第 1 章

预 备 知 识

1.1 对比试验

实验者无法控制的诸多变动往往遮蔽了所观测到的效应，而本书正是关于如何在这种情况下设计试验的。大量不可控的变化在科技试验和多种类型的生物科学工作中普遍存在，也正是在这些领域中本书描述的方法使用最为频繁。不过，了解其中一些简单的方法对于绝大多数的实验科学分支而言都是很有价值的。

以下给出一些大量难以预测的变化存在的典型例子。

例 1.1 大多数农业产量试验的目的是比较某种农作物的若干品种，或者若干可选的施肥方式，或者若干管理体系，等等。试验区域被划分成地块，不同品种或任何需要比较的对象被一一分配到每一个地块上。通过测量或估计每个地块的产量（或其他属性），人们可以凭借这些观测值进行不同品种之间的比较。根据经验可知，即使每一个地块上种植的都是同一个品种，地块与地块之间的产量也可能有很大的差异，这种差异的主要特点如下：

（1）邻近地块的产量相比于间隔较远地块的产量更为相仿；

（2）田地上可能存在系统性的变化趋势或者局部的周期性变化；

（3）如果在不同的田地上或不同的年份重复该试验，平均产量可能会大不相同。

田地中单个地块的产量与总体平均值的偏差高达±30％都可能是常见的，而品种之间5％的系统性差异就可能具有相当大的实际意义了。我们将关注安排试验的方法，以便可以自信而准确地将我们感兴趣的品种差异与不感兴趣的不可控差异区分开。

这种试验的目的是比较不同品种，而不是在指定条件下确定某种品种每英亩地可能的绝对产量。这么说有两个原因。首先，品种之间的差异决定了基于试验的任何可能的实际建议，换言之，选择两个品种中的哪一个并不取决于绝对产量，而是取决于一个品种的产量比另一个高多少，以及在其他认为重要的特性上的不同。其次，即使(3)中提到的平均产量出现了实质性的变化，品种之间的差异通常也保持相对恒定。这意味着直接比较品种比在单独的试验中为每个品种估计代表性条件下的平均产量然后再比较估计值要经济得多。

总结本例的讨论，在一个试验中我们需要考虑以下几方面的内容：

（1）目的是比较一些品种（或处理方法）；

（2）在没有品种差异的情况下，不同地块上的产量也有很大差异；

（3）品种之间的差异相对稳定，即使平均响应水平可能会有所波动。

为方便起见，我们引入一些标准术语。这些地块被称为**试验单元**（experimental unit），或更简洁地称为**单元**（unit），将想要比较的品种、肥料等称为**处理**（treatment）。试验单元的

正式定义是：它对应于试验材料的最小划分，使得任何两个单元在实际试验中都可以接受不同的处理。例如，假设为了估计地块上的产量，在每个地块上分别选取两个子区域，其上的农作物被收割且称重。这些子区域不是试验单元，因为一个地块上的两个子区域始终接受相同的处理。

例 1.2 工业技术中的许多试验都具有与例 1.1 相似的形式，目的可以是比较几个可选的处理方法，或者是评估对标准处理方法进行修改后的效果。此类试验包括将原料分为若干批，然后在第一个周期(天、小时等)用一种处理方法处理一个批次，在下一个周期一般使用另一种处理方法来处理另一个批次，以此类推；或者也可能使用多套设备同时处理。对每个批次进行观测(平均强度、产品的产量等)。在不考虑处理方法的不同所带来差异的情况下，观测值也会因批次而异，除了显然存在的随机变化外可能还会出现某些稳定的变化趋势，例如，温度和相对湿度依每小时、每天的变化以及由于加入刚运来的原料所产生的突然间断。

例 1.3 当藤壶(Balanus balanoides)附着的板岩暴露在海水中时，同类的藤壶会迅速附着上来。Knight-Jones(1953)在研究附着的机理时，分别使用了未经处理的板岩和多种化学试剂处理过的板岩。通过探究哪种试剂可使附着量大大减少，他可以对涉及的化学过程做出部分推断。

