PROBLEM SET 2

**Due on Tuesday, Sept 15, 2020, 8:30 am.**

I - INSTRUCTIONS

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

1. **Download this Word document file into your computer**
2. **Insert all your answers into this Word document**. Guidance [here](https://www.dropbox.com/s/ox9fhmbpvy2viw5/How%20to%20incorporate%20handwritten%20work%2C%20Stata%20output%2C%20and%20screenshot%20images.pdf?dl=0) on how to insert non-Word objects such as handwritten work or screenshot images in your answers.
3. **Once your document is complete, please save it as a PDF**. This is important to make sure all your work is preserved in the process of submission to Canvas.
4. **Please submit an electronic copy of the PDF (and any separate requested files) to the Canvas course page**.

II - IDENTIFICATION

1. **Your information**

|  |  |
| --- | --- |
| Your Last Name: |  |
| Your First Name: |  |

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

|  |
| --- |
|  |

1. **Compliance with Harvard Kennedy School Academic Code[[1]](#footnote-1) (mark with an X below)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Yes** | **No** | | I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code |  |  | |

**Table of Contents**

[PART 1 – ELECTIONS IN VENEZUELA 3](#_Toc50615492)

[PART 2 – IN SEARCH OF A COVID VACCINE 5](#_Toc50615493)

[PART 3A – ONLINE MODULE ON ESTIMATORS 7](#_Toc50615494)

[PART 3B – ONLINE MODULE ON SAMPLING DISTRIBUTIONS 7](#_Toc50615495)

[PART 4 – LEARNING ABOUT SAMPLING THROUGH COUNTRY-LEVEL COVID CASES 8](#_Toc50615496)

[PART 5 - LEARNING ABOUT INEQUALITY IN THE WORLD 10](#_Toc50615497)

[Appendix 1 - Venezuelan Elections – Simulating Election Day 1,000 times 16](#_Toc50615498)

[Appendix 2 – Making histograms in Excel 17](#_Toc50615499)

[Appendix 3 – Tips to code in R for World Inequality Question 18](#_Toc50615500)

# PART 1 – ELECTIONS IN VENEZUELA

Venezuelans voted on a referendum to recall President Hugo Chavez on August 15th, 2004. The official results gave the victory to Pres. Chavez (59% in favor of keeping him in office and 41% against). Opposition supporters argued that there was fraud. This question asks you to put yourself in the position of an objective statistical analyst trying to find out whether the opposition claims are supported by the evidence.

This problem set question is based on real events, but we will make a few simplifying assumptions. The voting process worked for the most part in the following manner: Each voter was assigned to one of the 8,394 voting centers nationwide. In each voting center, there were two electoral machines. Voters had to select “Yes” (the option favoring recalling the president) or “No” (the option favoring keeping the president in office).

(a) The pro-opposition governor of a state declared fraud shortly after the official results were announced. According to these results, two machines in one of the centers located in this governor’s state reported the exact same number of “Yes” votes (207). The table below describes the voting results for this center. The governor argued that it was extremely unlikely for this event to occur and said this was evidence of fraud.

|  |  |  |
| --- | --- | --- |
|  | Number of votes favoring each option | |
| Voting Center X | Yes | No |
| Machine #1 | 207 | 293 |
| Machine #2 | 207 | 293 |

Calculate the probability that both machines in this center reported exactly 207 “Yes” votes. To do this you may make the following assumptions:

* + Voters were distributed randomly across the two machines in each center, so you may assume that voting outcomes in both machines are independent.
  + For any given person, the probability of voting “Yes” is 0.40 and this probability (denoted as p) is the same for the two machines in each voting center.

|  |
| --- |
| *Show your calculations here [type them or paste a legible picture]* |

1. Calculate the probability that both machines in this center reported exactly the same number of “Yes” votes (regardless of whether it’s 207 or some other number). Call this probability q. You may use Excel to help you in these calculations, but make sure you specify in your answer the formula you used to compute the probability. You may use the same assumptions as in part (a).

|  |
| --- |
| *Show your calculations here [type them or paste a legible picture]* |

1. Other people started noticing that in their voting center both machines had the same number of “Yes” votes. Calculate how many centers you would predict would have the same number of “Yes” votes. You can make the same assumptions as in part (a), and assume that the two machines in each center received the same total number of votes (500 votes per machine) and that voting centers are independent.

|  |
| --- |
| *Show your calculations here [type them or paste a legible picture]* |

