第3章 数据获取

明确问题和目标后，特别是数据科学层面的问题和目标确定后，我们需要明确必要的前提假设(pre-assumption)，基于前提假设来设计数据的构成、明确总体和抽样方案，再搜集数据。

3.1 前提假设与数据方案设计

在这个过程中，我们首先需要根据任务提出前提假设，即我们要研究的问题或要进行的任务可能与哪些因素相关，然后根据前提假设设计数据方案，对所设计的数据方案进行可行性分析，如果不通过，就需要重新审视、调整方案；最后，根据可行的方案确定数据构成。数据方案的设计流程如图 3-1-1 所示。

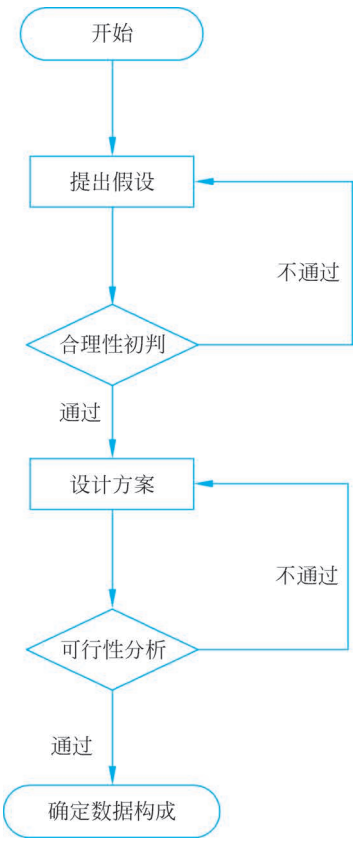

图 3-1-1 数据方案的设计流程

我们以第 2 章中思考题 2-1 为例进行示范。

思考题 2-1 第二次世界大战期间，为了提高战斗机在战场上的生存率，同盟国决定为战斗机装上更厚的装甲，以防被敌方击落。但是，为了不过多增加战斗机重量(重量太大会影响灵活性并增加油耗)，最好只给部分部位增加装甲。军方的需求是要确定对战斗机的哪个部位增加装甲。

3.1.1 前提假设

思考题 2-1 中,我们已经确定这是一个排序问题,即对飞机各部位中弹后的危险性进行排序,对最危险的部分施以保护。那么对于这个排序问题,我们可以采取哪些假设并以此为前提进行数据搜集呢?在以往的课堂上,学生们提出了各种假设,我们来看看其中三种有代表性的假设。

例 3-1-1 针对思考题 2-1 提出的前提假设。

A. 飞机各部位中弹的概率不一样,中弹概率越高的部位,带给飞机的危险越大,因此越需要保护。

B. 飞机各部位中弹概率不一样,但飞机中弹后的危险程度,除了与中弹概率有关,还与中弹对于该部位的破坏性,以及该部位对于飞机的重要性有关,应该综合上述三个因素对各部位中弹给飞机带来的危险打分,对最危险的部位进行保护。

C. 飞机各部位中弹概率是随机且均匀的,飞机中弹后的危险程度,与中弹对于该部位的破坏性和该部位对于飞机的重要性有关,应该就这两个因素对各部位打分。

显然,三种前提假设是不一样的。同时,我们也要认识到,既然是“假设”,其是否符合真实情况还需要后续实验数据验证。

3.1.2 数据方案设计

那么,上述三种前提假设各自需要什么样的数据方案或实验来支持呢?换句话说,我们需要搜集哪些数据?如何获得这些数据?搜集的范围又是怎样的?

例 3-1-2 针对例 3-1-1 的数据方案。

A. 既然我们怀疑中弹概率的高低与飞机中弹后的危险性高度相关,就应该对所有参与实战的飞机,搜集每架飞机各部位中弹概率和对应的危险程度数据。这里需要明确两个量(中弹概率和飞机危险程度)的衡量。如何衡量中弹概率?可以用单位面积的弹孔数,如每平方分米面积的弹孔数来衡量。飞机危险程度又如何衡量?可以尝试对飞机最终的残损程度进行打分评估。这样,在 A 假设下,我们确定了要搜集的数据——单位面积的弹孔数,以及飞机的残损程度,我们搜集数据的范围则是所有参与实战的飞机。这就是在 A 假设下的数据实验方案。

B. 在 A 假设的基础之上还增加了两个因素。因此,这两个因素也需要量化参数的衡量。中弹对于各部位的破坏性要怎么衡量?有学生提出,可以用各部位能承受的中弹次数的倒数来衡量。那么各部位对于飞机的重要性怎么衡量呢?有学生提出,各部位中弹后飞机能继续受控飞行的时间和里程可以用来衡量该部位的重要性。

C. 与 B 假设很接近,区别只在于认为各部位的中弹概率一样,所以 C 假设的数据方案可以参照 B 假设。

通过例 3-1-2 可见，不同的前提假设会导致不同的数据方案或研究内容。因此，提出假设后，应认真审视，特别是在有可能出现不同假设前提时，要确定与事实最符合的假设。

