Cultural Analytics

ENGL 64.05

Fall 2019

Prof. James E. Dobson

Cultural Analytics

Probability and Statistics

- Basic statistics (mode, mean, median, standard deviation, etc) are needed to understand the distribution of data in our datasets.
- Hypothesis testing is not presently widely used in cultural analytics.
- These tests (chi-squared, t-test, z-test, ANOVA, etc) can help us determine if our results are valid.
- We'll use probability functions, including one based on Bayes's rule, to help *classify* texts or sections of text.

Narrative and Statistics

“But undue emphasis on the damned lies can obscure the extraordinary power of statistical methods. If they are good at deceiving, they are likewise good at yielding insights in fields ranging from biology to linguistics. The connection between statistics and narrative may well be the most important point of intersection between mathematics and the humanities. And in the case, it may be that the former has as much to glean from the latter as the other way around” (Juola and Ramsay, 195).

Reporting Confidence

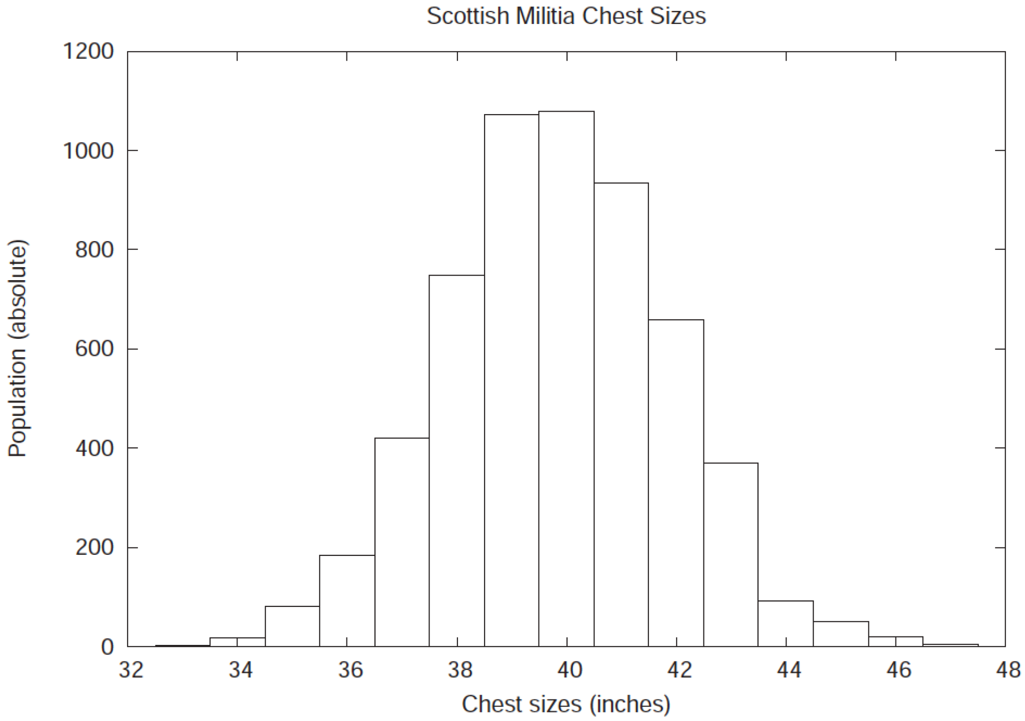
- For many experiments, we'll report our confidence in the conclusion / classification.
- A likelihood of 95% is pretty damn confident.
- The t-test returns a p-value, and the gold standard is 0.05.

Student's T-Test

- Used to compare and determine how different two datasets/populations might be.
- Stats packages (like Scipy) return t scores and p-values.
  - Larger **t score**, greater difference; smaller score, smaller difference.
    - The t score tells us how many times different datasets are from each other as within each other.
  - Larger **p-value**, greater likelihood the difference appears by chance.
- p-value (probability) is reported as percentage: 0.05 = 95% likely that this wasn't a chance difference.
- Both returned values are important. We want to know if things are different and if that difference is significant.

Guinness brewery chemist William Sealy Gosset aka "Student"

The Normal Distribution?



