B1

Emma Terreni

2023-09-19

B1 Homework

Indiscriminate Violence and Insurgency

```
## # A tibble: 6 x 6
##   village      groznyy  fire deaths preattack postattack
##   <chr>          <dbl> <dbl>  <dbl>     <dbl>      <dbl>
## 1 Elistanzhi         0     0     NA         4          3
## 2 Malye Shuani       0     1      0         0          1
## 3 Belgatoy           0     1     34         1          0
## 4 Oktya'brskoe       0     0     NA         0          0
## 5 Chiri-Yurt         0     0     NA         4          5
## 6 Gansolchu          0     1      0         0          0
```

```
## [1] 318   6
```

```
## # A tibble: 318 x 6
##    village       groznyy  fire deaths preattack postattack
##    <chr>           <dbl> <dbl>  <dbl>     <dbl>      <dbl>
##  1 Elistanzhi          0     0     NA         4          3
##  2 Malye Shuani        0     1      0         0          1
##  3 Belgatoy            0     1     34         1          0
##  4 Oktya'brskoe        0     0     NA         0          0
##  5 Chiri-Yurt          0     0     NA         4          5
##  6 Gansolchu           0     1      0         0          0
##  7 Achkoi-Martan       0     0     NA         4          8
##  8 Khidi-Khutor        0     1      0         3          1
##  9 Agishbatoi          0     0     NA         2          0
## 10 Shirdi-Mokhk        0     1      0         1          0
## # i 308 more rows
```

Question 1

Wrong answer, but I gave you credit still. See code comment for more

---

```
    chechen %>%
        count(fire)
```

```
## # A tibble: 2 x 2
##    fire     n
##   <dbl> <int>
## 1     0   159
## 2     1   159
```

```
# Incorrect because each row isn't a unique village. It's an event where the Russian's fired, and the `
```

159 villages in the data were attacked by Russians and an equivalent 159 villages were not attacked by Russians.

Question 2

```
chechen %>%
  filter(fire == 1) %>%
  group_by(groznyy) %>%
  summarise(
            meandeaths = mean(deaths, na.rm = TRUE),
            mediandeaths = median(deaths, na.rm = TRUE),
            n = n()
            )
```

```
## # A tibble: 2 x 4
##   groznyy meandeaths mediandeaths     n
##     <dbl>      <dbl>        <dbl> <int>
## 1       0       1.57            0   152
## 2       1       3.71            3     7
```

When only looking at the villages where there were Russian attacks, it is apparent that the villages in Groznyy experienced a greater number of deaths compared to the villages outside of Groznyy, bot when looking at the mean (a difference of 2.2 deaths on average) and the median (a difference of 3 deaths). It is important to note, that there were only 7 villages in Groznyy exposed to Russian attacks compared to the 152 outside of Groznyy.

Comment: Good!

Question 3

```
chechen %>%
  group_by(fire) %>%
  summarise(
            meanpostattacks = mean(postattack, na.rm = TRUE),
            postattack_q25 = quantile(x = postattack, probs = 0.25, na.rm = T),
            postattack_q50 = median(postattack, na.rm = TRUE),
            postattack_q75 = quantile(x = postattack, probs = 0.75, na.rm = T),
            n = n()
            )
```

```
## # A tibble: 2 x 6
##    fire meanpostattacks postattack_q25 postattack_q50 postattack_q75     n
##   <dbl>           <dbl>          <dbl>          <dbl>          <dbl> <int>
## 1     0            2.05              0              0              2   159
## 2     1            1.50              0              0              1   159
```

The average number of insurgent attacks after Russian fire for villages hit by artillery fire is 1.5 while the corresponding number is 2.05 for the villages not hit.

The first and second quartile are identical for villages hit and villages not hit, while the 3rd quartile is one death higher for the villages that were not hit by Russian fire. This suggests that the villages most severely affected, that were not hit, have higher numbers of deaths than the most severely hit villages, that were not affected by Russian artillery fire. That might also be what is reflected in the difference in means between the two groups.

