Report 1

Wei Zhong

February 12, 2017

Introduction

According to the World Health Organization (WHO), ebola virus disease (EVD), formerly known as Ebola haemorrhagic fever, is defined as a severe, often fatal illness in humans. The virus is transmitted to people from wild animals and spreads in the human population through human-to-human transmission. The average EVD case fatality rate is around 50%. Case fatality rates have varied from 25% to 90% in past outbreaks.[1]

Between 1 September and 24 October 1976, 318 cases of acute viral haemorrhagic fever occurred in northern Zaire. The outbreak was centred in the Bumba Zone of the Equateur Region and most of the cases were recorded within a radius of 70 km of Yambuku, although a few patients sought medical attention in Bumba, Abumombazi, and the capital city of Kinshasa, where individual secondary and tertiary cases occurred. There were 280 deaths, and only 38 serologically confirmed survivors. The index case in this outbreak had onset of symptoms on 1 September 1976, five days after receiving an injection of chloroquine for presumptive malaria at the outpatient clinic at Yambuku Mission Hospital (YMH). He had a clinical remission of his malaria symptoms. Within one week several other persons who had received injections at YMH also suffered from Ebola haemorrhagic fever, and almost all subsequent cases had either received injections at the hospital or had had close contact with another case. Most of these occurred during the first four weeks of the epidemic, after which time the hospital was closed, 11 of the 17 staff members having died of the disease. All ages and both sexes were affected, but women 15-29 years of age had the highest incidence of disease, a phenomenon strongly related to attendance at prenatal and outpatient clinics at the hospital where they received injections. The overall secondary attack rate was about 5%, although it ranged to 20% among close relatives such as spouses, parent or child, and brother or sister.[2]

Nearly after 20 years, ebola broke out again in Zaire. In May 10 of 1995 an appeal from Zaire reached the WHO about an epidemic outbreak of haemorrhagic fever in the city of Kikwit. The information was in the beginning fragmentary and contradictory. Viral samples were sent to USA where they were rapidly identified as the Ebola virus, known to be responsible for earlier epidemics in the 1970s. with a very high mortality rate. Several International Organizations arrived early at Kikwit in order to help the local health authorities to control the epidemic. These organisations did establish under the direction of WHO a well functioning collaboration with the local authorities. An intensive epidemiologic mapping was performed and the propagation and the extent of the epidemic could rapidly be evaluated. The spread of the disease did probably occur as early as January 1995 but continued and remained at a small extent around the city of Kikwit. In April an infected patient was operated at Kikwit General Hospital causing an explosive spread of the disease. A large number of other patients at the hospital were infected and approximately 50 hospital staff died due to Ebola fever. Not until then had the health authorities been aware of the outbreak of a serious epidemic. When the epidemic ended in July 1995 a total of 316 cases were reported, 245 of these died (78%). Despite substantial action by different organisations there was no adequate care offered to the Ebola patients.[3]

Compared to the outbreak in 1976, ebola virus was well known and studied by epidemiologists, and the local governments and international organization were more alarmed. However, in Zaire the dealth rate caused by ebola virus in 1995 were still pretty high, close to 80.6%, compared to the death rate of 88.1% in 1976 outbreak. Therefore, in this paper, I would like to test first whether there exists significant difference of ebola dealth rate in between the 1976 outbreak and the 1995 outbreak. Second, I tend to test whether this difference if exists, are larger than 5%.

[1] http://www.who.int/mediacentre/factsheets/fs103/en/

[2] International Commission. "Ebola haemorrhagic fever in Zaire, 1976." Bull World Health Organ 56.2 (1978): 271-293.

