从零做 DID：手算、回归与独立迁移

第一课：用四格均值理解反事实

对照组告诉我们一种假设的参照趋势：如果新泽西州没有这次政策，并且它的未处理变化与宾州东部一致，那么政策后平均就业的反事实约为20.4394−2.1656=18.2738。实际值为21.0274，与这个反事实相差约2.7536个全职当量。这与“两次差分”是同一算式的另一种看法。

平行趋势要求的是“没有政策时两组的变化可比”，不要求政策前水平相同。这里只有两轮均值，不能凭四格数检验处理前趋势；而且前后有效门店构成有变化。因此这一步先学会构造比较，尚不能只靠2.7536就排除同期冲击或样本变化的解释。

主例复用 Card 与 Krueger 1992 年两州快餐门店调查的既有四格汇总。反事实情景和另一份8家门店、2期的16行面板均为明确标注的教学设定。

开始之前：文件、软件与一次完整运行

本课先用纸笔完成四个均值，再用软件核对。下载完整课堂包并全部解压，文件夹内应有两州就业汇总.csv、反事实情景.csv、独立练习_模拟门店.csv，以及课堂练习.do和课堂练习.py。不要直接在压缩包浏览窗口里双击脚本，也不要只复制一份代码到另一个空文件夹；程序需要在同一位置找到输入。

在自己的授权Stata中，先保存当前工作，再在命令窗口输入cd "实际解压目录"。Windows路径可以写D:/课程/DID；路径必须换成自己的。输入pwd确认当前位置，再用Do-file Editor打开课堂练习.do。先通读开头的clear、set more off和日志命令，确认清空的是本次可重新读取的内存数据，随后运行整份do文件。不要把终端命令python或Windows双击说明输入Stata命令窗口。

第一次运行按顺序观察四行数据、两组变化、交互项和模拟结果，并查看目录中新生成的Stata四格均值结果.csv、Stata模拟四格结果.csv及文本日志。尚无Stata时，可先按网站的本课Python环境准备入口运行Python独立复核，输出包括DID核对结果.csv和趋势图。在 Stata 中用同一份数据核对结果。每次练习都保留自己的代码与结果。

先确定谁被处理、谁作参照以及什么时候比较

DID的第一步是明确比较设计。主例中处理组是新泽西州调查门店，对照组是宾夕法尼亚州东部调查门店，政策前和政策后是两轮调查。背景是1992年4月新泽西最低工资提高；本课不重新开展调查，也不把这个案例自动推广到其他行业或年代。处理组身份必须依据制度事实确定，不能看哪一组结果上升就把哪一组叫处理组。

在纸上画一个两行两列的格子。行写处理组、对照组，列写政策前、政策后。每一格应对应同一结果概念、明确分母和相应时期。随包真实资料每行是地区与调查时期的汇总；独立作业每行则是一家模拟门店在一个时期的记录。这两个文件的行数不能当作相同层级的样本量。

主程序先assert合法州别和调查时期，再建立treat与post。表达式(州别 == "新泽西州")在条件成立时生成1，否则为0；因此先检查允许值非常重要，拼错州名不能默默被当成正确对照。post为1表示政策后，不表示是否属于处理组。isid treat post要求四种组合各只有一行，重复行或缺格应先回查输入。

读懂四格：为什么不能直接比较就业总量

两州门店数不同，各轮有效记录也可能不同。新泽西政策前的均值是6561.05除以321，约为20.43940810；政策后是6707.75除以319，约为21.02742947。宾州东部两格分别是1796.50除以77和1629.75除以77，约为23.33116883和21.16558442。用计算器逐格复算一次，再与CSV中总量和分母相对照。

这里的结果单位是每家门店的平均全职当量。它由全职员工、经理及副经理与兼职人数的一半构成，不等于“实际多雇了多少个人”。6561.05也不是门店均值，321更不是两轮合并后的独立门店数。直接把两州合计作差，会混入门店数量不同的影响。遇到单位变更，先统一结果定义，再考虑估计方法。

代码用generate double avgfte = 全职当量合计 / 有效门店数保存均值，用format调整显示位数。format不会把底层数据截断；手算时尽量保留足够精度，到展示时再四舍五入。先把四格都截成两位再相减，会得到与完整精度不同的数字。真实汇总按各轮就业分项完整记录整理，前后分母321和319不同，不应称为已经锁定同一批完整追踪门店。

第一次相减看变化，第二次相减扣除参照变化

始终沿同一方向计算“政策后减政策前”。新泽西的变化为21.02742947减20.43940810，约增加0.58802137；宾州东部的变化为21.16558442减23.33116883，约减少2.16558442。DID再计算处理组变化减对照组变化，即0.58802137减去负的2.16558442，得到约2.75360578。先在纸上保留括号，可以避免把减负数误写成加负数。

