

试讲课程 · 教案

用 AI 做经营分析，三个常见的误判

副题：从毛利、回款、费用三个案例，看 AI 的结论错在哪

Three Common Misjudgments in AI-Assisted Business Analysis

杨浩 (August) | 2026 年 10 月

课名为工作标题。对外宣传用的标题，在口播稿和宣传阶段另定。

教案 | 2026-10 v2 变更：教学过程的章节时刻表与交付物规格已同步实际成片（16 分 36 秒，5 章）；内容框架与 v1 一致 本稿是后续口播稿、视频、网页、学员讲义的唯一上游。案例数据均为教学虚构，计算已程序验算；学术出处见文末参考文献（APA 第 7 版）。

一、课程卡片

项目	内容
课程名	用 AI 做经营分析，三个常见的误判
受众	在职财务 BP；希望转岗 BP 的会计、财务分析人员
时长	视频成片 16 分 36 秒（5 章）；现场试讲 30 分钟以内，余量用于互动与问答；讲不完的内容引导学员看讲义
形式	口播 + 数据动画；每章一个“先押答案”的暂停点
先修	读得懂毛利率、费用率；用过 Excel；不需要统计学基础
交付物	教案（本稿）、逐字口播稿、视频、互动网页、学员讲义 PDF
核心产出	学员带走三句追问、两种拆解方法、一份给 AI 的复核指令

课程定位：本课是“AI+财务 BP 能力体系”中数据推理模块的第 1 讲。

模块	主题	与本课的关系
一	用 AI 出数：大模型与表格工具的财务用法	前置：学员已经在用 AI 出表
二	数据分析复核：口径与推断	本课是第 1 讲
三	搭建财务分析智能体：提示词、工作流、RAG 知识库	后续：三个追问写进智能体的复核节点
四	经营分析会实战：把口径讲给业务听	后续：口径披露进入会议材料模板

二、选题依据

一份经营分析报告，过去要三天。这三天里，BP 有充裕的时间在底稿里发现问题。现在 AI 十分钟就能出稿，图表漂亮，结论通顺，留给 BP 发现问题的时间，往往只剩会上那十分钟。

这类报告很少错在算术上，错在算术之前的三个决定：数字按什么方式合并，哪些对象被算了进来，拿什么作对比基准。财务把这三个决定叫“口径”；研究方法论把它叫操作性定义（operational definition）：一个指标在被计算之前，必须先规定它怎样被测量。

AI 工具会替你默默做掉这些决定。pandas 的 `mean()` 默认跳过缺失值（`skipna=True`），Excel 的 `AVERAGE` 默认忽略空白单元格、但计入数值 0（pandas development team, 2026; Microsoft, n.d.）。决定被写进了默认参数，人就意识不到那里曾经有一个决定。人因工程研究把这种风险称为自动化偏差（automation bias）：人把自动化系统的输出当作省力的替代品，系统出错时，要么漏掉本该发现的问题，要么照着错误的建议去做（Parasuraman & Riley, 1997; Parasuraman & Manzey, 2010）。输出越流畅，越容易滑过去。

BP 的职责定义恰好落在这里。CIMA 与 AICPA 把财务业务伙伴定义为：将会计专业与业务理解结合，提供分析与洞察，以影响决策和绩效管理（CGMA, 2015）。财政部《管理会计基本指引》第三条也写明，管理会计要“参与单位规划、决策、控制、评价活动并为之提供有用信息”（财政部, 2016）。当出数的边际成本被 AI 压到接近零，BP 的价值就集中到一件事上：判断这个数能不能拿来做决策。

所以本课只讲一件事：用 AI 做经营分析时，结论上三个常见的误判。三个误判各对应方法论文献里一个成熟的问题族：聚合偏差、选择偏差、因果识别。课上把它们翻译成三句 BP 在会上就能问出口的话。

三、学情分析

维度	判断	教学上的对策
已有基础	熟悉报表与毛利率、费用率；多数人在中级或注会《财务管理》里学过因素分析法（连环替代法）	误判一从连环替代法切入，把旧知识直接升级为结构分解，降低陌生感
薄弱环节	缺少统计推断与因果识别训练；会算，不会追问	不讲公式推导，讲“该问什么”；学术名词只当锚点，放在讲解之后
新变量	已经或即将用 AI 出表，不清楚工具的默认口径	每个误判都配一个“AI 在这里替你做了什么决定”的拍点
学习动机	成人学习者以问题为中心，愿意学马上能用在工作上的东西（Knowles et al., 2025）	全课锚定一份真实感的 AI 经营分析报告，案例即工作场景
风险	学完变成“什么都怀疑”，在会上与业务对立	收束段明确：三个追问是程序，问过没毛病的数就放心用

四、教学目标

按修订版布鲁姆分类（Anderson & Krathwohl, 2001）设定，课后练习逐条对应。

层级	目标：学完后，学员能够	对应检测
理解	说明聚合、筛选、对照三类口径决定，如何让“每个数都对”的分析得出错误结论	练习 1-5 的追问环节
应用	用连环替代法或对称分解，把毛利率变动拆成构成效应与率效应	练习 1
分析	识别报表中被静默排除的样本（右删失、幸存者偏差），说出它对结论的偏向	练习 2、3
评价	用对照组与双重差分判断一项“改善”有多少是干预带来的	练习 4
创造	为 AI 助手写一份复核指令，要求它在下结论前先披露口径	练习 5