3.1.3 数据获取的可行性分析

到目前为止，我们已针对前提假设提出了数据方案，接下来要考虑非常重要的一步，即怎么获取我们需要的数据？这个时候，我们需要具体的获取方法，并认真评估方法的可行性。

例 3-1-3 针对例 3-1-2 的可行性分析。

A. 该方案的参数能获取吗？单位面积的弹孔数(或称弹孔密度)、飞机残损程度，就参数本身而言，派人员对参加实战并返航的飞机去检查、去评估是可以的，但是，之前确定数据搜集范围是所有参战飞机，对未返航的飞机获取数据存在困难。

B. 该方案中，各部位能承受的中弹次数怎么获取？有学生提出通过模拟战争场景的仿真实验获取。有条件当然是可以的。但是如果没有实验条件呢？对于已经参战的飞机，我们能获取这样的数据吗？同样，各部位中弹后飞机能继续受控飞行的时间和里程，如果没有足够的技术条件，我们能事后进行数据获取吗？不要忘记场景设置是在第二次世界大战中。

C. 方案 C 的参数与方案 B 类似。

所以，对应于假设 B 和 C 的方案，我们虽然设计了数据方案，但缺乏立等可行的数据获取办法，只能放弃。方案 A 能对部分参战飞机获取到数据，但依然存在一些问题，我们后续再审视。

可见，即便设计好了数据方案，在真正进入数据搜集环节之前，还须认真思考各数据方案的可行性，如果方案中的数据是项目期限内无法获得的，则必须及时作出调整。事实上，我们在面向实际应用时，常常会遇到设想很美好实际却不可行的情况，因此，对任何方案设计都不要忘记进行可行性分析。

3.1.4 确定数据构成

通过了可行性分析之后，就可以确定数据的构成了。通常而言，我们后续方便处理的数据都是“结构化”的数据。结构化的数据可以理解为一张不能再细分的二维表，表中一行代表一个存在且唯一的个体，一列代表一个属性。例如，例 3-1-2 的方案 A 中，一架飞机就是一个个体，飞机各部位的弹孔密度、飞机总体残损程度则是相应的属性。对一架具体的飞机，其对应的总体残损程度和各部位弹孔密度分别填入一行中的各列，不同的飞机填入不同的行，如表 3-1-1 所示，这样就构造出飞机数据的二维表格，也就是结构化的数据了。

表 3-1-1　例 3-1-3 中方案 A 的结构化数据

样　本	残损程度打分	机翼弹孔密度	机身弹孔密度	引擎弹孔密度	…
战斗机 1					
战斗机 2					
…					

我们再看一下降低银行不良贷款率任务的例子。第 2 章中,我们已明确这是一个对申请贷款的用户做出“普通客户”/“高风险客户”二分类的任务。接下来,你会提出怎样的前提假设?如果你认为放贷风险与申请客户的历史信用、贷款目的、贷款金额、还款能力等有关,那你能确定需要哪些数据吗?对于这些数据,你有切实可行的获取方法吗?确定好可行的数据方案后,我们构造一张申请贷款的客户资料二维表格(见表 3-1-2),表格中每一行代表一位客户,每一列则反映一个我们想要获取的客户数据,如历史信用、贷款期数(月)、贷款目的(购车、购房、教育或者普通消费)、贷款金额、储蓄或理财账户的余额、可支配月收入与月供比、房产、担保、抚养赡养人数、职业、婚姻状况、已有贷款金额和月供,等等,这样,贷款客户甄别任务的数据构成就确定了。

表 3-1-2　贷款客户甄别任务中的结构化数据

客　户	历史信用	贷款期数	贷款目的	贷款金额	可支配月收入与月供比	…
客户 1						
客户 2						
…						

例 3-1-4　某机构想研究大学的教育水平是否影响了其毕业生 20 年后的成就。请问数据方案如何设计?

不难理解,该问题是一个因果判断问题,假设的“因”是大学的教育水平,“果”是大学所培养学生的 20 年后成就。

第一步先确定前提假设:大学的教育水平可能与哪些因素相关?个人成就又可以由哪些特征来体现或衡量?例如,我们可能认为大学的教育水平既体现在其在某个公信度高的榜单上的排名,又体现在学校每年在学生身上投入的教育资金数量上;而个人成就可以由年收入、同行中的声望或在所从事行业的业内知名度或影响力(如学者的 H-index)等来体现。目前来看,这个前提假设有一定的合理性。

在第一步前提的基础上,我们确定对现今 40 岁左右的人进行调查,搜集被调查对象约 20 年前所就读大学时的“大学排名”和“平均每生每年教育经费”,以及被调查对象目前的“年收入”和“行业内影响力所属等级”等信息。我们注意到不同行业的业内影响力因子定义可能有差别,所以将影响力因子转换为等级可能更适合跨行业的比较。

第三步，可行性分析，这里主要关注第二步中的信息是否能获取。如果确定这些信息可以通过街头问卷或线上搜索调查获得一部分，同时还可以向一些专业数据公司购买获取，那么基本是可行的。

