

# 医学遗传统计分析与应用

# SAS

# 应用

主编 胡良平 郭 晋

AN UNOFFICIAL SAS FORUM

SAS FORUM

1998 1999 2000 2001

# 医学遗传统计分析与

# SAS 应用

主 编 胡良平 郭 晋

编 委(以姓氏笔画为序)

王 琪	军事医学科学院	周诗国	军事医学科学院
毛 玮	军事医学科学院	胡良平	军事医学科学院
伍亚舟	第三军医大学	胡纯严	军事医学科学院
刘惠刚	首都医科大学	柳伟伟	军事医学科学院
关 雪	军事医学科学院	贾元杰	军事医学科学院
李 霞	哈尔滨医科大学	高 辉	军事医学科学院
李子建	济南军区疾病预防控制中心	郭 晋	军事医学科学院
李长平	天津医科大学	陶丽新	军事医学科学院
李伍举	军事医学科学院	葛 毅	中国人民解放军 后勤指挥学院
张瑞杰	哈尔滨医科大学	鲍晓蕾	军事医学科学院
易 东	第三军医大学		

人民卫生出版社

## 图书在版编目 (CIP) 数据

医学遗传统计分析 with SAS 应用/胡良平等主编.

—北京: 人民卫生出版社, 2011. 2

ISBN 978-7-117-13904-5

I. ①医… II. ①胡… III. ①医学遗传学-医学统计-统计分析-应用软件, SAS IV. ①R394-31

中国版本图书馆 CIP 数据核字 (2010) 第 258374 号

门户网: <a href="http://www.pmph.com">www.pmph.com</a>	出版物查询、网上书店
卫人网: <a href="http://www.ipmph.com">www.ipmph.com</a>	护士、医师、药师、中医师、卫生资格考试培训

版权所有, 侵权必究!

## 医学遗传统计分析 with SAS 应用

主 编: 胡良平 郭 晋

出版发行: 人民卫生出版社 (中继线 010-59780011)

地 址: 北京市朝阳区潘家园南里 19 号

邮 编: 100021

E - mail: [pmph@pmph.com](mailto:pmph@pmph.com)

购书热线: 010-67605754 010-65264830

010-59787586 010-59787592

印 刷: 三河市富华印刷包装有限公司

经 销: 新华书店

开 本: 787×1092 1/16 印张: 17 插页: 4

字 数: 413 千字

版 次: 2011 年 2 月第 1 版 2011 年 2 月第 1 版第 1 次印刷

标准书号: ISBN 978-7-117-13904-5/R·13905

定 价: 36.00 元

打击盗版举报电话: 010-59787491 E-mail: [WQ@pmph.com](mailto:WQ@pmph.com)

(凡属印装质量问题请与本社销售中心联系退换)

## 内 / 容 / 提 / 要

本书介绍了医学遗传数据的统计分析方法与 SAS 实现、简明遗传学的基本概念和原理以及医学遗传统计分析的计算原理。详细介绍了如何使用 SAS/Genetics 模块和其他相关模块实现统计计算、输出结果的解释,并结合具体问题、统计和专业知识,作出令人信服的专业结论。

对医学遗传资料,首先应根据分析目的、资料所具备的前提条件正确判断其数据结构,从而选择分析方法并准确地向 SAS/Genetics 中录入数据。本书第 1 章对遗传数据结构做了详细介绍;而后针对拟分析的遗传资料,由简单到复杂的顺序,介绍各种遗传统计分析方法及遗传数据分析的 SAS 实现:基因、基因型频率的测定、哈代 - 温伯格平衡定律的验证;连锁不平衡检验、单体型分析;一般人群病例 - 对照遗传数据分析;遗传结果比较与校正;家系数数据分析、连锁分析、芯片数据的分析和物种遗传关系确定分析等。本书最后还提供了 SAS/Genetics 中全部过程的语法结构和用法简介以及与遗传有关的基本概念和基础知识,供读者参考。

本书能满足基础医学科研人员、生物试验室人员对遗传资料、遗传试验数据分析的需要,可作为高等院校遗传专业研究生工具书,也可作为医学院校、高等院校生物系本科生和研究生参考书,适用于遗传专业、生物医学专业和统计专业学者学习和借鉴。

