API - 209 | Problem Set 10

Prof. Dan Levy

Due on Tuesday, November 8, 2022 at 10:00 am.

GENERAL INSTRUCTIONS

- **THIS IS A GROUP PROBLEM SET** and you should work with your Final Exercise group.
- The only exception is the survey at the end of Question 2. **You need to submit this form individually.**
- Submissions to Canvas will be done by team (i.e., only one submission per team). Instructions on how to do so are **here**.
- We expect that it will take your team several hours to complete the problem set, and we think you will learn more (and have more fun) if you allow enough time to do so.

INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. **Download this RMarkdown document file into your computer.**
2. **Insert all your answers into this document.** Guidance **here** on how to insert objects such as handwritten work or screenshot images in your answers.
3. **SAVE your work frequently**.
4. To make things easier to visualize in RStudio, you can set the view mode as "Visual" instead of as "Source" in the top left of your screen (just below the Save button).
5. Once your document is complete, please save it as a PDF by clicking the **KNIT** button.
6. Please submit an electronic copy of the PDF (and any separate requested files) to the Canvas course page.

   6.a) If you want to check a PDF version of this problem set before starting to work on it, you can always knit it. In fact, you can knit the document at any point.

   6.b) If you cannot Knit and it's time to submit the problem set, submit the RMarkdown file and make an appointment with a member of the teaching team
7. Remember to consult the R resources from math camp, particularly the HKS R cheat sheet (available **here**, which contains many of the commands needed to answer the questions in this problem set.

IDENTIFICATION

1. Group Members (please list below the classmates you worked with on this problem set):

```
Group members: Luisa Leite, Carlos Viquez, Kwang Jun Lee, Bharath Ram, Shreya Chaturvedi, Masato Takahash
```

2. Compliance with Harvard Kennedy School Academic Code: Do you certify that my work in this problem set complies with the Harvard Kennedy School Academic Code[1] (mark with an X below)?

```
[ X ] YES                  [   ] NO
```

[1] We abide by the Harvard Kennedy School Academic code (available here) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus.

QUESTION 1 – PREDICTING US MIDTERM ELECTIONS[2]

The goal of this problem set question is to help improve your understanding of how some of the quantitative tools and simulations you have learned in this course can be used to forecast results of the election for the U.S. House of Representatives on Election Day, November 8[3]. The stakes of the election are high for the parties, democracy, and the future of the country. These and many other important aspects of the election are not addressed in this problem set question.

Part 1 explores several important statistical concepts using just two congressional districts to help ground your learning for the subsequent parts. In **Part 2**, you will run your own (relatively) basic simulation model to characterize the distribution of possible outcomes and to predict the number of seats that members of the Republican Party will win in aggregate on November 8[4]. In **Part 3**, you will enhance your basic simulation model by incorporating correlations among the outcomes of the elections in different congressional districts. Finally, in **Part 4**, you will further enhance your simulation model however you see fit to make a prediction of the aggregate election outcome.

PART 1 - A Simple Model of Two Congressional Districts

To get started, let's consider just two congressional districts – the California 22nd Congressional District and the California 27th Congressional District. These congressional districts are currently represented by Republicans David Valadao and Mike Garcia, respectively. According to the website FiveThirtyEight.com, the Democratic candidate Rudy Salas has a 43.1 percent chance of winning the seat in California 22nd, and the Democratic candidate Christy Smith has a 36.4 percent chance of winning the seat in California 27th[5].

Assume for now that the Democratic candidate winning in California 22nd and in California 27th are independent events. (Later, we will explore what happens if the outcomes are correlated.)

(a) What is the expected number of seats that the Democrats win in total in these two districts? Show your work[6].

Answer: Assuming independence, the probability of the Democrats winning in both districts is P(Democrat wins California 22nd) × P(Democrat wins California 27th) = 0.431 × 0.364 = 0.157. Similarly, the probability of the Democrats winning in either of the two districts is P(Democrat wins California 22nd) × P(Democrat loses California 27th) + P(Democrat loses California 22nd) × P(Democrat wins California 27th) = 0.431 × (1 - 0.364) + (1 - 0.431) × 0.364 = 0.481. Finally, the probability of the Democrats losing in both districts is P(Democrat loses California 22nd) × P(Democrat loses California 27th) = (1 - 0.431) × (1 - 0.364) = 0.362. Therefore, the expected number of seats for the Democrats in these two districts is 2 × 0.157 + 1 × 0.481 + 0 × 0.362 = 0.795

(b) What is the most likely number of seats the Democrats will win in total in these two districts? Show your work.

Answer: Since the probability of winning one seat is the largest among the three possible outcomes, the Democrats are most likely to win one seat.

Let's explore this question further. Our goal is to explore two ways in which we might enhance our simple model for these two congressional districts: (1) Adjusting for any systematic bias in the polls; and (2) Adjusting for potential correlations among outcomes in the two districts.

[2] My deepest gratitude to my dear colleague Jonathan Borck, who is the key intellectual architect behind the question in this problem set, and to a student who will remain anonymous for providing the impetus for producing this question.

[3] Relevant background: The U.S. House of Representatives has 435 voting members, each of which represents a specific geographic area of the country (a "congressional district"). The current U.S. House of Representatives has 220 members who are Democrats, 212 members who are Republicans, and 3 vacant seats. On November 8, voters in each of the 435 congressional districts across the United States will each elect one representative to represent their congressional district. Some will re-elect their current representative, while others will elect a new representative, perhaps from the other political party. The winning candidates from each of these 435 individual elections will make up the new U.S. House of Representatives. For the majority party in the U.S. House of Representatives to switch from Democrats to Republicans, Republicans would need to win 218 of the 435 seats on November 8.

[4] Because the Republican Party is considered to be favorite to win enough seats on November 8 to take control of the U.S. House of Representatives, we will phrase our predictions from the Republicans' perspective.

[5] FiveThirtyEight.com house "deluxe" forecast, Oct. 28, 2022, 8:12 p.m. EDT. (Note that, by the time you work on this problem set, the probabilities will almost certainly have changed, at least somewhat. You may use the probabilities given here to solve the problem. You do not need to obtain FiveThirtyEight's updated forecasts.)

[6] For this and similar questions, please feel free to write your answer by hand and insert a picture of your handwriting if this proves to be easier. If you do this, please make sure your handwriting is legible and the picture/scan is clear.

Adjustment 1 – Systematic Bias in the Polls

We might be concerned that the polls that underlie our prediction model are biased. In other words, the polls might systematically understate or overstate support for the Democratic candidate (or, equivalently, the Republican candidate).

