Poli 5D Social Science Data Analytics
Introduction

Shane Xinyang Xuan
ShaneXuan.com

January 11, 2017

Contact Information

Shane Xinyang Xuan

xxuan@ucsd.edu

The teaching staff is a team!

Professor Roberts	M	1600-1800 (SSB 299)
Jason Bigenho	Th	1000-1200 (Econ 116)
Shane Xuan	Th	1200-1400 (SSB 332)

Supplemental Materials

UCLA STATA starter kit

http://www.ats.ucla.edu/stat/stata/sk/

Princeton data analysis

http://dss.princeton.edu/training/

Housekeeping Notes (1)

– Section attendance is mandatory. We take attendance for every single section meeting.
– I have zero discretion on: late assignments, exam rescheduling, re-grading, section absence, ... Professor Roberts is going to make the final call.

Housekeeping Notes (1)

- Section attendance is mandatory. We take attendance for every single section meeting.
- I have zero discretion on: late assignments, exam rescheduling, re-grading, section absence, ... Professor Roberts is going to make the final call.
- What we care is if you try your best to learn, and if you gain something useful from Poli 5D.
- This is a difficult class, and it is completely acceptable to get lost at some point during the quarter. If that happens, please let me know.

Housekeeping Notes (1)

- Section attendance is mandatory. We take attendance for every single section meeting.
- I have zero discretion on: late assignments, exam rescheduling, re-grading, section absence, ... Professor Roberts is going to make the final call.
- What we care is if you try your best to learn, and if you gain something useful from Poli 5D.
- This is a difficult class, and it is completely acceptable to get lost at some point during the quarter. If that happens, please let me know.
- I post my slides and other resources on my website (ShaneXuan.com). In case you miss a section, check the "Teaching" page first. If you do not find the things you are looking for, please email me to let me know.
- Please give me 24–48 hours to get back to you. If your question is urgent, post it on the TritonEd course website.

One last thing, and you would appreciate knowing this:
Syllabus is your best friend.

One last thing, and you would appreciate knowing this:
Syllabus is your best friend.

– Office hours & locations

– Grading rubric

– Required readings for lectures (and exams)

– Late homework policy

– Academic integrity

  ⋮

Intuition

We are going to use an example from Daniel Kaplan's *Statistical Modeling: A Fresh Approach*. Here is a collection of 103 sea shells:

Intuition
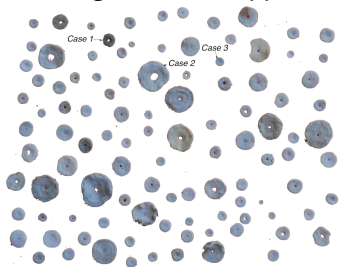
We are going to use an example from Daniel Kaplan's *Statistical Modeling: A Fresh Approach*. Here is a collection of 103 sea shells:
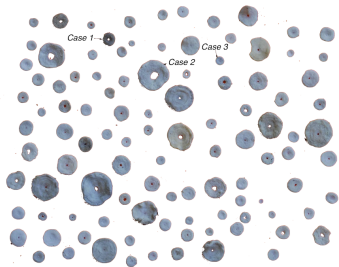
The data frame looks like:

| Case	diameter	weight	color	hole
1	4.3 mm	0.010 mg	dark	medium
2	12.0 mm	0.050 mg	light	very large
3	3.8 mm	0.005 mg	light	none

and so on — there are 103 cases altogether

Variable Types

- Nominal (categorical)

  i.e. Hillary, Donald, Gary, Jill

- Ordinal (can rank)

  i.e. strongly agree > agree > neutral > disagree > strongly disagree

- Interval (different by how much?)

  i.e. grade in school, happiness index, election fraud index

Figure: Hierarchy of measurement levels (Trochim & Donnelly 2006)
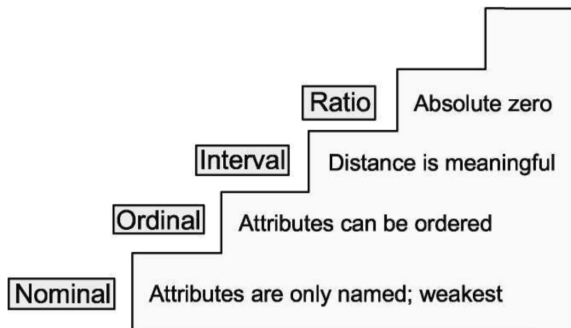
Terminology

– Longitudinal data track the same sample at different points in time
  – Marry-2002
  – Marry-2003
    ⋮
  – Marry-2008
  – Jake-2002
    ⋮
  – Jake-2008
    ⋮

Terminology

- Longitudinal data track the same sample at different points in time
  - Marry-2002
  - Marry-2003
    ⋮
  - Marry-2008
  - Jake-2002
    ⋮
  - Jake-2008
    ⋮
- Cross sectional data observe different subjects at the same point of time

Terminology

– Longitudinal data track the same sample at different points in time
  – Marry-2002
  – Marry-2003
    ⋮
  – Marry-2008
  – Jake-2002
    ⋮
  – Jake-2008
    ⋮
– Cross sectional data observe different subjects at the same point of time
– Unit of Analysis: Entities from which measurements are taken

Terminology

- Longitudinal data track the same sample at different points in time
  - Marry-2002
  - Marry-2003
    .
    .
    .
  - Marry-2008
  - Jake-2002
    .
    .
    .
  - Jake-2008
    .
    .
    .
- Cross sectional data observe different subjects at the same point of time
- Unit of Analysis: Entities from which measurements are taken
  - ▶ Country-year
  - ▶ Student-quarter
  - ▶ State-month

Terminology

– Longitudinal data track the same sample at different points in time
  - Marry-2002
  - Marry-2003
    $\vdots$
  - Marry-2008
  - Jake-2002
    $\vdots$
  - Jake-2008
    $\vdots$
– Cross sectional data observe different subjects at the same point of time
– Unit of Analysis: Entities from which measurements are taken
  - Country-year
  - Student-quarter
  - State-month
– Number of cases: Total units of analysis

Intro to Excel

- Shortcuts
  - Quickly move to bottom (⌘ ↓)
  - Quickly move to right (⌘ →)
  - Select column (control + space)
  - Select row (shift + space)
  - Change data into table (control + T)
- Practice: Variable type
- Note on missing data

Intro to Excel (2)

– Caution: Create a new column while generating variables for the sake of replication

– Scenario 1: Make negative values to NA's

    ▸ Formulas ⇝ Logical ⇝ IF

    ▸

Intro to Excel (2)

– Scenario 2: "4␣never " ⇝ "4"

> ▶
>
> **FIND**
>
> **find_text** = " "
>
> | " " |
>
> **within_text** = "5 Never"
>
> | A2 |
>
> **start_num** = number
>
> | |

> ▶
>
> **LEFT**
>
> **text** = "5 Never"
>
> | A2 |
>
> **num_chars** = 2
>
> | D2 |

– Scenario 3: Special paste (paste value)

See you next week!

Have a good weekend and I will see you next week!

Poli 5D, to many of you, is not an easy class. It requires some mathematical and computer science background, and a lot of hard work. We want you to try your best. If you get lost at any time, please contact me (xxuan@ucsd.edu). I prefer meeting with you during my office hours (Thursday 12-2 pm), but please email me and set up an appointment if you absolutely cannot make it to my office hours.