# 前 / 言

很多疾病与遗传有关,人们简称其为遗传病。前辈遗传了什么给他们的后代?遗传了细胞中染色体上的某些基因。若前辈属于某些遗传病的患者或其虽属非显性遗传病患者但却携带着遗传病的患病基因,则他们的后代患这些遗传病的概率就很大。经过多年的医学实践和探索,人类已经弄清一小部分遗传病的遗传规律,但还有很多遗传病的遗传规律根本不清楚。然而,随着全世界医学科技事业的蓬勃发展,特别是近年来蛋白质组学、代谢组学、基因组学和后基因组学的迅猛发展,人类对基因有了更全面深入的了解,越来越多的与疾病发生、发展、预后、康复和长寿有关的基因陆续地被发现。

在研究基因与疾病关系的过程中,一系列不可回避的问题被提出来了:如何进行遗传科研课题的研究设计、如何正确地收集医学遗传资料、如何识别遗传资料的数据结构、如何针对不同的遗传资料和分析目的选用相应的遗传资料分析方法、如何方便快捷地使用国际上公认的统计软件实现遗传数据的统计分析、如何结合遗传专业知识对计算结果作出合理的解释和给出科学的专业结论。本书正是在这样一个强大医学需求的背景下应运而生的。

本书分为以下3篇,第1篇医学遗传统计实例分析与SAS实现,包含10章内容,第1章医学遗传资料数据结构与分析方法选择,提纲挈领地呈现出遗传资料中四种类型的数据结构;第2章到第9章由浅入深地介绍了遗传资料的实例分析与SAS实现;第10章SAS/Genetics模块概述,较详细地介绍了此模块的功能、语法、过程和语句。第2篇简明遗传学基本概念与原理,包含3章内容,它们分别为第11章遗传学基本概念、第12章孟德尔遗传原理与遗传病和第13章群体遗传学的基本原理。第3篇医学遗传统计计算的原理,包含

7 章内容,分别为第 1 篇实例分析中所涉及的各种遗传资料统计分析方法的计算原理。

笔者有幸于 2007 年招收了一位具有扎实生物学知识基础的硕士研究生,他的名字叫郭晋。在三年的研究生学习生涯中,他刻苦钻研,勇于拼搏,不仅在基础统计学、多因素试验设计、多元统计分析和 SAS 软件应用等方面取得了长足进步,而且,较为全面地掌握了医学遗传资料统计分析的理论和方法,并能巧妙地运用 SAS/Genetics 模块实现遗传资料的统计分析和结果解释。本书中绝大部分内容的初稿都是由他完成的。

在本书即将出版之际,笔者真挚地感谢为本书作出过很多贡献的来自哈尔滨医科大学的李霞和张瑞杰教授、第三军医大学的易东教授和军事医学科学院基础医学研究所的李伍举教授等;还要感谢所有为本书付出过辛勤劳动的人们,他们不仅直接和间接地参与了某些章节的编写工作,还认真地为全书作了审阅和校对工作;胡纯严为本书制作了方便用户调用 SAS 软件的 SAS 引导程序,即 SASPAL 软件。正是由于他们的积极参与、不懈努力和真心奉献,才使这部具有明显特色的遗传统计学专著得以问世!

由于笔者水平有限,书中难免会出现这样或那样的不妥,甚至错误之处,恳请广大读者不吝赐教,以便再版时修正。

胡良平

于北京军事医学科学院生物医学统计学咨询中心

2010 年 12 月

# 目 / 录

## 第 1 篇 医学遗传统计实例分析与 SAS 实现

第 1 章 医学遗传资料数据结构与分析方法选择 .....	3
1.1 一般人群实验研究的数据结构 .....	3
1.1.1 位点基因型数据结构 .....	3
1.1.2 一般人群病例 - 对照研究单个位点基因型数据结构 .....	4
1.1.3 一般人群病例 - 对照研究多个位点基因型数据结构 .....	5
1.2 家系实验研究数据结构 .....	6
1.2.1 家系病例 - 对照研究数据结构 .....	6
1.2.2 单个家系研究的数据结构 .....	7
1.3 基因芯片数据结构 .....	8
1.4 物种遗传关系确定分析的数据结构 .....	9
1.5 小结 .....	9
第 2 章 Hardy-Weinberg 平衡定律验证的 SAS 实现 .....	10
2.1 验证哈代 - 温伯格平衡定律是否成立的实例计算与 SAS 实现 .....	10
2.2 小结 .....	18
第 3 章 病例 - 对照研究关联分析的 SAS 实现 .....	19
3.1 一般人群作病例 - 对照研究关联分析的 SAS 实现 .....	19
3.1.1 $\chi^2$ 检验与 Armitage 检验的实例分析与 SAS 实现 .....	19
3.1.2 高维表资料 CMH $\chi^2$ 检验、CMH 校正的秩和检验的 SAS 实现 .....	26



