公共卫生研究中群随机试验设计与分析方法

〔加〕A.唐纳　N.克拉　著
刘　沛　译

科学出版社
www.sciencep.com

东南大学科技出版基金资助

公共卫生研究中群随机试验设计与分析方法

〔加〕A. 唐纳　N. 克拉　著

刘　沛　译

科学出版社

北京

图字:01-2005-6190 号

图书在版编目(CIP)数据

公共卫生研究中群随机试验设计与分析方法/(加)唐纳(Donner,A.),(加)克拉(Klar,N.)著;刘沛译.—北京:科学出版社,2006.6

ISBN 7-03-017198-5

Ⅰ.公… Ⅱ.①唐…②克…③刘… Ⅲ.公共卫生-统计试验法 Ⅳ.R1-33

中国版本图书馆 CIP 数据核字(2006)第 043728 号

责任编辑:王 晖/责任校对:张 琪
责任印制:刘士平/封面设计:黄 超

First published in Great Britain in 2000 by Arnold, a member of the Hodder Headline Group, 338 Euston Road, London NW1 3BH

ISBN:0-340-69153-0

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

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2006 年 6 月第 一 版 开本:B5(720×1000)
2006 年 6 月第一次印刷 印张:10 1/2
印数:1—2 000 字数:195 000

定价:39.00 元

(如有印装质量问题,我社负责调换〈新欣〉)

序

——为中文版而写

当我们的《公共卫生研究中群随机试验设计与分析方法》一书于2000年出版时，我们引用了Stuart Pocock教授在1996年说过的一段话："随着群随机试验的应用，特别是在发展中国家干预试验中应用的日益增加，我们需要对其方法学基础进行更深入的了解。"本书正是为了提供这一基础而写。

自我们的书出版以来，群随机试验的研究特别是在发展中国家的应用有了明显增加(Bland，2004)。我们仅以设在瑞士日内瓦的世界卫生组织(WHO)和设在韩国汉城的国际疫苗研究所(IVI)最近进行的两项以地理区域为单位的大规模群随机试验为例说明这一问题。

世界卫生组织开展的产前照料试验是为了比较两种产前照料方式对母亲和新生儿健康的影响(Villar等，2000)。世界卫生组织将泰国、古巴、阿根廷和沙特阿拉伯的一些产科诊所按照群随机试验的方法随机分配到新型护理模式组(试验组)和现行护理模式组(对照组)中。在经过严格的群随机对照试验后，结果表明：新型护理模式是有效和可行的。

国际疫苗研究所开展的伤寒疫苗人群试验是一个旨在控制传染性疾病的多学科协作研究项目。国际疫苗研究所将中国、巴基斯坦、越南和印度尼西亚等亚洲国家的一些地区随机分为试验组和对照组，应用群随机试验方法对伤寒疫苗进行效果评价。我们看到，Yang等最近(2005年)报道了在中国广西壮族自治区进行的以107个地理区域为群随机单位的伤寒疫苗效果评价的初步研究结果。目前，这一试验仍在进行中。

最后，我们衷心感谢东南大学公共卫生学院刘沛教授和科学出版社将我们的这本书介绍给中国同道。我们殷切期望本书对提高中国群随机试验研究的设计和数据分析水平能有所裨益。

Allan Donner，Neil Klar

2005年7月

译 者 序

第一次接触“群随机试验”(cluster randomization trails)是在2000年本书作者Allan Donner教授举办的专题系列讲座上。那时他刚刚校对完本书的清样,我正以访问学者身份在他任主任的西安大略大学流行病与生物统计学系研修。他对群随机试验这一有别于传统随机试验方法的精辟讲解和坚信这一方法将在公共卫生领域特别是在发展中国家的卫生干预试验中得到广泛应用的预言,使我萌发了学习这一方法并将这一方法介绍给我国同道的想法。短短几年时间,群随机试验迅速发展的事实就证明了他当时这一预言的正确性。这是促使我最终决定翻译此书的主要原因。

群随机试验是将研究对象以群体为单位,随机分配到不同组进行干预试验的方法。在目前见到的这类试验中,群随机单位多种多样,可大到社区、工厂、医院、学校、性倾向群体,小到邻里、车间、病房、家庭和运动队。