这样，我们就可以设计出一张由数据构成的二维表格，如表 3-1-3 所示。

表 3-1-3　大学水平对毕业生成就影响任务中的结构化数据

	大学校名	大学排名	投入资金/(年·生)	目前年收入	行业内影响力所属等级	…
受访者 1						
受访者 2						
…						

请读者深入思考一下，这样的数据构成是否真的能导向因果推断呢？如果不能，需要如何改进呢？

总结一下，在明确任务之后，我们需要提出前提假设。不同的假设很可能会涉及不同的数据，最终导致不同的研究内容，因此，在真正进入数据搜集环节之前，应认真审视，确定与事实最符合的假设。然后，根据前提假设，给出对应的数据方案，并认真思考获取方案中数据的可行性，如果方案中的数据是在项目期限内无法获得的，就必须及时作出调整。最后，根据确定下来的数据方案，就可以确定数据构成了。

3.2　总体和抽样

在确定数据构成后，就面临具体的数据搜集。一般而言，由于我们面临的预算、时间、人力等客观条件的限制，所能搜集数据的范围常常是有限的。这时，我们需要明确地知道我们想研究的总体是什么，在无法获得总体时应选择什么样的抽样方案。所以，这里先介绍几个统计学的基本概念。

3.2.1　总体和个体

总体(population)是指待研究对象的全体，总体中的每个对象即是个体(individual)。在例 3-1-1 中，所有参战并中弹的战斗机就是潜在的研究总体。其中每架单独的战斗机就是个体。

然而，通常总体是难以获得的。例如，在例 3-1-1 的战斗机问题中，军方无法获得所有的参战并中弹的飞机的数据，因为会有一些被严重破坏的飞机根本无法返航。即便是当前的大数据时代，也不能轻易认为我们已经拥有总体的数据。事实上，数据科学项目的多数任务某种程度上都可以理解为从已知去推断未知，既然存在着未知，那么我们已知的就不能称为总体。因此，数据搜集是一种从总体中抽样的过程。

3.2.2　样本

在无法获得总体的情况下，从总体中抽取出来的子集称为样本(sample)。样本包含

的个体数量一般称为样本容量。

既然获得的数据只是抽样而不是全体,我们就一定要保持警惕,提醒自己回答这个问题:“样本的特点确实能真实反映总体的特点吗?”

这个问题要回答“是”,至少要满足两个条件:

(1) 样本容量不能过小,传统统计学认为小于 30 的样本容量不具备统计学意义,也就不能有效反映总体特点,大数据时代,这个条件容易满足;

(2) 抽样时不能有预设偏见,也就是必须无偏抽样。

3.2.3 无偏抽样

无偏抽样(an unbiased(representative)sample),又称为代表性抽样,是指抽样的过程不受个体性质(特别是与所研究问题有关的性质)的影响。

怎么理解“抽样过程不受个体性质的影响”?我们依然以例 3-1-1 的战斗机案例为例。其实,这是发生在第二次世界大战中的一个真实事件。当时军方在短期内获得的数据有限,只对返航飞机的弹孔密度进行了搜集,汇总后得到了如表 3-2-1 所示的统计表格。

表 3-2-1 例 3-1-1 中军方获得的数据

飞 机 部 位	弹孔密度/(个/平方英尺)
引擎	1.11
机身	1.73
油料系统	1.55
其他	1.80

请读者想一想:表中的弹孔密度能真实反映总体,也就是所有参战飞机的中弹情况吗?

并不能!因为表中的数据仅仅是从返航飞机获得的,对于损坏严重而未能返航的飞机则完全没有考虑。或者说,抽样过程本身受到飞机残损程度的影响,不符合无偏抽样的原则,因此,这里获得的数据并不是总体的真实反映,而是存在抽样偏差。更严重的是,残损程度直接与待研究的问题相关,从而引入的抽样偏差可能会完全导向错误的结论。

3.2.4 抽样偏差

抽样偏差(sampling bias)是指从总体中非随机性抽样带来的系统性错误。抽样偏差使得个体被抽样的概率不一样,有些个体可能根本没有被抽样的机会。

上述例子中,只对幸存下来成功返航的飞机进行抽样,这种抽样偏差有一个被广泛接受的名称——幸存者偏差。事实上,幸存者偏差是一种常见的抽样偏差,我们或多或少都遇到过。例如,调查成功人士的品质或成功公司的特点,归纳总结为成功的要素;采

访百岁老人，将他们的生活习惯总结下来作为长寿秘诀；询问得某种疾病之后又康复的人，认为他们吃的药品或食物对该疾病有疗效，等等。

从上述这几个例子中，读者能看出抽样偏差吗？抽样过程中它们分别漏掉了哪些群体呢？

例 3-2-1 找出下述抽样过程中的抽样偏差。

1. 调查成功人士的品质或成功公司的特点，归纳总结为成功的要素，其中漏掉了(　　)。

2. 采访百岁老人，将他们的生活习惯总结下来作为长寿秘诀，其中漏掉了(　　)。

3. 询问得某种疾病之后又康复的人，认为他们吃的药品或食物对该疾病有疗效，其中漏掉了(　　)。

该如何避免抽样偏差呢？要避免抽样偏差，通常的做法是随机抽样。

随机抽样是指总体中的个体是否被抽样并非确定的，即不因为个体的某个或某些性质一定被抽中或一定不被抽中，而是每个个体都以一定的概率被抽样。更一般地，当这个概率不受个体本身性质的影响而在所有个体上均匀分布时，即为简单随机抽样。

