STAT 426 — Sections 1GR, 1UG — Spring 2017

Data Report Exam 1 Sample Solution

1 Introduction

Ebola virus[1] (scientifically designated EBOV) causes a contagious and usually fatal disease in humans. Since 1976, it has been responsible for several disease outbreaks, mainly in Africa, most recently the West African Ebola virus epidemic (2013-2016).

In 1976, the Republic of Zaire (now called Democratic Republic of the Congo) experienced the first recorded outbreak[2] of EBOV in the region near Yambuku. It experienced its second such outbreak in 1995 in the region near Kikwit.

This report aims to answer the following:

- What was the human case fatality rate in each outbreak?
- Were the human fatality rates the same for both outbreaks? If so, what is the overall rate? If not, how did they differ?

2 Data

The following table summarizes the human fatality case data for the two outbreaks:

	Fatal	Nonfatal	Total
1976 outbreak	280	38	318
1995 outbreak	254	61	315
Total	534	99	633

The apparent fatality rates were approximately 88% for the 1976 outbreak and 81% for the 1995 outbreak, a difference of about 7%. The estimated risk of a fatality in the 1976 outbreak was about 1.092 times that of the 1995 outbreak — approximately 9% higher. The estimated *odds ratio* of fatality in the 1976 outbreak to fatality in the 1995 outbreak was about 1.77.

3 Model

We use an independent binomial model for this analysis. Specifically,

$$\begin{aligned} Y_1 &= \text{fatalities in 1976 outbreak} \sim \text{binomial}(n_1 = 318, \pi_1) \\ Y_2 &= \text{fatalities in 1995 outbreak} \sim \text{binomial}(n_2 = 315, \pi_2) \end{aligned}$$

where Y_1 and Y_2 are independent. The population fatality rates for the two outbreaks are π_1 and π_2.

Unlike the independent Poisson model (or the multinomial model), this model cannot make inference about the mean total numbers of cases in the two outbreaks. However, research interest centers on the fatality rates, which can be analyzed with the binomial model.

[1] Ebola virus. (2017, January 27). In Wikipedia, The Free Encyclopedia. Retrieved 01:50, February 3, 2017, from `https://en.wikipedia.org/w/index.php?title=Ebola_virus&oldid=762264957`

[2] List of Ebola outbreaks. (2017, January 25). In Wikipedia, The Free Encyclopedia. Retrieved 23:06, February 3, 2017, from `https://en.wikipedia.org/w/index.php?title=List_of_Ebola_outbreaks&oldid=761969353`

The independent binomial model reasonably assumes independence of fatalities between the two outbreaks (conditional on their fatality rates). However, it also assumes that, for each outbreak, the number of fatalities is binomial, as if deaths occurred independently of each other and with the same probability. Such an assumption could be called into question if victims were related, or if they were given different levels of medical care.

4 Analysis

Wald approximate 95% confidence intervals for the case fatality rates of the outbreaks are

$$\text{1976 outbreak:} \quad (0.845, 0.916) \qquad \text{1995 outbreak:} \quad (0.763, 0.850)$$

A transformed Wald approximate 95% confidence interval for the relative risk (1976 versus 1995) is

$$(1.021, 1.168)$$

This interval contains only values exceeding 1, indicating a greater fatality rate for the 1976 outbreak (between 2% and 17% more deadly).

A transformed Wald approximate 95% confidence interval for the odds ratio (1976 versus 1995) is

$$(1.141, 2.745)$$

This interval contains only values exceeding 1, again indicating a greater fatality rate for the 1976 outbreak.

More reliable tests for homogeneity (equal fatality rates) are the Pearson and likelihood ratio chi-squared tests. The Pearson chi-squared test gives

$$X^2 \approx 6.595 \qquad \text{df} = 1 \qquad P \approx 0.01022$$

indicating moderate evidence for inhomogeneity — the fatality rates of the outbreaks differ.

The likelihood ratio chi-squared test gives

$$G^2 \approx 6.645 \qquad \text{df} = 1 \qquad P \approx 0.00994$$

yielding the same conclusion as the Pearson test.

5 Simulation

The parametric bootstrap can be used instead of the chi-squared approximation for the Pearson and likelihood ratio tests of the previous section. The same test statistics (X^2 and G^2) will be used, but the reference distributions will be their empirical distributions simulated under the null hypothesis of homogeneity ($\pi_1 = \pi_2$).

