第 10 讲
通过数学建模解决“混合物转化为有机肥最佳过程”问题

□ 苏　淳

本讲的目的是介绍关于 AMCM-93 问题 A 的一篇优秀论文，材料取自 *The UMAP Jourani*，第 14 卷(1993 年)，第 3 期，第 211—222 页，并参阅了天津大学数学系边馥萍的译文(发表于《工科数学》专辑：“数学建模教育与国际数学建模竞赛”，叶其孝主编，1994 年 8 月，合肥)．这篇论文中有不少值得学习和借鉴之处，可以用作培训参赛选手的典型范文．我们将在本讲中着重介绍论文本身，而对有关的数学知识的介绍则放在第 11 讲中进行．

首先介绍一下问题的由来和内容．

本题是由美国东华盛顿大学数学系的 Yves Nievegelt 提供的，他在题中所叙述的情况与数据来自于华盛顿州 Medical Lake 地区监狱的餐厅，他还写了一篇评论文章 *The outstanding optimal composing papers*，评述这篇优秀论文(载同一杂志第 227-228 页)．下面便是这道题目的内容：

AMCM-93　问题 A：加速餐厅剩菜堆肥的生成

一家注重环境的学校餐厅正用微生物把顾客没吃完的食物再循环生成堆肥．每天餐厅把吃剩的食物和泥浆(粘结剂)混合，再把它们和厨房里容易弄碎的色拉菜以及少量扯碎的纸片混合，并把

混合物喂给一种真菌培养物和土壤细菌，它们把泥浆、绿叶菜、纸片消化形成有用的堆肥. 易碎的绿叶菜为真菌培养物提供氧气，而纸片则吸收过量的湿气. 但有时真菌培养物显得不能或不肯消化顾客留下的那么多的剩饭菜. 餐厅并没有因为真菌培养物没有胃口而责怪厨师长. 餐厅收到要大量购买他们生产的堆肥的报价. 所以餐厅正在研究增加堆肥产量的方法. 由于无力营建一套新的堆肥设备，因此餐厅首先寻求能加速真菌培养物活力的方法，例如，通过优化真菌培养物的环境(眼下大约是在 120F 和 100%温度的环境下生成堆肥的)，或通过优化喂给真菌培养物的混合物组成，或同时优化两者(而达到加速真菌培养物的活力).

试决定在喂给真菌培养物的混合物中泥浆、绿叶菜和纸片的比例与真菌培养物把混合物生成堆肥的速度间是否存在任何关系. 若你认为不存在任何关系，试说明理由. 否则，试决定什么样的比例会加速真菌培养物的活力.

除了按竞赛规则说明中规定的格式写的技术报告外，请为餐厅经理提供一页长的用非技术术语表示的实施建议.

作为数据，表 10-1 列出了分别存放在不同的箱子中用磅表示的混合物组成中各种原料的数量，以及把混合物喂给真菌培养物的日期以及完全生成堆肥的日期(以表示生成堆肥所需的时间).

表 10-1

泥浆(磅)	绿叶菜(磅)	纸片(磅)	喂人日期	生成堆肥的日期
86	31	0	19900713	19900810
112	79	0	19900717	19900813
71	21	0	19900724	19900820
203	82	0	19900727	19900822
79	28	0	19900810	19900912
105	52	0	19900813	19900918

续 表

泥浆(磅)	绿叶菜(磅)	纸片(磅)	喂人日期	生成堆肥的日期
121	15	0	19900820	19900924
110	32	0	19900822	19901008
82	44	9	19910430	19910618
57	60	6	19910502	19910620
77	51	7	19910507	19910625
52	38	6	19910510	19910628

下面我们就来具体介绍这篇优秀论文.

显然,要解答问题,首先先要明确问题.这篇论文的作者们(以下简称作者们)将堆肥(也称为有机肥)的生成过程视为一个生化过程:"餐厅将剩饭菜搅拌成半流体,混人不能食用的青菜根及少量碎报纸,形成混合物,再倒入混有土壤细菌的真菌培养液,经过生化作用,使混合物转化为可用的有机肥(堆肥)."

在这样的提法之下,原题所提供的数据表发生了一些微妙的变化:①成份不明确的"泥浆"被明确为由剩饭菜搅拌而成的"半流体";② "绿叶菜"被明确为不能食用的"青菜根";③ "纸片"被明确为"碎报纸".这些变化为后面的 C/N(碳与氮的含量比)的计算提供了依据.

在上述提法之下,作者们将问题明确为:"餐厅需要我们确定在混合物的比例与真菌培养液分解混合物的速度之间是否存在某种关系.如果不存在任何关系,需说明原因,否则要确定混合物中半流体、青菜根和碎报纸的比例,以这种比例构成的混合物可以提高真菌培养液的活动能力."并将原题所提供的 12 组观察数据作为建模的参考.

建模工作的第一步是对模型提出一些合理的假设条件.这些

假设条件首先应当合理，其次应当有利于模型的建立. 而要做到合理，就必须认真观察原始数据，并注意结合当地的环境、条件，还应具备一定的生物化学知识.

通过观察原题所提供的 12 组原始观察数据，可以明显地看出：第 1—4 组数据中的“生成堆肥的日期”分别地吻合于第 5—8 组数据中的“喂入日期”. 因此可以假定：

1. 堆肥设施是由四个独立的集装箱组成. 第 1—4 次试验，是把混合物分别倒入这四个箱内，当堆肥完全形成后，分别从箱内清除堆肥，再进行第 5—8 次，9—12 次试验.

此外，作者们还依据具体情况，提出了如下一系列假定：

2. 箱内生化反应过程中的温度不受人为限制，但受外界环境温度的影响. 餐厅没有在控制环境温度方面投资.

3. 在反应过程开始之前，真菌培养液与混合物是分别贮藏的. 真菌培养液贮存在温度为 120F，湿度为 100%的容器中，以利于真菌生存.

4. 操作过程是先把真菌培养液倒入箱内，然后再加入混合物. 由于餐厅工作人员缺乏堆肥方面的理论与经验，餐厅不能控制真菌培养液与混合物的适当比例.

5. 当箱内的 99%混合物转化为有机肥时，我们认为堆肥过程全部完成. 这时要把箱内所有物质清除. 99%的假设是合理的，因为操作者是凭观察而判断堆肥是否完成，混合物是否被“消化”，所以假设存在 1%的观察误差. 此外要从堆肥中分离真菌需付出较高的代价，所以我们认为当反应完成时，所有物质应从箱内清除.

6. 从 7 月中旬到 8 月中旬的平均温度要比从 8 月中旬到 9 月底的平均气温高. 这个假设来自于进行堆肥实验的地区的气象观察.

7. 适当的营养和环境，可使真菌培养液中真菌的数量迅速增长. 营养是指充足的食物(由混合物提供)，环境是指空间、温度和

湿度.迅速增长是指按一定比例增长.

8. 根据所给的问题,可以假设餐厅雇员不具有优化堆肥的知识.

9. 培养液中的真菌与细菌分布均匀,并且充满其所在容器.

10. 真菌培养液的活动能力与培养液中所含真菌的数量成正比,含微生物较多的培养液消化分解混合物的能力较强.

11. 在一个给定的周期内,环境温度取其平均值,并且是一常量.

这些假定一般来说是比较合理的,有些假定(例如第 11 条)虽然有些"粗糙",但却是在建立初步模型时所必须的,它们可以起到减少变量的作用.

建模工作的第二步是找出影响堆肥形成速度的因素,为此需对混合物的成份以及周围环境作定量或定性的分析.

作者们首先利用统计方法,借助于统计软件包对混合物各成分的比例与所需的时间之间的关系进行了相关分析.他们发现:

"半流体与青菜根的重量比"同"堆肥形成时间"之间的相关系数是 $r_1 = -0.27$.

"混合物总重量"同"堆肥形成时间"之间的相关系数是 $r_2 = -0.38$.

"碎报纸与混合物的重量比"同"堆肥形成时间"之间的相关系数是 $r_3 = 0.81$.

