Paper one

Course: Stat2180W

Name: Emma yang

Date: September 30, 2020

Instructor: Joshua spark

**Introduction**

Democratic government listen to its citizens is most basic standard requirement. The most basic criterion for a democratic government is to listen to the citizen. However, only public opinion can be reflected in specific policies can it be of practical significant, which is called “responsiveness” of the government. In American politics, this can be done through two methodologies. First is mechanism elections. Firstly, citizens select their policy preferences political parties. And then, political parties make policies that reflect public selection. Second is skill can be described as “absorption”, that is, governmental parties take the initiative to make corresponding adjustments in policies based on their judgment of public opinion. In other words, public opinion can influence policy even when there is no party rotation.

In 2008, the organization named American National Election Study (AMES) researched approach 1800 respondents on various facets of their identity in order to excavate patterns in the voting behavior of the United States. The survey variables is as follows:

|  |  |  |
| --- | --- | --- |
| Variable | Label | Values |
| Id | Id Number |  |
| Region | Census Region | 1 = Northeast |
|  |  | 2 = Midwest |
|  |  | 3 = South |
|  |  | 4 = West |
| Ideology | Liberal/conservative self-placement | 0 = No Answer |
|  |  | 1 = Extremely Liberal |
|  |  | 2 = Liberal |
|  |  | 3 = Slightly Liberal |
|  |  | 4 = Moderate |
|  |  | 5 = Slightly Conservative |
|  |  | 6 = Conservative |
|  |  | 7 = Extremely Conservative |
|  |  | 8 = Don’t Know |
|  |  | 9 = Refused |
| Ethnic | Race/ethnicity of Respondent | 1 = Black |
|  |  | 2 = Asian |
|  |  | 3 = Native American |
|  |  | 4 = Hispanic |
|  |  | 5 = White |
|  |  | 6 = Other |
|  |  | 7 = Don’t Know |
|  |  | 8 = Refused |
|  |  | 9 = No Answer |
| Gender | Respondent gender | 1 = Man |
|  |  | 2 = Woman |

Based on the result of this research, the company named “PolitiTACT” wants to know, what effect various categorical factors have on political ideology, with the intent on utilizing such patterns to revise public policy, one facet including whether specific claims widely assumed to be true have any substantive merit. The details about the goal of “PolitiTACT” will be introduced in next part. This paper will use statistic method to help “PolitiTACT” firm to know their problem. This report will illustrate following three claims:

Claim 1: 60% of the population identifies as politically moderate; 20% of the population identifies as politically liberal; and 20% of the population identifies as politically conservative.

Claim 2: Southerners and Midwesterners have a different ideological makeup than Northeasterners and Westerners.

Claim 3: Men are more likely than women to identify with the conservativism.

Firstly, this report will describe the report background and terminology. And then, this paper will introduce the methodology we used in this survey analysis. Finally, this paper will show our survey results and do analysis.

**Statistical Background & Terminology**

The survey result data of American Election Study in 2008 is including five attributes: Id, ideology, ethnic, and gender. Id denotes different individuals number; ideology is a survey questions which including 10 different answers: 0 denotes no answer, 1 denotes extremely liberal, 2 represents liberal, 3 represents slightly liberal, 4 denotes moderate, 5 denotes slightly conservative, 6 represents conservative, 7 means extremely conservative, 8 means do not know, 9 means refused; ethnic is one of survey question which including nine different answers: 1 mean Black, 2 means Asian, 3 denotes Native American, 4 denotes Hispanic, 5 represents White, 6 means Other, 7 denotes do not know, and 8 means refused; gender is also a survey question which including 2 answers: 1 for male and 2 for female.

In order to analyze American National Election Study survey data and get all results for three claims, we will use R programming language to do statistic analysis. The statistical methods we used is as follows:

* Hypothesis testing:

Hypothesis testing is a kind of behavior in statistics, through which analysts test hypotheses about population parameters. The method used by analysts depends on the nature of the data used and the reasons for the analysis. Hypothesis testing is to evaluate the credibility of hypotheses by using sample data. These data may come from a larger population or from the data generation process [5].

