America’s Nicotine Problem

(name redacted)
12/15/17

Abstract

Hypothesis tests are a great way of predicting differences between samples and populations. I decided to apply one test to a big characteristic of our economy today: tax collections. What not better to study tax collections than those of tobacco products, merchandise that bring in billions of tax dollars a year? The goal of this particular study is to determine whether the annual mean cigarette tax collections from years 1985 to 2014 is greater than the same years' annual mean tax collections from other tobacco products, such as chewing tobacco, hookah, cigars, and many more. These two groups of products will be the populations we're testing today, with p_1 being cigarettes and p_2 being other tobacco products.

Body

INTRODUCTION

Since its first commercial planting in the United States, tobacco has taken over the country. It has been publicly dubbed the official leading cause of preventable death, yet still sells millions of products every day. The amount of money being spent on these poisonous products is incredibly high and troublesome, especially considering the fact that all of the buyers are well aware of the dangers and health effects. For this reason, I decided to crunch the numbers to see how much money the country really is spending on these and how much of that cash goes back into the country's system, and to hopefully calculate some numbers that will be so extreme it's appalling and discouraging to those who choose to participate in the worldwide tobacco pandemic that they'll quit. I predict that a right-tailed test will conclude that the mean annual tax collections from cigarettes is greater than that of other tobacco products ($h_a : \mu_1 > \mu_2$).

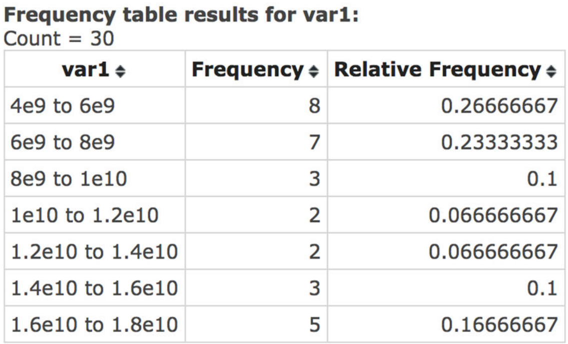
DATA COLLECTION METHODOLOGY

Collecting data for this project was surprisingly easy because the government and some of its organizations release their official tax reports yearly. There are also other people like me who are interested in the subject and wish to publish their studies to jolt the readers with the realization of how crazy these numbers are, so there are lots of compilations and studies accessible by the general public. The study I chose to supply my data is “The Tax Burden on Tobacco” by the Tobacco Tax Council whose report I found online. This historical compilation gives data on tobacco-related tax revenues for each year from 1865 to 2014, along with several graphs and maps to provide visual aid. The compilation’s table that I pulled my numbers from gives us the annual tax collections from every year since 1865 which guaranteed that my data are random because each year was equally as likely to be included in the samples as the next. The only tinkering I had to do with the numbers was converting them because the report gave the data in thousands of dollars, so I multiplied each observation by 1000 to get the actual dollar amount.

DATA ANALYSIS

Cigarettes

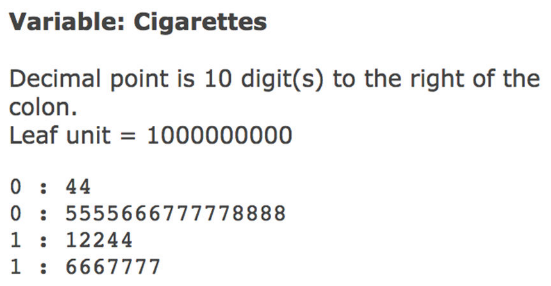
- **Mean ≈** $9,707,825,300
- **Standard deviation ≈** $4,547,429,800
- **Five number summary** = $4,314,268,000 , $5,924,192,000 , $7,939,764,500 , $14,285,200,000 , $17,107,803,000
- **Range** = $1,279,353,500
- **Frequency table** =



Frequency table results for var1:
Count = 30

var1 ⬍	Frequency ⬍	Relative Frequency ⬍
4e9 to 6e9	8	0.26666667
6e9 to 8e9	7	0.23333333
8e9 to 1e10	3	0.1
1e10 to 1.2e10	2	0.066666667
1.2e10 to 1.4e10	2	0.066666667
1.4e10 to 1.6e10	3	0.1
1.6e10 to 1.8e10	5	0.16666667

