Measurement: Survey Sampling

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

Sample Surveys

- Probability sampling to ensure representativeness
- Definition: every unit in the population has a known non-zero probability of being selected
- Simple random sampling: every unit has an *equal* selection probability

Population	Sample	Potential Bias
Target population		
↓		Frame bias
Frame population →	Sample	Sampling bias
	↓	Unit non-response
	Respondents	
	↓	Item non-response
	Completed items	
		Response bias

Difficulty of Obtaining a Representative Sample

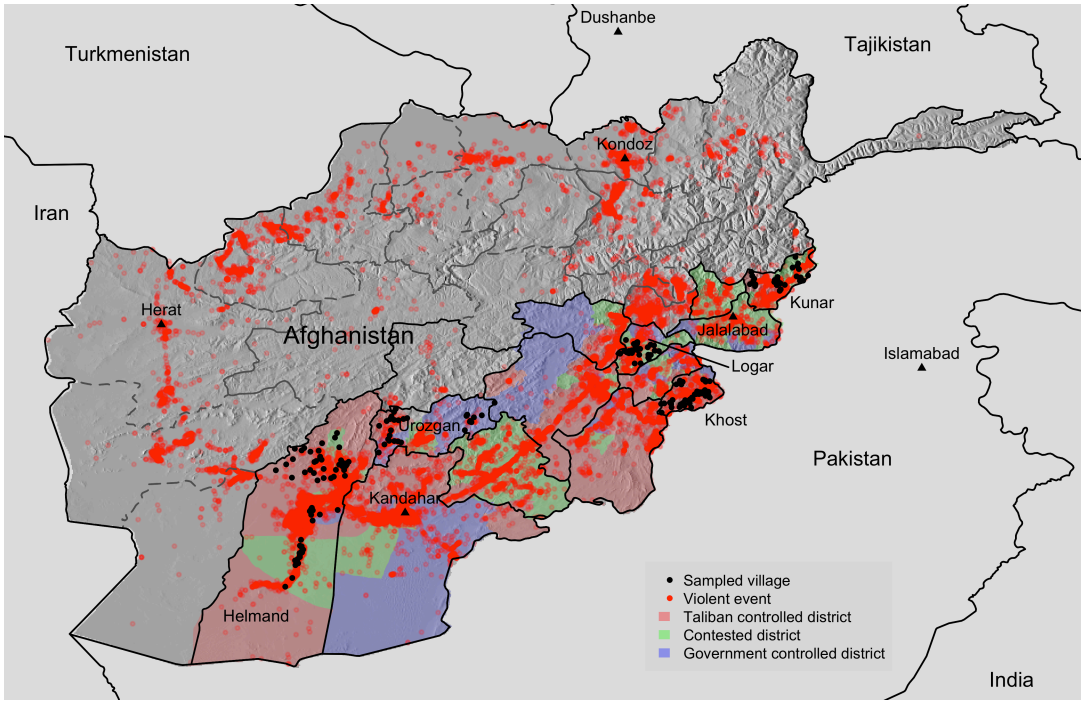
- Problems of telephone survey
  - Random digit dialing from phone book
  - Cell phones (double counting for the wealthy)
  - Caller ID screening (unit non-response)

- An alternative: Internet survey
  - Opt-in panels, respondent driven sampling $\rightsquigarrow$ non-probability sampling
  - Cheap but non-representative
  - Digital divide: rich vs. poor, young vs. old
  - Correct for potential sampling bias via statistical methods

- "Report of the AAPOR (American Association for Public Opinion Research) Task Force on Non-probability Sampling" (2013)

Civilian Attitudes and War against Insurgency

- Conventional war: military against another military
- Counter-insurgency war: military against insurgents
  - From Vietnam war to war in Afghanistan
  - Key to victory: winning hearts and minds of civilians
  - aid provision, information campaign, minimizing civilian casualties
- Afghanistan: arguably the world's most heavily surveyed population



