PSC4375: Missing Data

Week 3: Lecture 6

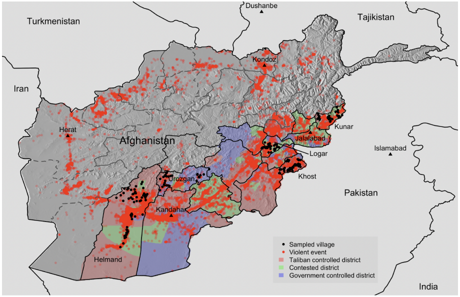
Prof. Weldzius

Villanova University

Slides Updated: 2025-02-05

Civilian attitudes and war against insurgency

- War in Afghanistan: counter-insurgency war
  - Military against insurgents
  - Key to victory: winning hearts and minds of civilians
  - Aid provision, information campaign, minimizing civilian casualties
- How does exposure to violence affect support for Taliban coalition?



Afghan study

```r
library(tidyverse)
data(afghan, package = "qss")
head(afghan[,1:8])
```

```
##   province     district village.id age educ.years employed
## 1    Logar Baraki Barak         80  26         10        0
## 2    Logar Baraki Barak         80  49          3        1
## 3    Logar Baraki Barak         80  60          0        1
## 4    Logar Baraki Barak         80  34         14        1
## 5    Logar Baraki Barak         80  21         12        1
## 6    Logar Baraki Barak         80  18         10        1
##          income violent.exp.ISAF
## 1 2,001-10,000                0
## 2 2,001-10,000                0
## 3 2,001-10,000                1
## 4 2,001-10,000                0
## 5 2,001-10,000                0
## 6         <NA>                0
```

Missing data

- **Nonresponse**: respondent can't or won't answer question
  - Sensitive questions ⇝ **social desirability bias**
  - Some countries lack official statistics like unemployment
  - Leads to missing data
- Missing data in R: a special value `NA`
- Causes problems with calculating statistics:

```r
## prop. of those who got hurt by ISAF
mean(afghan$violent.exp.ISAF)
```

```
## [1] NA
```

Handling missing data in R

- Adding `na.rm = TRUE` to some functions removes missing data

```
afghan %>% summarize(mean(violent.exp.ISAF, na.rm = TRUE))
```

```
##   mean(violent.exp.ISAF, na.rm = TRUE)
## 1                            0.3748626
```

- Or, you can remove missing values using `na.omit()` function:

```
afghan %>% summarize(mean(na.omit(violent.exp.ISAF)))
```

```
##   mean(na.omit(violent.exp.ISAF))
## 1                       0.3748626
```

Handling missing data in R

- See number of NAs with count() + group_by()

```
afghan %>%
  group_by(violent.exp.ISAF) %>%
  count()
```

```
## # A tibble: 3 x 2
## # Groups:   violent.exp.ISAF [3]
##   violent.exp.ISAF     n
##              <int> <int>
## 1                0  1706
## 2                1  1023
## 3               NA    25
```

Available-case vs. complete-case analysis

- **Available-case analysis**: use the data you have for that variable:

```r
afghan %>%
  summarize(sum(!is.na(violent.exp.ISAF)))
```

```
##   sum(!is.na(violent.exp.ISAF))
## 1                          2729
```

```r
afghan %>%
  summarize(mean(violent.exp.ISAF,na.rm=TRUE))
```

```
##   mean(violent.exp.ISAF, na.rm = TRUE)
## 1                            0.3748626
```

Available-case vs. complete-case analysis

- **Complete-case analysis**: only use units that have data on all variables
  - Also called **listwise deletion**

```r
dim(na.omit(afghan))
```

```
## [1] 2554   11
```

```r
afghan %>%
  na.omit() %>%
  summarize(mean(violent.exp.ISAF))
```

```
##   mean(violent.exp.ISAF)
## 1              0.3719655
```

Non-response and other biases

- Nonresponse can create bias
- More violent areas ⇝ more non-response:

```
afghan %>%
  group_by(province) %>%
  summarize(
    violent.exp.taliban = mean(is.na(violent.exp.taliban)),
    violent.exp.ISAF = mean(is.na(violent.exp.ISAF)))
```

```
## # A tibble: 5 x 3
##   province violent.exp.taliban violent.exp.ISAF
##   <chr>                  <dbl>            <dbl>
## 1 Helmand              0.0304           0.0164
## 2 Khost                0.00635          0.00476
## 3 Kunar                0                0
## 4 Logar                0                0
## 5 Uruzgan              0.0620           0.0207
```

- ⇝ oversampling citizens with less exposure to violence!