Problem Set 1 (Sept 22)

Jamie Duncan

18/09/2020

#Problem Set 1

library(here)

## here() starts at /Users/jamied/Documents/GitHub/qss/CAUSALITY/ProblemSet

here()

## [1] "/Users/jamied/Documents/GitHub/qss/CAUSALITY/ProblemSet"

library(tidyverse)

## ── Attaching packages ────────────────────

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ──── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Section 1

options(tinytex.verbose = TRUE) #not sure what this does...  
parlgov <- read.csv("parlgov.csv", stringsAsFactors = TRUE) #loading the data set, which I downloaded into my working directory

### Question 1

#### How many countries are there in the data?

dim(table(parlgov$country))#created a table with the variable of interest and used the dimension function to tell me how many table entries there are

## [1] 35

#### And how many party families?

dim(table(parlgov$party\_family))#same logic as above

## [1] 9

#### What is the average seat share (seats divided by seat total) in the data?

parlgov$seatshare <- parlgov$seats / parlgov$seats\_total   
##created new column "seatshare" as the result of seats divided by total seats  
mean(parlgov$seatshare, na.rm = TRUE)

## [1] 0.1120598

##calculated mean, removing null entries (thanks to the 'help' function)

#### Was the halfdecade variable coded correctly?

table(parlgov$halfdecade)

##   
## (1949,1954] (1954,1959] (1959,1964] (1964,1969] (1969,1974] (1974,1979]   
## 218 207 175 205 259 419   
## (1979,1984] (1984,1989] (1989,1994] (1994,1999] (1999,2004] (2004,2009]   
## 380 450 673 633 667 725   
## (2009,2014] (2014,2019]   
## 804 763

##took a peek at the variable using a table, Crtl+F'd in the textbook to verify accuracy

Yes. The rounded parentheses exclude the first number and the square brackets include the last. This is a slightly counter-intuitive but functional (in way I can’t yet explain) way of expressing 1950-54, 1955-1959, and so on.

### Question 2

#### Create a separate data frame that only includes the 2015 Canadian election.

can2015 <- parlgov[parlgov$country == "Canada" & parlgov$year == 2015, ]   
##used indexing to create subset with results related to the 2015 election in Canada

#### Are there any coding decisions the data set’s authors made with which you disagree?

Some entries in the “party\_name” field are labeled with characters that hold significance in R like “|” “[”

### Question 3

#### Create a separate data frame that only includes elections after 2014.

after2014 <- parlgov[parlgov$year > 2014, ]  
##used indexing to create subset with elections after 2014

#### Across the whole range of elections included here, which party got the highest vote share? And what was their vote share? How about seat share? What kind of election was that?

sortedafter2014 <- after2014[order (after2014$vote\_share, decreasing = TRUE), ] ## ordered by vote share   
head(sortedafter2014) #show me the top of the data frame

## election\_id country year  
## 6034 1010 Malta 2017  
## 6460 1066 Malta 2019  
## 6547 1076 Hungary 2019  
## 6160 1023 Hungary 2018  
## 6091 1015 Romania 2016  
## 6464 1067 Poland 2019  
## party\_name  
## 6034 Malta Labour Party  
## 6460 Malta Labour Party  
## 6547 Fidesz -- Hungarian Civic Party / Christian Democratic People's Party  
## 6160 Fidesz -- Hungarian Civic Party / Christian Democratic People's Party  
## 6091 Social Democratic Party  
## 6464 Law and Justice  
## vote\_share seats seats\_total left\_right state\_market liberty\_authority  
## 6034 55.04 37 67 4.2105 5.1128 6.3158  
## 6460 54.29 4 6 4.2105 5.1128 6.3158  
## 6547 52.56 13 21 6.5432 3.9949 7.5275  
## 6160 49.28 NA 199 6.5432 3.9949 7.5275  
## 6091 45.48 154 329 3.2325 2.4518 6.2794  
## 6464 45.38 26 51 7.6997 4.0466 8.2514  
## halfdecade party\_family seatshare  
## 6034 (2014,2019] Social democracy 0.5522388  
## 6460 (2014,2019] Social democracy 0.6666667  
## 6547 (2014,2019] Conservative 0.6190476  
## 6160 (2014,2019] Conservative NA  
## 6091 (2014,2019] Social democracy 0.4680851  
## 6464 (2014,2019] Conservative 0.5098039

In the 2017 Maltese general election, the Malta Labour Party got the highest vote share across the whole range of elections since 2014 with 55.04% of the votes. This afforded the party 55.22 percent of seats in the Maltese House of Representatives.