这些计算结果表明,只有 $\frac{\text{碎报纸重量}}{\text{混合物总重量}}$ 与堆肥形成时间的相关性显著,且为正相关($r_3 = 0.81 > 0$),意即该比值越大,则堆肥的形成时间越长.这些结果自然没有反映出问题的本质,而只是问题的一些"表象".

具有良好生物化学知识的作者们没有被表象所困惑,而是果

断地想到：影响堆肥形成时间的可能只是其中的某些化学成份的比例，这正是他们的高明之处. 他们想到：如果混合物中存在过量的碳(C)，则培养液中的真菌将消耗掉所有的氮(N)，并且从尚未反应的混合物中分离出 N 来，影响生化反应；反之，若含有过量的 N，则将使培养液释放出过量的氨气，亦不利于生化反应的进行. 这使他们猜想，影响堆肥形成时间的真正因素是 $R=C/N$(即混合物中碳与氮的含量之比).

为了证实上述的想法，他们调阅了有关文献，并根据 *Sussman, Vic*(1982), Easy composting Emmaus, *PA*: *Rodale Press* 计算出了餐厅混合物中各种成分中的 C/N，结果如下：

半流体中　C/N＝15∶1；

青菜根中　C/N＝17∶1；

碎报纸中　C/N＝170∶1.

这就是说，碎报纸含有极高的 C/N 比率，无怪乎在前面所述的相关分析中出现了显著的相关系数 $r_3=0.81$！接着作者们用加权平均法计算出了 12 组堆肥数据中的 C/N 比率，计算的公式是：

$$C/N=(15a_1+17a_2+170a_3)/a,$$

其中 a_1 是半流体重量(磅)，a_2 是青菜根重量(磅)，a_3 是碎报纸重量(磅)，$a=a_1+a_2+a_3$ 是混合物的总重量. 计算的结果表明，前 8 组数据中的 C/N 比率大致相等，皆在 15.2 到 15.8 之间；而后 4 组数据中的 C/N 比率则皆在 23.5 到 26.0 之间.

这种计算结果说明，混合物中的 C/N 比率是影响堆肥形成时间的重要因素(事实上，我们可以利用相关分析，算得 C/N 同"堆肥形成时间"的相关系数为 $r\approx0.81$)，但是，却不是唯一的因素. 因为如果是唯一的因素的话，那么由于前 8 组数据中的 C/N 比率是大致相等的(皆在 15.2 到 15.8 之间)，因此它们的堆肥形成时间也应当大致相等. 但事实上，在这 8 组数据中，属于第一周期的 4 次试验(即第 1—4 组数据)与属于第二周期的 4 次试验(即第

5—8 组数据)中的堆肥形成时间却有着显著性的差异(这一点可以通过直接观察数据看出,也可以通过方差分析来检验其差异的显著性),这说明环境温度的改变也是影响堆肥形成时间的重要因素.此外,作者们还从生物学的角度并结合实际观察,为温度因素的重要性找到了依据:

第一周期内四次制堆肥的日期是从 7 月 13 日到 8 月 22 日,第二周期的四次制堆肥的日期是从 8 月 10 日到 10 月 8 日.在第一周期,平均每次堆肥过程需用 27 天,而在第二周期,平均每次大约需用 38 天.根据我们的观察,第二周期的平均气温低于第一周期的平均气温.因此,环境温度是影响混合物分解率的重要因素.微生物学家 W. D. Bellamy 博士曾有过如下的评论:"在堆肥过程中,在高温情况下那些细菌进行化学反应的速度要比在低温情况下快得多.通常在低温情况下需要进行几个月的反应,而在高温情况下只需几个星期就可以完成."

作者们还指出:

在堆肥箱内提高温度的最好方法是要使混合物中 C/N 的比率达到最优,并在堆肥箱内适当充气.因为在此之前餐厅没有使用有效的堆肥技术,因此环境温度的变化就成为影响堆肥时间的重要因素.

基于上述分析,作者们根据堆肥的形成过程是一个生化过程这一本质,提出了如下看法:堆肥的形成速度主要取决于真菌培养液的活动能力,而活动能力又是与培养液中的真菌数量成正比,因此可假设混合物形成为堆肥的速度与培养液中微生物的数量成正比.但是微生物的数量又取决于其增长率,而增长率则由环境温度、混合物中的 C/N 比例以及混合物的重量所决定.

这样一来,作者们便为模型的设计和建立明确了思路,因而可以进入最重要的建模阶段了.不过应当指出的是:虽然看起来微生物的增长率似乎应当与混合物的总重量有关,但前面的相关分析

却表明 $r_2=-0.38$，即总重量与堆肥的形成时间并无明确的相关关系，因此作者们在这一点上的处理未必妥当. 而且作者们后来也发现：只要混合物的重量达到下界 100 磅以上时，若再增加混合物的重量并不显著影响堆肥的形成时间，因此可以根据反应箱的大小尽量多投入混合物. 但是混合物的投入重量有下界，如果少于 80 磅，则会因为真菌的“营养不足”而延长堆肥的形成时间. 这就是说投入少了反而不好.

建模工作的第三步是模型设计.

作者们认为他们的首要目的就是要为堆肥的形成过程建立数学模型，模型应与所给出的数据相匹配，力求达到高度准确. 他们根据前面所明确了的思路，定义出一系列函数及参数如下：

$w(t)$：混合物的重量，以磅计算.

k：比例常数，表示混合物的分解率.

$f(t)$：堆肥箱内的培养液中所含微生物的数量，设 1 单位 $=10^7$ 微生物体.

T：环境温度的平均值.

$m(T)$：每磅混合物中所含培养液中的微生物体数量的上界，依赖于温度 T.

R：C/N 比率.

$g(T,R)$：倒入堆肥箱内的培养液中所含微生物的增长或死亡系数，依赖于 T 和 R.

并且构造了两个常微方程做为堆肥模型：

$$\begin{cases} w'(t)=-kf(t), & (1) \\ f'(t)=g(T,R)f(t)[m(T)w(t)-f(t)]. & (2) \end{cases}$$

方程(1)表示混合物分解率 $w'(t)$ 与培养液中微生物数量 $f(t)$ 之间的关系. 方程(2)是修正的 Logistic 模型，描述了培养液中微生物的增长或死亡情况.

作者们根据第一步中对模型所作的 11 条假定，对出现在模型中的 3 个常数 $g(T,R)$，$m(T)$和 k 作了如下的处理：由于 T 是每个周期内环境温度的平均值，而比率 $R=C/N$ 在每个周期内保持不变，所以增长系数 $g(T,R)$在任一个已给的周期内保持为常量，在不同的周期内可以取不同的值；$m(T)$也是如此；至于 k，则在 3 个周期内都取为同一值. 具体的确定办法将在下一步中介绍.

由于微分方程系统很难得到精确解，他们采用数值计算，求近似解逼近微分方程的解. 为此，构造计算机程序，采用此程序，计算时间大约需要 15 分钟，并且当仅有 1%混合物尚未分解时，计算终止.

他们还指出，如果令 $M=m(T)w(t)$，设 c 是微生物体的存活时间，则可将方程(2)修改为如下方式：

$$f'(t)=g(T,R)f(t)[M-f(t)]-f(t-c). \qquad (3)$$

使用(3)式，可以使系统稳定在某一水平，得到控制. 因为较低的温度给出较低的 $g(T,R)$值，所以在短时间内，死亡率可以超过增长率，这样可以调节整个系统，使微生物数量稳定于低于上界 M 的一个水平. 而当温度较高时，情况恰好相反.

为简化计算，他们仍采用(2)式，模型并不因删去死亡项$-f(t-c)$而改变其基本性质.

作者们指出，(2)式中的 $m(T)w(t)$描述了培养液受制于环境温度，并连续依赖于剩余混合物的重量. 混合物的作用是为培养液提供营养，当混合物减少到一定程度时，将影响微生物的活动能力. 实际上，(2)式存在一个动态上界.

建模工作的第四步便是利用所建立的模型对所给的 3 个周期计 12 组数据进行计算，求出理论上所需的“形成堆肥的时间”(即模型预测天数)，以对模型进行检验并作误差分析.

