Stat 744 Assignment 3

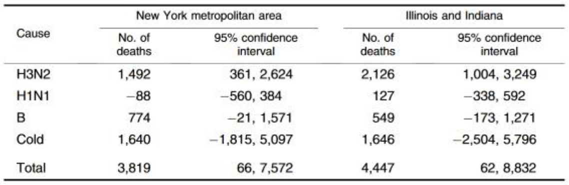
Yusang Hu 400177333

Sep. 23, 2019

```
knitr::opts_chunk$set(echo = TRUE,fig.width=15,fig.height=12)
```

Introduction

The following table is sourced from Dushoff et al. (2005). This table includes 4 variables: cause (H3N2, H1N1, B, Cold, Total), area (New York metropolitan area, Illinois and Indiana), number of deaths, 95% confidence interval. I'll split this table into two graphs. The first graph will include number of deaths and 95% confidence interval for H3N2, H1N1, B and cold in both two areas. The second graph will include number of death and 95% confidence interval for only Total in both two areas.



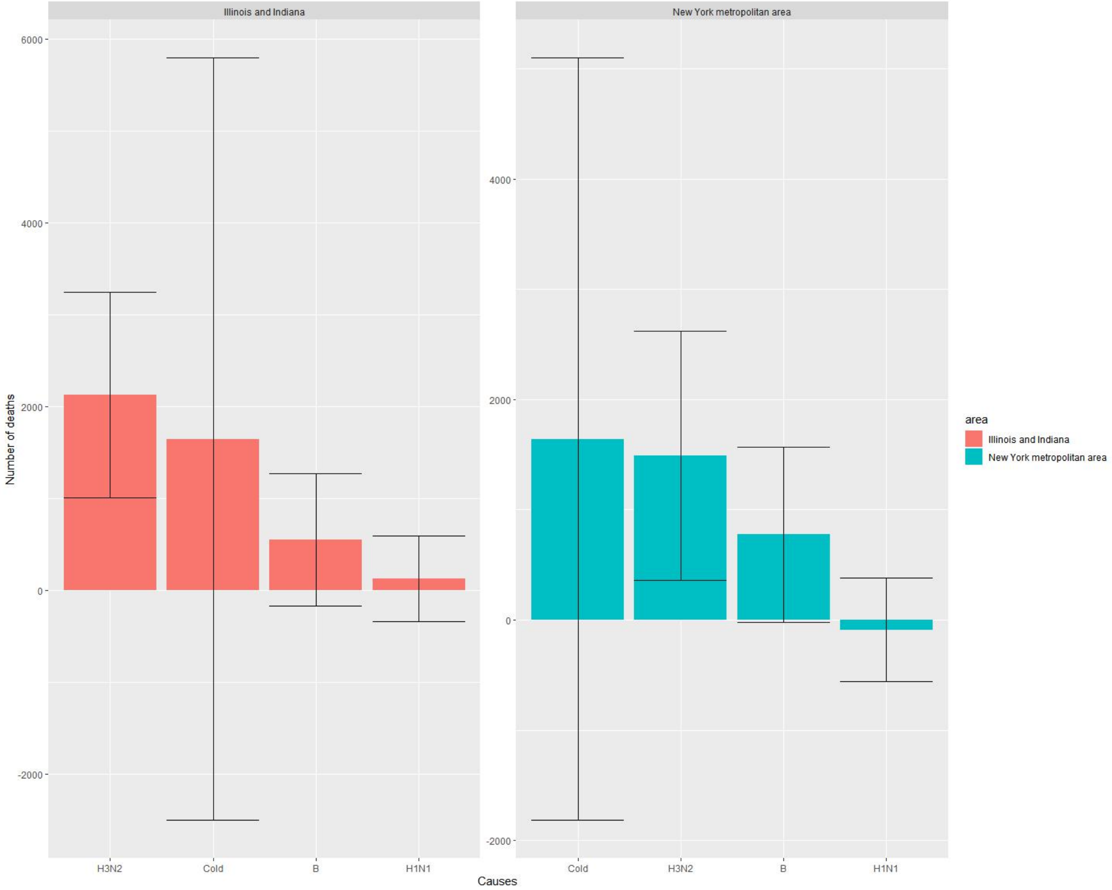
Cause	New York metropolitan area		Illinois and Indiana	
	No. of deaths	95% confidence interval	No. of deaths	95% confidence interval
H3N2	1,492	361, 2,624	2,126	1,004, 3,249
H1N1	−88	−560, 384	127	−338, 592
B	774	−21, 1,571	549	−173, 1,271
Cold	1,640	−1,815, 5,097	1,646	−2,504, 5,796
Total	3,819	66, 7,572	4,447	62, 8,832