这个试验与例 1.1 和例 1.2 相比增加了一个特征：只有当比较不同处理方法有助于揭示所研究现象的本质时，这种比较才有意义。该试验是关于比较的，因为它明智地包含了一系列未经处理的板岩以作为对照。这是为了确保经过处理后所观测到的附着速率的任何下降都不是由于自然附着速率的变化而引起的，后者属于不稳定的波动。

这里的试验单元是板岩，观测的是在三天的时间内藤壶附着的数量，处理指的是对照和各种化学试剂。

例 1.4 确定药物效力的一种方法是通过以下方式与公认的标准品进行直接比较：将药物以恒定速率作用于受试动物，并记录下发生动物死亡或其他可识别事件时的剂量。该临界剂量称为耐受量或阈值。分别使用待分析的药物和标准品对一些动物重复此步骤。尽管耐受性因动物而异，但是通过比较药物和标准品的平均对数耐受性(参阅 2.2 节)可以确定效力。在这里，每只动物都是接受两种可能的处理方法(待分析的药物和标准品)之一的试验单元。

一种替代方法是无须使用标准品而直接使用平均对数耐受性来测量药效。这通常是不能令人满意的，因为不同动物组之间的耐受性差异很大，因此在不同实验室和不同时间进行的试验结果只能大致作对比。经验表明，药物和合适的标准品之间的对数耐受性差异通常很少受到动物组之间系统性差异的影响，因此如果将标准品引入试验，则不同时间、地点重复试验所得到的药效量度变化不大。

Finney(1952)充分讨论了这种比较生物测定的简单形式。

例 1.5 使用新药物的临床研究提出了类似的试验设计问题。在研究中除了新的治疗方法，把对照治疗也包含进来几乎永远是明智之举，因为除了极个别的病例外，治愈比例的较小改变往往就能展现新方法的疗效。之所以将确定对照治疗的治愈比例作为试验的一部分而不是仅仅基于过去的经验，是有令人信服的理由的，我们将在后文详细讨论。在此应用中，每个患者都是一个试验单元，他们接受两种或多种可能的治疗方法之一。

在严重疾病的治疗中有一个复杂的问题，即不提供可能会增加生存机会的治疗被认为

是不道德的。因此，一旦存在合理的证据表明某种特定的治疗方法实际上是优越的，就必须立即结束试验（Armitage，1954）。

上述试验与许多物理或化学试验之间的本质区别在于，对于后者，一旦掌握了试验技术并且仪器工作正常，就能获得几乎可重复的结果。更准确地说，与对系统进行更改时预期产生的效果相比，不可控的变化很小。因此，如果系统发生更改并且观测值出现变化，则可以安全地假定所进行的更改是引起观测值变化的原因。在这种情况下，本书描述的方法几乎没有什么价值，除非作为防止设备缺陷引起误差的保护措施。但是，一旦所关心的效应与不可控的变化大小具有可比性，我们将要考虑的问题就变得很重要。

例 1.1～例 1.5 都是相同的形式。我们有一些试验单元和一些可供选择的处理方法。试验包括对每个单元应用一种处理方法并进行一次（或多次）观测，处理方法在单元上的分配受实验人员控制。当这种试验的目的是比较不同处理而不是确定绝对值时，该试验称为**对比**（comparative）。*

不是对比试验的有计划的观测主要与探究已定义事物的性质有关，例如，一批羊毛的平均纤维直径，或特定区域中甲虫的种类数量，或某地区儿童看电视的特点（经常看与不经常看）等。

区分上述最后一个例子中涉及的比较（看电视频次的对比）和在对比试验中要进行的比较尤其重要。关键的区别点在于，对比试验中每个单元的处理是由实验者选择的，而在有计划的调查中，观测者根本无法对使得特定个体归为某一组而不是另一组的原因进行控制。从有计划的调查中可以得出有趣的结论，特别是在相似个体的群组之间进行比较，例如在相同年龄、受教育背景、社会阶层等的儿童群体中进行比较时。不过，从对比试验中得到的结果比从有计划的调查中得到的结果在因果关系上具有强大得多的说服力。基于此，我们后面几乎完全将注意力集中在对比试验上。