在前述例子中，所有参加实战的飞机(返航的和未返航的)应当有同等的被抽样机会；所有有待证明品质的人或公司(成功的和非成功的)要有同等的被调查机会；所有有待证明生活习惯的人(长寿者和非长寿者)应有同等的被采访机会；所有患某种疾病后吃某种食品或药品的人(康复的和未康复的)要有同等的发声机会。这样获得的数据，才能真正体现总体、代表总体。

再来看一个例子。

在我国，地震是一种非规律性发生的自然灾害，各种大大小小的地震在各地偶见报道。而每每地震发生后，民间都会有一些传言，谈论地震发生之前自然界的各种异象，认为其可以用来预测地震，近年来比较有代表性的是“地震云”(见图 3-2-1)。而科学界和权威机构则反复强调“目前没有有效证据表明云可以用于预测地震”。我们要如何理解这个说法呢？或者说，我们要证明某种特定形状的“地震云”能预测地震，需要提供怎样的“有效”或科学证据？根据前面的介绍，我们相当于提出了一个前提假设：“地震云”能预测地震，或更明确地，“地震云”出现后 X 天之内会发生地震。对应于该假设，我们应搜集

图 3-2-1 传闻中的“地震云”

(图片源自百度百科)

所有出现该形状“地震云”后 X 天内是否发生地震的数据。只有当出现“地震云”时发生地震的条件概率显著大于同区域发生地震的先验概率时,才能下结论说“地震云”可预测地震。然而,很可惜,气象和地质等专业机构迄今没有获得这样的科学证据。

从抽样的角度来看这个问题,发生地震后去找地震前的自然界异象,属于回溯性研究,很有可能引入幸存者偏差,因为我们只关注了那些同时有地震、有异象的情况,而那些异象后并无地震的情况都被忽略了。有趣的是,作为智人,归因分析可能是我们区别于绝大多数物种的特点和优势,但归因过程中的回溯性研究却又很可能把我们带入认知的误区甚至陷阱。请仔细想想,你在生活中有没有遇到过这一类的问题呢?

总结一下,我们搜集数据前,必须明确研究的总体,无法获得总体时,抽样不能有抽样偏差,这样得到的数据才能作为总体的可靠代表。

本节的最后,让我们讲完战斗机的故事。

第二次世界大战中,同盟国采用的前提假设是“弹孔密度越高,给飞机带来的危险越大,越需要被保护”,所以,基于对返航飞机弹孔密度的数据,军方做出了对机身进行保护的决策。当时,美国哥伦比亚大学有一个统计研究小组,他们秘密地为同盟国服务。小组中有一位从德国逃到美国的统计学家亚伯拉罕·瓦尔德(见图 3-2-2)敏锐地觉察出了其中的问题,他提出:不应该给弹孔密度高的部位加装装甲,恰恰相反,应该给弹孔密度最低的部位,也就是引擎加装装甲。他的前提假设是,在远距离交战场合,飞机各部位中弹的概率应该是一样的。他认为,军方对返航飞机调查的数据显示出引擎弹孔密度低,恰恰是因为引擎中弹是致命的,引擎中弹的飞机很多根本都无法返航,那些消失的弹孔就在未能成功返航的飞机上。军方仔细评估了自己的假设和瓦尔德的假设,最终认为瓦尔德的假设更合理,并迅速将瓦尔德的建议付诸实施。而瓦尔德也正是基于这个案例首次提出了“幸存者偏差”。

图 3-2-2　亚伯拉罕·瓦尔德

(图片源自维基百科)

3.3 混杂因素及其避免方法

3.3.1 混杂因素和辛普森悖论

要依赖数据获得可靠结果，除了要做到无偏抽样，还要特别注意混杂因素的影响。先看一个例子。

为了比较两个网站的吸引力或受欢迎程度，我们构造了一个“回头率”参数，即一天之内，至少两次访问该网址的人数与至少一次访问该网址的人数之比，即

$$\text{回头率}=\frac{\text{一天之内至少两次访问该网址的人数}}{\text{一天之内至少一次访问该网址的人数}} \tag{3-3-1}$$

我们对两个网址（某明星微博和我们的课程网址）某一天的访问“回头率”进行了统计和对比，得到如表 3-3-1 所示表格。

表 3-3-1 两个网址的回头率参数统计

某明星微博回头率	我们的课程网址回头率
77%	83%

看起来很不错哦！我们的课程比明星微博更具有吸引力了呢！

真的吗？稍等。如果我们再仔细一点，先把数据分组再来看呢？根据用户登记的学历信息，我们把所有人分成了两组：大学及以上学历组，中学及以下学历组（见表 3-3-2）。

表 3-3-2 分组后两个网址的回头率参数统计

学 历 信 息	某明星微博回头率	我们的课程网址回头率
大学及以上学历	95%	92%
中学及以下学历	71%	34%
全部	77%	83%

这时，奇怪的事情发生了——在每一组里，我们的课程都不如明星微博受欢迎。对数据分组或不分组，得到的结论居然是相反的。这究竟是怎么回事呢？存在计算错误吗？我们仔细检查一下吧。

