Memo:

In the Unit 1 Paper version 2, I detect the technical illustration on why I choose the AC interval and try to use simple and understandable words to replace some technical words. For the method part, avoid using abundant math notes because this paper is not for people who study in stats. Avoiding assuming the reader already know some concept and result from stats 485. I also try to start from our actual study, using more solid words to illustrate the statistical ideas. For example, replacing p into cases rate caused by covid. At the same time, I try to avoid directly applying variable names from R in my tables. I generate more readable tables for readers rather than directly copy tables from R. At the discussion part, I spend more words on explain why age-stratified is more realiable.

I choose the question style introduction, just because I think the question can arise the curiosity of readers and the readers can have a clear mind on what this paper is trying to say. In the discussion part, I also should address the question which I state in the introduction part, which can make my essay more concise and keep the whole structure in a good shape.

Xiangsen Yang

Professor Hansen

Stats 485

31 January 2022

Vaccination and Severe COVID in Israel

INTRODUCTION

Two medical companies, Pfizer and Modern, received authorization for emergency use in the U.S. for providing vaccines for fighting with Covid-19 pandemic. Officials announce that those two companies' vaccines were found to be more than 90% effective in randomized clinical trials, which is a piece of good and exciting news around the world, but these trials were based on the original specie of Coronavirus, not on variant lately came out (Delta variant). With the more distributions of mRNA vaccines around the world, many publics may have questions: does the mRNA vaccine still work effectively for new variants? Some extreme anti-vaxers may even strongly bluff that vaccines are not efficient at all. Although we do not have first-hand experimental data to answer these doubts, we are able to use the online source (public health data from a single country) to try to approximate a randomized trial and get the answer.

This paper investigates public health data from Israel's national health service. This Israel's health data reflect Israel's population and Covid hospitalization counts as of August 15, 2021, by age and vaccination status. We will try to use Agresti-Coull Confidence Interval to create a 95% confidence interval for underlying rates of Covid severe disease in the vaccinated and unvaccinated group. By calculating point estimates of case rates, we can drive point

estimates of vaccine estimates and their corresponding confidence intervals for vaccine efficacy. By doing these steps above, we may safely arrive at an answer to the question: does vaccine has a positive effect on protecting people from severe diseases caused by covid-19 variants?

METHODS

Data from Israel include different age groups, from young teenagers to elderly (12 to 90+, ten age ranges), data also includes "number of cases" per age group by vaccination status and "population size" for each age group. In this research, we also introduce the concept of vaccine efficacy, which is defined by: $1 - \frac{p^{(v)}}{p^{(u)}}$ ($p^{(v)}$ and $p^{(u)}$ noted as the severe Coivd rate on vaccinated and unvaccinated subpopulations). Firstly, as each person's Covid severe diseases' outcome (infected or not infected) is independent from everyone else's, the binomial distribution should apply in our research. So the cases of severe disease caused by covid is the "successful cases" in our research, "population size" for each age group is n.

For building 95% confidence interval, as a serious statistical researcher, we also need to consider this confidence interval whether actually contain the unknown rate of severe diseases caused by Covid. In a professional statistics paper called "Interval Estimation for a Binomial Proportion", the authors proved that for binomial proportion, the standard interval (Wald interval) has inadequate coverage, which means if we use standard interval, the 95% confidence interval we constructed is unlikely to contain the actually case rate(Brown et al. 101). The question is arised here, what we should do under this situation?

Luckily, in the same paper, Brown, Lawrence D., et al mention other alternative intervals to replace the wald interval. In our study, we choose Agresti-Coull Interval to construct 95%

interval for capture the case rates. The reason is that as Brown, Lawrence D., et al suggest, the Agresti-Coull Interval has a better performance on coverage than Wald interval (Brown et al. 107), which means Agresti-Coull Interval are more likely to contain true case rates .