态度目标：形成“先审口径、再下结论”的职业习惯，同时不滑向“凡数皆疑”。

五、教学重点与难点

重点有三：三个误判各自的追问；构成效应与率效应的分解；固定分母的队列口径与对照组口径。

难点有三，各配一个突破手段。

难点	为什么难	突破手段
合并数与分组数谁对	直觉想要一个“正确答案”，而答案取决于决策问题与因果结构（Pearl, 2014）	同一张表算出三个数，再用“你要回答什么问题”把三个数各自归位
回归均值的机制	违反“有变化就有原因”的直觉	拆成“实力 + 运气”：挑出最差的一刻，挑中的往往是运气最差的一刻
删失为什么有方向	“已回款的平均天数”看上去没毛病	一个具体问题：9 月 25 日开的票，要在 9 月 30 日前出现在“已回款”里，它得多快

六、教学设计依据

每个设计决定都有出处。下表是本课的“设计说明书”，也是面向面试官的教研能力说明。

设计决定	依据	在本课中的落地
一个贯穿全课的锚定情境	抛锚式教学：把学习锚定在真实、足够复杂的问题情境中，让学习者明白知识何时有用（Cognition and Technology Group at Vanderbilt, 1990）	全课只用一份 AI 生成的《禾川饮品 Q3 经营分析》，三个误判各拆其中一条结论
每个误判先押答案再揭晓	预测测试效应：即使预测测试答错，也能提升随后的学习（Richland et al., 2009; Kornell et al., 2009）；有效失败（Kapur, 2008, 2016）	每章先由讲者口播设问，学员押完“信不信这条结论”再揭晓（视频成片按此执行；现场试讲改为举手表决）
揭晓按“预测—观察—解释”组织	POE 策略（White & Gunstone, 1992）	押答案 → 看拆开的数据 → 解释机制 → 给学名
误判—完整演示，练习逐步放手	样例效应（Sweller & Cooper, 1985）；认知负荷理论（Sweller, 1988）	误判—逐步演示分解；练习 1 换场景、只给框架

设计决定	依据	在本课中的落地
视频分 5 章	MOOC 数据：观看时长中位数约 6 分钟，与视频总长无关（Guo et al., 2014）；分段原则（Mayer, 2020）	章节标记 + 学员自控进度
口语化讲解，学术名词后置	个性化原则：对话式语言优于正式语言（Mayer, 2020）；教学视频设计指南（Brame, 2016）	先讲人话，讲清后才说“这叫什么”
画面只留必要信息	一致性原则与提示原则（Mayer, 2020）	屏幕常驻“三个误判”进度条；无装饰性动画
课后检索练习	测试效应：测验比重复阅读带来更好的长期保持（Roediger & Karpicke, 2006）	网页 5 题即时反馈；讲义附练习与解析
过程按教学事件编排	加涅九大教学事件（Gagné et al., 2005）	第九节时间轴逐段标注事件编号
评估分四级	柯氏四级评估（Kirkpatrick & Kirkpatrick, 2016）	第十一节

七、核心框架：三个误判

一个指标在被计算出来之前，已经被三个决定塑形。三个误判，就是这三个决定各自最容易做错的地方。

误判	BP 追问	方法论问题族	拆解工具	AI 在这里默认替你做了什么
误判一	这个数怎么合起来的？	辛普森悖论、构成效应	构成效应与率效应分解（Kitagawa, 1955）	选了一种平均方式（加权或不加权），且不告诉你
误判二	这个数算了谁，漏了谁？	幸存者偏差、样本选择偏差、右删失	固定分母的队列口径；意向性治疗原则	缺失值、空白、未完成的记录被静默跳过
误判三	这个数跟谁比？	回归均值、反事实、内部效度威胁	对照组与双重差分	默认拿“前后对比”当效果，不构造反事实

三个误判与 Shadish et al.（2002）列出的内部效度威胁有直接对应：误判二对应选择（selection）与流失（attrition），误判三对应回归假象（regression artifacts）与历史事件（history）。误判一则对应统计学里最古老的悖论之一，Simpson（1951）把它写清楚，Blyth（1972）给它定名；更早的 Yule（1903）已经记录过合并后关联消失的现象，所以文献里也叫尤尔-辛普森效应。

课上不展开这张对应关系，只在每个误判讲完时点一句学名。讲义里给全表。

八、锚定情境：一份 AI 写的经营分析

禾川饮品（虚构）是一家区域性饮料公司，两个品类：功能饮料与包装水；全国 40 个销售区域，经销商渠道，按票据回款。周一 9 点经营分析会，大屏上是 AI 助手生成的《Q3 经营分析》，摘要三条：

