Data Analysis Report on Ebola in Zaire

Zhao Liu, Josh

February 8, 2017

Introduction

Ebola, previously known as Ebola hemorrhagic fever, is a rare and deadly disease caused by infection with one of the Ebola virus species. Ebola can cause disease in humans and nonhuman primates (monkeys, gorillas, and chimpanzees). Ebola viruses are found in several African countries. Ebola was first discovered in 1976 near the Ebola River in what is now the Democratic Republic of the Congo (then Zaire) . Because of this, the virus that caused Ebola is scientifically named Zaire Ebolavirus.

The first outbreak of Ebola in Zaire caught the medical community off guard given its fast spread and deadliness. During a period as short as 55 days (from September 1 to October 24, 1976), as many as 318 cases of acute viral hemorrhagic fever occurred in northern Zaire, of which 280 were killed as a result of the disease. Medical community and pharmaceutical industry have focuses considerable amount of resources on the research of Ebola ever since, in an attempt to curb fatality caused by Ebola by treating or curing it. In 1995, another major outbreak of Ebola occurred in Zaire, infecting nearly as many as patients as in the outbreak in 1976: 315 cases were reported to the authorities. This time, 254 people were killed.

In this report, I will analyze the data listed above, i.e. total number of cases reported in each of the two Ebola outbreaks, and the number of deaths that occurred in each outburst. Specifically, my research questions are: 1. Are death rate in the 1995 Ebola outbreak and the death rate in the 1976 outbreak independent? 2. If they are not independent, what is their relationship?

Data

First, I list all data available in this contingency table. As the table shows, total numbers of Ebola cases are 318 and 315, for the 1976 outbreak and the 1995 outbreak, respectively. Of those cases, 280 people died in 1976, and 254 people died in 1995.

```
##       Death Non-Death Total
## 1976    280        38   318
## 1995    254        61   315
## Total   534        99   633
```

Since the total numbers of cases in the two outbreaks are different (although very close), I also list death rate of each outbreak in the table that follows:

```
##         Death Rate Non-death Rate
## 1976        0.8805         0.1195
## 1995        0.8063         0.1937
## Overall     0.8436         0.1564
```

The table shows that the death rate in the 1976 outbreak is 88.05%, and the figure for the 1995 outbreak is

80.63%. The overall death rate for the two outbreaks is 84.36%.

The difference between the death rate of 1995 and 1976 is $0.8063 - 0.8805 = -0.0742$. The results tell us the death proportion in 1995 is 0.0742 less than that of 1995. From only the difference, We cannot infer more information.

The relative risk is evaluated as $0.8036 \div 0.8805 = 0.9127$. It tells us the death rate in 1995 is 91.27% of the death rate in 1976.

Odds ratio is evaulated as $(280 \times 61) \div (254 \times 38) = 1.7696$. It means for an Ebola patient in 1976, the odds to die is 1.7696 as much as a patient in 1995.

Model

Since there are only two categories in the response variable (death or non-death), I will assume independent binomial model to test the death rate for the two. There are alternative models that we can use, such as multinomial model and Poisson model. Given the setup of the data, we know that the total number of deaths, which happened in the past, is a set number. Therefore, it is not appropriate to assume Poisson model.

I will use the following labels to represent data that we already have:

Define year is 1 for 1976, and 2 for 1995. Define result is 1 for Death, and 2 for non-Death. For the cell in the contingency table, let $n_{ij}, i = 1, 2, j = 1, 2$ represent the number in the cell in Row i and Column j. Row totals are represented by $n_{i+}, i = 1, 2$, and columns totals are represented by $n_{+j}, j = 1, 2$. We use N to denote sample size, which is 633. $

Let $\pi_{j|i}, i = 1, 2, j = 1, 2$ represent the theoretical probability, given a subject is in the i-th row, of being in the j-th column. Since there are only two attributes in the response variable, it is clear that $\pi_{1|i} + \pi_{2|i} = 1, i = 1, 2$. Therefore, we only need to know the distribution of $\pi_{1|i}$ to know $\pi_{2|i}$. To further simplify the matter, π_1 is used to denote given a subject is an Ebola patient in 1976, the probabiliby of death; and π_2 is used to denote given a subject is an Ebola patient in 1995, the probabiliby of death. Therefore, in this model, the number in cell (1, 1) $\sim bin(318, \pi_1)$, and the number in cell (2, 1) $\sim bin(315, \pi_2)$.