他们所算出的结果如表 10-2 所列：

表 10-2 堆肥实验数据、模型预测数据

周期	次数	反应箱	半流体（磅）	青菜根（磅）	碎报纸（磅）	总量（磅）	C/N	起始日期	天数		
									实际数据	模型预测天数	误差
1	1	1	86	31	0	117	15.5	19900713	28	27	1
1	2	2	122	79	0	191	15.8	19900717	27	27	0
1	3	3	71	21	0	92	15.5	19900724	27	27	0
1	4	4	203	82	0	285	15.6	19900727	26	28	−2
2	5	1	79	28	0	107	15.6	19900810	33	35	−2
2	6	2	105	52	0	157	15.7	19900813	36	35	1
2	7	3	121	15	0	136	15.2	19900820	35	35	0
2	8	4	110	32	0	142	15.5	19900822	47	35	12
3	9	1	82	44	9	135	26.0	19910430	49	48	1
3	10	2	57	60	6	123	23.5	19910502	49	48	1
3	11	3	77	51	7	135	23.8	19910507	49	48	1
3	12	4	52	38	6	96	25.5	19910510	49	51	−2

可以看出，他们所计算出的模型预测天数与实际数据中的天数，除了第 8 组(即 8 月 22 日)这一组有 12 天的误差外，其余的都吻合得很好.

当然，要想数据相吻合，一个重要的关键是合理选配参数 k，$g(T,R)$和 $m(T)$. 作者们的具体做法如下：

首先，他们利用第二周期的数据，确定参数 k，$g(T,R)$和 $m(T)$，结果如下：

$$k=0.0012 \text{ 磅}\times\text{单位}^{-1}\times\text{天数}^{-1},$$

$$g=0.00009 \text{ 天数}^{-1}\times\text{单位}^{-1},$$

$$m=80 \text{ 单位}\times\text{磅}^{-1}.$$

将这些数值输入模型中，检验模型的准确性. 除了 8 月 22 日这次

试验外，其余的试验数据与预测结果相吻合．根据前面的假设 k 在三个周期内应取同一值．他们推测第一周期的 $g(T,R)$ 和 $m(T)$ 的值需要提高，因为第一周期的平均气温高于第二周期，而第三周期的 $g(T,R)$ 和 $m(T)$ 的值应低于第二周期的值．

在确定必要参数 k，$g(T,R)$ 和 $m(T)$ 后，在第一周期内，他们测定不同重量的混合物与反应时间的关系．模型预测结果表明，不同的起始重量，对堆肥时间仅有小的影响，也就是说评价混合成分因素，应注重 C/N 比率．

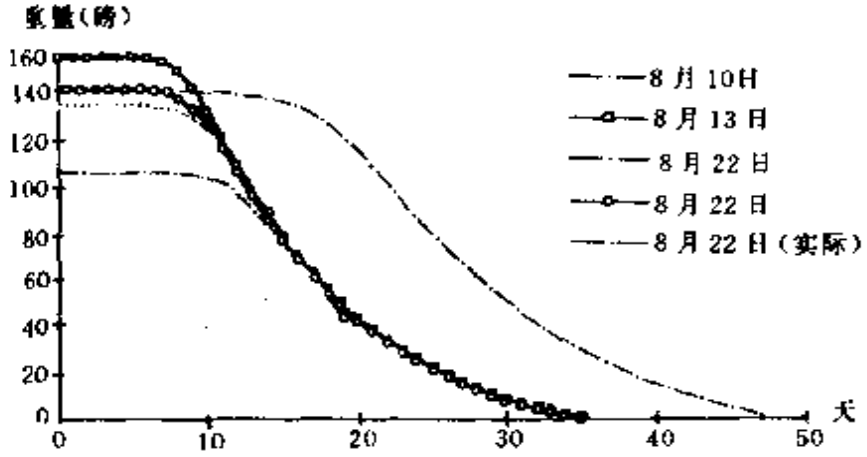

图 10-1　第二周期内的混合物整体重量与反应时间的关系

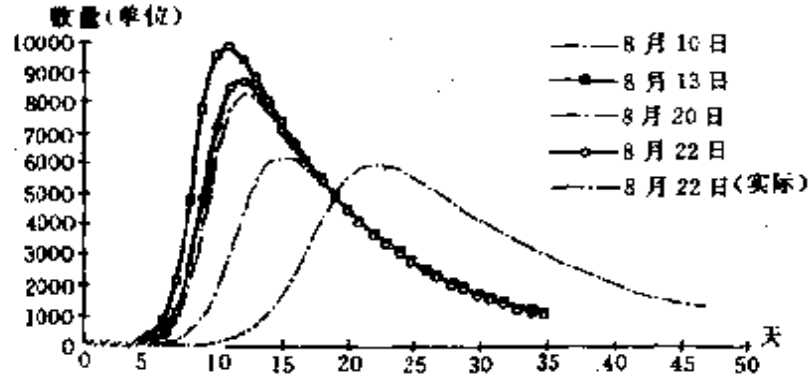

图 10-2　第二周期内微生物体的数量与反应时间的关系

他们还变化了数值计算中时间间隔的取法以检验模型的灵敏度. 开始,时间间隔长度为 1 天,但结果不能令人满意,而后,将长度取为 0.01 天,这样对第一周期,他们得出堆肥时间的平均值为 27 天. 这个间隔长度也适用于其他周期. 为了测定可接受的时间间隔的下界值,他们曾用 0.0005 天为间隔长度进行预测,所得的结果与用 0.01 天为间隔长度近似相同,为减少计算量,0.01 天是较优的间隔长度.

尽管多数试验模型预测结果与已给数据高度准确相匹配,但是不可避免地存在某些例外情况的误差,8 月 22 日开始进行反应的混合物,堆肥时间用了 47 天,而模型预测是 35 天,偏差较大. 这不能完全归于环境温度及 C/N 比率的小波动. 一种可能是 9 月份的最后一周到 10 月份的第一周这个期间内,天气变化急剧,下雨或降温. 另一种可能是工作人员的操作造成的,延误了清除堆肥的时间.

由于微分方程的解是数值解,时间间隔的取法不可能达到很精确. 当混合物仅剩 1%时,他们认为堆肥过程已经完成. 而在实际中,当从反应箱内清除堆肥时,很难准确测定这个数据.

由于 8 月 22 日数据的特殊情况,模型预测与实际反应天数的误差为±2 天(见表 10-2). 作者们认为:考虑到在一个周期内,两

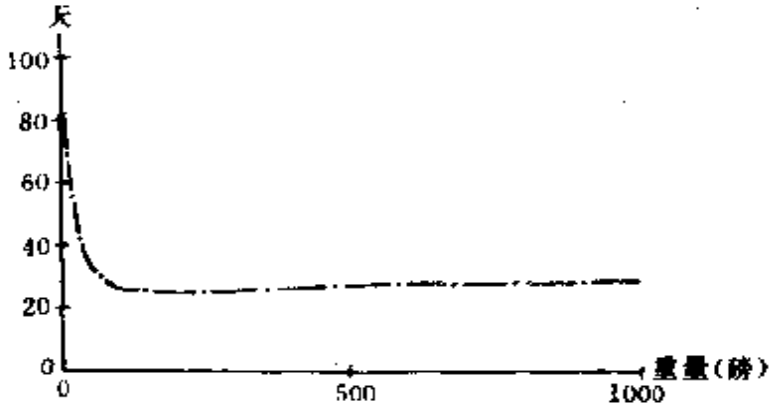

图 10-3　第一周期内的混合物的初始重量与堆肥时间的关系

次堆肥起始时间之间的温度波动，他们可能降低这个误差；再有，如果考虑到每一周期内 C/N 比率的微小变化，也可以使模型更精确.

