50.034 Introduction to Probability and Statistics 1D Project

CI03 Velusamy Sathiakumar Ragul Balaji
CI03 James Raphael Tiovalen
CI03 Shoham Chakraborty

"Our Zeroth Law of Correlation"

For this study, we will utilize the publicly-available dataset of the latest worldwide COVID-19 per-country statistics provided by Our World in Data. The dataset can be obtained from here. We processed a version of this dataset using our correlation finder script to obtain the results reported in this document.[1]

Before we begin our analysis, let us formally define the random variables that we will consider for this study.

The COVID-19 pandemic is currently an ongoing global pandemic caused by the SARS-CoV-2 coronavirus. For a specific country, we define the following random variables:

- Let X (hosp_patients_per_million) be the number of COVID-19 patients in hospital on a given day per 1,000,000 people.[2]
- Let Y (female_smokers) be the percentage or share of adult women who smoke in percent (ranging between 0.0 and 100.0) out of all adult women (ages 15 and over), at the most recent year available.[3]
- Let Z (human_development_index) be the composite index (ranging between 0.0 and 1.0) measuring average achievement in three basic dimensions of human development — a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506.[4]

[1] We retrieved the dataset on 11 March 2021 (with the last updated date being 10 March 2021). The dataset that we utilized for this study, along with our complete code implementation, can be found here, and a pseudocode snippet can be found in this document's Appendix section.
[2] Source: European CDC for European countries / UK Government / HHS for the United States / COVID-19 Tracker for Canada.
[3] Source: World Bank World Development Indicators, sourced from World Health Organization, Global Health Observatory Data Repository.
[4] Source: United Nations Development Programme (UNDP).

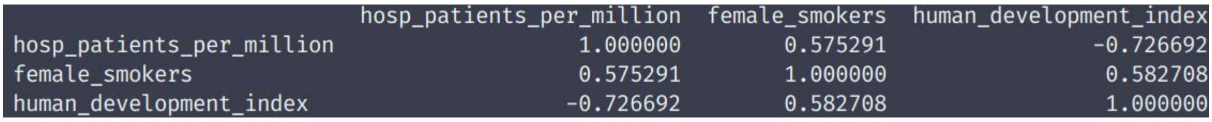
Our domain or sample space would be all of the countries in the world (on Earth) that are listed on the OWID dataset. The time period that we consider would be from 21 January 2020 until 10 March 2021 (on a 2 weeks increment interval window). X would be a discrete non-negative random variable taking on integer values between 0 and 1,000,000 inclusive, whereas Y and Z would be continuous non-negative random variables. In general and in real-life, we do not know neither the exact probability distributions and their parameters (and their parameter spaces), the cumulative distributions nor the conditional distributions of all of these three random variables and whether they are independent or dependent. In general, we also do not know neither the multivariate, the joint nor the marginal probability distributions of any pair/triple combinations of these 3 random variables. In this study, we will not assign or assume any prior or posterior distributions to these random variables (since we do not need to). Since in the real world, there is a finite number of countries and a finite number of humans, X, Y and Z all have their respective finite means and finite variances (whose actual values are unknown to us), and thus the correlation statistics of any pair of random variables exist and are well-defined.

Using the `pandas` library in Python,[5] we obtain the correlation matrix between the actual observed values of these 3 random variables that we have collected from the dataset. These are the corresponding observed Pearson's correlation coefficient values (to 6 decimal points):