1. The number of voting centers that exhibit the same number of “Yes” votes that you calculated in (c) is an expected value. But because of random fluctuations, the actual number could be higher or lower on any given election day. This question asks you to “simulate” 1,000 times what might have happened on election day (i.e. how many voting centers would exhibit the same number of “Yes” votes across the two machines if Election Day had happened 1,000 times). To do this, use the second tab of the Excel sheet titled “simulation”. Use Appendix 1 (at the end of the problem set) for help on how to do this in Excel. Once you have 1,000 estimates for the number of voting centers that have the same number of “Yes” votes, report the following:
2. Average number of voting centers with same number of yes votes:
3. Is this average similar to the number you calculated in (c)? Why or why not?
4. Draw a histogram with the 1,000 estimates. You may do so in Excel or R (whichever is easier for you). Use Appendix 2 (at the end of this problem set) if you need help drawing histograms in Excel)

|  |
| --- |
| *Please enter your answers here* |

1. As indicated above, the opposition governor argued that it was extremely unlikely to observe two machines in a voting center report the exact number of “Yes” votes. When other voting centers reported the same, some opposition leaders started counting such centers. Suppose they counted 225 centers that exhibited the same number of “yes” votes in their two machines. Would you conclude that this is likely the results of random fluctuations (i.e. due to chance) or such an unlikely event that you think something suspicious might be going on? Explain.

|  |
| --- |
| *Please enter your answer here* |

1. The opposition wants you to tell them how large does the number of voting centers that report the same number of “Yes” votes has to be for you to conclude that something suspicious might be going on (i.e. that the deviations from what you would expect are unlikely to be explained by chance). How would you respond?

|  |
| --- |
| *Please enter your answer here* |

# PART 2 – IN SEARCH OF A COVID VACCINE[[2]](#footnote-2)

With the outbreak of COVID-19, scientists rapidly began the search for a vaccine. Many nations are on pace to release vaccines in record time; Russia has already begun inoculations. This problem set question helps you apply what you have learned about probability and decision trees to understand some of the current debates.

1. Suppose you are a pharmaceutical firm in March 2020 contemplating whether to try to develop a vaccine. You estimate it would take an investment of $100 million to do so. Assume the probability that you will be successful at developing a COVID-19 vaccine is 0.01. If you succeed, you will earn $4 billion[[3]](#footnote-3) in net profits from future sales (note this $4 billion number accounts for the initial investment plus all the costs involved in producing the vaccine). Should you try to develop the vaccine? Draw a decision tree to represent your options. Assume that you are risk neutral and your only criterion is financial.

|  |
| --- |
| *Show your decision tree and calculations here [type them or paste a legible picture]* |

1. You are doubtful about the above estimate of 0.01 probability of being successful at developing a COVID-19 vaccine. How high would the probability of success need to be for your pharmaceutical firm to want to invest in the vaccine? Make the same assumptions as above.

|  |
| --- |
| *Please enter your answer here* |

1. Now put yourself in place of governments around the world. You are worried about the health and economic consequences of COVID-19. Read [this recent article](https://www-economist-com.ezp-prod1.hul.harvard.edu/leaders/2020/08/06/the-world-is-spending-nowhere-near-enough-on-a-coronavirus-vaccine)[[4]](#footnote-4) from The Economist. Draw a decision tree that the writers in The Economist might have had in mind from the perspective of governments around the world considering how many doses of vaccine to purchase at this point. There is not a single ‘right’ answer for this decision tree, but think about the key aspects of the problem you would like to address.

|  |
| --- |
| *Please enter your answer here* |

1. The Economist claims “Instead of seeing unproven vaccines as an extravagance, the world needs to think of them as an insurance policy.” Explain what this means in the context of the decision tree you drew in (c).

|  |
| --- |
| *Please enter your answer here* |

1. Assuming the claims of The Economist are true, why do you think the world is not spending nearly enough on a coronavirus vaccine? Use concepts and ideas you have learned in other courses in your program.

|  |
| --- |
| *Please enter your answer here* |

A key question for policy makers is whether normal vaccine safety protocols should be followed, or if the severity of COVID-19 and the costs of shutdowns merit approving vaccines with less data from clinical trials. The next few questions consider the likelihood of finding rare vaccine complications as a function of the number of people enrolled in a trial. Unless otherwise specified, assume throughout that all individuals in the trial have an independent probability, **p**, of having a severe adverse complication.

A vaccine being developed by Moderna is currently undergoing Phase III trials.[[5]](#footnote-5) This trial is expected to enroll 30,000 people, but assume so far 2,000 people have been enrolled.

1. Calculate the probability the first 2,000 enrollees show **no** severe adverse complications for the following values of **p**.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | P | Probability that the first 2,000 enrollees show no complications | | 0.01 = 1/100 |  | | 0.001 = 1/1000 |  | | 0.0001 = 1/10000 |  | |

1. A recent vaccine trial has been paused after the discovery of an adverse event.[[6]](#footnote-6) Suppose this happened during the enrollment of the first 2,000 patients. A friend of yours argues this is evidence that the incidence rate is 1/2000 and is therefore relatively low. Do you find this argument convincing? Could different risk-levels of an adverse event (represented by **p**) produce this result?

|  |
| --- |
| *Please enter your answer here* |

1. What is the range of values for **p** for which we can say that the probability of seeing zero adverse complications in the first 2,000 enrollees is above 5%? Using the upper value of this bound as your estimate of **p**, and assuming 80% of your country’s total population receives the vaccine, what is the expected number of people in your country who could suffer severe adverse complications?

|  |
| --- |
| *Please enter your answer here* |

# PART 3A – ONLINE MODULE ON ESTIMATORS

The goal of this problem set question is to help you prepare you for the class on **Estimators** that will be held on the week of **September 16**. The idea is to get everyone familiar with the basics of estimators and their key properties so that we can delve deeper in class on this topic than we would be able to do if we had to go through the basics in class.