应当指出，还有一个提高模型精确度的有效途径，就是在建立模型时不考虑 8 月 22 日的这一组数据。这在统计上叫做剔除“异常(outlier)值”，正如在歌咏比赛中“去掉一个最高分，去掉一个最低分”一样．作者们未能想到这一点，是由于他们统计知识不足之故．

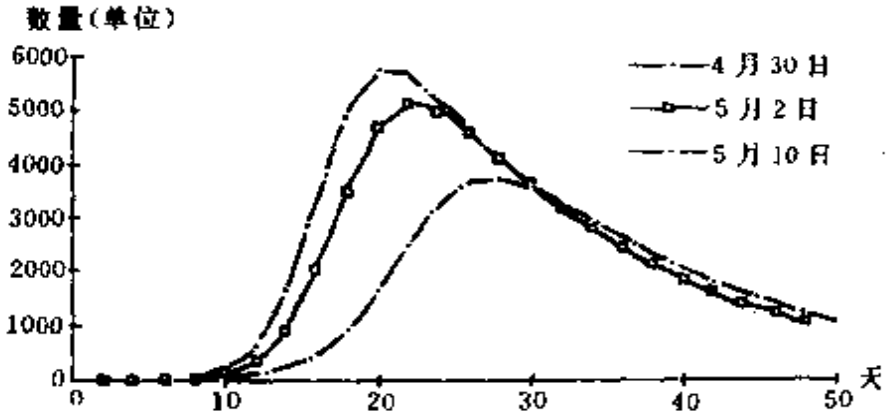

图 10-4　模型预测第三周期内微生物数量与反应时间的关系．4 月 30 日与 5 月 7 日的曲线相同

模型的另一特点是指出改变培养液中真菌的数量，则对堆肥时间的影响有限(见图 10-6)．当初始真菌数量减少到 1/10 单位时，堆肥时间的长度仅增加 10%，当真菌数量增加到 10 单位时，堆肥时间仅缩短 13%．应当说这是一个有趣的发现，可惜作者们未能从生物化学的理论角度予以解释。

建模工作的第五步是“模型评价”.

站在旁观者的角度，我们认为作者们所建立的模型有两个方面是极为成功的：首先，通过这一模型，进一步明确了堆肥的形成时间与混合物的总重量关系不大，而改变培养液中的真菌数量对堆肥形成时间的影响也有限，从而更加使人们认识到影响堆肥形

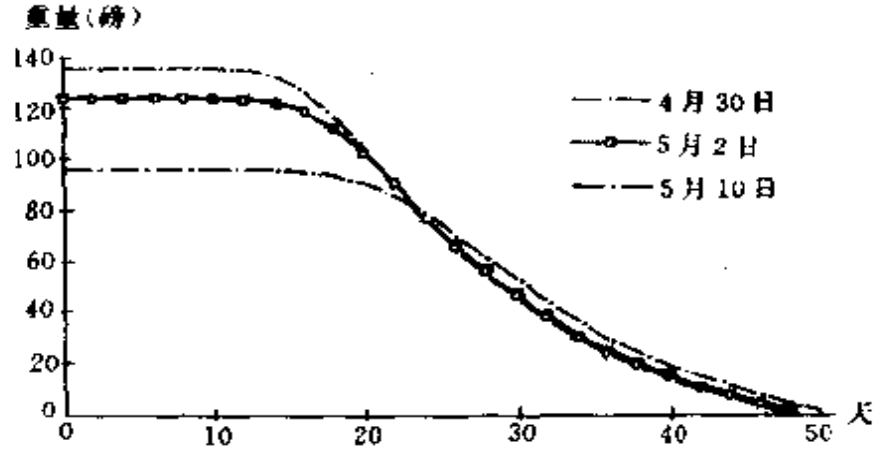

图 10-5　模型预测第三周期内混合物重量与反应时间的关系.
4 月 30 日与 5 月 7 日的曲线相同

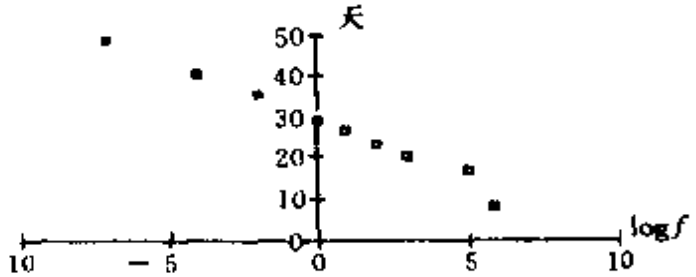

图 10-6　模型预测的反应时间长度与初始微
生物体数量的对数值之间的关系

成时间的关键因素是混合物中的 C/N 比率以及环境温度，因而为如何加速堆肥的形成指出了途径. 其次，模型预测结果与已给数据高度准确地相匹配，这是模型最有价值的一点. 模型近似模拟出真菌培养液的活动能力与环境温度及 C/N 比率的关系，微分方程解的曲线给出了在堆肥箱内已给有限资源和环境条件下，真菌数量如何增长或死亡(见图 10-4). 混合物与时间的关系图给出了混合物的分解率在开始是慢的，曲线斜率的绝对值增大，在达到一定时间后，斜率的绝对值逐渐下降，趋于零(见图 10-5).

对于模型所存在的缺点，作者们的看法是：

模型存在的缺点是当我们确定常数 k，$g(T,R)$和 $m(T)$时，生

物实验数据没有起作用.换言之,建模唯一依赖已给出的实验数据,而这些数据不能更精确确定模型中的参数.

模型具有弹性,如果我们掌握更多的数据(例如,每日测定反应物内的温度),我们可以调整模型中的常量,精确参数.

关于常数 k,$g(T,R)$和 $m(T)$如何确定,确实是一个问题,作者们没有给出一个确定的办法,而且给人的一个印象是:他们是单纯由实验数据"凑"出来的.这样,就局限了这一模型的作用,难以用它来对以后的实验作预测.

其次,在作者们的工作中,没有寻找有利于堆肥形成的最佳的C/N 比例,也没有给出如何随外界温度 T 的变化而改变 C/N 比率的办法,这一缺陷是作者们未能意识到,更未能指出的.

此外,作者们自己认为:在温度问题上,模型过分强调了气候条件和分解率密切相关.同时,他们设想,为描述堆肥箱内的热量向外界环境的扩散,他们应该增加第三个方程,使模型更完善.

最后,作者们根据题目的要求,提出了若干条改进堆肥形成的建议.应当说这些建议是中肯的,并且也是切实可行的.作者们的建议是:

由于餐厅无力增添新的堆肥设施,且用于堆肥的经费有限,为此,我们对堆肥改进提出以下建议:

1. 增加每批要处理的混合物重量.

2. 控制混合物中报纸含量.

3. 合理保护堆肥箱,避免环境温度大幅度变化而影响到堆肥时间.尽量防止外界与箱内之间的热扩散.

4. 箱内应有充足的氧气,供给微生物的生化反应.

5. 混合、搅拌反应物,有助于提高堆肥速度.至少每隔一天,搅动一次.

6. 反应箱内的最生温度为 120F,湿度为 100%.

7. 改变混合物成分.建议投入较多的青菜叶,果皮核,玉米梗

等，可加速分解.

作者们还在文末附上了他们所使用的参考文献：

参 考 文 献

[1] Johnson, George B., and Peter H. Raven. 1992, *Biology*. Boston, MA: Mosby-year Book.

[2] Koepf, H. H. 1986, *Compost: what it does*. Bio-Dynamics (Bio-Dynamic Farming and Gardening Assoc. Ins) No. 77.

[3] Stewart, James. 1991, *Calculus*, Pacific Grove, CA: Brooks/Cole.

[4] Sussman, Vic. 1982. *Easy Composting Emmaus*, PA: Rodale Press.

第 11 讲
方差分析与相关分析

□ 苏　淳

§1　方差分析

1. 问题的提出

在堆肥生成问题(参阅第 10 讲)中，我们曾遇到过这样的问题：堆肥的形成时间的长短究竟与哪些因素有关？可以考虑的因素粗粗列出来就有不少：混合物的总量，其中各种成份的比例，环境温度，微生物的数量，等等. 那么，在这些因素中，哪些是主要的，哪些是次要的？除了我们所考虑到的因素之外，是否还有其他因素在起作用？此外，我们还应顾及各种非人为所能控制的随机因素的影响. 因此，我们需要有一定的统计方法来帮助处理试验数据，以便对上述问题作出一个较为合理的回答.

方差分析就是一种可以用来帮助回答上述问题的统计方法. 方差分析的内容很丰富，在本讲中只能就其中的几个重要而简单的方面作些粗浅的介绍.