This would suggest that the fires dampened the insurgent attacks, but since we have not compared to the number of attacks before fire, we cannot say anything causal about the descriptive statistics. It might just

be, that the villages not hit by Russian fire experience more attacks on average than the villages hit. The table above says nothing about the change in insurgent attacks. I would not conclude that indiscriminate violence reduces insurgent attacks yet.

Comment: Good! Though, quartiles usually includes the min and max (full points though). The main point remains the same, most villages see no insurgent attacks, regardless of whether they were shelled. The difference in means is driven by small differences in the 75th and 100th quantiles

Question 4

```
chechen %>%
  group_by(fire) %>%
  summarise(
            meanpreattacks = mean(preattack, na.rm = TRUE),
            preattack_q25 = quantile(x = preattack, probs = 0.25, na.rm = T),
            preattack_q50 = median(preattack, na.rm = TRUE),
            preattack_q75 = quantile(x = preattack, probs = 0.75, na.rm = T),
            n = n()
            )
```

```
## # A tibble: 2 x 6
##    fire meanpreattacks preattack_q25 preattack_q50 preattack_q75     n
##   <dbl>          <dbl>         <dbl>         <dbl>         <dbl> <int>
## 1     0           2.15             0             0             3   159
## 2     1           2.11             0             0             2   159
```

The average number of attacks before Russian fire was almost identical for villages hit by Russian fire and for villages not hit. This suggests that, as far as frequency of insurgent attacks, the two groups were similar before Russian artillery fire. Lyall (2009) states that the Russian artillery strikes are uncorrelated with key spatial and demographic characteristics, making it a sound instrument. If we believe that the Russian artillery fires were carried out randomly and that the villages experiencing Russian fire are comparable to the villages that didn't, we are able to draw causal conclusions. The drop in insurgent attacks in the villages affected by Russian artillery fire is, on average, notably larger than the drop in villages not affected.

Comment: Good!

Question 5

Among the villages shelled by Russians, the number of insurgent attacks did not increase after the artillery attack. In fact, the number of insurgent attacks decreased on average. When comparing to the average decrease in attacks in non-shelled villages, the decrease in the shelled villages is substantial. When looking at the distribution of the differences in attacks, we see that the difference in means is mainly driven by a few villages experiencing a vast drop in attacks after Russian fire, as well as a generally more skewed distribution compared to the villages not shelled.

Comment: Good! Another way to look at this would have been to calculate what percent of villages experienced an increase in attacks (this is a minority ~13%, which explains why the average difference in attacks is negative)

```
chechen %>%
  mutate(diffattacks = postattack - preattack) %>%
  group_by(fire) %>%
  summarise(meandiffattacks = mean(diffattacks, na.rm = TRUE),
            mediandiffattacks = median(diffattacks, na.rm = TRUE)
            )
```

```
## # A tibble: 2 x 3
##    fire meandiffattacks mediandiffattacks
```

```
##   <dbl>           <dbl>             <dbl>
## 1     0          -0.101                 0
## 2     1          -0.616                 0
```

