大数据统计与计量分析II

西安交通大学管理学院
信息系统与电子商务系
刘笑笑

第二章 大数据统计与计量分析II
—数据、变量与其他条件不变

西安交通大学管理学院
信息系统与电子商务系
刘笑笑

第一章 数据、变量与其他条件不变

- **§1.1 数据**
- §1.2 变量
- §1.3 计量经济分析中的因果关系与其他条件不变

§1.1 数据

- 一、数据类型
- 二、数据来源

- 计量经济分析必须需要数据
- 计量方法的选择取决于所使用的数据
- 使用不合适的方法可能会导致错误的结果

一、数据类型

- 横截面数据（Cross-sectional data）
- 时间序列数据（ Time series data）
- 混合横截面数据（Pooled cross-sections data）
- 面板数据（Panel data）

横截面数据

- 横截面数据（Cross-sectional data）
  - 横截面数据集：在给定时点对个人、家庭、企业、城市、国家或一系列其他单位采集样本所构成的数据集
  - 在总体中随机抽样得到
  - 若不是随机抽样，会带来样本选择等问题
  - 应用：应用微观计量经济学

横截面数据

- 横截面数据（Cross-sectional data）
  - ➢例子：有关医生个体特征和咨询量的横截面数据集

Id	SelfIn	Activities	Visits	Papers	Patients
24	1	550	45,329	3	59
25	1	0	16,092	0	0
26	1	955	141,886	7	120
27	0	0	497	0	0
28	1	0	5,787	0	0
29	1	360	389,587	12	63
30	1	235	13,114	2	4
31	1	1995	104,774	5	196

观测序号　虚拟变量　在线活动量　访问量　文章数　患者数

横截面数据

- 横截面数据（Cross-sectional data）
  - 例子：有关医生个体特征和咨询量的横截面数据集（不同的变量对应不同的时期）

ID	PatientsFeb	Title	SelfIn	VoteJan
1	21	1	1	12
2	3	1	1	6
3	1173	0	1	69
4	6	1	0	1
5	144	0	1	16
6	2	1	0	3
7	3	1	1	7
…	…	…	…	…

横截面数据

- 横截面数据（Cross-sectional data）
  - 例子：2019年部分省份GDP、人均GDP和常住人口相关数据

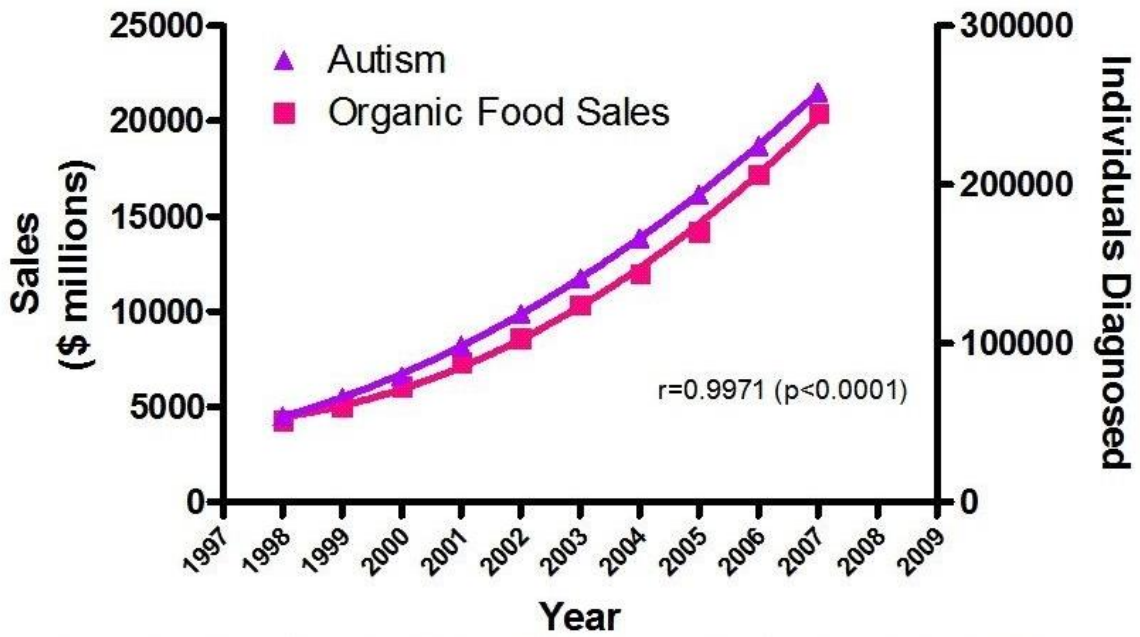
ID	Province	GDP	GDP_Capita	Per_Popu
1	陕西省	25793.2	66649	3876
2	北京市	35371.3	164220	2154
3	天津市	14104.3	90371	1562
4	河北省	35104.5	46348	7592
5	山西省	17026.7	45724	3729
6	内蒙古	17212.53	67852	2540
7	辽宁省	24909.5	57191	4352
…	…	…	…	…