The term "bias" is not meant to imply malicious intent. Bias can arise for perfectly innocent reasons, such as systematically failing to identify (or overidentifying) individuals who will actually vote, or including a larger portion of a certain party's supporters in a sample than exist in the voting population.

Suppose that, in the problem above, the polls are systematically biased in favor of the Democratic candidate. In other words, suppose that support for the Republican candidate in the population is higher in both congressional districts than the polls currently indicate, and as a result, the Democrat's chances of winning in California 22nd is actually 38 percent (not 43.1 percent) and the Democrat's chances of winning in California 27th is actually 33 percent (not 36.4 percent). Continue to assume these are independent events.

(c) How does correcting for this bias change (i) the expected number of seats the Democrats will win in these two congressional districts and (ii) the most likely number of seats the Democrats will win in these two congressional districts? Explain.

Answer: (i) Assuming independence, the probability of the Democrats winning in both districts is P(Democrat wins California 22nd) × P(Democrat wins California 27th) = 0.33 × 0.38 = 0.125. Similarly, the probability of the Democrats winning in either of the two districts is P(Democrat wins California 22nd) × P(Democrat loses California 27th) + P(Democrat loses California 22nd) × P(Democrat wins California 27th) = 0.33 × (1 - 0.38) + (1 - 0.33) × 0.38 = 0.459. Finally, the probability of the Democrats losing in both districts is P(Democrat loses California 22nd) × P(Democrat loses California 27th) = (1 - 0.33) × (1 - 0.38) = 0.415. Therefore, the expected number of seats for the Democrats in these two districts is 2 × 0.125 + 1 × 0.459 + 0 × 0.415 = 0.71

\textcolor{blue}{(ii) Since the probability of winning one seat is still the largest among the three possible outcomes, the Democrats are most likely to win one seat.

Adjustment 2 – Correlation Among Outcomes in Different Congressional Districts

Finally, consider correlation among the outcomes in different congressional districts. In the previous part of this question, we asked you to assume that the Democratic candidate winning in California 22nd and the Democratic candidate winning in California 27th are independent events. Instead, assume that the outcomes in these two congressional districts are correlated.[7] In particular, suppose:

- P(Democrat wins California 22nd | Democrat wins California 27th) = 0.781
- P(Democrat wins California 22nd | Democrat does not win California 27th) = 0.231

(d) Do these probabilities reflect a positive correlation or a negative correlation between the outcomes in California 22nd and California 27th districts? Explain.

Answer: Either the democrat in California 22nd or 27th disrict is more likely to win if the other democratic candidate has won the election. This implies a positive correlation.

(e) How does allowing for this type of correlation change

[7] You may assume the same original probabilities given initially in the prompt (assuming no polling bias), i.e., the Democratic candidate Rudy Salas has a 43.1 percent chance of winning the seat in California 22nd, and the Democratic candidate Christy Smith has a 36.4 percent chance of winning the seat in California 27th.

(i) the expected number of seats that the Democrats win in these two congressional districts and (ii) the most likely number of seats that the Democrats win in these two congressional districts? Explain.

Answer: (i) Assuming non-independence, the probability of the Democrats winning in both districts can be written as P(Democrat wins California 22nd | Democrat wins California 27th) × P(Democrat wins California 27th) = 0.781 × 0.364 = 0.284. Similarly, the probability of both Democrats losing is P(Democrat loses California 22nd | Democrat loses California 27th) × P(Democrat loses California 27th) = (1 - 0.231) × (1 - 0.364) = 0.489. Finally, the probability of the Democrats winning one of the two districts is 1 - 0.284 - 0.489 = 0.227. Thus, the expected number is 2 × 0.284 + 1 × 0.227 + 0 × 0.489 = 0.795.

(ii) Since the probability of winning no seats is now the largest among the three possible outcomes, the Democrats are most likely to win no seats in these two districts.

PART 2 - Simulating the Election for the U.S. House of Representatives (Basic Model)

The goal of this part is to predict the aggregate outcome of the election for the U.S. House of Representatives using a basic simulation model. By "aggregate outcome," we mean the total number of seats (out of 435) that one party's candidates will win across the country. Currently, the Democrats hold a majority of seats in the U.S. House of Representatives. To win a majority of seats in the new U.S. House of Representatives, the Republicans need to win a total of 218 seats on Election Day.

We will use FiveThirtyEight's estimates (as of October 28) of a Republican candidate's probability of winning in each congressional district to simulate the midterm elections on November 8, 2022. While FiveThirtyEight does not fully explain its methodology, its estimate of each candidate's win probability is based on aggregating election polls conducted by polling organizations and modeling based on these poll results[8].

You can find the necessary data for the following parts in the file `midterm_election_2022.csv`. The Republican candidate's probability of winning in each congressional district (`rep_win_prob`). These figures are taken from FiveThirtyEight.com's house "deluxe" forecast model and are current as of Friday, Oct. 28, at 8:12 p.m.

Using the approach to simulation we have followed in class and review sessions, **please simulate the election 1,000 times**.

[This involves generating a random number between 0 and 1 for each race, and comparing this number to the win probability listed. If the number is lower than the win probability, it means that the Republican candidate won the election in their district for that simulation (denote this as a 1). Otherwise, the Republican candidate lost the election in their district for that simulation (denote this as a zero). Do this for every race, and then repeat this process 999 more times. At the end of this process, you should have a matrix of zeroes and ones, corresponding to congressional districts in which the Republican candidate loses and wins, respectively.

```
rm(list = ls())

midterm_election_2022 <- read_csv("midterm_election_2022.csv") %>%
  as_tibble()
midterm_analysis <- midterm_election_2022

rep_win_prob <- midterm_analysis$rep_win_prob
simulation <- tibble()

set.seed(2)
for (i in 1:length(rep_win_prob)) {
  tmp <- rbinom(1000, 1, rep_win_prob[[i]])
  simulation <- simulation %>%
    rbind(tmp)
}
```

[8] Roughly speaking, FiveThirtyEight.com uses an average of recent polls and other information to predict the probability that each candidate wins each congressional district. If the website estimates, for example, that the Republican candidate is ahead in the vote in his or her congressional district, the probability that the Republican wins that district will be greater than 50 percent. The further ahead that the Republican candidate is, the larger his or her win probability will be. Please note that other polling and statistics sites provide different win probabilities. Estimating win probabilities is not a trivial task!

```
colnames(simulation) <- paste0("sim", rep(1:1000))

rm(list = c("i", "tmp"))
```

After simulating the election 1,000 times, please answer the following questions

(a) Report the percentage of iterations in which the Republican candidate wins in California 22nd. In other words, in what fraction of the 1,000 iterations does the Republican candidate wins in California 22nd? Does your answer approximately correspond to the Republican candidate David Valadao's win probability for California 22nd? Explain briefly.

```
midterm_analysis$sim <- rowMeans(simulation[, c(1:1000)])

midterm_analysis %>%
  filter(state == "CA", congressional_district == 22) %>%
  select(sim) %>%
  as.numeric()
```

```
## [1] 0.579
```

Answer: 57.9%. This approximately corresponds to Valadao's win probability for California 22nd.

