UMSS

大学数学科学丛书 —16

遗传学中的统计方法

李照海　覃红　张洪　编著

科学出版社
www.sciencep.com

大学数学科学丛书 16

遗传学中的统计方法

李照海 覃 红 张 洪 编著

科 学 出 版 社

北 京

内 容 简 介

本书系统阐述了遗传学中常用的统计学方法，其中包括最近的一些热点问题. 本书主要内容分成 7 章. 第 1 章介绍一些遗传学中的基本概念. 第 2 章和第 3 章介绍连锁分析方法. 第 4 章和第 5 章介绍关联分析方法. 第 6 章和第 7 章讨论多个位点的连锁分析和关联分析. 学习本书内容只需要具备概率论、数理统计和遗传学等方面的基础知识.

本书可用作数学、统计学和遗传学等专业的高年级本科生、研究生教材，亦可供有关科技工作者参考.

图书在版编目(CIP)数据

遗传学中的统计方法/李照海，覃红，张洪编著. —北京：科学出版社，2006
(大学数学科学丛书；16/李大潜主编)

ISBN 7-03-017832-7

I. 遗… II. ①李… ②覃… ③张… III. 数理统计-应用-遗传学 IV. Q3

中国版本图书馆 CIP 数据核字(2006) 第 094312 号

责任编辑：吕 虹 赵彦超 / 责任校对：张 琪
责任印制：安春生 / 封面设计：王 浩

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

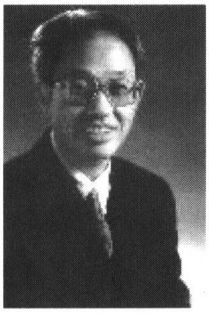
丽源印刷厂印刷

科学出版社发行 各地新华书店经销
*
2006 年 9 月第 一 版 开本：B5(720×1000)
2006 年 9 月第一次印刷 印张：15
印数：1—3 000 字数：277 000

定价：34.00 元

(如有印装质量问题，我社负责调换〈新欣〉)

作 者 简 介



李照海，美国乔治华盛顿大学统计系和流行病学与生物统计系教授、美国国家卫生局 (NIH) 癌症研究所的客座研究员. 1978 年毕业于华中师范学院 (现华中师范大学) 数学系，曾在武汉测绘学院 (现武汉大学) 工作一年，1981 年于华中师范学院获数理统计硕士学位并留校任教，曾访问斯坦福大学统计系一年，1989 年于美国哥伦比亚大学获统计学博士学位，曾在华盛顿大学医学院从事生物统计的研究与教学. 2006 年当选为美国统计协会(ASA)的 Fellow. 在“Journal of Royal Statistical Society”、“Journal of the American Statistical Association”、“Biometrics”、“American Journal of Human Genetics”、“Annals of Human Genetics”、“Human Heredity”等刊物上发表论文. 现主要从事生物统计的教学与研究，侧重于对遗传学中的统计方法的研究.

覃　红，华中师范大学统计系教授. 1988 年毕业于华中师范大学数学系，1991 年于华中师范大学获数理统计硕士学位并留校工作，2002 年于香港浸会大学获博士学位.曾应邀赴美国加州大学伯克利分校统计系、香港浸会大学做访问学者，2003 年 10 月至 2004 年 4 月于美国乔治华盛顿大学统计系从事博士后研究工作. 在“中国科学”、“Journal of Statistical Planning and Inference”、“Statistical and Probability Letters”、“Statistical Paper”、“Metrika”等国内外刊物上发表论文. 现主要从事试验设计、生物统计的教学与研究.

张　洪，中国科学技术大学讲师. 1997 年本科毕业于中国科学技术大学数学系；2000 年获中国科学技术大学授予的理学硕士学位，同年毕业留校；2003 年获得中国科学技术大学授予的理学博士学位；2004 年 3 月至 2005 年 4 月在美国乔治华盛顿大学从事博士后研究工作；曾先后应邀访问过香港科技大学和新加坡国立大学. 主要从事生存分析和遗传统计学的统计方法和理论研究.

