G5069_Data Challenge 1

Team 4

18 February 2017

```
#  ##############################################################################
#       File-Name:      DataChallenge1_Team4.Rmd
#       Version:        R 3.3.2
#       Date:           Feb 18, 2017
#       Author:         MM
#       Purpose:        Verify lethality index figures with reported figures
#       Input Files:    ConfrontationsData_170209.csv (processed data on confrontations)
#       Output Files:   NONE
#       Data Output:    NONE
#       Previous files: NONE
#       Dependencies:   NONE
#       Required by:    NONE
#       Status:         IN PROGRESS
#       Machine:        Mac laptop
#  ##############################################################################

# Load Libraries and Data
rm(list=ls(all=TRUE))

library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(ggplot2)

path <- "~/Documents/Columbia/5069_Applied Data Science/Data Challenge 1/ConfrontationsData_170209.csv"
dataset <- read.csv(path)
```

1. Can you replicate the 86.1% number? the overall lethality ratio? the ratios for the Federal Police, Navy and Army?

Perfect Lethality Events

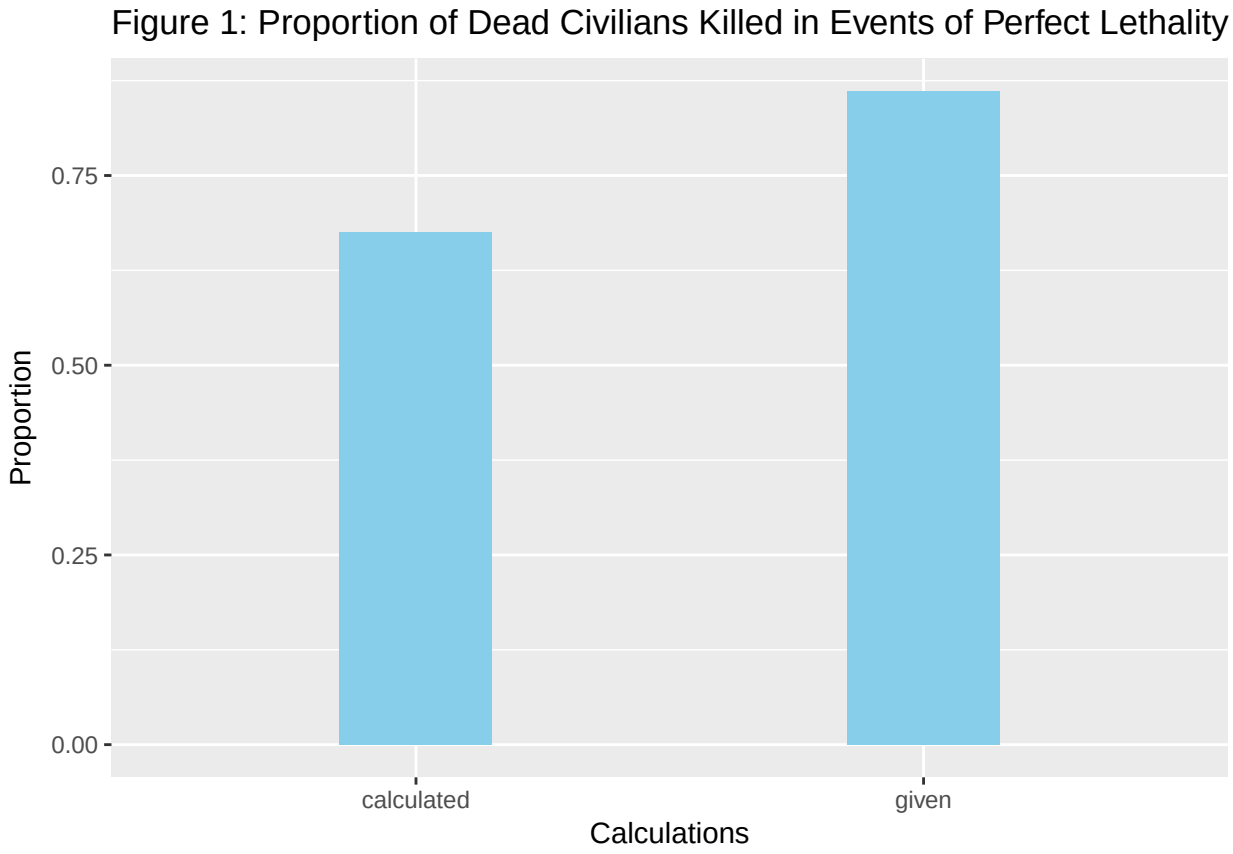
```
pl.count <- dataset %>%
  # filter events with 'perfect lethality' i.e. only civilian deaths and no civilian wounded
  filter(civilian.dead != 0) %>%
  filter(civilian.wounded == 0) %>%
  summarize(count = sum(civilian.dead))
dataset.count <- dataset %>%
  summarize(count = sum(civilian.dead))
pl.count/dataset.count
```

```
##       count
## 1 0.6755319
```

```
# Plot Comparison of Proportions
pl.proportion <- data.frame(result = c("calculated", "given"), proportion = c(0.676, 0.861))
ggplot(pl.proportion, aes(result, proportion)) +
```

```
  geom_col(width = 0.3, fill = "skyblue") +
  labs(x = "Calculations", y = "Proportion") +
  ggtitle("Figure 1: Proportion of Dead Civilians Killed in Events of Perfect Lethality") +
  ggsave(file = "plot1.png")
```

```
## Saving 6.5 x 4.5 in image
```