R code is listed in the Appendix. A total of 100000 random samples were used. The approximate bootstrap P-values for the two test statistics are

$$X^2 \text{ statistic:} \quad 0.010 \qquad G^2 \text{ statistic:} \quad 0.010$$

These support the same conclusion as the chi-squared tests of the previous section: There is evidence against homogeneity. The X^2 and G^2 P-values are about the same as before.

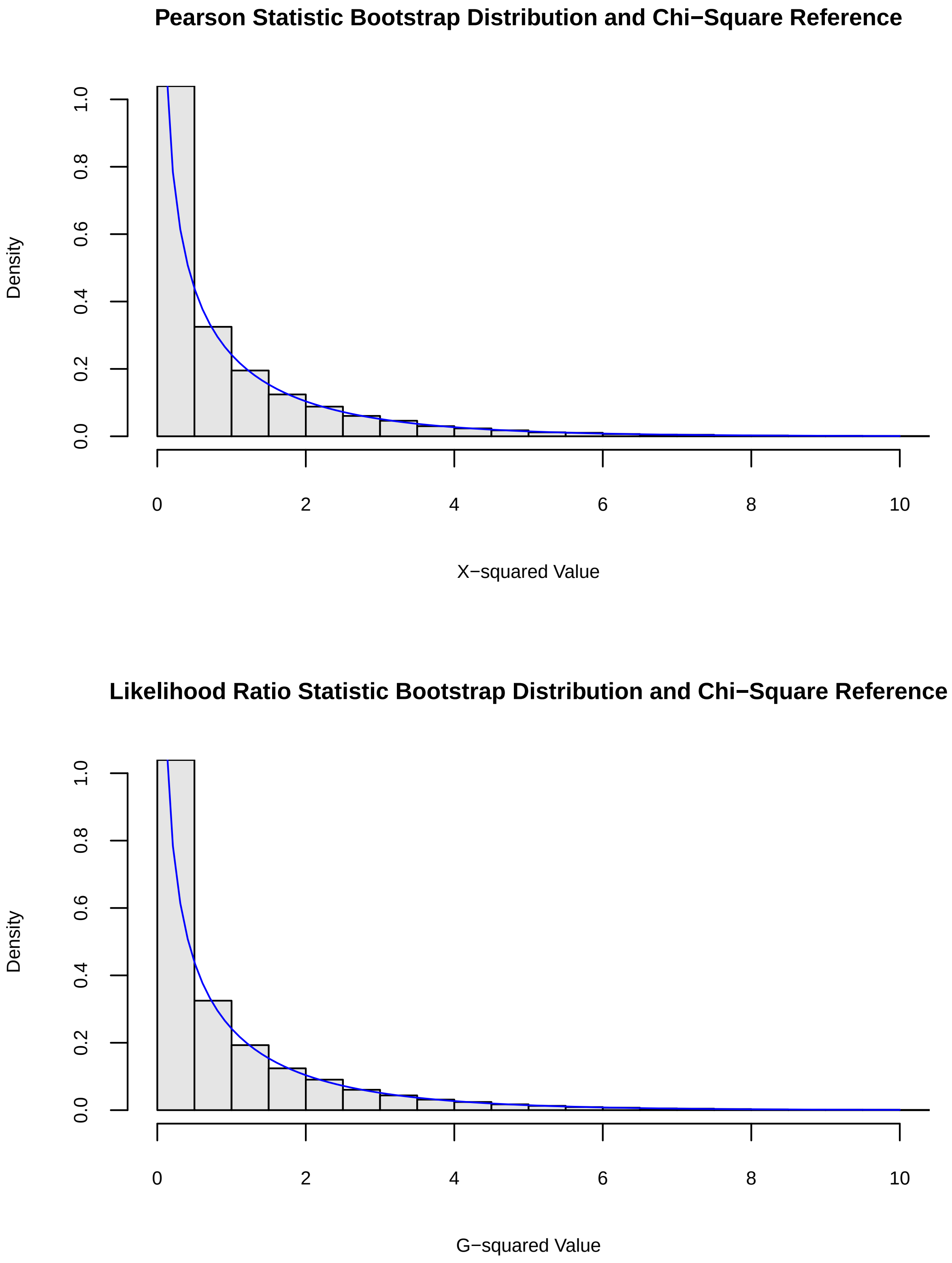
Histograms of the empirical bootstrap distributions appear in Figure 1, along with the reference chi-squared density with one degree of freedom. The bootstrap distributions of X^2 and G^2 both appear very similar to the reference chi-squared distribution, explaining why the bootstrap P-values are so similar.

