Poverty Rate and Smoking Prevalence

Bo Nappie

Statistics and Data Analytics

Introduction:

This research paper has been developed around the question, "Is there a relationship between poverty rate and smoking prevalence?". My parameter of interest is the differences in the proportion of smokers in a state's population where the poverty level is increased over a year, P_i, versus the proportion of smokers in a state's population where poverty levels decreased over a year, P_d, ($d=P_i - P_d$). This topic is of particular interest as both smoking and poverty are risk factors for a multitude of health problems. It is worth considering whether there is any correlation between the two. Insights can have a positive impact on society, future policies and interventions if needed.

Data Collection:

The data for this research topic has been curated from two datasets. One dataset from Infoplease, was used for data regarding the proportion of a population that was in poverty by state. The second data table had been pulled from Kaggle, which provided data on the prevalence of smokers in a population by state. Data from both tables that had accounted for the years 2005-2010 were combined to curate a new data table tailored to suit this study (See Appendix 1A for dataset). The population is the United States (including the District of Columbia). Each of the 51 states will act as one case. Thus, there are 51 cases in this study. The two variables, smoking rates in states where poverty increased, and smoking rates in states where poverty rates decreased are quantitative. The 2010 data will be analyzed. States have been grouped based on whether the poverty rate increased or decreased from 2005 to 2010.

Exploratory Data Analysis

There are two quantitative variables. One is the prevalence of smokers in states where poverty rates have increased. The other quantitative variable is the prevalence of smokers in states where poverty rates have decreased. The descriptive statistics shown below aim to provide a visually concise summary of the data's key features.

Summary Statistics

This five-number summary data analysis provides a visual of the distribution of data showing the 2010 data of the proportion of smokers in states where the poverty level increased from 2005 to 2010, compared to the proportion of smokers in states where the poverty level decreased from 2005 to 2010. The 2005 data (located in Appendix) serves as a baseline for contextual purposes when analyzing the 2010 data.

2010 Summary of Proportion of Smokers in States Where Poverty Increased

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.121	0.155	0.180	0.185	0.213	0.268

2010 Summary of Proportion of Smokers in States Where Poverty Decreased

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.091	0.150	0.169	0.167	0.185	0.237

Examining the two summaries for smokers in states where poverty rates increased, in 2005, the minimum smoking rate is 17.5% of the population of a state, while in 2010 the minimum value is 12.1% smoking rate in a state. In 2005, the 25th percentile value is 19.9%, while the 2010 value is 15.5%. The median value in 2005 is 21.5%, compared to 2010 where the

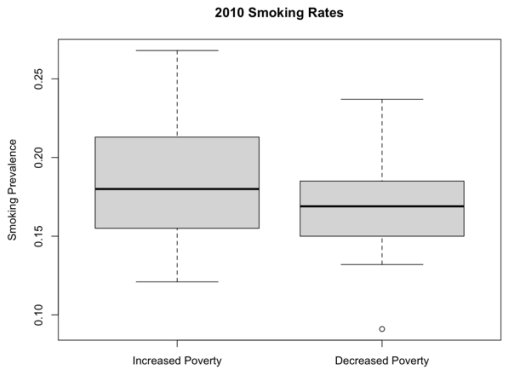
value is 18%. The 75th percentile value in 2005 is 22.6%, in 2010, it was 21.3%. In 2005, the maximum value is 26.7%, in 2010 it is 26.8%. The mean/average value for these states in 2005 is 21.4%, while in 2010 it is 18.5%. Subtracting the minimum from the maximum, we can compute the ranges for 2005 & 2010. In 2005 the range is, 26.7-17.5= *9.2%.* In 2010 the range is, 26.8-12.1=*14.7%.*