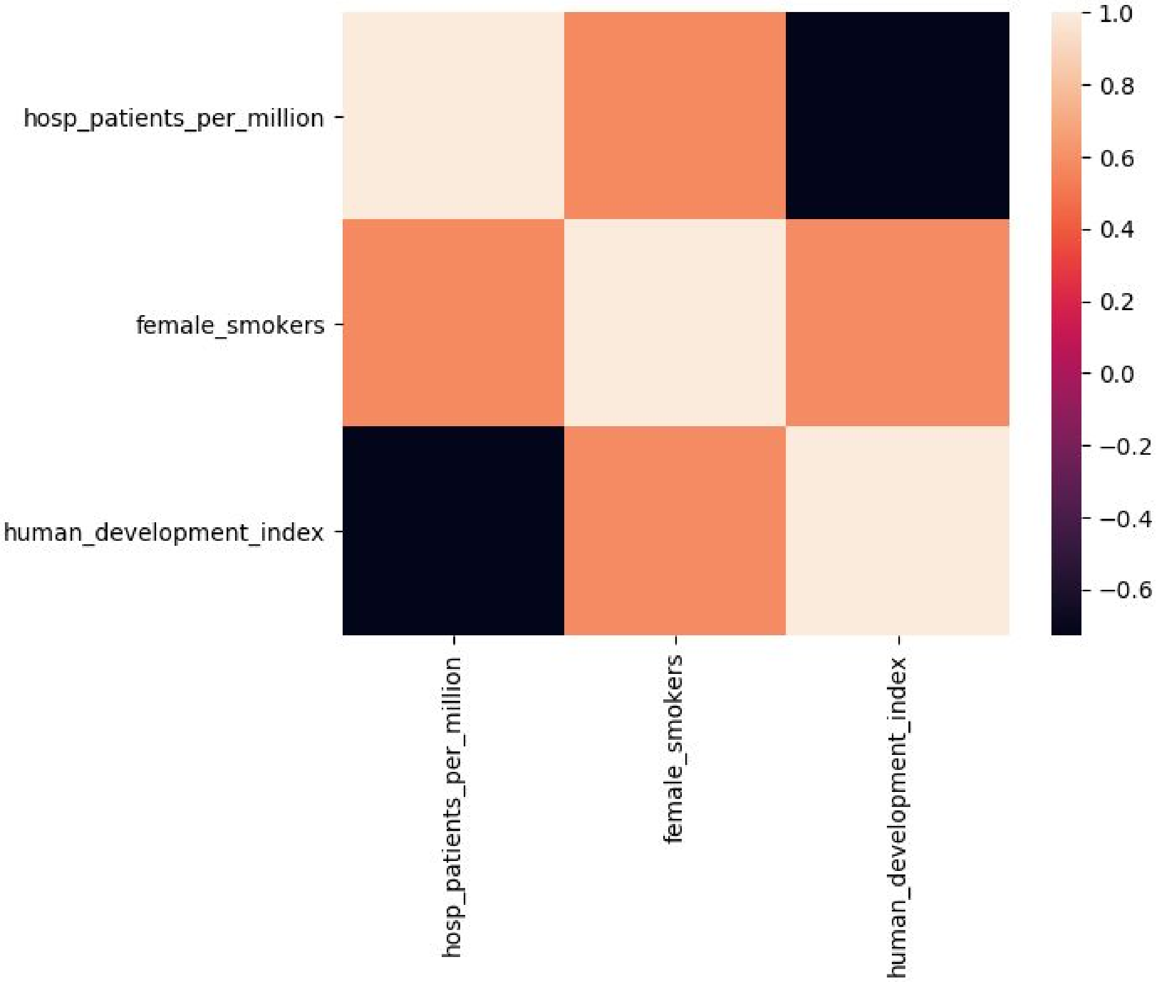
- ρ(X, Y) ≈ 0.575291 (> 0)
- ρ(X, Z) ≈ -0.726692 (< 0)
- ρ(Y, Z) ≈ 0.582708 (> 0)

[5] Technicalities: the `df.corr()` function in `pandas` automatically ignores any non-numeric data type columns and excludes any missing values in the dataframe being considered. Since the OWID dataset does have some missing values, this might lead to slightly inaccurate/unreliable conclusions, even though we tried our best and put in the effort to clean the data and minimize the errors, if there are any at all.

We can view and visualize the correlation matrix in different forms as well:



```
                          hosp_patients_per_million  female_smokers  human_development_index
hosp_patients_per_million                  1.000000        0.575291                -0.726692
female_smokers                             0.575291        1.000000                 0.582708
human_development_index                   -0.726692        0.582708                 1.000000
```



While at first glance, this seems bizarre, there is a possible explanation to this phenomenon:

1. X and Y are positively correlated. This seems quite obvious since a larger percentage of female smokers would imply that such a country would have generally a larger percentage of the population to be smoking and thus have more people admitted into the hospital due to higher susceptibility to COVID-19 as well as more serious

symptoms of COVID-19. It is a well-known fact among medical professionals that tobacco smoking is a risk factor that generally increases the severity of and susceptibility to respiratory diseases such as COVID-19.[6]

2. X and Z are negatively correlated. This also seems obvious since more-developed countries with a higher human development index (HDI) would be more capable, equipped, prepared and orderly in containing the spread of COVID-19. This would decrease the number of people who get infected by COVID-19 and would then decrease the number of people admitted into hospitals due to COVID-19. This data also suggests that the level of capability of countries in combating against COVID-19 seems to offset the availability of hospitals in less-developed countries in pure numbers.

