

Bush 631-600: Quantitative Methods

Lecture 3 (09.13.2022): Causality vol. II

Rotem Dvir

The Bush school of Government and Public Policy

Texas A&M University

Fall 2022

What is today's plan?

- ▶ Causality and deriving cause-effect relationship.
- ▶ Limitations of RCTs.
- ▶ Alternative designs: observational studies.
- ▶ Writing documents in R: data and text.
- ▶ Class task I: Using R and R Markdown
- ▶ Descriptive statistics: explore our data.
- ▶ R work: sub-setting data, spread of the data, quartiles, .

Causality

- ▶ Identify causes for outcomes of interest:
 1. Universal health care and better health status among poor.
 2. Drop in president approval during war.
- ▶ Establish causality:

Cause → Effect

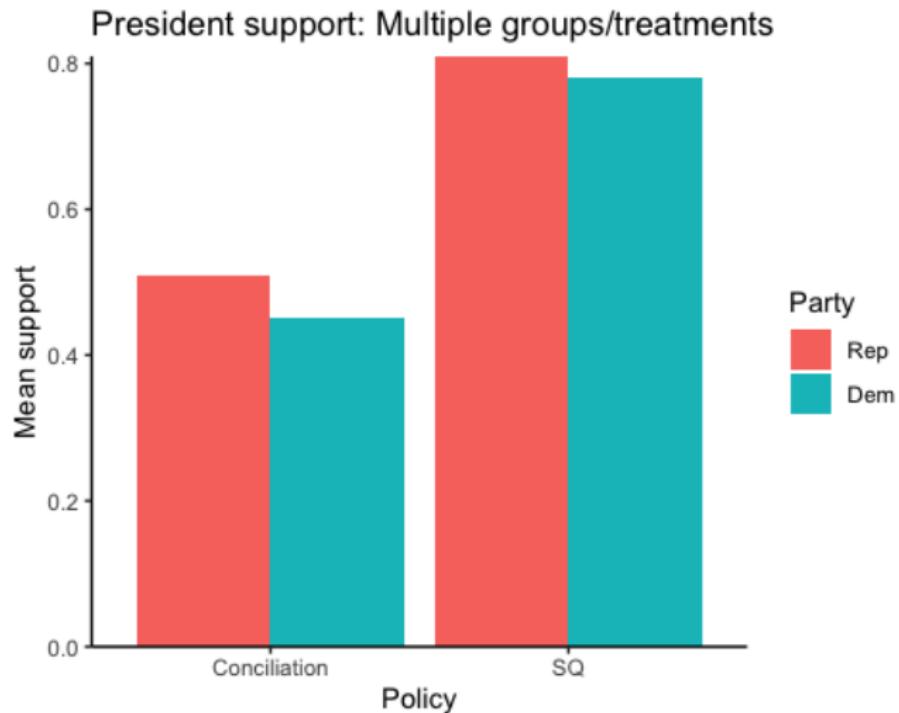
Experimental Research Designs

Mattes and Weeks (2019): FP actions and public opinion

- ▶ Elements of experiment:
 - ▶ Hypothetical scenario.
 - ▶ Adversary: China.
 - ▶ Important FP issue - access to arctic.
 - ▶ Outcome measured: approval of president's actions.
- ▶ Treatments:
 - ▶ Description of factors: leader type, party, policy enacted.
 - ▶ Vary between groups.
 - ▶ Compare outcome variables: approval (1-5 scale), proportion of support

Experimental Research Designs: RCTs

- ▶ Grouping treatments by president party and policy choice



RCT drive policy

- ▶ Behavioral insights & public policy:
 - ▶ Green initiatives: information structure.
 - ▶ Nudges - defaults and opt-out of retirement.
 - ▶ Pay your taxes: UK, 2014
- ▶ Get out to vote campaigns:
 - ▶ Comparing types of promotions (phone, in-person, etc.)

RCTs: Limitations

Ethical:

- ▶ Problematic treatments: manipulate police officers behavior.
- ▶ Deceit.

Logistical:

- ▶ Limited samples: students == elites ?? recruit world leaders ??