讨论对比试验的设计分为两个几乎完全不同的部分，其处理准则应考虑以下问题：

（1）选择要比较的处理方法、要进行的观测以及要使用的试验单元；

（2）将处理方法分配到试验单元的方法，以及决定应使用多少个单元。

本书的大部分内容是关于问题（2）的，但是在第 9 章中试着对问题（1）进行了讨论。

为方便起见，我们首先讨论能够成为好试验的要求。

1.2　好试验的要求

在本节中我们假定已经确定了处理、试验单元和观测属性。那么，一个好试验的要求是处理的对比应尽可能没有系统性误差且足够准确，结论应具有广泛的有效性，试验安排应尽可能简单，并且结论的不确定性应可以评估。

下面依次讨论这些要求。

（i）没有系统性误差

这意味着，如果使用大量试验单元进行指定设计的试验，则几乎肯定可以给出每个处理

---

* 所有测量（包括计数）在某种意义上都是对比，但这不会影响对比试验与其他试验之间的区别，因为在特定试验的框架内，通常可以将测量视为绝对测量。

对比的正确估计。下面这些例子可以说明此点。

例 1.6 考虑一个工业试验，在同一台机器上比较两个略有不同的过程 A 和 B，其中 A 总是在上午使用，B 总是在下午使用。无论处理多少批次，都不可能仅凭试验结果就将过程 A 和 B 之间的差异与单纯上午较之下午机器性能或操作人员的任何系统性变化区分开来，以示两者无关。有时确实存在这种系统性的变化。计算统计显著性时倒是不会遇到困难；它可能告诉我们，A 和 B 之间的表观差异并不像是纯粹随机的，但也无法确定对差异的两种或多种可能的解释中哪一种是正确的。

当然，认为这样的试验无用是不明智的。先前的试验工作，或者关于该过程的一般认知，或者对相关变量（例如温度、相对湿度）的补充测量可能说明，上午和下午的任何差异都不重要。于是，只要清楚地了解到试验的解释基于此额外的假设，就不会造成太大的问题。但是，假设试验获得了令人惊讶的结果，或者与以后的试验结果明显矛盾呢？那么除非存在强有力的证据证明上午和下午没有差异，否则该试验的说服力就可能大打折扣。

因此，一条合理的原则就是规划试验以尽可能地规避这种困境，即，确保接受一种处理的试验单元与接受另一种处理的试验单元不存在系统性的差别。

每当待检验的对比与试验材料的不同批次、不同观测者、不同试验方法之间的差异完全混杂在一起时，就会出现与刚才讨论相类似的难题。当所有接受同一种处理的单元被集中在一组，而不是独立进行响应时，这种问题也很容易发生。

例 1.7 在动物饲养试验中，一个可行的计划是让同一个围栏中的所有动物一起接受同一种处理。这在一定程度上模拟了实际条件，且试验工作组织起来也非常方便。但是如果我们让一大围栏的动物接受试验规定的量，则不可能将定量差异与围栏之间的系统性差异区分开，比如某一围栏中存在某些完全与试验处理无关的疾病时。

例如，Yates(1934)描述了在猪身上进行的试验，其中猪被分成单独饲养的小群，使得每种喂养方式都能在几组完全独立的猪身上进行测试。试验中发现没吃绿色食物的猪生病了。Yates 指出，如果当初没吃绿色食物的猪被养在了同一个围栏中，那得出的结论很可能就是疾病是由外来原因引起的，特别是先前试验表明了绿色食物并不是必需的。但是，相互独立的几组没有吃绿色食物的猪生病了，而其他的猪并没有生病，这一事实就有力地证明了是该喂养方式致病。

采用另一种方式来说明困难：在单一围栏试验中，根据 1.1 节的定义，试验单元是一围栏的动物，而不是单独的每一个动物。因此，这是一个没有重复的试验，在得出有效结论之前需要作进一步的假设。

在这种试验中做出使用哪种设计方法的决定并不容易，引用该例主要是为了说明所涉及的逻辑要点。Lucas(1948)和 Homeyer(1954)对动物饲养试验进行了进一步的讨论。