(b) Does the approach you used to simulate the election (specifically, the selection of random numbers and the use of those random numbers to determine the outcome of each election) implicitly reflect an assumption that the outcomes of elections in different congressional districts are independent, or not? Explain.

Answer: This simulation assumes independence across congressional districts because a simulation in each district is run solely based on the win probability of the Republican candidate in that district and nothing else.

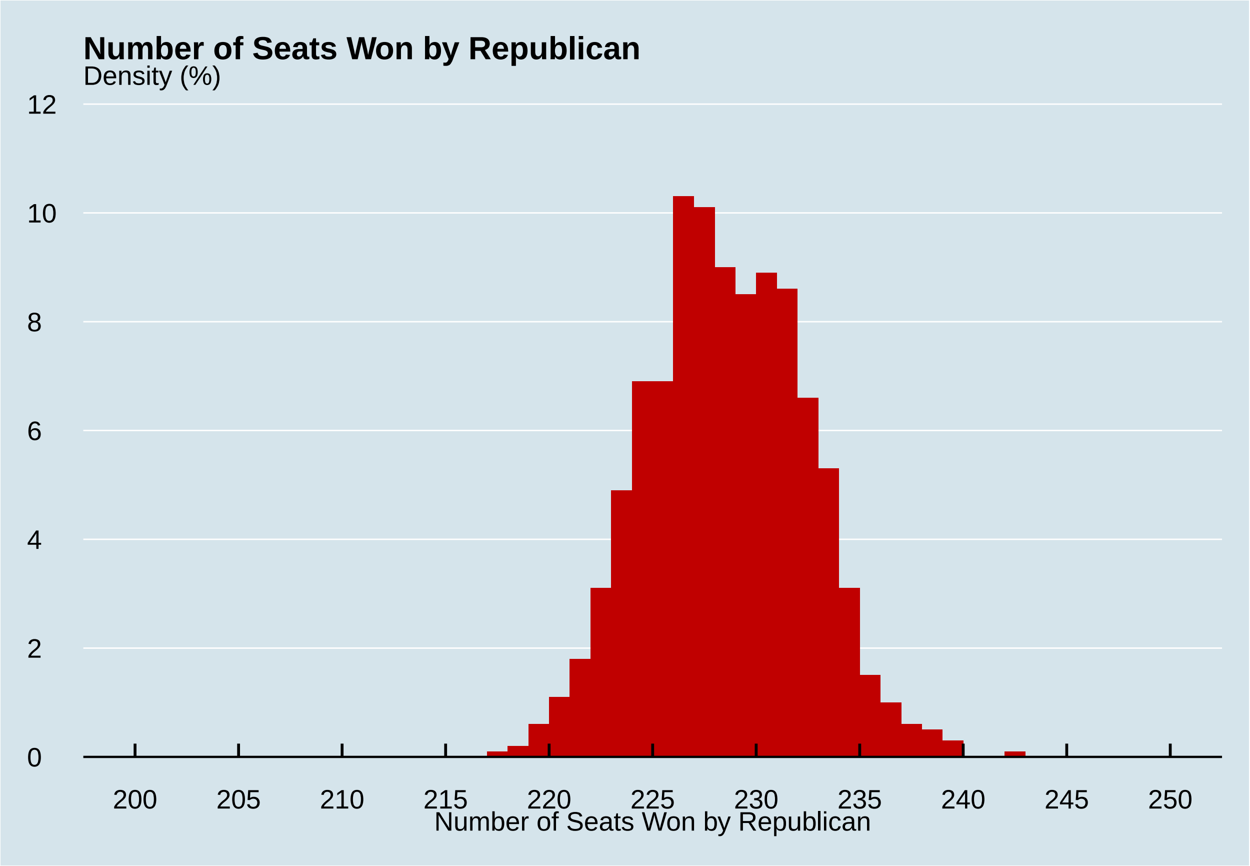
(c) Based on the 1,000 iterations of your simulation, generate a histogram (probability distribution) of the number of seats the Republican party will win in aggregate in the election.

```
rep_win <- colSums(simulation) %>% as_tibble()

histogram <- rep_win %>%
  ggplot() +
    geom_histogram(aes(x = value, y = ..density.. * 100),
                   binwidth = 1, boundary = 0, color = NA, fill = "#C00000") +
    scale_x_continuous(breaks = seq(200, 250, 5), labels = seq(200, 250, 5),
                       limits = c(200, 250)) +
    scale_y_continuous(breaks = seq(0, 12, 2), labels = seq(0, 12, 2),
                       limits = c(0, 12), expand = c(0, 0, 0.025, 0)) +
    labs(title = "Number of Seats Won by Republican",
         subtitle = "Density (%)",
         x = "Number of Seats Won by Republican") +
    theme_economist() +
    theme(plot.title = element_text(size = 12, hjust = 0),
          plot.subtitle = element_text(size = 10, hjust = 0),
          axis.title.x = element_text(size = 10),
          axis.title.y = element_blank(),
          axis.text.x = element_text(size = 10),
          axis.text.y = element_text(size = 10),
          panel.grid.major.y = element_line(color = "white", size = 0.25, linetype = "solid"))

# ggsave("histogram.pdf", histogram,
#             width = 10, height = 10, units = "cm",
#             dpi = 300, device = cairo_pdf)

histogram
```



(d) Make two key and distinct observations about your histogram.

Answer: Two key observations about the data are: The histogram is approximately normal. Also, almost the entire distribution is above 218 (seats needed to win) so it is extremely likely that the Republicans win.