时间序列数据

- 时间序列数据（Time series data）
  - 时间序列数据集：由对一个或者几个变量不同时间的观测值所构成
  - 例如：股票价格、国内生产总值
  - 一般按时间顺序排序
  - 数据频率：天，周，月，季节，年度
  - 典型特征：趋势和季节性
  - 应用：应用宏观计量经济，金融

时间序列数据

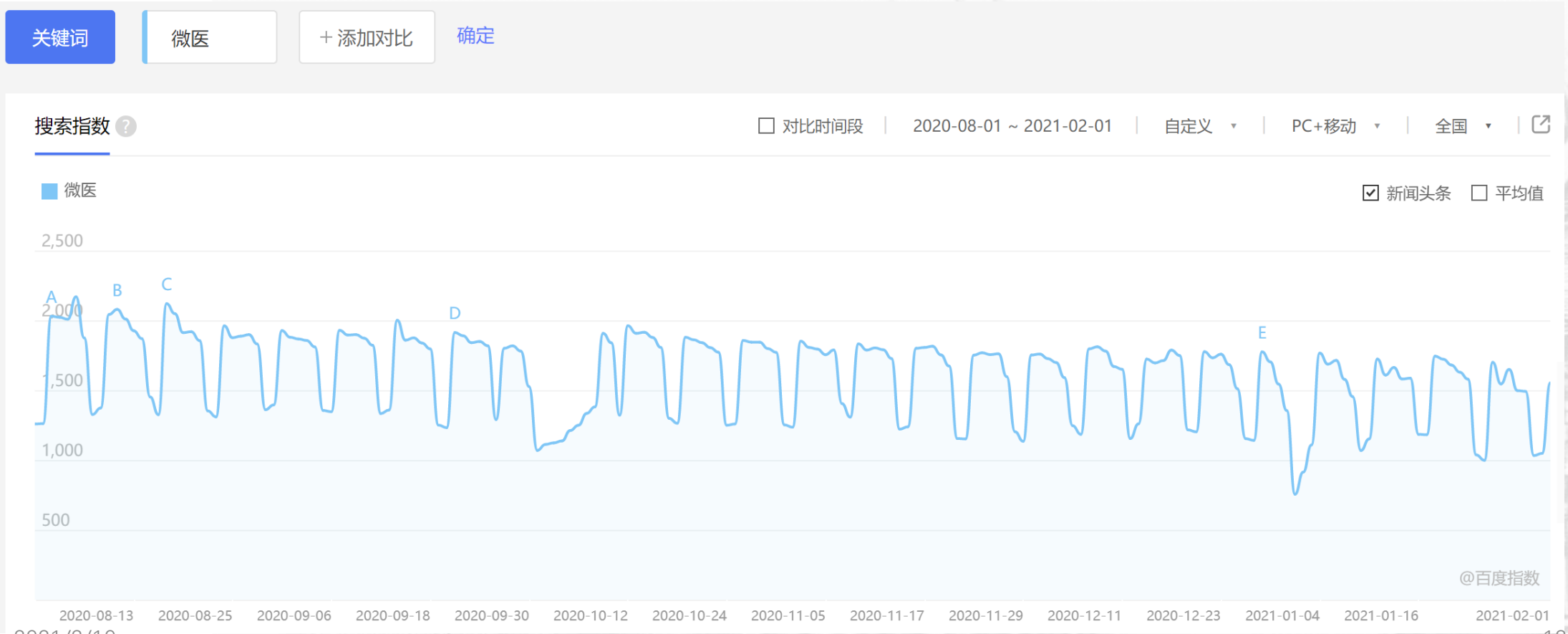
- 时间序列数据（Time series data）



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

时间序列数据

- 时间序列数据（Time series data）

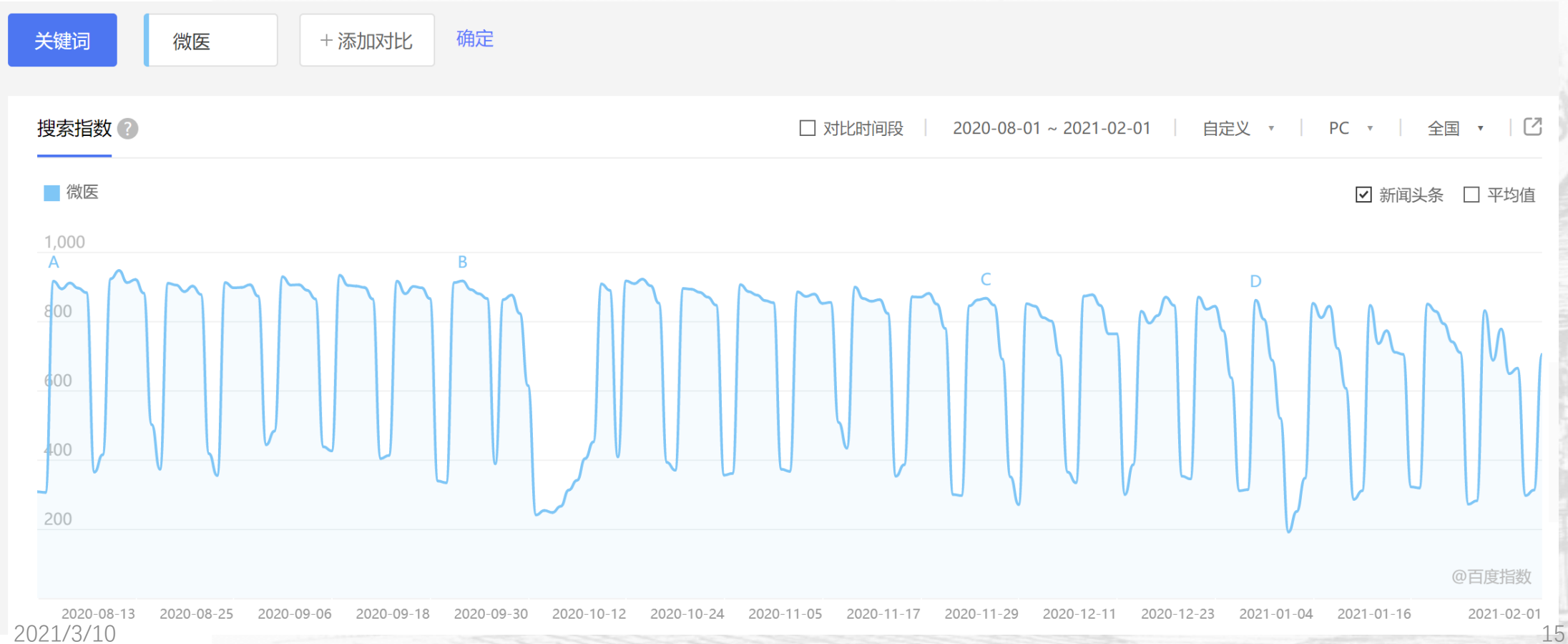


时间序列数据

- 时间序列数据（Time series data）

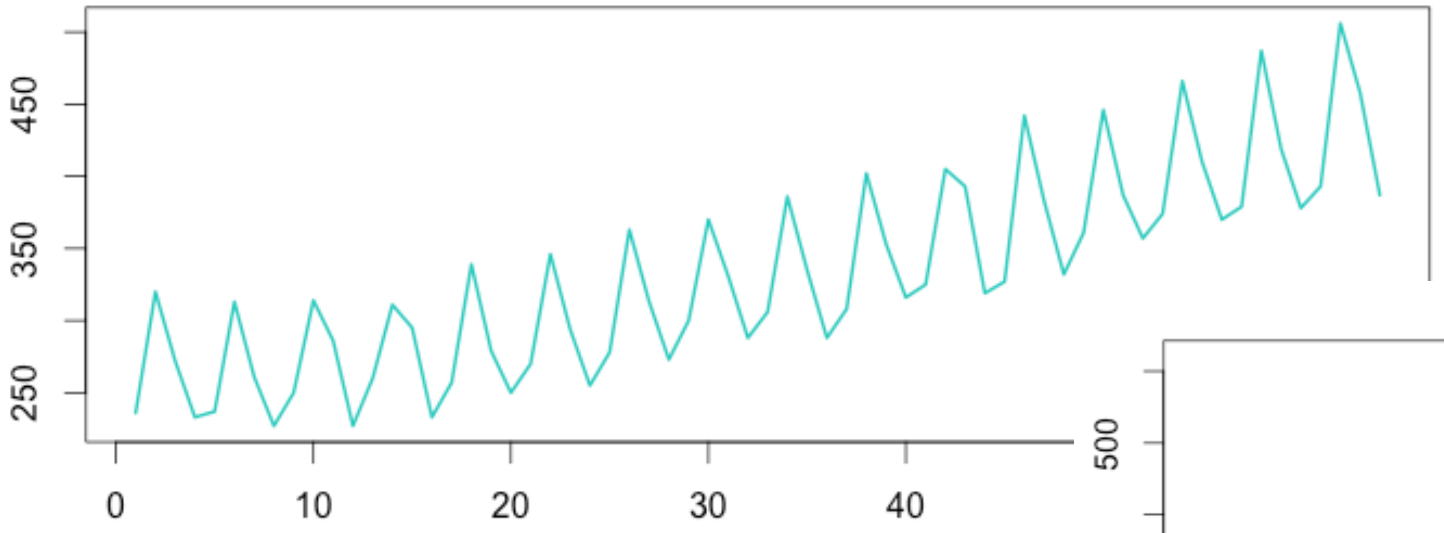
时间序列数据

- 时间序列数据（Time series data）

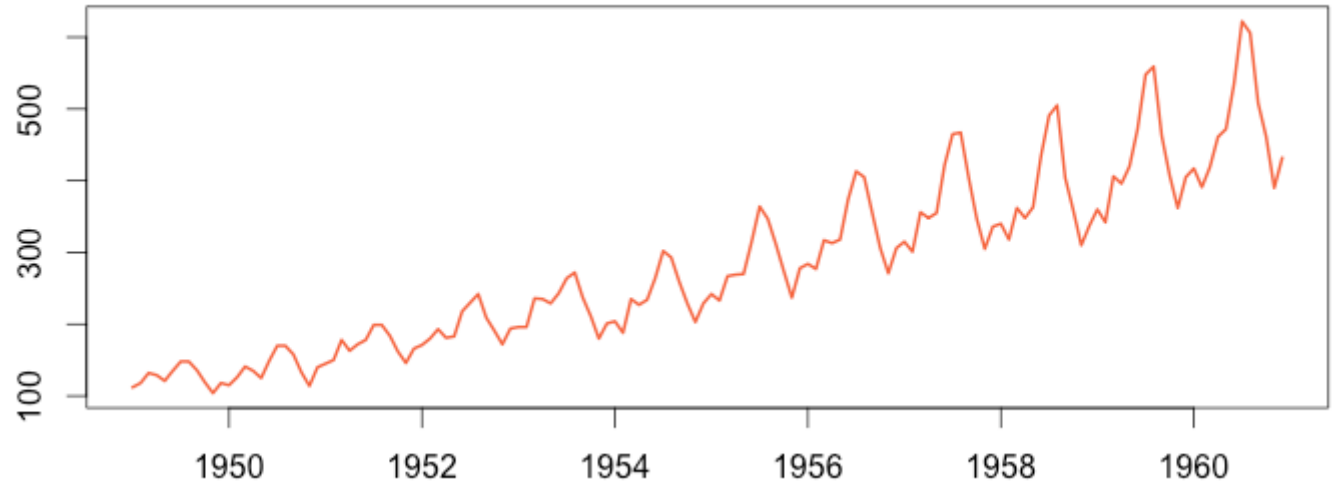


时间序列数据

- 时间序列数据（Time series data）



某国啤酒生产量



航空公司旅客数量

https://anomaly.io/seasonal-trend-decomposition-in-r/

时间序列数据

- 时间序列数据（Time series data）