3.1.3 高维表资料对数线性模型分析的 SAS 实现 .....	28
3.1.4 Logistic 回归分析的 SAS 实现 .....	30
3.2 家系病例 - 对照研究关联分析的 SAS 实现 .....	36
3.3 小结 .....	41
<b>第 4 章 遗传分析结果校正和输出的 SAS 实现 .....</b>	<b>43</b>
4.1 遗传分析结果校正的实例分析与 SAS 实现 .....	43
4.2 结果图形输出的 SAS 实现 .....	48
4.3 小结 .....	50
<b>第 5 章 连锁不平衡与单体型分析的 SAS 实现 .....</b>	<b>51</b>
5.1 单体型频率估计与连锁不平衡分析的 SAS 实现 .....	51
5.1.1 两位点估计的实例分析与所对应的 SAS 实现 .....	51
5.1.2 多位点估计实例分析与所对应的 SAS 实现 .....	55
5.2 多位点基因型与疾病关联研究实例分析与 SAS 实现 .....	57
5.3 标签 SNP 的确认实例分析与 SAS 实现 .....	61
5.4 单体型与疾病关联的回归分析与 SAS 实现 .....	64
5.5 小结 .....	68
<b>第 6 章 近交系数和亲缘系数估算的 SAS 实现 .....</b>	<b>69</b>
6.1 通过实例展示近交系数和亲缘系数的估算方法 .....	69
6.2 小结 .....	74
<b>第 7 章 遗传资料连锁分析的 SAS 实现 .....</b>	<b>75</b>
7.1 三代家系两位点连锁分析的 SAS 实现 .....	75
7.2 多位点连锁分析及遗传图谱构建的 SAS 实现 .....	78
7.3 小结 .....	88
<b>第 8 章 基因芯片数据分析的 SAS 实现 .....</b>	<b>89</b>
8.1 数据来源、数据结构与分析流程 .....	89
8.2 差异表达基因的筛选 .....	90
8.3 样品聚类分析方法 .....	91
8.4 判别分析 .....	94
8.5 变量聚类分析 .....	100
8.6 主成分分析 .....	105

8.7 基因调控网络分析 .....	107
8.8 小结 .....	112
<b>第 9 章 物种遗传关系确定分析的 SAS 实现 .....</b>	<b>113</b>
9.1 通过实例展示如何用 SAS 实现物种遗传关系确定的分析 .....	113
9.2 小结 .....	117
<b>第 10 章 SAS/Genetics 模块概述 .....</b>	<b>118</b>
10.1 GENETICS 模块简介 .....	118
10.1.1 ALLELE、HAPLOTYPE 和 HTSNP 过程简介 .....	118
10.1.2 CASECONTROL 和 FAMILY 过程简介 .....	118
10.1.3 INBREED 过程简介 .....	119
10.1.4 PSMOOTH 过程和 %TPLOT 自定义宏函数简介 .....	119
10.2 ALLELE、HAPLOTYPE 和 HTSNP 过程的语法结构及用法简介 .....	119
10.2.1 数据格式 .....	119
10.2.2 ALLELE 过程的语法结构及用法简介 .....	121
10.2.3 HAPLOTYPE 过程的语法结构及用法简介 .....	126
10.2.4 HTSNP 过程的语法结构及用法简介 .....	128
10.3 CASECONTROL 和 FAMILY 过程的语法结构及用法简介 .....	130
10.3.1 CASECONTROL 过程的语法结构及用法简介 .....	130
10.3.2 FAMILY 过程的语法结构及用法简介 .....	131
10.4 INBREED 过程的语法结构及用法简介 .....	134
10.5 PSMOOTH 过程的语法结构和 %TPLOT 自定义宏函数简介 .....	136
10.5.1 平滑处理和多重检验校正 .....	136
10.5.2 PSMOOTH 过程的语法结构及用法简介 .....	136
10.5.3 %TPLOT 自定义宏函数简介 .....	138