结论一：Q3 整体毛利率 34.5%，同比下降 3.5 个百分点，主因渠道价格战侵蚀单品毛利。建议申请 Q4 价格补贴预算。

结论二：Q3 平均回款周期 38 天，较 Q2 的 41 天缩短 3 天，回款管理持续改善。

结论三：Q3 对差旅超支最严重的区域实施专项管控，超支率由 36% 降至 14%，降幅 22 个百分点。建议 Q4 推广至全国。

三条结论里的每一个数字，都核得上底稿。课程的全部工作，是找出三条结论背后各自藏着的那个决定。

底稿数据如下（单位：万元；超支率 = （实际差旅费 - 预算）÷ 预算）。

表 A 品类毛利（对应结论一）

品类	去年 Q3 收入	去年 Q3 毛利率	今年 Q3 收入	今年 Q3 毛利率
功能饮料	6,000	50%	5,400	51%
包装水	4,000	20%	6,600	21%
合计	10,000	38.0%	12,000	34.5%

表 B 票据回款（对应结论二，截至 9 月 30 日）

队列	开票张数	已回款	已回款平均天数	未回款	未回款平均已挂账天数
Q2 开票	100	100	41	0	—
Q3 开票	100	70	38	30	55（仍在增加）

表 C 差旅专项管控（对应结论三）

Q2 超支最严重的 20 个区域，因管控人力有限分两批：按区域编号单双号，10 个在 Q3 实施，10 个排期 Q4。

组别	区域数	Q2 超支率	Q3 超支率	变动
管控组（Q3 实施）	10	36%	14%	-22 pt
待管控组（Q4 实施）	10	35%	21%	-14 pt

AI 报告只引用了表 A 的合计行、表 B 的“已回款”列、表 C 的管控组。没有一处算错，三处都只看了表的一部分。

九、教学过程

口播稿约 3,700 个汉字，实际录音 15 分 35 秒，加片头、片尾与展示留白，成片 16 分 36 秒（下表时刻与成片章节标记一致）。本节的讲解要点比口播多：口播只保留每个误判的主线（押答案、揭晓、机制、一个学名、一句追问），其余细节（肾结石数据、固定观察窗与 Kaplan-Meier、意向性治疗原则、Galton 原始数据、反事实与双重差分的文献）全部放进学员讲义。每章标注对应的加涅教学事件（G1-G9）。“关键句”是口播稿必须保留的原话，其余由口播稿展开。

章	时段	时长	内容	教学事件
0	0:00-3:32	3:32	开场：十分钟的报告	G1 引起注意、G2 告知目标、G3 激活旧知
1	3:32-8:00	4:29	误判一：合计跌了，分项都在涨	G4-G7
2	8:00-11:08	3:08	误判二：只算了回来的钱	G4-G7
3	11:08-14:29	3:21	误判三：把自然回落当成管控效果	G4-G7
4	14:29-16:36	2:07	收束：三句话，一份指令	G8 评估、G9 保持与迁移

第 0 章 开场：十分钟的报告（0:00-3:32）

画面：讲者出镜做一句话自我介绍；随后三条 AI 结论逐条上屏。

讲解要点：

1. 一句话自我介绍。“我是 August，一个喜欢研究 AI 的财务 BP。”不讲履历，不铺场景，口吻像财经科普博主那样平易近人。
2. 先亮报告。把禾川 AI 报告的三条结论直接念给学员：有数字，有原因，有建议，看起来更美好，数字也全都核得上。然后问“这三条算不算业务洞见，能不能照着去做？”，答案是不能。三个误判各拆一条，收尾回到这个问题。
3. 点明为什么是 AI 时代的问题。AI 让从取数到清洗合并的这套 ETL 变得很容易，也能很快标出数字上的异常；难的是解释“为什么变”。两件事同时出现，AI 很容易把相关写成果。不加分辨地照单全收，分析就会误判。
4. 告知目标。课后带走三句话：怎么合的？算了谁？跟谁比？外加两种拆法和一份给 AI 的指令。
5. 激活旧知。“你在财管里学过的连环替代法，就是今天的第一把刀。”

关键句：「以前写一份经营分析，BP 得自己拉数对数，一页一页翻底稿，忙上三天，很多问题就是在这个过程里发现的。现在 AI 十分钟就出稿，这个过程省掉了，问题只能靠你看稿的时候发现。」

设计意图：先亮出“看起来更美好”的三条结论，再问它们算不算业务洞见，用认知冲突引起注意；锚定情境同时完成告知目标；连环替代法是受众普遍学过的方法，提前点名，降低误判一的陌生感。

第 1 章 误判一：合计跌了，分项都在涨（3:32-8:00）

暂停卡 1：「结论一说‘价格战侵蚀单品毛利’。押一个：信，还是不信？」（倒计时 5 秒）

画面与讲解：

1. 摊开表 A。两个品类毛利率都涨了 1 个点，合计却跌了 3.5 个点。三个数都对，放在一起互相矛盾。
2. 同一张表算出三个答案：按收入加权，毛利率 38.0% → 34.5%，跌 3.5 个点；两个品类简单平均，35% → 36%，涨 1 个点；毛利额 3,800 万 → 4,140 万，多赚 340 万。哪个是真相？先别选，三个数分别在回答三个问题。
3. 用连环替代法拆开。先把结构换成今年、毛利率仍用去年： $0.45 \times 50\% + 0.55 \times 20\% = 33.5\%$ ，比 38.0% 低 4.5 个点，这是构成效应；再把毛利率换成今年：34.5%，比 33.5% 高 1.0 个点，这是率效

应。 $-4.5 + 1.0 = -3.5$ ，缺口分毫不差。本例两个品类涨幅相同，换替代顺序或用对称分解结果一致；一般情况下顺序会影响结果，讲义练习 1 专门演示。