接着，我们把各组的样本容量都列出来（见表 3-3-3），仔细验算，并没有计算错误。不过，读者看出什么问题了吗？

表 3-3-3 列出了访问人数的分组回头率统计

学 历 信 息	某明星微博回头率	我们的课程网址回头率
大学及以上学历	95%(76/80)	92%(231/250)
中学及以下学历	71%(193/270)	34%(17/50)
全部	77%(269/350)	83%(248/300)

原来,在访问某明星微博的350个用户中,主体是中学及以下学历的,很可能是中小学生。而访问我们课程的300个用户中,主体是大学及以上学历的都是成年人。中小学生每天上网的时间可能受到父母、学校的严格控制,无论对哪种网站,每天的二次访问率都要远低于成年人。亦即,表3-3-3中,回头率参数除了受到网站(体现在表格水平方向)的影响,还同时受到学历或年龄(体现在表格竖直方向)的影响。当我们不分组做比较时,其实是在比较两个主体,即某明星微博的用户主体中小学生以及我们课程网站的用户主体成年人的回头率,从表3-3-3来看,这是一个沿斜向(反对角线)的比较。这种沿斜向的比较,混淆了两个特征维度(即不同网站、不同学历)的影响,并不单纯反映网站本身的影响,所以导致了分组与不分组的结论反转。这个现象其实还有一个专门的名称,叫作"辛普森悖论"。

在这个案例中,当我们用"回头率"参数作为依据比较两个网站的吸引力时,要考虑到,除了网站本身的吸引力之外,用户的年龄、上网的便利程度都会对我们的"回头率"参数造成影响。这些因素不是我们的考察对象(网站和回头率的关系),但却可能对结果造成影响,即我们通常所说的"混杂因素"。要排除混杂因素的影响,常见的做法就是对两相比较的样本集,做好潜在混杂因素,甚至所有非考察因素的匹配。表3-3-3中分组的做法就是一种匹配方式,即高学历和高学历比,低学历和低学历比,其结果就是学历(其潜在是年龄和上网便利程度)差别带来的影响(即混杂因素)被排除了。

我们再看看第2章中汽车销售的案例。

某个汽车销售门店对过去三个月的销售业绩不满意,想要找到原因并做出改善。

前面的课程已经明确,这是一个关联问题中的归因任务。对于关联任务特别是归因任务,混杂因素的排除是非常关键的。通常的做法绝对不是把所有的数据混在一起下一个笼统的结论,而是提出假设确定好待考察因素之后,把所有假设以外的可能混杂因素找出来,并分组,再对比假设(待考察)因素的影响。例如,如果怀疑是促销方案改变带来了业绩下滑,那么,为排除其他混杂因素,如主推的汽车品牌、排量等的影响,应该按汽车品牌、排量等进行分组,分组后再对比促销方案改变前与改变后的销售业绩(见图3-3-1),这样得到的结论,才是可靠的。

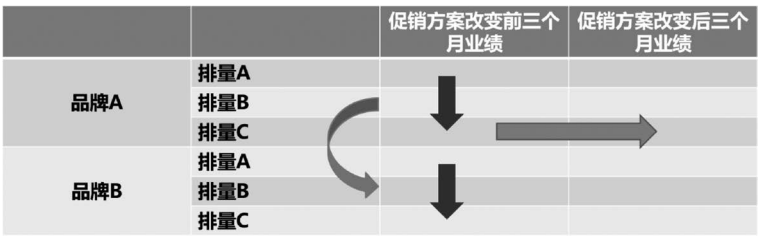

图3-3-1　汽车销售例子中的分组统计示意图

借助图3-3-1,我们其实可以考察三个不同因素(品牌、排量、促销方案)对销售业绩的影响:如果考察促销方案的影响,我们需要固定品牌和排量,在表格内做水平方向(X方向)的比较;如果考察排量的影响,我们需要固定品牌和促销方案,在表格中做竖直方向

(Y 方向)的比较；如果考察品牌的影响，则需要固定促销方案和排量，在表格内做如弧线所示的比较(如果三维显示，可以把品牌赋为 Z 方向)。这一思路我们应该并不陌生，这不就是科学实验中的控制变量法吗？科学实验中，控制变量法就是排除混杂因素的基本手段。举一个经典电学实验的例子，我们现在都知道导体电阻与导体的材料(电阻率 ρ)、导体长度 L 和横截面积 S 三个因素有关，那么，当初电阻公式

$$R=\frac{\rho L}{S} \tag{3-3-2}$$

是如何通过实验确定的呢？简单来说，就是匹配混杂因素(即固定控制变量)后，只观测考察因素的影响。例如，如果要确定电阻 R 与导体长度 L 的关系，则必须先匹配导体的材料与横截面积，即固定导体的材料和横截面积，只考察不同长度对于电阻的影响；如果要确定另外两个因素的影响，做法也是类似的(见图 3-3-2)。科学实验中的“控制变量”是指在实验过程中被固定下来保持不变的变量。例如，要研究 L 对 R 的影响，ρ 和 S 就是控制变量，以此类推。