Looking at the two summaries for smokers in states where poverty rates decreased, in 2005, the minimum smoking rate is 11.5% of the population of a state, while in 2010 the minimum value is 9.1% smoking rate in a state. In 2005, the 25th percentile value is 18.7%, while the 2010 value is 15.0%. The median value in 2005 is 20.4%, compared to 2010 where the value is 16.5%. The 75th percentile value in 2005 is 21.5%, in 2010, it was 18.5%. In 2005, the maximum value is 27.2%, in 2010 it is 23.7%. The mean/average value for these states in 2005 is 20.2%, while in 2010 it is 16.7%. Subtracting the minimum from the maximum, we can compute the ranges for 2005 & 2010. In 2005 the range is, 27.2-11.5= *15.7%.* In 2010 the range is, 23.7-9.1=*14.6%.*

In summative, the 2005 & 10 data for smoking rates in states where the poverty levels increased were both higher than the data for smoking rates in states where the poverty levels decreased.

Visualization:

Below is a boxplot visual of the 2010 data of smoking rates in states where poverty rates increased, and smoking rates in states where poverty rates decreased. Boxplots are beneficial for identifying outliers, which have the potential to significantly impact data analysis. They also help better visualize the five number summaries, which include the minimum, maximum, median, and quartiles, and compare the differences between the two groups.

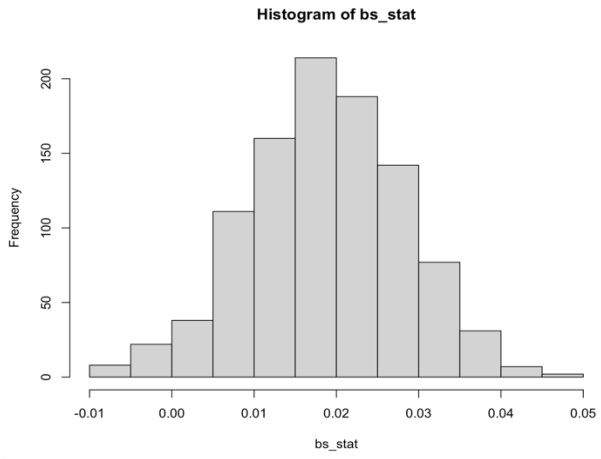


Upon analyzing the two groups, it is interesting to observe that among the two, the decreased poverty group has a minimum value of 9.1%, which has been deemed as so significantly different from the other values in the dataset that it is marked as an outlier (circle on the plot). Outliers have a notable influence on calculated mean values. It is apparent that there are no outliers within the increased poverty rate dataset, as there are no circles marked on the plot to indicate such a data value. The minimum value for the increased poverty group is 12.1%. The boxplots also reveal the increased poverty group has a median value of 18%, which is larger than the median value of the decreased poverty group, which is 16.9%. The 3rd quartile value is notably higher for the increased poverty group, 21.3% compared to 18.5%. Additionally, the maximum value is higher for the increased poverty group. Due to the outlier of the decreased poverty group, the range of smoking prevalence is nearly identical. The increased poverty group has a range of 14.7%, while the decreased poverty group has a range of 14.6%. With consideration to all of these observations, it can still be said that smoking rates are higher in states where poverty rates increased, compared to the smoking rate in states where poverty rates decreased.

Confidence Interval

It's necessary to first create a bootstrap distribution in order to generate a 95% confidence interval, which is useful for estimating the range of plausible values for the population parameter. Bootstrapping is a resampling technique that generates samples with replacement from the original sample, that stays the same size, and typically is resampled around 1000 times. Bootstrapping allows us to estimate the variability of the sample statistic. The primary parameter of interest is the difference, d, in the mean smoking prevalence among states where poverty rates have increased, P_i, versus the mean smoking prevalence in states where poverty levels decreased, P_d.

An empty vector will be initialized with intent to store to store the bootstrap statistics for the difference in means. Following, 1000 bootstrap samples will be generated by randomly sampling with replacement from the original samples, P_i and P_d. The bootstrap statistics are computed by taking the difference in means for each bootstrap sample. The resulting bootstrap distribution is of the bootstrap statistics for the difference in means.