RESULTS

Figure 1 below includes a 95% confidence interval for the underlying rates of severe disease in the vaccinated and unvaccinated subpopulations. Because of the relatively small population of severe disease, all of the rates are not high, but it can be seen that with the increase of age, the rates of severe disease increases significantly. The rate is almost negligible between the agex of 12-15, but the people over the ages of 90 with a maximum rate. The table below also contains the interval for the vaccinated subpopulations. It is obvious that the rates of severe disease have the same trend in both vaccinated and unvaccinated groups. In other words, the rate increases significantly with age. At the same time, by comparing the two results, it can be found that the rates of severe disease of the vaccinated “populations” are much lower than that of the unvaccinated “population”, there is almost a ten-fold difference between the two “popluation” at each age group, which means if we take each 100000 people from vaccinated and unvaccinated people at ages 60-69, we could prevent hundreds of people from severe diseases.

Ages	Lowest Case Rate for Unvaccinated	Highest Case Rate for Unvaccinated	Lowest Case Rate for Vaccinated	Highest Case Rate for Vaccinated
12-15	0.0000000	0.0000163	0.0000000	0.0000251
16-19	0.0000003	0.0000611	0.0000000	0.0000108
20-29	0.0000043	0.0000402	0.0000000	0.0000047
30-39	0.0000340	0.0001093	0.0000000	0.0000081
40-49	0.0001097	0.0002470	0.0000048	0.0000188
50-59	0.0002859	0.0005638	0.0000192	0.0000448
60-69	0.0005797	0.0010127	0.0000672	0.0001128
70-79	0.0013841	0.0026051	0.0001614	0.0002432
80-89	0.0017751	0.0035723	0.0003932	0.0005825
90+	0.0030765	0.0083531	0.0002398	0.0006151

Figure 1: For each age group, the 95% confidence interval for the underlying rates of severe disease in the vaccinated and unvaccinated subpopulations

Ages	estimated vaccine efficacy	Lowest vaccine efficacy	Highest vaccine efficacy
12-15	1.000	0.000	1.000
16-19	1.000	0.000	2.035
20-29	1.000	0.199	1.684
30-39	1.000	0.492	1.434
40-49	0.956	0.734	1.146
50-59	0.933	0.849	1.005
60-69	0.884	0.858	0.915
70-79	0.883	0.747	1.043
80-89	0.778	0.436	1.179
90+	0.922	0.897	0.952

Figure 2: a point estimate and corresponding confidence interval for the vaccine efficacy

In other words, vaccines have a significant effect on preventing hospitalization. The effect of the vaccine is particularly pronounced for people over 30 years old. The rates of severe disease of these people who have been vaccinated are ten times less than that of people who have not been vaccinated. From the point estimation results of vaccines, we can conclude that vaccines have a significant preventive effect on the disease.

In the Figure 2, we might be astonished at the results of vaccine efficacy on two ages groups (12-15 and 16-19). The values of lowest vaccine efficacy are zero, this does not means

vaccine is not work for these two ages groups, just because there isn't enough data to construct a small confidence interval.

It can be seen from the following two pictures that with the increase of age, the minimum value of vaccine effectiveness gradually increases, while the maximum value of vaccine effectiveness gradually decreases. However, in the younger age groups, the minimum value of vaccine efficiency may be small because the number of patients is very small or even zero. So we might suppose that with the increase of age, the efficiency of the vaccine should gradually decline. However, there is a relatively slight increase in vaccine efficacy between ages 70-79 and 80-89, as can be seen from the data.

The vaccine efficiency of all age groups was calculated, and then get the confidence interval of the vaccine efficiency was. It can be seen from the results that for all groups, the minimum vaccine efficiency is also close to 60% (the process of calculation is at the end of Appendix). But this result might be misleading, we will discuss this in the next section.

DISCUSSION

As indicated in the results, the vaccine actually work comparing with people who did not get vaccine before, because the rates of severe disease of the vaccinated subpopulations are much lower than that of the unvaccinated population, there is almost a ten-fold difference between the two columns. Moreover, except for younger age groups (people under 19), the vaccine efficacies for each other age group is not bad at all. This is a solid study result for us to persuade adults who do not believe that vaccine could protect themselves from severe diseases caused by covid-19 variants.