Fisher在创立实验设计理论时假定:在实验研究中,被随机化的单位就是分析单位,并在此基础上建立了一系列传统的统计分析方法。然而,群随机试验却并非如此,这一方法的随机单位是群体,而统计推断单位常常是个体。如在评价社区干预对重度吸烟者的戒烟效果时,随机单位是社区,而统计分析单位则为吸烟者的年龄、性别、教育水平、吸烟状态等个体指标。这种随机单位和统计推断单位间的不一致是群随机试验和传统随机试验间鲜明而重要的区别。而群随机试验和整群随机抽样调查的本质区别则为,后者是从确定的总体中随机抽取一定数量的群,但并不涉及将这些群随机分配到不同组进行干预试验。

与传统的个体随机试验相比,群随机试验由于能有效控制实验污染并可显著降低试验费用,近年来在国外得到了较为广泛的应用。然而,我国公共卫生工作者对这一方法的认识远不如对传统的个体随机试验那样深入。因此,介绍和引进这一方法对提高我国公共卫生研究水平具有现实意义。

作为国际上第一本系统介绍群随机试验方法的专著,本书具有两个显著的特点:一是系统性和完整性,第一至五章对群随机试验作了一般性介绍,包括群随机试验的基本概念、发展史、研究设计、样本含量估计、群随机试验涉及的伦理学问题等;第六至八章论述了群随机试验的统计分析方法,重点介绍了二项分类数据、计量数据、等级资料和随访资料的分析方法;最后(第九章),还详细介绍了群随机试验研究报告的撰写原则与注意事项。相信阅读过本书的读者将会对群随机试验有一个完整的认识。本书的另一显著特点是理论紧密联系实际,在阐述新的概念和

介绍统计方法时，引用了大量的公共卫生实例，这些实例对实际工作者具有重要的启发和引导作用。因此，本书既可作为公共卫生研究工作者的参考书，也可作为公共卫生专业学生的教科书。

在本书的翻译过程中，得到了东南大学及公共卫生学院各级领导、同事、学生们的关心和支持，特致以谢意。特别感谢 Allan Donner 教授给予的指导和帮助。感谢邹光勇博士仔细阅读了译稿并提出了许多宝贵意见。

由于译者水平有限，缺点和错误在所难免，敬请同行专家和广大读者批评指正。

刘　沛

2005 年 10 月于东南大学

前　言

群随机试验（cluster randomization trial）是指将一些完整的社会群体，而不是将单个的观察个体随机分配到不同干预组的研究方法。群随机试验有时也被称为成组随机试验（group randomization trail）。这一方法被广泛应用于评价生活方式、卫生照料制度、健康教育等非治疗性干预的效果。在这些研究中，群随机的单元多种多样，可以小到家庭、病房、班级，大到学校、医院、工厂乃至整个社区，也已见到将运动队（Walsch et al，1999）、部落（Glasgow et al，1995a）、宗教团体（Lasater et al，1997）和性倾向群体（Fontanet et al，1998）作为群随机单元的报道。

二十多年来，群随机试验在公共卫生研究中得到了日益广泛的应用，她已经形成并正在丰富着其特有的一套方法学。然而，这些研究成果大都散在于统计学、社会学、医学等不同学科的文献中。本书的目的是对这一方法进行系统而完整的介绍，为使用这一方法的研究工作者提供一本在研究设计和结果分析方面的参考书。本书也可作为公共卫生专业学生的教科书。本书的读者应具备初等生物统计学基础[Arimitage，Berry（1994）；Rosner（1995）]，并了解在一般教材[Friedman et al（1996）；Pocick（1983）]中所介绍的临床试验设计和结果分析的基本原则。

也许有人会问：在临床试验设计和结果分析方面已有许多优秀的教科书，是否需要一本专门的教科书用以介绍群随机试验？我们的观点是：在临床试验设计和结果分析的方法学方面，现有的教科书并不完善。在某些情况下，将现在教科书中介绍的方法应用于群随机试验，将会导致错误结论。需要强调的是，当研究单位由个体变为群体时，那些在一般教科书中介绍的方法已不能用于估计样本含量、分析试验数据和报告研究结果。