The alternative

OBSERVATIONAL STUDIES

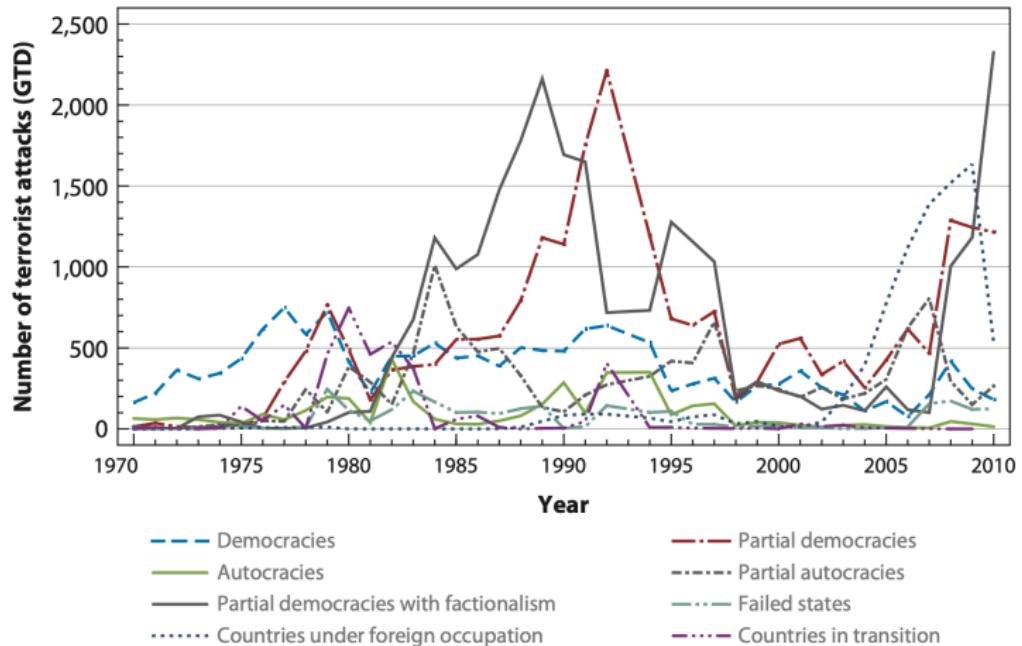
Do democracies experience more terror attacks than other regimes?

How to study?

- ▶ Observe actual events: record terror incidents.
- ▶ Treatment is 'assigned naturally' - countries are either a democracy or non-democracy.
- ▶ Study our collected data: does regime type matter for the frequency of terrorism?

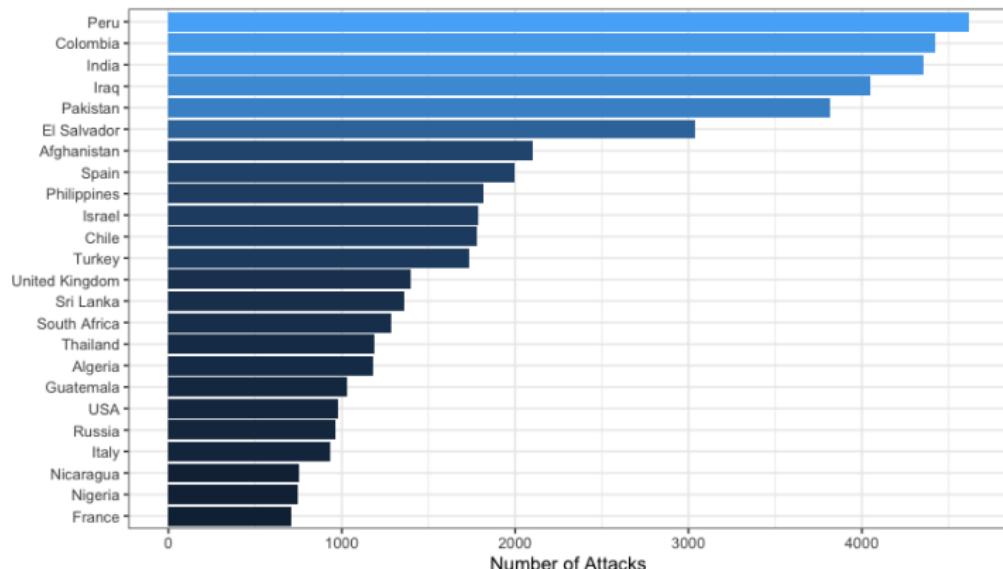
Terrorism and regimes type

The answer?