Under the null hypothesis H_0, Ebola result is independent from which year an outbreak takes place. Therefore, given empirical data, the expected value of the number in cell (1, 1)

$$\hat{\mu_{11}} = E[n_{11}] = N\pi_{11} = N\pi_{1+}\pi_{+1} = N\frac{n_{1+}}{N}\frac{n_{+1}}{N} = \frac{n_{1+}n_{+1}}{N} = \frac{318 \times 534}{633} = 268.27$$

Similarly:

$\hat{\mu_{12}} = 49.73$

$\hat{\mu_{21}} = 265.73$

$\hat{\mu_{22}} = 49.27$

Correspondingly, $\hat{\pi_1}$ and $\hat{\pi_2}$ are the estimators of the two probability. Clearly, $\hat{\pi_1} = n_{11} \div n_{1+} = 280 \div 318 = 0.8805$, and $\hat{\pi_2} = n_{21} \div n_{2+} = 254 \div 315 = 0.8063$

Analysis

First, we find confidence intervals of the difference of proportions. In this report, confidence intervals are at 95% level unless otherwise specified.

The confidence interval for the difference of proportions is given by

$$(\pi_2 - \pi_1) \pm z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_{1+}} + \frac{\pi_2(1-\pi_2)}{n_{2+}}}$$

I replace all π with $\hat{\pi}$, and can calculate the confidence interval for the difference between the two proportions is

```
## [1] -0.13050374 -0.01780414
```

Second, we find the confidence interval of the relative risk. Relative risk is given by

$$r = \frac{\hat{\pi_2}}{\hat{\pi_1}}$$

.

The natural logarithm of r has confidence interval

$$ln(r) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{\pi_1}}{n_{11}} + \frac{1-\hat{\pi_2}}{n_{21}}}$$

First, we find the lower and upper bounds of the natural logarithm of r:

```
## [1] -0.15556493 -0.02038826
```

Then we exponentiate the lower and upper bounds, and find the confidence interval for relative risk r is

```
## [1] 0.8559315 0.9798182
```

Third, we want to find the confidence interval for the odds ratio. Odds ratio is defined as

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

estimated by

$$\hat{\theta} = \frac{n_{11} \times n_{22}}{n_{21} \times n_{12}}$$

It is said that the natural logarithm of θ, $ln(\theta)$ has the confidence interval of

$$ln(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

So we calculate the confidence interval for $ln(\theta)$, which is

```
## [1] -0.5271908  0.3512376
```

Again, we exponentiate the lower and upper bounds, and get the 95% confidence interval for odds ratio, which is 0.5902608, 1.4208249.

As the last step of analysis, I will use Pearson's Chi-Squared and likelihood ratio Chi-Squred test to test independence between π_1 and π_2 .

Pearson's Chi-Squared test

According to Pearson's Chi-Squared test, the test statistic

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu_{ij}})^2}{\hat{\mu_{ij}}}$$

is said to asymptotically follow $\chi^2(I-1)(J-1)$ distribution. Therefore, in my analysis, I calculate the test statistic, and compare its p-value.

```
## [1] 0.01022494
```

P-value is lower than 0.05. Therefore, we reject our hypothesis that the death rate and year of occurrence are independent.

Likelihood Ratio Chi-Squared test

In likelihood ratio Chi-Squared test, it is said the test statistic,

$$G^2 = 2\sum_i \sum_j n_{ij} \ln \frac{n_{ij}}{\hat{\mu_{ij}}}$$

is also asymptotically follows $\chi^2(I-1)(J-1)$ distribution.

```
## [1] 6.644771
```

