Political Methodology

- **Political Methodology**
  - Using research designs and statistical methods to analyze data and test hypotheses derived from theory.
  - Name of the game: *making inferences about unknown quantities of interest based on known quantities of interest.*
    - Most frequently: From *sample* to *population*.
    - Generalization; knowledge claims about political phenomena
  - *Causal inference*: What is the causal effect of X on Y?
  - Communicating *uncertainty*; probabilistic modeling, so there is *sampling uncertainty*.
    - Notion of infinite sampling; classical probability theory

Causality

- **Causality and the notion of ceteris paribus**

  Definition of causal effect of x on y:

  > How does variable y change if variable x is changed but all other relevant factors are held constant?

- **x is our *independent variable* (explanatory, causal variable); y is our *dependent variable* (the outcome variable we want to explain).**
  - **X explains y; x causes y**

Causality

- **Causality and the notion of ceteris paribus**

Definition of causal effect of x on y:

> How does variable y change if variable x is changed but all other relevant factors are held constant?

- **Most questions are ceteris paribus questions**
- **It is important to define which causal effect one is interested in**
- **It is useful to describe how an experiment would have to be designed to infer the causal effect in question**

Causal Inference

- **Experimental Data**
- **Observational Data**

Data Structures

- **Different kinds of data sets in political science**
  - Cross-sectional data
  - Time series data
  - Pooled cross sections
  - Panel/Longitudinal data
  - Multilevel (hierarchical) data
- **Methods depend on the nature of the data used**
  - Use of inappropriate methods may lead to misleading results

Data Structures

- **Cross-sectional data sets**
  - Sample of individuals, households, firms, cities, states, countries, or other units of interest at a given point of time/in a given period
  - Cross-sectional observations assumed to be **independent**
  - For example, **pure random sampling** from a population
  - Sometimes pure random sampling is violated, e.g. units refuse to respond in surveys, or if sampling is characterized by clustering
  - Cross-sectional data in political science

Data Structures

- **Cross-sectional data set on wages and other characteristics**

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Indicator variables (1=yes, 0=no)

Observation number

Hourly wage

Data Structures

- **Cross-sectional data on growth rates and country characteristics**

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2.30	17	6

Growth rate of real per capita GDP

Government consumtion as percentage of GDP

Adult secondary education rates

Data Structures

- **Time series data**
  - Observations of a variable or several variables over time
  - For example, stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, automobile sales, …
  - Time series observations are typically **serially correlated**
  - Ordering of observations conveys important information
  - Data frequency: daily, weekly, monthly, quarterly, annually, …
  - Typical features of time series: **trends and seasonality**

Data Structures

- **Time series data on minimum wages and related variables**

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Average minimum wage for given year

Average coverage rate

Unemployment rate

Gross national product

Data Structures

- **Pooled cross sections**
  - Two or more cross sections are combined in one data set
  - Cross sections are drawn independently of each other
  - Pooled cross sections often used to evaluate policy changes
  - Example:
    - Evaluate effect of change in property taxes on house prices
    - Random sample of house prices for the year 1993
    - A **new** random sample of house prices for the year 1995
    - Compare before/after (1993: before reform, 1995: after reform)

Data Structures

- **Pooled cross sections on housing prices**

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Property tax

Size of house in square feet

Number of bathrooms

Before reform

After reform

Data Structures

- **Panel or longitudinal data**
  - The **same** cross-sectional units are followed over time
  - Panel data have a **cross-sectional and a time series dimension**
  - Panel data can be used to account for time-invariant unobservables
  - Panel data can be used to model lagged responses
  - Example:
    - City crime statistics; each city is observed in two years
    - Time-invariant unobserved city characteristics may be modeled
    - Effect of police on crime rates may exhibit time lag

Data Structures

- **Two-year panel data on city crime statistics**

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Each city has two time series observations

Number of police in 1986

Number of police in 1990

Hierarchical Data

Student	School	Y	X1	X2	X3	X4
1	1	54	2	32	1	44
2	1	64	4	25	1	44
3	1	87	9	45	1	44
4	2	24	4	44	0	36
5	2	98	7	32	0	36
6	2	65	6	22	0	36
7	3	45	9	19	0	22
8	3	32	5	15	0	22
9	3	37	2	25	0	22
10	4	84	7	30	1	45
11	4	45	4	38	1	45
12	4	65	3	36	1	45
13	5	21	8	41	1	18
14	5	65	6	22	1	18
15	5	98	1	18	1	18

- X1 and X2 are level-1 variables
- X3 and X4 are level-2 variables.
- Balanced data: cluster sizes are equal

Levels of Measurement

- Levels of measurement
  - three levels of measurement:
    - interval (continuous)
    - ordinal
    - nominal
  - Type of variable(s) you have dictates how you **analyze, summarize, and generalize** about it.
    - E.g., scatterplots, bar graphs, histograms

- Type of **dependent variable** you have dictates the type of method you use.
  - Interval/continuous: linear regression (OLS).

- Other types of dependent variables:
  - Binary: logit/probit via MLE
  - Ordinal: ordered logit/probit via MLE
  - Nominal: multinomial logit/probit via MLE
  - Duration (time until event): duration, event history models via MLE