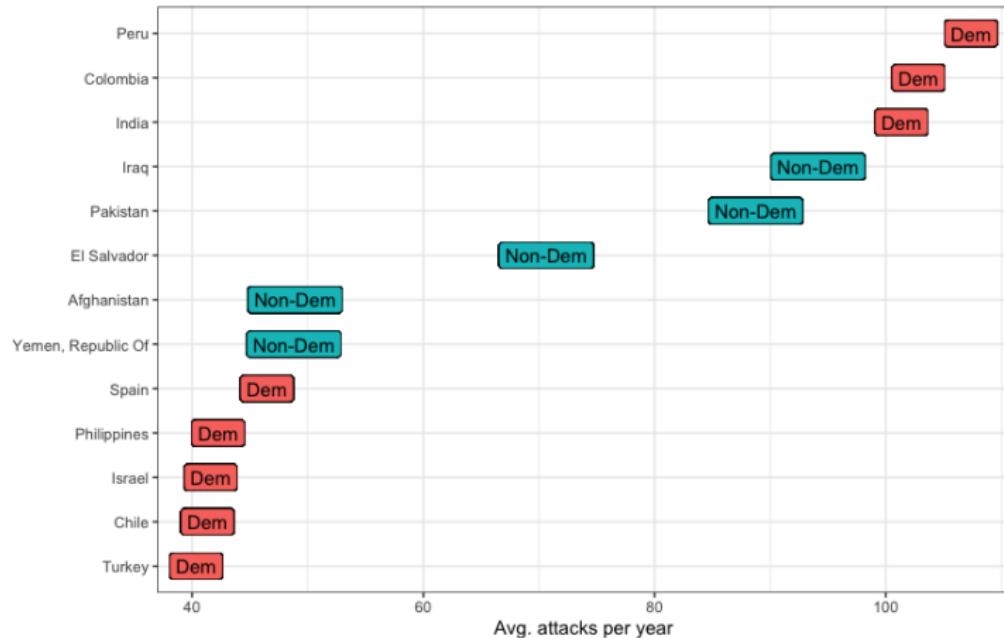
Domestic Terrorism and regimes type

Total terror incidents (1970-2012)



Domestic Terrorism and regimes type

Terror incidents per year and regime types



Domestic Terrorism and regimes type

- ▶ Regime type measure: Polity IV score
- ▶ Three groups: democracies, anocracies, autocracies.

```
## Difference in means between groups  
mean(a1$at, na.rm = T) -  
  mean(a2$at, na.rm = T) # Democracy - Anocracy
```

```
## [1] -3.036781  
mean(a2$at, na.rm = T) -  
  mean(a3$at, na.rm = T) # Anocracy - Autocracy
```

```
## [1] 10.6224  
mean(a1$at, na.rm = T) -  
  mean(a3$at, na.rm = T) # Democracy - Autocracy
```

```
## [1] 7.585618
```

Observational studies: INTA

STUDY LEADERS



Studying leaders

Fuhrmann and Horowitz (2015):

- ▶ Personal background and military technology.
- ▶ Nuclear weapons.

Rebel background → pursue nuclear weapons?

Leaders and nuclear tech

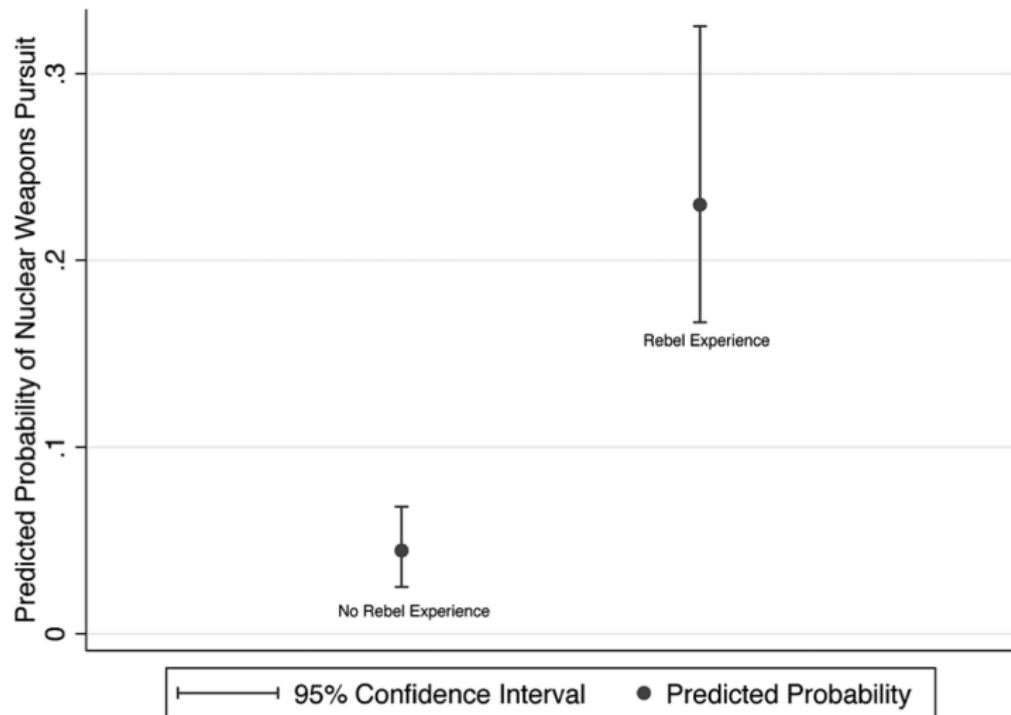
Why?

- ▶ Life experiences shape perceptions.
- ▶ Ensure national independence, discount allies.
- ▶ Underestimate financial and political costs.
- ▶ High risk tolerance.

How?

- ▶ Data on global leaders (1945-2000).
- ▶ 1342 leaders.
- ▶ Data on nuclear proliferation programs.
- ▶ Indicator for rebel participation.