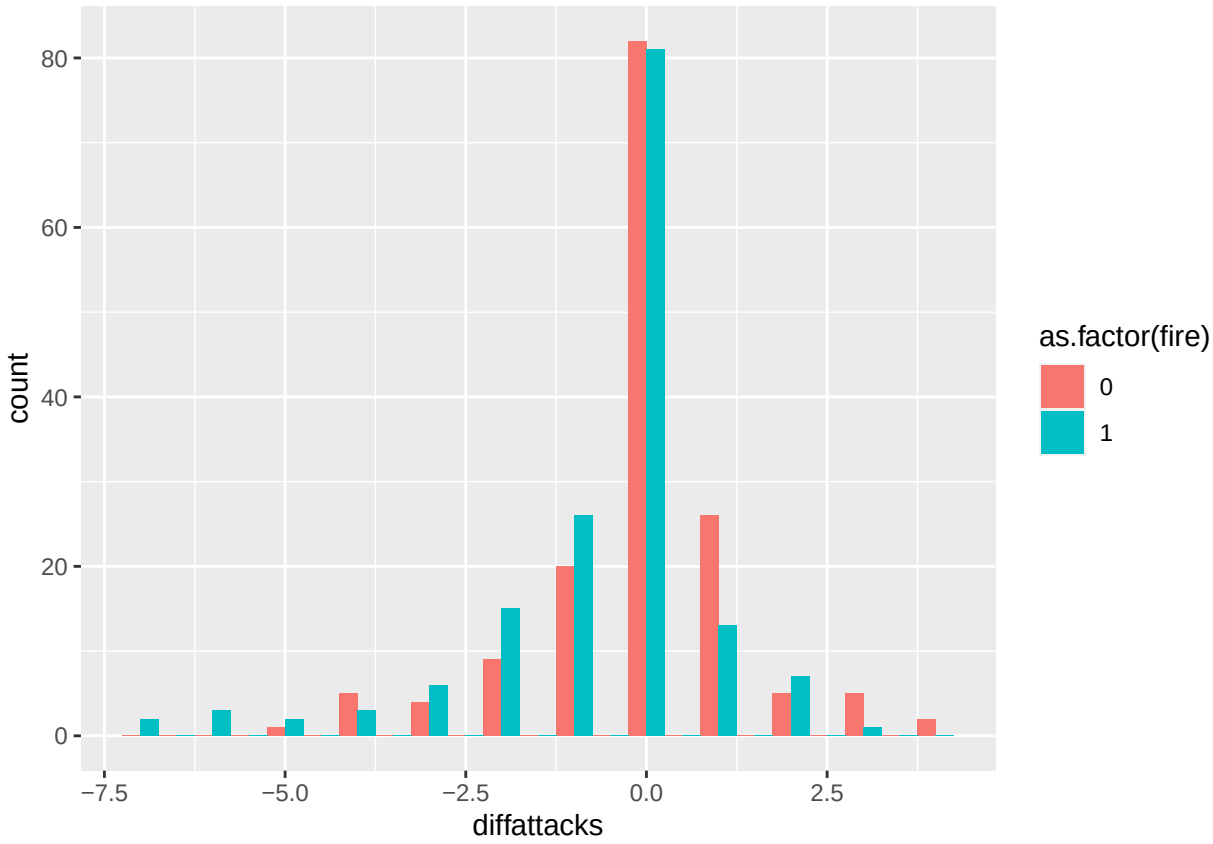
```
chechen_new <- chechen %>%
  mutate(diffattacks = postattack - preattack)

head(chechen_new)
```

```
## # A tibble: 6 x 7
##   village      groznyy  fire deaths preattack postattack diffattacks
##   <chr>          <dbl> <dbl>  <dbl>     <dbl>      <dbl>       <dbl>
## 1 Elistanzhi         0     0     NA         4          3          -1
## 2 Malye Shuani       0     1      0         0          1           1
## 3 Belgatoy           0     1     34         1          0          -1
## 4 Oktya'brskoe       0     0     NA         0          0           0
## 5 Chiri-Yurt         0     0     NA         4          5           1
## 6 Gansolchu          0     1      0         0          0           0
```

```
hist2 <- ggplot(
  data=chechen_new,
  mapping = aes(
    x = diffattacks,
    group = as.factor(fire),
    fill = as.factor(fire)
    )
  )

hist2 + geom_histogram(
  position = "dodge",
  binwidth = 0.5
  )
```



Question 6

The mean difference in the `difattacks` variable between villages shelled and villages not shelled is 0.52 attacks, meaning that there were, on average .52 fewer insurgent attacks in the villages shelled than the villages not shelled after the Russian fire. The difference-in-differences estimator has shown that the change in insurgent attacks in shelled villages is greater in magnitude than the change in insurgent attacks in villages not shelved. Accepting the assumption that villages shelled and not shelled do not vary systematically and that the Russian artillery fire was carried out randomly, i.e. not correlated with any characteristics specific to the villages, this analysis supports the claim that indiscriminate violence reduces insurgency attacks.

The validity of this analysis is improved compared to the previous questions since this analysis takes the change into account and not just the post Russian fire data. Using information on insurgency attacks before the Russian fire allows us to compute whether the Russian fires actually made a difference to the number of insurgency attacks or not. We needed pre-treatment data, so to speak, to evaluate any type of effect before even beginning to talk about causality.

```
mean(chechen_new$diffattacks[chechen_new$fire == 1])-mean(chechen_new$diffattacks[chechen_new$fire == 0
```

```
## [1] -0.5157233
```

Comment: Good!