- **Stem & leaf plot** =

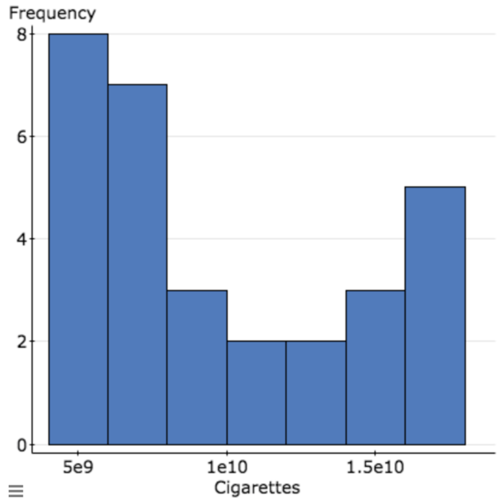


Variable: Cigarettes

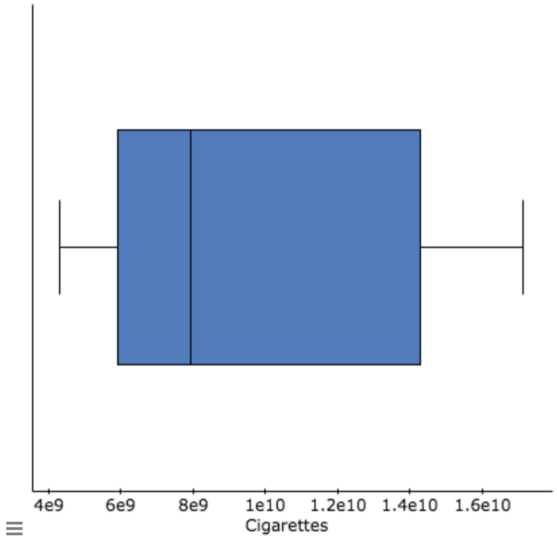
Decimal point is 10 digit(s) to the right of the colon.
Leaf unit = 1000000000

```
0 : 44
0 : 5555666777778888
1 : 12244
1 : 6667777
```

- **Histogram** =



- **Boxplot** =

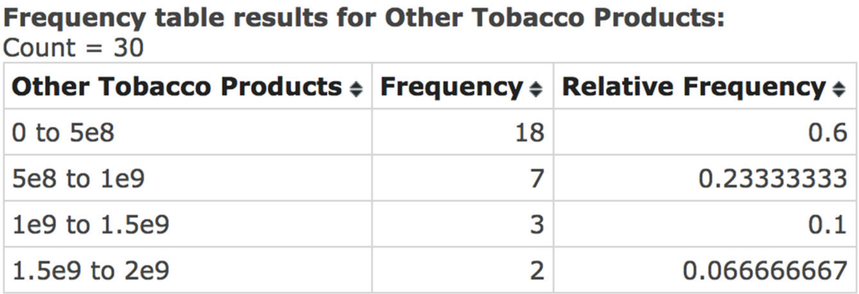


- 56.67% of the observations lie within 1 standard deviation of the mean
    - 70% lie within 2 standard deviations
    - 100% lie within 3 standard deviations

- **Conclusions** = As seen in the 5 number summary, the intervals of observations spread wider and wider as they go on so the rate of people buying cigarettes is getting higher, though we should also consider population growth and economic inflation. We can see from our histogram that our data is right-skewed which is supported by our relative frequency table showing that the higher numbers occurred less often, also due to the same reason of more people buying cigarettes now than ever and the rate is only increasing. The histogram does show a little dip in the curve, however, and my only guess with that is that everybody used to smoke but cigarettes didn't cost much, then the news came out that they're unhealthy so a lot of people quit (that's the dip), and then the price of them increased so the people still smoking are bringing in more money to the state. We also may be to a certain point in our country's state of health where a lot of people just stopped caring and went back to smoking. According to our calculated means of approximately 9.7 billion and 574.7 million dollars, I hypothesized correctly: cigarettes bring in more tax money to the states than other tobacco products. Our percentages that lie between standard deviations are interesting when compared to the second population: ours are around 60%, 70%, and 100%, while the other data set's are 70%, 90%, and 100%. I think that this is because cigarettes have been more popular for much longer so there are more numbers on them but I'm not really sure.