  - 例子：陕西省的出生人口数，GDP和人均GDP相关数据

ID	Year	Births_Num	GDP	GDP_Capita
1	2000	454000	1804	4968
2	2001	384000	2010.62	5511
3	2002	384000	2253.39	6161
…	…	…	…	…
19	2018	411000	24438.32	63477
20	2019	408000	25793.2	66649

混合横截面数据

- 混合横截面数据（Pooled cross-sections data）
  - 既有横截面数据的特点，也有时间序列的特点
  - 不同时间点上的横截面独立形成
  - 扩大样本容量，分析基本关系如何随时间变化
  - 通常用来评估政策
  - 例如：
    - 评估某国1994年财产税的下调对住房价格的影响
    - 来自1993年住房价格的随机样本
    - 来自1995年住房价格的新随机样本
    - 比较前后

混合横截面数据

- 混合横截面数据（Pooled cross-sections data）
  - ➢例子：混合横截面数据：两年的住房价格

ID	year	house_price	Property_tax	house_size	bedrooms	bathrooms
1	1993	85500	42	1600	3	2
2	1993	67300	36	1440	3	2
3	1993	134000	38	2000	4	2
…	…	…	…	…	…	…
250	1993	243600	41	2600	4	3
251	1995	65000	16	1250	2	1
252	1995	182400	20	2200	4	2
253	1995	97500	15	1540	3	2
…	…	…	…	…	…	…
520	1995	57200	16	1100	2	1

面板数据

- 面板数据（Panel data）
  - 面板数据集：由数据集中每个横截面单位的一个时间序列组成
  - 同一横截面数据的数据单位在不同时期重复观测
  - 优势：能够控制不随时间变化的观测单位观测不到的特征
  - 优势：能够研究决策行为或结果中滞后的重要性
  - 例如：
    - 顾客使用天猫精灵对其购买行为的影响
    - 每个顾客都观测30天
    - 顾客观测不到的特征能够被控制
    - 使用天猫精灵对购买行为的影响可能有滞后性

面板数据

- 面板数据（Panel data）
  - ➢例子：有关医生咨询量的面板数据集

每个医生有三个时间序列观察值

Id	Doctor_ID	T	SelfIn	Activities	Visits	Papers	Patients
1	1	1	1	210	45,317	3	56
2	1	2	1	310	16,062	0	0
3	1	3	1	955	141,686	7	120
4	2	1	0	0	267	0	0
5	2	2	0	0	5,687	0	0
6	2	3	0	360	389,187	12	63
7	3	1	1	235	13,014	2	4
8	3	2	1	1995	104,674	5	196
9	3	3	1	866	210,323	6	96