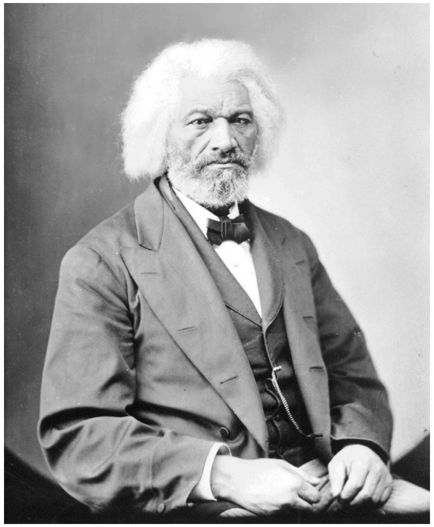
“Lots of things—especially biological and social things with complex causes—distribute in this way. For this reason, the distribution has come be called the normal distribution and its distinctive shape, the bell curve” (218-220).

Descriptive Statistics

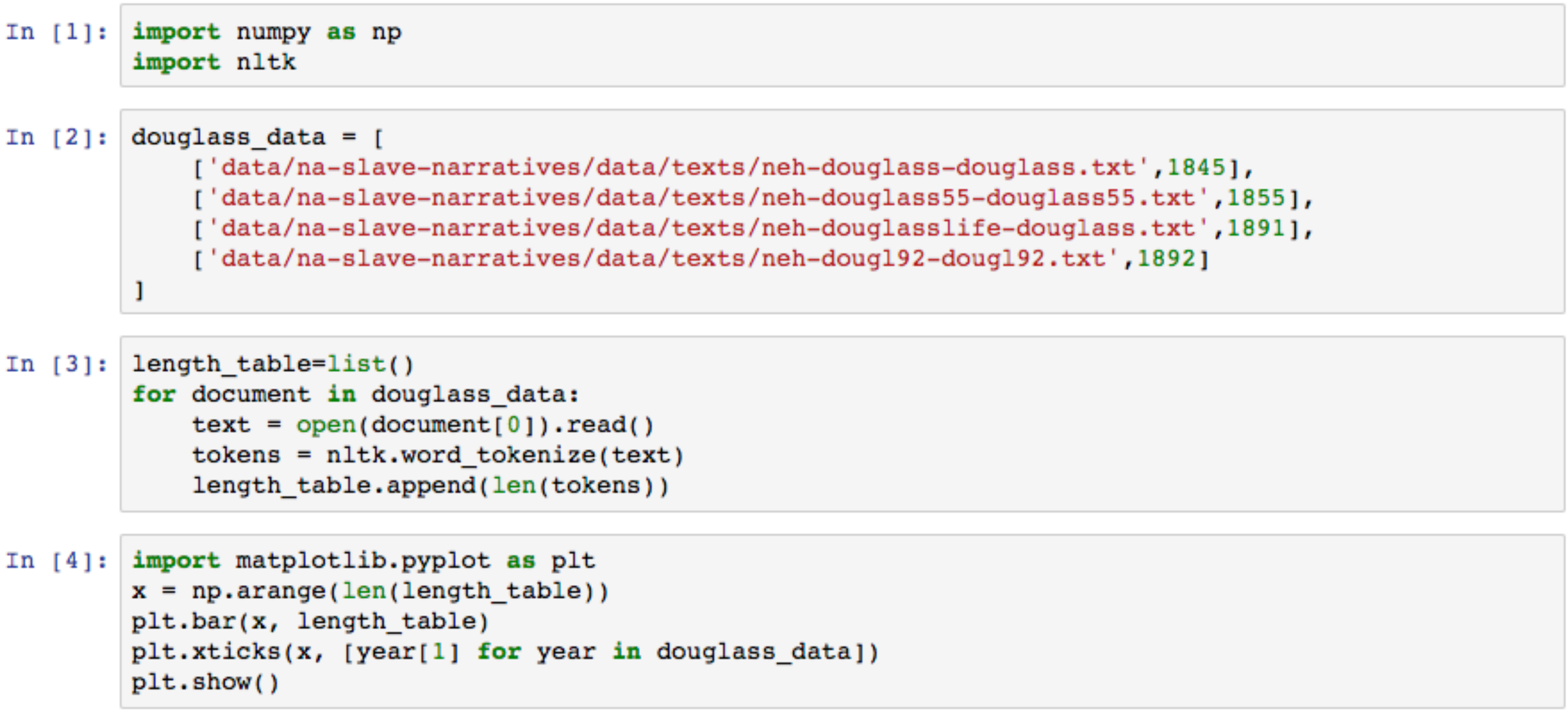
- **Average/mean**: approximate measure of the center of distribution.
- **Mode**: most frequently occurring measure/value/etc.
- **Median**: the value at which half the values are higher and half lower.
- **Standard Deviation**: measurement of variance found within the data.

