Understanding Large-N Analyses

Design Political Research: Week 15

Yue Hu

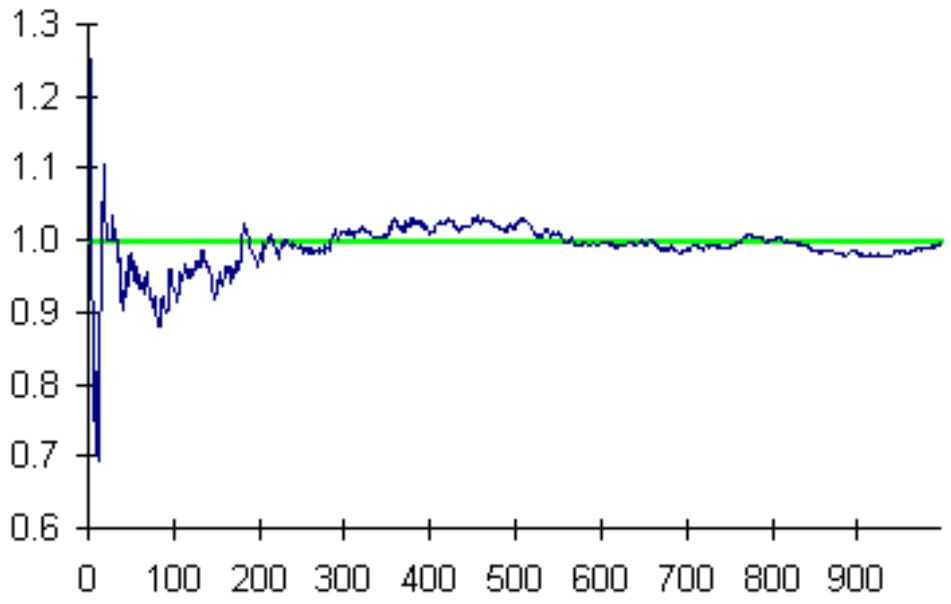
Welcome to the large-N world!

How large is a Large-N N?

Let's revisit the toss of coin:

- Tossing once? Head or Tail?

Tossing 1 million times? How many heads?

- 

Law of large numbers

As the number of experiments (sample) increases, the ratio of outcomes will converge to the theoretical (population) average.

- Rule of thumb: > 100

How to analyze large-N data?

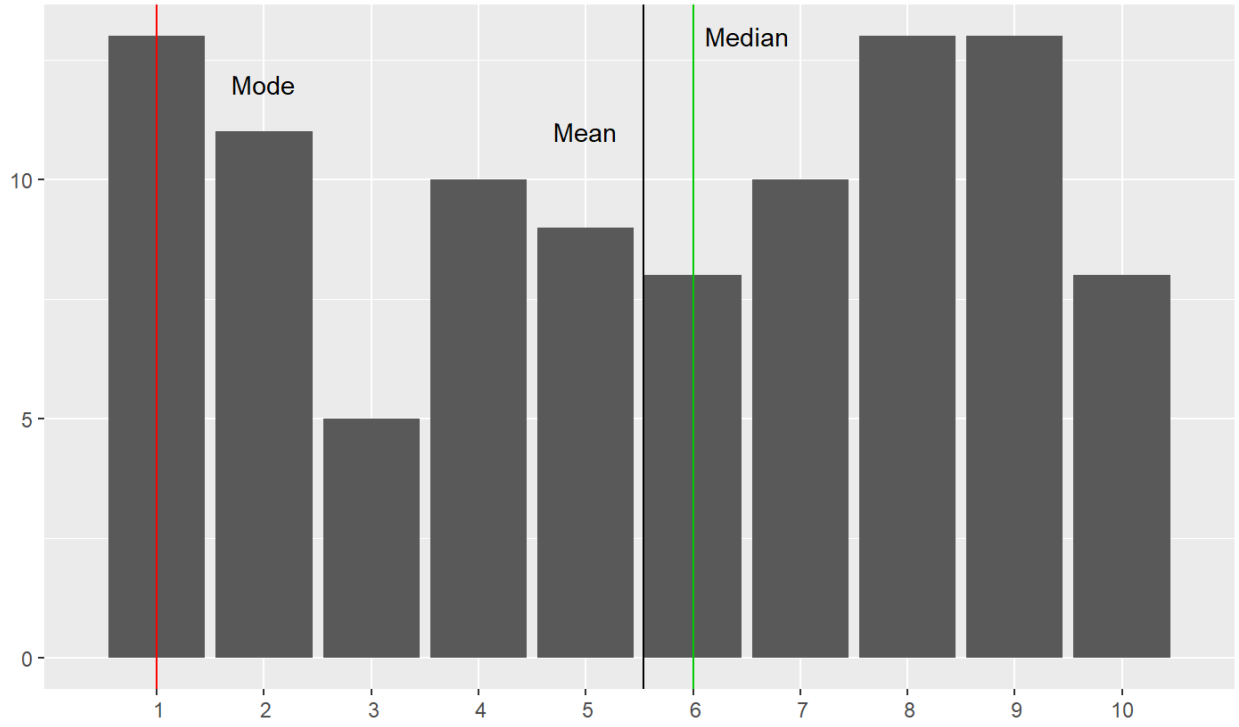
- Univariate analysis
- Bivariate analysis
- Multivariate analysis

