Gov 51: Boxplots and QQ-plots

Matthew Blackwell

Harvard University

Assassination attempts

- Load the assassination attempts data see the possible attempt results.

```
## see the categories of the results variable
leaders <- read.csv("data/leaders.csv")
lev <- unique(leaders$result)
lev
```

```
##  [1] "not wounded"
##  [2] "dies within a day after the attack"
##  [3] "survives, whether wounded unknown"
##  [4] "wounded lightly"
##  [5] "plot stopped"
##  [6] "hospitalization but no permanent disability"
##  [7] "dies between a day and a week"
##  [8] "dies, timing unknown"
##  [9] "survives but wounded severely"
## [10] "dies between a week and a month"
```

Creating an attempt fatal variable

- Use `ifelse` to create a `fatal` variable.

```
leaders$fatal <- ifelse(leaders$result %in% lev[1:4], 1, 0)

## rate of fatal
head(leaders$fatal)
```

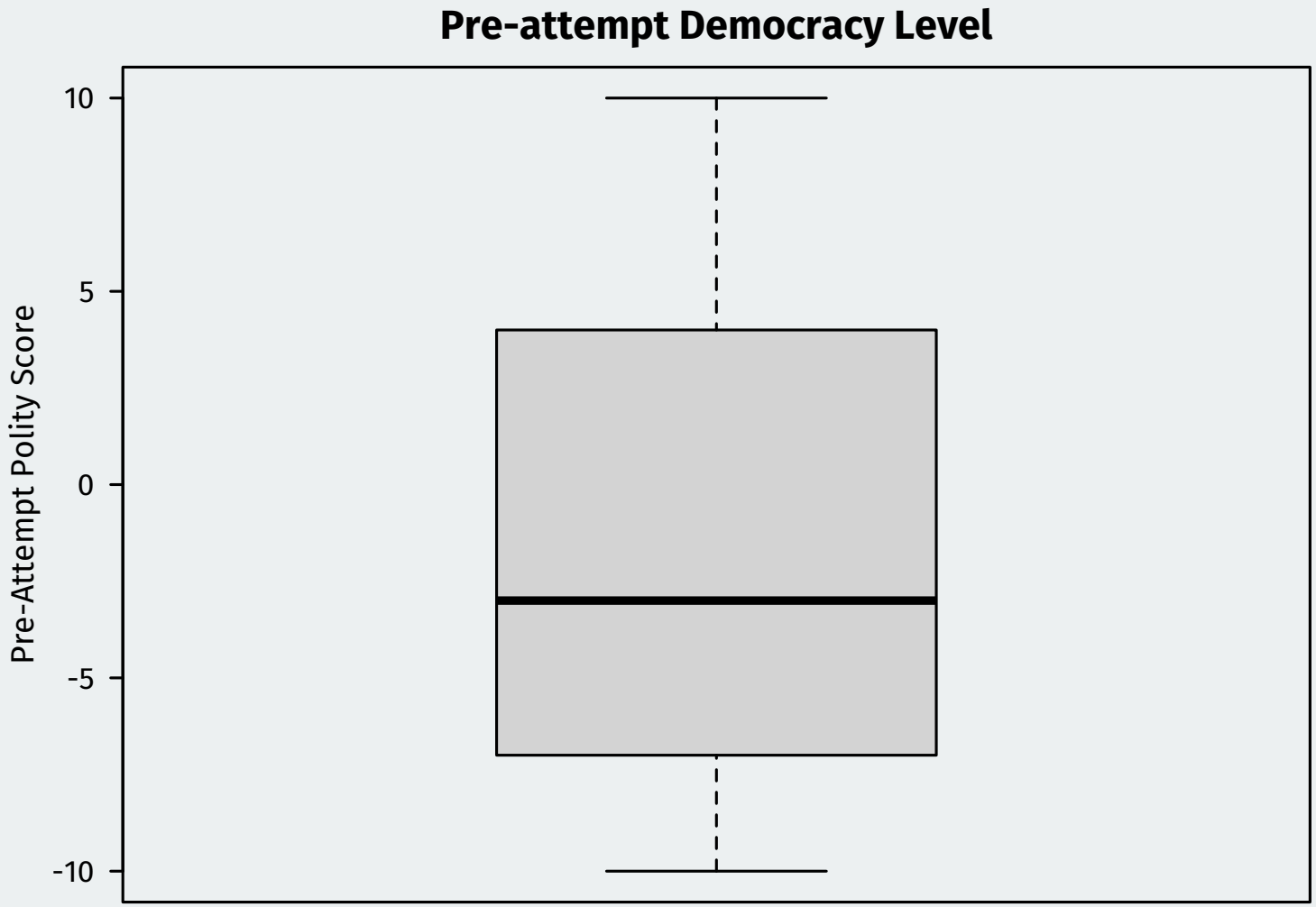
```
## [1] 1 1 1 1 1 1
```

```
mean(leaders$fatal)
```

```
## [1] 0.724
```