You will be asked to watch a short module and answer some questions in a quiz. The quiz results will give me information about overall performance of the class that I will use to prepare for class; your individual performance in the quiz will be registered in the system but will not count towards your grade in any way.

To get full credit for this question, you need to watch the module and complete the quiz.

The module is available [here](https://canvas.harvard.edu/courses/78089/modules/items/705076).

|  |
| --- |
| *Please enter “Done” in this field once you have completed the quiz.* |

# PART 3B – ONLINE MODULE ON SAMPLING DISTRIBUTIONS

The goal of this problem set question is to help you prepare you for the class on **Sampling Distributions** that will be held on the week of **September 14**. Rationale and Instructions are the same as for part 3A above.

The module is available [here](https://canvas.harvard.edu/courses/78089/modules/items/705079).

|  |
| --- |
| *Please enter “Done” in this field once you have completed the quiz.* |

# PART 4 – LEARNING ABOUT SAMPLING THROUGH COUNTRY-LEVEL COVID CASES

In this problem set question, you will learn some important statistical concepts related to statistical inference. The file “COVID Data - Our World in Data - Sept 2 2020.xls” contains several country-level measures related to the prevalence and severity of COVID, while also including some measures of possible determinants for outbreak severity, such as population density. Throughout this exercise, assume that this is our population of interest. Denote Y as a variable representing total cases.

1. **Explore the data set**: As always, take some time to explore the data set. Import the data set into R and do the following:
2. What are the top 5 countries in the data in terms of total COVID cases? Display their name and their cases.

|  |
| --- |
|  |

1. **Distribution of Y in the population:** Construct a relative frequency histogram of the total cases data (i.e. show proportions on the y-axis, instead of counts – see R cheat sheet for recommended method). Describe the distribution of total cases in the population, noting any unusual features. Indicate the number of observations that were used to draw this histogram. Calculate the following basic summary statistics for the population:
   1. Mean ()
   2. Standard Deviation ()

|  |
| --- |
| *Please paste your histogram here* |

|  |
| --- |
| *Please enter your answers describing histogram here* |

|  |
| --- |
| *Please enter your summary statistics here* |

1. **Distribution of Y in the sample:** Draw a random sample of 30 countries from this population.[[7]](#footnote-7) Construct a relative frequency histogram of these data. Describe the distribution of total cases in this sample. Is this distribution what you would expect? Is it similar to the distribution of Y in the population? Indicate the number of observations that were used to draw this histogram. Calculate the following basic summary statistics for the sample:
   1. Mean ()
   2. Standard Deviation ()

|  |
| --- |
| *Please paste your histogram here* |

|  |
| --- |
| *Please enter your answers describing histogram here* |

|  |
| --- |
| *Please enter your summary statistics here* |

1. **Population vs Sample:** Is the mean total cases in the above sample (e.g. 2.a) the same as the mean total cases in the population (e.g. 1.a)? Is this what you expected? Why or why not?

|  |
| --- |
|  |

# PART 5 - LEARNING ABOUT INEQUALITY IN THE WORLD[[8]](#footnote-8)

This question asks you to explore two broad views of development and characterize the extent of inequality in the world. It connects with the first class of the DEV-101 course, in which you discussed two different perspectives on economic development: (1) income/growth, and (2) capabilities.

Nobel Prize-winner Robert Lucas is a proponent of the former, centering discussion of development almost exclusively on GDP growth:

*“By the problem of economic development I mean simply the problem of accounting for the observed pattern, across countries and across time, in levels and rates of growth of per capita income.”*

The numbers involved in the cross-country variation in GDP growth, and the implications for the wealth of a country and its citizens, are staggering: Lucas said that, after wrapping your head around the scale of these divergences, it is “hard to think about anything else.” In our first API-209 problem set, by analyzing the distribution of per capita GDP worldwide, you started to do just that.

Proponents of the second perspective, though, would question whether the story is as simple as Lucas implies. Amartya Sen, another Nobel Prize winner, is the leading voice of the capabilities approach, which puts human freedom and self-realization, broadly defined, at the core of development and progress:

*“The elimination of ignorance, of illiteracy, of remediable poverty, of preventable disease, and of needless inequalities in opportunities are to be seen as objectives that are valued for their own sake. They expand our freedom to lead the lives we have reason to value, and these elementary capabilities are of importance on their own.”*

A higher GDP per capita alone might not imply improvements on all (or any) of these fronts. The goal of this problem set question is to help you learn about economic inequality in the world while you develop your R skills.[[9]](#footnote-9) You will be asked to do two things:

1. Delve deeper into the extent of income inequality in the world across countries.
2. Explore links between the two perspectives by exploring relationships between income and some basic development indicators.