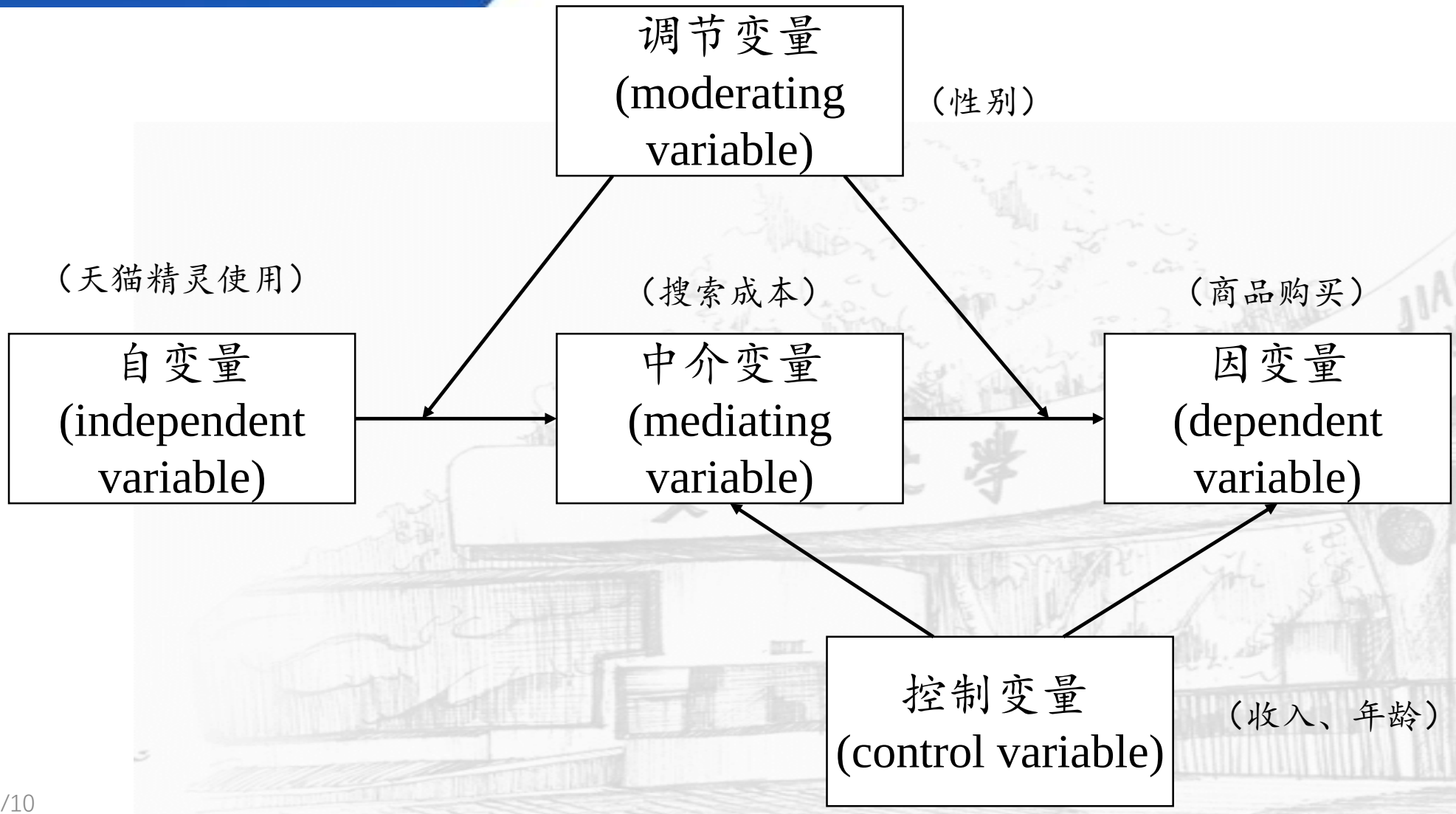
二、数据来源

- 实验，访谈，问卷…
- 脑电波，眼动仪，虚拟现实…
- 数据库：Wind，统计数据…
- 互联网
- 网站、社交媒体：淘宝，微博…
- 合作公司企业
- 各类信息系统记录的数据
- 多来源数据
- …

