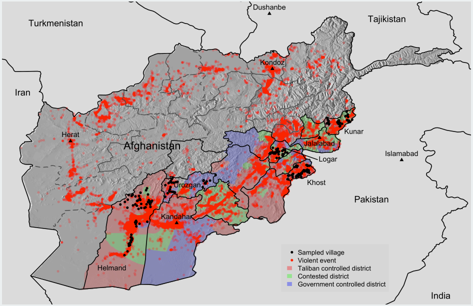
Gov 51: Missing Data

Matthew Blackwell

Harvard University

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
afghan <- read.csv("data/afghan.csv")
head(afghan[, 1:8])
```

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

```
## 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.

```r
mean(afghan$violent.exp.ISAF, na.rm = TRUE)
```

```
## [1] 0.375
```

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

```r
mean(na.omit(afghan$violent.exp.ISAF))
```

```
## [1] 0.375
```

- Add `NA` to `table()` with `exclude = NULL`:

```r
table(ISAF = afghan$violent.exp.ISAF, exclude = NULL)
```

```
## ISAF
##    0    1 <NA>
## 1706 1023   25
```

Available-case vs complete-case analysis

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

```
sum(!is.na(afghan$violent.exp.ISAF))
```

```
## [1] 2729
```

```
mean(afghan$violent.exp.ISAF, na.rm = TRUE)
```

```
## [1] 0.375
```

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

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

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

```
mean(na.omit(afghan)$violent.exp.ISAF)
```

```
## [1] 0.372
```

Non-response and other biases

- Nonresponse can create bias.

- More violent areas ⇝ more non-response:

```
tapply(is.na(afghan$violent.exp.taliban), afghan$province,
       mean)
```

```
## Helmand   Khost   Kunar   Logar Uruzgan
## 0.03041 0.00635 0.00000 0.00000 0.06202
```

```
tapply(is.na(afghan$violent.exp.ISAF), afghan$province,
       mean)
```

```
## Helmand   Khost   Kunar   Logar Uruzgan
## 0.01637 0.00476 0.00000 0.00000 0.02067
```

- ⇝ oversampling citizens with less exposure to violence.