Stat Final Paper Part1

Jingzhao Huang

6/4/2021

In the first part, we want to explore whether there is difference between the ratio of the being diseased (COVID-19) in the vaccine group and that of the placebo group.

Methods:

Let X denote the number of the being diseased in the vaccine group and let Y denote the number of being diseased in the placebo group. We will use binomial models to assume

$$X \sim Binom(17419, \pi_1)$$

$$Y \sim Binom(17673, \pi_2)$$

Let $n_1 = 17419$ represent the number of people in the vaccine group and let $n_2 = 17673$ represent the number of people in the placebo group. Let $x = 8$ represent the number of the being diseased in the vaccine group and let $y = 162$ represent the number of the being diseased in the placebo group.

The hypothesis are :

$$H_0 : \pi_1 - \pi_2 = 0$$

$$H_1 : \pi_1 - \pi_2 \neq 0$$

Based on the maximum likelihood method, we know the MLE for π_1 is $\frac{x}{n_1}$ and the MLE for π_2 is $\frac{y}{n_2}$.

We will use Likelihood Ratio Test, Wald confidence interval and Score Z-test with the significance level $\alpha = 0.05$ to test the hypothesis.

Procedure

First let us use Wald confidence interval.

Based on the maximum likelihood method, we know the MLE for $\pi_1 - \pi_2$ is $\hat{\pi_1} - \hat{\pi_2}$ where $\hat{\pi_1} = \frac{x}{n_1}$ and $\hat{\pi_2} = \frac{y}{n_2}$.

Then the standard error of sample difference of proportion under H_1 is :

$$SE_{\pi_1 - \pi_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

The observed standard error of sample difference of proportion under H_1 is :

$$SE_{\hat{\pi_1} - \hat{\pi_2}} = \sqrt{\frac{\hat{\pi_1}(1-\hat{\pi_1})}{n_1} + \frac{\hat{\pi_2}(1-\hat{\pi_2})}{n_2}}$$

So the Wald-Based 95% confidence interval (under H_1) based on the observed sample is calculated as

$$(\hat{\pi_1} - \hat{\pi_2} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi_1}(1-\hat{\pi_1})}{n_1} + \frac{\hat{\pi_2}(1-\hat{\pi_2})}{n_2}}, \hat{\pi_1} - \hat{\pi_2} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi_1}(1-\hat{\pi_1})}{n_1} + \frac{\hat{\pi_2}(1-\hat{\pi_2})}{n_2}})$$

Since we know $\hat{\pi_1} = x/n_1 = 8/17411$, $\hat{\pi_2} = y/n_2 = 162/17511$ and $z_{1-\alpha/2} = 1.96$, the Wald-Based 95% confidence interval (under H_1) based on the observed sample is

$$(\frac{8}{17411} - \frac{162}{17511} - 1.96\sqrt{\frac{\frac{8}{17411}(1-\frac{8}{17411})}{17411} + \frac{\frac{162}{17511}(1-\frac{162}{17511})}{17511}}, \frac{8}{17411} - \frac{162}{17511} + 1.96\sqrt{\frac{\frac{8}{17411}(1-\frac{8}{17411})}{17411} + \frac{\frac{162}{17511}(1-\frac{162}{17511})}{17511}})$$

```
##                est      lwr.ci       upr.ci
## [1,] -0.008791848 -0.01024514 -0.007338557
```

which is about

$$(-0.01024514, -0.007338557)$$

Since the Wald-Based 95% confidence interval (under H_1) based on the observed sample doesn't contain 0, we can reject the null hypothesis.

Now we use Z-score test:

Then the estimated standard error of sample difference of proportion under H_0 is

$$SE_{\hat{\pi_1}-\hat{\pi_2}} = \sqrt{\frac{\hat{\pi_1}(1-\hat{\pi_1})}{n_1} + \frac{\hat{\pi_2}(1-\hat{\pi_2})}{n_2}} = \sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}$$

where $\hat{\pi} = \frac{x+y}{n_1+n_2}$.

Then we get the observed z-value :

$$z = \frac{\hat{\pi_1} - \hat{\pi_2}}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{\frac{8}{17411} - \frac{162}{17511}}{\sqrt{\frac{8+162}{17411+17511}(1-\frac{8+162}{17411+17511})(\frac{1}{17411} + \frac{1}{17511})}} = -11.80274$$

```
## [1] -11.80274
```

Thus the approximated two-sided p-value is

$$pvalue = 2 \times [1 - \Phi(|z|)] = 0$$

```
## [1] 0
```

Since the p-value is 0, which lower than the set significance level $\alpha = 0.05$, we reject the null hypothesis.