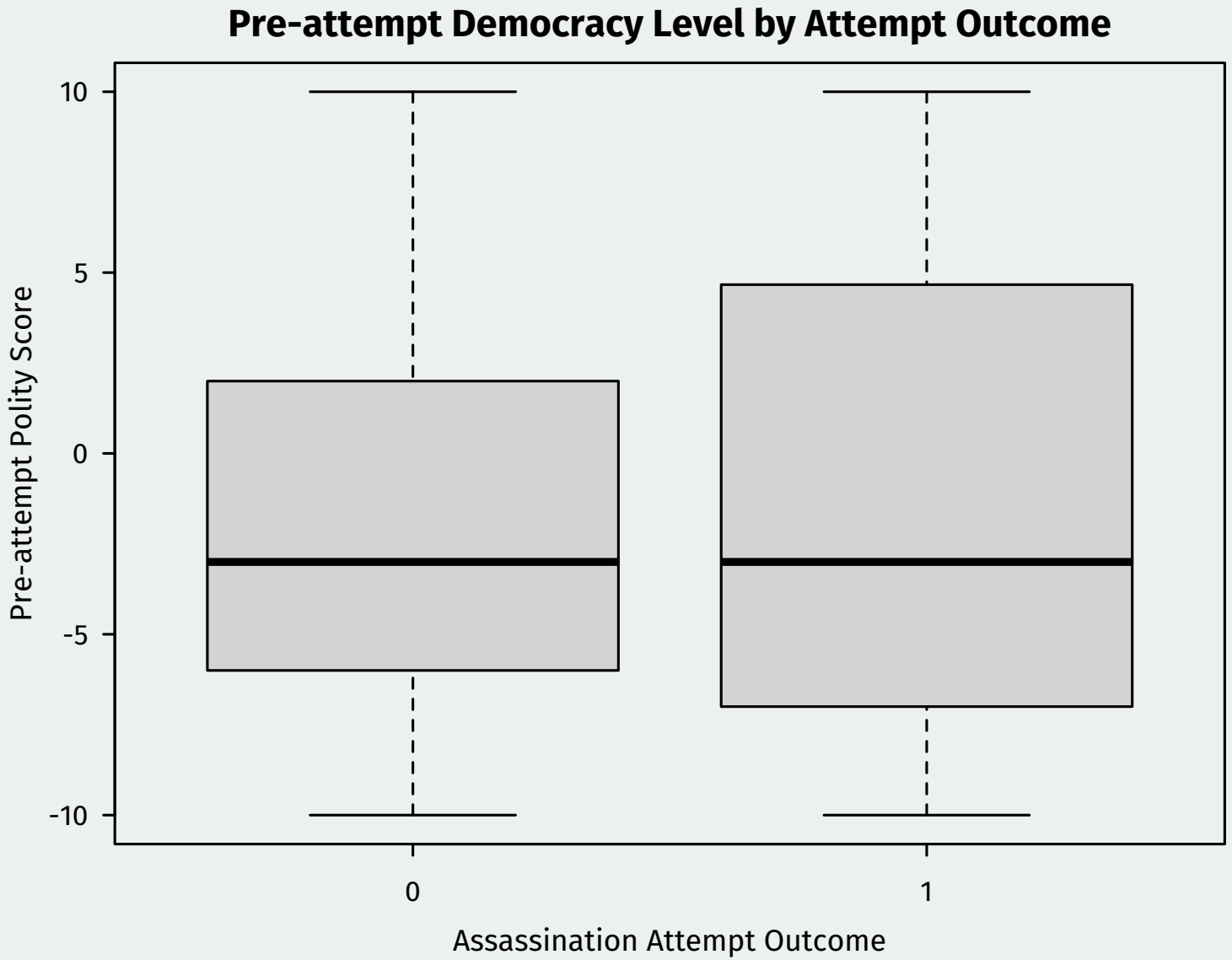
Remember boxplots?

- Boxplots were a tool to help visual continuous data.



Comparing distribution with the boxplot

- What if we want to know how the distribution varies by success?



Boxplot comparisons in R

```
boxplot(politybefore ~ fatal, data = leaders,
        names.arg = c("Survived", "Died"),
        main = "Pre-attempt Democracy Level by Attempt Outcome",
        ylab = "Pre-attempt Polity Score",
        xlab = "Assassination Attempt Outcome")
```

- 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.

Quantile-Quantile Plot

- How do we compare distributions of two variables that are not in the same dataset?
    - Could use boxplots, but it's only a crude summary of the distributions.
- **Quantile-quantile plot (Q-Q plot)**: scatterplot of **quantiles**.
    - (min of X, min of Y)
    - (median of X, median of Y)
    - (25th percentile of X, 25th percentile of Y)
- Intuitions:
    - If distributions are the same $\rightsquigarrow$ all points on a 45-degree line.
    - Points above 45° line $\rightsquigarrow$ y-axis variable has larger value of the quantile.
    - Points below 45° line $\rightsquigarrow$ x-axis variable has larger value of the quantile.
    - Steeper slope than 45° line $\rightsquigarrow$ y-axis variable has more spread.
    - Flatter slope than 45° line $\rightsquigarrow$ x-axis variable has more spread.

QQ-plot example

```
leaders$polity_change <- leaders$polityafter -
  leaders$politybefore
lived <- subset(leaders, fatal == 0)
died <- subset(leaders, fatal == 1)
qqplot(lived$polity_change, died$polity_change,
       xlab = "Change in Polity, Leader Survived",
       ylab = "Change in Polity, Leader Died")
abline(a = 0, b = 1, lty = 3)
```

QQ-plot example