第一章 数据、变量与其他条件不变

- §1.1 数据
- **§1.2 变量**
- §1.3 计量经济分析中的因果关系与其他条件不变

§1.2 变量

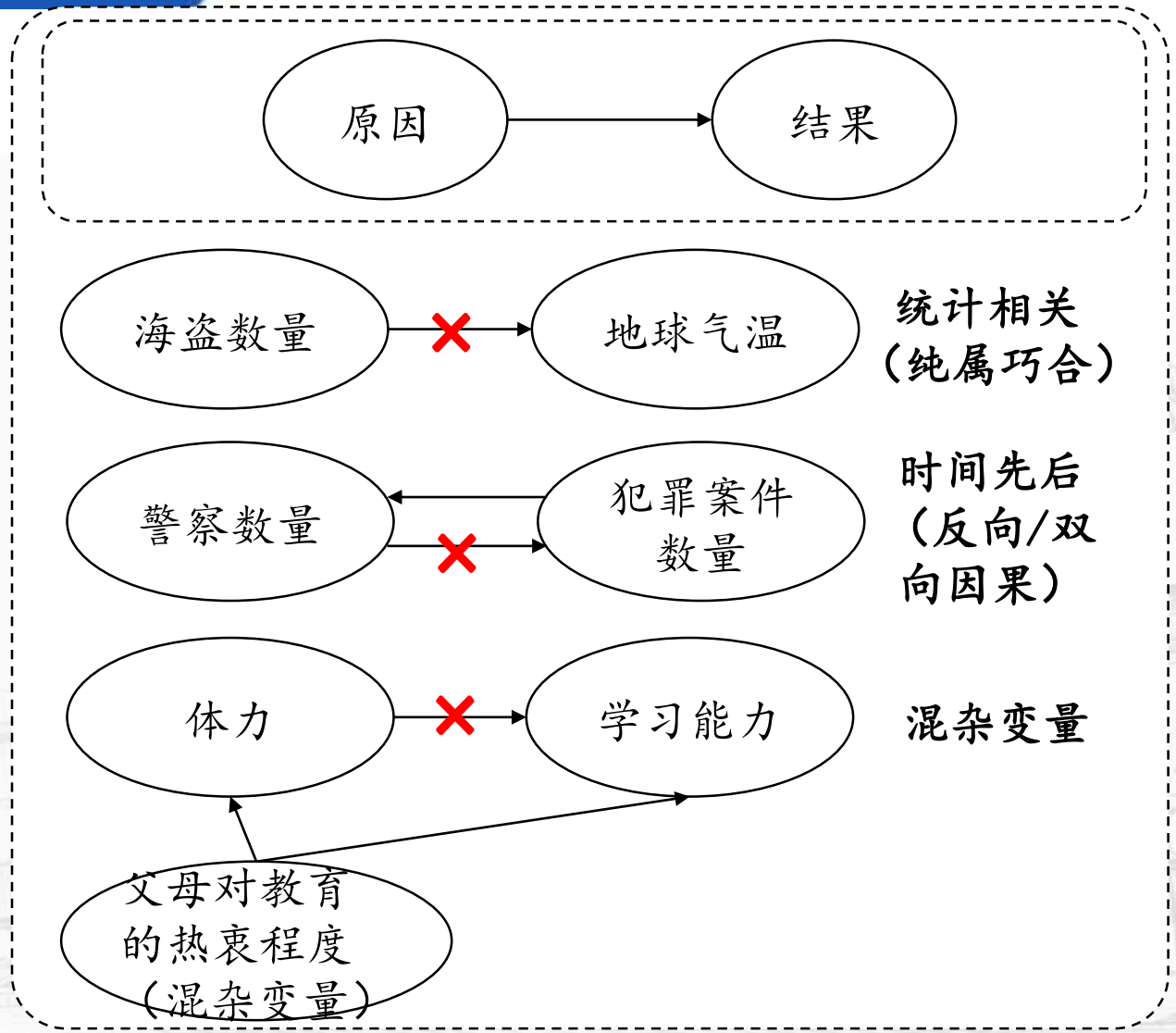

调节变量
(moderating variable)
（性别）
（天猫精灵使用）
（搜索成本）
（商品购买）
自变量
(independent variable)
中介变量
(mediating variable)
因变量
(dependent variable)
控制变量
(control variable)
（收入、年龄）

第一章 数据、变量与其他条件不变

- §1.1 数据
- §1.2 变量
- §1.3 计量经济分析中的因果关系与其他条件不变

回顾：相关与因果

- 相关vs因果
  - 相关≠因果



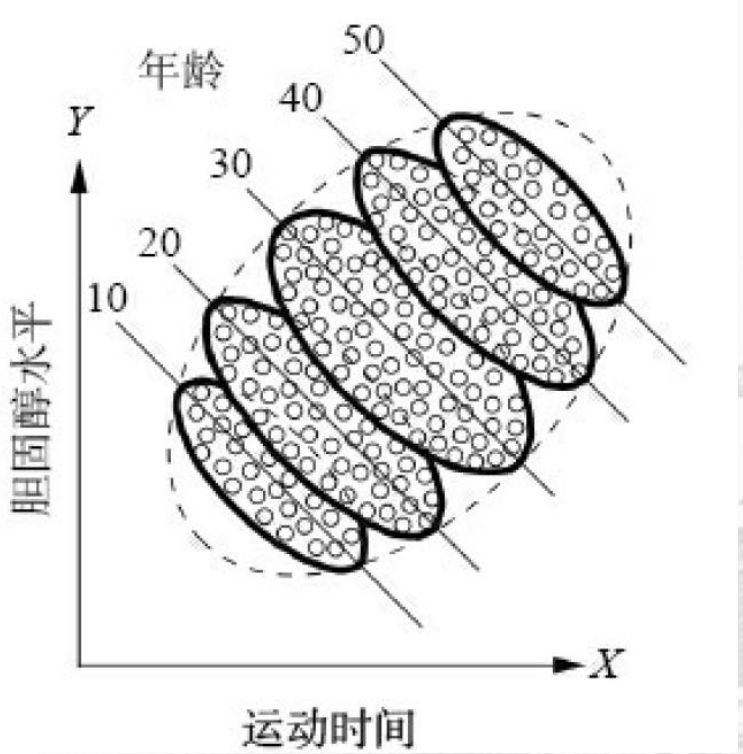
其他条件不变

- x对y的因果关系：

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

- 其他条件不变（ceteris paribus）——其他（相关）因素保持不变
- 多数实证研究中的一个关键问题：对于做出一个因果推断而言，是否有足够多的其他因素保持不变？