Univariate analysis

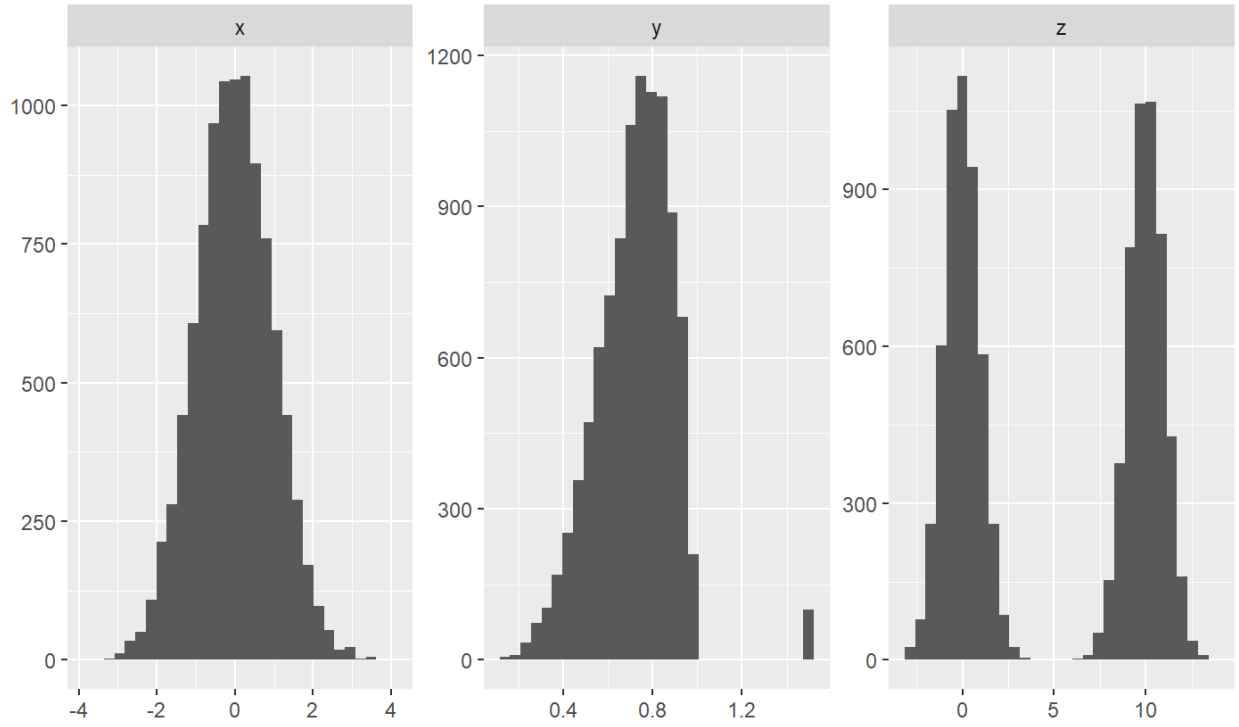
How can we describe a variable?

- Creating the expectation of a variable
  - Given a list: $(1, 1, 1, 2, 3, 3, 4)$
  - Mean: $\frac{1+1+1+2+3+3+4}{7} = \frac{15}{7} \approx 2.143.$
  - Median: 1, 1, 1, **2**, 3, 3, 4
  - Mode: three 1s, one 2, two 3s, and one 4.