## 第 2 篇 简明遗传学基本概念与原理

<b>第 11 章 遗传学基本概念 .....</b>	<b>143</b>
11.1 染色体 .....	143
11.2 基因 .....	144
11.2.1 基因的概念 .....	144
11.2.2 等位基因 .....	145
11.2.3 复等位基因 .....	146
11.2.4 致死基因 .....	147

11.2.5 非等位基因	147
11.3 细胞分裂	148
11.3.1 有丝分裂	148
11.3.2 减数分裂	150
11.4 突变	152
11.4.1 染色体突变	152
11.4.2 基因突变	155
11.5 小结	157
<b>第 12 章 孟德尔遗传原理与遗传病</b>	<b>158</b>
12.1 孟德尔遗传定律	158
12.1.1 分离定律(law of segregation)	158
12.1.2 自由组合定律(law of independence assortment)	161
12.2 摩尔根的遗传连锁定律	162
12.3 单基因常染色体遗传病	164
12.3.1 常染色体显性遗传病	164
12.3.2 常染色体隐性遗传病	166
12.4 伴性遗传	167
12.4.1 X连锁遗传病	168
12.4.2 Y连锁遗传病	169
12.5 多基因病	169
12.6 小结	170
<b>第 13 章 群体遗传学的基本原理</b>	<b>171</b>
13.1 哈代 - 温伯格平衡定律	171
13.1.1 哈代 - 温伯格平衡定律的原理与证明	171
13.1.2 拟合优度检验	174
13.2 连锁不平衡及单体型的概念与原理	174
13.2.1 连锁不平衡的概念	174
13.2.2 连锁不平衡公式的推导	175
13.2.3 产生连锁不平衡的原因	180
13.2.4 单体型的定义	181
13.3 小结	181

## 第 3 篇 医学遗传统计分析的计算原理

<b>第 14 章 病例 - 对照研究设计资料的计算原理</b> .....	185
14.1 一般人群作对照研究的方法与计算原理 .....	185
14.1.1 二维表资料 $\chi^2$ 检验与 Armitage 检验 .....	185
14.1.2 高维表资料 CMH $\chi^2$ 检验、CMH 校正的秩和检验的方法与计算原理 .....	187
14.1.3 高维表资料对数线性模型方法概述 .....	189
14.1.4 logistic 回归分析方法与计算原理 .....	190
14.2 家系病例 - 对照研究方法方法与计算原理 .....	193
14.2.1 父母亲作对照的关联分析原理介绍 .....	193
14.2.2 同胞作对照的关联分析原理介绍 .....	194
14.3 小结 .....	196
<b>第 15 章 遗传结果校正的计算原理</b> .....	197
15.1 平滑法 .....	197
15.1.1 Simes 法 .....	197
15.1.2 Fisher 法 .....	198
15.1.3 TPM 法 .....	198
15.2 多重比较 $P$ 值的校正 .....	198
15.3 小结 .....	198
<b>第 16 章 连锁不平衡与单体型分析方法的计算原理</b> .....	200
16.1 单体型频率估计与连锁不平衡分析的假设检验 .....	200
16.1.1 最大似然法估计单体型概率 .....	200
16.1.2 E-M 算法估计单体型概率及对连锁不平衡进行假设检验 .....	201
16.2 多位点基因型与疾病关联分析的计算原理 .....	202
16.3 标签 SNP 确认的计算原理 .....	202
16.4 单体型与疾病关联的 logistic 回归分析的计算原理 .....	203
16.5 小结 .....	203
<b>第 17 章 近交系数和亲缘系数的计算原理</b> .....	204
17.1 近交的概念与计算方法 .....	204
17.1.1 一般算法 .....	204
17.1.2 通径分析 .....	205
17.1.3 X 连锁基因近交系数的计算 .....	205