First Plot

Excess deaths in winter may be caused by couple of reasons including influenza (i.e. H3N2, H1N1, virus B) and cold. In my first plot, I'm going to study which one is the most important cause of excess deaths in winter in the New York metropolitan area and the Illinois and Indiana separately. I'll focus on number of death, causes (i.e. H3N2, H1N1, virus B, cold) and 95% confidence interval in this plot. First I'll use geom_bar() to make a grouped bar plot for two areas. Then I'll use facet_wrap to separate the grouped bar plot into two individual plots, one plot corresponds to one area. Also I'll set scales="free" to make each facet looks better. Next, I'll reorder x-axis from highest number of deaths to lowest number of deaths for each plot. Finally, I'll use geom_errorbar() to draw 95% confidence interval for each cause.

```
disease=data.frame(cause=c("H3N2","H3N2","H1N1","H1N1","B","B","Cold","
Cold"),area=c("New York metropolitan area","Illinois and Indiana","New
York metropolitan area","Illinois and Indiana","New York metropolitan ar
ea","Illinois and Indiana","New York metropolitan area","Illinois and In
diana"),
                   numofdeaths=c(1492,2126,-88,127,774,549,1640,1646),L=c(3
61,1004,-560,-338,-21,-173,-1815,-2504),
                   U=c(2624,3249,384,592,1571,1271,5097,5796))
library(ggplot2)
```

```
(g1<-ggplot(disease,aes(factor(-numofdeaths),numofdeaths,fill=area))
  +geom_bar(stat="identity",position="dodge")
  +geom_errorbar(aes(ymin=L,ymax=U),position="dodge")
  +facet_wrap(~area,scales="free")
  +labs(x="Causes",y="Number of deaths")
  +scale_x_discrete(labels=c("-2126"="H3N2","-1646"="Cold","-549"="B","
-127"="H1N1","-1640"="Cold","-1492"="H3N2","-774"="B","88"="H1N1"))
)
```