Example in large-N data

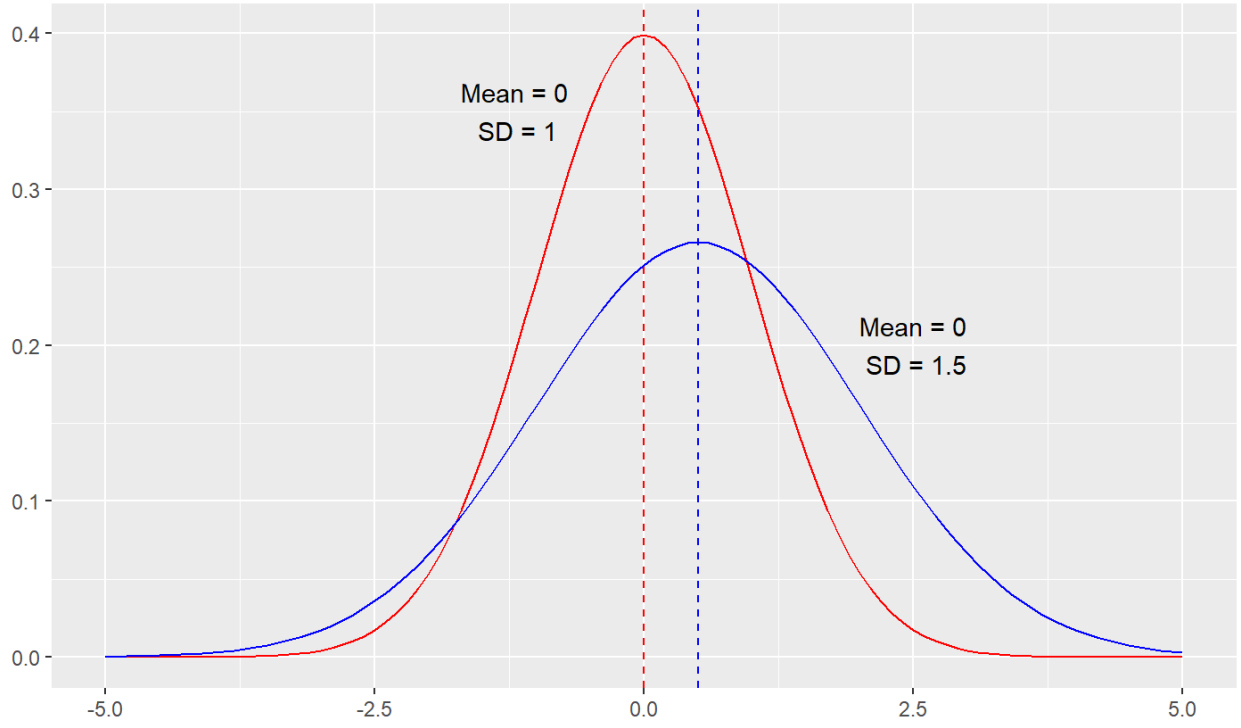


Which one should we choose?