一类常见的试验（以例 1.3 为例）包括：实施处理方法，注意观测值与不进行这种处理时预期的观测值之间的差别，从而得出是处理方法引起了该差别的结论。为了使这种试验本身具有说服力，必须将处理过的单元与一系列对照单元进行比较，其中对照单元不经过任何处理，但要在与接受处理单元相同的条件下进入试验，使两者没有系统性的差别。例如，过去已经获得了某种观测值，而现在处理单元给出了不同的观测值，这本身并不一定是处理效应非零的有力证据，因为试验单元之间可能存在系统性的差异或者外部条件发生了系统性的变化。如果过去的经验表明在未经处理单元上的观测值以稳定的方式变化，则该经验

(特别是在前期工作中)可能是为了省去特殊的对照单元。但是,这种做法等同于允许试验中各单元之间存在可能的系统性差异,例如例 1.6,在大多数情况下最好避免。

下面给出一个经典的由于缺少对照而在很大程度上无效的试验范例。

例 1.8　McDougall(1927)训练了一些大鼠在有灯和没有灯的出口之间进行选择以检验大鼠中可能存在的拉马克效应(Lamarckian effect)。然后,他繁育了这些大鼠,并测量了之后的每一代大鼠学会上述任务的速度。如果拉马克效应显现,则学习速度会随着遗传代数而稳定增加,且实际发现确实如此。试验排除了其他某些解释(例如选择),然而没有对照单元,即没有在相同条件下由未经训练的大鼠所繁殖的后代进行试验。因此,这种效应可能是由于试验条件中系统性的不受控制的变化所致。

Crew(1936)用对照单元重复了该试验,未发现明显的拉马克效应。Agar 等(1954)在一项持续了 20 年的试验中,发现了类似于 McDougall 的初始速度增加,但对照组和受训大鼠的情况相同。他们得出的结论是,这种效应是由于鼠群健康状况的长期变化所致。

我们可以总结如下:接受一种处理的试验单元与接受包括对照在内的任何其他处理的单元应仅显示出随机性差异,并应允许彼此独立做出反应。当此举不可能或不可行时,应明确辨认任何关于不存在系统性差异的假设,并尽量通过补充测量或以往的经验加以核实。

稍后我们将介绍,随机化手段是如何确保消除系统性误差的主要来源的。

(ii) 精度

如果通过随机化(第 5 章)实现了没有系统性误差,则从试验中获得的处理对比的估计值与其真实值*的差异仅仅来自随机误差。应当注意,术语"随机"在其严格的统计学意义下将被广泛使用。大致说来,这意味着它指的是那种不具有可再现模式的变异。如例 1.1 中简短描述的田地产量的变化不是随机的,因为它具有变化趋势以及相邻地块的产量之间具有相关性等。

通常可以使用**标准误差**(standard error)来量度处理的对比估计值中随机误差的可能大小。对此的精确定义和计算方法在关于统计方法的教科书中有所描述,例如文献(Goulden,1952,p.17-20),但就眼前的目标而言,以下的说法就足以使人理解其意了:

在大约 1/3 的情况下,估计值的误差将超过正负一个标准误差;

在大约 1/20 的情况下,估计值的误差将超过正负两个标准误差;

在大约 1/100 的情况下,估计值的误差将超过正负 2.5 个标准误差。

这些叙述需要一定的限定条件,具体取决于误差分布的形式和标准误差的准确性,而标准误差本身也必须进行估计。对这些问题目前我们还不需要关注。

标准误差的取值——由此任何一个指定试验的精度——取决于:

(1) 试验材料的内在差异性和试验工作的准确性;

(2) 试验单元的数量(以及对每个试验单元重复观测的数量);

(3) 试验的设计(如果效率不高,还取决于分析方法)。

在统计设计能够提供帮助的大多数试验中,改进试验材料或提高测量设备的精度只能实现非常有限的精度提高。其原因一部分是通常存在很难消除的固有变异性,一部分是在

---

* 第 2 章将精确地定义真实值。

极度受控条件下的试验将不再能代表实际条件，例如在温室、小规模工厂等。这一点将于第9章再进行讨论。

如果每个试验单元观测到一个观测值，则在其他条件相同的情况下，两种处理之间差异估计值的标准误差与每种处理的单元个数的平方根成反比。实际上，标准误差为

$$\text{标准差} \times \sqrt{\frac{2}{\text{每种处理包含的单元数}}} \tag{1.1}$$

若A、B两种处理得到的观测个数不同，则为

$$\text{标准差} \times \sqrt{\frac{1}{\text{A的单元数}} + \frac{1}{\text{B的单元数}}} \tag{1.2}$$

