

第 1 章 群体的遗传组成

基因和基因型的频率

描述一群生物个体的遗传组成，我们应详细说明它们的基因型，并说出每一种基因型有多少。假若我们不在意基因型间表型差异的性质，那么这将是一个完整的描述。为简单起见，假设我们关注某一个常染色体位点 A ，并且这个位点上的两个不同的等位基因 A_1 和 A_2 存在于个体之间，于是将有三种可能的基因型 A_1A_1 、 A_1A_2 和 A_2A_2 （这里我们只关注二倍体的生物体，全书均如此）。群体的遗传组成可通过属于每一种基因型的个体比例或百分率，或换句话说，通过个体之间三种基因型的频率进行充分的描述。这些比例或频率被称为基因型频率，一种特定基因型的频率就是它在个体之间的比例或百分率。例如，如果我们发现群体中四分之一的个体是 A_1A_1 型的，则这种基因型的频率就是 0.25 或 25%。当然，所有基因型的频率加在一起一定等于 1 或 100%。

例 1.1

人类 M-N 血型是由一个位点上两个等位基因决定的，三种基因型对应三种血型 M、MN 和 N。以下的数据取自 Mourant (1954) 的资料 (表 1-1)，显示了东格陵兰岛爱斯基摩人和冰岛人的血型分布频率。

表 1-1 东格陵兰岛爱斯基摩人和冰岛人的血型分布频率

		血 型			个体数目
		M	MN	N	
频率 (%)	格陵兰人	83.5	15.6	0.9	569
	冰岛人	31.2	51.5	17.3	747

很明显，这两个群体在这些基因型频率上不同，N 血型在格陵兰人中稀少而在冰岛人中相对普遍。这个位点不仅是两个群体各自变异的一种来源，而且也是群体间遗传差异的一种来源。

在遗传学意义上，一个群体不仅仅是一群个体，而且是一个能繁殖的类群；一个群体的遗传学不仅涉及个体的遗传组成，而且也涉及从一个世代到下一世代的基因传递。在传递过程中，亲本的基因型分开了，后裔中由配子传递的基因组成了一组新的基因型。这样，群体携带的基因在世代间具有连续性，但它们构成的基因型却没有这种连续性。一个群体的遗传组成，以它携带的基因为依据，可通过基因频率的阵列来描述；也就是说，通过在每一个位点上存在的等位基因以及在每一个位点上不同等位基因的数目或比例加以详细说明。例如，如果 A_1 是位点 A 上的一个等位基因，那么 A_1 基因的频率，或 A_1 的基因频率是这个位点上所有基因中 A_1 等位基因所占的比例或百分率。任何一个位点上所有等

位基因的频率加起来一定等于 1 或 100%。

一组个体之间某特定位点上的基因频率可根据已知的基因型频率来确定。举一个假想的例子，假定有两个等位基因 A_1 和 A_2 ，我们对 100 个个体进行分类并计数每一种基因型的数目（表 1-2）。

表 1-2 根据基因型频率确定基因频率

		A_1A_1	A_1A_2	A_2A_2	总数
个体数		30	60	10	100
基因数	$\left\{ \begin{array}{l} A_1 \\ A_2 \end{array} \right.$	60	60	0	120
		0	60	20	80

每一个体包含两个基因，这样我们计数了这个位点上的 200 个基因的代表。每一个 A_1A_1 个体包含两个 A_1 基因，每一个 A_1A_2 个体包含一个 A_1 基因。因此，样本中有 120 个 A_1 基因和 80 个 A_2 基因。进而 A_1 的频率是 60%或 0.6， A_2 的频率是 40%或 0.4。为了用一种更一般的形式表示这种关系，可以设定基因和基因型的频率（表 1-3）。

表 1-3 基因频率和基因型频率的关系

	基因		基因型		
	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
频率	p	q	P	H	Q

这样， $p + q = 1$ 以及 $P + H + Q = 1$ 。因为每一个体包含两个基因， A_1 基因的频率为 $\frac{1}{2}(2P + H)$ ，而计数过的个体间基因频率和基因型频率之间的关系如下：

$$\left. \begin{array}{l} p = P + \frac{1}{2}H \\ q = Q + \frac{1}{2}H \end{array} \right\} \quad [1.1]$$

例 1.2

为了说明从基因型频率中计算基因频率，我们可以取例 1.1 给出的 M-N 血型频率。M 和 N 血型表示两个纯合基因型，MN 血型表示杂合基因型。根据等式 [1.1]，格陵兰人中 M 基因的频率是 $0.835 + \frac{1}{2}(0.156) = 0.913$ ，N 基因的频率是 $0.009 + \frac{1}{2}(0.156) = 0.087$ ，频率总和等于 1.000。

对冰岛人样本进行同样的运算，我们发现两个群体中有如表 1-4 所示的基因频率（以百分率表示）。这样，两个群体在基因频率和基因型频率上都不同。

表 1-4 格陵兰人和冰岛人血型基因的频率

	基因频率 (%)	
	M	N
格陵兰人	91.3	8.7
冰岛人	57.0	43.0

自然群体中的孟德尔变异

我们对自然群体中作为孟德尔单位分离的不连续性状能够观察到很多不同水平的遗传变异。对表型（例如植物花色、蜗牛壳色和图案）具有巨大影响的可见变异或者像侏儒这样的巨大突变都在变异系谱的一端。但是，个体之间的很多变异在表型上并不产生明显的差异。这种隐蔽的变异已被研究蛋白质以及 DNA 本身差异的方法所揭示。例 1.1 和例 1.2 中的 MN 血型变异例释了一种通过抗体反应可检测出来的隐蔽变异。在电场存在的情况下，蛋白质电泳是检测可溶性蛋白质在凝胶上流动性差异的一种技术。这种方式检测出的变异是作为共显性等位基因遗传的，被称为等位酶。它们是由引起蛋白质电荷改变的氨基酸替代所造成的。因此，蛋白质电泳可检测蛋白质之间大约 25% 的氨基酸差异，因为 20 种氨基酸有 5 种是带电荷的。