4. 结论翻转。单品毛利一点没被侵蚀，两个品类的毛利率都在涨。变的是卖什么：包装水的收入占比从 40% 升到 55%。该追问的是：低毛利品类放量，是主动抢份额，还是被渠道推着走？申请价格补贴，是在补一个不存在的窟窿。
5. 点学名。这就是辛普森悖论：分组看一个方向，合并看另一个方向。医学统计里被引用最多的例子是肾结石治疗：按结石大小分开统计，开放手术两组都胜出（小结石 93% 对 87%，大结石 73% 对 69%）；不分大小合起来算，经皮肾镜取石反倒领先（83% 对 78%）。原因在病例分配：医生倾向于把更难处理的大结石交给开放手术（Charig et al., 1986; Julious & Mullee, 1994）。
6. 谁对？Pearl（2014）的回答是：看因果结构，公式本身判断不了。落到 BP 场景：问定价与成本管控做得好不好，看分组；问公司整体赚钱能力，看合并；问要不要调结构，看分解。
7. AI 拍点。让 AI “算一下各区域平均毛利率”，它可能给加权平均，也可能给简单平均，然后照着那个数写出一段流畅的归因，不会告诉你它选了哪种。

关键句：「要看你想回答什么问题。」

设计意图：误判一是完整样例，分解步骤逐步演示，为后两个误判建立“拆开再看”的模式。构成效应与率效应的分解源自人口学的 Kitagawa（1955），在管理会计里对应销售组合差异，在 FP&A 里常叫价格—数量—结构分析（PVM）；课上只说“结构变了还是效率变了”，名称留给讲义。

第 2 章 误判二：只算了回来的钱（8:00–11:08）

暂停卡 2：「回款周期从 41 天缩短到 38 天。押一个：真改善，还是假改善？」

画面与讲解：

1. 摊开表 B。38 天只来自已回款的 70 张票，还有 30 张没回款，平均已挂账 55 天，而且每天都在加。
2. 问一个具体问题：9 月 25 日开的票，要出现在“已回款”那一列，它得在 5 天内回款。离截止日越近，能进表的门槛越高，进来的只能是回款快的。表里的 38 天，是被截止日筛过一遍的 38 天。
3. 补回分母。把 30 张未回款按今天的挂账天数算，假设它们明天全部回款： $(70 \times 38 + 30 \times 55) \div 100 = 43.1$ 天。这只是下限，已经比 Q2 的 41 天更长。Q2 的 41 天是 100 张票全部回款后的完整数，Q3 的 38 天是一个还没长完的队列。拿一个完整的数跟一个没长完的数比，“改善”是口径制造出来的。
4. 正确的比法：同一套规则看两个队列。比如“开票后 45 天内回款率”，Q2、Q3 都只统计开票满 45 天的票。需要精确估计未完成的部分，统计学有专门的工具：右删失（right censoring）与 Kaplan-Meier 乘积极限法，它们本来就是为“观察期结束时事件还没发生”设计的（Kaplan & Meier, 1958）。
5. 点学名。这是选择偏差的一种，与幸存者偏差同源：只看留下来的，结论系统性偏好。金融研究里的经典证据：只用仍存续的基金做业绩研究，会把基金业绩每年高估约 0.9 个百分点，因为业绩差的早已清盘、退出了样本（Elton et al., 1996; Brown et al., 1992）。
6. AI 拍点。回款日期为空的行，相减得到缺失值；pandas 的 `mean()` 默认跳过缺失值，Excel 的 `AVERAGE` 默认忽略空白单元格。不报错，不提示，30 张票安静地离开了分母。

7. 一条可以直接搬走的原则。临床试验的意向性治疗原则（intention-to-treat）：所有被随机分组的对象，都按最初分组纳入分析，不管中途是否退出（Hollis & Campbell, 1999）。翻译成 BP 的话：分母在开头定下，谁都不许中途离场。

关键句：「看一张表，别只看表里有什么，还要问问哪些东西没进来。」

设计意图：删失是三个误判里最反直觉的一个，用“9 月 25 日那张票”把抽象机制压缩成一个可以心算的具体问题。工具默认口径放在本章讲，因为缺失值被跳过正是筛选的技术实现。