这里的标准差(standard deviation)是经过相同处理的试验单元上的观测值的随机散度的统计量度(Goulden，1952，p.17)。*

由式(1.1)，可通过增加到4倍数量的试验单元来使标准误差缩小一半，而要想将标准误差缩小为原来的1/10，则需要将单元数量增加到原来的100倍。尽管理论上可以通过增加单元数量来使标准误差任意减小，但这样提高精度的方法代价太大。

对试验单元进行重复观测而获得的收益小于或等于单元数量相应增加所产生的收益。其可以根据类似于式(1.1)和式(1.2)但稍微复杂一点的公式进行评估。

提高精度的第三种方法是通过改进设计，这正是我们最应该关注的。总体思路是，应使用任何有关试验单元的可用信息来减小式(1.1)和式(1.2)中的有效标准差。这种方法有时候有可能达到与大规模增加试验单元数量相当的精度提高。

大体说来，我们对精度的要求是标准误差应该足够小，以使得我们能够得出有说服力的结论，但标准误差又不能太小。如果标准误差很大，则试验本身几乎是无用的，而不必要的过小标准误差则意味着试验材料的浪费。在大多数情况下，目标是对处理差异的估计，此时，利用式(1.1)和式(1.2)能够在设计试验时预测出任何指定数量的单元将会达到的精度，或者指定精度时所需的单元数。为此，我们必须了解一些有关标准差的信息，即单元间的变异性，不过近似信息通常是可以从先前的类似试验中获得的。有时候目标不是估计处理差异，而是做出一个不可更改的决定，例如确定哪一种处理方法是最佳的。在这种情况下，如果一种处理方法比其他处理方法好得多，且单元是依次进行测试的，那么即使估计的准确性低也可以在少量观测后结束试验。这带来了一些特殊的问题。单元个数的选择问题将在第8章详细讨论。

(iii) 有效范围

当我们估计两种处理之间的差异时，得出的结论是基于试验所使用的特定单元以及试验所考察的条件。如果希望将结论应用于新的条件或单元，则除了标准误差所测得的不确定性以外，还涉及其他一些不确定性。此说法的唯一例外是当试验单元通过合适的统计抽样过程选取自定义明确的单元总体时。

试验所考察的条件范围越广，我们对结论进行外推的信心就越大。因此，如果能够在不降低试验准确性的情况下安排考察大范围的条件会是很理想的。这在决定某些实际操作过

---

* 应注意，标准差是指各个单元上观测值的变化；而标准误差是指从整个试验得到估计值的随机变化。

程的试验中尤其重要，而在目标纯粹是要了解某种现象的试验中则没那么重要。

例 1.9　“Student”(1931)提到了爱尔兰农业部做的与引进 Spratt-Archer 大麦有关的一些试验。引入工作几乎在所有地方都取得巨大的成功，但是一个地区的农民却拒绝种植，声称自己的本地大麦更加优良。一段时间后，农业部为了展示 Spratt-Archer 大麦的优越性，在相应地区以单行的形式种植了本地大麦，并与 Spratt-Archer 大麦进行了对比。“Student”报道说，令农业部惊讶的是，农民们完全正确：当地的大麦产量更高。同时，原因很明显：本地大麦生长更快，能够消灭在该地区盛行的杂草；但是 Spratt-Archer 大麦从一开始就没有那么强大，于是成了杂草的受害者。因此，最初在良田上进行的试验，当结论应用于其他地方时肯定会产生误导。

其他类型的试验也会得出类似的观点。一种在特定条件下效果很好的新试验技术可能不适合常规使用。试验过程中，在特殊监管下效果良好的新工业流程也可能无法在常规生产中获得成功。举一些具体的例子，在一批同源原料上测试的纺织工艺的改良实际上可能关键取决于原料的含油量；小麦品种之间的差异可能取决于土壤和天气条件；等等。