[3] http://www.socialstyrelsen.se/publikationer1997/theebolavirusepidemicinzaire1995kamedo-report69

Table 1: The Contigency Table of Ebola Outbreaks in 1976 and in 1995

	Outbreak in 1976	**Outbreak in 1995**
Dead	$N_{11} = 280$	$N_{12} = 254$
Not dead	$N_{21} = 38$	$N_{22} = 61$
Total	$n_{+1} = 318$	$n_{+2} = 315$

Data

In Zaire, in the 1976 outbreak of Zaire, the 318 total recorded human cases resulted in 280 deaths; and in the 1995 outbreak, the 315 total recorded human cases resulted in 254 deaths. Based on the record, I create a 2 by 2 contigency table, which shows the number of death and of non-death cases in 1976 and in 1995. According to the records, the estimates of proportions of 1976 and 1995 outbreaks are $\hat{\pi}_{D1} = 280/381 \approx 0.88$ and $\hat{\pi}_{D2} = 254/315 \approx 0.81$ respectively. Thus, in 1976 the death rate of ebola rivus in Zaire was about 88% while in 1995, the death rate was 81%. The relative risk is $RR = \hat{\pi}_{D1}/\hat{\pi}_{D2} \approx 1.09$, which means in 1976 the people who was infected by ebola virus had 1.09 times the risk of dying compared to victims infected in 1995. Furthermore, the odds ratio $\hat{\theta} = \frac{\hat{\pi}_{D1}(1-\hat{\pi}_{D1})}{\hat{\pi}_{D2}(1-\hat{\pi}_{D2})} \approx 1.78$. It means patients infected by ebola virus are associated with higher odds of death than patients in 1995.

Model

Decide upon a probability model for the analysis. Clearly specify the model and the meaning of each parameter. Discuss any limitations of the model (especially in comparison to other possible models).

In this case, I use independent binomial models as following:

$$N_{11} \sim Binomial(n_{+1}, \pi_{D1})$$

where n_{+1} is the number of total reported cases who were infected by ebola virus in 1976, and π_{D1} is the probability of ebola death in 1976 outbreak.

$$N_{12} \sim Binomial(n_{+2}, \pi_{D2})$$

where n_{+2} is the number of total reported cases who were infected by ebola virus in 1995, and π_{D2} is the probability of ebola death in 1995 outbreak.

The limitation of this model is that we assume two outbreaks are independent. In reality, it might not be the case.

Analysis

First Research Question

To start with, I calculate the standard errors for estimates of proportions, the relative risk and the odds ratio, then compute the corresponding confidence interval at the 0.05 level.

The standard error for $\hat{\pi}_{D1}$:

$$\hat{se}(\hat{\pi}_{D1}) = \sqrt{\hat{\pi}_{D1}(1-\hat{\pi}_{D1})/318} \approx 0.018$$

Thus the 95% CI for $\hat{\pi}_{D1}$ is:

$$\hat{\pi}_{D1} \pm 1.96 \times \hat{se}(\hat{\pi}_{D1}) = (0.8448510, 0.9161553)$$

The standard error for $\hat{\pi}_{D2}$:

$$\hat{se}(\hat{\pi}_{D2}) = \sqrt{\hat{\pi}_{D2}(1-\hat{\pi}_{D2})/315} \approx 0.022$$

The 95% CI for $\hat{\pi}_{D2}$ is:

$$\hat{\pi}_{D2} \pm 1.96 \times \hat{se}(\hat{\pi}_{D2}) = (0.7627105, 0.8499879)$$

The standard error for log of the relative risk RR is:

$$\hat{se}(\log(RR)) = \sqrt{\frac{38}{280 \times 318} + \frac{61}{254 \times 315}} \approx 0.034$$

And the 95% CI for RR is:

$$\exp(\log(rr) \pm 1.96 \times \hat{se}(log(RR))) = (1.020596, 1.168319)$$

The standard error for log of the odds ratio $\hat{\theta}$ is:

$$\hat{se}(\hat{\theta}) = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}} \approx 0.224$$

The 95% CI for the odds ratio $\hat{\theta}$ is:

$$\exp(\log(\hat{\theta}) \pm 1.96 \times \hat{se}(\hat{\theta})) = (1.140562, 2.745506)$$

As following, I calculate the "theoretical frequency", under the null hypothesis $H_0 : \pi_{D1} = \pi_{D2}$, which is against the alternative $H_a : \pi_{D1} \neq \pi_{D2}$. Then under the null hypothesis, $\tilde{\pi}_{D1} = \tilde{\pi}_{D2} = \frac{280+254}{318+315} \approx 0.844$. Thus the expected observation should be $E = (318 \times 0.844, 318 \times (1 - 0.844), 315 \times 0.844, 315 \times (1 - 0.844))$

Then the Pearson chi-square test for two outbreaks:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx 6.595282$$

And its corresponding p-value is 0.01022494.