The resulting bootstrap distribution is generally symmetric, and bell shaped. There are two possible ways of generating the 95% confidence interval. One way is to use the statistic +/- 2x the standard error, SE. The other way is to find the values at the 2.5^{th} and 97.5^{th} percentiles. The resulting 95% confidence interval for the mean difference in smoking prevalence between states with increasing poverty rates and states with decreasing poverty rates is (-.0005, .0371). The confidence interval is for the true difference in means between the two groups of states. With this information, we can be 95% confident that the true difference μ_d , of the mean smoking prevalence between states where poverty rates increased versus states where poverty rates decreased in 2010, is between -.05% and 3.71%. The mean difference is 1.884%, which is consistent with earlier findings that indicate that smoking rates are higher in states where poverty rates have increased.

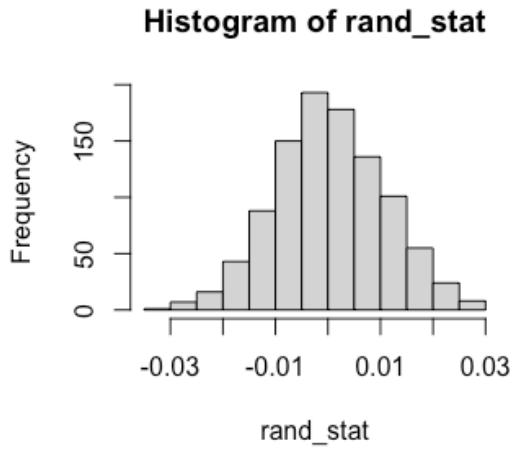
Hypothesis Test

A two tailed hypothesis test is useful for determining if the observed difference is statistically significant or if it is due to chance, at a 5% significance level. The null hypothesis, H_0, would mean that the mean difference between smoking prevalence in states where poverty rates increased, P_i, and smoking prevalence in states where poverty rates decreased, P_d, is zero; in other words, the mean smoking prevalence in states where poverty rates increased is equal to the mean smoking prevalence in states where poverty rates decreased, $\mu_d=0$. The alternative hypothesis, H_a would state that the mean difference between P_i and P_d is not zero; in other words, there is a difference in smoking prevalence among the two groups, $\mu_d \neq 0$.

The sample statistic was calculated by taking the difference between the mean smoking prevalence in states where poverty rates increased and the mean smoking prevalence in states where poverty rates decreased. The mean difference from the sample statistic is 1.867%. A

randomization distribution has been created by sampling with replacement to create 1000 random samples to simulate the null hypothesis, if there were no difference in smoking rates between states with increased poverty and states with decreased poverty. The random samples were based off of the original sample and sample size. The resulting randomization distribution, which has been visualized in the histogram below, is generally symmetric, bell shaped, and centered around zero.

Since the p-value of 0.04 is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is sufficient evidence of a statistically significant difference in smoking rates between states with increased poverty and states with decreased poverty.



Conclusion:

Considering the results from the analyzed data, it can be concluded that smoking rates are higher in states where poverty rates increased, compared to the smoking rate in states where poverty rates decreased, and we can reject the null hypothesis, H_0, which suggests that there is no difference between the two. These findings are significant as they suggest that poverty and smoking rates are associated and highlight the need for targeted interventions to reduce smoking

prevalence in high-poverty areas. Based off of the generated 95% confidence interval, we can be 95% confident that the true mean difference, μ_d, in smoking prevalence between states where poverty rates increased and decreased is between -.05% and 3.71%. Future research could explore potential confounding variables, such as education level or access to healthcare, to better understand the relationship between poverty and smoking rates.