Handling Missing Data in **R**

- Missing data in **R**: a special value `NA`
- Adding `na.rm = TRUE` to some functions removes missing data

```
afghan <- read.csv("data/afghan.csv")
## prop. of those who got hurt by ISAF
mean(afghan$violent.exp.ISAF)
## [1] NA
mean(afghan$violent.exp.ISAF, na.rm = TRUE)
## [1] 0.375
```

- Or, you can explicitly remove it by the `na.omit()` function

```
mean(na.omit(afghan$violent.exp.ISAF))
## [1] 0.375
```

- complete-case analysis (listwise deletion) vs. available-case analysis

```
sum(is.na(afghan$violent.exp.ISAF))
## [1] 25
dim(na.omit(afghan))
## [1] 2554   11
mean(na.omit(afghan)$violent.exp.ISAF)
## [1] 0.372
```

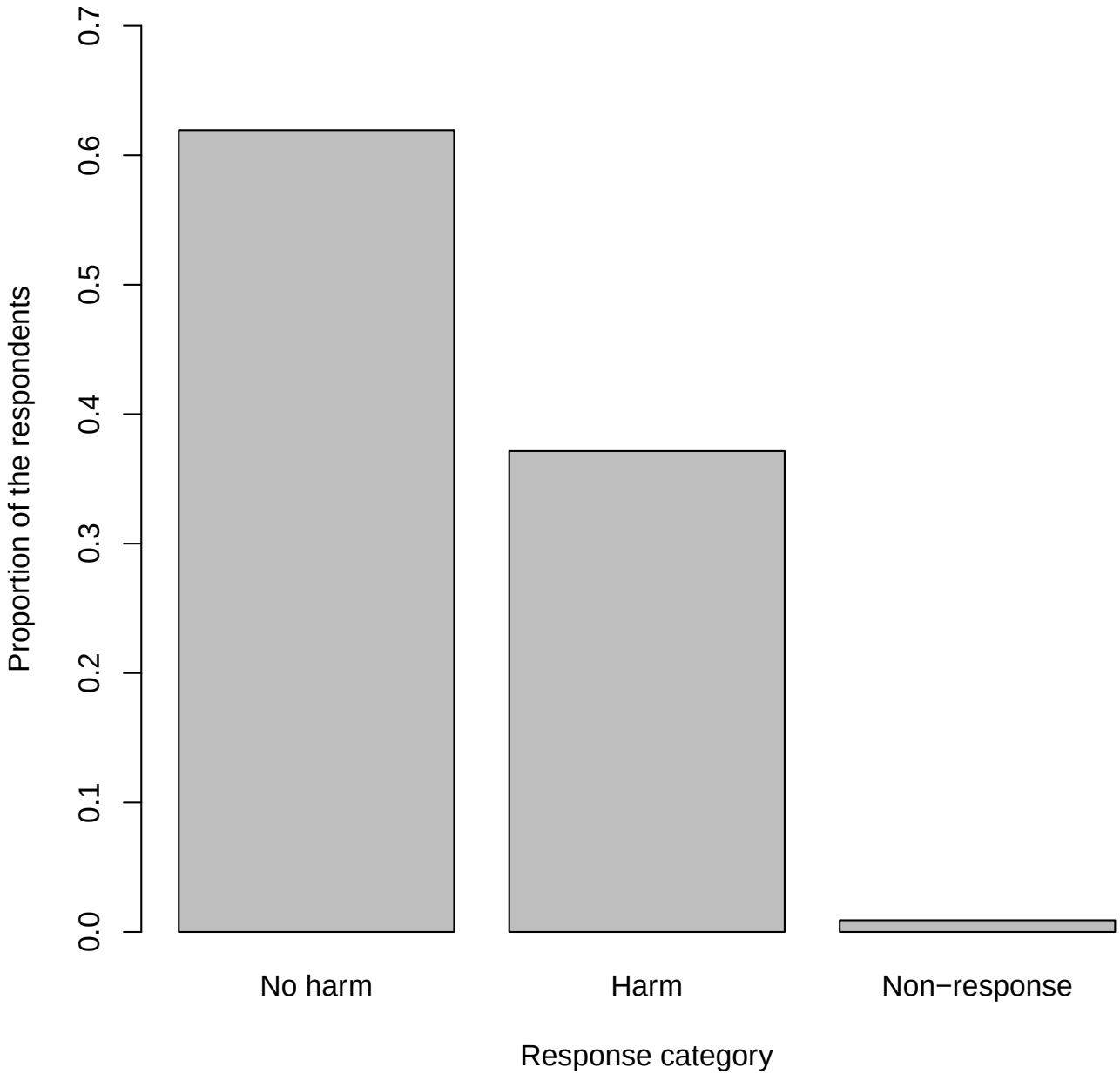
- Including NA as an additional category in a contingency table

```
table(ISAF = afghan$violent.exp.ISAF,
      Taliban = afghan$violent.exp.taliban, exclude = NULL)
##       Taliban
## ISAF      0    1 <NA>
##   0    1330  354   22
##   1     475  526   22
##   <NA>    7    8   10
```