The odds ratio chi-square statistic G^2 is:

$$G^2 = 2 \sum_{ij} O_{ij} \times \log(\frac{O_{ij}}{E_{ij}}) \approx 6.644771$$

And its corresponding p-value is 0.009944725.

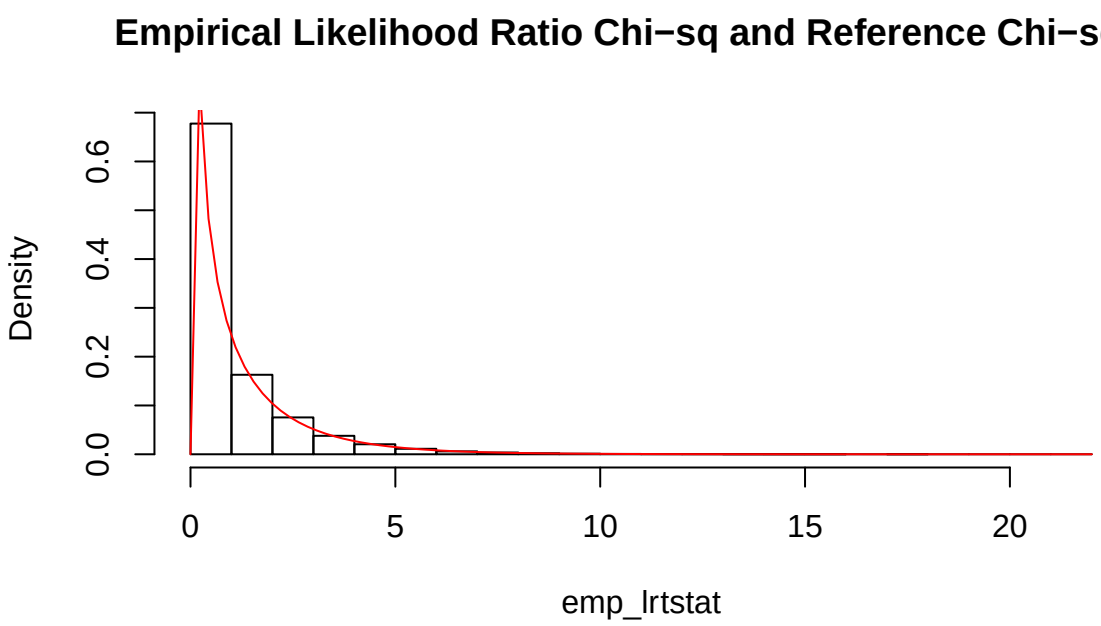
Both odds ratio chi-square statistics and Pearson chi-square statistics show that we can reject the null hypothesis and conclude that the death rate of two breaks in Zaire are signficantly different at the 5% level, though the number of death cases are pretty close.