设想有一项我们感兴趣的指标 Y，例如：堆肥的形成时间，某

种工业产品的数量、质量，农作物的亩产等. Y 之值受到一些因素 $X_1,\cdots,X_p$ 的影响，此外还受到随机误差的影响. 我们希望了解因素 $X_1,\cdots,X_p$ 对 Y 的影响的具体情况，也需要了解随机误差的影响有多大.

这些因素按其性质可以分为两类：

(1) 属性的(即分类性的)，例如，种子品种甲，品种乙，品种丙，等等，这时，我们就称"种子"这一因素有 3 个"水平"，每一品种称为一个"水平"，可将它们编号为水平 1，2，3.

(2) 数量性的，例如，堆肥问题中的 C/N 比例，在每一次试验中都有一个具体的数值.

当然，我们可以把数量性的因素"属性"化，即将它们分类，例如，将 C/N 比在 15—16 之间的归入"水平 1"，在 23—24 之间的归入"水平 2"，在 25—26 之间的归入"水平 3"，等等.

如果一个问题中，影响指标 Y 的各种因素都可以"属性"化，那么就可以采用方差分析的方法来作统计分析了.

2. 单因素方差分析(1)

如果我们希望考察某一个因素 A 对指标 X 的影响，那么就可以考虑采用单因素方差分析. 这时，在安排试验时，应当将其他因素都适当地固定下来. 如果我们面对的是别人所提供的试验数据，例如堆肥形成问题，由于试验已经做过，而做时其他因素并未固定，那么也可以采用单因素方差分析，但这时在分析结果的解释上与前有所不同.

我们来看"堆肥形成问题"中温度因素 A 对堆肥形成天数 X 的影响分析.

首先将温度因素 A"属性"化，可以分为 3 个水平：水平 1：发酵起始时间在 7 月份；水平 2：发酵起始时间在 8 月份；水平 3：发酵起始时间在 4 月底和 5 月初. 于是，根据题目中所提供的试验数

据知，在每种水平之下都做了 4 次试验，相应的“堆肥形成天数”为：

水平 1：28，27，27，26——平均值为 27.

水平 2：33，36，35，47——平均值为 37.75

水平 3：49，49，49，49——平均值为 49

这种情形可以一般性地描述为：有一个因素 A，它有 k 个水平，在每个水平之下都做了 n 次试验，得试验数据如下：

水平 1：$X_{11}, X_{12}, \cdots, X_{1n}$ ——$\overline{X_1}=\sum_{j=1}^{n} X_{1j}/n$

水平 2：$X_{21}, X_{22}, \cdots, X_{2n}$ ——$\overline{X_2}=\sum_{j=1}^{n} X_{2j}/n$

…… ……

水平 k：$X_{k1}, X_{k2}, \cdots, X_{kn}$ ——$\overline{X_k}=\sum_{j=1}^{n} X_{kj}/n$

由于在每一种水平 i 之下，因素 A 是固定的，因此可以认为 $X_{i1}, X_{i2}, \cdots, X_{in}$ 之间的差异与因素 A 无关. 当其他因素也都适当固定时，我们可以认为这种差异是由随机误差造成的；如果其他因素未能固定，则可认为这种差异是由众多因素（包括随机误差）共同作用的结果. 经验表明，如果我们假定 X_{ij} 的分布服从正态 $N(\mu_i, \sigma^2)$，那么在大多数情况下是合理的，且往往能得出合理的结论（注意，在我们的假定中，不同的水平之下，μ_i 可能不同，但 σ^2 却是相同的）. 于是我们的问题就转化为如下的假设检验问题：

H_0：$\mu_1=\mu_2=\cdots=\mu_k$

$\leftrightarrow H_1$：$\mu_1, \mu_2, \cdots, \mu_k$ 不全相同.

我们来计算出水平 i 之下的样本方差：

$$S_i^2=\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\overline{X_i})^2, \tag{1}$$

其中 $\overline{X_i}=\frac{1}{n}\sum_{j=1}^{n} X_{ij}$. 于是可得关于 σ^2 的 k 个估计值 $S_1^2, S_2^2, \cdots, S_k^2$. 将它们平均之，得到 σ^2 的估计

$$S^2=\frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X_i})^2. \tag{2}$$

下面来分析各水平不同所造成的影响. 水平 i 之下的期望值 μ_i 的估计为$\overline{X_i}=\frac{1}{n}\sum_{j=1}^{n}X_{ij}$. 易知，$\overline{X_1},\overline{X_2},\cdots,\overline{X_k}$之间的差异越大，那么 $\mu_1,\mu_2,\cdots,\mu_k$ 之间的差别就越大，从而各水平之间的差异也就越大. 我们算出$\overline{X_1},\overline{X_2},\cdots,\overline{X_k}$的样本方差

$$S_*^2=\frac{1}{k-1}\sum_{i=1}^{k}(\overline{X_i}-\overline{X})^2, \tag{3}$$

其中，
$$\overline{X}=\frac{1}{k}\sum_{i=1}^{k}\overline{X_i}=\frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}X_{ij}, \tag{4}$$

显然，$\overline{X}$ 就是全部 kn 个数据 X_{ij}的算术平均值，且 S_*^2 反映出了各水平差异的影响.

通过简单的计算，可以发现在上述各量之间有如下的关系式：

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X})^2=\sum_{i=1}^{k}n(\overline{X_i}-\overline{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X_i})^2. \tag{5}$$

在统计上习惯地把上式写为

$$SS_T=SS_A+SS_e. \tag{6}$$

其中 SS 表示平方和，是英文 Sums of Squares 的缩写，$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X})^2$称为总平方和；$SS_A=\sum_{i=1}^{k}n(\overline{X_i}-\overline{X})^2$ 称为“因素 A 平方和”，它反映了因素 A 的水平差异所造成的影响；$SS_e=\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X_i})^2$称为“误差平方和”，如前所说，它反映了随机误差以及其他各种因素所造成的综合影响. “方差分析”这一名词的由来及其精髓就反映在(6)式中，也就是说，它将“总平方和”分解为因素 A 的影响 SS_A 和随机误差与其他因素的影响 SS_e 这样两个部分.

我们现在所面临的问题是：因素 A 是否对指标 X 有较大的影响？这个问题可以通过考察由“由因素 A 所造成的‘总方差’是否较之于‘误差总方差’显著地大”来回答. 为此我们考察比值

$$F=\frac{nS_*^2}{S^2}=\frac{SS_A/(k-1)}{SS_e/k(n-1)}. \tag{7}$$

由数理统计学有关知识可知，F 是一个随机变量，它服从自由度为 $((k-1),k(n-1))$ 的 F 分布.

剩下来的问题便是：F 大到什么地步才算足够大呢？这就需要根据有关知识并结合实际需要来确定了.一般来说，当我们确定下来某个值 C 以后，便在 $F>C$ 时认为因素 A 的作用足够大，因此认为 $\mu_1,\mu_2,\cdots,\mu_k$ 不全相同，因而拒绝 H_0；相反，如果 $F\leqslant C$，则认为不能拒绝 H_0，从而认为因素 A 对指标 X 不产生较大的影响. 从数理统计学假设检验的一般理论来说，倾向于不轻易否定 H_0，因此都把 C 定得足够大，使得 $F>C$ 的概率很小，如果记

$$P(F>C)=\alpha, \tag{8}$$

那么通常将 α 定为 0.05，甚至 0.01，并通过查 F 分布表定出相应的 $C=F_\alpha(k-1,k(n-1))$ 来.

这时，如果我们所算得的 F 值大于 $F_{0.05}(k-1,k(n-1))$，则称“因素 A 在水平 $\alpha=0.05$ 下显著”，并在相应的 F 上加一个 * 号，表示显著；如果 $F>F_{0.01}(k-1,k(n-1))$，则称“因素 A 在水平 $\alpha=0.01$ 下显著”，并在相应的 F 上加两个 * 号，表示高度显著.

现在，我们就可以回过头来检验堆肥形成问题中的温度因素的作用是否显著了.