《大学数学科学丛书》序

按照恩格斯的说法, 数学是研究现实世界中数量关系和空间形式的科学. 从恩格斯那时到现在, 尽管数学的内涵已经大大拓展了, 人们对现实世界中的数量关系和空间形式的认识和理解已今非昔比, 数学科学已构成包括纯粹数学及应用数学内含的众多分支学科和许多新兴交叉学科的庞大的科学体系, 但恩格斯的这一说法仍然是对数学的一个中肯而又相对来说易于为公众了解和接受的概括, 科学地反映了数学这一学科的内涵. 正由于忽略了物质的具体型态和属性、纯粹从数量关系和空间形式的角度来研究现实世界, 数学表现出高度抽象性和应用广泛性的特点, 具有特殊的公共基础地位, 其重要性得到普遍的认同.

整个数学的发展史是和人类物质文明和精神文明的发展史交融在一起的. 作为一种先进的文化, 数学不仅在人类文明的进程中一直起着积极的推动作用, 而且是人类文明的一个重要的支柱. 数学教育对于启迪心智、增进素质、提高全人类文明程度的必要性和重要性已得到空前普遍的重视. 数学教育本质是一种素质教育; 学习数学, 不仅要学到许多重要的数学概念、方法和结论, 更要着重领会到数学的精神实质和思想方法. 在大学学习高等数学的阶段, 更应该自觉地去意识并努力体现这一点.

作为面向大学本科生和研究生以及有关教师的教材, 教学参考书或课外读物的系列, 本丛书将努力贯彻加强基础、面向前沿、突出思想、关注应用和方便阅读的原则, 力求为各专业的大学本科生或研究生 (包括硕士生及博士生) 走近数学科学、理解数学科学以及应用数学科学提供必要的指引和有力的帮助, 并欢迎其中相当一些能被广大学校选用为教材, 相信并希望在各方面的支持及帮助下, 本丛书将会愈出愈好.

李大潜

2003 年 12 月 27 日

前　言

统计方法在遗传学的发展过程中一直起着极为重要的作用, 现代统计学的奠基者 K. 皮尔逊和 R. A. 费歇尔就是研究遗传统计学的专家. 检验遗传学之父 G. J. 孟德尔发现的孟德尔第一定律和孟德尔第二定律, 就需要用到离散分布的拟合优度检验. 又如哈代 - 温伯格平衡定律, 其本身在数学上没有多少困难, 但是它在《科学》上的发表对以后人们利用关联分析来寻找致病基因提供了强有力的理论依据. 用于试验动植物的区间定位方法, 已经成为一类非常成熟的统计方法. 两类用于寻找疾病基因的主要的统计方法, 连锁分析和关联分析, 现在已经被遗传学家广为采用. 随着科研工作者对基因研究的深入进行, 统计方法必将继续发挥其不可替代的作用.

本书侧重于介绍遗传学尤其是人类遗传学中常用的统计学方法, 也涉及当前一些热点问题的探讨. 阅读本书不需要高深的数学知识, 因此非常适合于需要了解相关统计学方法但又畏惧艰深数学的读者, 所涉及到的概率统计基本知识大部分列在附录中. 另外, 为了方便读者学习, 我们对大部分关键公式给出了简明易懂的推导过程, 这是本书的一大特色. 在书的最后, 读者可以找到本书正文中出现的一些专有名词的索引.

本书的出版得到了科学出版社和吕虹编审的大力支持和关心. 美国哥伦比亚大学的应志良教授、乔治华盛顿大学的 J. L. Gastwirth 教授和美国国家卫生局 (NIH) 的 M. H. Gail 博士对本书的出版也倾注了关心. 华中师范大学的谢民育教授、中国科学技术大学的杨亚宁教授、北卡罗来纳大学公共卫生学院统计系的朱宏图教授、美国 NIH 的田欣博士和郑刚博士, 以及其他一些同事和学生在阅读本书的初稿时纠正了不少错误并提出了宝贵意见. 另外, 本书还得到 NIH 基金的资助. 编者借本书的出版之际向他们一并表示诚挚的谢意.

本书由李照海、覃红和张洪编著. 李照海教授在乔治华盛顿大学的研究生课程上讲授过本书的前五章的主要内容, 他还曾先后于中国科学技术大学、香港科技大学、北京大学和武汉大学的短期讲座上讲授过这部分内容. 覃红和张洪分别在华中师范大学和中国科学技术大学的讨论班上也讲授过本书的大部分内容. 由于编者水平有限, 难免会有一些错误和不当之处, 恳请同行和广大读者批评指正.

