

## Problem Set 2

Due 09/23/2021 by 12pm. Submit through canvas.

This problem set is due **September 23rd at 12pm (noon)**. Assignments should be submitted through Canvas and include i) a PDF document, and ii) an .R or .Rmd code file. The PDF document should include **all** your work (code, output, written typed answers, hand-written math answers). Please see the Problem Set [Submission Guidelines](#) page on Canvas for useful tips.

# 1 Theoretical Questions

Remember to show all your work. Also, if you use any probability rules, specify which ones and where they are used.

## 1.1 Vaccines

Age	Population		Hospitalized		Efficacy
	Not Vax	Fully Vax	Not Vax	Fully Vax	
All ages	1,302,912	5,634,634	214	301	
< 50	1,116,834	3,501,118	43	11	
> 50	186,078	2,133,516	171	290	

The table above provides data on COVID vaccinations and hospitalizations of severe cases in Israel (as of August 15, 2021). Israel has been one of the leading countries in terms of vaccination rates, and many people view it as a case study of what widespread vaccination can, or cannot, achieve. Except for the last column, each cell represents the number of individuals who fall into a specific category. Use this information to answer the following questions.

1. Compute  $P(\text{Hospitalized} \mid \text{Fully Vax})$  and  $P(\text{Hospitalized} \mid \text{Not Vax})$  for the overall population, and display your result as a percentage with at least four decimal places.
2. Calculate conditional probabilities from part 1 for the two age groups displayed in the table. Again, show your result as a percentage with at least four decimal places.
3. Let's define the vaccine "efficacy" as

$$\text{Efficacy} = 1 - \frac{P(\text{Hospitalized} \mid \text{Fully Vax})}{P(\text{Hospitalized} \mid \text{Not Vax})}.$$

This measure captures the percentage reduction in hospitalizations in the vaccinated group relative to the unvaccinated group. Compute the efficacy among the overall population.

4. Also compute the efficacy for the two age groups. How does the efficacy among the overall population compare to the breakdown by age? What could explain the discrepancy between age groups and the overall population?
5. Suppose in a counterfactual world the conditional probabilities computed in part 1 remain the same, but now we swap the vaccination rate, i.e., only 1,302,912 individuals are fully vaccinated and 5,634,634 are not vaccinated. Compute the expected number of hospitalizations for the vaccinated and unvaccinated population implied by the conditional probabilities. How many more hospitalizations would there have been?
6. Suppose someone observes the data and argues: "Most of the hospitalized are fully vaccinated individuals. Therefore, vaccines are not very effective." Do you find this argument convincing? In 1-2 sentences, explain why or why not. What other factors besides vaccine efficacy could explain that the majority of the hospitalized are vaccinated?

## 1.2 True or False

*Please include a short explanation for your answer, even if it is true*

1.  $P(E|F)P(F) = P(E)$  for any event  $E$  and  $F$ .
2. Suppose  $P(A) = 1/4$  and  $P(\text{not } B) = 3/4$ . If  $A$  and  $B$  are independent, then they cannot be mutually exclusive.
3. If  $A$  and  $B$  are mutually exclusive, then  $P(A \cap B) = P(A)P(B)$ .

## 1.3 Conditional probability

A lie detector test is 95 percent effective in detecting lies when statements are lies. However, the test also yields a “false positive” result for 1 percent of the tested statements (That is, if a truthful statement is tested, then, with probability 0.01, the test result will imply that the statement is a lie.) So, if 0.5 percent of the population of statements are lies, what is the probability that a statement is a lie given that the test result is positive?

## 2 The Geography of Intergenerational Mobility

The growth in inequality in the developed world during the last few decades has increased public interest in the topic of inequality. One robust finding of this literature is that many factors that influence a person's ultimate socioeconomic status are determined long before their entry in the labor market. Since many of these factors depend on family and background, there is some degree of intergenerational persistence of inequality: The rich and educated can guarantee better inputs to their children, who then have a better chance of remaining in the upper echelons of society upon adulthood. The term intergenerational mobility is used to refer to the strength of this perpetuation mechanism. If perpetuation of inequality is intense, we say intergeneration mobility is low, and vice-versa.

Economist Raj Chetty and his co-authors are at the forefront of the study of such intergenerational inequality. In a series of papers, they use high-quality data from the IRS on the incomes of millions of American workers to examine the state of intergenerational mobility in the US. The remainder of this Problem Set focuses on their 2014 paper<sup>1</sup> which studies how intergenerational mobility measures vary across the US. We will use a variety of data sets made available by Chetty's team on the [website](#) of their "Equality of Opportunity" project. If you would like to, take some time to explore their interactive data visualization at [Opportunity Atlas](#).

**The following questions will require the use of R. You must provide the code for your solutions and the corresponding output.**

### 2.1 Income Percentiles

The file `transition_matrix_national.csv` represents a matrix where each cell indicates the probability that a child has a family income percentile in a range corresponding to its row, conditional on having parent family income percentile in a range corresponding to its column.<sup>2</sup> A person with income in the  $x$ th percentile has an income that is greater than  $x\%$  of the population and less than  $(100 - x)\%$  of the population, where  $x$  is a number between 0 and 100. To create the matrix, Chetty et al. divide the income observations into 100 intervals of equal size. The first interval contains observations with income percentile less than or equal to 1%; the second interval contains observations with income percentile greater than 1% and less than or equal to 2%; the third interval contains observations with income percentile greater than 2% and less than or equal to 3%, and so on. Each column in the matrix corresponds to the subpopulation whose parent's income percentile fall into one of these intervals.