采用限制性酶可检测 DNA 序列的变异。这些酶识别特异性的 4 或 6 个碱基的 DNA 序列，这些序列无论何时出现，这些酶都可以切割它。然后，这些 DNA 片段通过电泳按大小被分开，并通过称为 Southern blotting 的程序与一种标记的 DNA 探针杂交而被观察。在探针可识别的 DNA 区间内，如果个体之间在限制性位点上存在变异，那么这种变异会被斑点内限制性片段的大小变化所揭示，这被称为限制性片段长度多态性，或 RFLPs。与只能检测功能蛋白质变化的蛋白质电泳不同，RFLP 变异可以位于基因组的非编码和编码区域。虽然克隆 DNA 对检测这种变异是必要的，但不必知道克隆探针的功能或染色体位置。限制性片段长度的变异还可由两个限制性位点之间的 DNA 序列的插入和缺失引起。大的插入通常是转座（位）因子：它的 DNA 序列在基因组中是以多个、分散的拷贝存在的，能够从一个位置移动到另一个位置。另外的长度变异是由“小卫星”或“微卫星”位点上串联重复 DNA 序列的数目变化所引起。前者，也称为 VNTR（可变量目的串联重复）位点，由 10~60 个碱基对长的重复单位组成。微卫星（或简单序列重复，SSR）位点由 1~6 个碱基对这样较短的重复单位组成，例如， $(CA)_n$ 或 $(AGC)_n$ ，其中 n 是重复单位数目，是可变的。最后，通过比较由直接测序所得的实际 DNA 序列来最终分辨个体间的变异。

对不连续性状的等位变异，不管是表型上可见的还是隐蔽的，都被称为多态性，关于这些，将在第 2 章和第 4 章进行更多的说明。多态位点导致数量性状的变异，这是本书的主题。

变化的原因

在基因从一个世代传递到下一世代过程中，几种作用影响基因和基因型频率。为了全面理解数量遗传变异，我们需要了解这些因素如何单独和共同影响群体随时间的遗传变异，以及作为基因频率改变的作用，它们的相对重要性有多大。这些作用形成了下面四章的主要内容，在此我们对这些作用做简要评述，以便说明本章对什么因素不予考虑的想法。可引起群体遗传特性变化的作用是：

群体大小 从一个世代传递到下一世代的基因是亲代基因的一个样本。因此，连续世代间的基因频率存在着抽样变异，而且亲本数量越小，抽样变异就越大。抽样变异的效应于第 3~5 章中给予考虑，但总是假定我们在处理一个“大群体”，这只意味着这个群体中的抽样变异小到可以忽略不计，因此，届时讨论中不涉及它。为实用目的，一个“大群

体”是成年个体数以百计而不是数以十计的一个群体。

生殖力和生活力的差异 虽然现在我们不涉及所讨论的基因表型效应，但我们不能忽视它们对生殖力和生活力的影响，因为这些效应影响下一代的遗传组成。亲本之间不同的基因型可能有不同的生殖力，如果是这样，则它们对形成下一世代用的配子贡献也不会相等。通过这种方式，基因频率在传递过程中可能会改变。而且，新形成的合子之间的基因型可能有不同的存活率，因此到个体成熟且它们本身成为亲本的时候，新的世代中基因频率也可能会改变。这些过程被称为选择，将在第 2 章描述。其时我们将假定这些过程不起作用。我们可以取人类血型基因来说明这一目的，因为对这些基因起作用的选择力量可能不强烈。相反地，产生一种与野生型相比是异常突变表型的基因，经常承受着更多的选择。

迁移和突变 群体的基因频率还可能因另一个群体中的个体迁入以及基因突变而改变。这些过程将在第 2 章描述，在这里也假定它们不起作用。

交配系统 后裔的基因型是由配子成对结合所形成的合子决定的，配子的结合又受到亲本交配的影响。因此，后裔世代中基因型频率受亲本世代中交配配偶的基因型的影响。我们将首先假定，就讨论的基因型而言，交配是随机的。随机交配（random mating 或 panmixia）意味着任一个体与群体中的任何其他个体都有相等的交配机会。重要之点在于交配的个体不应该有基因型的相似、或通过祖先而彼此有相关的特殊倾向。如果一个群体占有广大的地理区域，那么居住在同一地区的个体比居住在不同地区的个体似乎有更多的交配，因此，交配的配偶倾向于通过祖先而相关。为此一个分布很广的群体很可能被细分成很多地区群，并且仅在地区群内交配是随机的。细分群体的特性依赖于地区群的大小，这将在第 3 ~ 5 章群体大小的效应中再来描述。

哈迪—温伯格平衡

哈迪—温伯格法则

在一个没有选择、突变和迁移的大随机交配群体中，基因频率和基因型频率在世代间保持不变；而且，基因频率和基因型频率之间存在一种简单的关系。一个群体的这些特性是从众所周知的哈迪—温伯格法则这一定理（或原理）导出的，哈迪和温伯格各自于 1908 年证明了这些特性。一个具有稳定的基因和基因型频率的群体被认为处于哈迪—温伯格平衡。基因频率和基因型频率之间的关系是最重要的，因为有关群体遗传学和数量遗传学的很多推论都基于它。关系是这样的：如果亲本之间两个等位基因的基因频率是 p 和 q ，那么后裔之间的基因型频率是 p^2 、 $2pq$ 和 q^2 ，因此有：

亲本中基因		后裔中基因型			
A_1	A_2	$A_1 A_1$	$A_1 A_2$	$A_2 A_2$	
频率	p	q	p^2	$2pq$	q^2

[1.2]

（上面的关系指的是常染色体基因；性连锁基因并不这么简单，将在以后解释。）保持哈迪—温伯格法则所需要的随机交配和没有选择的条件仅仅指的是所考察的基因型。关于其他属性可能有优先交配，其他位点的基因型可能受到选择，但不影响正在讨论的问题。

两个附加的条件是基因在配子发生过程中正常分离以及雄性和雌性中基因频率相同。这些要求的理由将在证明过程中看出来。

哈迪-温伯格法则的证明包含 4 个步骤，它们汇总在表 1-5 中，同时也给出了确保每一步推导有效所必须保持的条件。4 个步骤详述如下。

表 1-5 在证明哈迪-温伯格法则过程中推导的步骤以及必须保持的条件

步骤	推导（起：止）	条 件
1a	亲本基因频率	(1) 正常的基因分离
	所有配子的基因频率	(2) 亲本具有相等的生殖力
1b	形成合子的配子的基因频率	(3) 配子有相等的受精能力
		(4) 大群体
2	合子的基因型频率	(5) 随机交配
		(6) 雄亲和雌亲有相等的基因频率
3	后裔中基因型频率	(7) 相等的存活力
4	后裔中基因频率	

1. 从亲本基因频率到配子基因频率 假设亲本世代具有如下的基因和基因型频率 (表 1-6):