With respect to the statistical methods in this study, it is very important for researchers to not ignore that there are other potential methods that may suitable for this statistical analysis. While the Agresti-Coull interval ensures the coverage and well performance on large sample sizes, the Agresti-Coull interval also is known to overfit the confidence level, when cases rate is around to 0 or 1, which causes greater coverage than we actually need and leads deficiencies in length. For some other methods, the Willson score interval may be a substitute way when cases rate is around 0 or 1, such as for the younger age groups in our dada.

Age is an unmistakable and significant confounder in our data. At the different age groups, the vaccination rates have significant differences, especially between young and elders. When vaccination rates are low, using raw trials can exaggerate the effectiveness of the vaccine. Instead, when vaccination rates are high, using such raw trials can weaken the effectiveness of the vaccine by making it appear less effective than it actually is. Age-specific vaccine efficacy does not care about differences in vaccination rate. In short, Age-specific vaccine efficacy ensures each cases of severe diseases within each age group receives proper representation within the sample. As a result, age-stratified random sampling provides better coverage of the population since the researchers have control over the different age groups to ensure all of them are represented in the sampling. Therefore, Unvaccinated versus vaccinated people within a given age range would make a much more fair comparison, and The vaccine efficiency of all age groups might cause bias.

Works Cited

Brown, Lawrence D., et al. “Interval Estimation for a Binomial Proportion.” *Statistical Science*, vol. 16, no. 2, 2001, https://doi.org/10.1214/ss/1009213286.

Unit 1 Paper Appendix, Version 2

Xiangsen Yang

1/17/2022

obtaining the data and package

```
library(confintr)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
data = read.delim("https://dept.stat.lsa.umich.edu/~bbh/s485/data/israel_severe_covid_2021_08_15.tsv")
list(data)
```

```
## [[1]]
##     ages cases_unvax cases_vax pop_unvax pop_vax
## 1  12-15           1         0    383649  184549
## 2  16-19           2         0    127745  429109
## 3  20-29           4         0    265871  991408
## 4  30-39          12         2    194213  968837
## 5  40-49          24         9    145355  927214
## 6  50-59          34        22     84545  747949
## 7  60-69          50        58     65205  665717
## 8  70-79          39        92     20512  464336
## 9  80-89          32       100     12683  208911
## 10   90+          16        18      3132   46602
```

Overview

The computations in this document explore the vaccinated and unvaccinated hospitalizations rate from different age groups in Israel, which includes the calculation for Agresti-Coull confidence intervals for different binomial variables.

CI of the Agresti-Coull interval

Combining BCD equation (5) on p. 108 with notation defined in section 3.1.2 on p.108, the Agresti-Coull interval is:

$CI_{AC} = \tilde{p} \pm \kappa\sqrt{\tilde{p}(1-\tilde{p})}\sqrt{\frac{1}{n+\kappa^2}}$,

Denote $\tilde{X} = X + \kappa^2/2$, $\tilde{n} = n + \kappa^2$.

1.

AC interval might includes negative number, but the rate is on the interval [0,1], so we limit our Agresti-Coull interval on[0,1]