0.5880是处理组自身的前后差，2.7536是两组的变化差。两者回答的问题不同。若把0.5880直接解释成政策效应，就隐含无政策时处理组平均结果不会变化；DID改用参照组同期变化作为反事实变化的依据。这个依据是否可信需要制度和样本论证，计算公式本身不能证明。

Stata每执行一次summarize，r(mean)就更新为这一次的均值。脚本在每次计算后立即保存到不同scalar中：tb、ta、cb、ca分别是处理组前后和对照组前后。quietly只是不显示命令结果，不会跳过计算。如果只运行最后一行did表达式而没有运行前面的标量赋值，可能出现未定义或沿用旧内存值；重新从导入开始顺序运行最容易定位问题。

把看不见的反事实画出来

我们观察到了新泽西政策后的实际均值，却观察不到同一批门店在同一时期“没有这次调薪”时的结果。两期DID用处理组政策前均值加上对照组同期变化，构造反事实后期均值。本例约为20.43940810加负2.16558442，即18.27382368。实际后期21.02742947与它的距离仍是2.75360578。

趋势图中黑实线连处理组观察值，灰虚线连对照组观察值，点线表示平行趋势假设下的处理组反事实。点线不是额外采到的第三组数据。纵轴注明每店平均全职当量，并从16起绘制可读刻度；读图时应看刻度与数值，不能把线条角度或线间距离当成脱离坐标轴的证据。

平行趋势要求无政策时两组的平均变化相近，并不要求政策前水平相同。起点20.44和23.33不同本身不能否定DID；只看两条观察线在政策后分开也不能验证假设。当前只有一个政策前时点，没有前期趋势序列，不能用这四格做政策前趋势检验。补充更多政策前资料有助诊断，但“前期未显著不同”也不能证明不可观察的反事实一定成立。

用交互回归核对同一个点估计

模型包含四项：常数、处理组指示、后期指示、处理组与后期的乘积。把对照组政策前作为基准，常数对应对照组前期均值；处理组项对应前期组间差；后期项对应对照组变化；交互项就是两组的变化差。可以把四种0/1组合代入模型，按四格逐一检验每项承担的作用。

Stata命令regress avgfte ib0.treat##ib0.post显式把0设为基准。双井号会加入两个主项及交互项；只写一个井号或只手工加入treat乘post，可能漏掉用于区分初始组间差和共同时间变化的项。运行后读取_b[1.treat#1.post]，与前面保存的did比较。脚本的assert要求二者差小于很小的数值容差，Python也用四行设计矩阵独立核对。

本课四个汇总观察拟合四个参数，残差自由度是0。回归表中的标准误、t值或区间缺失是这个饱和练习的结构，不是通过换一个vce选项就能补出的信息。不要复制四行很多遍制造样本量，也不要将有效门店数当频数权重就声称恢复了原始门店的不确定性。四格均值没有提供组内方差、前后协方差和抽样设计所需信息。

从算对数字到写对研究结论

可先写一条描述性结果：“按本课随包的四格汇总，新泽西门店的平均全职当量变化比宾州东部高约2.7536。”如果进一步写成政策影响，应说明平行趋势、无差异性同期冲击、无影响比较的溢出、无提前反应，以及样本构成和测量口径可比等条件。这些条件不能靠一个显著性标记代替。

本课已有汇总算得2.75360578，原论文表3报告2.76。两者保留各自口径，不能为了让数字完全相同改动原数据。2.75360578按两位小数通常显示2.75，因此差异不能简单说成这一数值四舍五入就等于2.76。当前包只交付既有完整分项口径的汇总，没有重做论文全部微观筛选、平衡样本和标准误程序。核对精度、分母和样本定义比抹平差异更重要。

样本范围是这次调查的连锁快餐门店和观察期。用“所有餐饮企业都多雇了2.75人”表述，会同时扩大适用总体、混淆全职当量与人数，并省去识别条件。作业中的文字说明需要人工判断；网页数值核对通过，只说明所填结果与指定练习相符，不能自动证明因果设计有效或论文结论成立。

改做：换一个明确标注的对照趋势

现在把计算中的反事实假设拿出来观察。打开反事实情景.csv，看到假设对照组变化为-1。它不是从真实调查估计出的新值。保持新泽西两格不变，假设宾州东部每店只减少1个全职当量，新的反事实后期就是20.43940810减1，新的DID就是0.58802137减负1，约为1.58802137。

