·医学数据分析与信息处理方法·

医学遗传学与遗传流行病学数据分析

郭政　李霞　何颖　著

黑龙江科学技术出版社

· 医学数据分析与信息处理方法 ·

医学遗传学与遗传流行病学数据分析

郭政　李霞　何颖　著

黑龙江科学技术出版社

内 容 简 介

本书前八章对遗传学群体与家系资料分析方法进行了系统的介绍，包括研究基因频率与群体分化、相关与遗传率、疾病关联、遗传异质性、数量性状的混合分布、AGFAP遗传方式、简单分离与复合分离、发病年龄与发病风险、受累同胞(亲属)对连锁、基因连锁与连锁异质性、通径分析等大量的遗传学研究的实验设计与数据分析方法．第九章介绍PPAP(Population and Pedigree Analysis Programs)使用指南．可供医学遗传学、遗传流行病学、计划生育等领域的研究者参考使用．

责任编辑：张日新
封面设计：刘连生

医学遗传学与遗传流行病学数据分析

Yixue Yichuanxue Yu Yichuan Luxingbingxue Shuju Fenxi

郭政　李霞　何颖　著

黑龙江科学技术出版社出版发行
(150001 哈尔滨南岗区建设街 41 号)
哈尔滨电工学院印刷厂印刷

787 × 1092 毫米 1/32 开本　9.75 印张　195 千字
1996 年 11 月第 1 版 · 1996 年 11 月第 1 次印刷
印数：1 ~ 3000 册　定价 8.60 元
ISBN 7 - 5388 - 3052 - 9/N · 144

序

近年来，医学遗传学的发展非常迅速，追溯其原因，除引人分子生物学方法进行探索外，用生物数学方法对所获数据进行分析，也是一个重要的手段，对遗传流行病学的研究更是如此．郭政等同志所著《医学遗传学与遗传流行病学数据分析》一书，就遗传流行病学研究中的基因频率与群体分化分析、相关与遗传率分析、疾病关联分析、通径分析、遗传异质性分析、遗传方式分析、受累同胞(亲属)对连锁与基因定位分析、遗传咨询中复发风险估计等几个方面的问题均有扼要论述和简要的数学原理介绍，并列出了各种计算机程序．最后一章中，设计了包括上述各方面程序的PPAP系统，介绍了该系统的功能和操作方法，而且全部实现了微机化，在IBM486微机上即可运行．这对我国的医学遗传学研究，特别是对国际上重视的复杂多基因病的相关基因的探索必将起到推动作用．

本书手稿完成后，我粗略阅读了一遍，受益非浅，觉得是一本有创造性新意的好书，对医学遗传学研究者是一本不可或缺的参考书，特为之序．

李璞于哈尔滨医科大学

1996年10月

前言

遗传性疾病的发生具有各种独特的家族传递规律、种族或群体差异等时空表现特征。根据群体与家系资料，应用统计分析技术可以研究民族的进化与亲缘关系、是否有遗传因素影响疾病或性状及其影响程度、遗传因素的影响方式与传递方式、发病年龄与发病风险、疾病的遗传学分型(遗传异质性分析)、致病基因的连锁定位等问题。统计与信息处理技术也是人类基因组计划研究的关键技术之一。

我国民族与人口众多，家系较大，居住亦相对集中，并且有许多隔离群体，具有利用群体与家系法进行遗传学研究的有利条件。另一方面，群体与家系资料的分析方法一般都涉及复杂的计算，这样便限制了许多重要的遗传学研究工作在国内的开展。我们对遗传学群体与家系资料的分析方法进行了系统的分析比较，在此基础上研制了有关的计算机程序，并在各章中分别介绍，同时配有应用实例。即使对于从未使用过计算机的研究者，也只需要1～2个小时的学习就能掌握这些程序的应用。利用这些程序，我们协助多项课题进行了应用研究。实践证明，采用方便高效的数据分析技术，可以从传统的实验数据中充分地提炼信息，不但能提高工作效率，还可以拓广研究领域，达到事半功倍的目的。

本书前八章以介绍各程序的使用为主，对遗传学群体与家系资料分析方法进行了系统的介绍，包括研究基因频率与群体分化、相关与遗传率、疾病关联、遗传异质性、数量性状的混合分布、遗传方式(分离分析等)、发病年龄与发病风险、受累同胞(亲属)对连锁、Lods连锁与连锁异质性、通径分析等遗传学研究的实验设计与数据分析方法。

第九章介绍PPAP使用指南。PPAP是由前七章介绍的独立程序整合而成的一个遗传学群体与家系资料分析的计算机系统，由12个功能子系统组成。系统采用汉字提示菜单，操作灵活、方便。该汉化的系统可靠实用，是人类与医学遗传学、遗传流行病学等研究的一个有效的工具。