We achieved a 67.6% of civilians killed in pure lethality events, which is smaller than the 86.1% reported. A possible reason is that the 86.1% is calculated based on events involving federal armed forces, which in this case, is difficult to classify (for reasons we will discuss next).

Lethality Index

```
# Add additional variables that help identify agency presence
dataset1 <- dataset %>%
  mutate(navy.casualty = navy.dead + navy.wounded,
         military.casualty = military.dead + military.wounded,
         federal.police.casualty = federal.police.dead + federal.police.wounded,
         nonfed.wounded = civilian.wounded + organized.crime.wounded,
         nonfed.dead = civilian.dead + organized.crime.dead)

# Check for mutual exclusivity among agency enforcement in each event
dataset2 <- dataset1 %>%
  filter(military.casualty != 0) %>%
  summarize(navy = sum(navy.casualty),
            military = sum(military.casualty),
            federal.police = sum(federal.police.casualty))
```

```
dataset2
```

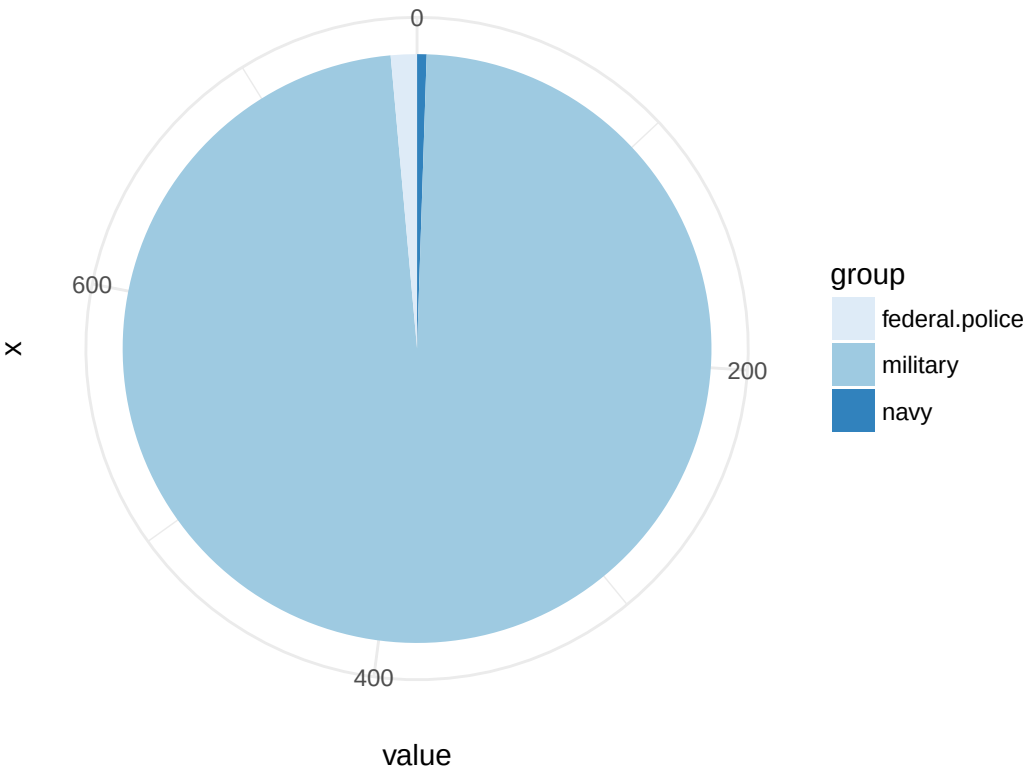
```
##   navy military federal.police
## 1    4      753             11
```

```
# Data Wrangling & Plotting of Graph
dataset3 <- data.frame(
  group = c("navy", "military", "federal.police"),
  value = c(4, 753, 11)
  )

ggplot(dataset3, aes(x = "", y = value, fill = group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  scale_fill_brewer(palette ="Blues") +
  theme_minimal() +
  ggtitle("Figure 2: Pie Chart of Composition of Casuality") +
  ggsave(file = "plot2.png")
```

```
## Saving 6.5 x 4.5 in image
```