表 1-6 亲本世代的基因和基因型频率

	基 因		基因型		
	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
频率	p	q	P	H	Q

产生两种类型的配子，一些配子携带 A_1 ，另一些配子携带 A_2 。 A_1A_1 个体只产生 A_1 配子。正常分离的 A_1A_2 个体产生同等数目的 A_1 和 A_2 配子。于是，只要所有基因型都具有相等的生殖力，整个群体产生的所有配子之间， A_1 的频率就是 $P + \frac{1}{2}H$ ，由等式 [1.1] 可知，它是产生配子的亲本中 A_1 基因的频率。这样整个配子产出中基因频率与亲本的相同。这是表 1-5 中的步骤 1a。只有一些配子形成将成为下一世代中个体的合子。只要携带不同等位基因的配子在其受精能力上没有差异，并且只要形成的合子代表亲本配子的一个大样本，合子中基因频率就不会改变。这是步骤 1b。

2. 从配子基因频率到合子的基因型频率 个体间随机交配等于它们的配子间随机结合。于是合子（受精卵）之间的基因型频率就是结合起来以产生它们的配子类型的频率之积。因此，随机交配产生的后裔之间的基因型频率可简单地通过将每一性别亲本产生的配子类型频率相乘来确定。只要两性中配子频率相同，产生的合子就如表 1-8 所示。

表 1-7 合子的基因型频率

	基因型		
	A_1A_1	A_1A_2	A_2A_2
频率	p^2	$2pq$	q^2

A₁ 卵子和 A₂ 精子的结合不必与 A₂ 卵子和 A₁ 精子的结合加以区别；因此得到如表 1-7 所示的合子基因型频率。

表 1-8 由两性配子得到的合子基因型频率

		雌性配子及其频率	
		A ₁ <i>p</i>	A ₂ <i>q</i>
雄性配子 及其频率	A ₁ <i>p</i>	A ₁ A ₁ <i>p</i> ²	A ₁ A ₂ <i>pq</i>
	A ₂ <i>q</i>	A ₁ A ₂ <i>pq</i>	A ₂ A ₂ <i>q</i> ²

3. 从合子到成年 正如 [1.2] 表述的那样，上面推导的合子基因型频率是哈迪—温伯格频率。可是，这不是证明的彻底结束，因为除非合子有相等的存活力，至少直到能够区分它们的基因型时，才能观察到这些频率。这个步骤可能是不重要的，但是如果要了解不同的存活力影响，就必须认识这一步骤。

4. 后裔中从基因型频率到基因频率 这个最后步骤证明基因频率没有改变。只要后裔中不同基因型相等地存活到成年，此时它们能成为亲本，它们的频率就是上面给出的那样。于是成熟后裔中基因频率可通过等式 [1.1] 找到。A₁ 的频率是 $p^2 + (2pq) / 2 = p(p + q) = p$ ，它与亲代的相同。这证明基因频率在世代间保持不变。

现在可以陈述哈迪—温伯格法则两个深层次的方面。首先，因为亲本和后裔中基因频率相同，[1.2] 中的基因频率和基因型频率之间的关系可适用于一单个世代。其次，后裔的基因型频率仅取决于亲本的基因频率而与亲本的基因型频率无关。这一点可以从上面步骤 1 中看出，其中配子类型的频率与亲本的基因频率相等，而不管基因型频率是多少。因而，具有任何基因型频率的亲本，只要它们随机交配并且雄亲和雌亲中的基因频率相同，就以哈迪—温伯格比例产生后裔。如果已知雄亲和雌亲中基因频率不相同，通过在步骤 2 中将适宜的配子频率代入表 1-8，可很容易地推导出后裔的基因型频率。一个常染色体基因的基因频率在后裔的两个性别中是相等的，并且随机交配的下一代产生与原始基因频率平均值对应的哈迪—温伯格基因型频率。

处于哈迪—温伯格平衡的一个群体中基因频率和基因型频率之间的关系显示在图 1-1 中。基因型频率的曲线显示了两个重要的特征。首先，杂合体的频率不可能超过 50%，当基因频率为 $p = q = 0.5$ 时，这个最大值才出现。其次，当一个等位基因的频率低时，这个稀有的等位基因主要出现于杂合体中，纯合体几乎没有。这对选择的有效性具有重要的影响，这一点将在下一章看到。

哈迪—温伯格法则的应用

哈迪—温伯格法则在三个方面特别有用，现在来说明它们。

隐性等位基因的基因频率 本章开始我们在等式 [1.1] 中看到，一群个体之间的基因频率可如何从它们的基因型频率中来确定；但对此知道所有 3 种基因型的频率是必要的。继而，当杂合体不能与显性纯合体区分时，等式 [1.1] 的关系不适用于隐性等位基

因的情况。可是，如果基因型符合哈迪—温伯格的比例，我们就不必知道所有 3 种基因型的频率。例如，假定用 a 表示频率为 q 的一个隐性基因；那么 aa 纯合体的频率就是 q^2 ，并且基因频率是纯合体频率的平方根。例 1.3 说明了这种计算。要确保这种估计基因频率的方法有效，在计数纯合体之前没有选择淘汰显然是必需的。还应当注意到采用这种方法对基因频率进行估计时，对非随机交配的影响是相当敏感的。