Barplot

- Visualize the distribution of a factor (categorical) variable

```
barplot(prop.table(table(ISAF = afghan$violent.exp.ISAF,
                         exclude = NULL)),
        names.arg = c("No harm", "Harm", "Non-response"),
        main = "Civilian victimization by the ISAF",
        xlab = "Response category",
        ylab = "Proportion of the respondents",
        ylim = c(0, 0.7))
```


Civilian victimization by the ISAF
Proportion of the respondents
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
No harm
Harm
Non−response
Response category

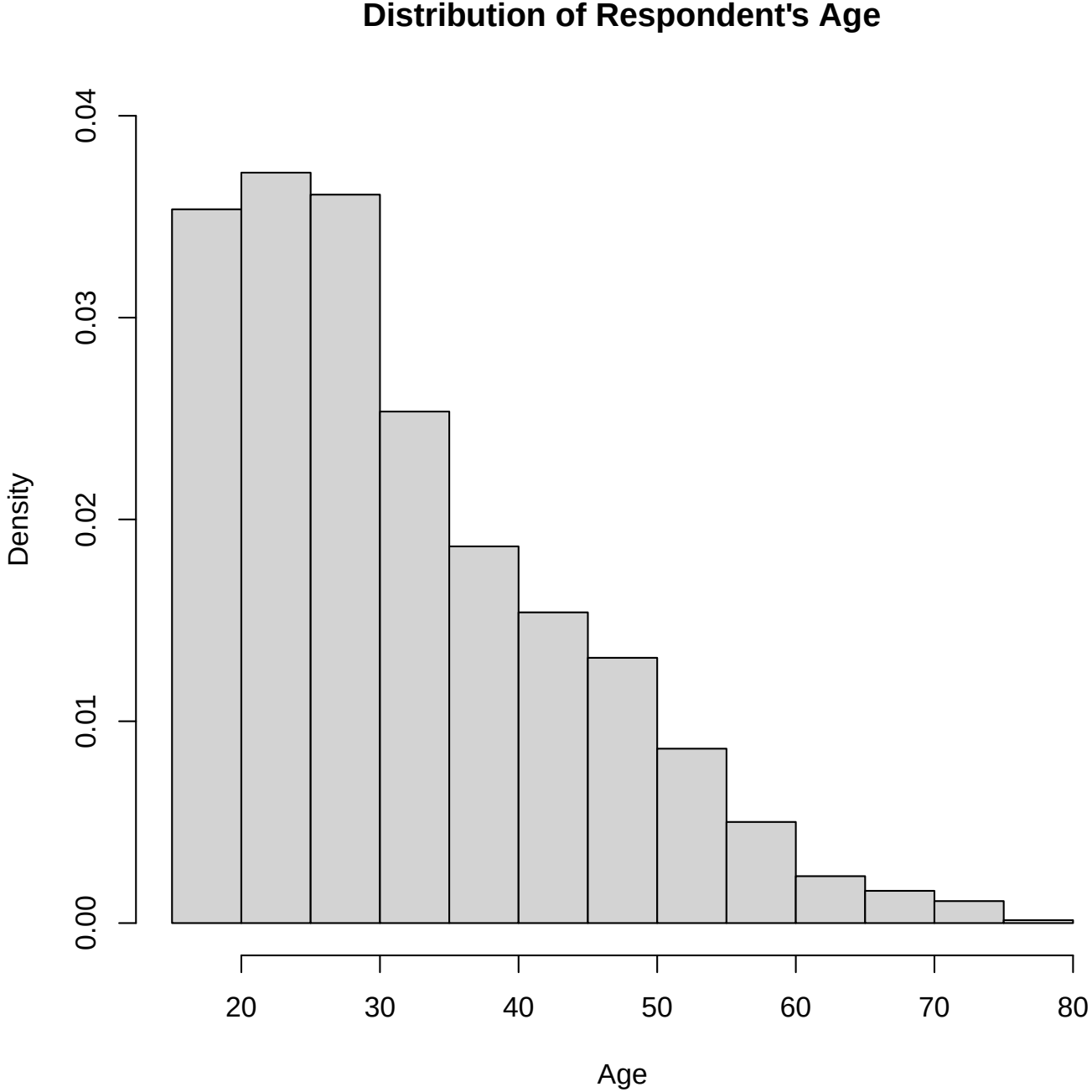
Histogram

- Visualize the distribution of a continuous variable
- How to create a histogram by hand:
  1. create bins along the variable of interest
  2. count number of observations in each bin
  3. density = bin height

$$\text{density} = \frac{\text{proportion of observations in bin}}{\text{bin width}}$$

- In **R**, we use the `hist()` with `freq = FALSE`

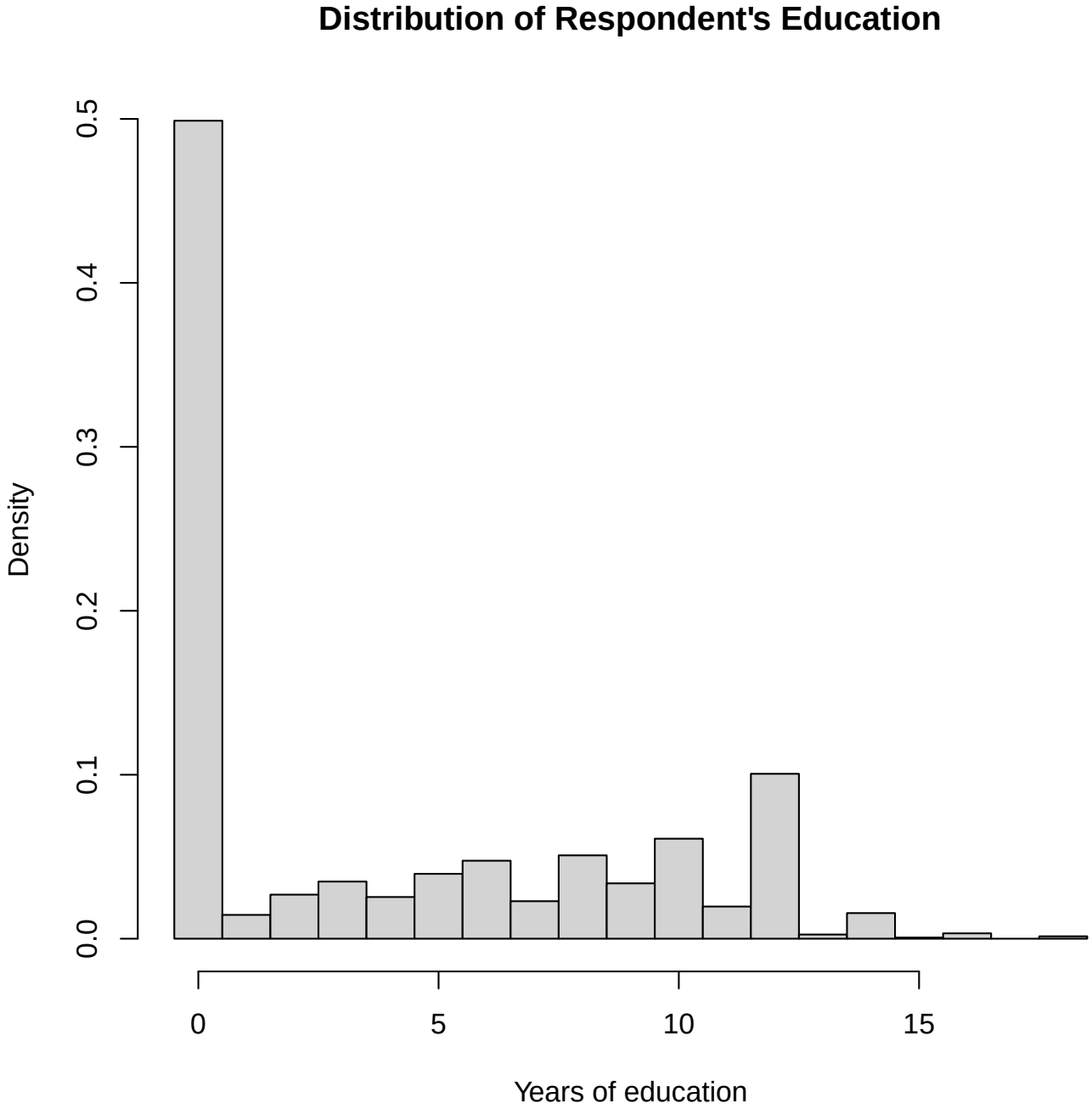
```
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04),
     xlab = "Age", main = "Distribution of Respondent's Age")
```