Frederick Douglass

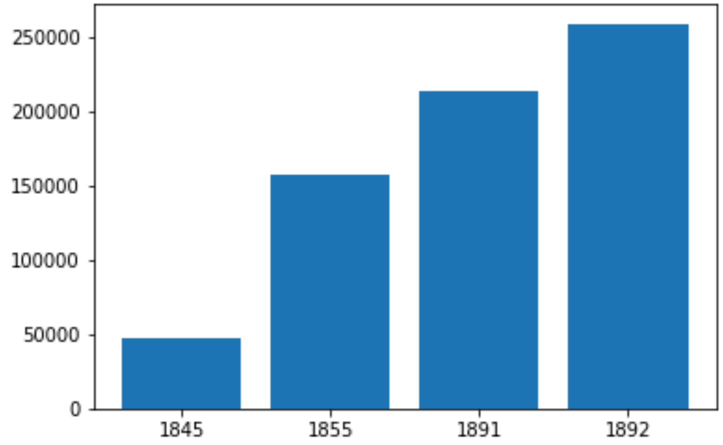
- Writes first autobiographical narrative (*Narrative of the Life of Frederick Douglass, an American slave, written by himself*) in 1845.
- Publishes another narrative in 1855 (revising first).
- Publishes yet another in 1891.
- Shortly after, produces a second revision of previous in 1892.



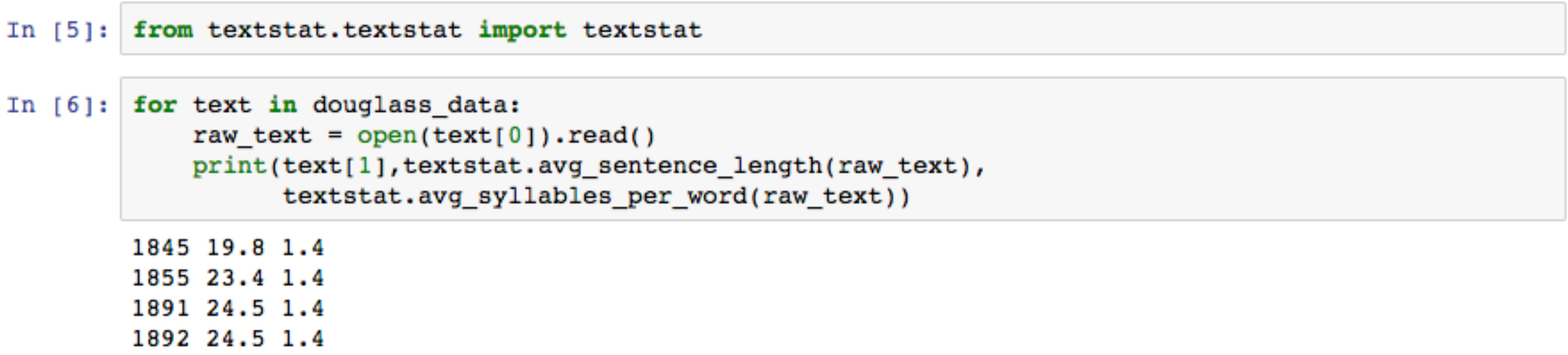
What Happens?



```
In [1]: import numpy as np
        import nltk

In [2]: douglass_data = [
            ['data/na-slave-narratives/data/texts/neh-douglass-douglass.txt',1845],
            ['data/na-slave-narratives/data/texts/neh-douglass55-douglass55.txt',1855],
            ['data/na-slave-narratives/data/texts/neh-douglasslife-douglass.txt',1891],
            ['data/na-slave-narratives/data/texts/neh-dougl92-dougl92.txt',1892]
        ]

In [3]: length_table=list()
        for document in douglass_data:
            text = open(document[0]).read()
            tokens = nltk.word_tokenize(text)
            length_table.append(len(tokens))

In [4]: import matplotlib.pyplot as plt
        x = np.arange(len(length_table))
        plt.bar(x, length_table)
        plt.xticks(x, [year[1] for year in douglass_data])
        plt.show()
```