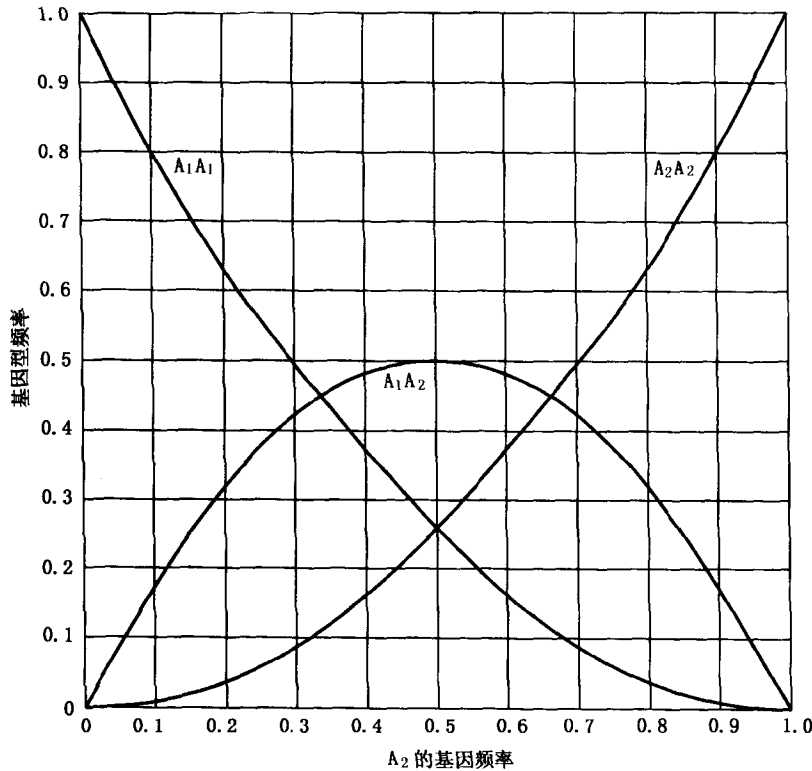


图 1-1 处于哈迪—温伯格平衡的一个群体中两个等位基因的基因型频率和基因频率之间的关系

例 1.3

苯丙酮尿症 (PKU) 是由单个隐性基因引起的一种人类代谢病。纯合体可在出生后几天检测出来，在那以前假定选择淘汰是可以忽略不计的。对英国伯明翰出生的婴儿经过 3 年的测验，从 55 715 个婴儿中检测出 5 例 (Raine 等, 1972)。样本中纯合体的频率是 90×10^{-6} 或约为 $1/11\ 000$ 。纯合体的哈迪—温伯格频率是 q^2 ，因此基因频率是 $q = \sqrt{90 \times 10^{-6}} = 9.5 \times 10^{-3} = 0.0095$ 。

整个群体中杂合体的频率是 $2q(1 - q)$ ，正常个体内杂合体的频率是 $2q/(1 + q)$ 。两者计算出来的答案近似为 0.019。因此，大约正常人的 2% 或每 50 人中有 1 人是 PKU 基因携带者。大多数人在发现一个稀有的隐性畸形体的杂合体竟如此的普遍时，往往会感到吃惊。作为从图 1-1 中得出的一个结论，已经注意到了这一点。

“携带者”的频率 了解隐性畸形体杂合体或“携带者”的频率经常是重要的，如果

基因频率已知就可以计算它。假定存在哈迪-温伯格平衡，则所有个体（包括纯合体）中的杂合体频率由 $2q(1-q)$ 给出。可是，往往了解正常个体内杂合体的频率是更妥当的，虽然这在纯合体稀少时不会有很大的不同。正常个体中杂合体的频率，用符号 H' 表示，是基因型频率 $Aa/(AA + Aa)$ 的比率，其中 a 是隐性等位基因。因此，当 a 的频率是 q 时，有：

$$H' = \frac{2q(1-q)}{(1-q)^2 + 2q(1-q)} = \frac{2q}{1+q} \quad [1.3]$$

哈迪-温伯格平衡的测验 如果能识别一个位点上的所有基因型，其所得数据可用，那么可以检验基因型的观察频率是否与处于哈迪-温伯格平衡的群体相一致。根据哈迪-温伯格法则，后裔的基因型频率由它们亲本的基因频率决定。如果群体处于平衡，则亲本与后裔的基因频率相同，因此，后裔中观察到的基因频率，就像是亲本的基因频率，能用来计算依哈迪-温伯格法则期望的基因型频率。这种方法例证在例 1.4 中。

例 1.4

例 1.1 给出了冰岛人 M-N 血型频率。样本的观察数在下表中。样本的基因频率首先通过等式 [1.1] 从观察数中计算。然后，通过等式 [1.2] 从基因频率中计算出哈迪-温伯格基因型频率 p^2 、 $2pq$ 和 q^2 ，每一个与总数相乘得到期望值。例如，MM 的期望值是 $(0.5696)^2 \times 747$ 。比较观察数和期望值表明，两种纯合体不足，杂合体过量。用 χ^2 测验观察数和期望值一致程度的好与坏。不一致性不显著，这在抽样过程中很容易偶然产生。注意这个 χ^2 只有一个自由度，因为从数据中已估计了基因频率，这样观察数和期望值在它们的基因频率和总数上必定一致（表 1-9）。

表 1-9 冰岛人 M-N 血型频率的 χ^2 测验

	基因型				基因频率	
	MM	MN	NN	总数	M	N
观察数	233	385	129	747	0.5696	0.4304
期望值	242.36	366.26	138.38	747		
$\chi^2 = 1.96$	$P \sim 0.2$					

对符合平衡群体的测验就是对哈迪-温伯格基因型频率产生的条件是否满足的测验。可是，从测验中能获得的结论却有限。当发现良好符合时，测验没有理由怀疑所有条件的满足。对血型基因作出的测验几乎总是显示相当良好的符合，正如例 1.4 所示。但是有一个条件，它的不满足并不导致不符合，这就是亲本之间相等的生殖力。其理由将马上给予解释。如果测验显示出观察和期望频率之间不一致，那么我们能得出结论：一个或更多的条件没有满足。但是不一致性不允许我们鉴别出其来源，或确定哪一个条件没满足。其理由是同样的不一致性可由不同的原因引起。例如，杂合体过多可由纯合体的选择淘汰引起，也可由亲本世代中雄性和雌性的基因频率不同引起。这种测验并不像它表现的那么简单，我们必须更仔细地观察这种测验能够做什么。

哈迪-温伯格法则将亲本的基因和后裔的基因型关联起来。因此，要全面测验它，我们需要知道亲本的基因频率并从亲本基因频率中计算出后裔的期望基因型频率。除描述的测验外，我们只有后裔。我们通过计数找出它们的基因频率。于是我们认为：如果这就是产生这些后裔的配子之间的基因频率，那么基因型应当像从观察的基因频率中计算出的那样处于哈迪-温伯格比例。如果亲本的基因频率和后裔的不同，那么我们已经使用了错误的基因频率来计算期望值。参考表 1-5 将指明测验的条件是随机交配，两性亲本具有相等的基因频率和后裔之间具有相同的生活力；但没有测验亲本之间相等的生殖力。因此，选择可通过生殖力起作用，并且不会被这种测验检出。通过对后裔生活力起作用的选择将导致观察和期望频率之间的不符合。可是，鉴定出降低生活力的基因型是不可能的。在论及选择的效应之后，下一章将解释其理由。关于测验局限性的更全面的讨论见 Wallace (1958, 1968), Prout (1965)；其统计方面的更全面考察见 Smith (1970)。

交配频率和哈迪-温伯格法则的另一种证明

现在让我们更仔细地观察一个随机交配群体的繁殖结构,按照配偶的基因型来区别交配类型，以及观察不同交配类型的后裔之间的基因型频率是多少。这将为叙述连续世代中的基因型频率提供一个一般的方法，这将在下一章中用到。它还为哈迪-温伯格法则提供了另一种证明；这种证明比已经给出的更麻烦，但更清晰地显示了哈迪-温伯格频率是如何从孟德尔分离法则中产生的。这个过程首先是要根据亲本之间基因型的频率获得所有可能交配类型的频率，然后要根据孟德尔比率获得每一种交配类型的后裔之间的基因型频率。