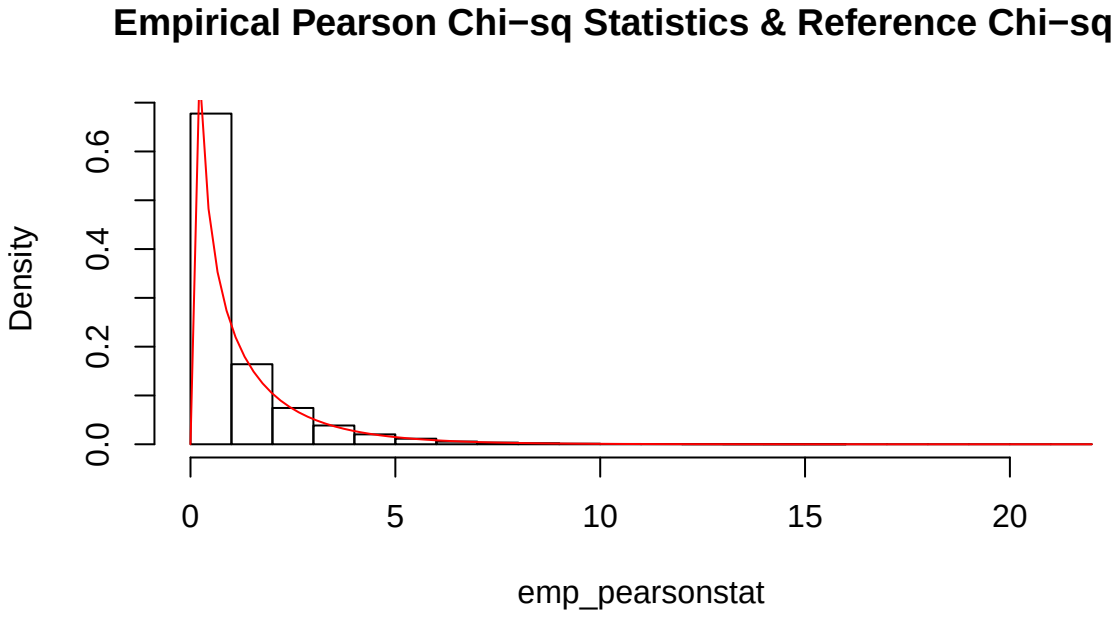
Second Research Question

In the second research question, I test whether the death rate in 1976 is at least 5% higher than the death rate of 1995 outbreak. Thus the null hypothesis is $H_0 : \pi_{D1} = \pi_{D2} + 0.05$, which is against the alternative hypothesis $H_a : \pi_{D1} > \pi_{D2} + 0.05$. By using the asymptotic normality, the z-score is 0.8401245, and its corresponding p-value is 0.2004193. It means we cannot reject the null hypothesis, thus the death rate in 1976 outbreak is at most 5% higher than that in 1995 outbreak. In other words, the difference of death rates between two outbreaks is not higher than 5%.

Simulation

Through a 100000 simulation, assuming the null hypothesis $H_0 : \pi_{D1} = \pi_{D2}$ is true, I generate the empirical Pearson chi-square statistics and empirical likelihood ratio G^2 statistics. The two empirical distributions are shown below. The correponding p-values of both Pearson chi-square and likelihood ratio is 0.01032, which is higher than 0.01022494 of Pearson chi-square and 0.009944725 of odds ratio chi-square test.





Conclusion

In this paper, I exmaine two research question. One is whether there is statistically significant difference between death rates of two ebola outbreaks in 1976 and 1995. Another is whether the difference of two death rates is statsitically signficant larger than 5% if it exists.

Throught compare the confidence interval of two estimates of proportions π_{D1} and π_{D2}, it seems the death rate in 1997 is higher than in 1995. Additionally, the p-values of the relative risk and odds ratio suggest we

can reject the null hypothesis $H_0 : \pi_{D1} = \pi_{D2}$. In other words, the death rates between two outbreaks are significantly different. Furthermore, I test whether death rate in 1976 is at least 5% higher than in 1995. The Z value and its corresponding p-value suggest that the death rate in 1976 outbreak is at most 5% higher than that in 1995 outbreak. The potential reasons for this difference might be: first, people in Zaire have better knowledge about the ebola virus, thereby decreasing the death rate. Second, international organizations and local governments are more alarmed about the ebola outbreak and more effectively and effeciently at dealing with the outbreak. Nevertheless, the fatality rate of ebola virus disease is still pretty high due to its difficult diagonsis and no proven treatment available.