6 Conclusions

There is evidence that Zaire's 1976 Ebola outbreak had a higher human case fatality rate (estimated to be between 85% and 92%) compared to the 1995 outbreak (estimated to be between 76% and 85%).

7 Appendix

```
### Fatality rate Wald 95% CIs:

n1 <- 318

n2 <- 315

y1 <- 280

y2 <- 254

pihat1 <- y1 / n1

pihat2 <- y2 / n2

pihat1 + c(-1,1) * qnorm(1-0.05/2) * sqrt(pihat1*(1-pihat1)/n1)

pihat2 + c(-1,1) * qnorm(1-0.05/2) * sqrt(pihat2*(1-pihat2)/n2)


### Relative Risk (Transformed) Wald 95% CI:

logr.CI <- log(pihat1/pihat2) + c(-1,1) * qnorm(1-0.05/2) *
             sqrt((1-pihat1) / y1 + (1-pihat2) / y2)

exp(logr.CI)


### Odds Ratio (Transformed) Wald 95% CI:

ORhat <- (pihat1 / (1-pihat1)) / (pihat2 / (1-pihat2))

logOR.CI <- log(ORhat) + c(-1,1) * qnorm(1-0.05/2) *
```



Pearson Statistic Bootstrap Distribution and Chi−Square Reference

Likelihood Ratio Statistic Bootstrap Distribution and Chi−Square Reference

Figure 1: Bootstrap empirical null distributions for the Pearson and Likelihood Ratio Chi-Squared statistics testing homogeneity, along with the reference χ^2_1 density.

```
              sqrt(1/y1 + 1/y2 + 1/(n1-y1) + 1/(n2-y2))

exp(logOR.CI)


### Pearson Chi-squared Test:

ebola.table <- rbind(c(y1,n1-y1), c(y2,n2-y2))

chisq.test(ebola.table, correct=FALSE)


### Likelihood Ratio Chi-squared Test:

muhat <- outer(margin.table(ebola.table,1), margin.table(ebola.table,2)) /
           sum(ebola.table)

G.sq <- 2 * sum(ebola.table * log(ebola.table / muhat))

G.sq

1 - pchisq(G.sq, df=1)
```

```
### Parametric Bootstrap Tests (Pearson and Likelihood Ratio)

obs.table <- rbind(c(280, 38),
                   c(254, 61))

pihat <- sum(obs.table[,1])/sum(obs.table)  # estimated pi assuming homogeneity

ns <- apply(obs.table, 1, sum)

X.sq.obs <- chisq.test(obs.table, correct=FALSE)$statistic

muhat <- outer(ns, c(pihat,1-pihat))

G.sq.obs <- 2 * sum(obs.table * log(obs.table / muhat))

Nsim <- 100000

X.sq.sim <- numeric(Nsim)

G.sq.sim <- numeric(Nsim)

for(i in 1:Nsim){
  y1sim <- rbinom(1,ns[1],pihat)
  y2sim <- rbinom(1,ns[2],pihat)
  sim.table <- rbind(c(y1sim, ns[1]-y1sim),
                     c(y2sim, ns[2]-y2sim))

  X.sq.sim[i] <- chisq.test(sim.table, correct=FALSE)$statistic

  ns.sim <- apply(sim.table, 1, sum)

  pihat.sim <- sum(sim.table[,1]) / sum(sim.table)

  muhat.sim <- outer(ns.sim, c(pihat.sim,1-pihat.sim))

  G.sq.sim[i] <- 2 * sum(sim.table * log(sim.table / muhat.sim))
}

mean(X.sq.sim >= X.sq.obs)  # X^2 bootstrap p-value

mean(G.sq.sim >= G.sq.obs)  # G^2 bootstrap p-value

pdf("empiricalplots.pdf",width=6,height=8)
par(mfrow=c(2,1), cex.axis=0.7, cex.lab=0.7, cex.main=0.8)

hist(X.sq.sim, breaks=50, freq=FALSE, col="grey90", xlab="X-squared Value",
     main="Pearson Statistic Bootstrap Distribution and Chi-Square Reference",
     xlim=c(0,10), ylim=c(0,1))
curve(dchisq(x, df=1), from=0.01, add=TRUE, col="blue")

hist(G.sq.sim, breaks=50, freq=FALSE, col="grey90", xlab="G-squared Value",
     main=paste("Likelihood Ratio Statistic Bootstrap Distribution and",
                "Chi-Square Reference"), xlim=c(0,10), ylim=c(0,1))
curve(dchisq(x, df=1), from=0.01, add=TRUE, col="blue")

dev.off()
```