**INSTRUCTIONS ON WHAT TO SUBMIT**

* A replicable .R file with proper syntax (please copy and paste in yellow box at the end of this question). The code should follow the style guide from math camp, be able to run from beginning to end (following the restart + re-run all procedure discussed in math camp), and reproduce all the results you submit in this problem set question.
* Your answers to all sub-questions below. Please paste R output (graphs, tables, etc.) into the yellow boxes below when appropriate

**GENERAL GUIDANCE FOR QUESTIONS INVOLVING R**

* Remember to consult the R resources from math camp, particularly the HKS R cheat sheet (available at [bit.ly/HKS-R](https://bit.ly/HKS-R)). The cheat sheet contains many of the functions needed to answer the questions in this problem set – some of which were not taught in math camp.
* Following the style guide for your code and routinely checking your intermediate code passes the “restart + re-run” test is not only a requirement for the final code, but also an effective way to save time. Cleanly-written and well-tested code is easier to debug and build up from.
* When pasting output or code into the yellow boxes, use a monotype font like Courier New, to ensure that it will be legible to the CA grading your work.

The data set available to you for this question comes from the World Development Indicators (WDI) produced by The World Bank. Below is the list of variables.

**Table 0 – Description of Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Values** | |
| country | Country name | Alphanumeric (Text) | |
| pop1992 | Population in 1992 (in millions) | Numeric | |
| pop2017 | Population in 2017 (in millions) | Numeric | |
| gdppc1992 | GDP per capita (PPP adjusted) in 1992 | Numeric | |
| gdppc2017 | GDP per capita (PPP adjusted) in 2017 | Numeric | |
| lifeexp | Life expectancy at birth (in years) in 2017 | Numeric | |
| nnmort | Neonatal mortality rate in 2017 | Numeric | |
| gini | Gini coefficient in 2012/2013 | Numeric (from 0 being complete equality to 1 being complete inequality) |
| corruption | Perception of corruption in 2017 | Numeric (from 0 being very corrupt to 100 being very clean) | |

Please answer the following questions:

1. First, start with similar metrics to what you worked with in the past problem set. Calculate the mean, median, 25th percentile value, and 75th percentile value for GDP per capita (PPP-adjusted) in 2017, using the WDI data. What does the fact that the mean is higher than the median tells us about the international distribution of per capita output?

|  |
| --- |
|  |

1. Inequality measures are a useful way to capture and compare information about the nature of a distribution in a single value. The inequality measure you are likely most familiar with is the Gini coefficient, which is commonly used to measure within-country inequality. The Theil index is one that can be used to measure inequality *across* countries.[[10]](#footnote-10) Using the dineq R package, calculate the Theil index for 1992 and 2017, based on PPP-adjusted GDP per capita. Has "international inequality” risen or fallen in the past 25 years?

|  |
| --- |
|  |

1. Now, let’s add weights to the countries in our dataset. Calculate the population-weighted Theil index for 1992 and 2017 based on PPP-adjusted GDP per capita. How do your results change compared to the unweighted values, and why?

|  |
| --- |
|  |

1. Now explore the Gini variable in the dataset. List the highest countries with the 3 highest and 3 lowest Gini scores. Also, find the mean and median of these values, and note any broad patterns you see browsing through.[[11]](#footnote-11)

|  |
| --- |
|  |

1. Thomas Piketty wrote in a recent paper that “China’s inequality levels used to be close to Nordic countries and are now approaching US levels,” and similarly referred to post-liberalization of India as the “billionaire Raj.” What does this quote and the analysis you did above (in b and c) suggest in terms of how inequality has evolved over time within countries vs how it has evolved across countries?[[12]](#footnote-12)

|  |
| --- |
|  |

1. Define as “corrupt” as a country that is in the group of the 20% most corrupt countries, and as “highly unequal” a country that has a Gini coefficient greater than 40. Using those binary classifications in a cross-tab, calculate P(corrupt | highly unequal) and P(highly unequal | corrupt) and explain the difference between these two conditional probabilities to someone intelligent but not well versed in statistics.

|  |
| --- |
|  |

1. Please insert your full reproducible R script below (see example in yellow box below using Courier New font; if you discover a workflow that works well for submitting R code, please contact us and/or share with your classmates in the discussion forum)