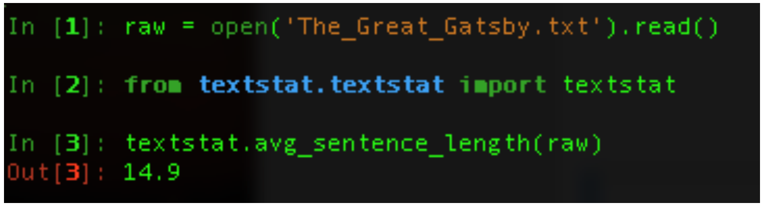


Does His Style Change?



```
In [5]: from textstat.textstat import textstat

In [6]: for text in douglass_data:
            raw_text = open(text[0]).read()
            print(text[1],textstat.avg_sentence_length(raw_text),
                  textstat.avg_syllables_per_word(raw_text))

        1845 19.8 1.4
        1855 23.4 1.4
        1891 24.5 1.4
        1892 24.5 1.4
```

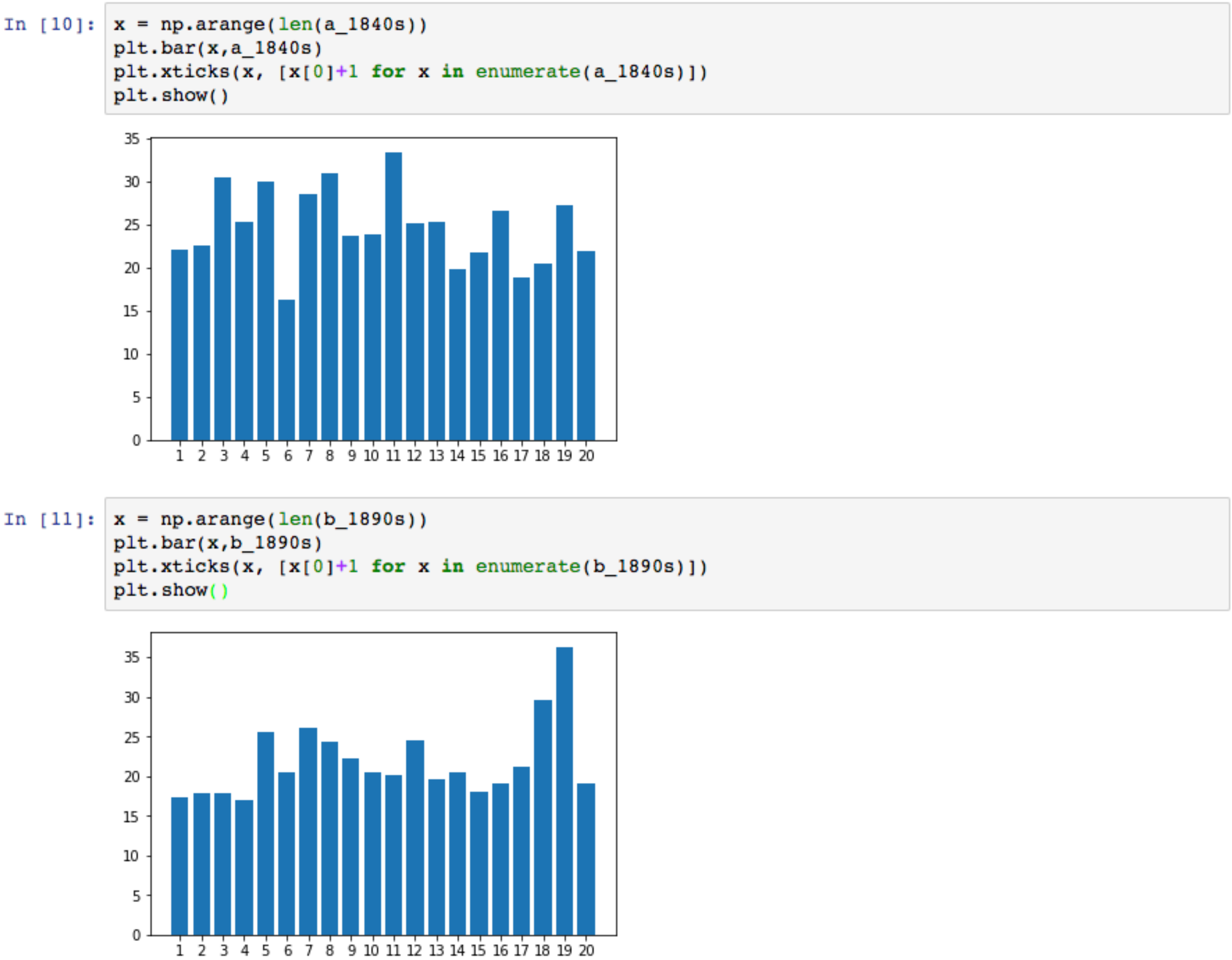
- Is this a good enough measure of "style?"
- What else might we want to measure?
- What could be the cause of the change?



```
In [1]: raw = open('The_Great_Gatsby.txt').read()

In [2]: from textstat.textstat import textstat

In [3]: textstat.avg_sentence_length(raw)
Out[3]: 14.9
```

