Gov 50: 7. Measurement: Visualizing Distributions

Matthew Blackwell

Harvard University

Fall 2018

1. Today’s agenda

2. Visualizing data

3. Anchoring vignettes

1/ Today's agenda

Where are we going?

- Last time: how to summarize data with numerical values
- This time: how to visually summarize data.
- Anchoring vignettes for cross-national surveys
  - King (in Govt Dept!), Murray, Salomon, and Tandon (2004). “Enhancing the validity and cross-cultural comparability of measurement in survey research.” American Political Science Review.
- Logistical issues:
  - DataCamp issues
  - Notetaker

2/ Visualizing data

Studying political efficacy

- 2002 WHO survey of people in China and Mexico.
- Goal: determine feelings of political efficacy.
- Question: “How much say do you have in getting the government to address issues that interest you?”
    1. No say at all
    2. little say
    3. some say
    4. a lot of say
    5. unlimited say

Data

- Load the data:

```
vignettes <- read.csv(”data/vignettes.csv”)
head(vignettes)
```

```
##   self alison jane moses china age
## 1    1      5    5     2     0  31
## 2    1      1    5     5     0  54
## 3    2      3    1     1     0  50
## 4    2      4    2     1     0  22
## 5    2      3    3     3     0  52
## 6    1      3    1     5     0  50
```

Contingency table

- The `table()` function shows us how many respondents are in each category of a categorical variable:

```
table(vignettes$self)
```

```
##
##   1   2   3   4   5
## 327 210 130  56  58
```

- We can use `prop.table()` to show what **proportions** of the data each response represents:

```
prop.table(table(vignettes$self))
```

```
##
##      1      2      3      4      5
## 0.4187 0.2689 0.1665 0.0717 0.0743
```

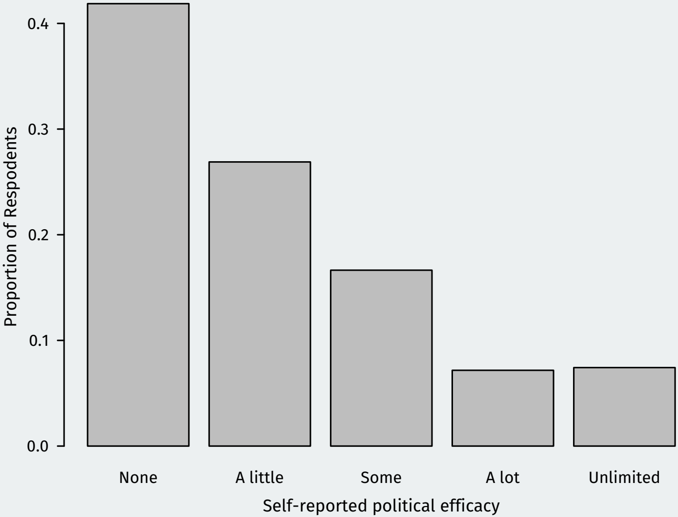
Barplot

- The barplot() function can help us visualize a contingency table:

```
barplot(prop.table(table(vignettes$self)),
        names = c(”None”, ”A little”,
                  ”Some”, ”A lot”, ”Unlimited”),
        xlab = ”Self-reported political efficacy”,
        ylab = ”Proportion of Respodents”)
```

- Arguments:
  - First is the height each bar should take (we’re using proportions in this case)
  - names are the labels for the each category
  - xlab, ylab are axis labels

Barplot


0.4
0.3
0.2
0.1
0.0
Proportion of Respodents
None
A little
Some
A lot
Unlimited
Self-reported political efficacy

Histogram

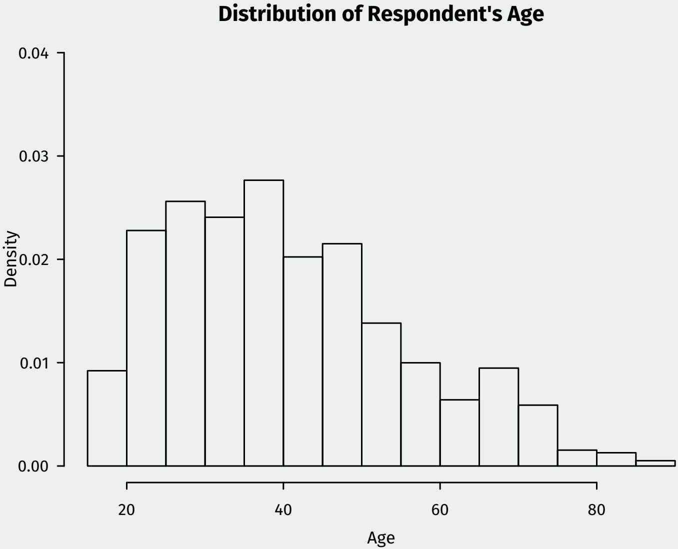
- Visualize density of continuous/numeric variable.
- How to create a histogram by hand:
  1. create bins along the variable of interest
  2. count number of observations in each bin
  3. **density** = bin height