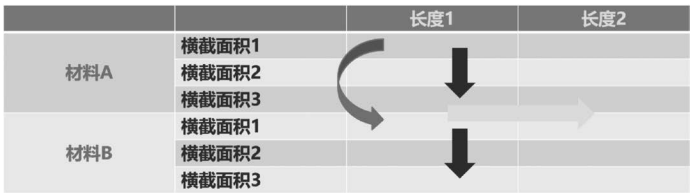

图 3-3-2　确定电阻公式时的控制变量法示意图

可见，找到潜在混杂因素并对其进行匹配的思想早已体现在实验室的科学研究中。该思想也应适用于现实中非实验室条件下的数据搜集和对比分析，只不过现实数据往往来自比实验室复杂得多的环境，更难以控制，混杂因素更难以鉴别或剥离。

总结一下，混杂因素是那些不在考察范围(即前提假设未包含)，但却有可能对结果造成影响的因素。辛普森悖论通常都是由于没有充分排除混杂因素影响所引起的。当我们提出了假设，即怀疑某因素对结果有影响时，必须严格审查其他可能对结果造成影响的混杂因素，并匹配这些混杂因素，也就是依据混杂因素进行分组，分组以后再考察我们的假设。

3.3.2　随机控制实验

即便是在非实验室环境下，也有一些可控或部分可控的实验情境。此时实验操作者会专门设计一些对比实验，在其他所有特征都匹配(或一致)的情况下，只观察一个可调控因素的不同取值对结果的影响。其中潜在混杂因素的匹配常常用对被实验者(也称被试)随机分组的手段来实现，这一类实验称为随机控制实验(randomized control trial，RCT)。随机控制实验在医学领域可体现为双盲实验。

例如在临床医学研究中，为检验某种新药物或新技术是否真的有效，研究人员会将

被试(此时即病患)随机分为年龄、性别、病史等都匹配的两组甚至多组,其中一组不被给予任何治疗,或者,为了排除心理因素的影响,被给予假称有疗效的安慰剂。这一组通常称为对照组。除对照组外,其他组则被给予待检验的药物或技术治疗,如果有多组,还可以分别给予不同的剂量。为进一步消除医生态度、病患心理等潜在混杂因素的影响,实验过程中病患和医生都不知道实际分组情况(只有研究人员了解具体分组情况),这就是"双盲"的意思。根据现代医学规范,一种新药或新治疗技术在被承认有效前,都必须经过双盲实验,只有实验组和对照组有显著差别时,药品或技术才会被官方承认是有效的。而我们在保健品促销广告中常见的"张大妈用我们的药治好了多年的老寒腿"这类说法,只是一种营销宣传,不能作为药品有效的科学证据。现在,你能够理解为什么市面上各种号称有这样那样疗效的保健品只是"食品",而不是"药品"了吧。

当随机控制实验中的可调控因素只有两个选项时,就是通常所说的 A/B Testing。近年来互联网行业凭借其数据搜集优势,也开始广泛使用 A/B Testing。例如,某电子商务网站将其所有注册会员随机地分成两组,给两组发送的促销广告基本相同,只有一项差别:一组发送"本促销将于本周六截止",另一组发送"本促销将于不久后截止",然后追踪两组会员在这次促销活动中的购买情况,作为今后选择上述两种中哪一种策略的依据。有报道显示,谷歌公司在 2011 年一年内就进行了超过 7000 次 A/B Testing 来帮助其制定各种决策。

3.3.3 自然实验

3.3.2 节介绍的随机控制实验依然是人为设计的实验,尽管不像化学、物理学实验那样在严格可控的实验室环境,但实验操作者仍然对可调控因素在各组的取值拥有控制权,例如,A/B Testing 中由实验操作者决定哪组给予 A 选项,哪组给予 B 选项。或者换句话说,实验被试在 A、B 两个选项的分配中,没有因为自己的主观意愿而引入混杂因素。然而,现实需求中我们还常常面临另一类(可能也是更普遍的)场景,即 A、B 选项是由被试主动选择的。我们来看一个例子。

例 3-3-1 某机构想了解规律性服用维生素补充剂对于健康的影响。该机构在设计好数据方案后,通过问卷调查,搜集了 100 人的相关信息,如表 3-3-4 所示。

表 3-3-4 "规律性服用维生素补充剂对健康的影响任务"中的结构化数据

	年龄	性别	每周至少 2 天服用维生素补充剂(是/否)	有无慢性疾病(有/无)	最近 1 次体检状况(异常/预警/正常)
受访者 1					
受访者 2					
…					

