# Extra Credit

Alice Ding

2023-03-02

## Overview

```
israeli_data <- read.csv("https://github.com/addsding/data607/raw/main/extra/ec3/israeli_vaccination_da
head(israeli_data)
```

```
##    Age Not.Vax.Population Not.Vax.Population.. Fully.Vax.Population
## 1 < 50            1116834                0.233              3501118
## 2 > 50             186078                0.079              2133516
##   Fully.Vax.Population.. Severe.Cases.Not.Vax.per.100k
## 1                  0.730                            43
## 2                  0.904                           171
##   Severe.Cases.Fully.Vax.per.100k
## 1                              11
## 2                             290
```

This data describes August 2021 data for Israeli hospitalization ("Severe Cases") rates for people under 50 (assume "50 and under") and over 50, for both un-vaccinated and fully vaccinated populations. The columns are:

- Age (50 and under, over 50)
- Not Vax Population (count)
- Not Vax Population (% of total population)
- Fully Vax Population (count)
- Fully Vax Population (% of total population)
- Severe Cases Not Vax per 100k
- Severe Cases Fully Vax per 100k

As the assignment is about Israel, here are some high-level facts I've discovered to start:

- 2021 Population of Israel = 9.451 million (source).
- Israel's population is younger on average than other developed countries (12% over the age of 65) (source).
- Fully vaccinated means having 2 doses of either the Pfizer or Moderna vaccines.
- People who were eligible were Israeli citizens of any origin as well as Palestinian residents of East Jerusalem; an estimated 5m Palestinians in west Bank and Gaza were not eligible though due to their healthcare being under the jurisdiction of the Palestinian Authority per the Oslo Records (source).
- There's also an age minimum for folks who can get the vaccine: 12 and over (source).

With this as a base, let's begin by tidying the data for analysis!

## Tidying the Data

When looking at this data, it's honestly a little hard to see how we could get this one .csv into one table. I honestly think it'd be cleaner to have these be separate tables:

- One with population counts and their respective percentages
- One with the per 100k numbers

We'll begin with the first one by pivoting each of the counts by age bucket.

```
# pivot the counts
israeli_data_pop_count <-
  pivot_longer(israeli_data[1:5]
               , cols=c(2, 4)
               , names_to="population_type"
               , values_to="count"
               )

# pivot the percentages
israeli_data_pop_count <-
  pivot_longer(israeli_data_pop_count
               , cols=2:3
               , names_to="population_type_"
               , values_to="percentage_of_population"
               )

# clean the names to remove the periods
israeli_data_pop_count$population_type <-
  gsub("\\.", " ", israeli_data_pop_count$population_type)
# this one needs white space removal as there were periods at the end
israeli_data_pop_count$population_type_ <-
  trimws(
    gsub("\\.", " ", israeli_data_pop_count$population_type_)
    )

# filter down the rows
israeli_data_pop_count <-
  israeli_data_pop_count |> filter(population_type == population_type_)

# remove the duplicate row
israeli_data_pop_count <-
  subset(israeli_data_pop_count, select=-c(population_type_))

head(israeli_data_pop_count)
```

```
## # A tibble: 4 x 4
##   Age   population_type        count percentage_of_population
##   <chr> <chr>                  <int>                    <dbl>
## 1 < 50  Not Vax Population   1116834                    0.233
## 2 < 50  Fully Vax Population 3501118                    0.73
## 3 > 50  Not Vax Population    186078                    0.079
## 4 > 50  Fully Vax Population 2133516                    0.904
```

Next is the data frame for counts per 100k.

```
# pivot the counts
israeli_data_100k <-
  pivot_longer(israeli_data[c(1, 6, 7)]
               , cols=2:3
               , names_to="population_type"
               , values_to="count"
               )

# clean the names to remove the periods
israeli_data_100k$population_type <- gsub("\\.", " ", israeli_data_100k$population_type)

head(israeli_data_100k)
```

```
## # A tibble: 4 x 3
##   Age   population_type                 count
##   <chr> <chr>                           <int>
## 1 < 50  Severe Cases Not Vax per 100k      43
## 2 < 50  Severe Cases Fully Vax per 100k    11
## 3 > 50  Severe Cases Not Vax per 100k     171
## 4 > 50  Severe Cases Fully Vax per 100k   290
```

Both data frames look clean and tidy!