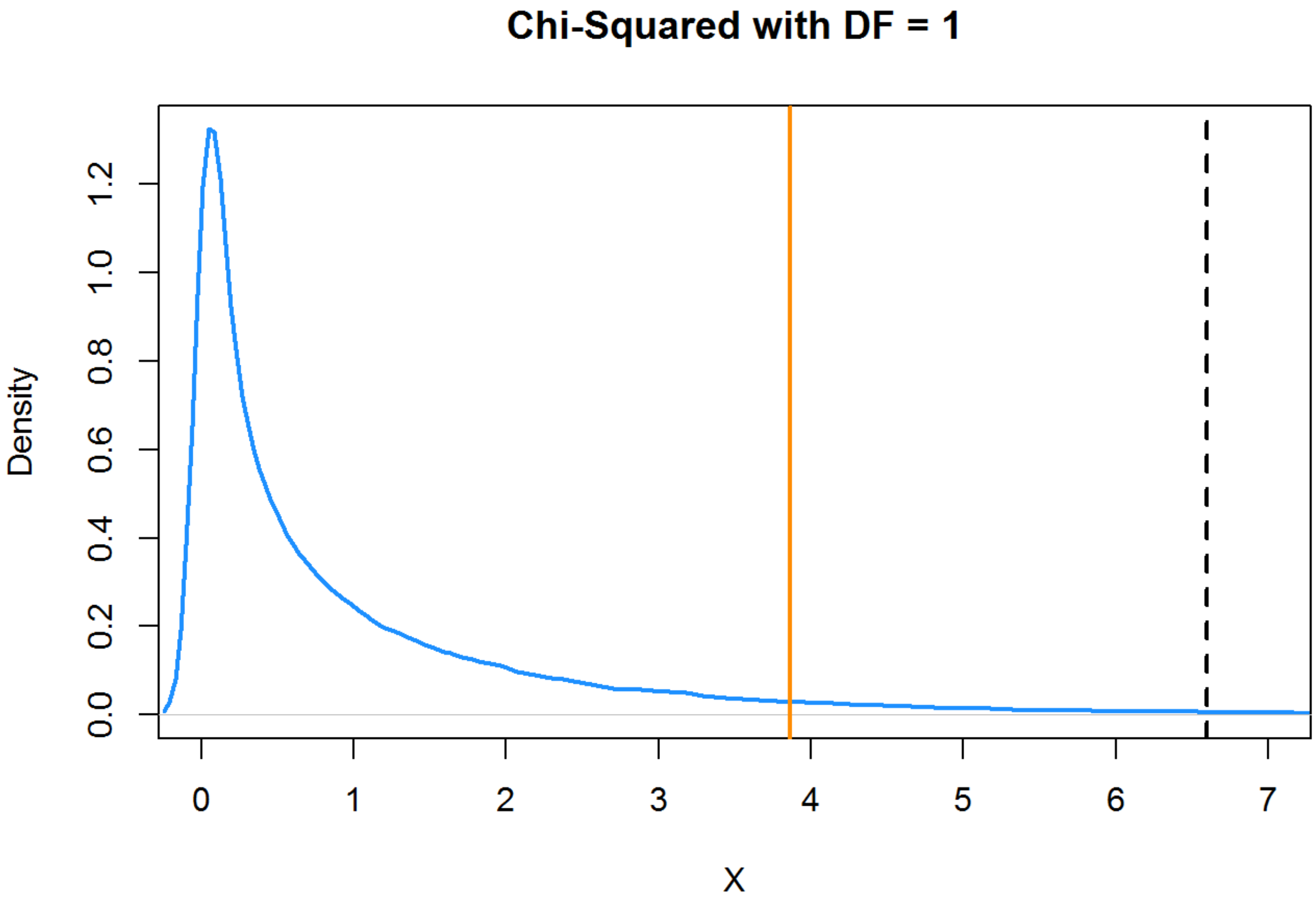
```
## [1] 0.009944725
```

This small p-value also indicates we should reject H_0.