编　者

2006 年 6 月 30 日

目　录

*　　*　　*

第 1 章　群体遗传学中的基本概念与原理

群体遗传学 (population genetics) 是研究群体 (population) 遗传受什么规律支配以阐明该群体生物进化机制的一门学科. 在这一章里我们对群体遗传学中的基本概念和原理作简单介绍, 其中我们参考了李景均 (C.C. Li) 于 1955 年出版的遗传学教材 "Population Genetics"[51]. 为了方便读者了解、掌握本书的主要内容, 我们试图用初等概率统计的语言和公式对群体遗传学中的概念与原理进行表述和推导.

§1.1　遗传学的基本概念与术语

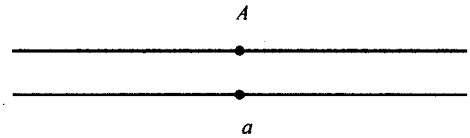


图 1.1.1

每一个正常的人有 23 对染色体 (chromosome), 其中有一对染色体起确定性别的作用, 称为性染色体 (sex chromosome), 它由染色体 X 和 Y 组成, 女性具有 XX 型染色体, 男性具有 XY 型染色体, 其余的 22 对称为常染色体 (autosome). 在本书中, 我们主要研究常染色体. 在研究常染色体时, 男性与女性没有区别, 父亲与母亲是同等的. 性染色体称为第 23 对染色体, 其余 22 对常染色体按长短顺序命名, 第一对染色体最长, 只有第 21 对染色体比第 22 对染色体短这一例外. 我们可以把一对染色体想象成为两条平行的直线 (见图 1.1.1). 染色体上一个给定的位置 (好比两条平行直线上的一点或一段) 叫做一个位点 (locus), 在同一位点上不同形式的脱氧核糖核酸 (deoxyribonucleic acid, 简称 DNA) 序列叫做等位基因 (allele), 通常用英文字母如 A, a, B, b 或数字 1, 2, 3 等来表示. 比如, 图 1.1.1 中的 A 和 a 就是两个不同的等位基因. 在分子生物学和遗传学中, 基因 (gene) 这个术语有时指位点, 有时指等位基因, 在本书中, 我们尽量回避这种用法. 在一个给定的位点上, 每个人有两个等位基因位于两条同源染色体 (homologous chromosome) 上, 例如图 1.1.1 中的 A 和 a, 同源染色体在同一个位点上的两个等位基因作为一个整体称为基因型 (genotype), 图 1.1.1 中的基因型为 Aa. 有时基因型的两个等位基因的顺序对研究没有关系而可以忽略, 因此可视 Aa 和 aA 为同一基因型, 在计算概率的时候需要注意这一约定. 如果某位点上的两个等位基因相同, 如 AA 或 aa, 则称此基因型为纯合的 (homozygous); 如果不同, 如图 1.1.1 中的 Aa , 则称此基因型为杂合的 (heterozygous). 同一位点上的两个等位基因可以通过实验测出, 这样也就确定了基因型. 在自然界中, 很多性状

(trait) 是受某一位点上的基因型控制的, 这些由基因型控制而又可以观察到的性状的不同形态称为表现型 (phenotype). 由基因控制的性状很普遍, 例如, 人的身高、眼睛颜色, 等等. 同一性状有可能由多个位点上的基因型所控制, 因此, 表现型与基因型之间的关系并不是一一对应的, 可能几种不同的基因型对应同一表现型或者更复杂. 如果基因型 AA 和 Aa 有同样的表现型, 但和 aa 的表现型不同, 这时称 A 相对于 a 是显性的 (dominant), 或者说 a 相对于 A 是隐性的 (recessive), 与 AA 和 Aa 对应的表现型称为显性的, 与 aa 相对应的表现型称为隐性的. 例如, 某种植物的颜色由某一位点上的两个等位基因 A 和 a 所控制, 基因型 AA 和 Aa 产生的表现型为红色, 基因型 aa 产生的表现型为黄色, 则 A 相对于 a 是显性的, 此时表现型红色为显性的, 黄色为隐性的. 如果杂合基因型 Aa 对应的表现型既不同于 AA 的表现型, 也不同于 aa 的表现型, 则称 A 和 a 是共显性的 (codominant). 例如, 在决定人类血型的位点上有三个等位基因: A、B 和 O, 基因型 AA 和 AO 有同样的表现型, 即血型 A; 基因型 BB 和 BO 有同样的表现型, 即血型 B; 基因型 AB 有不同于血型 A 和 B 的表现型, 即血型 AB; 而 OO 又有不同于前面三种血型的表现型, 即血型 O. 因此, A 相对 O 是显性的, B 相对 O 也是显性的, 而 A、B 为共显性的.

