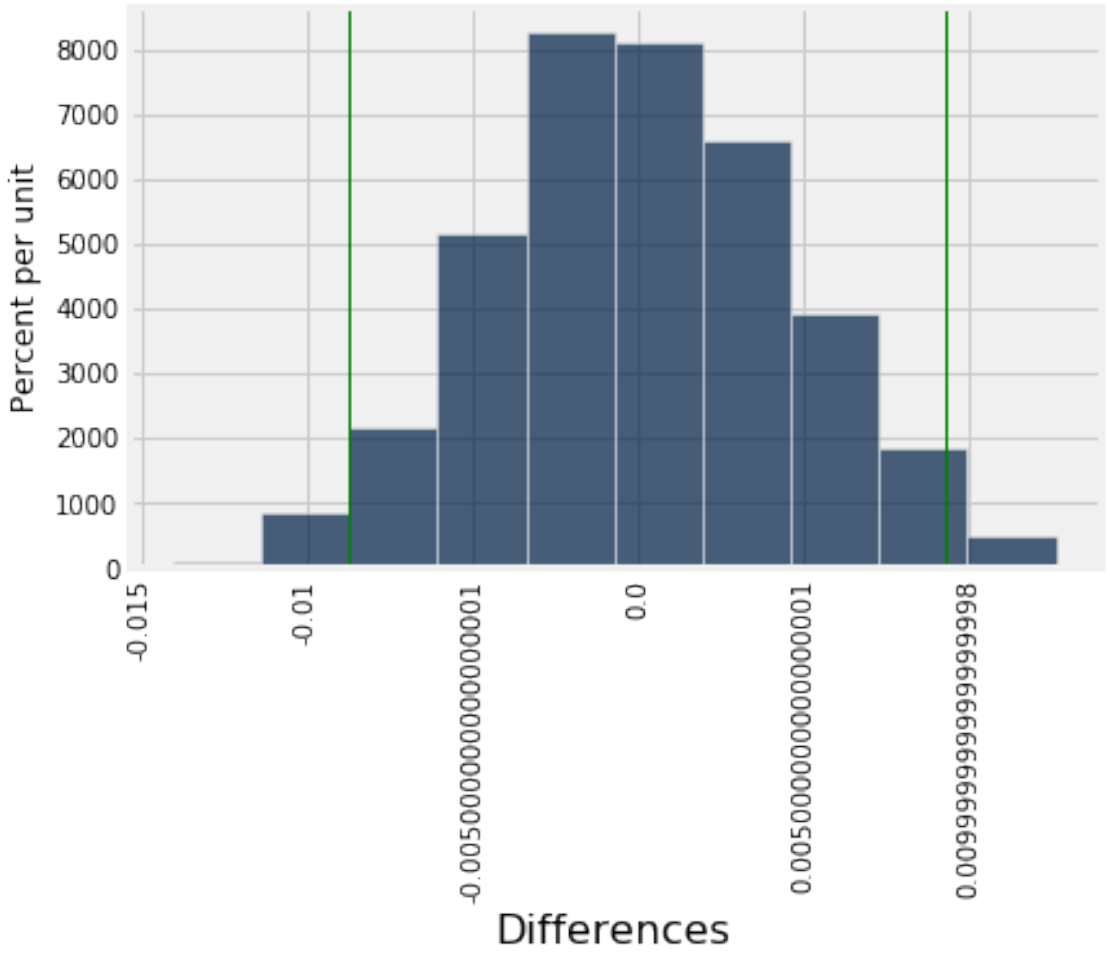

```
In [18]: #bootstrap hist 2017
         bootstrap_tbl.hist("Bootstrap 2017", unit="")
         p05_2017=Table().with_column('dind',bootstrap_tbl.column("Bootstrap 2017")).percentile
         p95_2017=Table().with_column('dind',bootstrap_tbl.column("Bootstrap 2017")).percentile
         plt.axvline(x=p05_2017,color='green',linewidth=1)
         plt.axvline(x=p95_2017,color='green',linewidth=1)

         left_2017 = percentile(2.5, bootstrap_tbl.column("Bootstrap 2017"))
         right_2017 = percentile(97.5, bootstrap_tbl.column("Bootstrap 2017"))
         print("The 95% confidence interval for the difference between the two data sets is:",
              left_2017 , "and", right_2017)
```

```
The 95% confidence interval for the difference between the two data sets is: -0.00547117223626
```




```
In [19]: #bootstrap hist differences
         bootstrap_tbl.hist("Differences", unit="")
         #confidence interval

         p05_diff=Table().with_column('dind',bootstrap_tbl.column("Differences")).percentile(2
         p95_diff=Table().with_column('dind',bootstrap_tbl.column("Differences")).percentile(9
         plt.axvline(x=p05_diff,color='green',linewidth=1)
         plt.axvline(x=p95_diff,color='green',linewidth=1)

         left_diff = percentile(2.5, bootstrap_tbl.column("Differences"))
         right_diff = percentile(97.5, bootstrap_tbl.column("Differences"))
         print("The 95% confidence interval for the difference between the two data sets is:",
               left_diff , "and", right_diff)
```

```
The 95% confidence interval for the difference between the two data sets is: -0.00878061593392
```



As we can see with this 95% confidence interval for the difference between average wage of African Americans in the labor force and percent of Latinos in the labor force in 2000 and 2017, 0 is within the confidence interval. This would mean that I fail to reject the null hypothesis, and conclude that there is no difference between the effects of Latino immigration on African American labor and wages in 2000 and 2017.

Now it's time to think about why this might be possible. First of all, for the 2000 and 2017 data themselves we see that the slopes are very close to 0, indicating that there is almost no correlation between the two variables I chose; choosing two slightly different variable may have made a higher correlation more evident and created more significant results. Perhaps one of the factors I listed earlier that could have made my test statistic prone to bias was true and did affect the end result of this test. It is possible that during the time period, the niches of job sectors that Latinos and African Americans occupied were different, meaning that the increased immigration of Latinos would not have made a difference at all. It may be that the metropolitan areas used happened to have either higher concentrations of Latinos or African Americans, which would affected the results of the differences.

Either way, I believe that in order to come to a decisive conclusion, more data extracts and tests would need to be done, possibly factoring for more of the biases that I mentioned. Based on my own results, I conclude that there was practically no difference between the Latino immigration

effect on African American labor and wages in 2000 and 2017.