通过对表 3-3-4 进行统计,该机构挑出年龄在 20～40 周岁,没有慢性疾病的 60 人的数据进行了统计,结果如表 3-3-5 所示。

表 3-3-5　20～40 周岁且无慢性疾病的受访者统计结果(单位：例)

	最近 1 次体检异常	最近 1 次体检预警	最近 1 次体检正常
每周至少 2 天服用维生素补充剂	1	3	23
每周最多 1 天服用维生素补充剂	6	15	12

根据表 3-3-5，该研究机构认为：在中青年无慢性疾病的人群中，规律性服用维生素(每周至少 2 天服用)的人明显比无规律性服用维生素习惯的人健康，因此规律性(每周至少 2 天)服用维生素有利于促进健康。请问，该结论可靠吗？或者说，目前的数据足以支撑这个结论吗？

这个例子中关键问题之一在于：是否规律性服用维生素补充剂是受访者的主动选择，该主动选择本身就蕴含着混杂因素，例如有服用维生素补充剂习惯的人，很可能本身就对健康重视程度更高，平常的生活习惯上也更讲究(作息规律、饮食健康等)。如果这些混杂因素没有被排除，我们无法下结论说“规律性服用维生素更利于保持健康”。

例 3-3-1 给了我们一个被试主动选择蕴含混杂因素的例子。还有些情况，即便不是被试主动选择，被试的被分组过程也可能蕴藏混杂因素。

例 3-3-2　某机构想研究大学的教育水平是否影响了其毕业生 20 年后的成就。该研究机构通过如例 3-1-3 所示的数据搜集方案获取了数据，经统计分析，发现世界大学排名前 50 的高校，其毕业生 20 年后的成就普遍高于排名在 50 名之后学校的毕业生，因此得出结论大学的教育水平确实能影响其毕业生的成就。请思考：该机构的推论过程是否足够可靠呢？

本例中读者可能已经发现了，被试的分组(曾就读世界排名前 50 的高校 vs. 曾就读世界排名 50 之后的高校)并不满足我们强调的随机分组。尽管就读哪所大学不完全是学生的主动选择，但这个分组本身仍蕴藏了不可忽视的混杂因素：在更好学校就读的学生可能智力、性格坚毅程度等个人能力或素质本来也更强，可能是这些因素导致了他们能被更好的高校录取，以及入职后个人成就更高。至少在当前给定的分析中，没有足够证据排除混杂因素而证明大学教育水平和其毕业生成就之间的因果关系。

例 3-3-1 和例 3-3-2 分别给出了现实情境中分组面临的两种常见场景，无论哪种场景，研究人员都无法控制分组，从而难以满足随机分组要求。这样的情况是否就完全不能做对比分析了呢？有聪明的读者提出，可以换个思路，做固定对象在不同选择下的对比分析。例如在维生素的例子中，专门挑选之前没有规律性服用维生素补充剂习惯的人，让他们规律性地服用维生素，然后对比前后的健康状况。这个思路很好。其实严格来说，在例如临床医学中，评价一个药物的疗效，是应该考虑到个体间差异，针对同一个人的治疗前后进行对比的。但即便是这样，也难以保证同一个人在接受实验期间没有其他改变。况且，在例 3-3-2 的问题中，我们根本无法获得同一个学生读顶尖排名高校和读普通高校后的两种结果。也就是说，我们想做对比的两个选项，对同一对象而言，其中始终只有一个是真实发生的，另一个，即被试没有选择或没有被分配的那个选项，是一种反事实的(counterfactual)性质，并不真实存在。更宽泛地说，由于选项间的互斥，个体不可能遍历选择中的不同选项，而只能选择其中一个选项，那么未选的选项即为反事实。很

多社会学或经济学领域的研究涉及反事实的对照,20 世纪 80 年代后期起,从这两个领域中发展出了一种新的“实验”体系——自然实验。自然实验一般来说并不是传统意义上的“实验”,而是一种观察性研究,主要特征在于其中因一些自然的(非目的性的)因素导致了类似随机实验中那样的分组。既然分组是由自然(非目的性)因素决定的,那么就可以认为其中的混杂因素基本被排除了。相应地,自然实验中的研究人员更多的是扮演观察者或调查者的角色,而非实验操作者的角色。如例 3-3-2 的问题中,一种自然实验的思路是缩小范围,只对比那些拿到了顶尖高校录取通知,但由于各种原因最后选择就读不同院校的人。其中暗含的思想是:能拿到顶尖高校录取通知的所有学生在个人能力上的差异并不显著(即个人能力这个混杂因素已匹配一致),当他们就读不同高校毕业后的个人成就差异就主要由大学的教育水平来决定了。

例 3-3-3 某机构针对目前 50 周岁左右的人群,搜集了他们受教育年限及当前年收入数据,想研究受教育年限是否决定了成年后的年收入水平。结果如图 3-3-3 所示。是否可以下结论说受教育年限决定了年收入水平,即两者之间是因果关系?