Leaders and nuclear tech



Working with observational data

Large-n data:

```
dim(mydata)
```

```
## [1] 8852    76
```

Time-series cross-sectional data (TSCS)

code OW numeric country code	idacr COW alpha country code	year Year	leadid30 LEAD Leader ID	leadername Leader Name	startdate LEAD Start Date	inday Leader Entry Day	l
2	USA	1995	A2.9-73	Clinton	1993-01-20		20
2	USA	1996	A2.9-73	Clinton	1993-01-20		20
2	USA	1997	A2.9-73	Clinton	1993-01-20		20
2	USA	1998	A2.9-73	Clinton	1993-01-20		20
2	USA	1999	A2.9-73	Clinton	1993-01-20		20
2	USA	2000	A2.9-73	Clinton	1993-01-20		20
20	CAN	1945	A2.9-118	King	1935-10-23		23
20	CAN	1946	A2.9-118	King	1935-10-23		23
20	CAN	1947	A2.9-118	King	1935-10-23		23
20	CAN	1948	A2.9-118	King	1935-10-23		23

Leaders data

Main variables we'll use:

```
# rebel experience: yes/no (coded 1/0)
table(rebels = mydata$rebel)
```

```
## rebels
##     0      1
## 5089 3743
```

```
# revolutionary leader: yes/no (coded 1/0)
table(rev_leaders = mydata$revolutionaryleader)
```

```
## rev_leaders
##     0      1
## 6816 1041
```

```
# pursue nuclear tech: yes/no (coded 1/0)
table(pursue_nukes = mydata$pursuit, exclude = NULL)
```

```
## pursue_nukes
##     0      1 <NA>
## 8257  225   370
```

Creating treatment & control groups

```
# subsets: rebel experience yes/no
lead_rebels <- subset(mydata, subset = (rebel == 1))
lead_norebels <- subset(mydata, subset = (rebel == 0))

dim(lead_rebels)

## [1] 3743    76

# subsets: revolutionary leaders yes/no
rev_leader <- subset(mydata, subset = (revolutionyleader == 1))
rev_noleader <- subset(mydata, subset = (revolutionyleader == 0))

dim(rev_leader)

## [1] 1041    76
```

Differnce-in-means

Does rebel experience matter?

```
# pursuit of nukes tech: diff-in-means (rebels - no rebels)
mean(lead_rebels$pursuit, na.rm = TRUE) -
  mean(lead_norebels$pursuit, na.rm = TRUE)

## [1] 0.0376728

# pursuit of nukes tech: diff-in-means (rev. leaders - no rev. leaders)
mean(rev_leader$pursuit, na.rm = TRUE) -
  mean(rev_noleader$pursuit, na.rm = TRUE)

## [1] 0.06781106
```

Differnce-in-means

Alternative measures: existing nuclear arsenals

```
# existing bomb program: yes/no (coded 1/0)
table(bomb_program = mydata$bombprgm)
```

```
## bomb_program
##     0      1
## 8258  594
```

```
# pursuit of nukes tech: diff-in-means (rebels - no rebels)
mean(lead_rebels$bombprgm, na.rm = TRUE) -
  mean(lead_norebels$bombprgm, na.rm = TRUE)
```

```
## [1] 0.02515995
```

```
# pursuit of nukes tech: diff-in-means (rev. leaders - no rev. leaders)
mean(rev_leader$bombprgm, na.rm = TRUE) -
  mean(rev_noleader$bombprgm, na.rm = TRUE)
```

```
## [1] 0.04400943
```

Why does it matter?

Policy Lessons??



Observational studies

Why study large-N data?

- ▶ Policy questions, real (sometime rare) events.
- ▶ Japan - Russia war (1905) \neq Gulf war (1991), right?

What does studying large-N means?

- ▶ Collect lots of observations.
- ▶ Apply stats methods to evaluate potential patterns in data.

So, Why?

Universe of cases:

- ▶ Better sense of phenomenon.
- ▶ Large variation.
- ▶ Identify *important cases*.

Rebels and Nuclear Weapons



So, Why?

Construct general theory of state behavior

- ▶ Social science overarching goal.
- ▶ One case? tough for general argument.
- ▶ Theory applies across time and space.

Vietnam (1965)



Afghanistan (2001)



Observational Studies

- ▶ Guiding assumption:

Treatment group (rebel leaders) = control group (no rebels)

Is it?



Kadar (Hungary):
1956-1988.
Leader of
Hungarian
rebellion (1956)

Did not pursue
Nuclear weapons



Bhutto (Pakistan):
1972-1977.