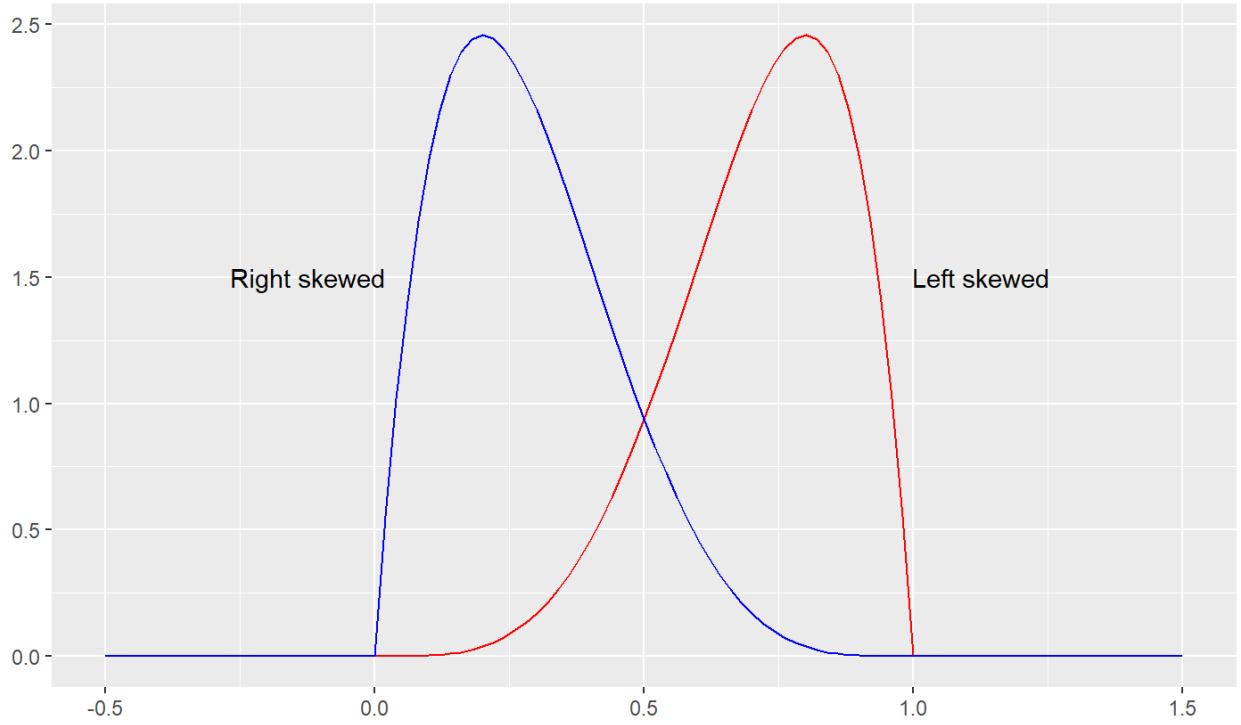


Moments of a variable

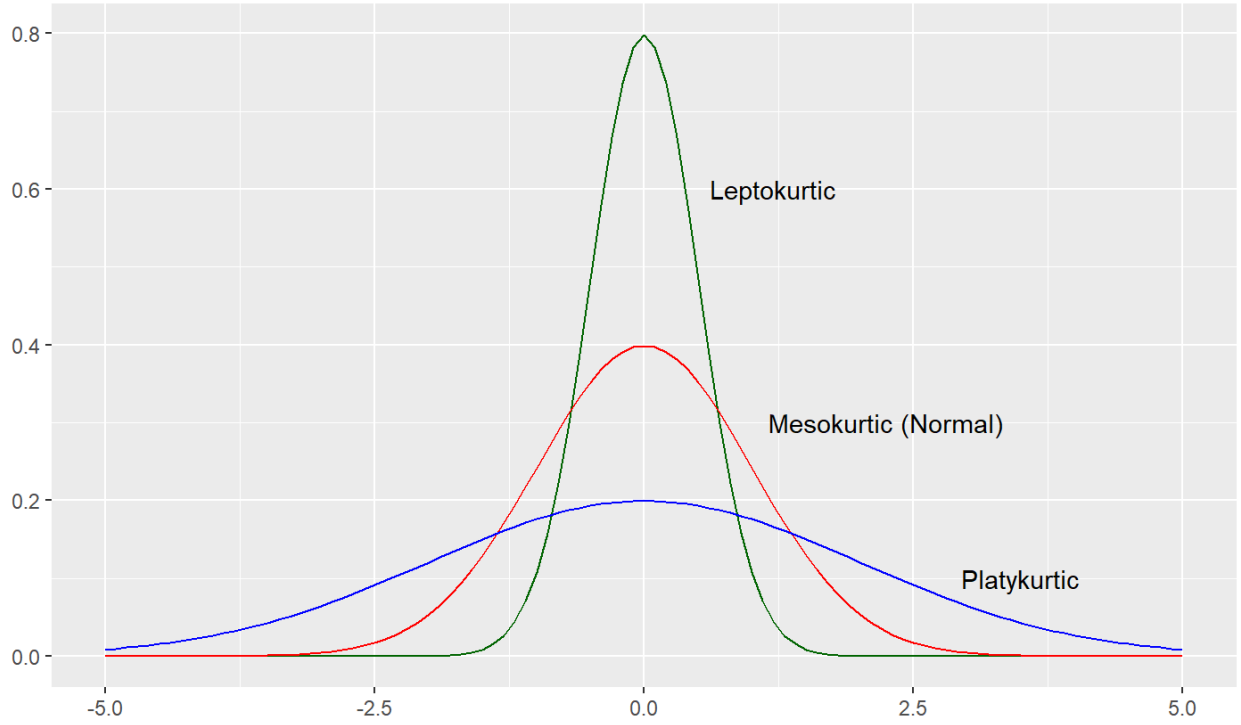
Mean: $\mu_x = \frac{\sum X_i}{N}$; Variance: $\sigma_x^2 = E[(X - \mu_x)^2]$.