本书前四章是对群随机试验的一般性介绍。第一章为绪论；第二章为群随机试验发展史；第三章论述群随机试验设计；而第四章则重点讨论群随机试验所涉及的伦理问题。

本书第五章至第八章论述群随机试验的数据分析方法。在这些章节，我们通过大量实例说明群随机试验在实际工作中的应用，从而使本书适合于仅具有统计学初等知识的读者阅读。第五章论述样本含量估计（sample size estimation）；第六章和第七章 分别介绍了二项分类数据（binary ouycome data）和计量数据（quatitative outcome）的分析方法；第八章对群随机试验中新近出现的随访资料和多项分类数据分析方法进行了讨论。各章排列顺序和篇幅大小按其在群随机试验中的重要性进行安排。

本书最后一章(第九章)论述了群随机试验报告的撰写原则,该章也对以前各章所出现的重点问题进行了总结。

就群随机试验的方法学而言,每周都会出现一些新进展,因此,在写这本书时,我们所面临的问题是,本书应在何处停笔。例如,群随机试验的广泛应用,使有关这一方法的 Meta 分析(Meta-analysis)大量增加,这些 Meta 分析试图对来自不同群随机单元的试验结果进行综合分析(Fawzi et al, 1993; Rooney & Murray, 1996);关于群随机试验经济效益的评价方法在今后将得到更多的关注,这些问题已超出了本书论述的范围。我们深信,在不远的将来,在这些问题上将会出现重要的进展。

本书有关样本含量估计和数据统计分析所需要的软件已由日内瓦世界卫生组织的 Alain 先生和 Gilda Piaggio 博士编写完成。通过 Arnold 出版公司的网页:http://www.arnoldpublishers.com/support/cluster 可获得这一软件,这一软件包含了其他软件所没有的有关群随机试验的统计计算方法。

Pocock 在 1996 年指出:"随着群随机试验的应用,特别是在发展中国家干预试验中应用的日益增加,我们需要对她的方法学基础进行更深入的了解。"本书正是为了提供这一基础而写。

目　录

第一章 绪 论

在本章中，我们将重点论述群随机试验的基本概念及其特征。首先，研究者应了解，在样本含量相等的条件下，群随机试验统计推断的效能低于个体随机试验，这个问题将在第一节中讨论。其次是群随机对试验方法学的影响，这些影响既涉及样本含量估计和试验结果分析，也涉及常用设计方法（如分层设计、配对设计）的应用，这些问题将在第二节和第三节论述。第四节说明为什么要进行随机试验；第五节介绍推断单位对研究设计和结果分析的影响；第六节解释我们采用“群随机”（cluster randomization）这一术语而不采用有关这一试验的其他术语的原因。

第一节 群随机试验及其特征

Cornfield 于 1978 年就明确提出：群随机试验的统计效能低于个体随机试验。这一观点引起了公共卫生研究工作者的广泛关注。然而，早在 1947 年 Walsh 就提出了这一观点（Walsh, 1947）。统计效能降低的原因是：同一群内个体的反应比不同群间个体的反应有更大的相似性。Mckinlay 等（1989）指出：这也许是为什么当 Ronald Fisher 爵士（Fisher, 1935）创立实验设计理论时就假定：毫无例外，被随机的试验单元也就是分析单元。然而，在群随机试验中，情况并不一定如此，这是群随机试验的一个重要特征。

通常，群内各个体间对某一干预反应的相似程度可用一个被称作群内相关系数（intracluster correlation coefficient）的参数来度量，我们用希腊字母 ρ 来表示这个参数，它可被解释为同一群内任何两个反应间的标准 Pearson 相关系数。$\rho>0$ 表示群间变异大于群内变异，在这样的情况下，我们可以说这个试验具有“群间变异”（between-cluster variation）的特征。用 Whiting-O'Keefe 等（1984）建议的术语这也等同于说：就实验结果而言，各个群不能被认为是“可互换”（interchangeable）的，缺乏互换性或出现群间变异的基本原因随试验的不同而不同，但在实际工作中，可有如下原因：