这些说法将得到以下结果。首先，即使是在纯粹的技术性试验中，重要的也不仅是关于处理差异为多少的经验知识，还有对差异产生原因的理解。这些认知将表明结论的哪些外推是合理的。其次，如果可以人为改变条件而不增大误差的话，我们应该在设计试验时这么做。例如，在比较两种拉伸羊毛的方法时，有时可能会期望这两种方法之间的差异不受羊毛含油量的影响。通常有利的做法是在试验中同时使用含油少的羊毛和含油多的羊毛，以直接检验拉毛方法的差异和含油量之间是否独立。症结所在当然是如果要包含几个这样的补充因子，试验就可能会变得难以组织。此外存在这样的可能性：如果系统很复杂，由于没有一组试验条件能得以完全考察，因此无法得出任何明确的结论。这导致了第三点：重要的是要明确认识到对任何特定试验结论的限制有哪些。

这些考虑因素在纯科学工作中不太重要，通常最好的办法是设法对某些非常特殊的情况获得透彻的理解，而不是一次试验就得出广泛的结论。

(iv) 简单性

这是一个非常重要的、必须牢记的问题，但是很难做出概括性的评述。它涉及多个注意事项。如果试验是由相对不熟练的人员完成的，则可能难以确保遵循复杂的变更计划。如果要在生产条件下进行工业试验，则重要的是尽可能少地干扰生产，即，对不同的工序进行长时间的操作，而不是频繁地进行更改。在科学工作中，尤其是在研究的初期阶段，保持灵活性可能很重要；试验最初可能会提出一系列有希望的好问题，因此，获得任何有价值的结果之前如果必须完成一个大型试验可不是好事。不过，当然存在某些情况下，相当复杂的安排是有好处的，只是要决定在任何特定应用中安排得多复杂为好则需要判断力和经验。

以上说明适用于简化设计。简单的分析方法也是人们想要的。幸运的是，高效设计和简便分析的要求是高度相关的，对于本书中的几乎所有方法，只要满足稍后描述的某些假设，完整统计分析的简单方案都是可供选用的。如果仅需要估计处理差异而无须估计精度，那么很少见设计除了简单平均以外有其他更多的要求了。

使用计算机来分析试验结果是最近的一项重要进展，特别是对于涉及大量数据或试验工作所花费的时间等于或小于用常规方法分析结果所需时间的领域。如果编写的程序合

适，在计算机上进行统计分析所需的时间一般都非常短。

(v) 不确定性的计算

前述的要求都不是统计学上的，最后介绍的不确定性的计算是统计学上的。我们希望，如果可能的话从数据本身出发，能够计算出估计处理差异时的不确定性。这通常意味着估计出差异的标准误差，由此可以在任何所需的概率水平上计算出真正差异的误差范围，并由此可以量度处理之间差异的统计显著性。

为了能够进行严格的计算，必须有一组试验单元对一种处理独立地做出响应，并且与接受其他处理的其他组单元之间仅存在随机性的差异。对接受相同处理的单元进行观测值的比较(不一定是直接的)可以得出误差的有效量度。使用随机化(在第 5 章中将详细讨论)以消除不同处理单元之间的系统性差异，会自动使差异随机化，并在较弱假设下证明统计分析的合理性。应当仔细注意这种分析与例 1.6 之间的区别。

在试验单元数量很少的试验中，可能无法由观测值本身获得对误差标准差的有效估计。在这种情况下，有必要使用先前试验的结果来估计标准差(参见 8.3 节)。这样做的缺点是我们需要假设随机变异的大小保持不变。

本书的一般原则是不介绍统计分析方法。其原因一方面是有许多关于此类方法的出色描述，另一方面是将其包括在内不仅会大大增加本书的篇幅，而且还会分散人们在设计方面进行考虑的注意力。

概　　要

我们主要分析以下形式的试验：有多种可选的**处理**，在每个**试验单元**上应用其中一种，然后进行**观测**。目的在于能够从假定存在的不可控变异中分离出处理之间的差异；当然，这可能只是向着了解所研究现象迈出的第一步。

一旦定下了处理、试验单元和观测的属性后，则主要的要求如下：

(1) 接受不同处理的试验单元之间不应存在系统性的差异，即至少从切实可行的角度，应尽可能避免假设某些变异来源不存在或可忽略不计；

(2) 估计的随机误差应适当小，这应当以尽可能少的试验单元来实现；

(3) 结论应当具有广泛的合理性；

(4) 试验的设计和分析应当简单；

(5) 在没有人为假设的情况下应当能够对结果进行适当的统计分析。

参 考 文 献*

Agar, W. E., F. H. Drummond, O. W. Tiegs, and M. M. Gunson. (1954). Fourth (final) report on a test of McDougall's Lamarckian experiment on the training of rats. *J. Exp. Biol.*, 31, 307.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Q. J. of Medicine*, 23, 255.