### Question 4

#### Write a function that takes as arguments a data frame with the results from a single election and the name of a party family. It should return the total votes for that party family in that election.

germany.2019 <- after2014[after2014$election\_id == 1055, ] ##data frame with the results of a single election  
drop\_na(germany.2019)

## election\_id country year party\_name vote\_share seats  
## 1 1055 Germany 2019 Free Voters 2.2 2  
## 2 1055 Germany 2019 Free Democratic Party 5.4 5  
## 3 1055 Germany 2019 Social Democratic Party of Germany 15.8 16  
## 4 1055 Germany 2019 Alliance 90 / Greens 20.5 21  
## 5 1055 Germany 2019 The Left / PDS 5.5 5  
## 6 1055 Germany 2019 Christian Democratic Union 22.6 23  
## 7 1055 Germany 2019 Christian Social Union 6.3 6  
## 8 1055 Germany 2019 Family Party of Germany 0.7 1  
## 9 1055 Germany 2019 Alternative for Germany 11.0 11  
## 10 1055 Germany 2019 Ecological Democratic Party 1.0 1  
## seats\_total left\_right state\_market liberty\_authority halfdecade  
## 1 96 7.4000 6.4000 7.0000 (2014,2019]  
## 2 96 5.9233 8.2173 3.0695 (2014,2019]  
## 3 96 3.6451 3.8443 3.9791 (2014,2019]  
## 4 96 2.9308 4.3268 1.3251 (2014,2019]  
## 5 96 1.2152 1.0617 3.2050 (2014,2019]  
## 6 96 6.2503 6.3498 6.9333 (2014,2019]  
## 7 96 7.2871 6.5615 7.8675 (2014,2019]  
## 8 96 7.4000 6.4000 7.0000 (2014,2019]  
## 9 96 8.8000 5.9000 8.5000 (2014,2019]  
## 10 96 2.5000 2.5000 1.8000 (2014,2019]  
## party\_family seatshare  
## 1 Conservative 0.02083333  
## 2 Liberal 0.05208333  
## 3 Social democracy 0.16666667  
## 4 Green/Ecologist 0.21875000  
## 5 Communist/Socialist 0.05208333  
## 6 Christian democracy 0.23958333  
## 7 Christian democracy 0.06250000  
## 8 Conservative 0.01041667  
## 9 Right-wing 0.11458333  
## 10 Green/Ecologist 0.01041667

##this is a function that takes the results from a single election and the name of the party family  
avg.perf <- function(df, parameter){   
 s1.avg.perf <- ifelse(df$party\_family == parameter, 1, 0)  
 return(aggregate(df$vote\_share , by = list(s1.avg.perf), sum, na.rm = TRUE)[2, 2])  
}  
 avg.perf(germany.2019, "Conservative")

## [1] 2.9

There is no way that I know to calculate total votes with this information but the above function calculates vote share by party family.

#### Use that function to calculate the total vote share of each party family in the 2019 European elections in Germany.

## I modified the above function for efficiency.  
avg.perf.all <- function(df){  
 for(i in unique(df$party\_family)){ #replaced the parameter input with a for() loop  
 return(aggregate(df$vote\_share, by = list(df$party\_family), sum, na.rm = TRUE)) #modified return to refer to df  
 }  
}  
  
avg.perf.all(germany.2019)

## Group.1 x  
## 1 Christian democracy 28.9  
## 2 Communist/Socialist 5.5  
## 3 Conservative 2.9  
## 4 Green/Ecologist 21.5  
## 5 Liberal 5.4  
## 6 Right-wing 11.0  
## 7 Social democracy 15.8  
## 8 Special issue 5.2

#### Describe conceptually how you would go about calculating the average performance of each party family in each half decade.

Similar to above, I would include an addition nested “for()” loop to run “avg.perf.all” for each unique halfdecade row entry.