Result of interval in the unvaccinated :

```
#interval for the underlying rates of severe disease in the unvaccinated
Interval_U <- as.data.frame(matrix(nrow=0,ncol=3))
names(Interval_U)<-c("ages","PU_interval_2.5","PU_interval_97.5")
a <- 1
while (a <= length(data[,1])){
  R_Agresti <- ci_proportion(data[a,'cases_unvax'],n =data[a,'pop_unvax'],
                             type = "Agresti-Coull" )

  Interval_U <- rbind(Interval_U,
                      data.frame(ages = data[a,'ages'] ,
                                 PU_interval_2.5 = round(R_Agresti$interval[1], digits = 7),
                                 PU_interval_97.5=round(R_Agresti$interval[2], digits = 7)))
  a <- a + 1
}
Interval_U
```

```
##     ages PU_interval_2.5 PU_interval_97.5
## 1  12-15       0.0000000        0.0000163
## 2  16-19       0.0000003        0.0000611
## 3  20-29       0.0000043        0.0000402
## 4  30-39       0.0000340        0.0001093
## 5  40-49       0.0001097        0.0002470
## 6  50-59       0.0002859        0.0005638
## 7  60-69       0.0005797        0.0010127
## 8  70-79       0.0013841        0.0026051
## 9  80-89       0.0017751        0.0035723
## 10   90+       0.0030765        0.0083531
```

Result of interval in the vaccinated :

```
#interval for the underlying rates of severe disease in the vaccinated
Interval_V <- as.data.frame(matrix(nrow=0,ncol=3))
names(Interval_V)<-c("ages","PV_interval_2.5","PV_interval_97.5")
b <- 1
while (b <= length(data[,1])){
  R_Agresti <- ci_proportion(data[b,'cases_vax'],
                             n =data[b,'pop_vax'],type = "Agresti-Coull" )
  Interval_V <- rbind(Interval_V,
                      data.frame(ages = data[b,'ages'] ,
                                 PV_interval_2.5 = round(R_Agresti$interval[1], digits = 7),
                                 PV_interval_97.5 = round(R_Agresti$interval[2], digits = 7)))
  b <- b + 1
}
Interval_V
```

```
##     ages PV_interval_2.5 PV_interval_97.5
## 1  12-15       0.0000000        0.0000251
## 2  16-19       0.0000000        0.0000108
## 3  20-29       0.0000000        0.0000047
## 4  30-39       0.0000000        0.0000081
## 5  40-49       0.0000048        0.0000188
## 6  50-59       0.0000192        0.0000448
## 7  60-69       0.0000672        0.0001128
## 8  70-79       0.0001614        0.0002432
```

```
## 9  80-89       0.0003932        0.0005825
## 10   90+       0.0002398        0.0006151
```

The overall result, as shown below:

```
PV_PU <- merge(Interval_U,Interval_V,by="ages")
PV_PU
```

```
##     ages PU_interval_2.5 PU_interval_97.5 PV_interval_2.5 PV_interval_97.5
## 1  12-15       0.0000000        0.0000163       0.0000000        0.0000251
## 2  16-19       0.0000003        0.0000611       0.0000000        0.0000108
## 3  20-29       0.0000043        0.0000402       0.0000000        0.0000047
## 4  30-39       0.0000340        0.0001093       0.0000000        0.0000081
## 5  40-49       0.0001097        0.0002470       0.0000048        0.0000188
## 6  50-59       0.0002859        0.0005638       0.0000192        0.0000448
## 7  60-69       0.0005797        0.0010127       0.0000672        0.0001128
## 8  70-79       0.0013841        0.0026051       0.0001614        0.0002432
## 9  80-89       0.0017751        0.0035723       0.0003932        0.0005825
## 10   90+       0.0030765        0.0083531       0.0002398        0.0006151
```

Ages	Lowest Case Rate for Unvaccinated	Highest Case Rate for Unvaccinated	Lowest Case Rate for Vaccinated	Highest Case Rate for Vaccinated
12-15	0.0000000	0.0000163	0.0000000	0.0000251
16-19	0.0000003	0.0000611	0.0000000	0.0000108
20-29	0.0000043	0.0000402	0.0000000	0.0000047
30-39	0.0000340	0.0001093	0.0000000	0.0000081
40-49	0.0001097	0.0002470	0.0000048	0.0000188
50-59	0.0002859	0.0005638	0.0000192	0.0000448
60-69	0.0005797	0.0010127	0.0000672	0.0001128
70-79	0.0013841	0.0026051	0.0001614	0.0002432
80-89	0.0017751	0.0035723	0.0003932	0.0005825
90+	0.0030765	0.0083531	0.0002398	0.0006151