Skewness: $\gamma_x = E[(\frac{X-\mu_x}{\sigma})^3]$.



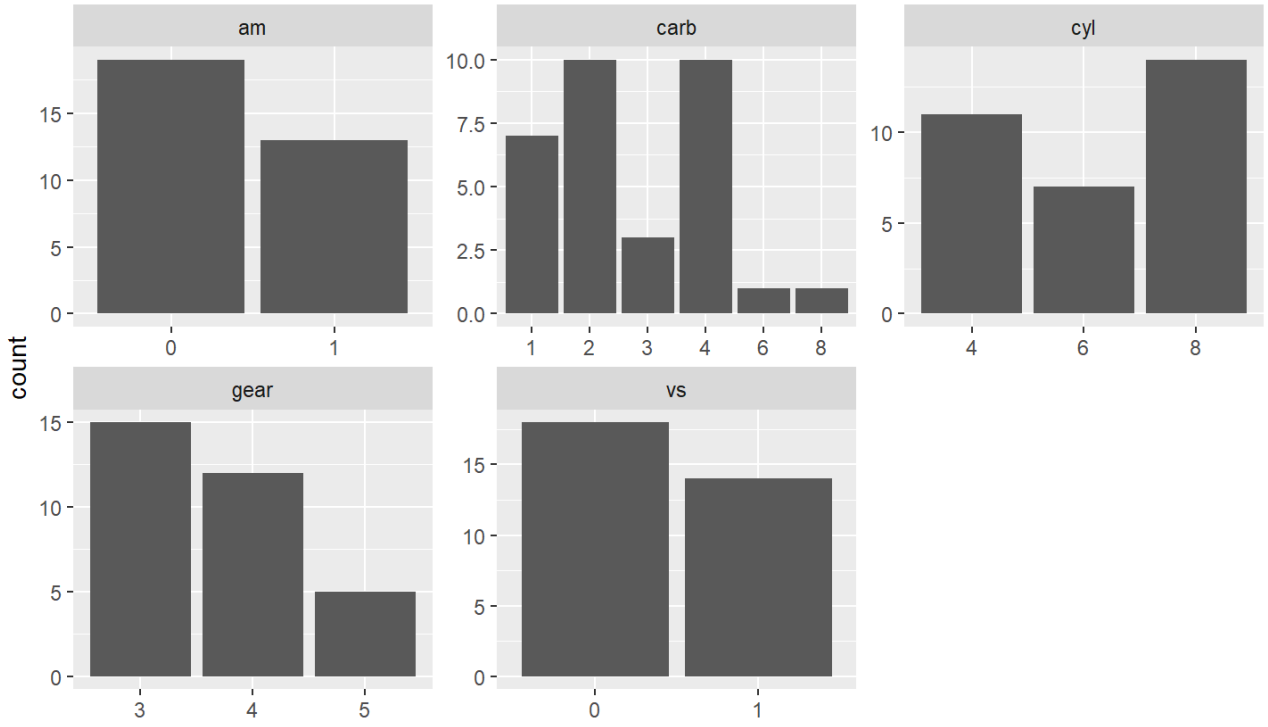
Kurtosis: $\kappa_x = E[(\frac{X-\mu_X}{\sigma})^4]$.



Descriptive Statistics

am	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.4062	3rd Qu.:1.0000	Max. :1.0000
carb	Min. :1.000	1st Qu.:2.000	Median :2.000	Mean :2.812	3rd Qu.:4.000	Max. :8.000
cyl	Min. :4.000	1st Qu.:4.000	Median :6.000	Mean :6.188	3rd Qu.:8.000	Max. :8.000
gear	Min. :3.000	1st Qu.:3.000	Median :4.000	Mean :3.688	3rd Qu.:4.000	Max. :5.000
vs	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.4375	3rd Qu.:1.0000	Max. :1.0000