在一群体中, 随机地抽取一个等位基因, 这个等位基因是 A 的概率记为 $P(A)$. 这一等位基因的概率可以通过随机抽样, 用等位基因的频率来估计. 例如, $p=P(A)=0.3$, 表示在给定位点上的等位基因是 A 的可能性为 30%. 注意到, 在人群中抽样时, 首先抽取的是一个个体. 每一个人有基因型, 基因型由两个等位基因组成. 因此, 可以把对等位基因抽样看为两步随机抽样: 第一步随机抽取一个个体; 第二步, 从被抽取到的人的两个等位基因中随机抽取一个等位基因, 每个等位基因有相等的概率被抽取到. 假定某一位点上有两个等位基因 A 和 a, 三个基因型概率为

$$P(AA)=p_{AA},\ P(Aa)=p_{Aa},\ P(aa)=p_{aa}.$$

利用两步随机抽样的思想和全概率公式 (见附录 B.1.3) 可以导出

$$P(A)=P(A|AA)P(AA)+P(A|Aa)P(Aa)+P(A|aa)P(aa)=p_{AA}+\frac{1}{2}p_{Aa}, \quad (1.1.1)$$

其中 $P(A|Aa)=1/2$ 表示给定已知抽取到的个体基因型为 Aa, 随机抽取到等位基因 A 的条件概率为 1/2.

类似地可以得到

$$P(a)=p_{aa}+\frac{1}{2}p_{Aa}. \quad (1.1.2)$$

考虑一个有等位基因 A 和 a 的位点, 研究者随机抽取了 n 个个体, 其中具有基因型 AA, Aa 和 aa 的个体数目分别为 n_{AA}, n_{Aa}, n_{aa} $(n_{AA}+n_{Aa}+n_{aa}=n)$, 则等

位基因概率 $p_A = P(A)$ 可由

$$\hat{p}_A = \frac{2n_{AA} + n_{Aa}}{2n} = \frac{n_{AA}}{n} + \frac{1}{2} \times \frac{n_{Aa}}{n} \tag{1.1.3}$$

来估计. 注意到, n_{AA}/n 是基因型 AA 的频率, n_{Aa}/n 是基因型 Aa 的频率, 由此可见, 公式 (1.1.3) 对应于公式 (1.1.1). 值得注意的是公式 (1.1.3) 中的分母为 $2n$ 而不是个体数 n, 是因为每一个体有两个等位基因. 不难计算等位基因概率的估计量 $\hat{p}_A$ 的方差为

$$\mathrm{Var}(\hat{p}_A) = \frac{p_A(1-p_A)}{2n}.$$

例 1.1.1([51]) 人类 MN 血型位点上有两个等位基因 M 和 N, 现有 MN 血型的数据: 1000 多个中国香港地区的居民接受了调查, 获得以下结果:

血型	MM	MN	NN	总数
数目	342	500	187	1029

等位基因 M 的概率可以估计为

$$\hat{p}_M = \frac{2 \times 342 + 500}{2 \times 1029} = \frac{1184}{2058} = 0.5753.$$

由 (1.1.1) 和 (1.1.2) 知, 某一群体的基因型概率决定了等位基因概率. 值得注意的是, 等位基因概率并不能决定基因型概率. 一个自然的问题是, 在什么条件下等位基因概率可完全决定基因型概率? 答案是满足哈代 - 温伯格平衡定律 (Hardy-Weinberg equilibrium).

§1.2 哈代 - 温伯格平衡定律

英国数学家哈代 (G.H. Hardy) 和德国生理学家温伯格 (W. Weinberg) 于 1908 年同时发表了遗传学中的平衡定律. 他们证明了基因型概率在一代随机交配后达到平衡, 而且以后一直保持这种平衡, 除非一些因素改变群体的等位基因概率. 哈代 - 温伯格平衡定律在遗传学的研究中起着重要作用. 李景均发现卡斯尔 (W.E. Castle) 早在 5 年前 (1903) 就提出了同样的定律, 并于 1967 年在纪念卡斯尔诞辰 100 周年时公布了这一细心的发现, 他将这一遗传平衡定律称为 "卡斯尔 - 哈代 - 温伯格定律"(见 [1]).