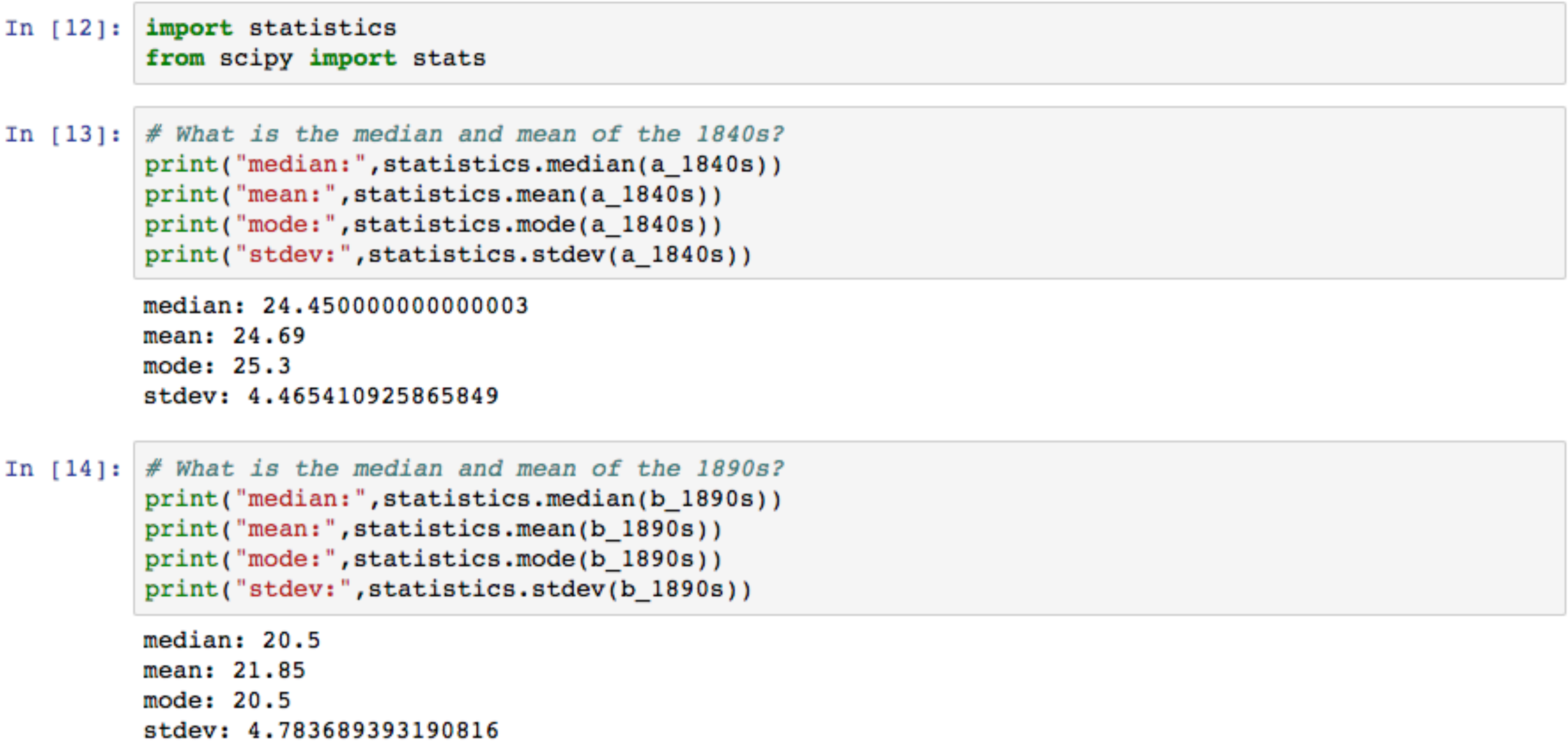
Cultural Analytics

Sampling: 1840s and 1890s



```
In [10]: x = np.arange(len(a_1840s))
         plt.bar(x,a_1840s)
         plt.xticks(x, [x[0]+1 for x in enumerate(a_1840s)])
         plt.show()
```

```
In [11]: x = np.arange(len(b_1890s))
         plt.bar(x,b_1890s)
         plt.xticks(x, [x[0]+1 for x in enumerate(b_1890s)])
         plt.show()
```



In [12]:

```
import statistics
from scipy import stats
```

In [13]:

```
# What is the median and mean of the 1840s?
print("median:",statistics.median(a_1840s))
print("mean:",statistics.mean(a_1840s))
print("mode:",statistics.mode(a_1840s))
print("stdev:",statistics.stdev(a_1840s))
```

```
median: 24.450000000000003
mean: 24.69
mode: 25.3
stdev: 4.465410925865849
```

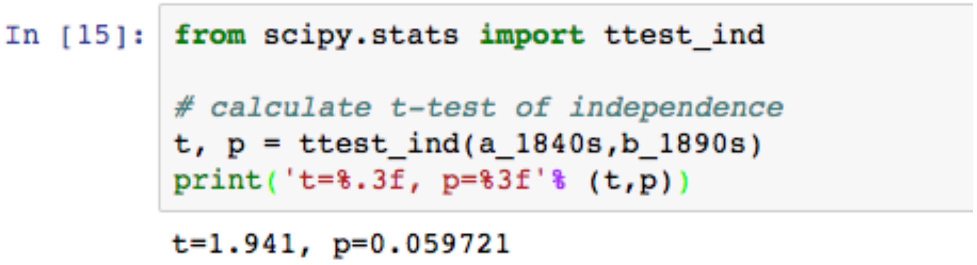
In [14]:

```
# What is the median and mean of the 1890s?
print("median:",statistics.median(b_1890s))
print("mean:",statistics.mean(b_1890s))
print("mode:",statistics.mode(b_1890s))
print("stdev:",statistics.stdev(b_1890s))
```