(e) Based on the 1,000 iterations of your simulation, report the mean, median, and mode of the number of seats that Republicans win in total.

```
# Mean
mean(rep_win$value)
```

```
## [1] 229.052
```

```
# Median
median(rep_win$value)
```

```
## [1] 229
```

```
# Mode
rep_win %>%
  group_by(value) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  .[1, 1] %>%
  as.numeric()
```

```
## [1] 227
```

Answer: Mean: 229, Median: 229, Mode: 227

(f) Based on your basic simulation in this part of the problem set, what is the probability that the Republican party wins a majority of seats in the U.S. House of Representatives?

```
## [1] 99.9
```

Answer: 99.9%.

(g) A friend of yours who (unfortunately) is not enrolled in API-209 opens your Rmd file and notices that there are only 214 congressional districts in which the Republican candidate is reasonably likely to win (with probability greater than 75 percent). Your friend observes, "Why does everyone predict that the Republicans will win at least the 218 seats necessary to take control of the U.S. House of Representatives, when they are strong favorites in only 214 individual elections?" Respond to your friend in a short paragraph (2-3 sentences).

Answer: Given that the republicans are strong favourites in 214 districts, they only need an additional 4 district wins to take the necessary control of the US House of Representatives. The odds of them winning at least 4 out of the remaining 221 seats is extremely high or almost certain.

(h) Compare your histogram of the election outcomes to FiveThirtyEight's histogram, available here. Note that FiveThirtyEight uses essentially the same individual win probabilities in its simulation as you did in your simulation. (That's because we downloaded those win probabilities directly from them!) Compared to your analysis, does FiveThirtyEight's simulation suggest greater uncertainty in the aggregate outcome, less uncertainty, or the same uncertainty? Explain in 1-2 sentences.

Answer: Greater uncertainty.The FiveThirtyEight histogram mentions that in 80% of the possible scenarios, republicans win between 213 and 245 seats whereas in our model 100% of possible scenarios involve republicans winning a number of seats in this range. In other words, our histogram has a narrower base and therefore lesser uncertainty.

Note that the reasons for the differences in the distributions of outcomes are likely related to the factors you explored in Part 1 and will explore again in Part 3. Let's go to Part 3 now.

Part 3 – Incorporating Correlations Among Election Outcomes into Your Basic Model

Your basic model in Part 2 assumed that the outcomes of individual elections were independent events. This implies that knowing that a Republican candidate won in one congressional district – even if she was a huge underdog – did not affect our assessment of the likelihood that the Republican candidate in a different, nearby district would win. In other words, each election was determined by a separate coin flip, where the coin was weighted in a unique way for each candidate but the outcome of one flip had no impact on the outcome of another flip.

Election modelers, such as the team at FiveThirtyEight, do not generally believe that the assumption of independence holds. Instead, they attempt to incorporate dependence, or correlation, into their models. For example, in discussion on FiveThirtyEight's modeling of the 2016 Presidential Election just before that election took place, Nate Silver wrote [9]:

> "State outcomes are highly correlated with one another, so polling errors in one state are likely to be replicated in other, similar states... [Y]ou shouldn't count on states to behave independently of one another, especially if they're demographically similar. If Clinton loses Pennsylvania despite having a big lead in the polls there, for instance, she might also have problems in Michigan, North Carolina and other swing states... [A]ssumptions about the correlation between states make a huge difference."

In this part, we'd like you to model correlation among election outcomes and explore how incorporating correlation affects your predictions for the aggregate outcome on Election Day.

- `census_division` provides the U.S. Census Division for each of the states. The U.S. Census Bureau defines nine divisions, ranging from "New England" (Division 1) to "Pacific" (Division 9).
- In Part 2, you simulated the election by selecting a unique and independent random number for each election in each congressional district. We are going to abandon that approach here and instead will select a random number for each Census division, not for each individual congressional district, in each iteration. By selecting one random number for each Census division, we will ensure that the outcomes of the elections within each division are correlated[10].
- Simulate the election 1,000 times! As before, you should have a matrix of zeroes and ones, corresponding to congressional districts in which the Republican candidate loses and wins, respectively.

After simulating the election 1,000 times, please answer the following questions:

```r
simulation_2 <- midterm_election_2022 %>%
  arrange(census_division, rep_win_prob) %>%
  group_by(census_division) %>%
  # Create within-group id
  # (... because congressional districtsare not distinct)
  mutate(id = row_number())

rep_win_prob_2 <- simulation_2 %>%
  mutate(swinger_id = id,
         swinger_prob = rep_win_prob) %>%
  select(census_division, swinger_id, swinger_prob)

for (i in 1:1000){

  tmp <- simulation_2 %>%
    summarise(no_of_districts = max(id)) %>%
    rowwise() %>%
    mutate(swinger_id = sample(1:no_of_districts, 1)) %>%
    select(-no_of_districts)

  simulation_2 <- simulation_2 %>%
    left_join(tmp, by = c("census_division")) %>%
```

[9] https://fivethirtyeight.com/features/election-update-why-our-model-is-more-bullish-than-others-on-trump/

[10] By selecting one random number for each Census Division, we ensure that, if a particular Republican candidate wins his or her election, all other Republicans in the same division who have higher win probabilities will also win their elections. Similarly, we ensure that, if a particular Republican candidate loses his or her election, all other Republicans in the same division who have lower win probabilities will also lose their elections. Our method still allows some Republicans to win and others to lose within the same division, but the outcomes will be highly correlated. Note that there are certainly other ways to model correlations among outcomes. You might consider them in Part 4.

```
    left_join(rep_win_prob_2,
              by = c("census_division", "swinger_id")) %>%
    mutate(sim = ifelse(rep_win_prob >= swinger_prob, 1, 0)) %>%
    select(-c("swinger_id", "swinger_prob"))

  colnames(simulation_2)[ncol(simulation_2)] <- paste0("sim", i)

}
```

