DATA 607 Tidying and Transforming Vaccination Data

David Simbandumwe

Data Preparation

```
# create the empty data frame
df <- data.frame( age = character(),
                  pop_not_vax = double(),
                  po_not_vax_per = double(),
                  pop_vax = double(),
                  pop_vax_per = double(),
                  case_not_vax = double(),
                  case_vax = double()
)

# add data from xls
df <- df  %>%
  add_row(
    age = "under 50",
    pop_not_vax = 1116834,
    po_not_vax_per = 0.233,
    pop_vax = 3501118,
    pop_vax_per = 0.730,
    case_not_vax = 43,
    case_vax = 11
  ) %>%
  add_row(
    age = "over 50",
    pop_not_vax = 186078,
    po_not_vax_per = 0.079,
    pop_vax = 2133516,
    pop_vax_per = 0.904,
    case_not_vax = 171,
    case_vax = 290
  )

# write the start file
write.csv(df, "/Users/dsimbandumwe/dev/cuny/data_607/DATA607/Homework/israeli_vaccination_data_analysis


# read csv from
vax_df <- read_csv( file = "/Users/dsimbandumwe/dev/cuny/data_607/DATA607/Homework/israeli_vaccination_
```

```
## Rows: 2 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## chr (1): age
## dbl (6): pop_not_vax, po_not_vax_per, pop_vax, pop_vax_per, case_not_vax, ca...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tiddy Data

Transform the wide data format into tiddy data using the gather function. Note the percentages for completeness but you could ignore those values since they can be calculated when needed. Created a boolean variable to track vaccination status. Updatd the cas counts so that each row reflects age, vax status, population and the case count.

```
#tiddy data


vax_df <- vax_df %>%
  gather(
    "pop_not_vax","pop_vax",
    key = "type", value = "population"
  ) %>%
  mutate (
    vax_status = ifelse("pop_vax" == type, TRUE, FALSE)
  ) %>%
  mutate (
    cases = ifelse(vax_status, case_vax, case_not_vax)
  ) %>%
    select(age, vax_status, population, cases)
```

(1) Israel's total population

```
vax_df %>%
  summarise(
    tot_pop = sum(population)
  )
```

```
## # A tibble: 1 x 1
##   tot_pop
##     <dbl>
## 1 6937546
```

According to the data Israel's total population is 6,937,546. I believe what they are referring to is the total population of individual eligible for a covid vaccine.

https://www.cbs.gov.il/en/pages/default.aspx Israel's Central Bureau of Statistics (July 2021) 9.378 million residents in isreal

https://www.cbs.gov.il/he/publications/doclib/2021/2.shnatonpopulation/st02_03.pdf Averages for 2020 9,215.1 - total population 2.123 - 11 and under 7.078 - 12 and over

(2) Who is eligible to receive vaccinations, and

According to the latest article that i could find from the times of isreal. Individual over the age of 12 are eligable for vaccination and the 3rd booster shot of the Pfizer vaccine.

https://www.timesofisrael.com/israel-offers-covid-booster-shot-to-all-eligible-for-vaccine/ Israel widens 3rd COVID booster shot to those aged 12 and over

(3) What does it mean to be fully vaccinated? Please note any apparent discrepancies that you observe in your analysis. People fully vaccinated shows how many people have received the full amount of doses for the COVID-19 vaccine. Since some vaccines require more than 1 dose, the number of fully vaccinated people is likely lower. Although it does not state this in the document I assume that they are using the standard definition of fully vaccinated. The number presented match the Google news COVID tracker for Israel 5.611 million (google) vs 5.635 million (assignment)

The analysis has 2 categories not vaccinated and fully vaccinated. It does not address individual that are partially vaccinated 1 shot of 2 or who have taken the booster shoot (3rd dose) will complicate that analysis. Also individual under the age of 12 are not included in the analysis

(1) Do you have enough information to calculate the total population. What does this total population represent?

No you do not have enough information to calculate the total population of Israel. The data is missing residents of Israel under the age of 12. The analysis seems to capture only the individuals eligible to be vaccinated.

(2) Calculate the Efficacy vs. Disease; Explain your results.

```
# calculate under 50 efficacy
u50df <- vax_df %>%
  filter (age == "under 50") %>%
  mutate (
    pop_percent = population / sum(population),
    per_100k = cases / population * 100000
  )

a <- subset(u50df,subset = vax_status,select = c("per_100k"))
b <- subset(u50df,subset = !vax_status,select = c("per_100k"))

u50eff <- 1 - (a/b)
u50eff
```

```
##   per_100k
## 1 0.918397
```

```
# calculate over 50 efficacy
o50df <- vax_df %>%
  filter (age == "over 50") %>%
  mutate (
    pop_percent = population / sum(population),
    per_100k = cases / population * 100000
  )

a <- subset(o50df,subset = vax_status,select = c("per_100k"))
b <- subset(o50df,subset = !vax_status,select = c("per_100k"))

o50eff <- 1 - (a/b)
o50eff
```

```
##    per_100k
## 1 0.8520888
```

```
# calculate overall efficacy
pop_df <- vax_df %>%
  group_by(vax_status) %>%
  mutate (
    per_100k = sum(cases) / sum(population) * 100000
  ) %>%
  select(vax_status, per_100k) %>%
  distinct()


a <- subset(pop_df,subset = vax_status,select = c("per_100k"))
b <- subset(pop_df,subset = !vax_status,select = c("per_100k"))


eff <- 1 - (a/b)
eff
```

```
##    per_100k
## 1 0.6747614
```

The overall efficacy rate is 67.5% which is substantially lower than the efficacy rate for under 50 (91.8%) and over 50 (85.2%). The cause of this discrepancy is Simpson's Paradox where the trend that appears in both groups in reversed when the groups are combined.

(3) From your calculation of efficiency vs. disease, are you able to compare the rate of severe cases in unvaccinated individuals to that in vaccinated individuals?

We cannot compare the rates for the entire population however we can compare the rates for the population that is eligible for vaccinations. * 0.2% cases for non vaccinated individual * 0.3% for vaccinated individual