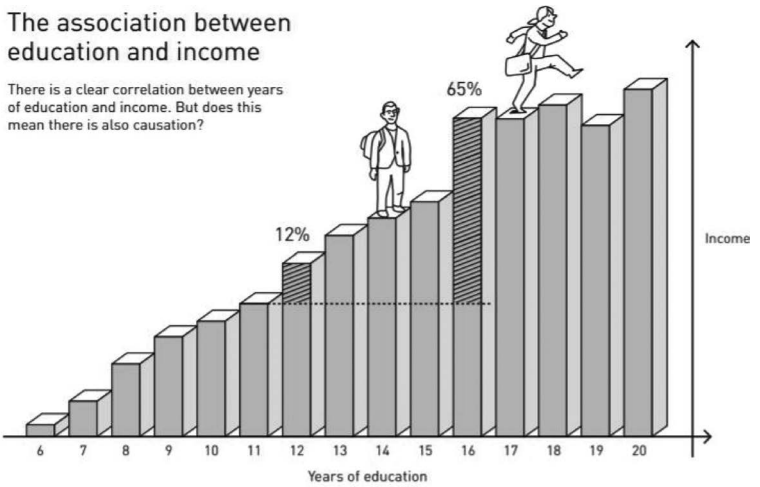

图 3-3-3 受教育年限和成年后年收入关系图

(图片源自网络)

图 3-3-3 显示,在受访人群中受教育年限和成年后的年收入之间确实存在着显著的正相关关系。但即便如此,是否可以下结论说受教育年限决定了成年后的收入水平呢?通过之前的例子我们不难理解受教育年限这一特征中很可能暗含了混杂因素(如学习能力、长期规划能力等),即那些学习能力更强、长期规划能力更强的人更容易坚持长时间接受教育,也更容易获得高收入。为排除这些混杂因素,Angrist 和 Krueger 采用了自然实验的研究方法。他们注意到美国的义务教育对入学年龄的认定只精确到年份,而对允许退学年龄的认定是精确到日期的。即小学新生统一 9 月 1 日入学时,只要当年能达到 6 周岁即可,而年满 16 周岁或 17 周岁退学时,却需要真的过完生日才被允许。在这样的规定下,以一个 2006 年 1 月 1 日出生的人和一个 2006 年 12 月 31 日出生的人为例,他们

都在2012年9月1日入学，入学时前者年龄为6周岁8个月，后者年龄为5周岁8个月。如果他俩都选择一满17周岁就退学，那么他们的受教育时长一个是10年4个月，一个是11年4个月，最大差别能到1年。于是Angrist和Krueger就专门针对生日在第一季度和生日在第四季度的选择主动辍学的学生进行了研究。结果他们发现，第四季度出生的人平均受教育时长确实比第一季度出生的人长，同时平均而言前者的收入也比后者高出约9%。按照Angrist和Krueger的观点，出生在第一季度或第四季度是由随机的自然因素决定的，因此两组人群间不存在能力或素质上的特别差异，那么两组人收入上的差异就只能解释为受教育时长所导致了。当然，Angrist和Krueger的这一结论有其特定的适用范围，即只适用于那些确实一到义务教育年限就主动辍学的人。任何科学的结论都有其适用范围，这也恰恰是科学的特点之一。不区分适用范围的终极真理，并非科学的追求或意义。

巧妙的自然实验法帮助人们理解了很多社会学和经济学的问题。2021年诺贝尔经济学奖就被授予Card、Angrist和Imbens三人，以表彰他们应用自然实验解答了很多社会重要问题，如移民的影响、最低工资水平对劳动力市场影响，以及教育的影响等。

本章中，我们主要针对数据获取，从提出前提假设、设计数据方案开始，接着介绍了数据搜集或实验中的基本科学原则和常用方法。事实上，数据的获取是非常关键的一步，本章介绍的基本流程和原则能帮助我们在实际应用中有效避免常见错误或徒劳无功。

至此，我们即将从现实世界转入计算机世界，了解数据在计算机中的表示与表达。

思考题

3-1 你打算开发一个小程序来估算房屋的合理租金。动手开发代码之前，你打算如何设计数据方案？

3-2 确定了新生宿舍分配(思考题2-2)的数据科学层面问题后，你打算如何设计数据方案？

3-3 你认为网购衣服由于不能试穿很不靠谱。可是你某位衣品高的朋友号称她的衣服都是网购而来的。你觉得难以置信，而她确实没有说谎。你能解释这是为什么吗？

3-4 请根据本章的内容判断一下“张大妈服用我们的产品治好了多年的老寒腿”，为什么不能作为产品有效的科学依据？

3-5 总结长寿老人的生活习惯作为长寿的秘诀既然不可靠，那么请你设计一种获得可靠长寿秘诀的数据搜集方案。

3-6 张三认为未婚人士不如已婚人士有责任感，因此他假设未婚人士贷款后违约的风险更高。他手上有一批银行以往客户的贷款数据，其中包括客户的性别、年龄、婚姻状况、贷款原因、贷款金额、月收入与月还款比及当次是否违约的信息。请问，要验证他的假设，他需要如何设计统计表格？