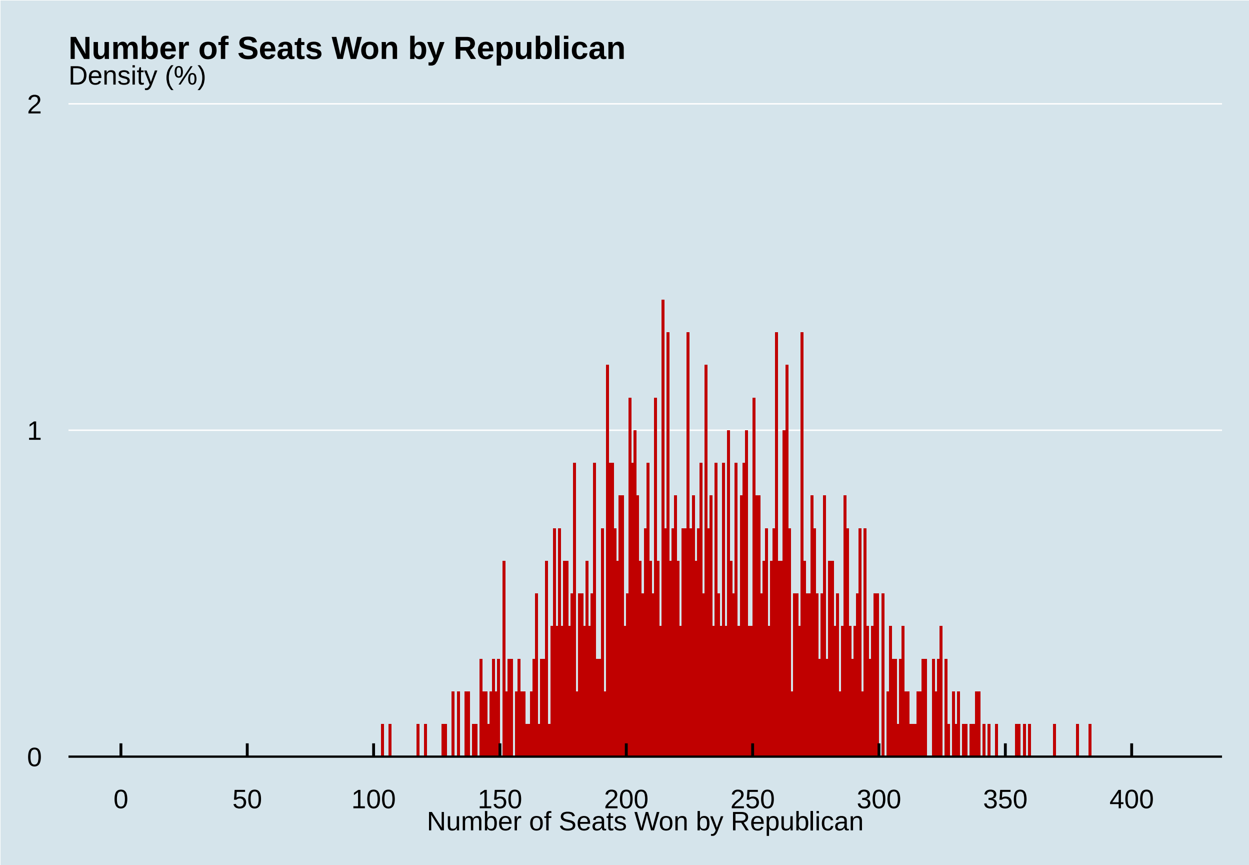
(a) Based on the 1,000 iterations of your simulation, generate a new histogram (probability distribution) of the number of seats the Republican party will win in aggregate in the election.

```
# Enter only code here.

simulation_2_gdata <- simulation_2[, 6:1005]
rep_win_2 <- colSums(simulation_2_gdata) %>% as_tibble()

histogram_2 <- rep_win_2 %>%
  ggplot() +
    geom_histogram(aes(x = value, y = ..density.. * 100),
                   binwidth = 1, boundary = 0, color = NA, fill = "#C00000") +
    scale_x_continuous(breaks = seq(0, 415, 50), labels = seq(0, 415, 50),
                       limits = c(0, 415)) +
    scale_y_continuous(breaks = seq(0, 10, 1), labels = seq(0, 10, 1),
                       limits = c(0, 2), expand = c(0, 0, 0.025, 0)) +
    labs(title = "Number of Seats Won by Republican",
         subtitle = "Density (%)",
         x = "Number of Seats Won by Republican") +
    theme_economist() +
    theme(plot.title = element_text(size = 12, hjust = 0),
          plot.subtitle = element_text(size = 10, hjust = 0),
          axis.title.x = element_text(size = 10),
          axis.title.y = element_blank(),
          axis.text.x = element_text(size = 10),
          axis.text.y = element_text(size = 10),
          panel.grid.major.y = element_line(color = "white", size = 0.25, linetype = "solid"))

histogram_2
```



Answer:

Histogram above

(b) Compare your new histogram to the histogram you generated in Part 2. Briefly describe the similarities and the differences.

Answer:

This histogram is also approximately normal. However, the second histogram is more spread out than the first one.

(c) Does incorporating correlations among outcomes within each Census District increase the uncertainty in the aggregate outcome, decrease the uncertainty in the aggregate outcome, or have no effect on the aggregate outcome? Explain.

Answer:

Increases the uncertainty in the aggregate outcome, as shown by the second histogram having a wider distribution. This is because we have now introduced additional variability (randomness) to the threshold level for which Republican win is determined.

(d) Based on your modified simulation in this part of the problem set, what is the probability that the Republican party wins a majority of seats in the U.S. House of Representatives?

```
# Enter only code here.

rep_win_2 %>%
  summarise(sum = sum(value > 218) / 1000 * 100) %>%
  as.numeric()
```

```
## [1] 60.1
```

Answer:

There is a 60.2% probability that the Republican party of seats in the US House of Representatives

---

Part 4 – Predicting Tuesday’s Election Results (with Prizes!)

It is now time to summon all your powers to predict what will happen on Election Day. It is your (and your group’s) moment to shine!

To make your prediction, you should start with the basic simulation model from Part 2 or the modified simulation model from Part 3 (your choice). But you should adapt it as you see fit. A few elements to consider and perhaps incorporate into your enhanced model:

1. Do you think the current polls – which form the basis of estimating the Republican candidates’ win probabilities – are systematically biased in favor of one party or the other? Perhaps current polls are missing something regarding the effectiveness of get-out-the-vote efforts, or the polls do not reflect the impact of recent news events[11].
2. Does the model adequately account for uncertainty due to sampling fluctuation? For example, does the model understate the degree of uncertainty? Or is the outcome more certain than the basic model expects[12]?
3. Should the model incorporate correlations among outcomes in different congressional districts? You could use the method from Part 3 directly, or modify the method as you see fit.