第 3 章 误判三：把自然回落当成管控效果（11:08–14:29）

暂停卡 3：「管控让超支率降了 22 个点。押一个：真实效果比 22 大、差不多，还是小很多？」

画面与讲解：

1. 先讲机制，用“实力 + 运气”。一个区域某季度的超支 = 结构性需求 + 一次性因素（一场展会、一批集中报销）。挑出 Q2 最严重的 20 个，挑中的往往是一次性因素最集中的那一刻。下一季运气回到平常，数字自己就会往回落。
2. 摊开表 C。AI 只引用了管控组。看没被管的那组：同样严重，一分钱没管，也从 35% 降到 21%，降了 14 个点。
3. 双重差分。管控组降 22 个点，待管控组降 14 个点，管控的净效果是 $(-22) - (-14) = -8$ 个点。22 个点里约 14 个点，不管控也会发生。待管控组同时吸收了两种干扰：回归均值，以及 Q3 本身可能是差旅淡季这类共同冲击。
4. 点学名。回归均值（regression toward the mean）。“回归”一词就出自 Galton（1886）的身高研究：子女身高偏离均值的幅度，平均约为双亲中值偏离幅度的三分之二。管理场景里最传神的例子来自 Tversky 和 Kahneman（1974）：飞行教官发现，表扬一次漂亮的降落，下一次往往变差；批评一次糟糕的降落，下一次往往变好，于是认定批评比表扬管用。其实不管表扬还是批评，极端表现之后都会往均值回落。专挑最差的部门帮扶，再拿帮扶后的回升当成效，是同一个结构。注意区分：金融里的“均值回归”（mean reversion）是另一个概念。
5. 反事实。效果的定义是“做了之后的结果”减去“同一批对象如果没做会怎样”（Rubin, 1974）。后一项永远观察不到，只能找最像的一组来代替，这就是对照组。用两组前后变化之差估计效果的做法叫双重差分，Card 和 Krueger（1994）用它研究最低工资对就业的影响，成了经济学的标准方法。
6. 回到决策。管控的真实效果是 8 个点。要拿去跟管控成本比的是这 8 个点。8 个点可能依然值得推广，结论变成“先算账再推广”。
7. 可以直接搬走的设计原则：先设计评估，再上线方案。分批上线就是一个免费的对照组，用抽签或编号决定先后，不要按“谁最差”排先后。
8. AI 拍点。AI 默认只做前后对比，你不要求，它不会去找对照组。

关键句：「没有对照组，这个“效果”就没法算。」

设计意图：本章不否定管控，只修正效果的大小，避免学员把“复核”理解成“拆台”。双重差分只用一次减法，财务人员可以当场心算。

第 4 章 收束：三句话，一份指令（14:29-16:36）

画面与讲解：

1. 把三条结论重写成“BP 版”，口径写在结论前面：

- 结论一：毛利率同比 -3.5 个点，其中构成效应 -4.5 个点（包装水收入占比 40% → 55%），率效应 +1.0 个点。单品毛利未被侵蚀。价格补贴暂缓，先复盘包装水放量是不是既定策略。
- 结论二：Q3 队列尚未完整，38 天不能与 Q2 比；计入未回款的下限为 43.1 天，已长于 Q2。改用“开票后 45 天回款率”按队列追踪，30 张未回款票据进入催收清单。
- 结论三：以待管控组为对照，管控净效果约 8 个点，约 14 个点为自然回落。按 8 个点测算收益后再决定是否推广；推广时继续分批，保留对照。

2. 三句话上屏（见第十节一页纸），每句配一个本课的数字：-4.5、43.1、8。

3. 展示给 AI 的复核指令（见第十节），说明它就是第三模块“财务分析智能体”复核节点的雏形。

4. 布置迁移任务：回到工作中，挑一份最近的经营分析，用三句话过一遍，记下哪一句问出了问题。

5. 引导读讲义。课上只讲主线，完整数据、算法步骤、概念出处、26 篇参考文献、两篇延伸阅读、AI 复核指令和五道练习都在讲义里。现场超时，第 2 章 Kaplan-Meier、第 3 章平行趋势假设各压成一句“讲义里有”。

关键句：「AI 出数越快，越要在用之前多问一句。以后会上再看到一条特别顺的结论，先把三句问一遍：怎么合的？算了谁？跟谁比？」

设计意图：重写结论即 G8 评估表现的示范，三个数字作为提取线索帮助保持；迁移任务对应柯氏第三级行为层的追踪起点。

十、一页纸与 AI 复核指令

三个误判一页纸

误判	一句话追问	看到什么就要警觉	怎么拆	本课的数
误判一	怎么合的？	合计数与分组数方向相反；“平均毛利率”“平均费用率”	构成效应 + 率效应	-4.5
误判二	算了谁，漏了谁？	比较的两期成熟度不同；空白、缺失、“已完成”口径	固定分母、固定观察窗	43.1
误判三	跟谁比？	“实施后改善了 X”；对象是按最差、最好挑出来的	对照组 + 双重差分	8

给 AI 的复核指令（可直接粘贴）

你在为财务 BP 准备经营分析。给出任何结论之前，先逐项披露下列口径。无法确定的写“未确定”，不要猜。

【聚合】

1. 每个比率指标的合并方式：加权平均（写明权重）还是简单平均。
2. 合并数与分组数方向不一致时，把变动拆成构成效应与率效应，分别报告。

【筛选】

3. 分析对象的起始名单与条数、最终进入计算的条数，以及差额的原因（缺失值、未完成、已剔除）。
4. 缺失值与空白的处理方式；不得静默跳过。
5. 被比较的两期是否处于同样的观察成熟度；不同时，改用固定观察窗口。

【对照】

6. 任何"改善""见效""拉动"类结论，写明对照组是谁；没有对照组时，标注"仅前后对比，不能归因"。
7. 分析对象若因极端表现入选（最差或最好的 N 个），提示回归均值的可能。

最后单列一段：结论的可信度，以及需要人工核实的地方。

十一、评估设计与课后练习

柯氏四级评估

层级	测什么	怎么测
一 反应	相关性与可用性	网页末尾 3 题问卷：与工作相关吗？一周内会用吗？有信心在会上问出口吗？
二 学习	能否识别与拆解	三张暂停卡即前测；练习 1-4 为同构变式后测，比较正确率
三 行为	是否迁移到真实工作	30 天后提交一条"追问记录"：一份真实分析、问了哪句、发现了什么
四 结果	对组织的影响	企业内训场景：分析报告因口径问题返工的次数、决策被修正的案例数