（1）群内观察个体的选择可造成群间变异。例如：在群单元为医疗单位的群随机试验中，选择到某一医疗单位就诊的病人，其特征可能与他们医生的年龄和性别有关，而病人的特征常常与治疗的效果有关。这样，由于病人对医疗单位的选择，在这些医疗单位间就产生了群间变异。Rhee 等（1980）指出：被同一个医生治疗的病人，治疗效果与那位医生的治疗方式有关。相似的情况也可出现在以工作

场所为单元的群随机试验中，在这类试验中，雇员和雇主根据某些特性，如社会经济状况或道德观念等进行双向选择(Kelder et al, 1993)。在以较大社区或县为单元的群随机试验中，地理因素也可能影响群内观察对象的选取，如老年人或患有呼吸道疾病的病人喜欢选择住在温暖地区。Koepsell(1998)指出：人们选择居住在某一社区往往是因为他们具有和那一社区其他居民所共有的“适合于”他们居住的特征，这些特征常常与他们所喜欢的健康行为有关。

(2) 群水平协变量的影响可产生群间变异。例如：在以外科医生为单元的群随机试验中，不同外科医生治疗的病人可在手术效果、手术并发症和术后死亡率上出现差别(McArdle & Hole, 1991)。Divine 等(1992)认为：因为医疗态度和医学知识上的差别，不同医疗工作者的医疗质量具有明显差别。Bland 和 Kerry(1997)注意到：在卫生照料试验中，群效应对研究结果可能有显著的影响，这是因为，在这些试验中，干预被应用到卫生照料的提供者而并非直接指向他们的病人。在这些试验中，不同卫生照料者之间的差异和不同病人间的差异共同构成了群间变异。为了说明这一观点，他们引用了由 White 等(1987)所报告的试验，这一试验是为了在社区医疗中，通过“内科医生教育项目”来改进哮喘病的治疗效果。

某些重要的环境特征也可产生群效应。例如，在托儿所内传染病的发生可能与该所的温度或其他环境条件有关；而不同地区的地方法规则能影响禁烟项目的成功率。Rice 和 Leyland(1996)指出：同一医院收治的病人可能受到某些共同因素，如病床周转率的影响。另外，当我们把家庭或家族作为随机单元时，环境和遗传的联合效应将造成群间变异。

(3) 传染病在家庭内比在家庭外传播更快的特征可造成群间变异。这是因为，传染病在某些群体内的流行率和暴发率、致病原的传播方式、致病原的毒力等均可造成不同人群间发病率之差异。

(4) 群内个体的相互影响可造成群间变异。例如，对同一组内研究对象进行健康教育，则可使群内个体共享信息相互影响而使其具有某种相同的倾向性，从而产生群效应。正如 Koepsell 所说：正像传染病的致病原能从一个人传播给另一个人一样，在那些密切接触的人群中，人们的行为、道德规范、处事方式也因相互影响具有一定的相似性。

没有丰富的经验，则很难分辨产生群间变异的原因。然而，无论何种原因，群间变异都会降低试验有效的样本含量。群越大，ρ 值越大，有效样本含量的减少就越明显，有效样本含量的减少导致研究效率的下降。

既然与个体随机试验相比，群随机试验的研究效率较低，那么为什么还要采用这一试验方法呢？常见的原因有三点：① 满足伦理学要求；② 降低研究费用；③ 控制实验污染(experimental contamination)。从现有文献看，控制实验污染是最常见的原因。例如，Sommer 等(1986)报道了补充维生素 A对儿童死亡率影响的

群随机试验。在他们的试验中,印度尼西亚的450个村庄被随机分配到维生素A补充组和对照组。他们采用群随机设计是因为从实际情况出发,对个体进行随机是不可行的。他们也注意到,把整个村庄作为随机单元将避免由随机个体所引起的实验污染,因为在这个试验中,若将同一村庄的村民一部分随机分配到服用维生素A的试验组,另一部分随机分配到服用安慰剂的对照组,则很可能发生试验组误服安慰剂,对照组误服维生素A的实验污染。

