Project Description: Analyzing Associations

Jyotishka Datta

11/9/2019

Final Project: Test of Association for Political Data

1. This intro will show you how to read the two data-sets and get basic summary. It will also suggest some of the analysis that you can perform.
2. You can use one of the methods taught in class (or try new methods - it's upto you (and your team)).
3. You do not need to present your work. Only submit a well-written report.

Contingency tables

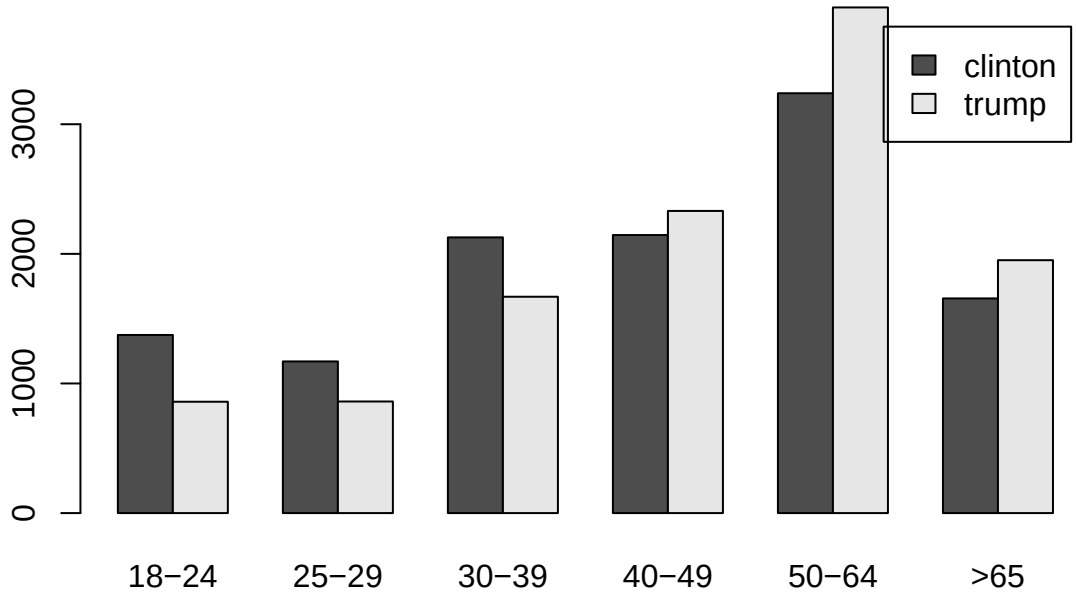
- Some of the best examples of contingency tables come from Political data analysis.
- Example: Analyzing Exit Poll Data from CNN. Total number of respondents = 24558. We start with the following table:

Party	18-24	25-29	30-39	40-49	50-64	65 and older
Clinton	56%	53%	51%	46%	44%	45%
Trump	35%	39%	40%	50%	53%	53%

Suppose we want to test if there is a significant association between political inclination and age of the respondant. This is easy using the chi-square test.

Test for association

```
clinton = c(1374,1170,2127,2145,3239,1656)
trump = c(859,861,1669,2331,3901,1951)
elect16= rbind(clinton,trump)
dimnames(elect16) = list(candidate = c("clinton","trump"),
                         agegp = c("18-24","25-29","30-39",
                                   "40-49","50-64",">65"))
barplot(elect16, beside=T, legend=T)
```



Chi-square test

```
chisq.test(elect16)
```

```
##
##  Pearson's Chi-squared test
##
## data:  elect16
## X-squared = 313.46, df = 5, p-value < 2.2e-16
```

- We can reject the null hypothesis that the proportions are not equal across different age groups.