考察具有两个等位基因的一个位点，并且像以前一样，假定亲本中的基因和基因型频率如表 1-10 所示：

表 1-10 亲本的基因和基因型频率					
频率	基 因		基因型		
	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
	p	q	P	H	Q

总共有 9 种交配类型，当交配是随机时，它们的频率可通过表 1-11 下角显示的频率相乘而获得。因为亲本的性别对此无关，一些交配类型是等价的，所以不同类型的数目简化成 6 种。通过对等价类型频率的求和，我们获得表 1-12 前两栏中所示的交配类型的频率。现在我们必须考察由每一种交配类型产生的后裔的基因型，并找出在整个后裔中每一种基因型的频率，当然得假定所有的交配类型具有相等的生殖力以及所有的基因型具有相等的生活力。这列在表 1-12 的右边栏中。这样，例如交配类型 $A_1A_1 \times A_1A_1$ 仅产生 A_1A_1 后裔。因此，整个后裔中， P^2 部分是从这种交配类型中获得的 A_1A_1 基因型。类似地， $A_1A_2 \times A_1A_2$ 交配的后裔中有 $\frac{1}{4}$ 是 A_1A_1 。因此，这种频率为 H^2 的交配类型贡献给总的 A_1A_1 后裔部分是 $\frac{1}{4}H^2$ 。为了求出整个后裔中每一种基因型的频率，我们将每种交配类型贡献的频率相

加。简化之后的总和列在表 1-12 的底部，从等式 [1.1] 给出的恒等式可知，它们等于 p^2 、 $2pq$ 和 q^2 。这些就是哈迪—温伯格平衡频率，并且我们已经证明它们可以通过一代随机交配而达到，不管亲本之间的基因型频率如何。

表 1-11 两个等位基因随机交配时各种基因型的频率

			雌性亲本的基因型及其频率		
			A_1A_1	A_1A_2	A_2A_2
			P	H	Q
雄性亲本	A_1A_1	P	P^2	PH	PQ
的基因型	A_1A_2	H	PH	H^2	HQ
及其频率	A_2A_2	Q	PQ	HQ	Q^2

表 1-12 两个等位基因随机交配基因型频率的计算

交 配		后裔的基因型及其频率		
类型	频率	A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	P^2	P^2	—	—
$A_1A_1 \times A_1A_2$	$2PH$	PH	PH	—
$A_1A_1 \times A_2A_2$	$2PQ$	—	$2PQ$	—
$A_1A_2 \times A_1A_2$	H^2	$\frac{1}{4}H^2$	$\frac{1}{2}H^2$	$\frac{1}{4}H^2$
$A_1A_2 \times A_2A_2$	$2HQ$	—	HQ	HQ
$A_2A_2 \times A_2A_2$	Q^2	—	—	Q^2
	总和	$\left(P + \frac{1}{2}H\right)^2$	$2\left(P + \frac{1}{2}H\right)\left(Q + \frac{1}{2}H\right)$	$\left(Q + \frac{1}{2}H\right)^2$
	=	P^2	$2pq$	q^2

复等位基因

当一个位点上有两个以上等位基因时，由基因频率确定基因型频率的方法，与两个等位基因的完全相同。如果 A_1 和 A_2 是任意两个等位基因，并且它们具有频率 q_1 和 q_2 ，那么处于哈迪—温伯格平衡之下的基因型频率如表 1-13 所示：

表 1-13 两个等位基因处于哈迪—温伯格平衡时的基因型频率

	基因型		
	A_1A_1	A_1A_2	A_2A_2
频率	q_1^2	$2q_1q_2$	q_2^2

这些频率也可通过一代随机交配而达到。

如果所有等位基因都是共显性的，这样所有基因型可在表型上识别，那么通过简单的计数便可从表型中估计基因频率。可是，如果一个或多个等位基因是隐性的，那么通过计数等位基因就不能获得基因频率。例如，在人类 ABO 血型系统中，O 等位基因对 A 和 B 是隐性的，因此 A 血型（表型）由两种基因型 AA 和 AO 组成，B 血型由 BB 和 BO 组成。

基因频率必须通过最大似然法来估计，正如 Weir (1990) 所描述的事例那样。

性连锁基因

对于性连锁基因，其情形比常染色体基因要复杂得多。同配性别中基因频率和基因型频率之间的关系与一个常染色体基因相同，但异配性别只有两种基因型，且每一个体仅携带一个基因而不是两个。由于这个原因，群体中性连锁基因的三分之二由同配性别携带，三分之一由异配性别携带。为简明起见，异配性别将被指定为雄性。考察两个等位基因 A_1 和 A_2 ，频率为 p 和 q ，假定基因型频率如表 1-14 所示：

表 1-14 性连锁的两个等位基因和基因型频率

	雌 性			雄 性	
	A_1A_1	A_1A_2	A_2A_2	A_1	A_2
频率	P	H	Q	R	S

于是雌性 A_1 的频率是 $p_f = P + \frac{1}{2}H$ ，雄性 A_1 的频率是 $p_m = R$ 。整个群体中 A_1 的频率是：

$$\left. \begin{aligned} \bar{p} &= \frac{2}{3}p_f + \frac{1}{3}p_m \\ &= \frac{1}{3}(2p_f + p_m) \\ &= \frac{1}{3}(2P + H + R) \end{aligned} \right\} \quad [1.4]$$

现在，如果雄性和雌性之间的基因频率不同，则群体处于不平衡。整个群体中基因频率不会发生变化，但随着群体接近平衡，基因频率在两个性别之间的分布会来回摆动。其理由可从下面的考察中看出来。雄性只从它们的母亲那儿获得其性连锁基因；因此， p_m 等于上一代的 p_f 。雌性从双亲那儿相等地获得其性连锁基因；因此， p_f 等于上一代的 p_m 和 p_f 的平均值。使用符号“'”表示后裔世代，有：

$$\begin{aligned} p'_m &= p_f \\ p'_f &= \frac{1}{2}(p_m + p_f) \end{aligned}$$

两个性别中频率的差异为：

$$\begin{aligned} p'_f - p'_m &= \frac{1}{2}(p_m + p_f) - p_f \\ &= -\frac{1}{2}(p_f - p_m) \end{aligned}$$

即为上一代差异的一半，但处于另一个方向。因此，两个性别之间基因的分布来回摆动，但在连续世代中差异减半，群体很快接近一个平衡，在平衡处两个性别的频率是相等的。当群体由一种类型的雌性（全为 A_1A_1 ）和另一种类型的雄性（全为 A_2 ）混合在一起而开始，并让它们随机繁殖时，图 1-2 说明了基因频率接近 $2/3$ 时的平衡情况。

