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 ---------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v 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 i
```

Question 1

```
dim(table(parlgov$country))#created a table with the variable of interest and used the dimension functi
```

How many countries are there in the data?

```
## [1] 35
```

```
dim(table(parlgov$party_family))#same logic as above
```

And how many party families?

```
## [1] 9
```

```
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)
```

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

```
## [1] 0.1120598
```

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

```
table(parlgov$halfdecade)
```

Was the halfdecade variable coded correctly?

```
##
## (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

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

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

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

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

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

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

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? 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

```
germany.2019 <- after2014[after2014$election_id == 1055, ] ##data frame with the results of a single el
drop_na(germany.2019)
```

```
##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")
```

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.

```
## [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.

```r
## 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
  }
}

avg.perf.all(germany.2019)
```

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

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

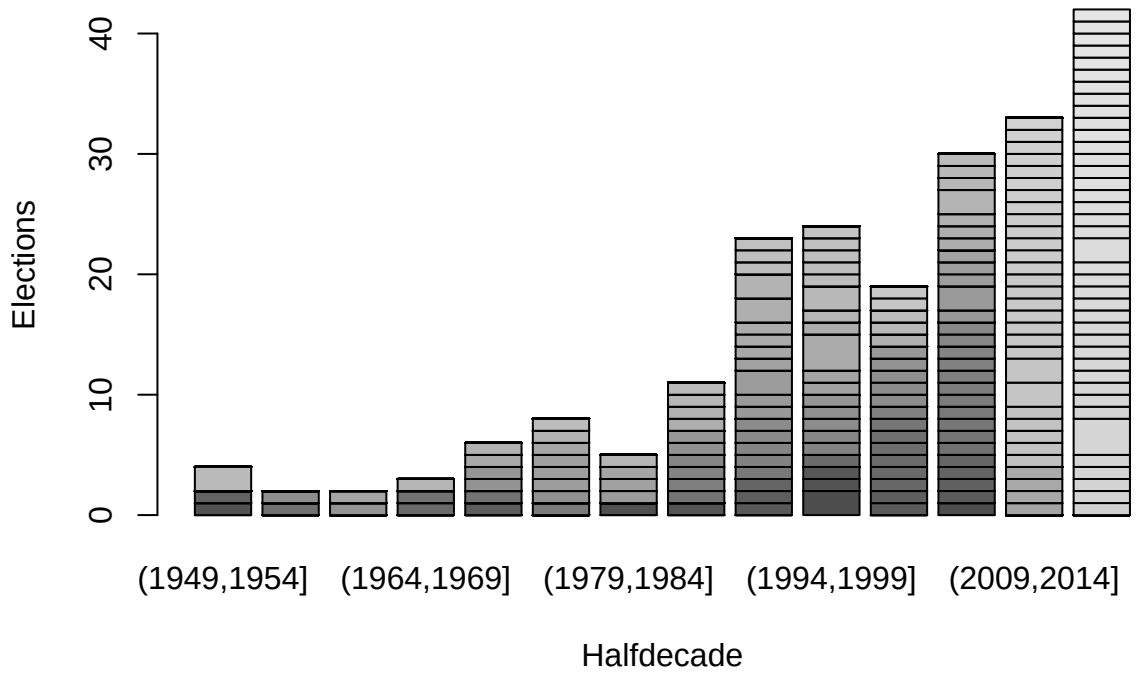
```r
rad.right <- parlgov[parlgov$party_family == "Right-wing" & parlgov$vote_share > 5, ]
rad.right <- na.omit(rad.right)
```

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

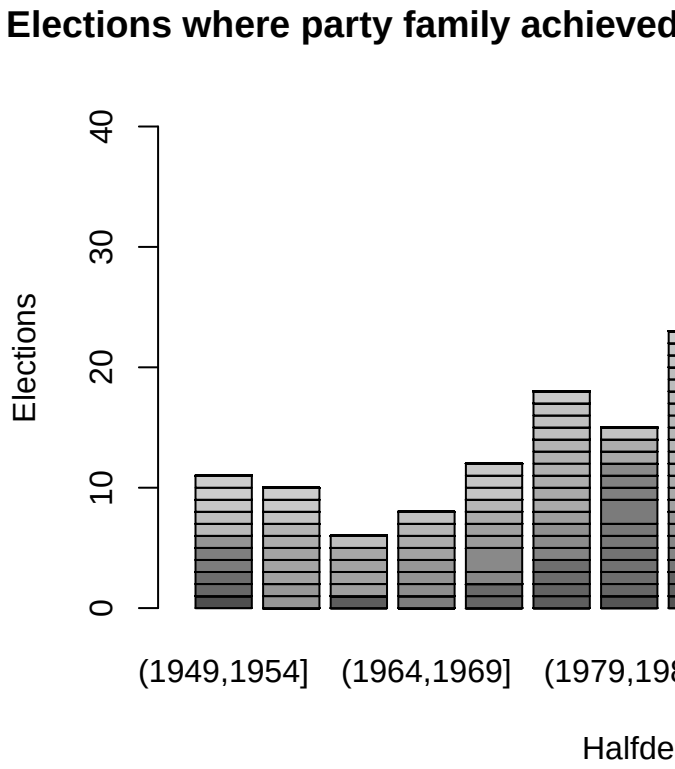
Create a barplot showing how many such party-elections fall in each half decade–label the y-axis. I interpret party-election to mean that multiple parties in a family could be counted in one election.