* 这些在文中明确提及。另有一些一般性的参考见本书“一般参考书目”部分。

Crew, F. A. E. (1936). A repetition of McDougall's Lamarckian experiment. *J. Genet.*, 33, 61.

Finney, D. J. (1952). *Statistical method in biological assay*. London: Griffin.

Goulden, C. H. (1952). *Methods of statistical analysis*. 2nd ed. New York: Wiley.

Homeyer, P. G. (1954). Some problems of technique and design in animal feeding experiments. Chapter 31 of *Statistics and Mathematics in Biology*. Ames, Iowa: Iowa State College Press. Edited by O. Kempthorne et al.

Knight-Jones, E. W. (1953). Laboratory experiments on gregariousness during setting in *Balanus balanoides* and other barnacles. *J. Exp. Biol.*, 30, 584.

Lucas, H. L. (1948). Designs in animal research. *Proc. Auburn Conference on Applied Statistics*, 77.

McDougall, W. (1927). An experiment for testing the hypothesis of Lamarck. *Brit. J. Psychol.*, 17, 267.

"Student" (1931). Agricultural field experiments. *Nature*, 127, 404. Reprinted in "*Student's*" *collected papers*. Cambridge, 1942.

Yates, F. (1934). A complex pig-feeding experiment. *J. Agric. Sci.* 24, 511.

第 2 章

一些关键假设

2.1 引言

在许多试验中，对每个试验单元会进行好几种类型的观测。例如，在比较甜菜的品种时，可以测量根的产量、顶部的产量、糖的产量以及（如果有可能的话）植物的数量，或许还可以观测疾病的发生率、抽薹的频率以及对糖的化学分析。在比较羊毛纱线的纺纱方法时，通常会测量纱线的不规则性、纱线强度和纺纱时的断头率，并可能对用纱线织造的织物进行测试。在初步的描述中，可以方便地假设每个试验单元上仅进行一次观测。该观测结果可以通过一些试验读数计算得出。例如，纱线不规则性的量度通常是通过沿着纱线长的厚度变化轨迹中计算出所谓的变异系数来获得的。再者，在实验心理学的学习试验中，待分析的观测通常是对学习速度的一种量度。这是从原始数据中得出的，原始数据包括如每次尝试完成试验任务时成功或失败的记录。

以下假设（或对其的一些简单修改）构成了本书中介绍的大多数设计的基础。假定将特定处理应用于特定试验单元时获得的观测值是

$$（只取决于特定试验单元的量）+（取决于使用处理的量） \tag{2.1}$$

且不受其他单元对处理的特定分配的影响。可以更生动地描述如下：用符号 $T_1, T_2, \cdots, T_t$ 表示可选的处理；假定使用 T_1 在任何一个单元上获得的观测值与使用 T_2 时获得的观测值相差一个常数 $a_1 - a_2$。每种处理对应 $a_1, a_2, \cdots, a_t$ 中的一个常数，试验的目的是估计形如 $a_1 - a_2$ 的差异，我们称这种差异为**真实的处理效应**（true treatment effect）。

这个假设的要点是：

（1）式（2.1）中处理所在项加到单元所在项上，而不是其他（如相乘）；

（2）处理效应是恒定的；

（3）一个单元上的观测不受其他单元如何处理的影响。

这三个要点将在后文分别讨论。

对观测结果进行全面的统计分析时以上假设尤其重要。即使仅通过计算简单的平均值来分析试验，这些假设仍然是必需的，因为与假设的重大背离将影响对结果的整体定性解释。通常可以从数据中对假设进行一定程度的检查，但永远无法完全避免做出这些或那些假设。读者初读时不必过多注意以下部分的细节。