Distribution of Respondent's Age
Density
0.00
0.01
0.02
0.03
0.04
20
30
40
50
60
70
80
Age

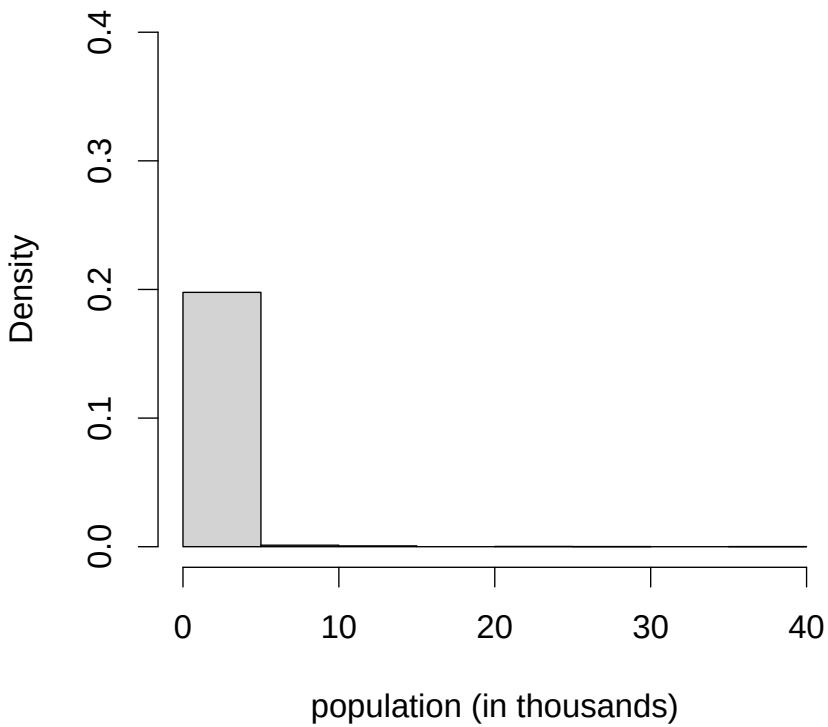
What is Density?