|  |
| --- |
|  |

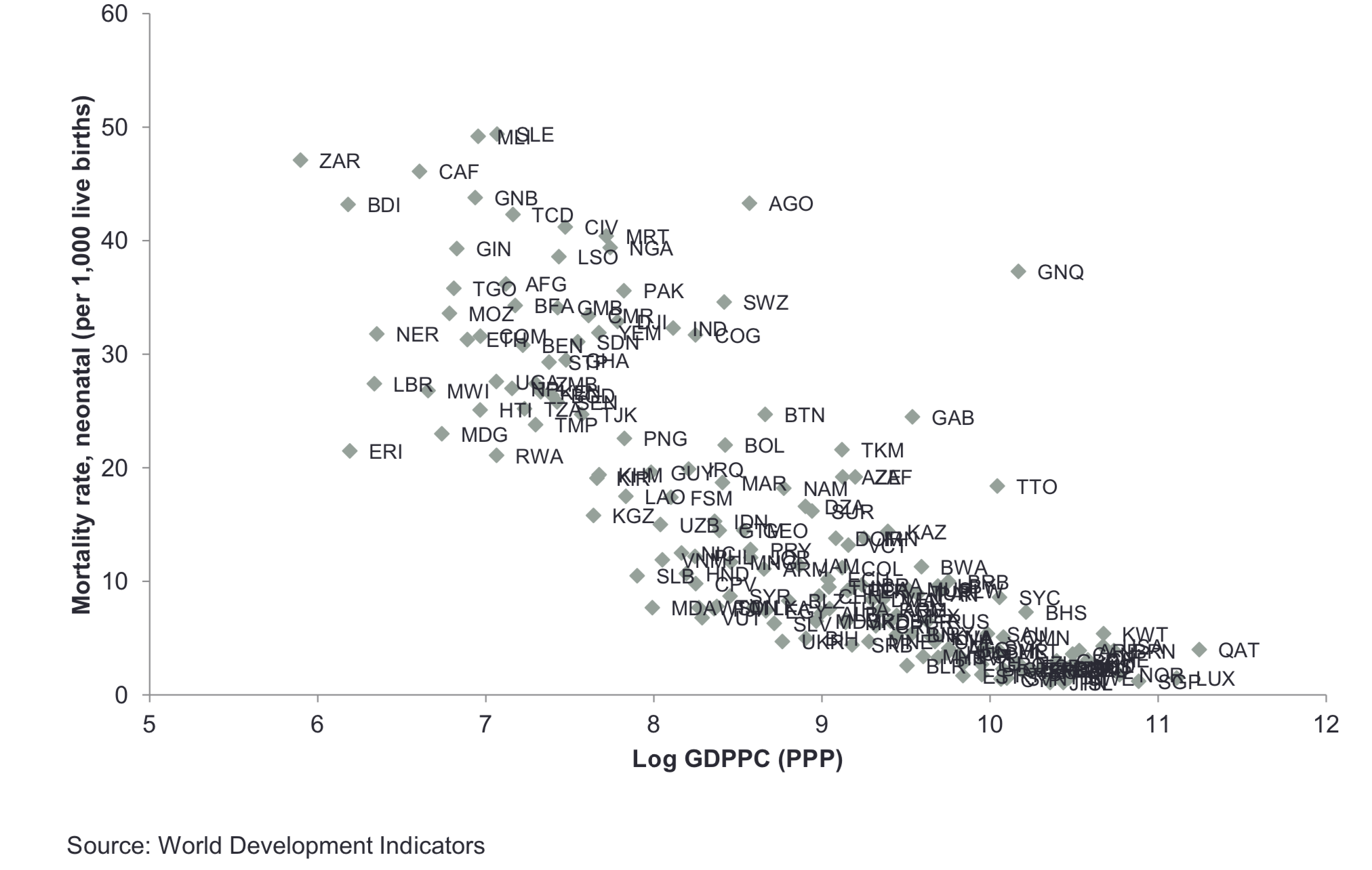
**OPTIONAL**

If you want to develop skills to produce cool and informative graphs in R and are not totally exhausted by now, here is an exercise to tackle:

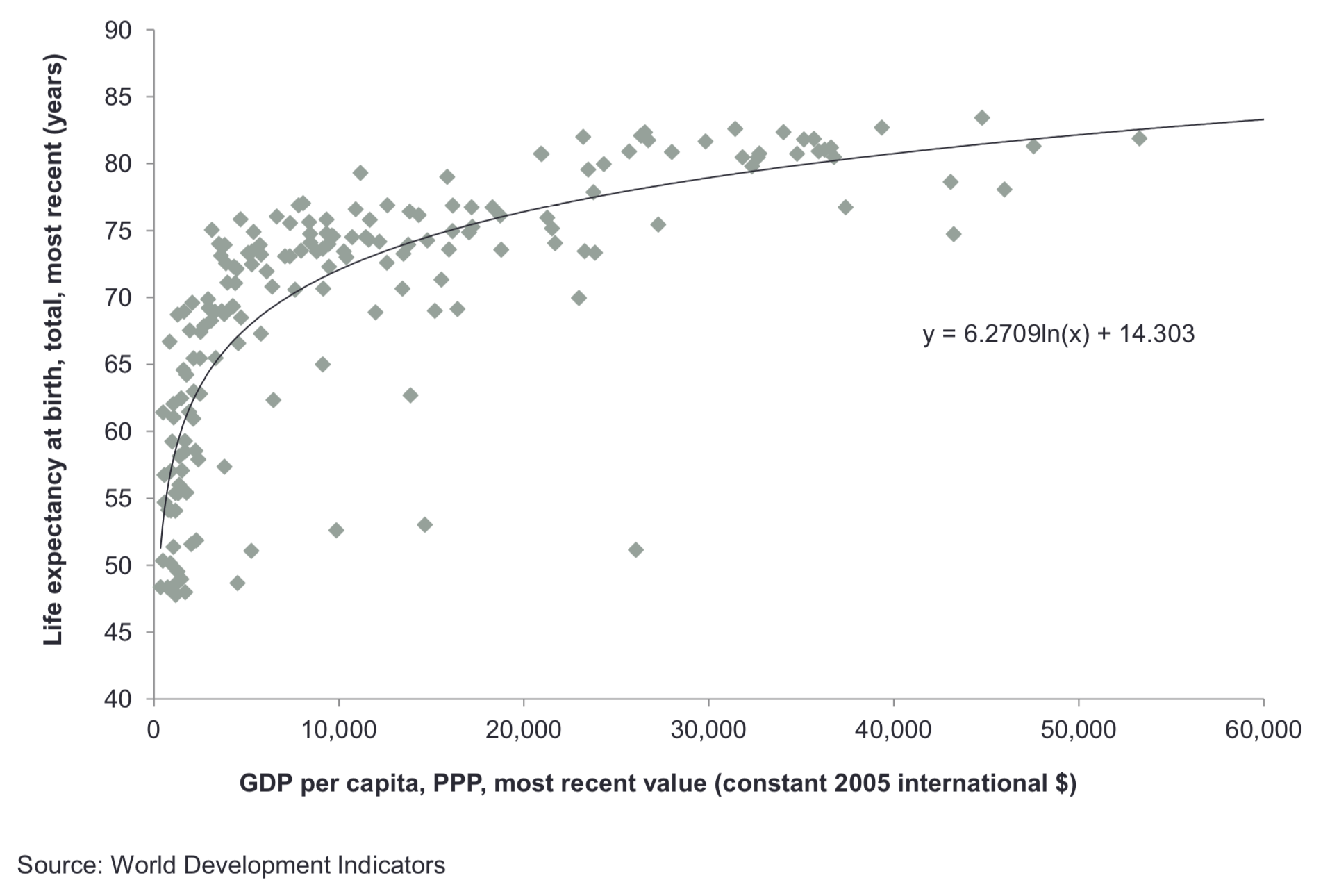
Look at the 3 scatterplots below related to your DEV-101 course (pre-class video for Dani Rodrik’s class #1). Choose one of them and produce it using the WDI data from this exercise. Note: The graph you produce does not need to be identical to the ones below; just similar enough to make the same point.[[13]](#footnote-13) Once you produce your graph, paste the PNG or PDF file below and explain in 2-3 sentences the main conclusion you draw from it.

|  |
| --- |
|  |

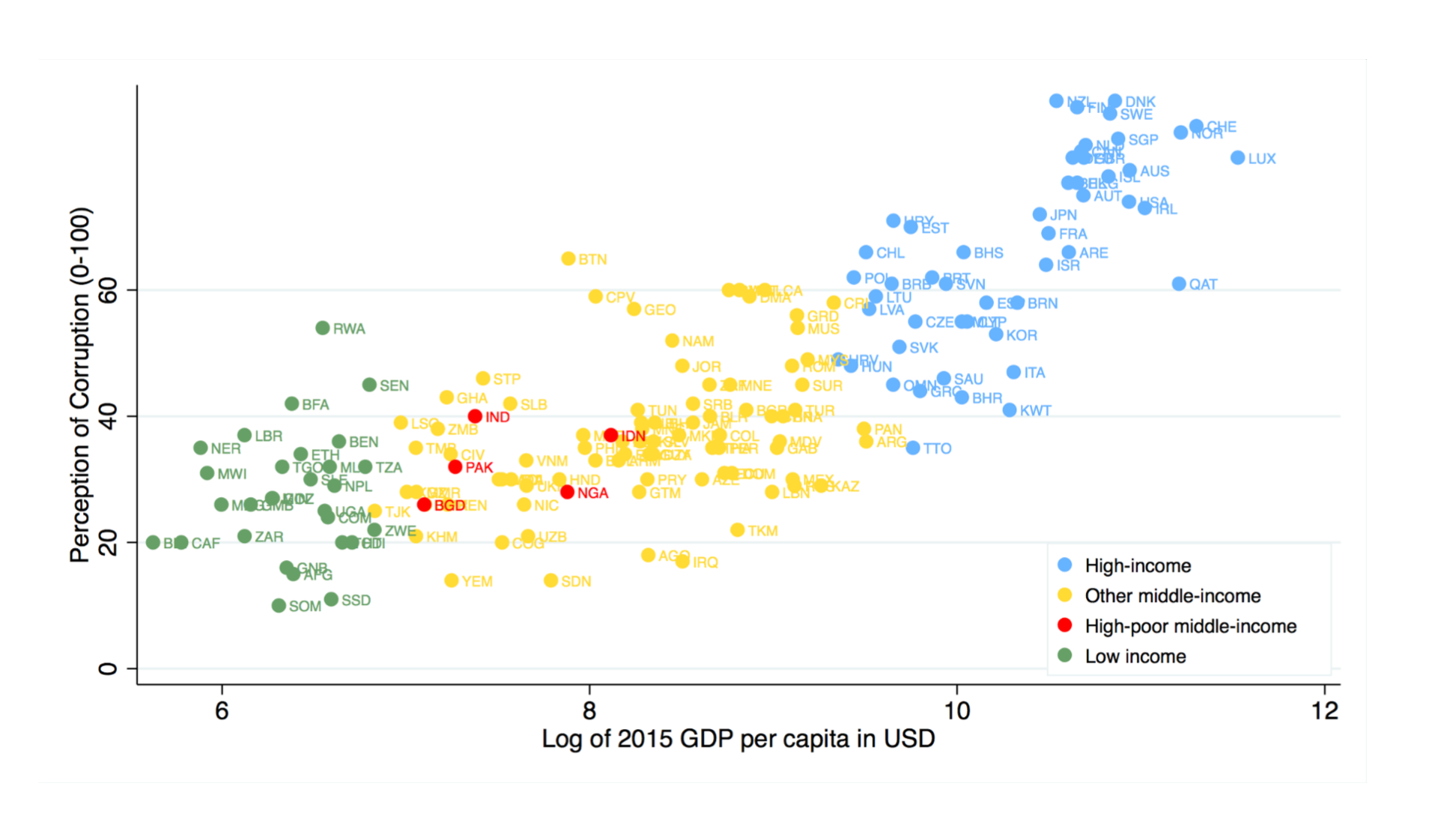
**Graph 1 – Relationship between Income and Infant Mortality**