在讨论哈代 - 温伯格平衡定律之前, 需要介绍两个概念: 一个是随机婚配 (random mating), 另一个是孟德尔第一定律 (Mendelian first law) 或独立分离原理 (the principle of independent segregation). 随机婚配是指任何一位女性有相同的可能性

与任何一位男性婚配, 即任一对夫妻的两个人作为随机变量是独立的, 这样婚配类型的概率等于女性和男性的基因型概率的乘积, 例如

$$P(AA \times Aa) = P(AA)P(Aa),$$

其中符号 $AA \times Aa$ 表示 AA 基因型个体和 Aa 基因型个体相婚配这一事件. 随机婚配这一概念是对某一个给定位点而言的. 比如说, 某位点控制人体的高度, 则对这一位点而言, 人群很可能不是随机婚配, 因为人们选择配偶时在身高上有一定的倾向. 然而, 对血型位点而言, 人群很可能是随机婚配, 人们一般不会考虑以血型相配作为择偶条件.

孟德尔 (G. J. Mendel) 于 1866 年发表了孟德尔第一定律和孟德尔第二定律 (Mendelian second law 或 independent assortment), 对孟德尔第二定律的讨论在后面进行. 孟德尔在种豌豆的实验中发现豌豆有黄和绿两种颜色, 如果基因型 AA 和 Aa 对应的表现型为黄色, 基因型 aa 对应的表现型为绿色, 则当基因型 Aa 与基因型 Aa 的个体交配 (crossing) 时, 产生的后代中颜色为黄和绿的比率为 3:1. 也就是说, 交配型 $Aa \times Aa$ 的后代基因型为 AA 或 Aa 的概率为 3/4, 为 aa 的概率为 1/4. 由此, 孟德尔总结出了如下猜想: 母本 (或父本) 遗传两个等位基因中的任何一个给后代的概率是相等的 (都是 1/2), 并且母本与父本的等位基因遗传是独立的. 这就是孟德尔第一定律. 可用概率统计的符号来表示这一定律:

$$P(\to A|AA) = 1,\ P(\to A|Aa) = \frac{1}{2},\ P(\to A|aa) = 0,$$

这里记号 $\to A|AA$ 表示事件 "基因型是 AA 的父本 (或母本) 遗传等位基因 A 给后代" 这一事件. "父本与母本传递等位基因给后代是相互独立的" 这一结论可以表示为

$$\begin{aligned}
&P(M \to A, F \to A|M = Aa, F = Aa) = P(M \to A|M = Aa)P(F \to A|F = Aa),\\
&P(M \to A, F \to a|M = Aa, F = Aa) = P(M \to A|M = Aa)P(F \to a|F = Aa),\\
&P(M \to a, F \to A|M = Aa, F = Aa) = P(M \to a|M = Aa)P(F \to A|F = Aa),\\
&P(M \to a, F \to a|M = Aa, F = Aa) = P(M \to a|M = Aa)P(F \to a|F = Aa),
\end{aligned}$$

其中 M 表示母本 (mother), F 表示父本 (father). 用孟德尔第一定律可以很好地解释孟德尔的实验数据. 费歇尔 (R.A. Fisher) 于 1936 年分析了孟德尔的实验数据, 由于拟合性太好, 引起他对数据的真实性的怀疑. 继费歇尔之后, 很多人对孟德尔的数据进行了重复分析 (见 [97]).

下面我们来介绍哈代 - 温伯格平衡定律. 考虑一个位点, 有两个等位基因 A 和 a, 假设两个等位基因在亲代 (parental generation) 群体中的概率分别为

$$P(A) = p,\ P(a) = 1 - p = q.$$

如果群体的三种基因型的概率与等位基因的概率的关系为

$$P(AA)=p^2,\ P(Aa)=2pq,\ P(aa)=q^2, \tag{1.2.1}$$

即

$$p_{AA}=p_A^2,\ p_{Aa}=2p_Ap_a,\ p_{aa}=p_a^2,$$

则称此群体在该位点处的基因型概率具有哈代 - 温伯格比例 (proportion). 在随机婚配条件下, 子代 (offspring generation) 的基因型和等位基因概率仍分别为