Appendix

```
rm(list = ls())

## Setup ##
Nsim = 100000
n_1976 = 318.0
n_1995 = 315.0

## define pearson chi-sq stat
statPearson = function(obs, exp){
  stat = sum((obs - exp)^2/ exp)
  return(stat)
}

## define likelihood ratio chi-sq stat
statLRT = function(obs, exp){
  g.sq = 2* sum(obs * log(obs/exp))
  return(g.sq)
}

## estimates of proportions
pi_d1 = 280.0/n_1976
pi_d2 = 254.0/n_1995
odds1 = pi_d1 / (1-pi_d1)
odds2 = pi_d2 / (1-pi_d2)

## relative risk
rr = (pi_d1/pi_d2)

## odds ratio
theta = (odds1/odds2)

print(paste("1976 Death Rate:", pi_d1))
print(paste("1995 Death Rate:", pi_d2))
print(paste("The relatie risk:", rr))
print(paste("The odds ratio:", theta))


## calculate CI for proportions, relative risk, and the odds ratio ##
se_pi_d1 = sqrt(pi_d1 * (1-pi_d1) / n_1976)
se_pi_d2 = sqrt(pi_d2 * (1-pi_d2) / n_1995)
se_log_rr = sqrt(38.0/(280.0*318.0) + 61.0/(254.0*315))
se_theta = sqrt(1.0/38 + 1.0/280 + 1.0/61 + 1.0/254)

pi_d1_ci = c(pi_d1 - 1.96 * se_pi_d1, pi_d1 + 1.96 * se_pi_d1)
pi_d2_ci = c(pi_d2 - 1.96 * se_pi_d2, pi_d2 + 1.96 * se_pi_d2)
rr_ci = exp(c(log(rr)-1.96*se_log_rr, log(rr) + 1.96*se_log_rr))
theta_ci = exp(c(log(theta) - 1.96 * se_theta, log(theta) + 1.96 * se_theta))

### Pearson Chi-Squared Test
obs = c(280.0, 38.0, 254.0, 61.0)
mle = (280.0 + 254.0) / (n_1976 + n_1995)
exp = c(318*mle, 318*(1-mle), 315*mle, 315*(1-mle))
```

```r
pearson_obs = statPearson(obs, exp)
lrt_obs  = statLRT(obs, exp)

## p-value for Pearson and LRT
pearson_pvalue = 1- pchisq(pearson_obs, df = 1)
lrt_pvalue = 1 - pchisq(lrt_obs, df = 1)

## simulation ##
pearson_extreme  = rep(0, Nsim)
lrt_extreme = rep(0, Nsim)
obser  = matrix(nrow = Nsim, ncol = 4, 0)
expec = matrix(nrow = Nsim, ncol = 4, 0)
MLE = rep(0, Nsim)
emp_pearsonstat  = rep(0, Nsim)
emp_lrtstat = rep(0, Nsim)

set.seed(111)
for(i in 1: Nsim){
  d_1976 = rbinom(1, 318, mle)
  d_1995 = rbinom(1, 315, mle)
  obser[i, ] =  c(d_1976, n_1976 - d_1976, d_1995, n_1995 - d_1995)
  MLE[i] = (d_1976 + d_1995) / (n_1976 + n_1995)
  expec[i, ] = c(n_1976*MLE[i], n_1976*(1-MLE[i]), n_1995*MLE[i], n_1995*(1-MLE[i]))

  pearson_extreme[i] = statPearson(obser[i, ], expec[i, ]) > pearson_obs
  lrt_extreme[i] = statLRT(obser[i, ], expec[i, ]) > lrt_obs

  emp_pearsonstat[i] = statPearson(obser[i, ], expec[i, ])
  emp_lrtstat = statLRT(obser[i, ], expec[i, ])

  empirical_lrt_pvalue = sum(pearson_extreme)/ Nsim
  empirical_pearson_pvalue = sum(lrt_extreme) / Nsim
}

## plot empirical density of pearson chi-sq stat ##
hist(emp_pearsonstat, prob = T, main = "Empirical Pearson Chi-sq Statistics & Reference Chi-sq (1)")
x = rchisq(Nsim, df = 1)
curve(dchisq(x, df=1), col='red', add=TRUE )

## plot empirical density of likelihood  ratio test ##
hist(emp_lrtstat, prob = T, main = "Empirical Likelihood Ratio Chi-sq and Reference Chi-sq(1)")
x = rchisq(Nsim, df = 1)
curve(dchisq(x, df=1), col='red', add=TRUE )

empirical_lrt_pvalue
empirical_pearson_pvalue

#research question: H0: pi_1d = pi_2d +0.05 vs Ha: pi_1d > pi_2d +0.05
#use Asymptotic Normality
test = (pi_d1 - pi_d2 - 0.05) / sqrt(se_pi_d1^2 + se_pi_d2^2)
ptest = 1 - pnorm(test, lower.tail = T)
ptest
```