**Graph 2 – Relationship between income and Life Expectancy**



**Graph 3 – Relationship between Income and Corruption**

****

# Appendix 1 - Venezuelan Elections – Simulating Election Day 1,000 times

This Appendix explains how to simulate 1,000 times “Election Day” in the Venezuelan Elections using **Excel**. When doing simulations of this size, sometimes Excel becomes very slow. If you are experiencing this slowness, you may address it by turning off automatic calculations temporarily (see details below).

* Simulate the election once.
  + The approach to simulating the election will be as follows. For each voting machine, you will randomly select a number from the uniform distribution between 0 and 1.[[14]](#footnote-14) If the random number is less than the probability you calculated in (b) above, then that voting center is one has the same number of “Yes” votes across the two machines. If the random number is greater than or equal to the probability you calculated in (b), then that voting center is one has a different number of “Yes” votes across the two machines.
  + To implement this approach, start by going to Cell C2 and typing the following formula: **=if(rand()<$B2, 1, 0)**. This command will select a random number from a uniform distribution, compare it to the probability you calculated in (b) and code as 1 the voting centers predicted to have same number of “Yes” votes” and as 0 the rest of them.
  + Now copy (or drag down) the formula across column C to fill in the rest of column C (i.e. cells C3 to C8395) for all other voting centers.
* Simulate the election 999 more times
  + Copy the formulas in Column C across to fill cells D3 through ALO395.
  + The result will eventually be a very large matrix (8,394 rows by 1,000 columns) of zeroes and ones. The ones will correspond to centers where the two machines yielded the same number of “yes” votes and zero otherwise.
* Calculate number of voting centers that have two machines with same number of “yes” votes.
  + In row 8398, sum up the 0s and 1s from each of the columns representing simulated election days.

Turning off automatic formula calculations.

* + If you are running Windows, go to File > Options > Formulas. Under Calculation Options > Workbook Calculations, select “Manual” and uncheck the box “Recalculate workbook before saving.” After doing these things, your formulas will not automatically recalculate. Instead, you will have to press F9 to evaluate any formulas. This may be disconcerting at first because you will not see results right away, but it will help address the slowness you might experience with Excel.
  + On a Mac, go to Excel > Preferences > Calculation. Under “Calculate sheets,” select “Manually” and uncheck the box “Always calculate before saving workbook.” After doing these things, your formulas will not automatically recalculate. Instead, you will have to press Command + = to evaluate any formulas. This may be disconcerting at first because you will not see results right away, but it will help address the slowness you might experience with Excel.

# Appendix 2 – Making histograms in Excel

For Question 1, sub-question (d), you are asked to draw histograms using the sets of estimates you generated earlier in the question. Here is some guidance for how to do this in Excel.

1. Before you draw the histogram, you need to specify the bin range of your estimates. This is the range of data values represented by any one bar of the histogram. As this varies depending on your results, you will need to use your own judgment for this. An example of a bin range is as follows:

|  |
| --- |
| -0.5 |
| -0.4 |
| -0.3 |
| -0.2 |
| -0.1 |
| 0 |
| 0.1 |
| 0.2 |
| 0.3 |
| 0.4 |
| 0.5 |

You can create many more bins to create a smoother graph. You will have to create this table in a column somewhere on the worksheet.