运动时间对胆固醇水平的影响



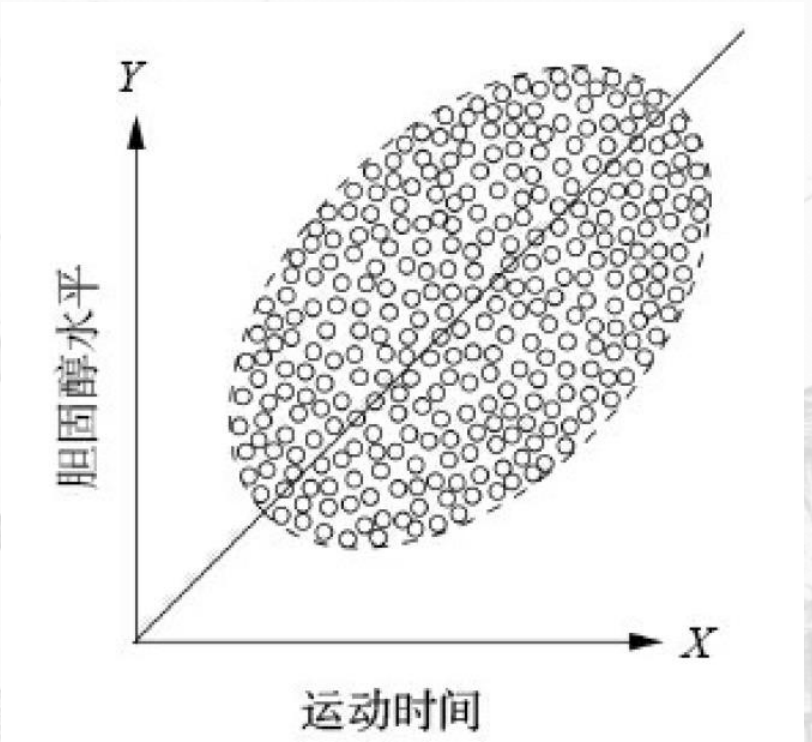


图6.6 辛普森悖论：对于每个年龄组来说，运动似乎都是有益的（向下的趋势线），但对整个总体而言，运动似乎是有害的（向上的趋势线）。

其他条件不变

- x对y的因果关系：

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

- 其他条件不变（ceteris paribus）——其他（相关）因素保持不变
- 多数实证研究中的一个关键问题：对于做出一个因果推断而言，是否有足够多的其他因素保持不变？
- 回想一下反事实框架