Even better…

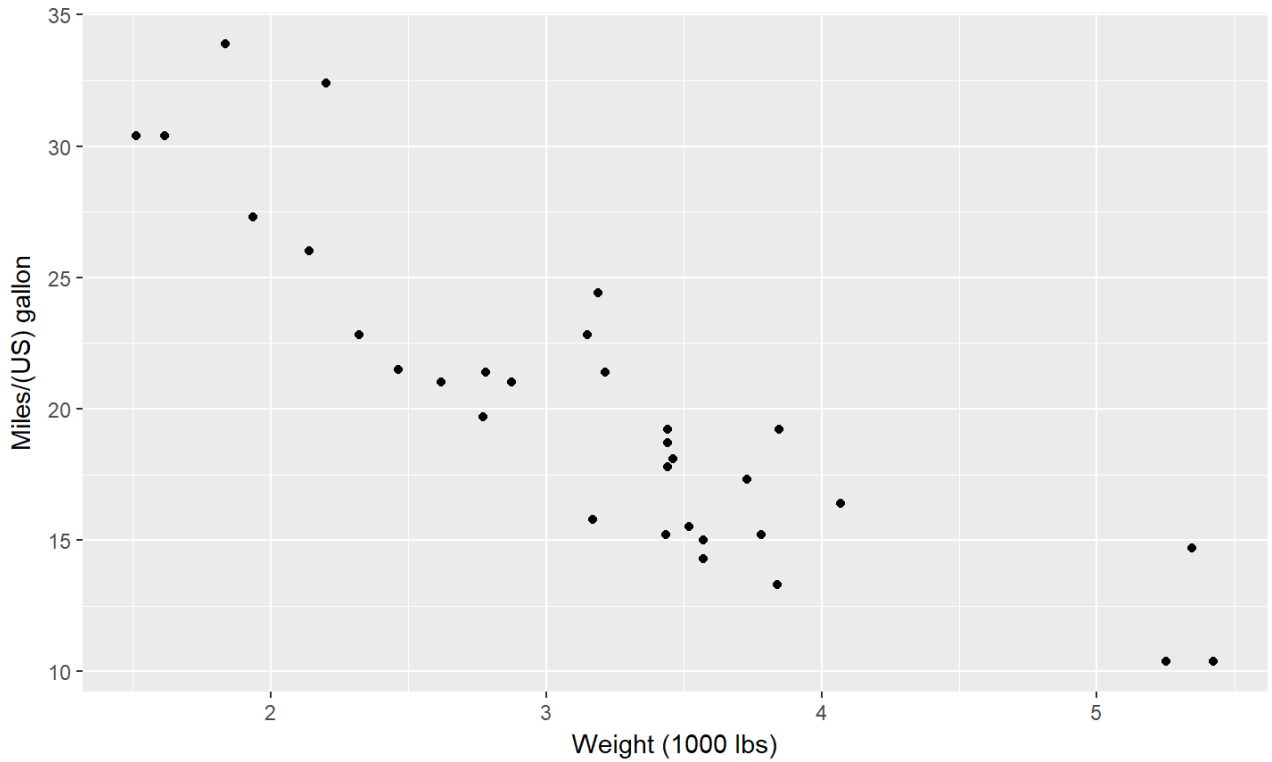


How to describe a relationship?

Contingency table:

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
## ============================================
##                mtcars$cyl
## mtcars$gear        4       6       8   Total
## --------------------------------------------
## 3                  1       2      12      15
##                0.031   0.062   0.375
## --------------------------------------------
## 4                  8       4       0      12
##                0.250   0.125   0.000
## --------------------------------------------
## 5                  2       1       2       5
##                0.062   0.031   0.062
## --------------------------------------------
## Total             11       7      14      32
## ============================================
```

Scatter plot



Do a multivariate analysis

Example: Tang, Hu, and Jin (2016)

Puzzle: Same education level, but difference in labor mobility between Han and Uyghur

Theory

Affirmative inaction language policy reduces Uyghurs' labor mobility.

Hypothesis:

- H_1: Education is fairly equal between the Han and the Uyghur groups.
- H_2: The linguistically distinctive Uyghurs are far less proficient in Mandarin than the Han majority.
- H_3: Hans enjoy a higher degree of socioeconomic status than the Uyghurs.
- H_4: Language proficiency plays a favorable role in improving the socioeconomic conditions for the Uyghurs.

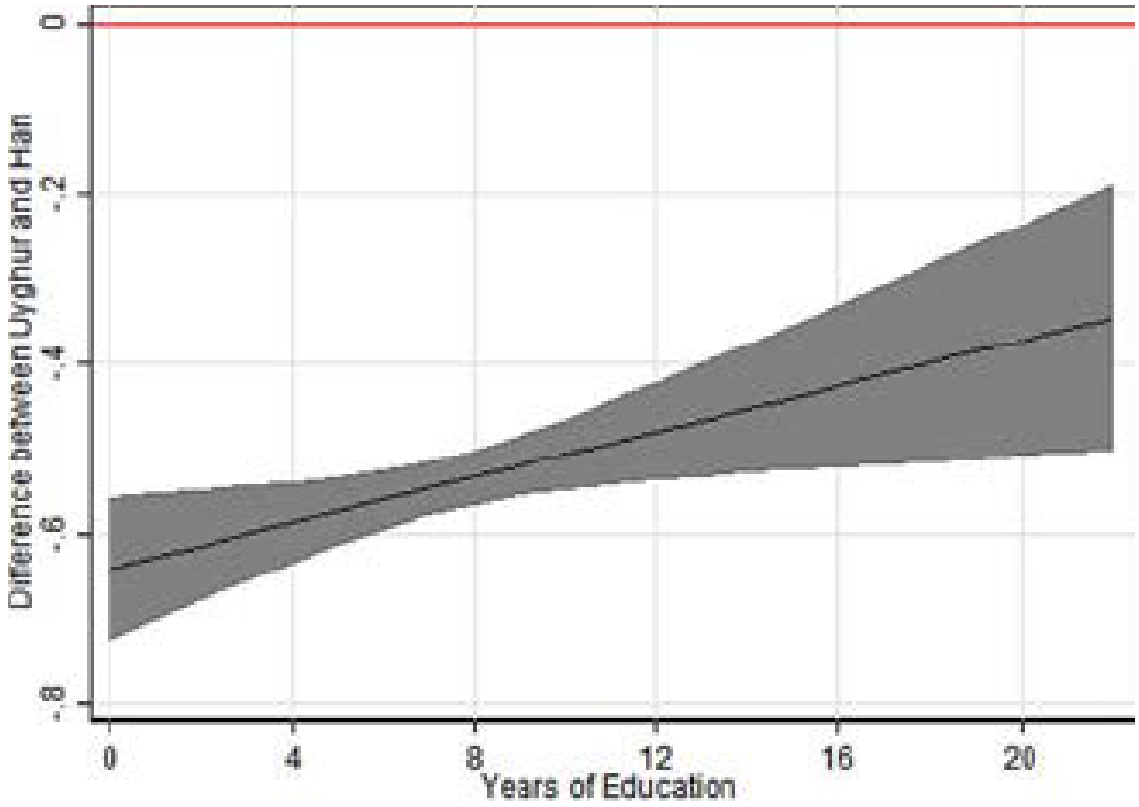
Examination

$H_{1,2}$: Education is fairly equal between the Han and the Uyghur groups.