Once you have enhanced the simulation model however you and your group consider appropriate, please run the simulation 1,000 times (or more if you think this would help). Then answer the following questions:

(a) [Primary contest question] **Report your best guess of the number of seats the Republican party will win on Nov. 8**. Important: Your response must be supported by your enhanced simulation model and your written response to question (d) below. For example, your group’s prediction could be the mean, median, or mode from your enhanced simulation model.

```
# Enter only code here.
midterm_election_2022$rep_win_prob_new <- midterm_election_2022$rep_win_prob + 0.01

simulation_4 <- midterm_election_2022 %>%
  arrange(census_division, rep_win_prob) %>%
  group_by(census_division) %>%
  # Create within-group id
  # (... because congressional districtsare not distinct)
  mutate(id = row_number())

rep_win_prob_4 <- simulation_4 %>%
  mutate(swinger_id = id,
         swinger_prob = rep_win_prob) %>%
  select(census_division, swinger_id, swinger_prob)
```

[11] To correct for systematic bias, consider increasing or decreasing the Republican candidate’s win probability in each congressional district by a certain amount.

[12] To incorporate greater uncertainty in your model, consider shifting the Republican candidate’s win probability in each congressional district closer to 50 percent. To reduce the uncertainty in your model, consider shifting the Republican candidate’s win probability in each district closer to 100 percent if she is the leader or closer to 0 percent if she is behind in that congressional district.

```
set.seed(2)
for (i in 1:1000){

  tmp <- simulation_4 %>%
    summarise(no_of_districts = max(id)) %>%
    rowwise() %>%
    mutate(swinger_id = sample(1:no_of_districts, 1)) %>%
    select(-no_of_districts)

  simulation_4 <- simulation_4 %>%
    left_join(tmp, by = c("census_division")) %>%
    left_join(rep_win_prob_4,
              by = c("census_division", "swinger_id")) %>%
    mutate(sim = ifelse(rep_win_prob_new >= swinger_prob, 1, 0)) %>%
    select(-c("swinger_id", "swinger_prob"))

  colnames(simulation_4)[ncol(simulation_4)] <- paste0("sim", i)

}

simulation_4_gdata <- simulation_4[, 7:1006]
rep_win_4 <- colSums(simulation_4_gdata) %>% as_tibble()

median(rep_win_4$value)
```

```
## [1] 298
```

Answer: Our model predicts that the republic party will win 298 seats in the midterm election.

(b) [Contest tie breaker] Based on your enhanced simulation model in this part of the problem set, **predict which of the following five congressional districts the Republican candidate will win**: Texas 34th, Pennsylvania 7th, New York 19th, Virginia 2nd, Rhode Island 2nd.

```
# Enter only code here.

S4_TX <- simulation_4 %>%
  filter(state == "TX", congressional_district == 34)
median(as.numeric(S4_TX[,7:1006]))
```

```
## [1] 0
```

```
S4_PA <- simulation_4 %>%
  filter(state == "PA", congressional_district == 7)
median(as.numeric(S4_PA[,7:1006]))
```

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

```
S4_NY <- simulation_4 %>%
  filter(state == "NY", congressional_district == 19)
median(as.numeric(S4_NY[,7:1006]))
```

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

```
S4_VA <- simulation_4 %>%
  filter(state == "VA", congressional_district == 2)
median(as.numeric(S4_VA[,7:1006]))
```

```
## [1] 0
```

```
S4_RI <- simulation_4 %>%
  filter(state == "RI", congressional_district == 2)
median(as.numeric(S4_RI[,7:1006]))
```

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

Answer:

Republics will win the seats PA 7, NY19, RI 2.

(c) Based on your enhanced simulation model in this part of the problem set, what is the probability that the Republican party will win a majority of seats in the U.S. House of Representatives? Please report the probability in the form of a number between 0 and 1.

```
# Enter only code here.

rep_win_4 %>%
  summarise(sum = sum(value > 218) / 1000 * 100) %>%
  as.numeric()
```

```
## [1] 97
```

Answer: There is a 97% chance of the Republican party winning a majority in the US House of Representatives

(d) In 1-3 paragraphs, please summarize the ways in which your group enhanced the basic simulation model from Parts 2 and 3 and how your team decided on your entry to question (a) just above (i.e., the predicted number of seats that Republican party candidates will win on Election Day).

Answer:

Our model assumes the correlation model, with an additional assumption to reflect 1%p higher win probability for Republicans. This additional assumption is based on the observation that political / election polls tend to understate actual support for the Republican party (e.g. during / following the 2016 election). It is difficult to quantify the exact extent of understatement, so we have conservatively assumed 1%p.

(e) **VERY IMPORTANT:** In addition to submitting answers to these questions with your problem set, one person from each group should submit your official entry into the contest by 8:00 a.m. on Tuesday, Nov. 8 (Election Day!).

https://harvard.az1.qualtrics.com/jfe/form/SV_9BqH8XNA2xkPEay

Prizes will be awarded based on your response to part (a) and, in the event of a tie, part (b). The biggest prize of all is perpetual bragging rights.

Good luck!

QUESTION 2 – WORKER SAFETY IN THE ERA OF MACHINE LEARNING

At the beginning of the semester, we distinguished between 3 uses of statistical inference: descriptive, causal, and predictive. The goal of this question is to help you get familiar with some of the key concepts and ideas behind predictive use of evidence. When you hear terms like machine learning, data science, big data, predictive analytics, they are all about using statistical techniques to predict some outcome of interest. While the techniques used, size of data sets, and corresponding demands of the programming languages vary from discipline to discipline, the key underlying ideas are similar.

The setting in which you will apply these concepts is the Occupational Safety and Health Administration (OSHA), the government agency in charge of ensuring safe and healthful working conditions in the United States. You will be asked to read a case and do some empirical analysis in preparation for the discussion about the case that we will have in our classes to be held on **November 8 and November 15**.

The plan for this prediction unit is as follows:

- **Problem set #10 (this one, due Nov 8)**: You will read the OSHA case and answer some questions related to it.
- **Class #21 (Nov 8):** Class will be an overview of key concepts about prediction and machine learning, with some discussion about how they apply to the case.
- **Problem set #11 (due Nov 15):** You will do the empirical work related to the OSHA case. To do this, you will be given an OSHA data set on injuries and inspections in work establishments. The goal is to help OSHA improve the way it selects establishments to be inspected. They are under increasing pressure because some policymakers are arguing that OSHA inspections frequently represent an unnecessary cost to businesses. In this problem set, you will be asked to use various algorithms designed to predict a key outcome variable that OSHA could take into account in deciding which establishments to select for inspections.
- **Class #22 (Nov 15):** Class will debrief your results from problem set #11, probe deeply into your findings, and step back to think about broader policy uses of prediction.

Note: Please have a **copy** of your answers to this question **handily available for classes on November 8 and November 15**. This will allow you to participate more fully in the discussion of the case in class.

The work for this problem set is as follows:

We will be talking about using data to make predictions. So instead of looking for causal links to estimate the impact of a program/intervention, we are interested in questions where we predict the characteristics or behavior of individual agents.