练习（讲义与网页共用，附解析）

练习 1 误判一 | 应用。某 SaaS 公司两类客户：大客户收入占比 70% → 50%，毛利率 62% → 64%；中小客户收入占比 30% → 50%，毛利率 42% → 45%。整体毛利率变动多少？拆成构成效应与率效应。

解析：整体 56.0% → 54.5%，-1.5 个点。先换结构后换率：构成 -4.0，率 +2.5；先换率后换结构：率 +2.3，构成 -3.8。两种顺序结果不同，这就是连环替代法“替代顺序影响结果”的性质；对称分解（取两期均值作权重，Kitagawa, 1955）给出构成 -3.9、率 +2.4，不依赖顺序。结论不受影响：两类客户毛利率都在涨，整体下滑来自中小客户占比上升。

练习 2 误判二 | 分析。创新项目复盘会上的汇报：“近两年结项的 12 个创新项目，平均 ROI 46%。”你要先补问哪个数？

解析：同期立项的项目总数，包括中途终止的与仍在进行的。终止项目的 ROI 多数为负，它们不在“结项”名单里。正确口径是按立项队列统计，终止项目计入分母，在途项目单列或按删失处理。

练习 3 误判二 + 误判一 | 分析。AI 助手给出两行代码。它们各替你做了什么口径决定？

```
df.groupby("区域")["毛利率"].mean()
(df["回款日期"] - df["开票日期"]).dt.days.mean()
```

解析：第一行对每张票的毛利率做简单平均，没有按收入加权，大单与小单一票一权（误判一）；区域毛利率应为毛利合计 ÷ 收入合计。第二行中未回款票据的回款日期为空，相减得到缺失值，`mean()` 默认 `skipna=True` 将其跳过（误判二）；应先报告缺失条数，再改用固定观察窗口。

练习 4 误判三 | 评价。客诉率最高的 15 家门店上月参加服务培训，本月客诉率下降 30%。区域经理建议全面推广。你会怎样设计评估？

解析：门店按“最差”入选，回归均值会制造一部分下降。方案：从客诉率最高的 30 家中抽签选 15 家先培训，另 15 家下月培训；比较两组同期变化之差（双重差分），同时排查新品上市、节假日等共同冲击。

练习 5 综合 | 创造。以第十节指令为底稿，为你所在的业务写一份 AI 复核指令，并补一条行业专属口径（如零售的同店口径、SaaS 的 ARR 口径）。

评分要点：三个误判各有至少一条可执行的指令；措辞是“披露”“报告”，不是“注意”；规定了无法确定时怎么处理；行业口径写清分子、分母与时间窗口。

十二、交付物规格

本稿定稿后，按下表生产其余四件交付物。四件共用同一套案例数据、三句追问和视觉识别。

交付物	规格	要点
逐字口播稿	约 3,700 字，开头一句话自我介绍，随即亮出三条 AI 结论；每个误判开头先用大白话重述 AI 的说法，指标先解释再给数；按第九节五章切分；细节移入讲义	按课堂讲授写，不写分镜和剪辑标注；停顿、提问、悬念直接写进说的话里；第二人称“你”；学术名词先讲现象、后给名字；保留第九节全部关键句；在对应位置引导学员看讲义
视频	16:9, 1080p25, 16 分 36 秒, 5 个章节标记	讲者口播 + 数据动画；片头片尾配乐；中文字幕内嵌；作者与版权信息写入文件元数据
互动网页	单文件 HTML，离线可用	三个误判各一屏，“押答案 → 揭晓数据 → 解释机制”；两个小工具：构成/率效应分解器、双重差分计算器；5 题即时反馈测验；反应层问卷；讲义下载入口；键盘可操作、对比度达标
学员讲义 PDF	22 页，A4；7 张数据图 + 5 张插图	封面；三个误判一页纸；三个案例的完整数据与拆解步骤；学术概念卡（中英术语 + 出处）；工具默认口径速查（pandas / Excel）；AI 复核指令；练习与解析；延伸阅读（第十三节）；术语表；参考文献；封底（作者与联系方式）

视觉识别（讲义已落地，视频沿用）：期刊编辑风格。浅灰纸底（#F0F0EE）配墨黑文字，单一强调色橙（#F5572F）；标题用衬线体（思源宋体），正文用 Inter 加思源黑体，数字用等宽数位；数据图只用灰阶加一处橙色，每张图只把橙色给结论所在的那一个数；插图同色系扁平编辑插画，不含文字，不用装饰性图标。

现场试讲版：暂停卡改为举手表决，揭晓时报出表决分布；时间超出时，第 2 章第 4 点（固定观察窗与 Kaplan-Meier）压缩为一句，细节移到讲义。

十三、延伸模块（只进讲义，不进视频）

延伸 A 平均值缺陷：按平均需求做计划

禾川饮品 Q4 一款新品，产能上限 10 万箱，每箱毛利 20 元。需求三种情景等概率：6 万、10 万、14 万箱，平均正好 10 万箱。

按平均需求做计划： $20 \times \min(10, 10) = 200$ 万元。

逐情景计算再取期望： $(20 \times 6 + 20 \times 10 + 20 \times 14) \div 3 = 173.3$ 万元。需求 14 万箱的情景里，多出的 4 万箱卖不出去，产能把好运气截掉了，坏运气却照单全收。