Appendix 1: Additional Figures:

1A: Data table

	Year	State	Smoker	Nonsmoker	Poverty.Rate	trend
6	2010	Alabama	0.219	78.10%	0.173	increase
12	2010	Alaska	0.203	79.70%	0.124	increase
18	2010	Arizona	0.151	85.00%	0.186	increase
24	2010	Arkansas	0.229	77.10%	0.155	increase
30	2010	California	0.121	87.90%	0.163	increase
48	2010	Delaware	0.173	82.80%	0.121	increase
54	2010	District of Columbia	0.16	84.40%	0.199	increase
66	2010	Georgia	0.176	82.40%	0.187	increase
72	2010	Hawaii	0.145	85.50%	0.121	increase
84	2010	Illinois	0.169	83.10%	0.141	increase
90	2010	Indiana	0.213	78.80%	0.163	increase
102	2010	Kansas	0.17	83.00%	0.143	increase
108	2010	Kentucky	0.248	75%	0.177	increase
114	2010	Louisiana	0.221	78%	0.216	increase
126	2010	Maryland	0.152	84.80%	0.108	increase
132	2010	Massachusetts	0.141	85.90%	0.106	increase
138	2010	Michigan	0.189	81.00%	0.155	increase
150	2010	Mississippi	0.229	77%	0.227	increase
156	2010	Missouri	0.21	78.90%	0.148	increase
174	2010	Nevada	0.214	78.60%	0.164	increase
192	2010	New Mexico	0.185	81.60%	0.186	increase
198	2010	New York	0.155	84.60%	0.16	increase
204	2010	North Carolina	0.197	80.30%	0.174	increase
210	2010	North Dakota	0.173	82.70%	0.122	increase
216	2010	Ohio	0.225	77.50%	0.153	increase
228	2010	Oregon	0.15	84.90%	0.142	increase
234	2010	Pennsylvania	0.184	81.60%	0.122	increase
240	2010	Rhode Island	0.157	84.30%	0.136	increase
246	2010	South Carolina	0.21	79.00%	0.17	increase
252	2010	South Dakota	0.154	85%	0.132	increase
258	2010	Tennessee	0.201	79.90%	0.167	increase
264	2010	Texas	0.158	84.20%	0.184	increase
264	2010	Texas	0.158	84.20%	0.184	increase
276	2010	Vermont	0.153	84.70%	0.108	increase
294	2010	West Virginia	0.268	73.20%	0.169	increase
36	2010	Colorado	0.16	84.00%	0.122	decrease
42	2010	Connecticut	0.132	86.80%	0.083	decrease
60	2010	Florida	0.17	82.80%	0.16	decrease
78	2010	Idaho	0.157	84.40%	0.14	decrease
96	2010	Iowa	0.162	83.80%	0.103	decrease
120	2010	Maine	0.182	81.80%	0.125	decrease
144	2010	Minnesota	0.149	85.10%	0.105	decrease
162	2010	Montana	0.188	81.20%	0.14	decrease
168	2010	Nebraska	0.172	82.80%	0.102	decrease
180	2010	New Hampshire	0.169	83.10%	0.066	decrease
186	2010	New Jersey	0.144	85.50%	0.107	decrease
222	2010	Oklahoma	0.237	76.30%	0.163	decrease
270	2010	Utah	0.091	90.90%	0.1	decrease
282	2010	Virginia	0.185	81.40%	0.107	decrease
288	2010	Washington	0.15	84.80%	0.115	decrease
300	2010	Wisconsin	0.191	80.90%	0.099	decrease
306	2010	Wyoming	0.195	80.50%	0.096	decrease

2005 Summary of Proportion of Smokers in States Where Poverty Increased

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.1750	0.1990	0.215	0.214	0.226	0.267

2005 Summary of Proportion of Smokers in States Where Poverty Decreased

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.115	0.187	0.204	0.202	0.215	0.272

Appendix 2:

Citations:

"Percent of People in Poverty, by State, 2002–2010." *Infoplease*, Infoplease, https://www.infoplease.com/business/poverty-income/percent-people-poverty-state-2002-2010.

"U.S. Tobacco Use Data." *Kaggle*, 24 Jan. 2023, https://www.kaggle.com/datasets/thedevastator/u-s-tobacco-use-data-1995-2010?resource=download.