2. a point estimate and corresponding confidence interval for the vaccine efficacy

Result of interval for the vaccine efficacy :

```
#Efficiency Mean Point Estimation
c <-1
while (c <= length(PV_PU[,1])) {
  point_U<- mean(PV_PU$PU_interval_2.5[c],PV_PU$PU_interval_97.5[c])
  point_V<- mean(PV_PU$PV_interval_2.5[c],PV_PU$PV_interval_97.5[c])
  PV_PU[c,'Point_est'] <- round(1-point_V/point_U,3)
  c <- c+1
}
PV_PU[which(is.nan(PV_PU$Point_est)),'Point_est']<-1

#Confidence Intervals for Vaccine Efficiency
d <-1
while (d <= length(PV_PU[,1])) {
  #Calculate the vaccine efficiency according to the rate of the first question
  efficacy_Lower <- 1-PV_PU$PV_interval_2.5[d]/PV_PU$PU_interval_2.5[d]
  efficacy_Upper <- 1-PV_PU$PV_interval_97.5[d]/PV_PU$PU_interval_97.5[d]
```

```
  #confidence interval
  Ci_P <- CI(c(efficacy_Lower,efficacy_Upper))

  if(is.nan(Ci_P[1])){
    Ci_P[1]<-1
  }
  #The vaccine efficiency cannot be negative, and the negative value is set to 0
  if(is.nan(Ci_P[3]) || Ci_P[3]<0 ){
    Ci_P[3]<-0
  }
  PV_PU$efficacy_Lower[d] <- efficacy_Lower
  PV_PU$efficacy_Upper[d] <- efficacy_Upper

  PV_PU$Eff_Lower[d] <- round(Ci_P[3],3)
  PV_PU$Eff_Upper[d] <- round(Ci_P[1],3)

  d <- d+1
}

PV_PU[,c('Point_est','Eff_Lower','Eff_Upper')]
```

```
##    Point_est Eff_Lower Eff_Upper
## 1      1.000     0.000     1.000
## 2      1.000     0.000     2.035
## 3      1.000     0.199     1.684
## 4      1.000     0.492     1.434
## 5      0.956     0.734     1.146
## 6      0.933     0.849     1.005
## 7      0.884     0.858     0.915
## 8      0.883     0.747     1.043
## 9      0.778     0.436     1.179
## 10     0.922     0.897     0.952
```

Ages	estimated vaccine efficacy	Lowest vaccine efficacy	Highest vaccine efficacy
12-15	1.000	0.000	1.000
16-19	1.000	0.000	2.035
20-29	1.000	0.199	1.684
30-39	1.000	0.492	1.434
40-49	0.956	0.734	1.146
50-59	0.933	0.849	1.005
60-69	0.884	0.858	0.915
70-79	0.883	0.747	1.043
80-89	0.778	0.436	1.179
90+	0.922	0.897	0.952

3. The estimated vaccine efficacy as calculated without adjustment for age, (Where pv and pu) denote severe COVID-19 rates across vaccinated and unvaccinated Israelis aged 12 and older.

```
#PV and PU for all age groups
PU <- ci_proportion(sum(data[,'cases_unvax']),n =sum(data[,'pop_unvax']),type = "Agresti-Coull" )
```

```r
PV <- ci_proportion(sum(data[,'cases_vax']),n =sum(data[,'pop_vax']),type = "Agresti-Coull" )

#Vaccine Efficiency at All Ages
efficacy_U_2.5 <- 1-PV$interval[1]/PU$interval[1]
efficacy_U_97.5 <- 1-PV$interval[2]/PU$interval[2]
all <- c(efficacy_U_2.5,efficacy_U_97.5)
#confidence interval
CI(all)
```

```
##     upper      mean     lower
## 0.7618178 0.6746826 0.5875475
```

The mean estimated vaccine efficacy is about 68%.