Educational and Language Differences Between Han and Uyghur in China (weighted)

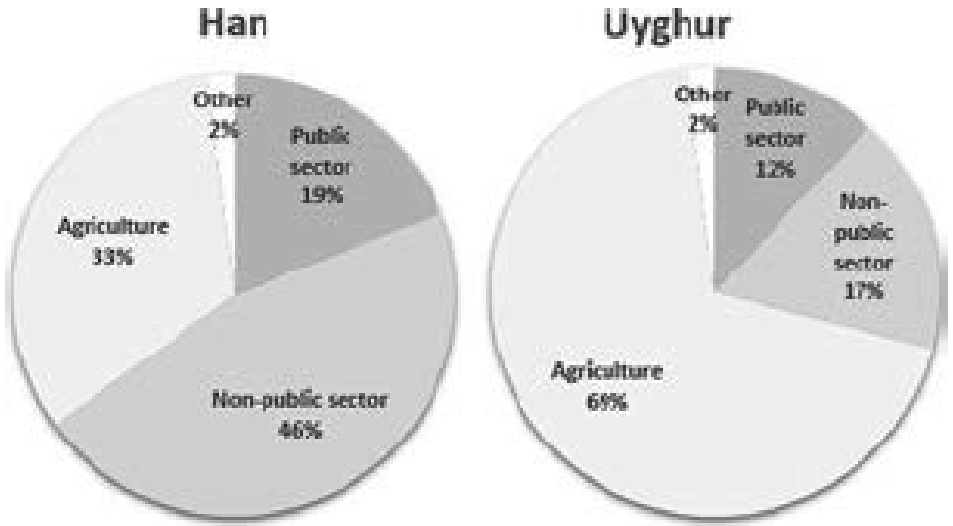
Ethnicity	Mean	Std. Err.	*T*-test (against Han)
A. Education (in year)			
Average	8.511	0.046	
Han	8.509	0.469	
Uyghur	8.363	0.175	−0.145(0.181)
B. Mandarin proficiency (0–1)			
Average	0.654	0.004	
Han	0.666	0.004	
Uyghur	0.109	0.016	−0.557(.016)***

H_2: The linguistically distinctive Uyghurs are far less proficient in Mandarin than the Han majority.



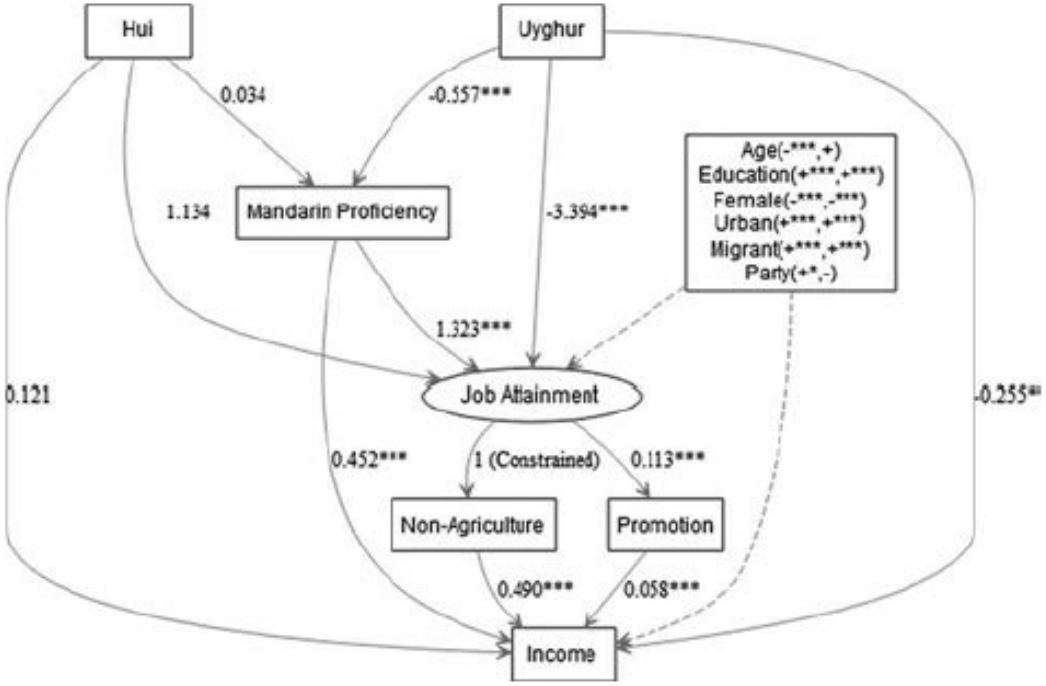
H_3: Hans enjoy a higher labor mobility than the Uyghurs.

Figure 2. **Job type by ethnicity (Han and Uyghur, weighted %).**



H_4: Language proficiency plays a favorable role in improving the socioeconomic conditions for the Uyghurs.

Figure 3. **The structural equation model of the effect of language on socio-economic attainment. $*p < 0.05$, $***p < 0.001$.**



Wrap up

Understand Large-N Analysis

	Dependent variable:	
	delay	
	(1)	(2)
temp	0.088**	0.088**
	(0.041)	(0.043)
wind	0.166	0.166
	(0.164)	(0.159)
precip	18.918***	18.918***
	(3.249)	(4.735)
Constant	7.263**	7.263**
	(3.099)	(3.053)
F Statistic (df = 3; 360)	12.879***	7.73***
Observations	364	364
R^2	0.097	0.097
Adjusted R^2	0.089	0.089
Residual Std. Error (df = 360)	13.248	13.248
Note:	*p<0.1;* ***p<0.05;*** p<0.01	

Do large-N analyses

- Elaborate the puzzle
- Set up the theory
- Imply hypotheses
- Design empirical examination
- Data analysis
- Result discussion