Now we use Likelihood Ratio test:

We set the parameter set under H_0:

$$\Omega_0 = \{(\pi_1, \pi_2) : \pi_1 = \pi_2 = \pi, 0 < \pi_1, \pi_2 < 1\}$$

Then we set the parameter set under H_1:

$$\Omega = \{(\pi_1, \pi_2) : 0 < \pi_1, \pi_2 < 1\}$$

Then we know

$$dim(\Omega_0) = 1$$
$$dim(\Omega) = 2$$

To compute the log form of the likelihood function $L(\hat{\Omega}_0)$, we need to maximize

$$L(\pi) = \pi^x(1-\pi)^{n_1-x}\pi^y(1-\pi)^{n_2-y} = \pi_{x+y}(1-\pi)^{n_1+n_2-x-y}$$

This max occurs at

$$\hat{\pi} = \frac{x+y}{n_1+n_2} = \frac{8+162}{17411+17511} = \frac{170}{34922}$$

To compute the log form of the likelihood function $L(\hat{\Omega})$, we need to maximize

$$L(\pi) = \pi_1^x(1-\pi_1)^{n_1-x}\pi_2^y(1-\pi_2)^{n_2-y}$$

This max occurs at

$$\hat{\pi_1} = \frac{x}{n_1} = \frac{8}{17411}$$

$$\hat{\pi_2} = \frac{y}{n_2} = \frac{162}{17511}$$

The the ratio of of the likelihood functions is

$$\lambda = \frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\hat{\pi}^{x+y}(1-\hat{\pi})^{n_1+n_2-x-y}}{\hat{\pi_1}^x(1-\hat{\pi_1})^{n_1-x}\hat{\pi_2}^y(1-\hat{\pi_2})^{n_2-y}}$$

The p-value is

$$p-value = P(Chisq(df=1) \geq -2\lambda) = 0$$

```
## [1] NaN
```

```
library(maxLik)
```

```
## Loading required package: miscTools
```

```
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

```
loglik0<-function(p,x,y,nobs,mobs){
  dbinom(x,size=nobs,prob=p,log=T)+dbinom(y,size=mobs,prob=p,log=T)
}
loglik1<-function(p,x,y,nobs,mobs){
  p1=p[1]
  p2=p[2]
  dbinom(x,size=nobs,prob=p1,log=T)+dbinom(y,size=mobs,prob=p2,log=T)
}
ml0<-maxLik(logLik=loglik0,start=0.5,x=8,y=162,nobs=17411,mobs=17511)
ml1<-maxLik(logLik=loglik1,start=c(0.5,0.5),x=8,y=162,nobs=17411,mobs=17511)
lrtstat=-2*(logLik(ml0)-logLik(ml1))
1-pchisq(q=lrtstat,df=1)
```

```
## [1] 0
## attr(,"df")
## [1] 1
```

Since the p-value is 0, which lower than the set significance level $\alpha = 0.05$, we reject the null hypothesis.

Conclusion: The three test statistics—Wald, Likelihood Ratio and score—are asymptotically equivalent, that is they will all have the same sampling distribution which is a chi-squre distribution with the same degrees

of freedom. Wald test uses the MLE and depends on curvature of likelihood at the MLE. The Score test depends on the slope and curvature at 0, while LR test uses information from both the MLE and 0. In theory, if the samples sizes are large, the statistics of Wald, Score and Likelohood Ratio methods are almost the same, which leads them to result in the same answer. The sample size 17411 of the vaccine group and the sample size 17511 of the placebo group are large, so the same results (failing to reject the null hypothesis that $\pi_1 - \pi_2 = 0$) of Wald, Score and Likelihood Ratio methods are resonable. Therefore, we can conclude that since we fail to reject the null hypothesis, it is significantlly important that the probability of being diseased after taking the vaccine is not equal to the probability of being diseased without taking the vaccine.

Reference: 1.6.6 -Hypothesis tests & related Intervals: STAT 504. PennState: Statistics Online Courses. (n.d.). https://online.stat.psu.edu/stat504/lesson/1/1.6/1.6.6.