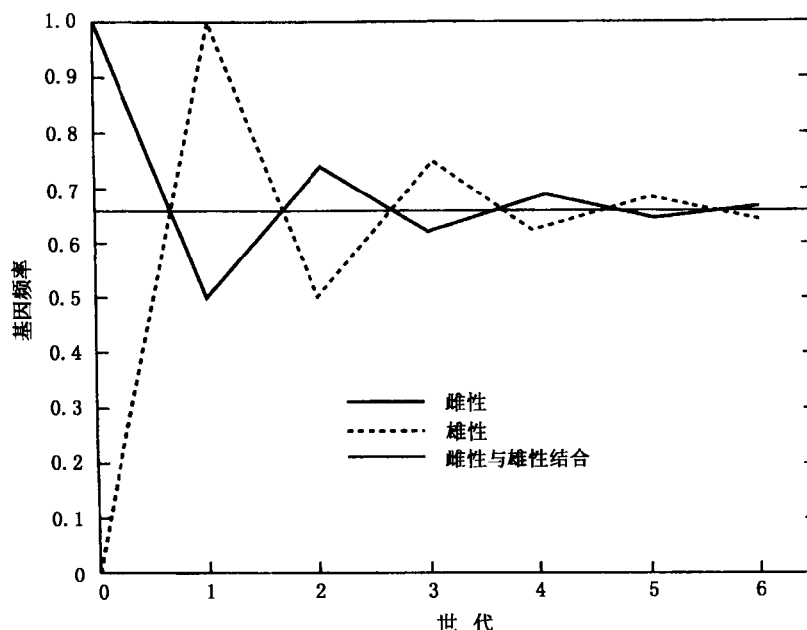


图 1-2 关于一性连锁基因在随机交配下接近平衡的情况

显示了雌性内、雄性内以及两性结合时的基因频率。这个群体从全是一种类型的雌性 ($q_f = 1$) 和全是另一种类型的雄性 ($q_m = 0$) 而开始的。

例 1.5

Searle (1949) 给出了伦敦猫一个样本中许多基因的频率。检查过的猫都是被送到诊所来杀死的；因此它们不一定是一个随机样本。在所研究的基因中间有以前称为“黄色”但现在叫“橙色”(O)的性连锁基因。雌性中所有 3 种基因型都可识别，杂合体是“玳瑁壳”或“杂色”猫 (表 1-15)。数据用来对哈迪-温伯格期望值进行了测验，尤其要看是否有非随机交配的任何证据。第一个测验是要看两个性别中的基因频率是否相同。然后用与例 1.4 相同的方式对雌性基因型测验哈迪-温伯格法则。每一表型类别的数目以及从它们中计算出来的基因频率显示在下表中。雄性的基因频率稍高一点，但不显著。因此，到目前为止，没有理由认为群体不处于平衡。为简化起见，将对雌性观察到的基因频率视为用于计算雌性期望基因型频率的适宜基因频率。用与例 1.4 相同方式计算出的期望值给在下表中。观察到的数目不与期望值良好吻合，但与小期望值的不一致仅是值得怀疑的显著。如果不一致是真实的，则它可归因于非随机交配，但它也可能归因于人们对颜色的偏爱，这使得抽样有偏差，从而使样本不代表繁殖群体。对猫群体的更广泛的分析和讨论，见 Metcalfe 和 Turner (1971)。

表 1-15 应用伦敦猫的资料进行的哈迪—温伯格平衡 χ^2 测验

	个体数						
	雌 性				雄 性		
	+ +	+ 0	00	总和	+	0	总和
观察数	277	54	7	338	311	42	353
期望值	273.4	61.2	3.4	338			
	$\chi^2_{[1]} = 4.6; \quad P = 0.04$						
	基因数				O 基因频率		
	+	0	总和		<i>q</i>		
雌性	608	68	676		0.101		
雄性	311	42	353		0.119		

多于一个位点

当分别考察所有常染色体位点时，经一代随机交配后基因型频率就达到平衡是真实的。但联合考察两个或更多个位点的基因型时，这不是真实的。为了说明这一点，假定有两个群体，一个完全由 $A_1A_1B_1B_1$ 基因型组成，另一个完全由 $A_2A_2B_2B_2$ 基因型组成。假定这两个群体以每一性别相等的数目进行混合，并允许随机交配。对两个位点的每一个位点上的两个等位基因来说，有 9 种可能的基因型，但只有 3 种将出现在第一代后裔中，它们是两个起始的双纯合体和双杂合体。由两个位点决定的性状之间存在完全的联合，两个性状似乎是由单个基因的差异所决定。随着连续的随机交配，丢失的基因型将在随后的世代中出现，但不会立即达到它们的平衡频率，并且性状之间起始的联合会逐渐减少。如果两个位点是连锁的，那么平衡频率的获得将花费更长时间，因为丢失的基因型出现取决于两个位点之间的重组。关于两个或更多位点的不平衡被称为配子期不平衡，或连锁不平衡，不管位点连锁与否。具有不同基因频率的群体混合或小群体内的偶然性可产生不平衡。选择较其他等位基因组合更为优势的等位基因组合也可产生并维持不平衡。表 1-16 是一个随机繁殖群体逼近平衡的速度的推导。

表 1-16 随机繁殖群体逼近哈迪—温伯格平衡速度的推导

基因	A_1	A_2	B_1	B_2
基因频率	p_A	q_A	p_B	q_B
配子类型	A_1B_1	A_1B_2	A_2B_1	A_2B_2
平衡时频率	p_Ap_B	p_Aq_B	q_Ap_B	q_Aq_B
实际频率	r	s	t	u
与平衡的差异	$+D$	$-D$	$-D$	$+D$

我们首先需要对不平衡数量的度量。这最好用配子类型的频率而不是合子基因型的频率来表示。考察两个位点，每一位点具有两个等位基因，基因频率显示在表 1-16 中。于是有 4 种类型配子。如果配子含有基因的随机组合，则群体处于平衡。因此，平衡时的配子频率只取决于基因频率，正如表 1-16 中显示的那样。假定实际的非平衡频率为 r 、 s 、 t 和 u 正如从表中看到的那样。这些中的每一个与平衡频率相差一个数量 D ，两个配子