Other tobacco products

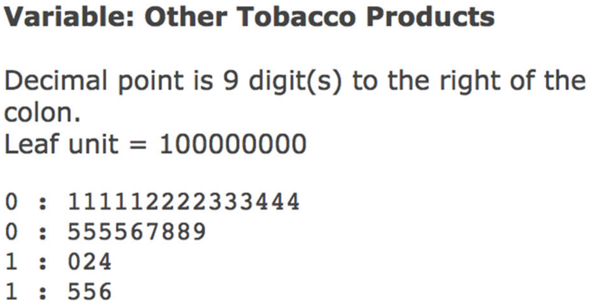
- **Mean ≈** $574,675,430
- **Standard deviation ≈** $463,767,070
- **Five number summary** = $61,286,000 , $207,735,000 , $444,841,000 , $802,488,000 , $1,583,603,000
- **Range** = $1,522,317,000
- **Frequency table** =



Frequency table results for Other Tobacco Products:
Count = 30

Other Tobacco Products ⬍	Frequency ⬍	Relative Frequency ⬍
0 to 5e8	18	0.6
5e8 to 1e9	7	0.23333333
1e9 to 1.5e9	3	0.1
1.5e9 to 2e9	2	0.066666667

- **Stem & leaf plot** =

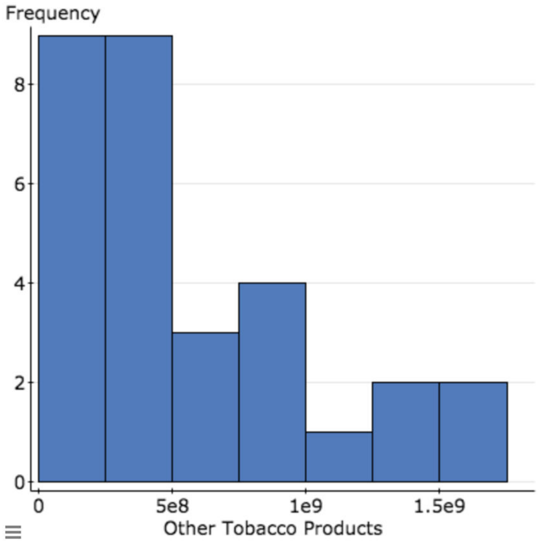


Variable: Other Tobacco Products

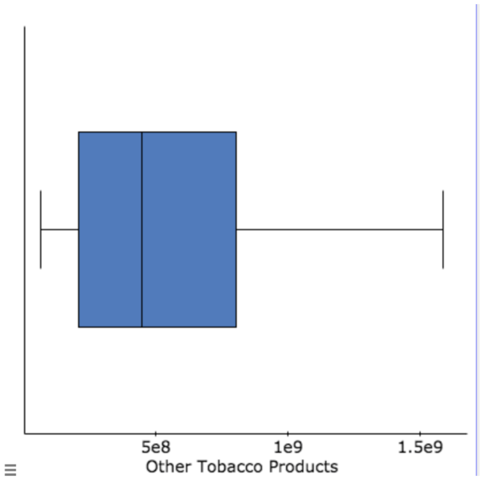
Decimal point is 9 digit(s) to the right of the colon.
Leaf unit = 100000000

```
0 : 111112222333444
0 : 555567889
1 : 024
1 : 556
```

- **Histogram** =



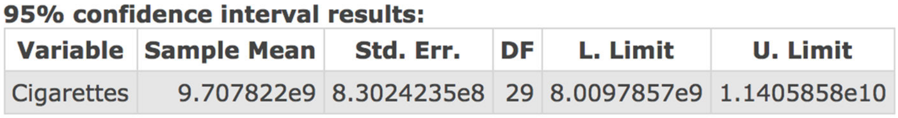
- **Boxplot** =



- 70% of the observations lie within 1 standard deviation of the mean
  - 93.33% lie within 2 standard deviations
  - 100% lie within 3 standard deviations
- **Conclusions** = Similar to our first population, the state tax collections from other tobacco products have been increasing. Our histogram shows that this set of data is also right-skewed as our first population but without the dip, so the increase has been consistent and more linear than p_1.