由前所述，我们有 $k=3,n=4$，

$$\overline{X}=\frac{1}{3}(\overline{X_1}+\overline{X_2}+\overline{X_3})=37.92,$$

$$SS_A=4[(\overline{X_1}-\overline{X})^2+(\overline{X_2}-\overline{X})^2+\overline{X_3}+\overline{X})^2]=968.2,$$

$$SS_e=\sum_{i=1}^{3}\sum_{j=1}^{4}(X_{ij}-\overline{X_i})^2=120.74,$$

将上述数据代入(7)式，算得

$$F=\frac{SS_A/2}{SS_T/9}=\frac{968.2\times 9}{120.74\times 2}=36.08.$$

查表得

$$F_{0.05}(2,9)=4.26, F_{0.01}(2,9)=8.02.$$

显然有 $F>F_{0.01}(2,9)$，这说明即使在 $\alpha=0.01$ 的水平下，温度因素对堆肥形成时间的影响也是显著的，因此为高度显著，故拒绝 H_0 承认 H_1.

这样一来，我们就从方差分析的角度支持了第 10 讲中所介绍的论文的作者们的看法，印证了他们所引用的微生物学家 W. D. Bellamy 博士的观点："在堆肥过程中，在高温情况下那些细菌进行化学反应的速度要比在低温情况下快得多."我们还可将如上所作的方差分析列成表 11-1.

表 11-1　三种温度水平下堆肥形成时间方差分析表

方差来源	平方和	自由度	均　方	F 值
温　度	968.2	2	484.1	36.08**
其他综合因素	102.74	9	11.42	
总　和	1 088.94	11		结论：高度显著

3　单因素方差分析(2)

在前面所讨论过的问题中，各个水平之下所作的试验次数是相同的，现在我们要来讨论试验次数不同的情况.

设因素 A 共分为 k 个水平，在水平 $1,2,\cdots,k$ 上分别作了 n_1，$n_2,\cdots,n_k$ 次试验，$N=n_1+n_2+\cdots+n_k$，得到了试验数据

$$X_{ij},\ j=1,2,\cdots,n_i;\ i=1,2,\cdots,k.$$

那么又该怎样作方差分析呢？

显然，现在我们有

$$\overline{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij},\ i=1,2,\cdots,k, \tag{9}$$

$$\overline{X}=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}=\frac{n_i}{N}\sum_{i=1}^{k}\overline{X_i},$$

并且(6)式仍然成立，不过其中

$$\begin{aligned}&SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X})^2,\ SS_A=\sum_{i=1}^{k}n_i(\overline{X_i}-\overline{X})^2,\\&SS_e=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X_i})^2.\end{aligned}\tag{10}$$

结合对自由度的分配，可以得到

$$F=\frac{SS_A/(k-1)}{SS_e/(N-k)}=\frac{SS_A}{SS_e}\cdot\frac{N-k}{k-1}.\tag{11}$$

由数理统计知识知，这里的 F 服从自由度为$(k-1,N-k)$的 F 分布.

利用(11)式，我们就可以作方差分析了.

现在来看堆肥形成问题中的 C/N 比对于形成时间的影响是否显著. 如同本讲 § 1.1 所述，可将试验数据中的C/N 比划分为 3 个水平，并且按照那里的分类办法，我们有：$n_1=8,n_2=n_3=2$，并且在各个水平下的发酵天数为：

水平 1：28,27,27,26,33,36,35,47　　$\overline{X_1}=32.375$

水平 2：49,49　　$\overline{X_2}=49$

水平 3：49,49　　$\overline{X_3}=49$

且总平均天数仍为 $\overline{X}=37.92$，于是可算得

$$SS_A=8(\overline{X_1}-\overline{X})^2+2(\overline{X_2}-\overline{X})^2+2(\overline{X_3}-\overline{X})^2=736.6,$$

$$SS_e=\sum_{i=1}^{3}\sum_{j=1}^{n_i}(X_{ij}-\overline{X_i})^2=315.59,$$

$$F=\frac{SS_A}{SS_e}\cdot\frac{N-k}{k-1}=\frac{736.6\times9}{315.59\times2}=\frac{6629.4}{631.18}=10.5,$$

查表得 $F_{0.05}(2,9)=4.26$，$F_{0.01}(2,9)=8.02$，故知 $F>F_{0.01}(2,9)$，表明“因素 C/N 比”对堆肥形成时间的影响高度显著(意即在 $\alpha=0.01$的水平下显著). 这一断言有力地支持了第 10 讲所介绍的

论文之作者们的论点，也表明了数理统计是生物学研究中的得力工具.

下面再来看一看“混合物总重量”对“堆肥形成时间”之影响的方差分析.

我们将总重量分为 5 个水平，其中：水平 1 为 91—115 磅；水平 2 为 116—140 磅；水平 3 为 141—165 磅；水平 4 为 191—215 磅；水平 5 为 266—290 磅. 由试验数据表知：$n_1=3, n_2=5, n_3=2, n_4=n_5=1$，且相应的发酵天数为：

水平 1：27，33，49	$\overline{X_1}=36.33$
水平 2：28，35，49，49，49	$\overline{X_2}=42$
水平 3：36，47	$\overline{X_3}=41.5$
水平 4：27	$\overline{X_4}=27$
水平 5：26	$\overline{X_5}=26$

注意到仍有 $\overline{X}=37.92$，于是不难算得

$$\begin{aligned}SS_A &= 3(36.33-37.92)^2+5(42-37,92)^2\\ &\quad +2(41.5-37.92)^2+(27-37.92)^2+(26-37.92)^2\\ &=376.57,\end{aligned}$$

$$SS_e=\sum_{i=1}^{5}\sum_{j=1}^{n_i}(X_{ij}-\overline{X_i})^2=713.17,$$

$$F=\frac{SS_A/k-1}{SS_e/N-k}=\frac{376.57\times 7}{713.17\times 4}=0.92,$$

查表得 $F_{0.05}(4,7)=4.12$，于是知 $F<F_{0.05}(4,7)$.

这就告诉我们，混合物总重量对堆肥形成时间的影响，即使在 $\alpha=0.05$ 的水平之下，也是不显著的. 从而再次证实了作者们的观点.

以上我们只是就单因素方差分析法作了介绍. 至于多因素方差分析方法，限于篇幅，不再涉及，有兴趣的读者可以参阅本讲末尾所附的参考文献. 我们在这里介绍方差分析的目的是想引起读

者对这一统计方法的重视，了解其基本思想，以便在今后的应用中作为参考.

§2 相关分析

1. 问题的提出

在方差分析中，我们讨论过一个指标 Y 可能受到多个因素 $X_1,\cdots,X_p$ 的影响，并特别就其中 $p=1$（单因素）的情形作了较详细的讨论. 在那里，我们要解决的问题是：因素 X 对指标 Y 的影响是否显著？因此，一般说来，我们是将“因素 X”看成“因”，而将指标 Y 看成“果”的，即是说 X 与 Y 处于不对称的位置.

在现实中，我们常常会碰到这样的两个量 X 与 Y，例如，一个人的身高 X 与他的体重 Y，一个人的抽烟量 X 与他的喝酒量 Y，一个家庭的月收入数 X 与该家庭的孩子数 Y，等等. 我们都知道，在这些 X 与 Y 之间都存在着一定的关系，但是却没有什么因果关系可言，即是说 X 与 Y 的地位是平等的，或言对称的. 并且在 X 与 Y 之间也没有某种确切的关系可言. 当我们知道某人的身高 X 之后，无法说出他的体重 Y 一定是多少；反之也一样，知道了体重 Y，也不能说出他的身高 X. 但是我们都知道，在身高 X 与体重 Y 之间有着一定的关系，个子高的人一般来说体重也大一些，个子矮的人一般来说体重也小一些. 因此我们可能会希望了解 X 与 Y 之间的关系究竟密切到什么程度.

当我们所遇到的两个变量 X 与 Y 的地位是平等的，并且只希望了解 X 与 Y 之间关系的密切程度时，就可以运用“相关分析”这一统计方法. 当然，这里只是说的两个变量的情形. 对于多个地位平等的变量 $X_1,\cdots,X_p$，也有相应的“相关分析”方法.