No rebel
background

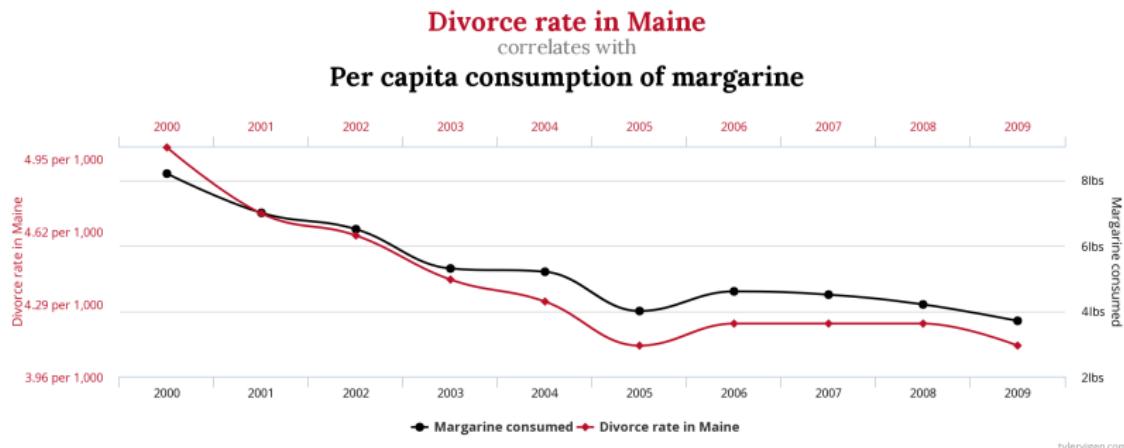
Pursued Nuclear
weapons

Confounders

- ▶ Pre-treatment variables → treatment & outcome.
- ▶ Realized 'before' treatment → who 'receive' treatment.
- ▶ **Selection bias:** cannot assign who gets treatment (assign rebel experience).
- ▶ Unobserved differences → is it rebel background.
- ▶ More examples:
 1. Terrorism and regime type (civil strife).
 2. Economic growth (international trade).
 3. Sanctions effective? (corrupt leader).
 4. Prevail in conflict - democracy (or military capacity).

Inference problems

Association does not imply causation



More? ([SpuriousCors_Link](#))

Our confounders

- ▶ **Superpower alliance:** no need to pursue nuclear weapons.
- ▶ Hungary & USSR: Kadar did not pursue nuclear weapons.
- ▶ UK & US: Churchill and Atlee pursue nuclear weapons.
- ▶ West Germany & US: Kohl did not pursue nuclear weapons.
- ▶ Both rebels and non-rebels pursue nuclear weapons.

Bias our causal explanation!!!

Confounders

- ▶ Ever present problem of observational studies.
- ▶ What do we do?
 - ▶ Ensure correct cases identification.
 - ▶ Statistical ‘control’ of confounding factors (we’ll get to it).
- ▶ **Sub-classification:**
 - ▶ Minimize similarities b-w treatment & control groups.
 - ▶ subsets of shared pre-treatment values.
 - ▶ Comparing main factor within subsets.

Sub-classification in R

- ▶ `prop.table()`: tabulate proportions of different levels of factor variables.

```
# Confounders: alliance with a superpower  
# Leaders with rebel experience  
prop.table(table(rebel_allies = lead_rebels$spally))
```

```
## rebel_allies  
##      0      1  
## 0.670247 0.329753
```

```
# Confounders: alliance with a superpower  
# Leaders with no rebel experience  
prop.table(table(no_rebels_allies = lead_norebels$spally))  
  
## no_rebels_allies  
##      0      1  
## 0.5848161 0.4151839
```

Subsetting alliance and rebel leaders

```
# subsets: rebel/non-rebel leaders and superpower alliance
rebel_ally <- subset(lead_rebels, subset = (spally == 1))
norebel_ally <- subset(lead_norebels, subset = (spally == 1))

# diff-in-means in nuclear weapons pursuit
mean(rebel_ally$pursuit, na.rm = TRUE) -
  mean(norebel_ally$pursuit, na.rm = TRUE)

## [1] 0.0231065
```

- ▶ Fuhrmann and Horowitz (2015):
 - ▶ Countries with no superpower alliance → 4.6 more likely to pursue nukes.
 - ▶ Other confounders for nuclear tech: nuclear cooperation agreement, rivalry, military disputes.

More research designs

BEFORE AND AFTER DESIGN

- ▶ *Longitudinal / Panel data*
- ▶ Collecting time series data.
- ▶ Time-related information for treatment and control groups.
- ▶ Better comparison of groups.

Before and after design: QSS textbook

- ▶ Topic: changes to minimum wage and levels of full-time employment.
- ▶ Method: compare fast food restaurants (NJ - PA).
- ▶ Longitudinal design: compare **within** NJ group
- ▶ Before and after (change in minimum wage).
- ▶ Result: $\text{diff-in-means} = 0.023$ (2.3% increase in employment).
- ▶ Benefit: control all NJ confounders.
- ▶ Cost: time trend factor may bias results.