Let  $P$  be the parent income percentile and  $K$  the child income percentile. The variable `k_centile` reports the child percentile interval, and the variable `p_cN` represents the  $N$ th parent percentile interval,  $(N - 1)\% < P \leq N\%$ . Thus, to find the probability

$$Pr(74 < K \leq 75 \mid 24 < P \leq 25),$$

---

<sup>1</sup>A link to this paper is provided on the page for Assignment 2 on Canvas.

<sup>2</sup>Thus, each column represents a conditional distribution and should add up to one. However, in practice, columns do not necessarily add up to 1 due to rounding.

we can use the following command:

```
subset(dataset, k_centile == 75, p_c25),
```

where `dataset` is the name of the data frame that students chose. This command returns the value of the variable `p_c25` when `k_centile == 75`. Based on this information, calculate the following objects.

1.  $Pr(49 < K \leq 75 | 49 < P \leq 50)$
2.  $Pr(49 < K \leq 75)$ . In other words, what is the marginal probability of a child's income being between the 49th and 75th percentiles?<sup>3</sup> You do not need to use R to answer this question.
3. The sample used in this paper tracks cohorts with a total of 9,867,736 children. If we know that there are 98,677 children with parents in the 50th percentile interval, estimate  $Pr(49 < P \leq 50)$ .
4. Use Bayes' Rule and your previous answers to estimate  $Pr(49 < P \leq 50 | 49 < K \leq 75)$ .

## 2.2 Measuring Intergenerational Mobility

To understand how parental income is related to children's income, Chetty et al. work with a couple of different measures. To define these measures, we will start by indexing observations for each child by  $i$ . Let  $K_i$  denote the national income percentile of child  $i$ , and let  $P_i$  denote her parent's percentile in the income distribution of parents.

- The Absolute Mobility measure (AM) is given by  $AM = E[K_i | 24 < P_i \leq 25]$ , i.e., it is the expected child income percentile given that her parent's income percentile is between 24% and 25%.
- Let  $Y_x = E[K_i | x - 1 < P_i \leq x]$  denote the expected income percentile of a child whose parents are at the  $x$ th percentile interval in the income distribution. The Relative Mobility measure (RM) is defined as  $RM = Y_{100} - Y_1$ , i.e., it is the difference between expected ranks of children whose parents are at the top and bottom of the distribution.

The file `stats_par_percentile_national.csv` contains a data set with values for  $Y_x$  at the national level. Each observation (row) corresponds to a parent income percentile interval, and it is indexed by the variable `par_bin`. The variable `kid_fam_rank` represents  $Y_x$  for the  $x$ th row. Load this data set and use it to answer the following questions.

1. Create a scatter plot where `par_bin` is on the horizontal axis and `kid_fam_rank` on the vertical axis.<sup>4</sup> Interpret the graph.
2. Describe how RM can be depicted in this figure and calculate its value. You do not need to plot a new graph.

---

<sup>3</sup>Note that by definition  $Pr(n - 1 < K \leq n) = 0.01$  for  $n = 1, 2, \dots, 100$ .

<sup>4</sup>To make a scatterplot for two variables `x` and `y`, one can use the following command: `plot(x, y, xlab = "x-axis title", ylab = "y-axis title", main = "title")`.

3. Describe how one can find the value of AM at the national level in this figure and calculate its value. You do not need to plot a new graph.

## 2.3 Variation across the US

The file `mobility_county.csv` is a data set with mobility measures for each county in the US. Load this data set and use it to answer the following questions.<sup>5</sup>

1. How many observations are in the data set? How many variables are in the data? What are the first 5 variables in the data set?
2. The variable `region` shows to which geographical region a county belongs to. Given that we randomly select a county from the North East Region, what is the probability that it lies in Connecticut?
3. The variable `k_count` is the number of children in the 1980-1982 cohort who grew up in a given county. What is the average cohort size conditional on being from the North East? How does this compare to the average cohort size conditional on being from West, South, and Midwest?
4. The variable `rm` reports the *Relative Mobility* measure. What is New Haven's RM? Compare this to the average RM of the North East, West, South, and Midwest. According to this metric, does New Haven have more or less intergenerational mobility than these regions?
5. The variable `am` reports the *Absolute Mobility* measure. What is New Haven's AM? According to the AM measure, does New Haven have more or less mobility than the US at the national level? (That is the value we computed in part 3 of the previous section.)
6. In the data, the variable `par_median` is the median income for parents of children who grew up in a given county, while `k_median` is the median income eventually achieved by the children. Make a scatter plot of `k_median` against `par_median`. Make sure the plot has a title and axis labels. Using the plot, comment on the relationship between these two variables (one sentence maximum).
7. Based on all your answers and any additional analysis with the data, comment on the relation between mobility and geographic region (three sentences maximum).

---

<sup>5</sup>The commands `dim()`, `names()`, `head()`, `tail()`, `order()`, `which()`, and `subset` may be useful in the following problems.