以医院或工作场所为单位的群随机试验也存在实验污染问题。例如,在一个预防冠心病的多工厂试验中(Rose,1970;世界卫生组织欧洲合作组,1986),把工厂作为随机单元是为了控制实验污染,也就是防止在不同的干预组内,某一组的观察对象受到另一组预防冠心病指导信息的影响。另一个例子是,产后立即给婴儿哺乳是否能降低产后出血的发生率(Bullough et al,1989),在这个试验中,接生员是群随机单元,研究者将接生员而不是产妇随机分配到干预组(早期哺乳组)和对照组,因为待分娩的产妇对早期哺乳的可能好处已有所了解,因此,这样做将减少干预组被污染的可能性。当一个研究人员既要负责试验组的观察对象又要负责对照组的观察对象时,特别容易发生这样的污染。在这个试验中,对群(接生员)随机而不是对个体(产妇)随机的另一个原因是,研究者担心接生员们并不理解随机化的意义和方法,因此很难保证她们能把她们负责接生的产妇进行随机分组。

权衡群随机试验所造成的研究效率降低和个体随机试验可能产生的实验污染之间的利弊是决定是否采用群随机设计的一个重要因素。Slymen和Hovell(1997)对这一问题进行了研究,给出了对群随机设计和个体随机设计的得失进行定量分析的方法。他们认为:对ρ值、群的大小和实验污染程度的综合分析是决定采用哪种设计方法的主要因素。例如,在对实验污染进行定量分析时,可采用下面的方法,当污染减弱了干预效果时,研究者可以通过对照组中有多大比例的个体其"行为"好像是受了只有试验组个体才能接受到的干预措施的影响,从而定量分析实验污染的大小。

当然,采用群随机设计并不能保证完全消除实验污染。然而,当将不同的地区作为试验的群时,选择具有明确地理分界的群体作为观察单位常常会降低实验污染。将群的范围增大,可进一步降低实验污染,然而这样做将会降低有效样本含量而进一步降低研究效能。Grosskurth等(1995)的试验说明了这一问题,他们研究了改进卫生服务对坦桑尼亚某一地区艾滋病病毒(HIV)感染率的影响,在这个试验中,他们仅仅选择那些较大的社区作为研究的群,在每一个群内,其人群的卫生服务仅由同一个卫生中心和它的下级机构负责,因此减少了如果在一些毗邻地区设置实验组和对照组可能发生的污染。另一个例子是COMMIT试验(Gail et al,1992),在设置配对社区时,他们既考虑到社区间的距离不要太远,以免增加调查、监测和干预上的难度,又不宜靠得太近,以免在试验社区中进行的健康教育活动影

响到不开展健康教育活动的对照社区，从而有效降低了实验污染。

另外，也可通过设立外对照的方法评价实验污染。所谓外对照，是在原有试验组和对照组以外的某个人群设立的对照。Schlegel(1977)采用这一方法评价了预防青少年饮酒和被虐待的健康教育项目。他将两个学校中的所有班级均随机分配到不同的干预组，为了避免一种称之为"处理溢出"(treatment spillage)的实验污染，他在未开展这一教育项目的另一所学校选取了两个班级作为外对照。

Bass 等(1986)应用群随机设计研究高血压筛检效果时发现，如果病例只来自一个医疗单位，那么最终的研究结果将没有推广价值，因此他们决定采用群随机设计，在试验组和对照组中各随机分配了 17 个医疗单位。相似的，在有关孕期照料的研究中，应该将产科医生、诊所或医院作为群随机单元。

在上述研究中，随机分配整个医疗单位，大大简化了试验的组织工作，同时也避免了可能出现的污染。例如，当一部分医生或其他医务人员知道了干预试验的内容后，将会有意或无意地影响到对对照组的观察。对同一个医生来说，他对接受新疗法的干预组病人比对接受一般疗法的对照组病人将会给予更多的关注和照料。因此，当干预是非盲法时，与随机分配个体有关的实验污染的危险性也许是非常大的。