先独立写出两个算式，再运行脚本的情景步骤。Stata使用preserve和restore暂存并恢复真实四格，情景表只有一行；读到的数值先存成scalar，再计算。Python单独读取情景文件，并把这一行结果明确标成课堂假设。检查真实两州CSV的分母和总量仍未改变，这是区分观察资料和反事实练习的基本习惯。

完成指定作业后，可以复制情景文件的备份，把假设变化改为0再运行，理解为什么此时DID等于处理组前后差。不要去改真实调查的宾州总量来假装这一情景发生过。网页指定作业仍按-1的原始情景核对，探索其他数值时应在自己的答案中记录改动参数、结果和假设性质，完成后恢复原情景以便复核。

独立做：换一份门店面板重建四格

独立练习_模拟门店.csv包含8家虚构门店、两期共16行。它由本课原创构造：试点组各店增加5，参照组各店增加2，意在检验你是否能从更细的记录重新计算，而不是提供一项真实政策证据。先复制迁移练习起步.do，按注释补命令，暂时不要打开迁移练习参考.do。

先用门店编号与时期检查联合唯一性，再确认每家门店有两期、组别不会在两期之间改变。这里每个格子都有4家门店，按组别和时期求简单平均即可。处理组均值从13变到18，对照组从11变到13。独立算出处理组变化5、对照组变化2和DID为3，再用交互回归核对。Stata参考脚本先collapse成四格，仍只作点估计演示。

如果得到2或5，分别检查是否只提交了对照变化或处理变化；如果得到-3，检查处理组编码和减法方向；如果记录数是8，检查是不是只统计了门店数而题目要求门店-时期记录数。本例没有缺失和进入退出。真实研究中这些问题可能改变观察组成，不能为凑成平衡面板就把未观察格填0，也不能默认删掉不齐全门店后仍估计同一个总体。

提交、复核与下一步学习

交付至少包含输入版本说明、可重跑的代码、四格结果表，以及说明单位、资料性质和识别边界的一段文字。真实主例与假设情景、模拟迁移分别命名，不能把它们叠在一张表里却省略性质。核对数值时保留足够精度，展示时可写四位小数。把代码中的路径设为实际解压目录，不把个人电脑的绝对路径当成别人也具备的环境。

按跟做、改做、独立做的次序完成三项作业。参考答案解释了为什么得到这些值；如果数字不同，先查分母和四格是否完整，再查处理组、后期及相减方向，最后检查是否提前取整。网页会核对数值并给出检查提示；能否说明反事实条件、是否混淆观察单位、代码能否从输入重跑，仍需按人工评分标准逐项检查。

完成后可以继续学习政策文件编码、更多前期的事件图，以及分批实施政策的比较组选择。两组两期的简洁公式不能直接替代多期研究设计。先把当前一个问题讲清楚：这次比较用谁作参照、凭什么作参照、估计量的单位是什么，以及有哪些问题这份资料还回答不了。保留出错日志与修复记录，会比只保留最后一张漂亮表更有助于下一次研究。

本课软件与五步学习说明

主软件：Stata

软件准备入口：/learning-setup#stata

1. 选择本课软件路线

本课主路线为Stata。软件准备指南：/learning-setup#stata。只准备本路线需要的工具。

2. 完成一次学习准备

先检查已有软件。Python主课或选做Python独立复核时，首次/修复使用prepare-python.cmd（Windows）或prepare-python.command（Mac）；日常打开使用open-python对应入口。其他软件按各自指南准备，不先装Python。

3. 第一次跑通并保存

先完成该软件指南里的小任务：打开示例数据、执行一次操作、看到结果、保存文件并找到它。只有安装成功还不能算跑通。

4. 打开本课项目

完整解压课程包。打开课堂练习.do、迁移练习参考.do、迁移练习起步.do。课程输入和作业分别保存；兼容课程复用已准备环境，不在每节课重复安装。

5. 按步骤学习并改做

按本课分步实操运行，逐项核对预期结果。再复制课程目录，修改参数或换用独立作业数据；保留基准版本，不覆盖原始资料。

可选Python独立复核

选做数值对照时，可以运行随包的 Python 代码。

本项目额外要求与平台范围

本轮Python原脚本的既有运行记录继续保留；新界面或其他平台的验证以本次验收记录为准。

商业软件需使用自己的有效授权；课程材料不含许可证。

Python固定依赖组：empirical-models

原脚本：课堂练习.py

分步笔记本：课堂学习.ipynb

原生文件

stata：课堂练习.do

stata：迁移练习参考.do

stata：迁移练习起步.do

参考资料

Card 与 Krueger（1994）论文及表 3 https://davidcard.berkeley.edu/papers/njmin-aer.pdf

作者公开数据入口 https://davidcard.berkeley.edu/data_sets.html