For example, credit card companies might use your credit card transaction data to predict the probability that you will default on your credit card. In the OSHA case, OSHA uses data about work establishments to predict which of them are at risk of safety violations and takes this information into account in deciding which establishments to inspect.

1. Read the case “Improving Worker Safety in the Era of Machine Learning” (HBS Case #N9-618-09). The case can be found on the readings column on the Canvas home page.
2. Summarize the key advantages and disadvantages of each of the four targeting approaches compared to OSHA’s current approach, which amounts to selecting randomly from the list of sites with high historical injury rates.

Answer:

Approach 1: Use local discretion

a. Description: "The 81 local OSHA Area offices could use their discretion about which workplaces to inspect each year."

b. Advantages:

· The local inspectors could leverage their local knowledge to decide which firms to inspect. They know who the local businessman/managers are and what are their attitudes towards safety.

· Apart from having access to formal accident reports, local inspectors know about accidents that were informally alluded to via their informal networks

· Local inspectors better understand the local culture in different industries, and they have direct access to workers and unions, who could constantly update their biases.

c. Disadvantages:

· Local inspectors are prone to be corrupt. They could avoid inspecting specific sectors or firms if they are given the monetary incentives to look the other way.

· Local inspectors already have preconceived biases about what are the most risky industries. Therefore, they may continue inspecting a handful of companies every year.

· Local inspectors are active members of the community. They may avoid inspecting a particular firm because they are perhaps related to someone in the board or in the management team. In turn, they could be motivated to inspect a firm if they have a personal issue with someone in its team.

Approach 2: Highest historical injury rates

a. Description: "OSHA could use the most recent data available, which tended to be two years old, to prioritize inspecting workplaces based on their historical injury rates, starting with the workplace with the worst record."

b. Advantages:

· With approach 2 we would inspect those companies that historically show high injury rates. It would allow OSHA to better understand what is going on in these specific firm that is triggering a large number of incidents

· Companies would be incentivized to maintain a low number of injuries in order to decrease the probability of being inspection.

c. Disadvantages:

Garbage-in-garbage-out: There would be an incentive to underreport the number of accidents. In that case, the data would no longer be reliable. The OSHA would have to invest in implementing checks and balance in order to ensure that all accidents are being accounted for.

· Firms that have not had accidents recently would know that they are off the hook. The probability of facing an inspection would be extremely low. Therefore, they may start reducing their efforts to improve safety at the workplace.

· Workplaces in high hazard industries are only required to report their injury data every three years. Therefore, the OSHA would have no visibility into current trends. If it is focusing on firms that had high injury rates 3 years ago, it may miss that in the present injuries are rising in order firms/industries.

Approach 3: Highest predictive injury rates

a. Description: OSHA could use regression and/or machine learning techniques applied to the most recent data available and develop an algorithm that would predict the injury rates likely to occur that year at each workplace. OSHA could then inspect the workplaces predicted to have the highest injury rates."

b. Advantages:

· If a good algorithm is developed then the OSHA could predict injury rates with a reasonable degree of certainty and they could inspect firms that show a high risk.

· If various variables are considered in the algorithm then it would allow the OSHA to adopt a more objective mechanism to select future inspections. This approach is way more objective than simply asking the local inspector to use their judgement to pick what firms to inspect.

c. Disadvantages:

· The regression model may not include the ideal variables to predict the injury rates likely to occur that year at each workplace.

· The regression model is likely to include human bias in its design. For instance, it may overweight industry type as one of the explanatory variables simply because the individuals designing the regression model believe industry type is more important than other variables.

Approach 4: Stratified random sampling of all workplaces

a. Description: "Workplaces could be categorized into groups ("stata") based on historical injury rates (approach 2) or predicted injury rates (approach 3). OSHA could then randomly select a percentage of workplaces randomly from each stratum, with higher percentages being drawn from higher-injury strata."

b. Advantages:

· A percentage of workplaces is randomly selected from each stratum, therefore, all companies have the chance of being selected for inspection. As a result, firms would in general feel the need to be better prepared in case they receive a visit from the inspector.

· Randomly selecting companies reduces human bias from the process.

c. Disadvantages:

· One could argue that resources are not entirely invested in companies with the highest number of injuries.

· The OSHA would not be able to target firms with a particular characteristic.

3. Please fill in this **survey**, based on your answers above. **THIS IS THE ONLY QUESTION THAT YOU SHOULD ANSWER INDIVIDUA**

OPTIONAL

The world of data science and machine learning is growing rapidly. If you are interested in exploring further, here are some recommended readings and videos:

Readings "A Guide to Solving Social Problems with Machine Learning" by Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. December 8, 2016. Accessible at https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning "Machine Learning: An Applied Econometric Approach" by Sendhil Mullainathan and Jann Spiess. Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106 Videos Select those that look most interesting to you from the appendix below

Appendix – More resources if you are interested in Data Science (compiled by Erich Nussbaumer)

Useful to watch before class:

- Non-technical introduction to prediction and what it is used for: https://www.youtube.com/watch?v=m30LxzzbRik
- What is prediction & example on prediction of spam email: https://www.coursera.org/learn/practical-machine-learning/lecture/116Tb/what-is-prediction
- What are key characteristics of prediction (Min 3:30 to 5:10): https://www.youtube.com/watch?v=IqJ2CQTLzZk
- Steps of building predictive models: https://www.coursera.org/learn/practical-machine-learning/lecture/YYmBu/relative-importance-of-steps
- Visual guide to machine learning (intro): http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Useful to watch after class:

- Introduction to different machine learning methods: https://www.youtube.com/watch?v=IpGxLWOIZy4
- Introduction to machine learning (MIT): https://www.youtube.com/watch?v=h0e2HAPTGF4
- Learn more about neural networks: https://www.youtube.com/watch?v=ILsA4nyG7I0 or https://www.youtube.com/watch?v=Q9Z20HCPnww

QUESTION 3 - FINAL EXERCISE

This part of the problem set is designed to be completed with your final exercise team. As opposed to other problem set questions where you are asked to write your answers in your own words, all team members can submit identical answers for this question. But please submit answers individually (as part of the problem set you submit) this time to facilitate the grading.

Note: The goal of this question is to help you advance in the final exercise, so you increase your chances of producing a final product you are proud of. Don't feel too constrained by the specific prompts you see below. You should try to answer each of the prompts but your team should decide how much time it is worth to spend at this time on each of the items below. Ultimate goal is to nudge you in the direction of making progress.