本书对涉及的数学原理一般只作简要介绍，比较复杂的数学公式并不列出。一些读者在阅读分析方法的数学原理时可能有困难，这时可以略而不读。我们尽量多列出一些有关方法与应用方面的参考文献，为有兴趣的读者进一步学习探索提供便利。数量性状的混合分布、发病年龄分布的最大似然函数分析与通径分析(涉及优化算法)程序有时存在收敛不稳定的问题，此时需要较多的数学经验通过调整参数的上下限与初值来解决，这对一般的使用者是较困难的，本书作者愿意提供这方面的帮助。

第八章中简要介绍的遗传病计算机辅助诊断与查询系统(IDDC)是在刘权章教授主持下研制完成的。

我们的研究工作从一开始就得到了张贵寅教授与安齐国教授的帮助与指导，他们分别为我们提供了大量的实验数据与良好的科研条件，这对我们研制PPAP系统具有很大的帮助。我们的研究工作也得到了《中国优生与遗传杂志》主编李崇高教授、《国外医学遗传学分册》编辑部戴德林教授、上海医科大学沈福民教授、哈尔滨医科大学李璞教授等遗传学家的帮助与鼓励。MC. Babron博士为我们提供了MASC程序，K. Lange博士为我们提供了MENDEL程序。谨在此表示深切的谢意！

作 者

1996年10月

目　录

第一章 基因频率与群体分化

本章介绍Hardy-Weinberg平衡与基因频率、单体型频率及连锁不平衡参数的估计方法，并介绍有关群体分化与群体间的遗传分化的度量方法等。

一 Hardy-Weinberg定律与基因频率估计

Hardy-Weinberg定律是群体遗传学基本定律，根据这个基本定律可以估计基因频率。

(一) Hardy-Weinberg定律

Hardy-Weinberg定律是由英国数学家Hardy与德国生理学家Weinberg于1906年分别提出的群体遗传学基本定律．其中心内容是：在随机交配的大群体内，若没有选择、突变或迁移因素的影响，基因频率或基因型频率在世代间保持恒定不变，基因频率与基因型频率之间存在着如下的简单关系：

若一个常染体基因座位上有两个等位基因A1与A2，其频率分别为P_1与P_2，则三种基因型A1A1、A1A2、A2A2的频率分别P_1P_1、$2P_1P_2$与P_2P_2(亲、子代间均恒定)．一般地，若一个常染色体基因座位有多个等位基因A1、A2、…、An，其频率分别为P_1、P_2、…、P_n，则AiAi(i=1，2，…，n)基因型的频率为P_iP_i，而AjAk(j≠k)的频率为$2P_jP_k$。对X染色体上的座位，女性群体中的基因频率与基因型频率仍符合前述关系。

上述关于基因频率与基因型频率的关系仅当群体已处于Hardy-Weinberg平衡时才成立。在这种条件下，男女两性中的基因频率无差异。若在初始群体中，男女两性中的某常染色体座位上的基因频率有差异，则经过一次随机交配后，两性中的基因频率即可变为一致，达到平衡；若男女两性中的某X染色体座位上的基因频率有差异，则每经过一次随机交配后，两性中的基因频率差异将减少一半，多次随机交配后，基因频率最终将在两性中相等。

要检验一个群体是否已处于Hardy-Weinberg平衡，首先依次计算各表型的期望值，表型期望值与观察值的吻合程度用χ^2来量度，对每一种表型求χ^2，然后相加即得总的χ^2，计算式为

$$\chi^2=\Sigma\ ((\text{观察值}-\text{期望值})^2/\text{期望值}$$

Σ 为对所有表型求和，自由度为$df=m-n$。其中m为表型数，n为等位基因个数(空白基因计为一个)。

在χ^2测验中，常会遇到所求得的表型期望值小于5。对此情况，一般合并几项表型使之大于5后再求χ^2，但这种合并方式多少有点主观因素，合并之后求得的χ^2值将下降。另外，不同资料由于合并方式不同也不便比较。故对于表型期望值小于5的项，也可不必合并，而是同样求χ^2，尽管这样处理会使χ^2升高，但如果在这种“升高”的情况下仍然得出吻合的结论说明该结论是比较可靠的。按此方法作检验，结果较为保守，但也是较为简单实用的方法。一般以$p\geq0.05$作为期望值与观察值无显著差异的界限。

对X染色体座位，只需根据女性表型的观察值与期望值进行平衡检验。

(二) 常染色体座位等位基因频率估计

根据常染色体座位等位基因之间不同的显隐关系，有不同的基因频率估计方法。

1. 常染色体座位的等位基因呈共显性关系

此时基因型完全能够根据表型区别出来。某等位基因的频率p可由其纯合体基因型频率加上含有该基因的所有杂合子频率总和的一半来估计。

例1 人的MN血型的表型有三种，由一对等位基因G^M与G^N决定。对69685人作调查，M型有21045人，N型有14262人，MN型有34378人，求该群体的基因G^M与G^N的频率。

G^M与G^N是共显性基因，所以基因G^M与G^N的频率分别是:

$$P(G^M) = p = \frac{21045}{69685} + \frac{1}{2} \times \frac{34378}{69685} = 0.5487$$

$$P(G^N) = q = \frac{14262}{69685} + \frac{1}{2} \times \frac{34378}{69685} = 0.4513$$