类型有正的离差，两个有负的离差。每一配子类型的 D 值必定是相同的，但符号除外。用 D 来度量不平衡的数量。通过比较相引和相斥双杂合体的频率、参照基因型可表示不平衡。不管两个位点连锁与否，基因型 A_1B_1/A_2B_2 可称为相引杂合体。它的频率是 $2ru$ 。相斥杂合体是 A_1B_2/A_2B_1 且它的频率是 $2st$ 。如果群体处于平衡，这两种基因型有相等的频率。与 D 的关系是：

$$D = ru - st$$

因此， D 等于相引和相斥杂合体间频率差异的一半。

例 1.6

黑腹果蝇 (*Drosophila melanogaster*) 的整个染色体可从自然群体的个体中抽取出来，并且可与含有经显性标记的平衡染色体的种蝇杂交，在这种染色体的抑制重组作用下，形成完全的纯合。一条染色体上任何位点数目的基因型被称为单倍型 (haplotype)，因为染色体基本上就是一种可能的单倍体配子类型。总数为 47 的第 2 号染色体从位于北卡罗来纳 Raleigh 的一个自然群体中抽取出来，用来评定 *scabrous* 基因区域内的限制性片段长度变异 (Lai 等, 1994)。下面的数据是两个 Pst I 限制位点每一个单倍型的染色体数目 (表 1-17)。 Pst I (-12.0) (位点 A) 定位在离 *scabrous* 基因转录起点的 5' 端 12kb 处， Pst I (5.8) (位点 B) 则在离转录起点的 3' 端 5.8kb 处。“+”表示位点存在，“-”表示位点不存在。我们想要：(i) 确定这两个限制性位点是否处于连锁 (不) 平衡，(ii) 计算连锁不平衡系数 D 。

表 1-17 两个 Pst I 限制位点每一单倍型的染色体数目

单倍型		观察数目	观察频率	期望频率	期望数目
A	B				
+	+	4	0.085	$p_A p_B = 0.06$	2.8
+	-	4	0.085	$p_A q_B = 0.11$	5.2
-	+	13	0.277	$q_A p_B = 0.30$	14.1
-	-	26	0.553	$q_A q_B = 0.53$	24.9

(i) 采用表 1-16 的术语，假定 p 表示每一个位点上 + 等位基因的频率， q 代表 - 等位基因的频率。在两个限制性位点座位上基因频率的估计值是 $p_A = 8/47 = 0.17$ ； $q_A = 39/47 = 0.83$ ； $p_B = 17/47 = 0.36$ ； $q_B = 30/47 = 0.64$ 。于是，期望的单倍型频率和数目就像表 1-17 的右边显示的那样。

期望数目与观察数目相配得很好，差异在统计上是不显著的 ($\chi^2_{[1]} = 0.93$ ； $P > 0.5$)。因此，对这些限制性位点，样本没有显示出群体处于连锁不平衡的证据。这些位点被 17.8kb 分开，这说明了非常紧密连锁的位点不一定处于连锁不平衡。

(ii) 上面给出了观察的单倍型频率，据此 $D = ru - st = (0.085) \times (0.553) - (0.085) \times (0.277) = 0.023$ 。 D 的绝对值取决于基因频率。 D 的绝对范围从 -0.25 到 0.25；如果 $p_A = q_A = p_B = q_B = 0.5$ ，或 $r = u = 0$ 以及 $s = t = 0.5$ ($D = -0.25$)，或 $s = t = 0$ 和 $r = u = 0.5$ ($D = 0.25$)，这些值就出现。如果基因频率不是精确的相等，那么 D 的绝对值将在这个范围之

内。因此, D 经常用 D/D_{MAX} 表示, 其中, D_{MAX} 是给定的观察基因频率下 D 所能取得的最大数值。表 1-16 中, $s = p_A q_B - D \geq 0$ 和 $t = q_A p_B - D \geq 0$, 所以, D 一定 $\leq p_A q_B$ 且 $\leq q_A p_B$ 。因此, D_{MAX} 是 $p_A q_B$ 或 $q_A p_B$ 中较小的一个。对这个例子, D_{MAX} 是 0.11, 因此 $D/D_{\text{MAX}} = 0.21$ 。 D 的计算值是它最大值的 21%, 它是重要的。大的样本大小对检测连锁不平衡是必要的, 特别是当基因频率处于极端值时。

当处于连锁不平衡的群体随机交配时, 不平衡的数量会随着每一个后续的世代而逐渐降低。这种情况发生的速率取决于两个连续世代中配子类型的频率。如果两个位点被认为在同一条染色体上连锁, 那么这种情况可能最容易具体化。后裔世代中不平衡系数 D , 可依 4 种配子类型的任一种的频率而获得, 因此, 让我们只考察 $A_1 B_1$ 类型。这种类型可以两种方式出现于后裔配子中。首先, 它可作为基因型 $A_1 B_1/A_x B_x$ 的非重组体而产生, 下标 x 意味着两个等位基因中任一个都可出现。以这种方式产生的 $A_1 B_1$ 的频率是 $r(1-c)$, r 是亲代配子中 $A_1 B_1$ 的频率, c 是重组频率。其次, 或者它可作为基因型 $A_1 B_x/A_x B_1$ 的重组体而产生。 $A_1 B_x$ 染色体的频率是 p_A , $A_x B_1$ 染色体的频率是 p_B 。那么, 以这种方式产生 $A_1 B_1$ 的频率是 $p_A p_B c$ 。因此, 后裔配子中 $A_1 B_1$ 的频率是:

$$r' = r(1-c) + p_A p_B c$$

并且后裔世代中的不平衡系数是:

$$\begin{aligned} D' &= r' - p_A p_B \\ &= r(1-c) - p_A p_B (1-c) \\ &= (r - p_A p_B)(1-c) \\ &= D(1-c) \end{aligned}$$

如果我们再考察一个世代的这种过程, 我们得到:

$$D'' = D'(1-c) = D(1-c)^2$$

这样, 在任何数目 t 世代之后, 不平衡系数由下式给出:

$$D_t = D_0(1-c)^t \quad [1.5]$$

不平衡中的位点不必是连锁的。对不连锁位点 $c = \frac{1}{2}$, 不平衡的数量随着每一世代的随机交配而减半。对连锁位点, 不平衡消失更缓慢。图 1-3 显示了不同程度的连锁在 12 个世代过程中不平衡数量是如何减少的。

只要 $(1-c)$ 定义为任何位点之间没有重组时一个配子传递一代的概率, 由上面等式给出的平衡途径就同等地适用于联合考察的任何数目位点的不平衡。位点数目越大, 没有重组的概率就越小; 对两个不连锁位点, 它是 $1/2$, 对 3 个不连锁位点, 它是 $1/4$, 对 4 个不连锁位点, 它是 $1/8$ 。这样, 多位点不平衡比 2 个位点衰退得更快, 不久就控制了许多位点间的总的不平衡。这种情形的实际后果是当许多位点可用于研究时, 不平衡更象是在成对位点中发现的, 而不是在联合考察的较多数目的位点中发现的。有关 3 个位点的细节见 Crow 和 Kimura (1970), 其估计方法见 Weir 和 Cockerham (1979)。