### Question 5

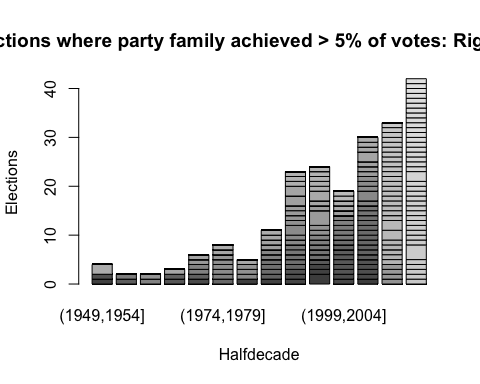
#### Create a data set containing only radical right parties and only elections in which those parties received more than 5% of the vote.

rad.right <- parlgov[parlgov$party\_family == "Right-wing" & parlgov$vote\_share > 5, ]  
rad.right <- na.omit(rad.right)

#### Create a barplot showing how many such party-elections fall in each half decade–label the y-axis.

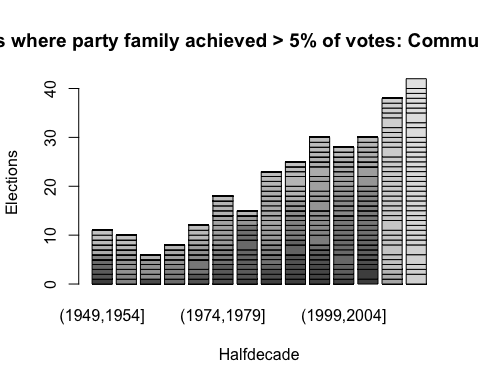
I interpret party-election to mean that multiple parties in a family could be counted in one election.

##wrote a function to increase efficiency for the next question  
fam.elec.plot <- function(df, parameter){ # function takes a data.frame (parlgov) and a parameter (party\_family)  
 fam.elec\_a <- df[df$party\_family == parameter & df$vote\_share > 5, ] #subsets the dataframe by the parameter, excluding instances with a vote share less than 5%  
 fam.elec\_b <- na.omit(fam.elec\_a) #omits NAs  
 fam.elec\_tab <- table(fam.elec\_b$election\_id, fam.elec\_b$halfdecade) #creating a table with election\_ids and halfdecades  
 return(barplot(fam.elec\_tab, main = paste("Elections where party family achieved > 5% of votes:", parameter), xlab = "Halfdecade", ylab = "Elections")) #outputting a barplot  
}  
  
fam.elec.plot(parlgov, "Right-wing") #using function to answer first part of this question

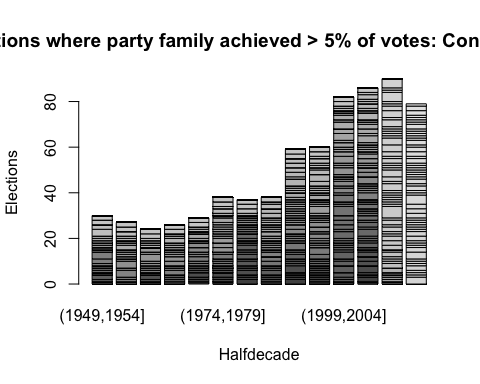


#### Do the same for the communist and conservative parties.

fam.elec.plot(parlgov, "Communist/Socialist") #using function for the second part of this question



fam.elec.plot(parlgov, "Conservative")



#### How would you interpret these? Does seeing the second change your interpretation?

Despite radical parties gaining traction in more elections over time, one can also observe through the third chart that in that more moderate Conservative parties are A) far more likely to earn at least 5% of votes in an election and B) are also gaining traction relative to past half decades.

### Question 6

#### Normalize the previous results using the total number of elections in each half decade.

I spent a long time trying to figure out how to do this. I asked a friend and he told me to look up the group\_by function in tidyverse. This would be part of the solution…I’m still lost.

parlgov %>%  
 group\_by(halfdecade) %>%  
 summarize(unique\_elecs = n\_distinct(election\_id))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 14 x 2  
## halfdecade unique\_elecs  
## <fct> <int>  
## 1 (1949,1954] 32  
## 2 (1954,1959] 31  
## 3 (1959,1964] 26  
## 4 (1964,1969] 30  
## 5 (1969,1974] 32  
## 6 (1974,1979] 48  
## 7 (1979,1984] 47  
## 8 (1984,1989] 52  
## 9 (1989,1994] 65  
## 10 (1994,1999] 67  
## 11 (1999,2004] 70  
## 12 (2004,2009] 79  
## 13 (2009,2014] 78  
## 14 (2014,2019] 73