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

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

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

- Other arguments:
  - `ylim` sets the range of the y-axis to show (if you don't set it, uses the range of the data).
  - `main` sets the title for the figure.

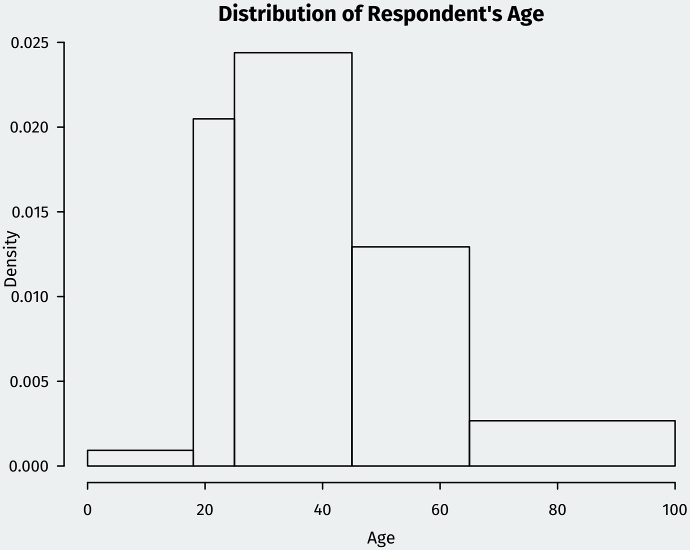

Distribution of Respondent's Age
Density
0.04
0.03
0.02
0.01
0.00
20
40
60
80
Age

What is density?

- The areas of the blocks = proportion of observations in those blocks.
- ⇝ area of the blocks sum to 1 (100%)
- Can lead to confusion: height of block can go above 1!
- We can also choose the bin locations on our own via the `breaks` (location of the bin breaks) or `nclass` (number of bins):

```
hist(vignettes$age, freq = FALSE,
     breaks = c(0, 18, 25, 45, 65, 100),
     xlab = ”Age”,
     main = ”Distribution of Respondent's Age”)
```

Creating our own bins



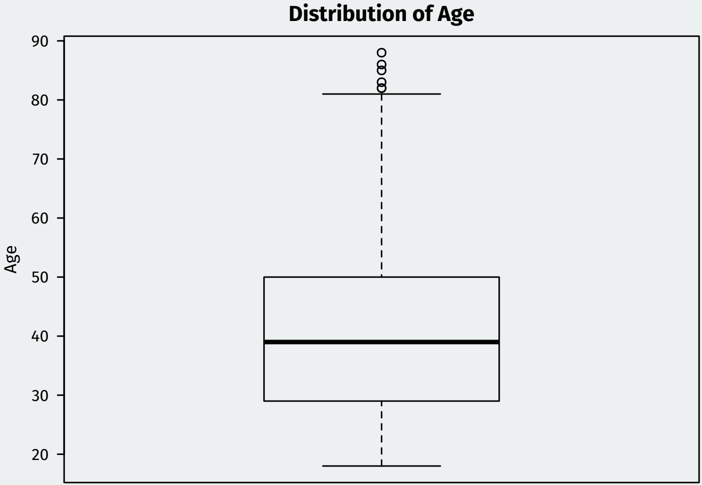
Boxplot

- A **boxplot** can characterize the distribution of continuous variables
- Use `boxplot()`:

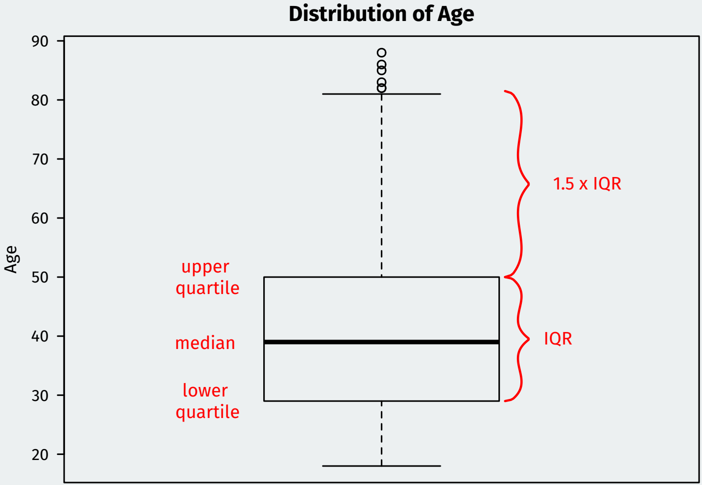
```r
boxplot(vignettes$age, main = ”Distribution of Age”,
        ylab = ”Age”)
```

- “Box” represents range between lower and upper quartile.
- “Whiskers” represents either:
  - 1.5 × IQR or max/min of the data, whichever is smaller.
  - Points beyond whiskers are outliers.

Boxplot



Boxplot

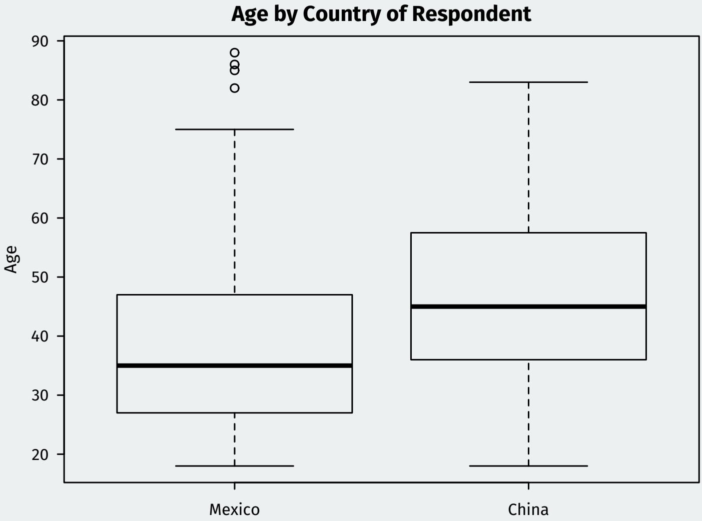

Distribution of Age
Age
90
80
70
60
50
40
30
20
upper quartile
median
lower quartile
1.5 x IQR
IQR

Comparing distribution with the boxplot

- Useful for comparing a variable across groups:

```
boxplot(age ~ china, data = vignettes,
        names = c(”Mexico”, ”China”),
        main = ”Age by Country of Respondent”,
        ylab = ”Age”)
```

- First argument is called a formula, `y ~ x`:
  - `y` is the continuous variable whose distribution we want to explore.
  - `x` is the grouping variable.
  - When using a formula, we need to add a `data` argument.


Age by Country of Respondent
Age
90
80
70
60
50
40
30
20
Mexico
China

3/ Anchoring vignettes

Possible biases

- Question: “How much say do you have in getting the government to address issues that interest you?”
  1. No say at all
  2. little say
  3. some say
  4. a lot of say
  5. unlimited say
- Problem? Different people interpret questions differently
  - Cross-cultural differences, vague questions.

Vignettes to the rescue

- Solution: try to anchor responses with **vignettes** with different levels of "objective" efficacy:

  Alison lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

- How much say does Alison have in getting the government to address issues that interest her?
  - Use the same scale as self-assessment.

Jane vignette

Jane lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.

- How much say does Jane have in getting the government to address issues that interest her?

Moses vignette

Moses lacks clean drinking water. He would like to change this, but he can’t vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

- How much say does Moses have in getting the government to address issues that interest him?
- “Objective” ranking: Alison $>$ Jane $>$ Moses.