3. However, Y and Z are *positively* correlated! This seems confusing, weird and contradictory at first glance. However, we can explain this by considering the factor of female empowerment. More-developed countries with a higher HDI would tend to have better gender equality and more female empowerment movements since their citizens would develop better critical-thinking, are equipped with better resources and are generally more cognitively capable to be concerned with gender equality and gender equity issues. Their citizens would be less concerned with pure survival and they would be able to take on these more abstract concepts. A higher HDI would also mean a higher availability of readily purchasable cigarette packs which would further explain the positive correlation. As such, a larger share or portion of their citizens (and female adults) would smoke compared to less-developed countries.[7] It also seems that based on the dataset that we have obtained, these effects seem to

[6] Source: https://www.who.int/news/item/11-05-2020-who-statement-tobacco-use-and-covid-19

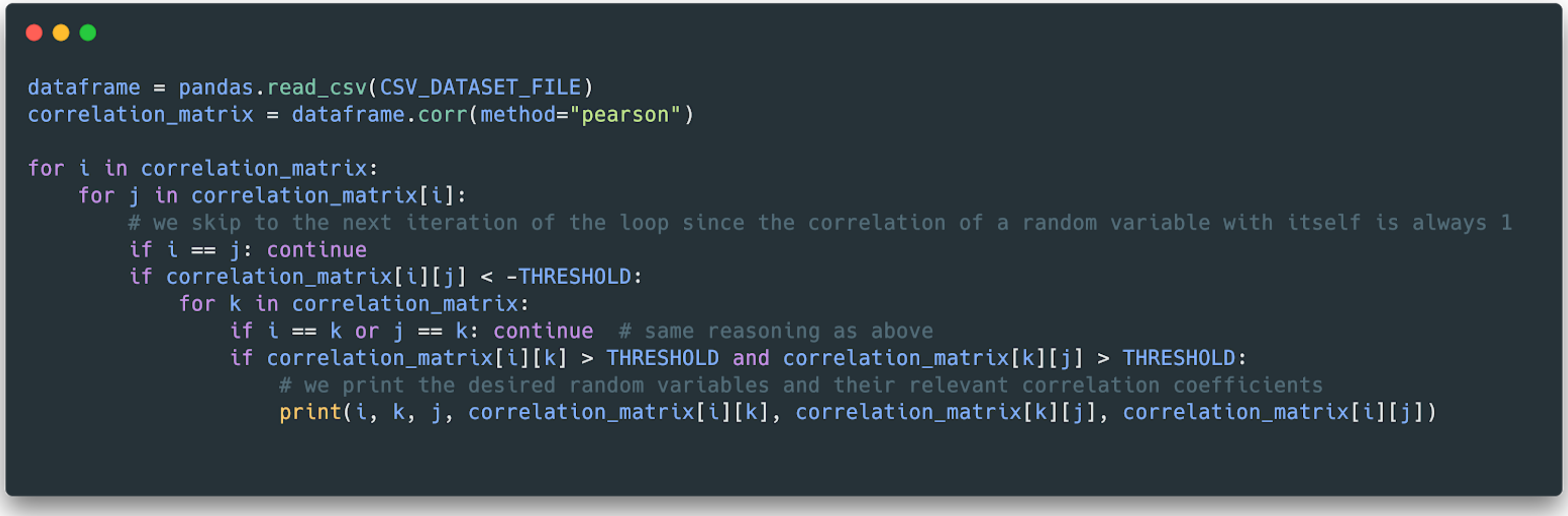
[7] Source: https://www.who.int/bulletin/volumes/89/3/10-079905/en/

substantially offset the level of awareness or education regarding the potential dangers and negative/adverse consequences and impacts of smoking. Thus, the larger the share of women who smoke in a country, the greater the tendency that such a country would have a higher HDI.

Hence, through this exploratory study, we have found a real-life example that showcases the non-transitivity property of Pearson’s correlation coefficient.

Appendix

Here is a pseudocode implementation of our correlation finder/mining algorithm:



```
dataframe = pandas.read_csv(CSV_DATASET_FILE)
correlation_matrix = dataframe.corr(method="pearson")

for i in correlation_matrix:
    for j in correlation_matrix[i]:
        # we skip to the next iteration of the loop since the correlation of a random variable with itself is always 1
        if i == j: continue
        if correlation_matrix[i][j] < -THRESHOLD:
            for k in correlation_matrix:
                if i == k or j == k: continue  # same reasoning as above
                if correlation_matrix[i][k] > THRESHOLD and correlation_matrix[k][j] > THRESHOLD:
                    # we print the desired random variables and their relevant correlation coefficients
                    print(i, k, j, correlation_matrix[i][k], correlation_matrix[k][j], correlation_matrix[i][j])
```