在这里，我们应当强调指出：相关分析只是用于“X 与 Y 的关

系密切到何种程度”的. 如果我们的目的不在于此，而是希望在知道了某人的身高 X 后去“预测”他的体重 Y 的话，那么就不属相关分析的内容，而是属于统计学上的另一类问题——回归分析了. 另外，什么叫做“关系密切”，“何种关系密切”，也都是需要明确的. 为此我们提请读者在下面注意这一点. 此外，我们还希望读者注意：对经相关分析而得出的相关指标作解释时需持慎重的态度. 例如，考察一个人的烟支出 X 和酒支出 Y 时，统计分析也许会显示两者显著相关. 如果我们据此就试图去寻找两者互相影响的学理上的影响，恐怕就会是徒劳无益的了. 事实上，这两者表面上的相关，可能只是因为它们都与一个人的收入有关. 因此，我们用“相关分析”方法得出的相关关系应称为“统计相关”. “统计相关”只是表示从统计资料上看，两个变量的取值存在着一定的关联，而不断言它们有因果关系，并且统计学本身也不试图去说明存在这种关联的原因.

2. 相关系数

设 X 和 Y 是两个随机变量，它们的均值(即数学期望)分别为 a 和 b，它们的方差分别为 σ_1^2 和 σ_2^2，记

$$\mathrm{Cov}(X,Y)=E[(X-a)(Y-b)], \tag{12}$$

即用 $\mathrm{Cov}(X,Y)$ 表示乘积 $(X-a)(Y-b)$ 的均值，称为 X 与 Y 的协方差. 在 X 与 Y 无关联时，就有

$$\mathrm{Cov}(X,Y)=0.$$

因此可以将 $\mathrm{Cov}(X,Y)$ 作为 X 与 Y 是否相关的一个度量. 但这样做有一个麻烦，例如，如果 X 表示一个人的身高，Y 表示一个人的体重，那么在 X 选用不同的重量单位(厘米或米)、Y 选用不同的重量单位(斤或公斤)时，$\mathrm{Cov}(X,Y)$ 的值是不同的. 为了克服这一困难，我们将它单位化，即令

$$r=r(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sigma_1\sigma_2},\tag{13}$$

并将 r 称为 X 与 Y 的相关系数. 显然有 $-1\leqslant r\leqslant 1$，从而消除了不同度量单位造成的影响. 可以想见，用 r 作为变量 X 同 Y 相关程度的一个度量是合理的.

现在，我们来对上面的话作些解释：

(1) r 之值与 X 和 Y 所选用的度量单位无关.

事实上，当身高由厘米换为米时，X 之值缩小了 100 倍，当体重由斤换为公斤时，r 之值缩小了 2 倍，于是 $\mathrm{Cov}(X,Y)$ 的值缩小了 200 倍. 但这时，乘积 $\sigma_1\sigma_2$ 的也相应地缩小了 200 倍，因此 r 之值不变.

(2) r 之值是 X 同 Y 线性相关程度的一个度量.

事实上，如果 X 与 Y 之间存在明确的线性关系，即有 $Y=cX+d$ 时，我们有

$$EY=cEX+d=ca+d,$$

其中 $EX=a$ 是 X 的均值. 如果记 $b=EY$，就有

$$b=ca+d,$$

从而就有

$$Y-b=cX+d-(ca+d)=c(X-a),$$

于是

$$\mathrm{Cov}(X,Y)=E[(X-a)(Y-b)]=cE(X-a)^2=c\sigma_1^2.$$

而因 $\sigma_2^2=E(Y-b)^2=c^2E(X-a)^2=c^2\sigma_1^2$，所以 $\sigma_2=|c|\sigma_1$，于是 $\sigma_1\sigma_2=|c|\sigma_1^2$，这样就有

$$r=\frac{\mathrm{Cov}(X,Y)}{\sigma_1\sigma_2}=\frac{c}{|c|}=\pm 1,$$

即 $r=1$ 或 -1. 这就是说，当 X 与 Y 有线性关系 $Y=cX+d$ 时，$|r|$ 之值达到 1，并且当 $c>0$ 时，$r=1$；当 $c<0$ 时，$r=-1$.

并且反之，如果有 $r=\pm 1$，即有

$$E(X-a)(Y-b)=\pm\sigma_1\sigma_2, \tag{14}$$

我们可以证明必相应地有

$$Y-b=\frac{\sigma_2}{\sigma_1}(X-a)\text{或}Y-b=-\frac{\sigma_2}{\sigma_1}(X-a).$$

事实上，如果记 $c=r\sigma_2/\sigma_1$，其中 $r=\pm1$，就有

$$\begin{aligned}&E[(Y-b)-c(X-a)]^2\\&=E(Y-b)^2+c^2E(X-a)^2-2cE(Y-b)(X-a)\\&=\sigma_2^2+\frac{\sigma_2^2}{\sigma_1^2}\cdot\sigma_1^2-2\cdot r\sigma_2/\sigma_1\cdot r\sigma_1\sigma_2\\&=\sigma_2^2+\sigma_2^2-2\sigma_2^2=0,\end{aligned}$$

这就表明

$$Y-b=c(X-a),\text{a. s.}$$

即 X 与 Y 之间几乎必然存在线性关系.

综合上述，我们可以这样认为：当且仅当 X 与 Y 之间存在线性关系时，有 $|r|=1$.

相应地，如果 $r=0$，我们就说 X 与 Y 不线性相关．应当注意这里的用词：不线性相关. 这是因为当 X 与 Y 之间存在某种曲线关系，例如 $Y=X^2$ 时，也可能有 $r=0$. 因此 $r=0$ 只是表明了 X 与 Y 之间不存在线性关系.

当 $r\neq0$ 时，我们就说 X 与 Y 之间存在某种程度的线性相关关系，并且 $|r|$ 之值越接近于 1，X 与 Y 的线性相关程度就越大，(参阅图 11-1).

另外，当 $r>0$ 时，我们还称 X 与 Y 线性正相关；当 $r<0$ 时，称线性负相关. 在不致于造成误解时，可略去用语中的“线性”二字.

3 相关性检验

考察两个变量 X 和 Y，设对它们作了 n 次观测，得到数据 $(X_1,Y_1),(X_2,Y_2),\cdots(X_n,Y_n)$，据此我们可以构造出如下的统计量

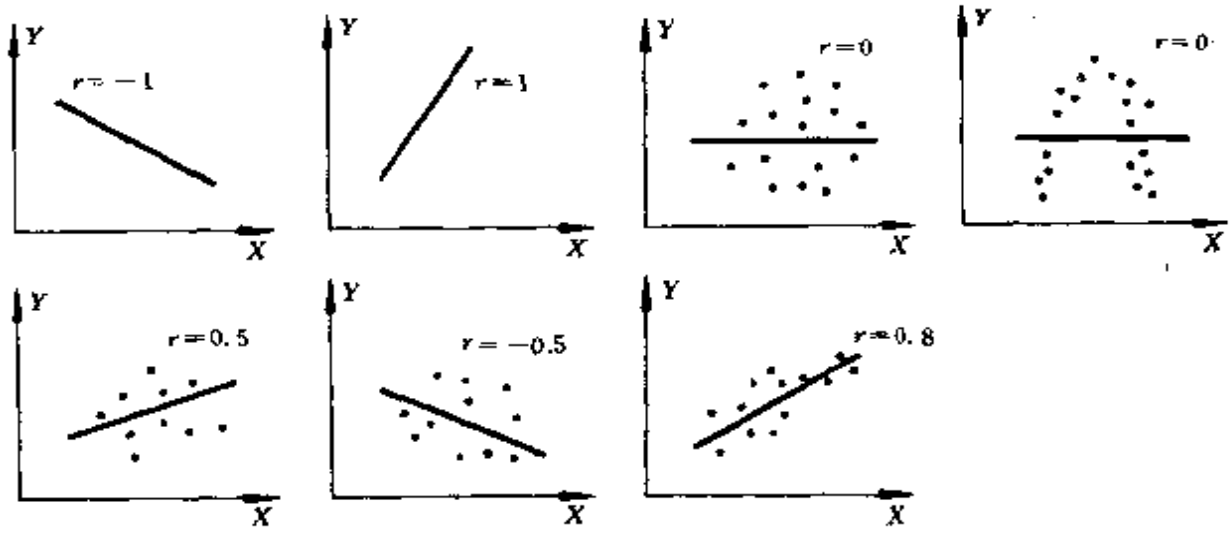

图 11-1

$$\hat{r}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\cdot\sum_{i=1}^{n}(Y_i-\overline{Y})^2}},\tag{15}$$

$\hat{r}$ 称为变量 X 和 Y 的通过样本$(X_1,Y_1),\cdots,(X_n,Y_n)$所算出的样本相关系数. 在数理统计上就以 $\hat{r}$ 作为对 X 和 Y 的相关系数 r 的估计. 所谓相关分析，就是通过对 $\hat{r}$ 作统计分析以判定 X 和 Y 的线性相关程度.