在自然群体中经常发现重组来不及消除起始不平衡位点之间存在的连锁不平衡, 例如疾病位点和紧密连锁的 RFLP 之间。找到一个多态位点和一个疾病基因之间完全联合(完全连锁不平衡)实际上是在分子水平上鉴定疾病基因特性的第一步, 并且对遗传诊断非常有用。

可是，紧密连锁并不自动意味着位点将是处于连锁不平衡之中。甚至紧密连锁位点却可能处于连锁平衡之中（见例 1.6）；如果位点上具有不同基因频率的两个群体混合在一起，完全不连锁的位点能够处于连锁不平衡，正如上面所列举的那样。因此，对群体不平衡的了解能够告诉我们的有关群体繁育史的一些情况，单个位点不能做到这一点，因为它们经单个世代随机交配之后便达到哈迪-温伯格频率（Crow, 1986）。连锁不平衡的其他原因是小群体中不同位点等位基因之间的偶然联合，以及对特定等位基因组合的选择。连锁不平衡具有的重要性，在以后各章尤其是关于数量性状的选择中必须加以考虑。

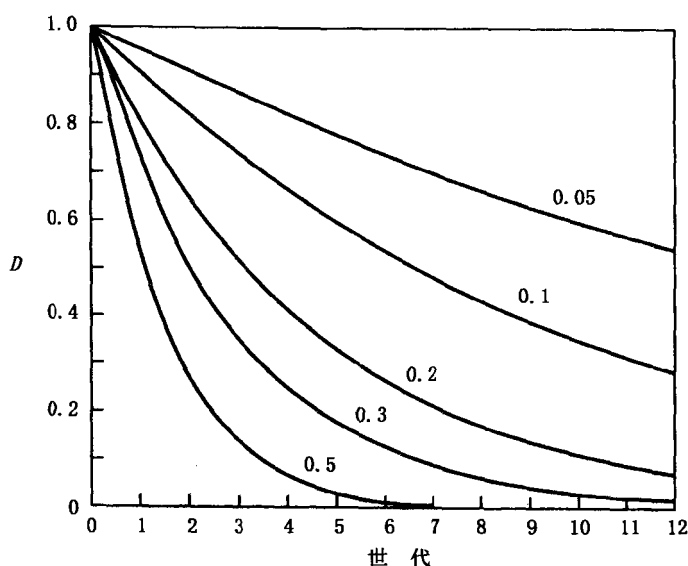


图 1-3 联合考察的两个位点随机交配之下接近平衡的情况

曲线显示了相对于 0 世代不平衡的不平衡数量 D 。5 条曲线指的是两个位点之间不同的连锁程度，它们由沿着每条曲线显示的重组频率来表示。标明 0.5 的曲线指的是不连锁位点。

非随机交配

有两种不同形式的非随机交配。第一种是交配个体因祖先血统而彼此有亲缘。这往往增加所有位点上的纯合体频率。它的效应将在第 3 章中描述。第二种是个体针对所考察的任一特定位点上的基因型倾向于优先进行交配。在此需简要地论及这种形式的非随机交配。

选型交配

如果配偶具有相同表型的机会比偶然发生的更频繁，就被称为选型交配，如果不频繁则被称为非选型交配。在表型反映基因型的程度上，选型或非选型交配影响基因型频率。Crow 和 Kimura (1970) 描述了这种影响，这里只简要列出。选型交配在人类群体中具有一些重要性，选型交配针对身高、智力和其他性状而发生。可是，这些不是单一的基因差异，不属本书讨论的范围。非选型交配在植物的自体受精体系中是普遍的。

通过适当修改交配类型的频率以允许增加相似表型之间的交配频率，可从表 1-12 中推

导出关于单个位点选型交配的后果。对后裔之间基因型频率的影响是增加纯合体的频率和减少杂合体的频率。实际上，群体只要部分地细分为两组，交配在组内就比在组间更经常发生。如果选型交配在连续世代中继续下去，那么群体接近一种平衡状态，在该状态下基因型频率保持不变。

非选型交配的后果，一般地说与选型交配的相反：它引起杂合体的增加和纯合体的减少。可是，非选型交配通常具有改变基因频率的额外后果。如果交配主要在不同表型之间进行，那么与普通表型相比，较稀少的表型在交配时有更多的成功机会。因此，较稀少的等位基因占优势，基因频率趋向中间值，在这个数值上，表型在频率上是相等的。一个熟悉的非选型交配例子是两性生殖方式，它直接导致频率为 0.5 的一个基因或染色体。植物的自体受精机制以复等位基因为基础，稀有等位基因的有利地位导致了大量的等位基因均以大致相等的频率并存。

习 题

1.1 在美国白人的一个样本中记录了人类 M-N 血型的下列数字。

M	MN	N
1 787	3 039	1 303

- (1) 在这个样本中观察到的基因型频率是多少？
- (2) 基因频率是多少？
- (3) 对观察到的基因频率，按哈迪—温伯格法则预测的基因型频率是多少？
- (4) 观察到的频率和期望值一致性程度怎样？

数据来自 Wiener, A.S. (1943) 由 Stern, C. (1973) 引用, *Principles of Human Genetics*. Freeman, San Francisco.

[解答序号: 1]

1.2 大约 30% 的人不能辨别苯氨基硫甲酸 (PTC) 的苦味。没有能力辨味是由于常染色体上—单个隐性基因引起的。假定群体处于哈迪—温伯格平衡，那么不能辨味基因的频率是多少？

[解答序号: 11]

1.3 白化病在欧洲人群体中发生的频率大约是每 20 000 人中有 1 人。假定它是由常染色体上—单个隐性基因引起的，并假定群体处于哈迪—温伯格平衡，那么携带者的比例是多少？只需给出一个近似答案。

[解答序号: 21]

1.4 作为一种代数练习，计算出将导致 $\frac{1}{3}$ 的正常个体为携带者的一个随机交配群体中一个隐性突变体的基因频率。

[解答序号: 31]

1.5 在 178 名英国人的一个样本中发现了红细胞酸性磷酸酯酶的 3 个等位变异 A、B 和 C。通过电泳可识别所有基因型，并且样本中的频率是：

基因型	AA	AB	BB	AC	BC	CC
频率 (%)	9.6	48.3	34.3	2.8	5.0	0.0