17.2 亲缘系数的概念与其计算方法	206
17.2.1 父与子	206
17.2.2 祖父与孙子	206
17.2.3 同胞	206
17.2.4 半同胞	206
17.2.5 叔侄	207
17.2.6 堂(表)亲	207
17.2.7 从堂(表)亲	207
17.3 小结	207
<b>第 18 章 遗传资料连锁分析的计算原理</b>	<b>209</b>
18.1 三代家系回交实验的连锁分析的计算原理	209
18.1.1 重组率直接计算方法介绍	209
18.1.2 贝叶斯方法与蒙特卡洛模拟法估计重组率原理介绍	210
18.2 两代家系两位点的连锁分析的计算原理	211
18.2.1 最大似然法估计重组率与 LOD 记分法	211
18.2.2 根据子代基因型估计重组率	213
18.2.3 根据子代表型估计重组率	222
18.3 三位点连锁分析的计算原理	223
18.4 连锁分析的特点与局限性	224
18.5 小结	225
<b>第 19 章 基因芯片数据分析方法与计算原理</b>	<b>226</b>
19.1 基因表达谱的概念	226
19.1.1 基因芯片	226
19.1.2 基因表达图谱与空间	226
19.1.3 基因表达数据的标准化	228
19.2 样品聚类分析方法	230
19.2.1 距离的定义	230
19.2.2 样品聚类分析原理与方法概述	231
19.2.3 样品聚类分析的 SAS 实现——CLUSTER 过程	233
19.3 判别分析	234
19.3.1 判别分析方法的种类	235
19.3.2 判别准则	237
19.4 变量聚类分析	239

19.4.1 变量聚类分析中相似系数的定义	239
19.4.2 变量聚类分析方法的概述	240
19.4.3 变量聚类分析的 SAS 实现——VARCLUS 过程	241
19.5 主成分分析	243
19.5.1 主成分分析方法与原理	243
19.5.2 主成分分析的 SAS 实现——PRINCOMP 过程	246
19.6 基因调控网络分析	249
19.7 小结	249
<b>第 20 章 物种遗传关系确定分析的计算原理</b>	<b>251</b>
20.1 分子标记技术的应用与基本原理	251
20.2 相似系数与距离的计算	252
20.2.1 根据 Nei 公式计算相似系数和遗传距离	252
20.2.2 计算 Jaccard 系数与距离	252
20.3 小结	252
<b>附录 1 胡良平统计学专著及配套软件简介</b>	<b>254</b>
<b>附录 2 <math>\chi^2</math> 分布临界值表</b>	<b>258</b>

# 第 1 篇

## 医学遗传统计实例分析与 SAS 实现

- 第 1 章 医学遗传资料数据结构与分析方法选择
- 第 2 章 Hardy-Weinberg 平衡定律验证的 SAS 实现
- 第 3 章 病例 - 对照研究关联分析的 SAS 实现
- 第 4 章 遗传分析结果校正和输出的 SAS 实现
- 第 5 章 连锁不平衡与单体型分析的 SAS 实现
- 第 6 章 近交系数和亲缘系数估算的 SAS 实现
- 第 7 章 遗传资料连锁分析的 SAS 实现
- 第 8 章 基因芯片数据分析的 SAS 实现
- 第 9 章 物种遗传关系确定分析的 SAS 实现
- 第 10 章 SAS/Genetics 模块概述





# 医学遗传资料数据结构与分析方法选择

在本书的开篇,我们详细地介绍一下常见的几种医学遗传资料的数据结构。数据结构对于认清数据的本质非常重要,同时,标准的数据结构也便于顺利地实现向标准统计软件(如 SAS 软件)中录入,从而便于后续的数据整理和分析。医学遗传数据大体分为一般人群实验数据、家系数据以及基因芯片数据等,在数据收集过程中,应尽可能地将数据整理或转化为本章所呈现的格式。

## 1.1 一般人群实验研究的数据结构

一般人群实验研究的数据,主要信息为个体的患病情况、位点基因型等,根据实验设计类型又可细分为下面几类。

### 1.1.1 位点基因型数据结构

这是一类最简单的数据结构,即将一般人群中个体的基本信息以及研究者所关心的基因或位点的基因型列入表中,成为数据库的形式,便于遗传资料信息的保存和提取。如表 1-1 所示。

表 1-1 可疑影响因素及位点基因型状态的数据结构

受试者编号	性别(M/F)	年龄	A16G	Gln27Glu	T164I	A523C
1	M	5	AA	GG	TT	CC
2	M	6	AA	GG	TT	CC
3	M	6	AA	GG	TT	CC
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>n</i>	M	44	AG	GG	TT	AA

对数据结构的分析:这种数据结构的录入格式不言自明,因为每行表达了一个受试者有关的全部信息,而每列代表一个变量。第 1 列为“编号”,仅表达数据所在行的位置;第 2 列