Figure 2: Pie Chart of Composition of Casuality



Events which had casualties from the Navy also saw some casualties from the Army. As such, the enforcement agency for each event is not mutually-exclusive. For example, looking at events where military personnel were killed/wounded, we see that there were also 4 navy casualties and 11 federal police casualties. Given that there is no variable indicating which federal enforcement agency participated in each event, we are unable to accurately calculate the lethality index sorted by the Army, Navy and Federal Police.

Nonetheless, we sought to achieve an approximate number for each federal enforcement agency by classifying events based on their respective casualty numbers. For example, if there were navy casualties from an event,

it is likely that the navy participated in the confrontation.

However, again in cases of joint enforcement, our figures may over/underattribute the number of deaths/wounded to a particular agency, when the perpetrator was otherwise. This is the first source of error. This method also excludes confrontations where the agency may have participated in, but had zero casualties. In these cases with zero casualties, the enforcement agency could potentially be more effective in their killing - as such, our calculated ratios are likely to underreport the actual lethality index. This is the second source of error.

```
# Calculating Lethality Indices for Criminals, Civilians & Both (Non-Fed)
li.total <- dataset1 %>%
  summarize(lethality.nonfed = sum(nonfed.dead)/sum(nonfed.wounded),
            lethality.organized.crime = sum(organized.crime.dead)/sum(organized.crime.wounded),
            lethality.civilian = sum(civilian.dead)/sum(civilian.wounded))

li.federal.police <- dataset1 %>%
  filter(federal.police.casualty != 0) %>%
  summarize(lethality.nonfed = sum(nonfed.dead)/sum(nonfed.wounded),
            lethality.organized.crime = sum(organized.crime.dead)/sum(organized.crime.wounded),
            lethality.civilian = sum(civilian.dead)/sum(civilian.wounded))

li.navy <- dataset1 %>%
  filter(navy.casualty != 0) %>%
  summarize(lethality.nonfed = sum(nonfed.dead)/sum(nonfed.wounded),
            lethality.organized.crime = sum(organized.crime.dead)/sum(organized.crime.wounded),
            lethality.civilian = sum(civilian.dead)/sum(civilian.wounded))

li.military <- dataset1 %>%
  filter(military.casualty != 0) %>%
  summarize(lethality.nonfed = sum(nonfed.dead)/sum(nonfed.wounded),
            lethality.organized.crime = sum(organized.crime.dead)/sum(organized.crime.wounded),
            lethality.civilian = sum(civilian.dead)/sum(civilian.wounded))

# Data Wrangling & Plotting of Graph
li.summary <- rbind(li.total, li.federal.police, li.navy, li.military)
li.summary$given <- c(2.6, 2.6, 17.3, 9.1)
row.names(li.summary) <- c("total", "federal.police", "navy", "military")
li.summary$type <- row.names(li.summary)
li.summary
```