按平均做计划高估 26.7 万元，约 15.4%。数学上这是琴生不等式（Jensen, 1906）：当结果是输入的凹函数时，函数的期望不超过期望的函数。Savage（2009）把它称为平均值缺陷的强形式（The Flaw of Averages, strong form）。与之相关的另一种情形是两列相关的数各自取平均再相乘： $E[XY] = E[X] \cdot E[Y] + \text{Cov}(X, Y)$ 。协方差为负，平均值之积大于积的平均；协方差为正则相反。实务对策是情景测算或蒙特卡洛模拟，用分布做计划，不用单点。

延伸 B 威尔·罗杰斯现象：调一家店，两级都进步

A 类门店 4 家，月销售额 500、400、300、120 万元；B 类 3 家，90、80、70 万元。把 120 万那家降为 B 类：

	调整前平均	调整后平均
A 类	330.0	400.0
B 类	80.0	90.0
7 家整体	222.9	222.9

两级平均都涨，整体分毫未动。条件只有一个：被挪动的对象低于原组平均、高于新组平均。这个现象由 Feinstein et al.（1985）命名，原始场景是肺癌的分期迁移（stage migration）：1977 年那批患者做了大量新式影像检查，查出了原本看不到的微小转移，许多人因此被改判为更晚的分期。他们比原晚期组的病情轻、比原早期组的病情重，于是早期组和晚期组的统计生存率双双提高，而任何一位患者的实际预后都没有改变。BP 遇到“组织调整后各单元指标普遍改善”时，先看总盘子。它和误判一是同一类问题：合并方式没变，变的是分组边界。

十四、常见误区与教学预案

学员可能的误解	纠正口径
分组数才是真相，合并数是骗人的	两者回答不同的问题。肾结石案例里分组数更可信，是因为病情轻重同时影响了术式选择和结果；换一个因果结构，结论可以相反（Pearl, 2014）
回归均值说明管控没用	净效果 8 个点是真实存在的。回归均值说明的是 22 个点高估了效果
学完以后什么数都不信	三个追问是一套程序。三句都问过、没问出毛病的数就放心用，多问一句是为了让结论站得更稳
“均值回归”和“回归均值”是一回事	统计现象写作回归均值（regression toward the mean）；均值回归（mean reversion）在金融里指价格向长期均值收敛的过程，二者不同
AI 会算错，所以要重算	AI 很少算错。要审的是它在算之前替你做的决定

课堂预案：现场有人问“能不能让 AI 自己避开这三个误判”，回答“能，这就是第十节那份指令，也是下一模块搭智能体时复核节点要做的事；但指令只能让 AI 披露口径，判断口径适不适合这个决策，仍然是 BP 的工作”。

十五、术语表

中文	英文	本课位置
操作性定义	operational definition	选题依据
辛普森悖论（尤尔-辛普森效应）	Simpson's paradox (Yule-Simpson effect)	误判一
构成效应 / 率效应	composition (mix) effect / rate effect	误判一
Kitagawa 分解；因素分析法（连环替代法）	Kitagawa decomposition; factor analysis by chain substitution	误判一
价格—数量—结构分析	price-volume-mix (PVM) analysis	误判一
混杂因素	confounder	误判一
幸存者偏差	survivorship bias	误判二
样本选择偏差	sample selection bias	误判二（Heckman, 1979）
右删失；Kaplan-Meier 法（乘积极限法）	right censoring; Kaplan-Meier estimator	误判二
队列分析	cohort analysis	误判二
意向性治疗原则	intention-to-treat (ITT) principle	误判二
回归均值	regression toward the mean	误判三
反事实；潜在结果	counterfactual; potential outcomes	误判三
双重差分法	difference-in-differences (DID)	误判三
内部效度威胁	threats to internal validity	核心框架
自动化偏差	automation bias	选题依据

中文	英文	本课位置
平均值缺陷；琴生不等式	the flaw of averages; Jensen's inequality	延伸 A
威尔·罗杰斯现象；分期迁移	Will Rogers phenomenon; stage migration	延伸 B
生态学谬误（区群谬误）	ecological fallacy	讲义概念卡（Robinson, 1950）

十六、参考文献

APA 第 7 版。全部条目已于 2026-10-12 通过 Crossref、PubMed、出版商页面或官方文档联网核对；核对中发现的问题见本节末“核对说明”。

统计、因果推断与方法论

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366. <https://doi.org/10.1080/01621459.1972.10482387>

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4), 553–580. <https://doi.org/10.1093/rfs/5.4.553>

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4), 772–793.

Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clinical Research Ed.)*, 292(6524), 879–882. <https://doi.org/10.1136/bmj.292.6524.879>

Elton, E. J., Gruber, M. J., & Blake, C. R. (1996). Survivorship bias and mutual fund performance. *The Review of Financial Studies*, 9(4), 1097–1120. <https://doi.org/10.1093/rfs/9.4.1097>

Feinstein, A. R., Sosin, D. M., & Wells, C. K. (1985). The Will Rogers phenomenon: Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *The New England Journal of Medicine*, 312(25), 1604–1608. <https://doi.org/10.1056/NEJM198506203122504>

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. <https://doi.org/10.2307/2841583>

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. <https://doi.org/10.2307/1912352>

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, 319(7211), 670–674. <https://doi.org/10.1136/bmj.319.7211.670>

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30, 175–193. <https://doi.org/10.1007/BF02418571>

Julious, S. A., & Mullee, M. A. (1994). Confounding and Simpson's paradox. *BMJ*, 309(6967), 1480–1481. <https://doi.org/10.1136/bmj.309.6967.1480>

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272), 1168–1194. <https://doi.org/10.1080/01621459.1955.10501299>

Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8–13. <https://doi.org/10.1080/00031305.2014.876829>

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. <https://doi.org/10.2307/2087176>

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. <https://doi.org/10.1037/h0037350>

Savage, S. L. (2009). *The flaw of averages: Why we underestimate risk in the face of uncertainty*. John Wiley & Sons.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2), 238–241. <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2), 121–134. <https://doi.org/10.1093/biomet/2.2.121>

财务业务伙伴与管理会计

CGMA. (2015). *Building a better business, together: Welcome to finance business partnering*. AICPA & CIMA. <https://assets.ctfassets.net/rb9cdnjh59cm/2NoLmrFOUB2Nv63FLVpLC6/7c0df6fe35a6a544186ce564828cffa7/cgma-finance-business-partnering.pdf>

财政部. (2016). 管理会计基本指引 (财会〔2016〕10号). 中华人民共和国国务院公报, 2016(33). https://www.gov.cn/gongbao/content/2016/content_5139840.htm

人因工程与 AI 使用

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>

教学设计与学习科学

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Complete ed.)*. Longman.

Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, 15(4), es6. <https://doi.org/10.1187/cbe.16-03-0125>

Cognition and Technology Group at Vanderbilt. (1990). Anchored instruction and its relationship to situated cognition. *Educational Researcher*, 19(6), 2–10. <https://doi.org/10.3102/0013189X019006002>

Gagné, R. M., Wager, W. W., Golas, K. C., & Keller, J. M. (2005). *Principles of instructional design (5th ed.)*. Thomson/Wadsworth.

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale* (pp. 41–50). Association for Computing Machinery. <https://doi.org/10.1145/2556325.2566239>

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379–424. <https://doi.org/10.1080/07370000802212669>

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, 51(2), 289–299. <https://doi.org/10.1080/00461520.2016.1155457>

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. ATD Press.

Knowles, M. S., Holton, E. F., III, Robinson, P. A., & Caraccioli, C. (2025). *The adult learner: The definitive classic in adult education and human resource development (10th ed.)*. Routledge. <https://doi.org/10.4324/9781003387671>

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 989–998. <https://doi.org/10.1037/a0015729>

Mayer, R. E. (2020). *Multimedia learning (3rd ed.)*. Cambridge University Press. <https://doi.org/10.1017/9781316941355>

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15(3), 243–257. <https://doi.org/10.1037/a0016496>

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59–89. https://doi.org/10.1207/s1532690xc0201_3

White, R. T., & Gunstone, R. F. (1992). *Probing understanding*. Falmer Press.

技术文档

Microsoft. (n.d.). AVERAGE function. Microsoft Support. Retrieved October 12, 2026, from <https://support.microsoft.com/en-us/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6>

pandas development team. (2026). `pandas.DataFrame.mean` (Version 3.0) [Software documentation]. Retrieved October 12, 2026, from <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mean.html>

学员延伸阅读（中译本）

卡尼曼, D. (2012). 思考，快与慢（胡晓姣, 李爱民, 何梦莹, 译）. 中信出版社.（第 17 章讲回归均值）

珀尔, J., & 麦肯齐, D. (2019). 为什么：关于因果关系的新科学（江生, 于华, 译）. 中信出版集团.

哈夫, D. (2009). 统计数字会撒谎（廖颖林, 译）. 中国城市出版社.

核对说明

1. Charig et al. (1986) 合计成功率（78% 对 83%）由 PubMed 摘要直接核实；分组数据（93%/87%、73%/69%）经 Julious & Mullee (1994) 的二次引用核实，原文 PDF 未能取得。
2. Elton et al. (1996) 印刷版标题为 “Survivorship bias...”，出版商元数据写作 “Survivor bias...”，本稿按印刷版。约 0.9 个百分点/年为该文三指数模型的基准估计。
3. Mayer 《Multimedia Learning》第 3 版，出版商标为 2020 年，作者本人有时引作 2021 年。
4. Guo et al. (2014) 的原文结论是“观看时长中位数最多约 6 分钟，与视频总长无关”，本稿按此表述，不写作“6 分钟后参与度显著下降”。
5. Savage (2009) 的中译本信息在两个来源间不一致（《被平均的风险》，中信出版社 2019；《平均值缺陷》，东方出版社 2011），制作讲义前需对照实体书核实，暂不列入中译本清单。
6. 因素分析法（连环替代法）为中级与注会《财务管理》《财务成本管理》教材通行内容，本稿未引具体版次；如需正式引用，需对照纸质教材确认版次与页码。