```r
##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 (part
  fam.elec_a <- df[df$party_family == parameter & df$vote_share > 5, ] #subsets the dataframe by the pa
  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_
  return(barplot(fam.elec_tab, main = paste("Elections where party family achieved > 5% of votes:", para
}

fam.elec.plot(parlgov, "Right-wing") #using function to answer first part of this question
```

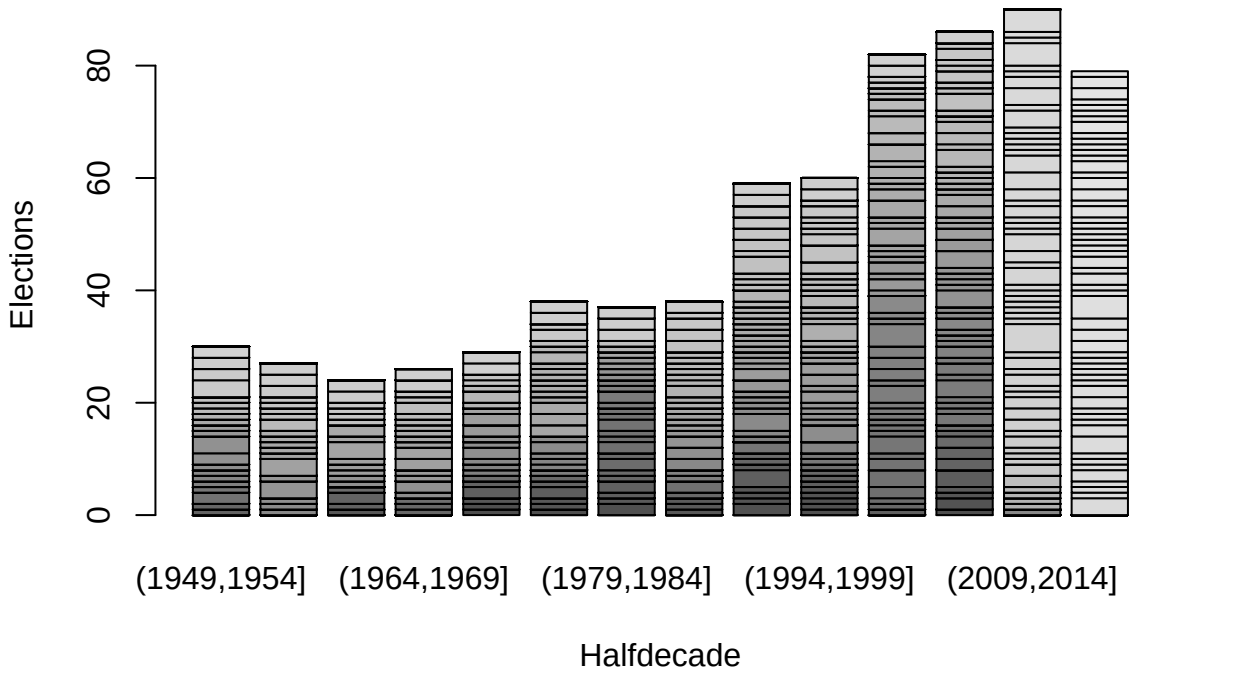



```
fam.elec.plot(parlgov, "Communist/Socialist") #using function for the second part of this question
```



Do the same for the communist and conservative parties.

```
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)
```

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

Produce new plots and save them as pdf files

```
## pdf
##   2
```

```
pdf("communist_socialist.pdf")
fam.elec.plot(parlgov, "Communist/Socialist")
dev.off()
```

```
## pdf
##   2
```

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

```
## pdf
##   2
```

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

Section 2

Question 1

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

Load the data & check with summary() funtion.

```
summary(survey)
```

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

```
##      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

```r
aggregate(survey$ideology, by = list(survey$year), mean, na.rm = TRUE) #aggregate ideology by year whil
```

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

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

```r
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 averag
print(dim.sate) #print result
```

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