Before and after design: Rebel leaders

- ▶ Slight diversion: pursue nuclear weapons over leader tenure
- ▶ Compare: year 1 vs. subsequent years

```
# subsets: rebel leaders, first year and subsequent years
reb_one <- subset(lead_rebels, subset = (nonpuryrs == 0))
reb_after <- subset(lead_rebels, subset = (nonpuryrs > 0))

# diff-in-means: nuclear weapons pursuit over time
mean(reb_one$pursuit, na.rm = T) -
  mean(reb_after$pursuit, na.rm = T)

## [1] 0.2263734
```

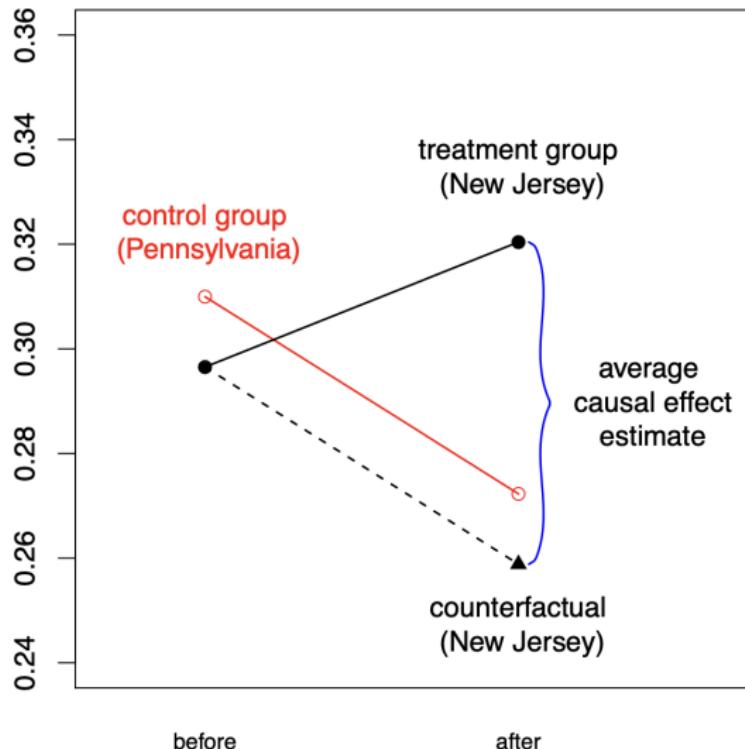
More research designs

DIFFERENCE IN DIFFERENCE DESIGN

- ▶ Extends the before-and-after design.
- ▶ Control for time trends (effects).
- ▶ Using control group before-and-after to infer on treatment group (the counterfactual).

Diff-in-diff design

- ▶ Minimum wage and full-time employment



Diff-in-diff design: QSS Textbook

- ▶ Quantity of interest:
 - ▶ *SATT*: Sample Average Treatment effect for the Treated.
 - ▶ Difference b-w observed outcome and counterfactual (no increase in NJ)

$$\text{DiD estimate} = \underbrace{\left(\bar{Y}_{\text{treated}}^{\text{after}} - \bar{Y}_{\text{treated}}^{\text{before}} \right)}_{\text{difference for the treatment group}} - \underbrace{\left(\bar{Y}_{\text{control}}^{\text{after}} - \bar{Y}_{\text{control}}^{\text{before}} \right)}_{\text{difference for the control group}} .$$

Research Designs

- ▶ Cross-sectional comparison:
 - ▶ Compare treated units with control units after treatment.
 - ▶ Assumption: treated and control groups are comparable.
 - ▶ Problems of confounders.
- ▶ Before-and-after comparison:
 - ▶ Compare the same units before and after treatment.
 - ▶ Assumption: no time-varying confounding.
- ▶ Differences-in-differences comparison:
 - ▶ Assumption: similar trend assumptions.
 - ▶ Design accounts for unit-specific and time-varying confounders.

Observational studies

Internal validity → weak:

- ▶ Pre-treatment variables.
- ▶ Can we show the effect of 'our' treatment?

External validity → strong:

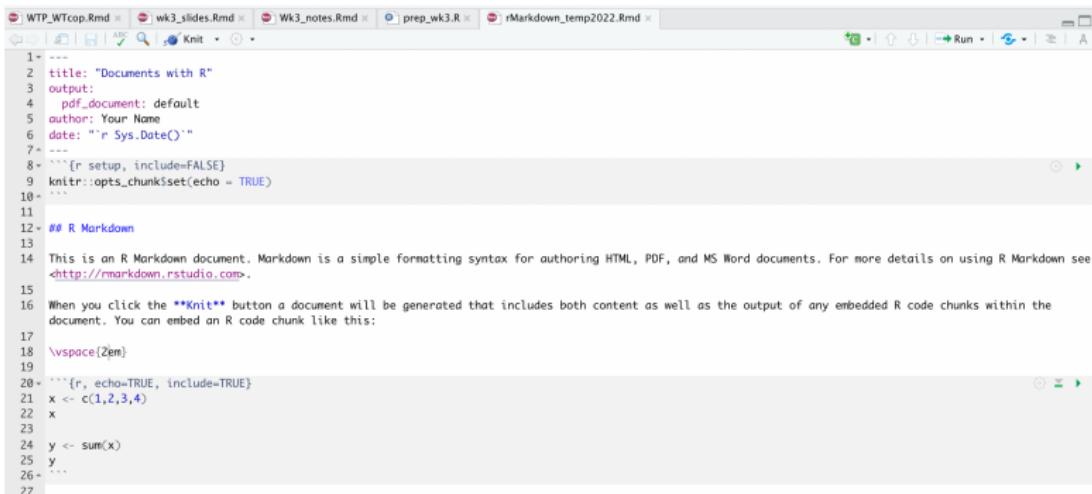
- ▶ Larger samples.
- ▶ Time series data.
- ▶ Elites, politicians can be easily collected.
- ▶ Results can be generalized.