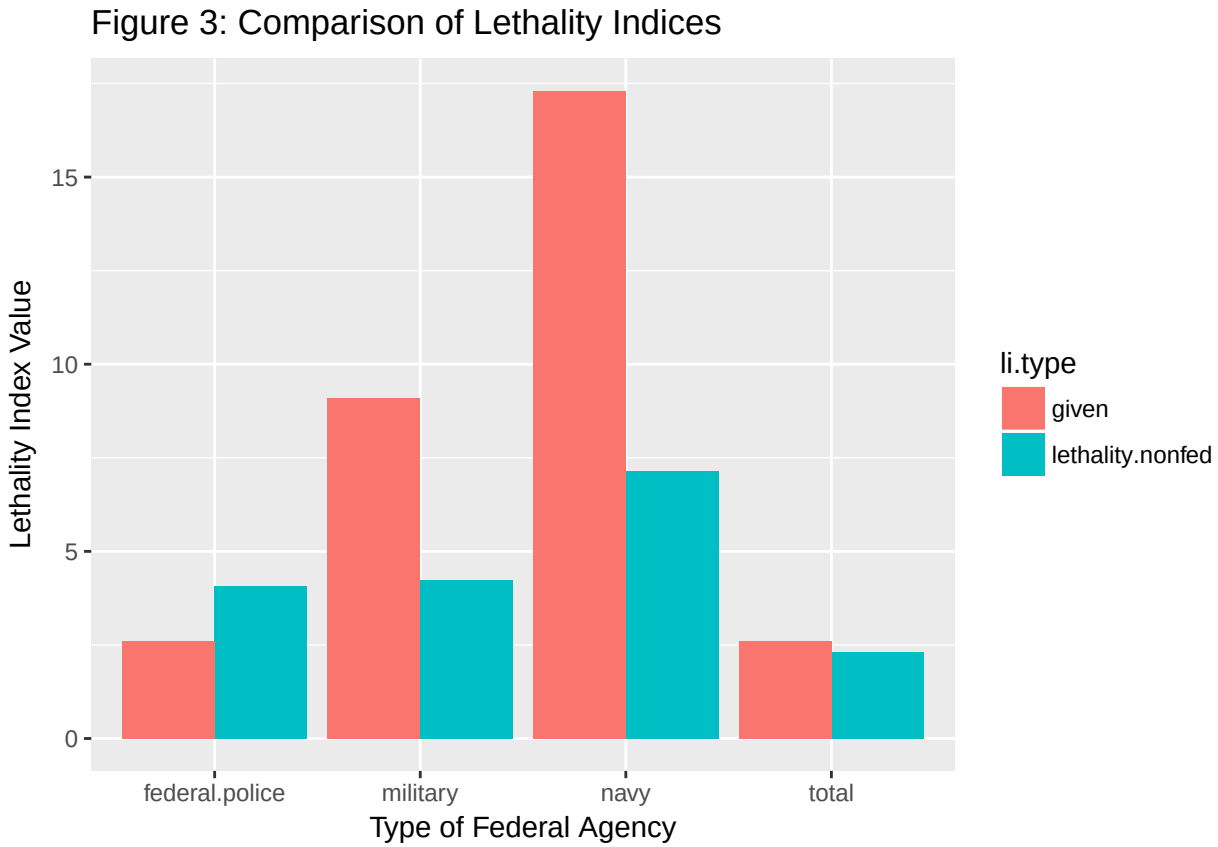
```
##                lethality.nonfed lethality.organized.crime
## total                  2.309972                  3.078751
## federal.police         4.078431                  6.032258
## navy                   7.142857                 11.111111
## military               4.246753                  6.494737
##                lethality.civilian given           type
## total                   0.5758040   2.6          total
## federal.police          1.0500000   2.6 federal.police
## navy                    0.0000000  17.3           navy
## military                0.6271186   9.1       military
```

```
li.summary <- li.summary %>%
  select(type, lethality.nonfed, given) %>%
  gather(lethality.nonfed, given, key = "li.type", value = "index.value")

ggplot(li.summary, aes(x = type, y = index.value)) +
```

```
  geom_bar(aes(fill = li.type), stat = "identity", position = "dodge") +
  labs(x = "Type of Federal Agency", y = "Lethality Index Value") +
  ggtitle("Figure 3: Comparison of Lethality Indices") +
  ggsave(file = "plot3.png")
```

```
## Saving 6.5 x 4.5 in image
```



We calculated our lethality index as the number of organized crime and civilians killed over the number of organized crime and civilians wounded. The difference between our ratios and the given ratios is shown in Figure 3. We see that the general pattern still holds - the navy has the highest lethality index among all agencies. However, the absolute difference could possibly be attributed to the two errors we have discussed above.