INFERENTIAL STATISTICS

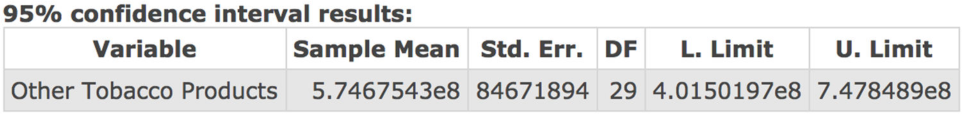
- **Confidence intervals** =
  - We are 95% confident that the mean dollar amount of one year's tax collections from cigarettes is between $8,009,785,700 and $14,058,580,000.



95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Cigarettes	9.707822e9	8.3024235e8	29	8.0097857e9	1.1405858e10

  - We are 95% confident that the mean dollar amount of one year's tax collections from other tobacco products is between $401,501,970 and $747,848,900.



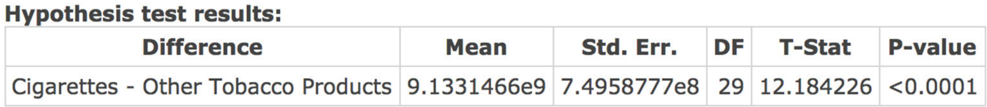
95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Other Tobacco Products	5.7467543e8	84671894	29	4.0150197e8	7.478489e8

- **Two-sample hypothesis** = I hypothesize that, at a 10% significance level, the mean dollar amount of one year's tax collections from cigarettes will be greater than that of other tobacco products.
    - $H_o : \mu_1 = \mu_2$
    - $H_a : \mu_1 > \mu_2$ (right tailed)
    - $\alpha = .10$
    - Non-pooled t test: $t = ({}_1 - {}_2)/\sqrt{(1/1) + (2/2)}$ w/ df = n - 1

      so, $t = (9{,}707{,}825{,}300 - 574{,}675{,}430)/$

      $\sqrt{(4{,}547{,}429{,}800/30) + (463{,}767{,}070/30)}$ w/ df = 29
    - Statcrunch results:



Hypothesis test results:

Difference	Mean	Std. Err.	DF	T-Stat	P-value
Cigarettes - Other Tobacco Products	9.1331466e9	7.4958777e8	29	12.184226	<0.0001

    - Now that statcrunch has given us a p-value of .0001 or lower, we can decide whether to reject our null hypothesis or not. If p ≤ a is true, then we should reject our null and vice versa. In this case, p ≤ a is true because .0001 ≤ .10. So, at a 10% significance level, the data provide sufficient evidence to conclude that the mean dollar amount of one year's tax collections from cigarettes is greater than that of other tobacco products.