Simulation

Simulation of Pearson's Chi-Squared Test

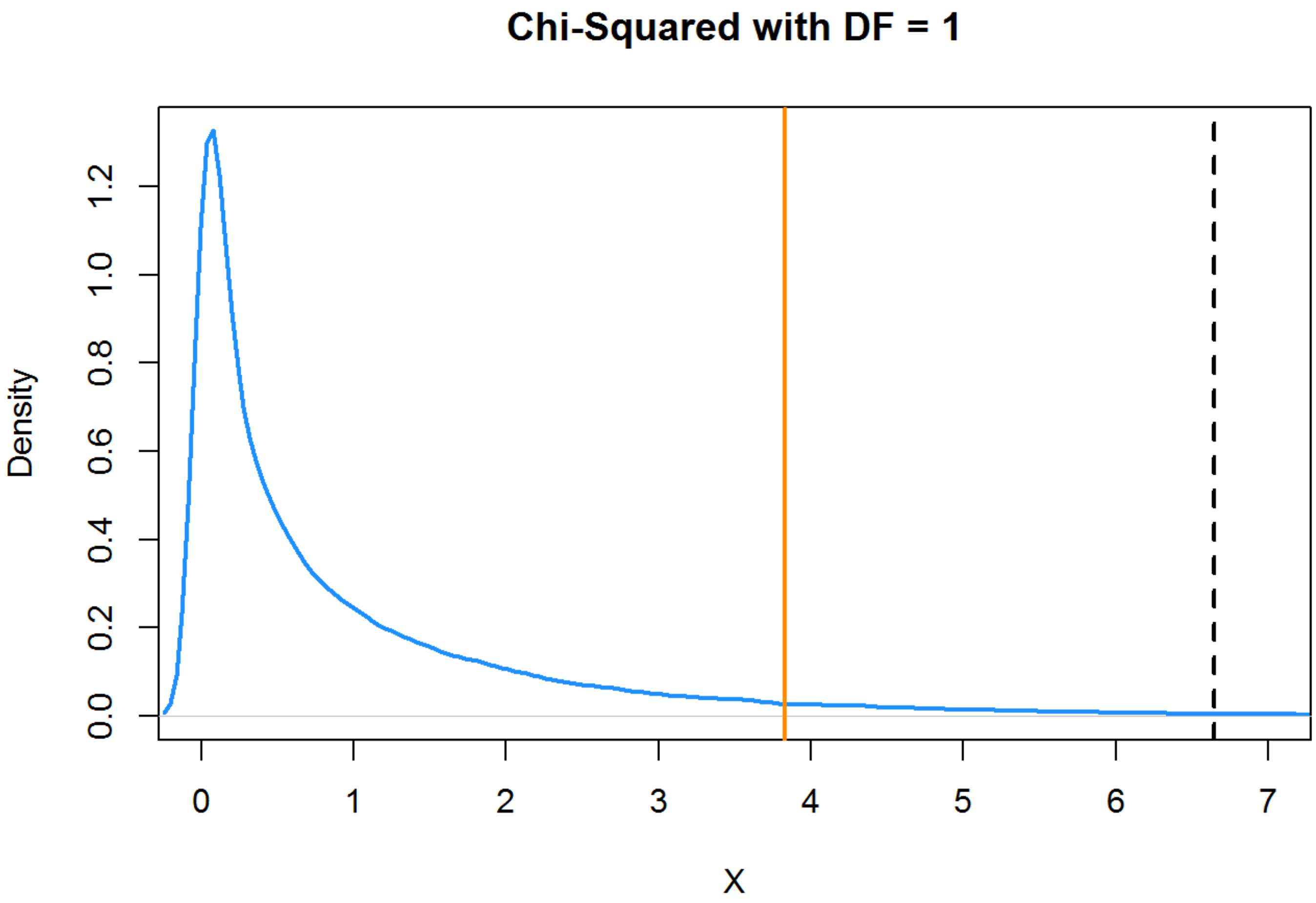
In the following simulation, I sampled 100,000 data points from χ^2 distribution with degree of freedom being 1. Then, I mark the cut-off point of the 0.05 critical region with an orange dashed line. The previously obtained X^2 test statistic is marked with a dotted line. It can be seen that the test statistic is way beyond the cut-off point.



Simulation of Likelihood Ratio Chi-Squared Test

In the following simulation, I sampled 100,000 data points from χ^2 distribution with degree of freedom being 1. Then, I mark the cut-off point of the 0.05 critical region with an orange dashed line. The previously obtained G^2 test statistic is marked with a dotted line. It can be seen that the test statistic is way beyond the cut-off point.

```
##      95%
## 3.829691
```


Chi-Squared with DF = 1
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
5
6
7
X