Data

```r
head(vignettes)
```

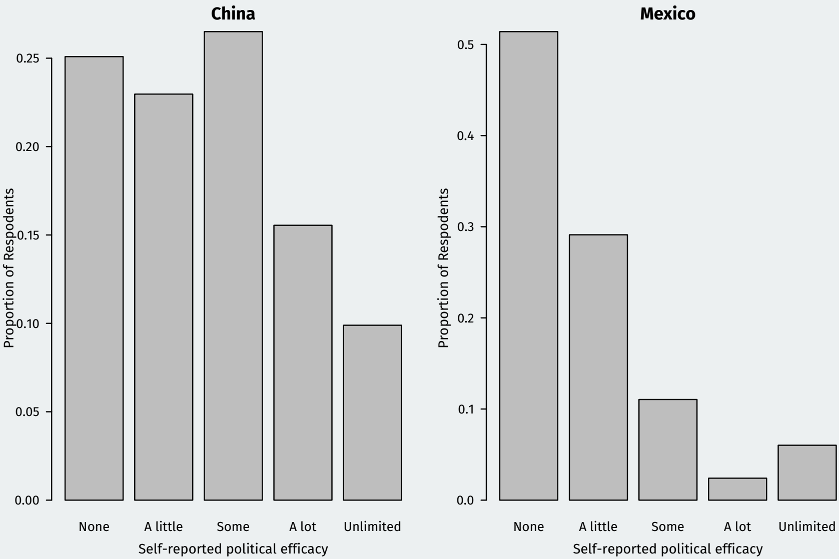
```
##   self alison jane moses china age
## 1    1      5    5     2     0  31
## 2    1      1    5     5     0  54
## 3    2      3    1     1     0  50
## 4    2      4    2     1     0  22
## 5    2      3    3     3     0  52
## 6    1      3    1     5     0  50
```

Self-reported efficacy

```
china <- vignettes[vignettes$china == 1,]
mexico <- vignettes[vignettes$china == 0,]

barplot(prop.table(table(china$self)),
        names = c(”None”, ”A little”,
                  ”Some”, ”A lot”, ”Unlimited”),
        xlab = ”Self-reported political efficacy”,
        ylab = ”Proportion of Respodents”,
        main = ”China”)

barplot(prop.table(table(mexico$self)),
        names = c(”None”, ”A little”,
                  ”Some”, ”A lot”, ”Unlimited”),
        xlab = ”Self-reported political efficacy”,
        ylab = ”Proportion of Respodents”,
        main = ”Mexico”)
```


China
Proportion of Respodents
0.00
0.05
0.10
0.15
0.20
0.25
None
A little
Some
A lot
Unlimited
Self-reported political efficacy
Mexico
Proportion of Respodents
0.0
0.1
0.2
0.3
0.4
0.5
None
A little
Some
A lot
Unlimited
Self-reported political efficacy

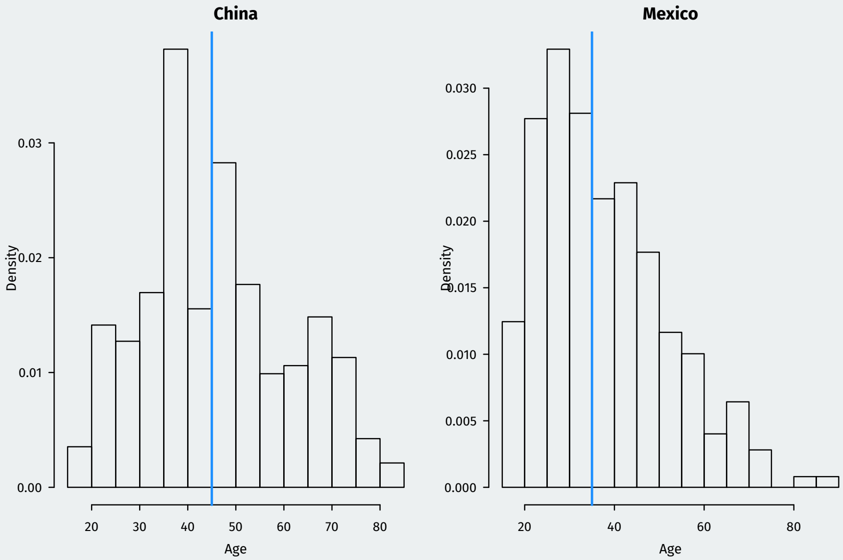
What’s going on?