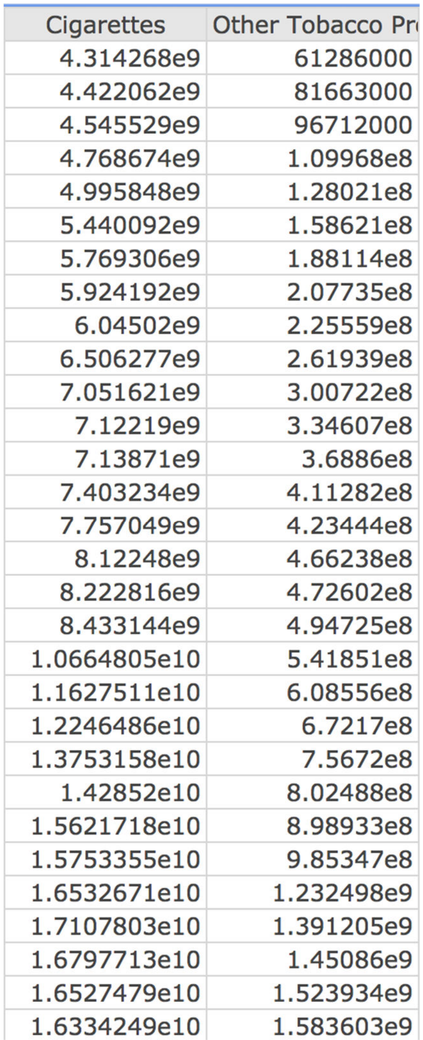
CONCLUSIONS

______Our goal today was to determine whether the mean tax collections from cigarettes or other tobacco products is greater. We performed a non-pooled t-test at a 10% significance level and concluded that the first population, cigarettes, was indeed greater. Both populations had means above 500 thousand, with cigarettes ranking at 9,707,825,300 dollars, and 574,675,430 for other tobacco products. Standard deviations were at 4,547,429,800 vs. 463,767,070. All in all, most observations and consequent data analyses were all higher for cigarettes than other tobacco products. I’m not surprised seeing as cigarettes have been more popular for a long time and the rate of purchases is still increasing. My hypothesis was therefore correct and the mean of all cigarette tax collections is higher than with other tobacco products.

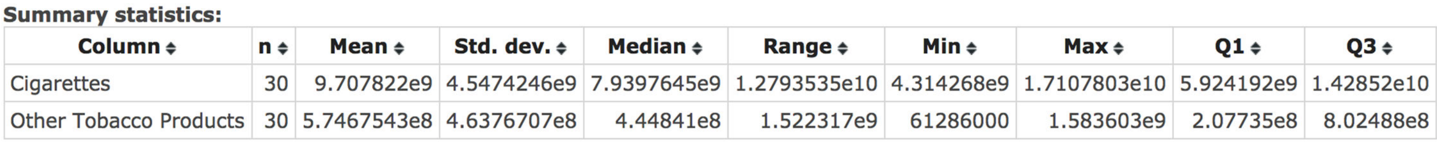
Sources

https://www.taxadmin.org/assets/docs/Tobacco/papers/tax_burden_2014.pdf - used table on page 5

Data

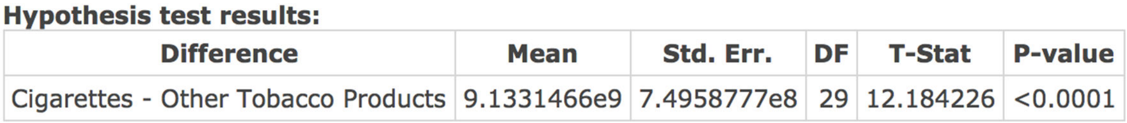


Cigarettes	Other Tobacco Pr
4.314268e9	61286000
4.422062e9	81663000
4.545529e9	96712000
4.768674e9	1.09968e8
4.995848e9	1.28021e8
5.440092e9	1.58621e8
5.769306e9	1.88114e8
5.924192e9	2.07735e8
6.04502e9	2.25559e8
6.506277e9	2.61939e8
7.051621e9	3.00722e8
7.12219e9	3.34607e8
7.13871e9	3.6886e8
7.403234e9	4.11282e8
7.757049e9	4.23444e8
8.12248e9	4.66238e8
8.222816e9	4.72602e8
8.433144e9	4.94725e8
1.0664805e10	5.41851e8
1.1627511e10	6.08556e8
1.2246486e10	6.7217e8
1.3753158e10	7.5672e8
1.42852e10	8.02488e8
1.5621718e10	8.98933e8
1.5753355e10	9.85347e8
1.6532671e10	1.232498e9
1.7107803e10	1.391205e9
1.6797713e10	1.45086e9
1.6527479e10	1.523934e9
1.6334249e10	1.583603e9



Summary statistics:

Column	n	Mean	Std. dev.	Median	Range	Min	Max	Q1	Q3
Cigarettes	30	9.707822e9	4.5474246e9	7.9397645e9	1.2793535e10	4.314268e9	1.7107803e10	5.924192e9	1.42852e10
Other Tobacco Products	30	5.7467543e8	4.6376707e8	4.44841e8	1.522317e9	61286000	1.583603e9	2.07735e8	8.02488e8



Hypothesis test results:

Difference	Mean	Std. Err.	DF	T-Stat	P-value
Cigarettes - Other Tobacco Products	9.1331466e9	7.4958777e8	29	12.184226	<0.0001