* Data distribution analysis:

Data distribution analysis is including mean value, the standard deviation, median value, data range, variation and data distribution diagram. This paper will use analysis methods above to analyze different attributes in American Election Study survey results.

* Data quality analysis:

Data quality analysis is including missing value analysis, outlier data analysis, and different value analysis. The missing of data mainly includes the missing of records and the missing of some field information in the records. Outlier analysis examines data for input errors and contains data that does not make sense. Data inconsistency is the contradiction and incompatibility of index data [7]. Direct mining of inconsistent data may produce mining results contrary to the reality.

* Data attributes analysis:

Data attributes analysis includes statistical analysis, periodic analysis, data contribution analysis, and correlation analysis. Quantitative data are described by statistical indicators, which are usually analyzed from two aspects: central tendency and off-center tendency. The indicator of average level is the measure of individual central tendency, and the most widely used are mean and median. The indicators that reflect the degree of variation are the measures of individuals leaving the average level, and the standard deviation and variance and four-digit spacing are widely used [8]. Contribution analysis, also known as Pareto analysis, is based on Pareto's law. The same inputs in different places produce different returns [9]. Correlation and regression methods can be used to analyze the degree and nature of relationships between different variables. Correlation analysis is used to understand the nature of the relationship between two independent variables. For example, if our goal is to study the impact of foreign direct investment (FDI) on Vietnam's level of economic growth, then two variables can be specified as FDI and GDP over the same period. Correlation coefficient can be calculated by following formula:

文本

描述已自动生成

Pearson product moment correlation is calculated by taking the ratio of the product of the sample of two variables and two standard deviations, which shows the strength of the linear relationship. In Pearson product moment correlation, the correlation coefficient is not robust because there is no strong linear relationship between variables. The correlation coefficient is sensitive to the edge points, so the correlation coefficient has no resistance.

* Chi-square test:

Chi-square test is a widely used hypothesis testing method for counting data. It belongs to the category of non-parametric test, which mainly compares two or more sample rates (composition ratio) and the correlation analysis of two categorical variables. The basic idea is to compare the coincidence degree or goodness of fit between the theoretical frequency and the actual frequency.

Its application in statistical inference of classified data includes chi-square test for comparison of two rates or two constituent ratios; Chi-square test of multi-rate or multi-component ratio comparison and correlation analysis of classified data.

**Methodology**

In our survey result analysis, we use R programming language to help us get all statistical results as above. Firstly we read American National Election Study survey data, and use different variables to records different attributes. The code is in appendix.

Variable named “ideology” records attribute ideology; variable named “region” records attribute region; variable ethnic records attribute ethnic; and variable gender records attribute gender. And then we use statistic analysis method to get claim 1 result. Thirdly, we use data quality analysis and Data attributes analysis to get results of claim2. At last we use correlation matrix to do hypothesis testing for claim 3.

**Results**

For statistical analysis we get results as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | **mean** | **median** | **range** | **std** | **variation** | **1/4 quantile** | **3/4 quantile** | **distance** |
| ideology | 4.842857 | 5 | 9 | 1.995602 | 0.4120712 | 3 | 6 | 3 |
| region | 2.557143 | 3 | 3 | 1.0108 | 0.3953 | 2 | 3 | 1 |
| ethnic | nan | 3 | 3 | 1.1338 | nan | 2 | 3 | 1 |
| gender | 1.5643 | 2 | 1 | 0.496 | 0.31717 | 1 | 2 | 1 |

Table 1: The statistic analysis for AMES 2008 survey

After doing data quality analysis we found that the survey result data has no null data or missing values. Therefore, we next doing data distribution analysis. The distribution for these four attributes are as follows:

|  |  |
| --- | --- |
| ideology | **图表, 直方图  描述已自动生成** |
| Region | 图表, 条形图  描述已自动生成 |
| Ethnic | 图表, 直方图  描述已自动生成 |

Table 2: The distribution for AMES 2008 survey

From table1 and table 2 we can get claim 1 result. we find that there is 60 percent of the population identifies as politically moderate; there is 18.09524 percent of the population identifies politically liberal; there is 27.5 percent of the population identifies as politically conservative. Consequently, the result is different with claim 1, but the result is similar. Thus the analysis result for claim1 is: much of the population is composed of moderates, with as small portion on each side who are strongly identify as “liberal” or “conservative”.