Link to final exercise is here.

OPTION 1 - Mongolia (Macro)

1. Review the list of final exercise items from the previous problem set and complete those items that you were not able to complete (skip this step if you did them all).
2. Identify the most important variables in the datasets you plan to work with. Familiarize yourself with their summary statistics.
3. Conduct some preliminary analysis for each of the 3 sectors you were asked to examine. Does it give you any initial clues about your diagnostic?
4. Choose 3-5 peer countries to compare to Mongolia. Why did you choose them? What metrics will you be comparing? Are they aspirational comparison countries or benchmark comparison countries? If they are aspirational, what time periods should you be looking at?
5. Choose some of the more straightforward tests you identified last week and run them. What do the results tell you? Interpret the results in technical terms (with the appendix in mind) and also in more practical language.
6. If you are struggling to find the data to run an ideal test, assign someone in your group to think of an alternate plan. Be creative, work with what you have, and remember that rarely does the perfect data exist in the real world.
7. Draft your analysis plan. This implies translating each of the points listed in the assignment into tasks that can be operationalized, i.e., list the tabulations, cross-tabulations, graphs, and regressions you envision doing for the final memo. You don't have to complete all tests/analyses, but you should have completed some, and have in mind the key things you have yet to analyze.
8. Decide how you plan to organize the rest of the work. Who will do what? Establish some deadlines.

Please enter your answers here

OPTION 2 - Health in Brazil (Micro)

1. Review the list of final exercise items from the previous problem set and complete those items that you were not able to complete (skip this step if you did them all).
2. Identify the most important variables in the datasets you plan to work with. Familiarize yourself with their summary statistics.
3. Think about for which geographical and demographic levels do you want to estimate the disease burden.
4. For the same geographic level you selected above, describe the health infrastructure both in 2010 and 2020. How it evolved?
5. In both SIH and SIH datasets you are going to find some municipalities that have "-" in certain columns. This could be due to missing data or could mean that the value is zero. What are the implications of each one of these cases?
6. Reflect about how are you going to use the available data to predict the burden of diseases at the municipal level.
7. If you are struggling to find the data to run an ideal test, assign someone in your group to think of an alternate plan. Be creative, work with what you have, and remember that rarely does the perfect data exist in the real world.
8. Draft your analysis plan. This implies translating each of the points listed in the assignment into tasks that can be operationalized, i.e., list the tabulations, cross-tabulations, graphs, and regressions you envision doing for the final memo. You don't have to complete all tests/analyses, but you should have completed some, and have in mind the key things you have yet to analyze.
9. Decide how you plan to organize the rest of the work. Who will do what? Establish some deadlines.

Please enter your answers here

1. Review the list of final exercise items from the previous problem set and complete those items that you were not able to complete (skip this step if you did them all).

    1. DONE

2. Identify the most important variables in the datasets you plan to work with. Familiarize yourself with their summary statistics.

    1. Variables to estimate disease burden

        1. Population
        2. Gender
        3. Years of education
        4. Age
        5. Cause of death
    2. Variables to predict disease burden in municipalities
        1. Type and number of health care workers
        2. Type and number of health faciltiies
        3. Type and number of health equipment

3. Think about for which geographical and demographic levels do you want to estimate the disease burden.
    1. Rurality
    2. Gender
    3. Education level
    4. Income level
    5. Age

4. For the same geographic level you selected above, describe the health infrastructure both in 2010 and 2020. How it evolved?
    1. In terms of health professionals, 2020 documented whether or not municipalities had nutritionists, which will be helpful to see in terms of nutritional issues. Otherwise, there are more of every kind of doctor except, interestingly, general surgeons. This was done by looking at the change in numbers between the two years. This same analysis will be done for equipment and facilities.

5. In both SIH and SIM datasets you are going to find some municipalities that have “-” in certain columns. This could be due to missing data or could mean that the value is zero. What are the implications of each one of these cases?
    1. If this means missing data that will skew our results as we will not include information about this area but then try to draw conclusions about it. If this means zero this means that we will have to rewrite the column to include a zero in place of that in order to run calculations (e.g. mean) on that column when we are trying to summarize the data. Notably, since we don’t know which is which we may incorrectly be taking some data to mean zero and others to mean missing data. We will also check how much of this information is labeled as “-” to see the extent to which it impacts our analysis.

6. Reflect about how are you going to use the available data to predict the burden of diseases at the municipal level.
    1. The existing data gives us information about the main causes of death among different demographic groups. Assuming the environment and situation in these municipalities has not drastically changed in 10 years (like what might happen with massive reform or a large natural disaster), we can predict what are the most likely causes of hospitalizations and death in municipalities in the coming years, which helps us prepare for them. We are going to look at the data to determine if there are trends (for example, children and older people may be more susceptible to infectious diseases) which helps us propose programs to curb this in the future.
    2. We are going to use the prediction algorithm on a subset of relevant independent variables to draw conclusions about the future disease burden and structure to address it.

7. If you are struggling to find the data to run an ideal test, assign someone in your group to think of an alternate plan. Be creative, work with what you have, and remember that rarely does the perfect data exist in the real world.
    1. Understood.

8. Draft your analysis plan. This implies translating each of the points listed in the assignment into tasks that can be operationalized, i.e., list the tabulations, cross-tabulations, graphs, and regressions you envision doing for the final memo. You don’t have to complete all tests/analyses, but you should have completed some, and have in mind the key things you have yet to analyze.
    1. Graphs on numbers of hospitalizations and deaths broken down by cause of death
        1. Those causes then individually broken down by different demographic factors
    2. Comparing the numbers of workers, facilities, and equipment across municipalities

   3. Utilize this information and correlate it with (a) to find relevant independent variables
   4. Run prediction algorithm on this subset of independent variables to generate predicted disease burden
   5. Determine which needs different municipalities have and how this helps us predict future disease burden and structural needs

9. Decide how you plan to organize the rest of the work. Who will do what? Establish some deadlines.

   1. Create graphs by compiling and summarizing data using R (Shreya, Masato, Luisa)
   2. Create models to determine future disease burden and areas of need (Shreya, Masato, Luisa)
   3. Draft policy proposals and memo based on model (Kwang Jun, Carlos, Bharath)
   4. Create proposal slide deck and notes ((Kwang Jun, Carlos, Bharath)
   5. Deadline for completion of R work and graph creation: November 13
   6. Deadline for policy memo draft: Nov 18
   7. Deadline for slide deck draft: Nov 22