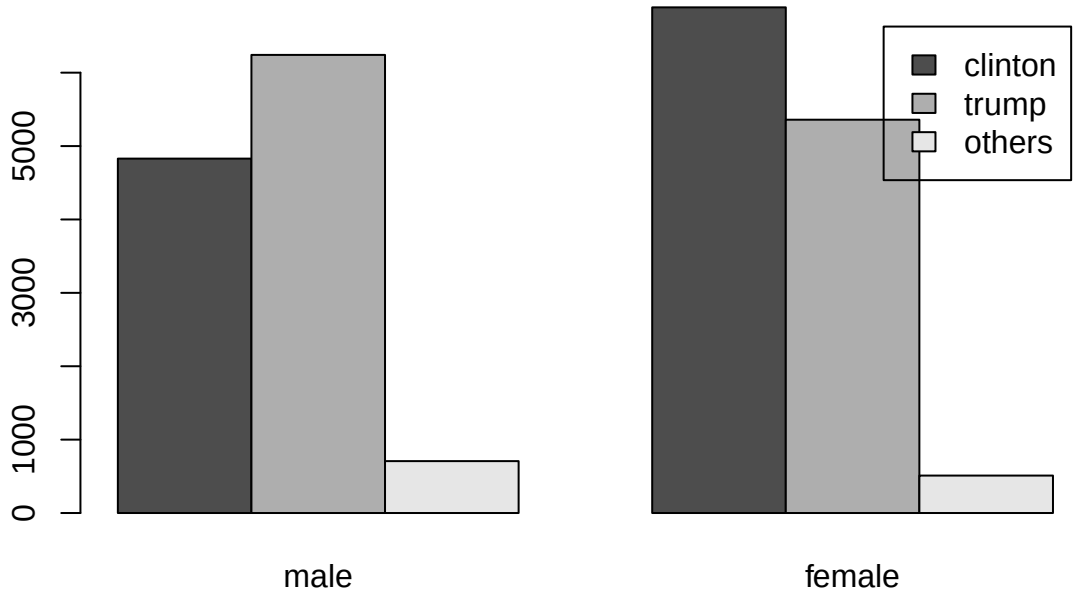
Effect of Gender?

- We can also look at the effect of Gender!
- The same CNN exit poll data : 24537 respondents.

Party	clinton	trump	others
male	41%	53%	6%
female	54%	42%	4%

Use R

```
gender <- matrix(c(4829,6242,707,6890,5359,510), byrow = T, ncol = 3)
dimnames(gender) = list(gender = c("male","female"),
                        candidate = c("clinton","trump","others"))
barplot(t(gender), beside=T, legend=T)
```



Chi-square test of association

```
gender <- matrix(c(4829,6242,707,6890,5359,510), byrow = T, ncol = 3)
dimnames(gender) = list(gender = c("male","female"),
                        candidate = c("clinton","trump","others"))
prop.table(gender, 1)
```

```
##         candidate
## gender     clinton     trump     others
##   male   0.4100017 0.5299711 0.06002717
##   female 0.5400110 0.4200172 0.03997178
```

```
chisq.test(gender)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  gender
## X-squared = 423.02, df = 2, p-value < 2.2e-16
```

Now, if you look at the main exit poll web page, you will find many such contingency tables across different demographic variables or survey questions, e.g.

1. Age
2. Race
3. Education
4. Inclome
5. Party ID.
6. Ideology
7. Marital status
8. religion
9. Served in the military
10. Were you born a US citizen? etc.

A natural question is which of these variables are associated with political association? Which are not?

Goal: Which variables are interesting or important?

- Should you perform the analysis for the whole national data or state-wise? Are the patterns different? Could a variable be significant for one but not so for another? Or change directions.
- You can consider any set of variables and their intersections and any level (state or country)
  - Education / Ideology / Religion
  - Opinion on Immigration / Insurance / Criminal justice / National Economy.

See the entire list at https://www.cnn.com/election/2016/results/exit-polls.

Goal: Multiplicity and Choice of Categories

- You need to perform this test of association for at least 10 different variables, get P-values for each of them and then apply multiple testing correction if needed (e.g. Bonferroni's).
- Pick a single variable but different granularities, e.g. multiple different categorizations of age, e.g four categories with 18-29, 30-44, 45-64, >65 versus six categories with 18-24, 25-29, 30-39, 40-49, 50-64, >65. How does this affect the strength of associations?
- *Write your conclusions clearly.* Submit your R codes along with your report.

Help

- If you get stuck with any of the steps, please let me know at jd033@uark.edu.