1. Make sure that the Analysis ToolPak is enabled.[[15]](#footnote-15) Instructions for doing so are [here](https://support.office.com/en-us/article/Load-the-Analysis-ToolPak-6a63e598-cd6d-42e3-9317-6b40ba1a66b4).
2. For ‘Input range’, select directly from the worksheet the 100 estimates for a particular sample size from your table of results. For ‘Bin range’, select the bin range you have created on your worksheet.
3. If you want your histogram on the same worksheet, click on a single cell on your worksheet for ‘Output range’. If you prefer it in a new worksheet, select ‘New worksheet ply’ and give the new worksheet a name (e.g. ‘Histogram for size 5’). Select ‘Chart output’, or else Excel will not generate the histogram.
4. There are numerous ways to enhance the appearance of the histogram. Right click on various parts of the histogram (bars, axes, legend, titles, plot area, etc.) to change these settings.

# Appendix 3 – Tips to code in R for World Inequality Question

Please note this appendix is meant to provide some hints and suggestions for the R part of the problem set, but do not represent the only (or sometimes even the best) way to answer the question at hand. They are meant to provide a bit of scaffolding if you are learning R and need some help.

1. Use the command summarize to calculate various statistics. For computing missing and non-missing values combine the sum function within summarize (and use is.na and !is.na respectively).
2. Use the theil.wtd command.
3. Many R commands (including *theil.wtd*) allow for the use of weights. But different commands have different syntax for incorporating weights. Check the document from (b) to see how theil.wtd incorporates weights.
4. Use the summarize command for calculating mean and median, and the *arrange* and slice commands for listing the top 3 and bottom 3 countries in terms of Gini.
5. No R involved
6. Use commands *mutate (for generating “corrupt” and “high income” variables) and xtabs (from cheat sheet, to generate the cross tabulations)*.

Optional: *Combine mutate and log to generate the log of GDP. Then use ggplot to generate graphs. For the first graph use geom\_point. For the second graph use geom\_point and geom\_smooth (the latter to generate the log curve). For the third graph, first create a factor variable for high, medium and low income, and then use that variable as the color aesthetic (i.e. color = income\_cat). If you would like to add country labels to your graph, use geom\_text\_repel in the ggrepel package.*

1. We abide by the Harvard Kennedy School Academic [code](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1) 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. [↑](#footnote-ref-1)
2. I am very grateful to Casey Kearney for his help in developing this question. [↑](#footnote-ref-2)
3. In the U.S, $4 billion is equal to $4,000,000,000 (i.e. four thousand million). [↑](#footnote-ref-3)
4. The Economist, “Vaccine Economics: A Bigger Dose”, August 8, 2020. [↑](#footnote-ref-4)
5. <https://www.nih.gov/news-events/news-releases/phase-3-clinical-trial-investigational-vaccine-covid-19-begins> [↑](#footnote-ref-5)
6. https://www.statnews.com/2020/09/08/astrazeneca-covid-19-vaccine-study-put-on-hold-due-to-suspected-adverse-reaction-in-participant-in-the-u-k/ [↑](#footnote-ref-6)
7. To do this in R, use the command *sample\_n*, e.g. type “*samp\_covid <- sample\_n(covid, 100)*” [↑](#footnote-ref-7)
8. I am very grateful to Michael Lopesciolo for his help in designing and writing this question. [↑](#footnote-ref-8)
9. If you are new to R, there are some hints in the Appendix to help you with the questions below. My advice is that you refer to the appendix only after having tried to figure out how to do it by yourself. You will learn more this way. [↑](#footnote-ref-9)
10. If you are interested in understanding differences between Gini and Theil indices, you might find it helpful to consult this [document](http://siteresources.worldbank.org/PGLP/Resources/PMch6.pdf) and this [paper](http://piketty.pse.ens.fr/files/BourguignonMorrisson2002.pdf). [↑](#footnote-ref-10)
11. Note: The data is from 2012/13 in order to expand the dataset, since Gini data is often not calculated annually. Data are not available for every country. [↑](#footnote-ref-11)
12. Note: You are not expected to do any new data analysis; just use your answers from (b) and (c), and the quotes to make your argument. Feel free to also bring some of what you learned in DEV-101. [↑](#footnote-ref-12)
13. For example, to keep things simple for the third graph, define as low income the bottom 15% of countries, high income the top 30%, and middle income ever other country. Ignore the category of High-poor middle income. [↑](#footnote-ref-13)
14. When drawing a random number from a uniform distribution, every value has the same probability of selection. [↑](#footnote-ref-14)
15. FOR MAC USERS: Earlier versions of Excel for Mac do not have the Analysis tool pack. If you are using a version of Excel for Mac that does not have access to the Analysis tool pack and you don’t want to download Excel 2016 and install the Analysis tool pack, follow the instructions below to draw a histogram. Otherwise, ignore the rest of this footnote. You can manually create the histogram using the same bin range suggested here using the function “FREQUENCY” and then use one of the bar graphs available. (The FREQUENCY function is an array function (i.e., it evaluates a range of cells simultaneously). Select all the cells to the right of your bin range, type “=FREQUENCY(“, specify your data range (e.g., B2:B101) and bin range and hit ‘Command’, ‘Shift’ and ‘Enter’ simultaneously). All selected cells will then be populated with the corresponding frequencies). [↑](#footnote-ref-15)