- Weird since:
  - Mexico just ousted long-ruling party (PRI) in 2000 election.
  - China has no free and fair elections.
- Could it be due to age differences between the samples?
  - Maybe Mexico sample is older and has more experience under 1-party rule?

```
hist(china$age, freq = FALSE, xlab = ”Age”, main = ”China”)
abline(v=median(china$age), col = ”dodgerblue”, lwd = 2)

hist(mexico$age, freq = FALSE, xlab = ”Age”,
     main = ”Mexico”)
abline(v=median(mexico$age), col = ”dodgerblue”, lwd = 2)
```

- `abline(v = 1)` adds a vertical line at 1, `abline(h = 1)` adds a horizontal line at 1.
  - `col` is the color of the line
  - `lwd` controls the width of the line


China
Mexico
Density
0.00
0.01
0.02
0.03
20
30
40
50
60
70
80
Age
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
20
40
60
80
Age

Relative self-efficacy

> Moses lacks clean drinking water. He would like to change this, but he can’t vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

- What proportion of respondents report **less** efficacy than Moses?

```r
mean(china$self < china$moses)
```

```
## [1] 0.562
```

```r
mean(mexico$self < mexico$moses)
```

```
## [1] 0.249
```

Adjust self-reported efficacy

- Use the vignettes to measure the respondent’s **relative** efficacy.
- First, subset to those who rank the vignettes in the correct order:

```
china.sane <- subset(china, alison >= jane & jane >= moses)
mexico.sane <- subset(mexico, alison >= jane & jane >= moses)
```

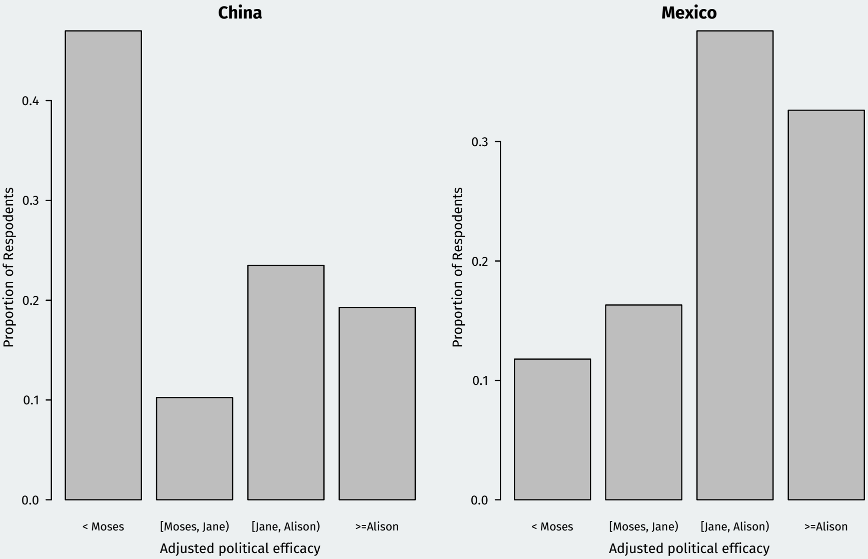
- Now, let’s create new measures with the following values:
  1. if self score $<$ Moses score
  2. if self score $>=$ Moses, but $<$ Jane
  3. if self score $>=$ Jane, but $<$ Alison
  4. if self score $>=$ Alison.

- Creating the adjusted scores:

```
china.sane$self.adj <- 1 +
  (china.sane$self >= china.sane$moses) +
  (china.sane$self >= china.sane$jane) +
  (china.sane$self >= china.sane$alison)


mexico.sane$self.adj <- 1 +
  (mexico.sane$self >= mexico.sane$moses) +
  (mexico.sane$self >= mexico.sane$jane) +
  (mexico.sane$self >= mexico.sane$alison)
```

- R converts `TRUE` to `1` and `FALSE` to `0` when adding.


China
0.0
0.1
0.2
0.3
0.4
Proportion of Respodents
< Moses
[Moses, Jane)
[Jane, Alison)
>=Alison
Adjusted political efficacy
Mexico
0.0
0.1
0.2
0.3
Proportion of Respodents
< Moses
[Moses, Jane)
[Jane, Alison)
>=Alison
Adjusted political efficacy

Wrap up

- Today:
  - Barplots for categorical variables
  - Histograms and boxplots for continuous variables.
- Datacamp Assignment 3:
  - Due by Thursday.
- Homework 2:
  - Will go out today on Canvas/rstudio.cloud.
  - Does having daughters (versus sons) affect a judge's rulings?
  - Get started early!