R logistics

- ▶ Writing reports and professional documents.
- ▶ R Markdown: mix of text and code to analyze data and produce content.
- ▶ Reproducible, automation, more efficient.
- ▶ Create PDF, html (and yes, also Word).

R Markdown

Replacing the R script



The screenshot shows the RStudio interface with several tabs open at the top: WTP_WTcap.Rmd, wk3_slides.Rmd, Wk3_notes.Rmd, prep_wk3.R, and rMarkdown_temp2022.Rmd. The main pane displays an R Markdown document with the following content:

```
1<---
2  title: "Documents with R"
3  output:
4    pdf_document: default
5  author: Your Name
6  date: "r Sys.Date()"
7<---
8<-- {r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10<---
11
12<# R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see
<a href="http://rmarkdown.rstudio.com">http://rmarkdown.rstudio.com.
15
16 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the
document. You can embed an R code chunk like this:
17
18 \vspace{2em}
19
20<-- {r, echo=TRUE, include=TRUE}
21 x <- c(1,2,3,4)
22 x
23
24 y <- sum(x)
25 y
26<--
27
```

R Markdown

The output

Documents with R

Your Name

2022-01-28

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
x <- c(1,2,3,4)
x

## [1] 1 2 3 4
y <- sum(x)
y

## [1] 10
```

R Markdown Intro

- ▶ Use the template for any document you write.
- ▶ Use the instructions file.

- ▶ Change in template:
 - ▶ Your name, date and project title.
 - ▶ Add data to work with (more packages if needed).
 - ▶ Main text - your project.
 - ▶ R code: change as needed.

- ▶ Do not change:
 - ▶ Setting in YAML & R code chunks.

Learn from data

Descriptive Statistics

- ▶ Cross-sectional comparison → average outcome of interest.
- ▶ General findings: 4.8% of all rebel leaders (1945-2000) pursue nuclear weapons.
- ▶ More? other numerical summaries (min, max values, range).
- ▶ *Quantiles*: divide data to groups based on magnitude.
- ▶ **Median**: the middle value when the data is divided to two groups.

The median of a variable x is defined as:

$$\text{median} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even,} \end{cases}$$

Rebel leaders data

```
# pursuit of nuclear weapons: all leaders  
median(mydata$pursuit, na.rm = TRUE)
```

```
## [1] 0
```

```
# pursuit of nuclear weapons: rebel leaders  
median(lead_rebels$pursuit, na.rm = TRUE)
```

```
## [1] 0
```

```
# Economic growth measures: GDP per capita  
median(mydata$gdpcap, na.rm = TRUE)
```

```
## [1] 3612
```

```
# Involvement in MID: 5 year average  
range(mydata$disputes, na.rm = TRUE)
```

```
## [1] 0.00 17.75
```

Descriptive Stats

THE MEAN - MEDIAN DEBATE

- ▶ Both describe center of distribution (data spread).
- ▶ Not always equal.

```
# Economic growth measures: GDP per capita  
median(mydata$gdpcap, na.rm = TRUE)
```

```
## [1] 3612
```

```
# Economic growth measures: GDP per capita  
mean(mydata$gdpcap, na.rm = TRUE)
```

```
## [1] 5808.161
```

The mean - median debate

- ▶ Why not equal?

```
v1 <- c(100,200,300)
mean(v1)
```

```
## [1] 200
median(v1)
```

```
## [1] 200
```

- ▶ mean → sensitive to *outliers* - extreme values.

```
v2<- c(100,200,4000)
mean(v2)
```

```
## [1] 1433.333
median(v2)
```

```
## [1] 200
```

Descriptive Stats

- ▶ **Quartiles:** more complete description of data.