For claim 2, the Southerners and Midwesterners ideological makeup is as follows:

图表, 饼图

描述已自动生成

The Northeasterners and Westerners ideological makeup is as follows:

图表, 饼图

描述已自动生成

Thus, Southerners and Midwesterners have a different ideological makeup than Northeasterners and Westerners.

For claim3, we use hypothesis testing to get final analysis result. Our hypothesis is:

Men are more likely than women to identify with the conservativism.

Firstly, we analyze ideological make up for male and female data:

文本

描述已自动生成

And then we do correlation analysis and get correlation matrix for male and female data as follows:

* Male:

图表, 瀑布图

描述已自动生成

* Female:

图表, 瀑布图

描述已自动生成

After doing correlation analysis and ideological make up analysis we find that men are not more likely than women to identify with the conservativism. In addition, we can use chi-square test to test: men are more likely than women to identify with the conservativism. The chi-test for male data and female data is as follows:

|  |  |
| --- | --- |
| chisq-statistic | 292.1525 |
| Confidence alpha | 0.05 |
| re | 0 |
| p-value | 0.998 |

From chi-square test we can see that in confidence value is 0.05, the re is equal to 0, so we confuse the original hypothesis. Consequently, men are not more likely than women to identify with the conservativism.

**Conclusion**

After analyzing American National Election Study survey data, we help political company named “*PolitiTACT*” what effect various categorical factors have on political ideology, with the intent on utilizing such patterns to revise public policy, one facet including whether specific claims widely assumed to be true have any substantive merit. The majority of the population is made up of moderates, and a small number on each side are strongly identified as "liberals" or "conservatives. The ideological structure of southerners and Midwesterners is different from that of northeast and Westerners. Male is more likely to identify with conservatism than female.