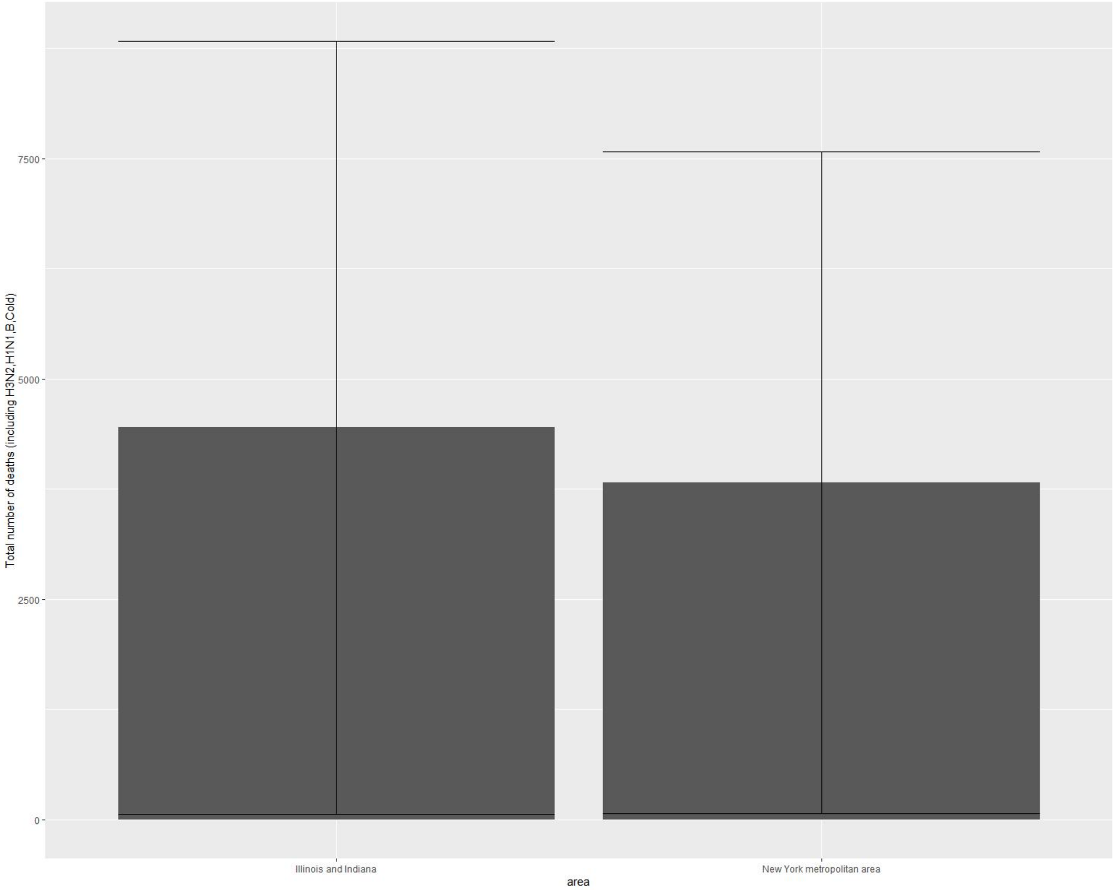
Conclusion for First Plot

In the Illinois and Indiana, we can observe that the highest number of death is “H3N1”. But confidence interval for “Cold”" is very wide which means the estimate of number of death caused by “Cold”" is unstable, therefore we cannot conclude that “H3N1”" is the most important cause of excess deaths in winter in the Illinois and Indiana. In the New York metropolitan area, we can observe that “Cold” has the highest number of deaths, but simultaneously it has the widest confidence interval, so there exists some uncertainty on the number of deaths caused by “Cold”. Therefore we cannot conclude that “Cold” is the most important cause of excess deaths in winter in the New York metropolitan area, we need further information about “Cold” to make a precise conclusion.

Second Plot

My second plot will only include information for “Total”. In my second plot, I’ll study whether the New York metropolitan area or the Illinois and Indiana has the highest number of deaths caused by influenza and cold. I’ll focus on total number of deaths and 95% confidence interval in two areas.

```
totaldisease=data.frame(totalnumofdeaths=c(3819,4447),area=c("New York metropolitan area","Illinois and Indiana"),totalL=c(66,62),totalU=c(7572,8832))
(g2<-ggplot(totaldisease,aes(area,totalnumofdeaths))
  +geom_bar(stat="identity")
  +geom_errorbar(aes(ymin=totalL,ymax=totalU))
  +labs(y="Total number of deaths (including H3N2,H1N1,B,Cold)")
)
```



Conclusion for Second Plot

From the second plot, we can observe that the Illinois and Indiana has the highest number of deaths, but confidence interval for both the New York metropolitan area and the Illinois and Indiana are very wide. Therefore we cannot conclude that the

Illinois and Indiana has higher number of annual pneumonia and influenza deaths than the New York metropolitan area, further information are needed.

Reference

Dushoff, J., Plotkin, J. B., Viboud, C., Earn, D. J. D., & Simonsen, L. (2005). Mortality due to Influenza in the United States—An Annualized Regression Approach Using Multiple-Cause Mortality Data. American Journal of Epidemiology, 163(2), 181–187. doi: 10.1093/aje/kwj024