```
# Quartiles and summary function  
summary(lead_rebels$gdpcap)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.    NA's  
##     281    1197   2476    3937    5026   41762     454
```

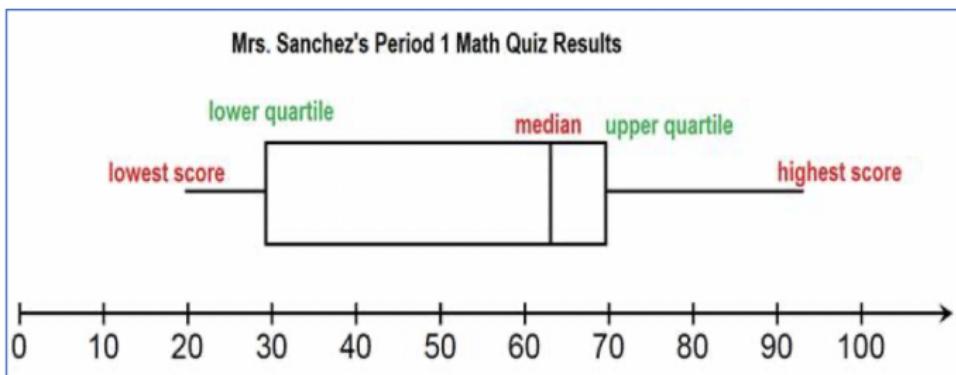
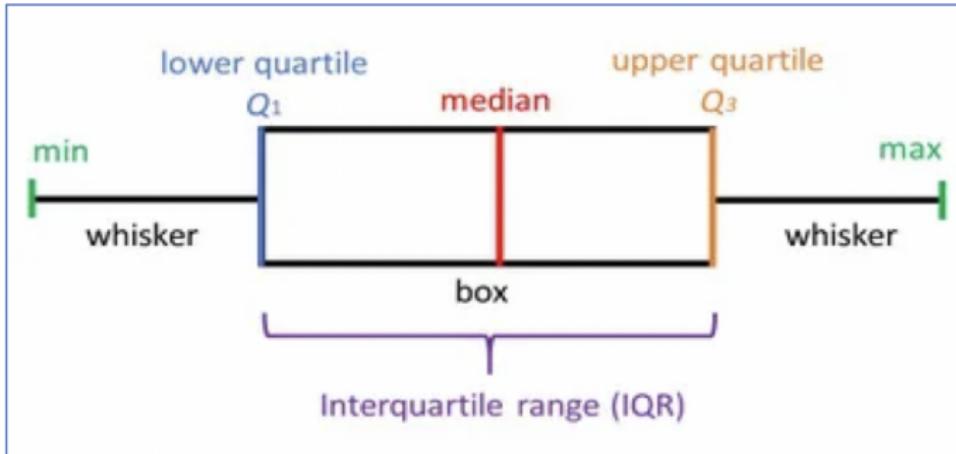
- ▶ **IQR:** range that contains 50% of the data (spread of distribution)

```
# IQR function: openness (economic measure)  
IQR(lead_rebels$openness, na.rm = TRUE)
```

```
## [1] 38.01  
IQR(lead_norebels$openness, na.rm = T)
```

```
## [1] 50.2475
```

Descriptive Stats



Descriptive Stats

- ▶ Other quantiles:

- ▶ terciles (3 groups)
- ▶ quintiles (5 groups)
- ▶ deciles (10 groups)
- ▶ percentiles (100 groups)

```
# deciles (10 groups) for dispute involvement
# compare rebels and non-rebels

quantile(lead_rebels$disputes, probs = seq(from = 0,
                                             to = 1, by = 0.1), na.rm = T)

##      0%     10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
##  0.00   0.00   0.00   0.00   0.20   0.40   0.60   0.80   1.40   2.20  17.75

quantile(lead_norebels$disputes, probs = seq(from = 0,
                                               to = 1, by = 0.1), na.rm = T)

##      0%     10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
##  0.0    0.0    0.0    0.0    0.0    0.2    0.4    0.4    0.8    1.4    9.4
```

Spread of data

- ▶ **RMS (Root Mean Square)**: magnitude of each observation.

$$RMS = \sqrt{\frac{entry_1^2 + entry_2^2 + entry_3^2 + \dots}{\sum entries}} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

- **SD (Standard Deviation)**: average deviation of each data point from mean ('distance' of points from average).

$$SD = \sqrt{\frac{(entry1 - mean)^2 + (entry2 - mean)^2 + \dots}{\sum entries}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Spread of rebel leaders data

```
# compare mean pursuit of nuclear weapons  
sd(lead_rebels$pursuit, na.rm = TRUE)
```

```
## [1] 0.2141889  
sd(lead_norebels$pursuit, na.rm = TRUE)
```

```
## [1] 0.1020045  
# compare mean dispute involvement  
sd(lead_rebels$disputes, na.rm = TRUE)
```

```
## [1] 1.516794  
sd(lead_norebels$disputes, na.rm = TRUE)
```

```
## [1] 1.006146
```

Wrapping up week 3

Causality vol. II:

- ▶ Assessing causality with observational studies.
- ▶ The problem of confounding bias and pre-treatment variables.
- ▶ Designs: before-and-after; diff-in-diff.
- ▶ Descriptive stats: median, quartiles, RMS, SD.
- ▶ R work: `prop.table()`, `subset()`, `median`, `summary`, IQR, SD.