Our calculated lethality index for Mexico, 2.3, is relatively close to that of the given number 2.6.

2. Now you know the data more intimately. Think a little bit more about it, and answer the following questions:

- Is this the right metric to look at? Why or why not?

The "lethality index" is the appropriate metric to look at if we want to gauge the effectiveness of killing of the enforcement agencies. This is the concern raised by the New York Times article and the PPD. Their concerns hint at potential excessive brutality by the enforcement agencies - i.e. a high lethality index indicates that more criminals/civilians are unnecessarily killed (rather than being wounded/detained) during a confrontation.

- What is the "lethality index" showing explicitly? What is it not showing? What is the definition assuming?

The "lethality index" shows the number of organized crime and civilians killed over the number of organized crime and civilians wounded. It is not showing possible criminals or civilians who were wounded but may have escaped, and in this case the lethality index would overstate the actual ratio.

The "lethality index" also assumes that the enforcement agencies know who killed/wounded who - i.e. that the killing of an individual can be attributed properly to the army, navy or police. This is unlikely to be accurate if there are two or more enforcement agencies participating in the event.

Finally, there is also the possibility of fratricide, which would overstate the effectiveness of the federal agencies, if some of the killings were internal.

- With the same available data, can you think of an alternative way to capture the same construct? Is it "better"?

An alternative way is to look at a "sacrificial index", that is the number of deaths/wounded/both of criminals and civilians over the number of deaths/wounded/both of enforcement agencies. This gives us an idea of how many criminals and civilians were killed/wounded for every enforcement agency personnel killed/wounded. It can be calculated as follows:

```
dataset4 <- dataset %>%
  mutate(
    enforcement.dead = military.dead + navy.dead + federal.police.dead + afi.dead +
      state.police.dead + ministerial.police.dead + municipal.police.dead +
      public.prosecutor.dead,
    enforcement.wounded = military.wounded + navy.wounded + federal.police.wounded +
      afi.wounded + state.police.wounded + ministerial.police.wounded +
      municipal.police.wounded + public.prosecutor.wounded,
    nonfed.dead = civilian.dead + organized.crime.dead,
    nonfed.wounded = civilian.wounded + organized.crime.wounded
  )
si.ratios <- dataset4 %>%
  summarize(
    si.dead = sum(nonfed.dead)/sum(enforcement.dead),
    si.wounded = sum(nonfed.wounded)/sum(enforcement.wounded)
  )
si.ratios
```

```
##    si.dead si.wounded
## 1 9.372137    1.27611
```

From the above, we can conclude that 9.37 criminals and civilians died for every death on the side of the enforcement agencies. This also gives us some idea of the effectiveness of the enforcement agenices - more criminals and civilians are killed for every enforcement agency personnel killed.

- What additional information would you need to better understand the data?

Clearer information on how the number for deaths and wounded are recorded/calculated will be good, since measurement error is likely to be high especially in large confrontations e.g. capturing people who escape etc..

In addition, having an idea of the total number of people present at the confrontation will provide us with a better idea of the scale of the deaths/wounded. Is the enforcement agency only killing/wounding a small proportion of individuals at the confrontation? Alternatively, the number of unwounded people can be provided in order for us to calculate the total number of people present.

Finally, we are also ignorant of possible geographical characteristics of Mexico. For example, a confrontation that does not occur on water would probably not involve the navy. Geographical traits could potentially impact/limit the counts of casualty and severity of the confrontation.

- What additional information could help you better capture the construct behind the "lethality index"?

We would need a metric on whether a particular enforcement agency participated in the confrontation in order to attribute a killing/wounding to a specific agency, instead of the loose metric we used in Part 1. We understand that this data is actually present, but only made available in the new dataset for Week 6.

Finally, the "lethality index" derived from this dataset allows us to only count the total people present at the confrontation based on counts of casualty, rather than the actual number of people present i.e. including those who were not wounded. The 'detained' variable is not ideal, given that it is not mutually exclusive with the number of people wounded.