2. 常染色体座位等位基因存在显隐关系

若某基因仅在纯合状态下才致病，调查样本为N人，发病者为n人，则该基因的频率q可按下法估计:

$$q = \sqrt{n/N} \qquad SE = \sqrt{(1-q2)/4N}$$

例2 Pearn在英国231370个新生儿中确定了39例Werdnig－Hoffmann病(隐性遗传)患儿，则该致病基因的频率为:

$$q = \sqrt{39/231370} = 0.0062, \quad SE(q) = 0.0010$$

计算机程序FRE

当一个基因座位上有几个共显性等位基因、一个隐性(或空白)等位基因时(人类ABO血型及HLA血型都是典型的这种遗传类型)，常采用Bernsterin公式等近似方法估计基因的频率，但这种近似估计的偏差较大。最大似然函数法是估计等位基因频率的一种较可靠方法，我们编制了有关的计算机程序FRE。该程序首先通过调查获得的表型频数，采用Bernsterin公式估计各基因频率的初值，再使用Yasuda与Kimura[1]的基因计数法，根据Nam与Gart介绍的改良算法[2]由对数最大似然函数迭代计算基因频率及其估值的标准误，同时计算用于Hardy-Weinberg平衡检验的χ^2值与自由度[3]。

例3 106名汉族个体中HLA－C座位上的抗原表型观察数见表1.1。

表1.1 HLA－C座位上的抗原表型观察值

表 型	1	2	3	4	1,2	1,3	1,4	2,3	2,4	3,4	空白
观察数	10	1	43	9	0	17	1	1	1	5	18

使用EDIT建立输入数据文件(文件名fre.in)输入数据．输入文件内容与格式如下:

```
5
18  10  1  43  9  0  17  1  1  1  5
```

注: m

```
n(m)     n(1)    n(2)   …    n(m-1)
n(1,2)   n(1,3)  …   n(1,m-1)
n(2,3)     …    n(2,m-1)
……
```

n(m-2, m-1)

m为该座位等位基因个数；n(i)为表型为Ai的例数(i=1, …, m)；而Am表示空白或隐性基因；n(i, j)为表型为AiAj的例数(i=1, …, m-1; j>i)。

输入数据文件建立好后，存盘退出字处理系统。运行程序fre，按屏幕提示逐项输入下述内容：

输入文件名：fre. in

输出文件名：fre. out

键入回车后程序即可运行计算分析结果如下：

常染色体座位之基因频率估计及H-W平衡检验

表型	观察值	期望值	xx	基因频率 ± SE
空白	18	15. 2502	0. 4958	p(0)=0. 3871 ± 0. 0422
1	10	13. 5132	0. 9134	p(1)=0. 1395 ± 0. 0234
2	1	1. 1899	0. 0303	p(2)=0. 0142 ± 0. 0081
3	43	46. 4402	0. 2548	p(3)=0. 3797 ± 0. 0275
4	9	7. 1897	0. 4558	p(4)=0. 0795 ± 0. 0165
1, 2	0	0. 4211	0. 4211	
1, 3	17	11. 2306	2. 9639	
1, 4	1	2. 3502	0. 7757	
2, 3	1	1. 1461	0. 0186	
2, 4	1	0. 2398	2. 4094	
3, 4	5	6. 3957	0. 3046	
总和	106	105. 3666	9. 0435	df=6 PIC=0. 6213

注：p(0)为空白基因频率．PIC(polymorphism information content)为该座位多态性信息含量．

(三) X染色体座位上的基因频率

在X染色体座位上，男性为半合子，所以男性中的某等位基因频率等于其相应的表型频率。根据女性的表型频数估计女性中基因频率的方法与根据常染色体座位表型频率估计基因频率的方法一样。假设X连锁座位上的共显性等位基因为S和s，在m例男性中，带有S的个体数为a，带有s的个体数为b；在n例女性中，SS个体数为c，Ss个体数为d，ss个体数为e，S与s基因频率分别为p与q。用基因计数法有

$$p=(a+2c+d)/(a+b+2c+2d+2e)$$

$$q=(b+d+2e)/(a+b+2c+2d+2e)$$

在随机婚配的群体中，女性纯合型的频率是男性相应性状频率的平方。因此由一个X连锁隐性基因决定的性状，在男性比较常见，而女性中却少得多。例如血友病，男性患者约为万分之一，而在女性中几乎见不到。

计算机程序FREX

当一个X连锁基因座位上有几个共显性等位基因、一个隐性(或空白)等位基因时，可用程序FREX综合两性别的表型频数，采用最大似然函数法迭代计算基因频率及各基因频率估值的标准误，同时计算用于Hardy-Weinberg平衡检验的χ^2与自由度$df=n-m-1$[4]，其中n为男性表型种数加女性表型种数，m为被估计的参数个数[4]。以下例说明frex的应用。

例4 调查X-连锁红绿色盲，男性9049人中，色盲者725人，非色盲者8324人；女性9072人中，色盲者40人，非色盲者9032人。