## Transforming & Analyzing

### Question 1: Do you have enough information to calculate the total population?

We have numbers and what percentage they represent, but I don't believe that'd be *total* population for Israel. Let's double check that though.

For the `< 50` group, let's see what the total number would be based on both population types.

```
print(paste("Population based on not vax population:"
            , toString(
              israeli_data_pop_count$count[1] /
                israeli_data_pop_count$percentage_of_population[1])
            )
      )
```

```
## [1] "Population based on not vax population: 4793278.96995708"
```

```
print(paste("Population based on vax population:"
            , toString(
              israeli_data_pop_count$count[2] /
                israeli_data_pop_count$percentage_of_population[2])
            )
      )
```

```
## [1] "Population based on vax population: 4796052.05479452"
```

These numbers are very close – we can assume that the difference is due to rounding, this is relatively accurate otherwise. From this, I'll say that the total population of the `< 50` group is 4.79m

Now, let's do the same for the `> 50` group.

```
print(paste("Population based on not vax population:"
            , toString(
              israeli_data_pop_count$count[3] /
                israeli_data_pop_count$percentage_of_population[3])
            )
      )
```

```
## [1] "Population based on not vax population: 2355417.72151899"
```

```
print(paste("Population based on vax population:"
            , toString(
              israeli_data_pop_count$count[4] /
                israeli_data_pop_count$percentage_of_population[4])
            )
      )
```

```
## [1] "Population based on vax population: 2360084.07079646"
```

Again, due to rounding, these are slightly off but otherwise close. I 'll say that for the `> 50` group, the population is 2.36m.

Together, this is 7.15m. This, however, does not come close to the 9.451m that I researched beforehand which is supposed to represent the population of Israel. This discrepancy likely comes down to those who are eligible or not and so to rephrase what we are able to find: we are able to calculate the total population of *people who are eligible for the vaccine*.

I will also say that if you add up all of the population counts, it does not equal 7.15m. The percentages also don't add up to 100% – I believe a group is missing here. One I can think of would be folks who are partially vaccinated, so for example they only have 1 of the 2 doses.

### Question 2: Calculate the Efficacy vs. Severe Disease

Given the formula: `1 - (% fully vaxed severe cases per 100k / % not vaxed severe cases per 100k)`, we can calculate it like so:

```
below_efficacy <- 1 - ((israeli_data_100k$count[2] / 100000) / (israeli_data_100k$count[1] / 100000))

above_efficacy <- 1 - ((israeli_data_100k$count[4] / 100000) / (israeli_data_100k$count[3] / 100000))

print(paste("Efficacy for < 50 population:", toString(below_efficacy)))
```

```
## [1] "Efficacy for < 50 population: 0.744186046511628"
```

```
print(paste("Efficacy for > 50 population:", toString(above_efficacy)))
```

```
## [1] "Efficacy for > 50 population: -0.695906432748538"
```

Based on these results, we can see that for folks fifty and under, being fully vaccinated proves to be 75% effective against severe cases. For folks above the age of fifty, the vaccine actually failed to be effective for these cases as its efficacy was negative. Folks over the age of fifty though are likely to have more health problems and be more susceptible to disease so it makes sense that the vaccine wasn't as effective.

**Question 3: From your calculation of efficacy vs. disease, are you able to compare the rate of severe cases in unvaccinated individuals to that in vaccinated individuals?**

I would say that past a certain age, it becomes much more individualized as to whether a vaccine would help someone or not due to illnesses and other health concerns that come with older age. I think comparisons would work best for people below fifty as it's easier to generalize that population than the older one.

## Conclusion

The data was pretty hard for me to picture as one data frame so I ended up splitting it into two, however I'm guessing there's a way to combine them – the data just wouldn't be as tidy in my opinion. The analysis was a little confusing as I wasn't entirely sure a negative efficacy percentage made sense, however I see now that it does since the fully vaccinated severe case count in the `> 50` group is higher than the non vaccinated one which means that the vaccine wasn't effective for this age group.

Next steps would perhaps be to take a look at other countries with similar size / conditions to see if this trend continues here or if Israel is an outlier with its performance.