Appendix Code:

FiveNumberSummary:

```
>View(FINAL.PROJECT.DATA)

> incPoverty<- subset(FINAL.PROJECT.DATA, FINAL.PROJECT.DATA$Year == 2010 &

FINAL.PROJECT.DATA$Poverty.Rate >

FINAL.PROJECT.DATA[FINAL.PROJECT.DATA$Year ==2005,]$Poverty.Rate)

> incPoverty_fivenum <- fivenum(incPoverty$Smoker)

> incPoverty_fivenum

[1] 0.121 0.155 0.180 0.213 0.268

> mean(incPoverty$Smoker)

[1] 0.1853824

> decPoverty<- subset(FINAL.PROJECT.DATA, FINAL.PROJECT.DATA$Year == 2010 &

FINAL.PROJECT.DATA$Poverty.Rate <

FINAL.PROJECT.DATA[FINAL.PROJECT.DATA$Year == 2005,]$Poverty.Rate)
```

```
> View(decPoverty)
> decPoverty_fivenum<- fivenum(decPoverty$Smoker)
> decPoverty_fivenum
[1] 0.091 0.150 0.169 0.185 0.237
> mean(decPoverty$Smoker)
[1] 0.1667059
incPoverty_2005_fivenum<- fivenum(incPoverty_2005$Smoker)
> incPoverty_2005_fivenum
[1] 0.1750 0.1990 0.2145 0.2260 0.2670
> mean(incPoverty_2005$Smoker)
[1] 0.2144091
> decPoverty_2005<- subset(FINAL.PROJECT.DATA, FINAL.PROJECT.DATA$Year ==
2005 & FINAL.PROJECT.DATA$Poverty.Rate <
FINAL.PROJECT.DATA[FINAL.PROJECT.DATA$Year == 2005,]$Poverty.Rate)
> decPoverty_2005_fivenum<- fivenum(decPoverty_2005$Smoker)
> decPoverty_2005_fivenum
[1] 0.115 0.187 0.204 0.215 0.272
> mean(decPoverty_2005$Smoker)
[1] 0.2022222
> difference<- incPoverty$Smoker - decPoverty$Smoker
> diff_fivenum<- fivenum(difference)
> diff_fivenum
[1] -0.067 -0.012  0.014  0.059  0.157
```

```
> mean(difference)

[1] 0.01867647
```

BoxPlot

```
>boxplot(incPoverty$Smoker, decPoverty$Smoker,

+        main="2010 Smoking Rates", ylab="Smoking Prevalence",

+        names=c("Increased Poverty", "Decreased Poverty"))
```

Bootstrap

```
>spli<-incPoverty$Smoker

> spld<-decPoverty$Smoker

for(ii in 1:bs_n){

+ bs_spli<-sample(spli,replace=TRUE)

+ bs_spld<-sample(spld,replace=TRUE)

+ bs_stat[ii]<-mean(bs_spli)-mean(bs_spld)}

> hist(bs_stat)
```

95% Confidence Interval

```
lower<- quantile(bs_stat, .025)

> upper<- quantile(bs_stat, .975)

> lower

        2.5%

-0.0005308824

>

> upper
```

97.5%

0.03709559

Sample Statistic

incmean<- mean(incPoverty$Smoker)

> decmean<- mean(decPoverty$Smoker)

> rstat<- incmean-decmean

> rstat

[1] 0.01867647

Randomization Distribution

all_data<- c(incdata, decdata)

+ rand_data<-sample(all_data)

+ rand_inc<- rand_data[1:length(incdata)]

+ rand_dec<- rand_data[(length(incdata)+1):length(rand_data)]

+ rand_stat[i]<- mean(rand_inc)

+ rand_stat[i]<- mean(rand_inc)-mean(rand_dec)

+ }

> hist(rand_stat)

> p_valr <- sum(rand_stat >= abs(rstat)) / rand_n

> print(p_valr)

[1] 0.04