$$P(AA)=p^2,\ P(Aa)=2pq,\ P(aa)=q^2$$

和

$$P(A)=p,\ P(a)=q.$$

我们利用全概率公式给出上述子代的基因型和等位基因概率的一个简单的推导. 在遗传学里经常用图表来给出推导的细节, 这里我们也用图表来推导. 用 MT 表示婚配类型, OG 表示子代基因型, 表 1.2.1 给出了 MT 的概率、OG 和 MT 的联合概率分布, 亲代基因型概率具有哈代 - 温伯格比例. 下面以计算 $P(OG=AA)$ 为例来说明子代基因型的概率的计算.

表 1.2.1 随机婚配条件下子代的基因型分布

序号	MT	$P(MT)$	$P(OG,MT)$		
			AA	Aa	aa
1	$AA\times AA$	p^4	p^4	0	0
2	$AA\times Aa$	$4p^3q$	$2p^3q$	$2p^3q$	0
3	$AA\times aa$	$2p^2q^2$	0	$2p^2q^2$	0
4	$Aa\times Aa$	$4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
5	$Aa\times aa$	$4pq^3$	0	$2pq^3$	$2pq^3$
6	$aa\times aa$	q^4	0	0	q^4
	总 和	1	p^2	$2pq$	q^2

注意表 1.2.1 的第三列由随机婚配假设推出, 例如

$$P(AA\times Aa)=2P(AA)P(Aa)=2p^2\times 2pq=4p^3q.$$

由于考虑的位点在常染色体上, 男性与女性地位是对等的, 因此视 $AA\times Aa$ 与 $Aa\times AA$ 为同一婚配类型. 实际上, 如果区分性别, $AA\times Aa$ 对应两种不同的婚配类型, 一种是男的基因型为 AA 而女的基因型为 Aa, 另一种是男的基因型为 Aa 而女的

基因型为 AA, 这就是在上式中为什么要乘 2 的原因. 表 1.2.1 中的第四、五、六列由孟德尔第一定律推出, 例如第二行第五列的式子可由下面推出

$$P(OG=AA, MT=AA\times Aa)=P(OG=AA|MT=AA\times Aa)P(MT=AA\times Aa)$$
$$=\frac{1}{2}\times 4p^3q=2p^3q,$$

其中, 根据孟德尔第一定律, 有

$$P(OG=AA|MT=AA\times Aa)=P(\rightarrow A|AA)P(\rightarrow A|Aa)=1\times\frac{1}{2}=\frac{1}{2},$$

这里 $\rightarrow A|AA$ 表示基因型为 AA 的个体传递等位基因 A 给下一代这一事件.

由各个婚配型的不相容性, 有

$$P(OG=AA)=\sum_{i=1}^{6}P(OG=AA, MT=i).$$

根据表 1.2.1 的第四列, 立即可得

$$P(OG=AA)=p^4+2p^3q+p^2q^2=p^2(p^2+2pq+q^2)=p^2.$$

表 1.2.1 的最后一行给出子代的基因型概率为

$$P(AA)=p^2,\ P(Aa)=2pq,\ P(aa)=q^2.$$

根据公式 (1.1.1), 子代的等位基因 A 和 a 的概率分别为

$$P(A)=p^2+\frac{1}{2}\times 2pq=p\quad 和\quad P(a)=1-P(A)=1-p=q.$$

这就是哈代 - 温伯格平衡定律的第一部分. 概括地讲, 哈代 - 温伯格平衡定律的第一部分内容为: 如果亲代的基因型概率具有哈代 - 温伯格比例 (1.2.1), 在随机婚配的假定下, 以后每一代的基因型概率不变, 因此等位基因概率也不变, 而且以后每一代基因型概率都满足哈代 - 温伯格比例 (1.2.1). 作为练习, 建议读者推导 $P(OG=Aa)=2pq$.

哈代 - 温伯格平衡定律的第二部分是说: 如果亲代的基因型概率并不具有哈代 - 温伯格比例 (1.2.1), 在亲代随机婚配的假定下, 子代的基因型概率将具有哈代 - 温伯格比例 (1.2.1). 下面我们来给出这一结论的数学推导.

如果我们假定亲代的基因型概率为

$$P(AA)=D,\ P(Aa)=H,\ P(aa)=R,$$