#### Produce new plots and save them as pdf files

pdf("rightwing.pdf")  
fam.elec.plot(parlgov, "Right-wing")  
dev.off()

## quartz\_off\_screen   
## 2

pdf("communist\_socialist.pdf")  
fam.elec.plot(parlgov, "Communist/Socialist")  
dev.off()

## quartz\_off\_screen   
## 2

pdf("conservative.pdf")  
fam.elec.plot(parlgov, "Conservative")  
dev.off()

## quartz\_off\_screen   
## 2

fam.elec.plot(parlgov, “Conservative”) fam.elec.plot(parlgov, “Right-wing”)

## Section 2

### Question 1

#### Load the data & check with summary() funtion.

options(tinytex.verbose = TRUE)  
survey <- read.csv("survey.csv", stringsAsFactors = TRUE)

#### Which variable has missing data? Also, check the number of missing observations (including missing values)

summary(survey)

## draft year ideology state   
## Min. :0.000 Min. :1950 Min. :1.000 CO:284   
## 1st Qu.:0.000 1st Qu.:1950 1st Qu.:2.000 OR:391   
## Median :1.000 Median :1951 Median :3.000   
## Mean :0.523 Mean :1951 Mean :3.047   
## 3rd Qu.:1.000 3rd Qu.:1952 3rd Qu.:4.000   
## Max. :1.000 Max. :1952 Max. :5.000   
## NA's :15

“ideology” has missing values.

### Question 2

#### Calculate the mean ideology score by year that respondents were born in.

aggregate(survey$ideology, by = list(survey$year), mean, na.rm = TRUE) #aggregate ideology by year while removing NAs

## Group.1 x  
## 1 1950 3.018957  
## 2 1951 3.052174  
## 3 1952 3.068493

#### Briefly interpret the result.

Over the three years of birth documented, left wing sentiment increases by about 0.05 out of 5 or roughly 1%. I would interpret this as a negligible change.

### Question 3

#### Estimate the sample average treatment effect on ideology. In this question we pool all years. Briefly interpret the result.

drafted <- survey[survey$draft == 1, ] #subset for drafted  
drafted.ideo <- mean(drafted$ideology, na.rm = TRUE) #average ideology for drafted  
safe <- survey[survey$draft == 0, ] #subset for non-drafted  
safe.ideo <- mean(safe$ideology, na.rm = TRUE)#average ideology for non-drafted  
dim.sate <- drafted.ideo - safe.ideo #difference in means SATE is equal to average treated minus average non-treated  
print(dim.sate) #print result

## [1] -0.02524286

Using the difference in means method of calculating sample average treatment effect (SATE), we can see a small negative treatment effect, whereby those drafted tend to be very slightly more conservative by under 0.03 out of 5 possible points. This very small difference could be complicated by confounding variables such as age or state.

### Question 4

#### Write a function to estimate the sample average treatment effect on ideology for each state (pooling all years). Briefly interpret the results.

## $drafted  
## Group.1 x  
## 1 CO 2.926667  
## 2 OR 3.118557  
##   
## $not.drafted  
## Group.1 x  
## 1 CO 2.881890  
## 2 OR 3.179894  
##   
## $sate  
## [1] 0.04477690 -0.06133748

For Colorado, the difference in means SATE is roughly 0.045 meaning those drafted were slightly more liberal than those not drafted in the state. We see the opposite in Oregon, with a difference in means SATE of -0.061, where those drafted were slightly more likely to be conservative than those not drafted in the state. In both cases the differences are not what I would intuitively imagine to be significant (though being able to definitively say so will come later in this course, I imagine.)

### Question 5

#### A politician from a country in Asia is planning to use this paper to discuss the effect of Draft lottery in her country. Is this a valid approach for policy making? Discuss briefly (two to four sentences will suffice).

This is not a good approach to policy-making as there are many possible confounding variables (some measurable and some not) that would make this study non-analogous to other parts of the world or even to the United States today. The Vietnam War and the consequent draft took place in a very particular cultural, political, and social setting. The study is interesting but not generalizable.