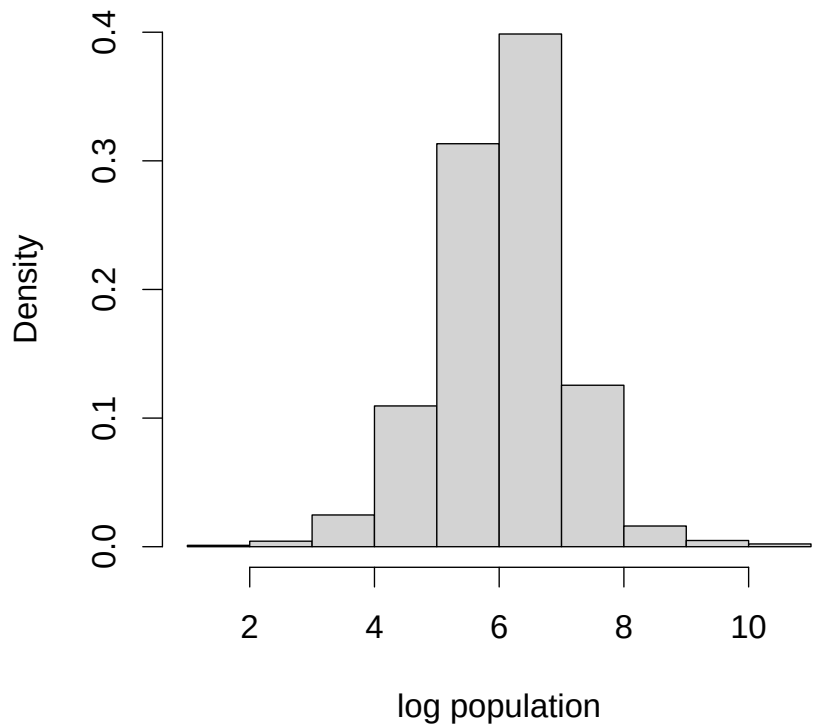
- The areas of the blocks sum to 1 or 100%
- Density $\neq$ Percentage (e.g., range $\neq [0, 1]$)
- The height of the blocks equals the percentage divided by the bin width: in this case, "percent per year"
- More generally, "percentage per horizontal unit"
- We can also choose the bin locations on our own via the `breaks` (locations of bin breaks) or `nclass` (number of bins) arguments

```
hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of education",
     main = "Distribution of Respondent's Education")
```


Distribution of Respondent's Education
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
Years of education

Afghan Village Population





Non-response and Other Sources of Bias

- Item non-response, like unit non-response, can create bias
- More violent areas $\rightsquigarrow$ 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
```

- Sensitive questions $\rightsquigarrow$ non-response, social desirability bias
- racial prejudice, corruption, even turnout
- Do you support ISAF? What about Taliban?

Public Nature of Interviews

List Experiments

- Script for the control group:

  I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

  Karzai Government; National Solidarity Program; Local Farmers

List Experiments

- Script for the treatment group:

  I'm going to read you a list with the names of different groups and individuals on it. After I read the entire list, I'd like you to tell me how many of these groups and individuals you broadly support, meaning that you generally agree with the goals and policies of the group or individual. Please don't tell me which ones you generally agree with; only tell me how many groups or individuals you broadly support.

  Karzai Government; National Solidarity Program; Local Farmers; **ISAF** (**Taliban**)

Analysis of List Experiment

- Proportion of those who support ISAF

```
mean(afghan$list.response[afghan$list.group == "ISAF"]) -
  mean(afghan$list.response[afghan$list.group == "control"]
## [1] 0.049
```

- The problem of list experiment: floor and ceiling effects

```
table("response" = afghan$list.response,
      "group" = afghan$list.group)
##          group
## response control ISAF taliban
##        0     188  174       0
##        1     265  278     433
##        2     265  260     287
##        3     200  182     198
##        4       0   24       0
```

Summary

- Importance of probability sampling:
  1. ensures representativeness
  2. enables the calculation of uncertainty (QSS Chapter 7)

- Sources of bias in survey sampling
  1. unit non-response
  2. item non-response
  3. social desirability bias
  4. differential item functioning

- Survey methods for sensitive questions
  1. list experiment
  2. randomize response method