Inequality of Success in Online Music Markets

Question 1

Proportion of users assigned separately for the Social Influence and the Independent condition.

```
users1_new <- users1 %>%
  mutate(expnr = 1)
```

```r
users2_new <- users2 %>%
  mutate(expnr = 2)

users_full <- rbind(users1_new,users2_new)

view(users_full)

  table1 <- prop.table(table(users_full$expnr, users_full$world_id), margin = 1)
  round(table1,3)
```

```
##
##         1     9
##   1 0.328 0.672
##   2 0.323 0.677
```

```r
  tabledata1 <- prop.table(table(users1$world_id))
  tabledata2 <- prop.table(table(users2$world_id))

  tabledata1
```

```
##
##         1         9
## 0.3275782 0.6724218
```

```r
  tabledata2
```

```
##
##         1         9
## 0.3227166 0.6772834
```

The proportions within each experiment are practically identical, with 33% assigned to the Social Influence condition and 67% to the Independence condition in Experiment 1. The equivalent proportions are 32% and 68% for experiment 2.

Comment: Good!

Question 2

The average number of downloads (assuming this is equivalent to purchasing the song) per user.

```r
usersfull_temp <- users_full %>%
  group_by(expnr, world_id) %>%
  summarise(
    mean_download = mean(download)
    )
```

```
## `summarise()` has grouped output by 'expnr'. You can override using the
## `.groups` argument.
```

```r
usersfull_temp
```

```
## # A tibble: 4 x 3
## # Groups:   expnr [2]
##   expnr world_id mean_download
##   <dbl>    <dbl>         <dbl>
## 1     1        1          42.2
## 2     1        9          47.2
## 3     2        1         301.
## 4     2        9         124.
```

```
# Attempt using just users1 and users2 separately
users1 %>%
  group_by(world_id) %>%
  summarise(
    mean_download = mean(download)
    )
```

```
## # A tibble: 2 x 2
##   world_id mean_download
##      <dbl>         <dbl>
## 1        1          42.2
## 2        9          47.2
```

The average number of downloads per user in Experiment 1 is 42.2 for the people assigned the social influence condition and 47.15 for the people assigned the Independence condition.

The average number of downloads per user in Experiment 2 is 301.2 for the people assigned the social influence condition and 124.1 for the people assigned the independence condition.

```
# calculating number of workers
usersnr <- users_full %>%
  group_by(expnr, world_id) %>%
  summarise(n_users = n())
```

```
## `summarise()` has grouped output by 'expnr'. You can override using the
## `.groups` argument.
```

```
head(usersnr)
```

```
## # A tibble: 4 x 3
## # Groups:   expnr [2]
##   expnr world_id n_users
##   <dbl>    <dbl>   <int>
## 1     1        1     702
## 2     1        9    1441
## 3     2        1     689
## 4     2        9    1446
```

```
# calculating total number of times a song was listened to
songs1_new <- songs1 %>%
  mutate(expnr = 1)
songs2_new <-songs2 %>%
  mutate(expnr = 2)
songs_full <- rbind(songs1_new,songs2_new)

songsnrtemp <- songs_full %>%
  group_by(expnr) %>%
  summarise(
            listen_soctotal = sum(listen_soc),
            listen_indeptotal = sum(listen_indep)
            )
songsnrtemp
```

```
## # A tibble: 2 x 3
##   expnr listen_soctotal listen_indeptotal
##   <dbl>           <dbl>             <dbl>
## 1     1            2461              5394
```

```
## 2     2            2455              5643
```

```
songsnr <- songsnrtemp %>%
    pivot_longer(
    cols = starts_with("listen"),
    names_to = "world_id",
    values_to = "listen_total"
    )

songsnr$world_id[which(songsnr$world_id=="listen_soctotal")] <- "1"
songsnr$world_id[which(songsnr$world_id=="listen_indeptotal")] <- "9"
songsnr$world_id <- as.numeric(songsnr$world_id)

# merging users and song data
totals <- merge(usersnr, songsnr, by=c("expnr","world_id"))
totals <- totals %>%
  mutate(listen_per_user = listen_total/n_users)
totals
```

```
##   expnr world_id n_users listen_total listen_per_user
## 1     1        1     702         2461        3.505698
## 2     1        9    1441         5394        3.743234
## 3     2        1     689         2455        3.563135
## 4     2        9    1446         5643        3.902490
```

The average listens per user in Experiment 1 is 3.51 for the people assigned the social influence condition and 3.74 for the people assigned the Independence condition.