事实：
使用天猫精灵

月购买量
1000元

原因
使用天猫精灵

方法
对“事实”中原因发生后的结果与“反事实”中原因未曾发生时的结果进行对比

结果
月购买量

事实和反事实的结果存在差异，因此使用天猫精灵和购买量之间存在因果关系

反事实：
不使用天猫精灵

月购买量
500元

其他条件不变

- x对y的因果关系：

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

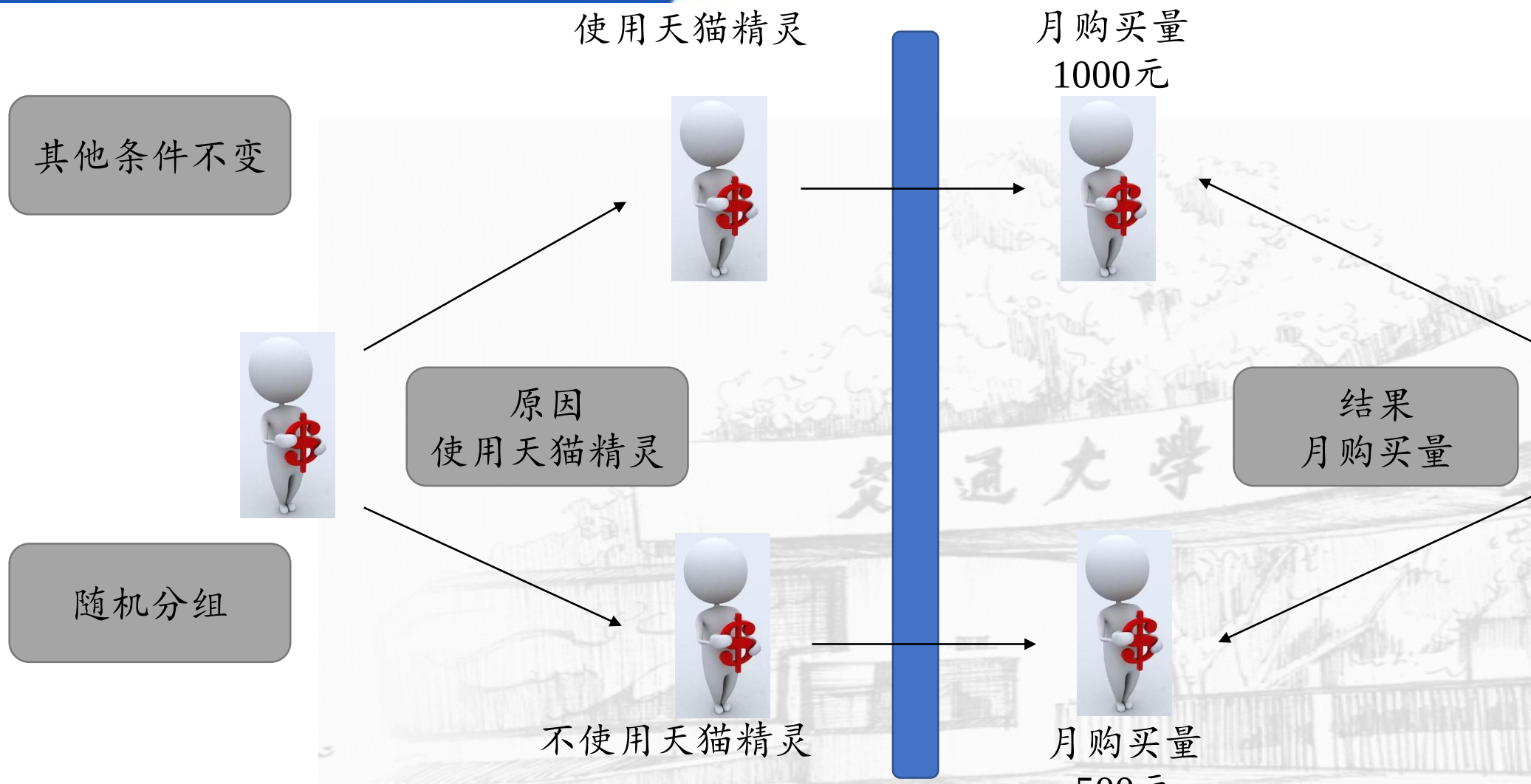
- 其他条件不变（ceteris paribus）——其他（相关）因素保持不变
- 多数实证研究中的一个关键问题：对于做出一个因果推断而言，是否有足够多的其他因素保持不变？
- 回想一下反事实框架
- 实验：其他条件不变

案例

- 教育回报率：教育水平对工资的影响
  - 因果关系：所有其他影响工资的变量都保持不变，例如：工作经历，家庭背景，智力等
- 实验:
  - 选择一组人，将他们随机分配到不同的教育年限，然后比较工资　　不允许！
  - 若从工作人群总体中随机地抽取一大群人，使用非实验数据，问题：教育程度与其他影响工资的因素相关（例如，智力）

计量经济方法的许多进展都试图对计量模型中的这些无法观测因素进行处理

案例


使用天猫精灵
月购买量
1000元
其他条件不变
原因
使用天猫精灵
结果
月购买量
两组的结果存在差异，因此使用天猫精灵和购买量之间存在因果关系
随机分组
不使用天猫精灵
月购买量
500元

练习题2-1

- 现在正在进行一项研究，以确定课余时间上网课是否会影响三年级学生成绩。
  - ➢你为什么预计课后上网课与学生成绩呈正（负）相关关系？
  - ➢假设你能做任何你想做的实验，你想做些什么？请具体说明。
  - ➢假设你能搜集到某地区几千名三年级学生的观测数据，你能得到他们在2020年课后上网课的小时数和三年级末的标准化考试分数。你发现上网课与学习成绩呈正相关关系，那么一定意味着学生利用课余时间上网课一定能够提高其学习成绩吗？

练习题2-2

- 企业进行工作培训的原因之一是其能够提高工人的生产力。现有某个城市制造企业的数据，即对每个制造企业，有人均工作培训小时数(training)和单位工时生产的合格产品数(output)方面的信息。要求你分析更多的工作培训是否会使工人更有生产力。
  - 请阐述这个问题背后其他条件不变的思维实验。
  - 一个企业培训其员工的决策是否独立于工人特征？工人可观测与不可观测的特征各有哪些？
  - 除工人特征外，请再列出其他影响工人生产力的因素。
  - 你若发现training和output之间成正相关关系，能否证明工作培训能够提高工人的生产力？为什么？

课后作业（小组）

- 请结合你的研究问题
  - 1. 针对你的研究问题，假设你能做任何你想做的实验，你想做些什么？请具体说明。
  - 2. 请列出其他可能影响因变量的因素（可观测与不可观测）。
  - 3. 请列出可能存在的混淆变量。
  - 4. 根据你的研究问题，请提出你的假设，并解释为什么？
- 问题1-3，小组课后讨论
- 问题4，结合理论构建和假设提出，补充到研究项目中
- 完善研究项目

第二章 总结

- 数据：类型、特点、区别
- 变量
- 因果关系与与他条件不变