**Appendix**

* **R code:**

|  |
| --- |
| setwd("./")  data = read.csv("./data.csv")  ideology = data[, 2]  region = data[, 3]  ethnic = data[, 4]  gender = data[, 5]  ##############################nan analysis  sum(!complete.cases(ideology))  sum(!complete.cases(region))  sum(!complete.cases(ethnic))  sum(!complete.cases(gender))  ##############################statistic analysis  # ideology statistic analysis  mean\_ = mean(ideology, na.rm = T)  median\_ = median(ideology, na.rm = T)  range\_ = max(ideology, na.rm = T) - min(ideology, na.rm = T)  std\_ = sqrt(var(ideology, na.rm = T))  variation\_ = std\_ / mean\_  q1 = quantile(ideology, 0.25, na.rm = T)  q3 = quantile(ideology, 0.75, na.rm = T)  distance = q3 - q1  a = matrix(c(mean\_, median\_, range\_, std\_, variation\_, q1, q3, distance), 1, byrow = T)  colnames(a) = c("mean", "median", "range", "std", "variation", "1/4 quantile", "3/4 quantile", "distance")  print(a)  # region statistic analysis  mean\_ = mean(region, na.rm = T)  median\_ = median(region, na.rm = T)  range\_ = max(region, na.rm = T) - min(region, na.rm = T)  std\_ = sqrt(var(region, na.rm = T))  variation\_ = std\_ / mean\_  q1 = quantile(region, 0.25, na.rm = T)  q3 = quantile(region, 0.75, na.rm = T)  distance = q3 - q1  a = matrix(c(mean\_, median\_, range\_, std\_, variation\_, q1, q3, distance), 1, byrow = T)  colnames(a) = c("mean", "median", "range", "std", "variation", "1/4 quantile", "3/4 quantile", "distance")  print(a)  # ethnic statistic analysis  mean\_ = mean(ethnic, na.rm = T)  median\_ = median(ethnic, na.rm = T)  range\_ = max(ethnic, na.rm = T) - min(ethnic, na.rm = T)  std\_ = sqrt(var(ethnic, na.rm = T))  variation\_ = std\_ / mean\_  q1 = quantile(ethnic, 0.25, na.rm = T)  q3 = quantile(ethnic, 0.75, na.rm = T)  distance = q3 - q1  a = matrix(c(mean\_, median\_, range\_, std\_, variation\_, q1, q3, distance), 1, byrow = T)  colnames(a) = c("mean", "median", "range", "std", "variation", "1/4 quantile", "3/4 quantile", "distance")  print(a)  # gender statistic analysis  mean\_ = mean(gender, na.rm = T)  median\_ = median(gender, na.rm = T)  range\_ = max(gender, na.rm = T) - min(gender, na.rm = T)  std\_ = sqrt(var(gender, na.rm = T))  variation\_ = std\_ / mean\_  q1 = quantile(gender, 0.25, na.rm = T)  q3 = quantile(gender, 0.75, na.rm = T)  distance = q3 - q1  a = matrix(c(mean\_, median\_, range\_, std\_, variation\_, q1, q3, distance), 1, byrow = T)  colnames(a) = c("mean", "median", "range", "std", "variation", "1/4 quantile", "3/4 quantile", "distance")  print(a)  #########################Claim 1  # politically moderate: moderate 4,slightly 3 ,liberal 2,slightly conservative 5  # politically liberal: liberal 2,extremely liberal 7  # politically conservative:conservative 6,extremely conservative 7  total\_num = length(ideology)  a = length(ideology[ideology == 4]) + length(ideology[ideology == 3]) + length(ideology[ideology == 2]) + length(ideology[ideology == 5])  politically\_moderate = a / total\_num  politically\_moderate  a = length(ideology[ideology == 2]) + length(ideology[ideology == 7])  politically\_liberal = a / total\_num  politically\_liberal  a = length(ideology[ideology == 6]) + length(ideology[ideology == 7])  politically\_conservative = a / total\_num  politically\_conservative  #########################Claim 2  # South 3, Midwest 2, Northeast 1  south\_mid = data[data[, "region"] == 3 | data[, "region"] == 2,]  north = data[data[, "region"] == 1,]  ideology\_south\_mid = south\_mid[, 2]  ideology\_north = north[, 2]  table(ideology\_south\_mid)/sum(table(ideology\_south\_mid))  table(ideology\_north)/sum(table(ideology\_north))  ########################Claim 3  # man 1, woman 2  men = data[data[, "gender"] == 1,]  women = data[data[, "gender"] == 2,]  women  men  # men analysis  men\_ideology = men[, 2]  women\_ideology = women[, 2]  men\_rst = table(men\_ideology)/sum(table(men\_ideology))  women\_rst = table(women\_ideology)/sum(table(women\_ideology))  men\_rst  women\_rst  # correlation analysis  cor(men\_ideology) |

* **Glossary:**

|  |  |
| --- | --- |
| Chi-Square test | Chi-square test is a widely used hypothesis testing method for counting data. It belongs to the category of non-parametric test, which mainly compares two or more sample rates (composition ratio) and the correlation analysis of two categorical variables. |
| Data distribution analysis | Data distribution analysis is including mean value, the standard deviation, median value, data range, variation and data distribution diagram. |
| Hypothesis testing | Hypothesis testing is a kind of behavior in statistics, through which analysts test hypotheses about population parameters. |
| Correlation analysis | Correlation analysis is a statistical analysis method to study the correlation between two or more random variables at the same position. |
| Data distribution analysis | The sampling distribution, also known as the statistical distribution and the function distribution of random variables, refers to the distribution of sample estimators. Descriptive statistical indicators are also referred to as parameters. |
| Data quality analysis | Data quality analysis is including missing value analysis, outlier data analysis, and different value analysis. |