The average listens per user in Experiment 2 is 3.56 for the people assigned the social influence condition and 3.9 for the people assigned the independence condition.

It seems odd that there are less listens than downloads per user. If we use the songs-dataset for downloads as well as listens, there will be lower downloads than listens, leading me to believe that the relevant downloads-data is from the songs-data and not the users-data.

Comment: Good! The answer you gave for average number of downloads per user is incorrect, but that is because the `download` variable is deceptively labeled (see answer key for more details). So, I didn't take points off for that.

Question 3

```
library(ineq)

songs_full %>%
  group_by(expnr) %>%
  summarise(
    gini_down_soc = ineq(down_soc),
    gini_down_indep = ineq(down_indep),
    gini_listen_soc = ineq(listen_soc),
    gini_listen_indep = ineq(listen_indep)
    )
```

```
## # A tibble: 2 x 5
##   expnr gini_down_soc gini_down_indep gini_listen_soc gini_listen_indep
##   <dbl>         <dbl>           <dbl>           <dbl>             <dbl>
## 1     1         0.334           0.244           0.169            0.121
## 2     2         0.457           0.187           0.420            0.0898
```

The results support the hypothesis that inequality in both listens and downloads is higher under the social

influence condition than the independent condition. Furthermore, in Experiment 2 where the songs were ranked according to popularity as well as displaying number of downloads for each song, the inequality is substantially larger. ***Comment: Good!***

Question 4

```
users_full %>%
  group_by(expnr, world_id) %>%
  summarise(
    mean_web = mean(web, na.rm = TRUE),
    mean_visit = mean(visit, na.rm = TRUE),
    mean_purchase = mean(purchase, na.rm = TRUE),
    n = n()
  )
```

```
## `summarise()` has grouped output by 'expnr'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 6
## # Groups:   expnr [2]
##   expnr world_id mean_web mean_visit mean_purchase     n
##   <dbl>    <dbl>    <dbl>      <dbl>         <dbl> <int>
## 1     1        1     4.50       1.87         0.721   702
## 2     1        9     4.51       1.90         0.733  1441
## 3     2        1     4.51       1.96         0.749   689
## 4     2        9     4.47       1.91         0.734  1446
```

The groups are on average practically identical across experiment and condition type. Additionally, the group sizes for each condition are similar.

Question 5

```
exp1effect_d <- ineq(songs1$down_soc)-ineq(songs1$down_indep)
exp2effect_d <- ineq(songs2$down_soc)-ineq(songs2$down_indep)

exp1effect_d
```

```
## [1] 0.09055846
```

```
exp2effect_d
```

```
## [1] 0.2696289
```

```
exp1effect_d - exp2effect_d
```

```
## [1] -0.1790704
```

```
exp1effect_l <- ineq(songs1$listen_soc)-ineq(songs1$listen_indep)
exp2effect_l <- ineq(songs2$listen_soc)-ineq(songs2$listen_indep)

exp1effect_l
```

```
## [1] 0.04774469
```

```
exp2effect_l
```

```
## [1] 0.3303566
```

```
exp1effect_l - exp2effect_l
```

```
## [1] -0.2826119
```

The difference in the change in Gini-coefficient is .18 and .28 gini-points for downloads and listens respectively. The between study comparison reflects the extra increase in inequality from exposing the users to an ordering of the songs by popularity. One could suspect this between study to have less internal validity simply from the fact that we cannot be sure whether the experiments were conducted similarly and thus are believably comparable. However, question 4 revealed that the groups in the two experiments look similar in measurable characteristics which strengthens the validity of the between study comparison.

Comment: Good!