应当注意，由于 $\hat{r}$ 是借助样本计算出来的，而样本观察值会受到随机性的影响，所以在算得 $\hat{r}\neq 0$ 时，我们不一定就能断言 X 与 Y 有关. 例如，一般人都会相信，一个人的姓氏笔画 X 与他的工资数 Y 无关. 但是若真去抽取 n 个人观察其 X、Y 值，由此所算出的 $\hat{r}$ 不一定就是 0.

所以，合理的办法是确定一个界限 c，当 $|\hat{r}|$ 超过 c 时，判定 X 与 Y 有关联；当 $|\hat{r}|$ 小于 c 时，就判定它们无关. 这样一来，我们的问题就成为一个假设检验问题：

H_0：X 与 Y 不相关↔H_1：X 与 Y 相关.

因此，c 值的选取与所选定的检验水平 α 有关. 如果我们想要得出的结论是 H_1，而所选定的 α 很小，就意味着我们要求有很强的证据才判定 X 与 Y 有关联. 一般取 $\alpha=0.05$ 或 $\alpha=0.01$. 当取 $\alpha=0.05$ 而作出结论 H_1 时，称 X 与 Y 的相关性“显著”；当取 $\alpha=0.01$ 而作出结论 H_1 时，则称 X 与 Y 的相关性“高度显著”. 另外，c 值的选定还与观察次数 n 有关.

数理统计上证明了，在假定 (X,Y) 服从正态分布且观察次数为 n，并选取水平为 α 时，应当将 c 取为

$$c=\frac{t_{n-2}\left(\frac{\alpha}{2}\right)}{\sqrt{n-2+t_{n-2}^2\left(\frac{\alpha}{2}\right)}}, \tag{16}$$

其中 $t_{n-2}\left(\frac{\alpha}{2}\right)$ 的意思如图 11-2 所示，即 $p(t_n>t_n(\alpha))=\alpha$，这里 n 为自由度，$t_n(\alpha)$ 之值可由 t 分布表查出.

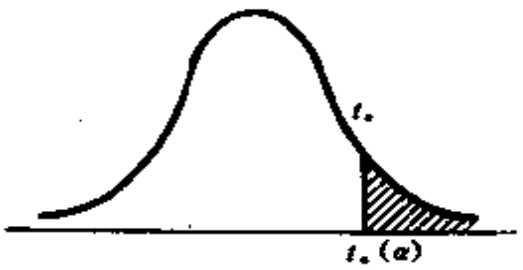

图 11-2

下面来看几个例子.

例 1 以 X 表示一个家庭的月收入，以 Y 表示其月支出，以元为单位，随机抽取了 10 个家庭，得观察值 (X_i,Y_i) 如下：

月收入 X_i	200	150	200	250	150	200	250	300	250	120
月支出 Y_i	180	160	200	250	140	230	210	250	230	140

我们来作 X 与 Y 的相关分析，容易算出

$$\Sigma X_i=2070,\ \Sigma Y_i=1990,\ \Sigma X_i^2=456900,$$

$$\Sigma X_iY_i=431300,\ \Sigma Y_i^2=412100.$$

注意到 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$，不难将(15)式变形为

$$\hat{r}=\frac{\sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i/n}{\sqrt{\left[\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2/n\right]\left[\sum_{i=1}^{n}Y_i^2-(\sum_{i=1}^{n}Y_i)^2/n\right]}}. \tag{17}$$

于是算得

$$\hat{r}=\frac{431300-2070\times1990/10}{\sqrt{456900-2070^2/10}\sqrt{412100-1990^2/10}}=0.917.$$

取 $\alpha=0.01$，查表得 $t_{n-2}\left(\frac{\alpha}{2}\right)=t_8(0.005)=3.355$，由(2.4)式算出 $c=3.355/\sqrt{8+3.355^2}=0.765$. 现在有 $\hat{r}=0.917>0.765$，故知 X 与 Y 相关高度显著. 显然，这一结论是合乎常情的，即收入多的人家一般来说支出也多.

例 2 分别以 X 和 Y 表示一个人的身高和体重，从某大学中随机地抽取了 10 名男生，得观察值如下(X 以米为单位，Y 以千克为单位)：

身高 X_i	1.71	1.63	1.84	1.90	1.58	1.60	1.75	1.78	1.80	1.64
体重 Y_i	65	63	70	75	60	55	64	69	65	58

为作 X 与 Y 的相关分析，算出

$\Sigma X_i=17.23$，$\Sigma Y_i=644$，$\Sigma X_i^2=29.7935$，$\Sigma Y_i^2=41790$，$\Sigma X_iY_i=1114.88$.

从而就有

$$\sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i/n=1114.88-17.23\times644/10=5.268,$$

$$\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2/n=29.7935-(17.23)^2/10=0.10621,$$

$$\sum_{i=1}^{n}Y_i^2-(\sum_{i=1}^{n}Y_i)^2/n=41790-644^2/10=316.4,$$

代入(2.5)，即得

$$\hat{r}=5.268/\sqrt{0.10621\times316.4}=0.91>0.765,$$

所以 X 与 Y 相关性高度显著.

4. 多个变量的情形

在一个问题中往往涉及多个变量 $X_1, X_2, \cdots, X_p$，为了研究它们之间的相关性问题，数理统计上给出了多种方法. 这些方法是从各种不同的角度来考虑问题的.

例如，以 X_1 和 X_2 分别记一个人在烟和酒上的月支出数，以 X_3 记他的月收入数. 如果去收集一些资料，按前面所述的方法计算 X_1 和 X_2 的相关系数 $\hat{r}$，则通常有 $\hat{r}>0$，即正相关. 这表明一般在烟上消费较多的人，倾向于在酒上消费也较多. 我们自然很难设想：抽烟会是造成好酒的原因. 主要的原因恐怕取决于月收入数 X_3：收入高的人有能力抽好烟喝名酒，收入低的人则相反. 因此我们自然希望了解：在去掉 X_3 这个因素后，X_1 与 X_2 的相关还有多大？为了解决这个问题，数理统计学上给出了有关的方法，并引入了“偏相关系数”这个量. 这样我们就得到了如下的概念：

如果在 X_1 和 X_2 的关系中，有一部分是因为受到 $X_3, \cdots, X_p$ 的共同影响而产生的. 当把这种影响按照一定的统计方法从 X_1 和 X_2 中消除后所作的相关性分析，称为“偏相关分析”，如果此时 X_1 与 X_2 仍相关，则称为 X_1 和 X_2 对 $(X_3, \cdots, X_p)$ 的“偏相关”.

除了偏相关之外，统计上还有多种相关性指标，其中较为重要的还有：

以 X_1 为一方，以 $(X_2, \cdots, X_p)$ 为另一方，考察两者的相关性，称为 X_1 对 $(X_2, \cdots, X_p)$ 的“复相关”.

由于这些问题都涉及较为专门的数理统计知识，限于篇幅，不便赘述.

参 考 文 献

[1] 陈希孺、倪国熙：《数理统计学教程》，上海科技出版社，1988 年.

[2] 陈希孺、苏淳：《统计学漫话》，科学出版社，1987 年.

[3] 傅权、胡蓓华：《基本统计方法教程》，华东师范大学出版社，1989 年.