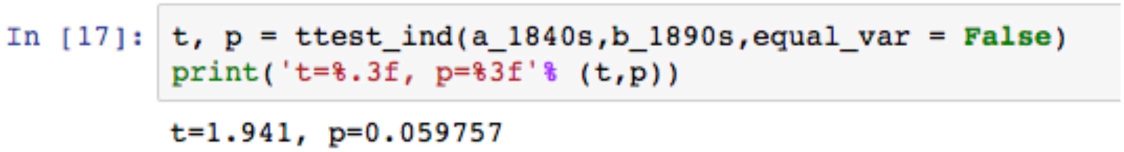
```
median: 20.5
mean: 21.85
mode: 20.5
stdev: 4.783689393190816
```

t-test of difference in sentence length

```
help(ttest_ind)
This is a two-sided test for the null hypothesis that 2
independent samples have identical average (expected)
values. This test assumes that the populations have
identical variances by default.
```



```
In [15]: from scipy.stats import ttest_ind

         # calculate t-test of independence
         t, p = ttest_ind(a_1840s,b_1890s)
         print('t=%.3f, p=%3f'% (t,p))

         t=1.941, p=0.059721
```



```
In [17]: t, p = ttest_ind(a_1840s,b_1890s,equal_var = False)
         print('t=%.3f, p=%3f'% (t,p))

         t=1.941, p=0.059757
```

- The sentence lengths used in the 1840s are **1.94 times** as different from the 1890s as they are within each period.
- And we can say that with **95% confidence** that this did appear by chance.

Next Class

- Adrian Mckenzie, "Vectorization and Its Consequences."
- Focus on the historical move from tables to vectors.