下面的问题出现在爱丁堡(Edinburgh)乳腺癌试验中。研究者随机选取社区医疗单位而不是个体病人作为他们的实验组和对照组(Alexander et al, 1989)。除了事先无法得到完整的病人名单用于抽样的实际问题外，他们也担心随机选取病人将会产生实验污染。在实验设计时，他们认为，如果对照组妇女知道试验组妇女可得到定期医学检查后，她们也将要求医生对她们进行定期医学检查。更重要的原因是，研究人员认为在这个调查中对每个妇女进行随机分配根本就不会被全科医生(general practitioner)和他们的病人所接受。Paci 和 Alexander(1997)更加强烈地坚持这一观点，他们认为：被邀请来参加乳腺癌筛检研究的妇女将希望与她们的全科医生(GP)讨论她们是否能够得到定期医学检查，假如一个全科医生的病人仅仅有一半被随机安排接受检查，另一半不安排检查，那么，对这位全科医生来说，这样做的结果必将导致另一半妇女对他不满，因而在这种情况下，对个体随机是不现实的。

另外，一些现实问题也会影响对群随机设计的选择。例如：Farr 等(1988)研究了某种滴鼻剂和安慰剂对降低呼吸道疾病发病率的效果，在这个研究中，随机分配的单元是家庭。将家庭而不是个人进行随机安排是为了提高顺从性，他们认为将家庭成员安排在同一治疗组内将会提高顺从性。

在传染病的人群研究中，常常需要采用群随机设计。因为，根据免疫学原理，有些没有注射疫苗的人，由于群体免疫(herd immunity)的效应而得到了免疫保护(Comstock, 1978；Hayes, 1998；Smith & Morrow, 1991)，在减毒活疫苗的人群试

验中这一现象尤其突出(Pollock,1966),因此,对这类试验,随机分配个体是不合适的。

在以上的例子中,研究者选择群随机设计是为了避免试验过程可能发生的伦理、可行性或方法学的问题。但是,在某些研究中,采用群随机设计只是一种必然选择,并不出自特别的考虑。Dunn(1994)指出:"在引进一种新的、更完善的卫生服务体系或改变某种卫生管理方法时,我们不可能使一部分人使用新方法而另一部分人仍保持老方法"。上面提到的 HIV 干预研究(Grosskurth et al,1995)就雄辩地说明了这一点,正如作者指出的那样,必须以社区为单位进行随机化,因为这一干预涉及向研究对象提供更好的卫生服务,这种更好的卫生服务必须使用到某一社区的整个人口。一个被称为 CATCH 的大规模预防心血管疾病的社区卫生研究,是这方面的又一个例子(Zucker et al,1995)。这个调查的研究者将学校随机分配到不同的干预组是因为,"CATCH 的干预最终要被应用到以学校为单位的学生中去"。

大规模健康教育的干预项目与上面的情况相似,因为一般情况下我们不能在同一社区内对一些人给予饮食、运动、戒烟等指导,而对另一些人不给予指导,在重度吸烟者中加快戒烟速度和降低吸烟率的 COMMIT 试验(COMMIT 研究组,1995)就是这方面的一个例子。在这个试验中,为配合健康教育和媒体宣传活动,卫生照料提供者和雇主被鼓励参与戒烟宣传活动。Gail 等(1996)指出,这种以社区为基础的干预对社区内的每个吸烟者都会有影响,所以不能使用个体随机的方法。Byau(1998)更明确地指出:"我们知道,使用通常的个体随机研究方法时,重度吸烟者为了戒烟将面临许多困难;我们采用群随机试验是基于下面的考虑:当重度吸烟者发现吸烟几乎不能被社会接受,而戒烟又能得到来自社会各方面的帮助,戒烟对他们就变得容易了"。关于干预必须应用到社区或某一特定区域时采用群随机试验的一些例子可见 Smith & Morrow(1991)和 Raudenbush(1997)的研究。

第二节 群随机对实验设计和结果分析的影响

上一节我们已指出,Fisher 的经典实验设计理论认为:毫无例外,被随机的实验单元也就是分析单元。而群随机试验对这一理论提出了挑战,因为这一方法的随机化单元是群,而统计推断单元则常常是个体。假如我们把统计推断只局限在群水平上,那么我们至少可以在样本含量估计和统计方法应用方面认为群随机试验和一般个体随机的临床试验相同。假如我们要将统计推断应用到个体水平,由于群内各个体间存在相互影响,是非独立的,这样就不能采用一般的统计方法进行样本含量估计和数据分析。一般地说,使用通常的样本含量估计公式将低估所需的样本含量,结果往往是抽取的研究对象太少,研究的把握度太低而不能得出预期

结果。另一方面,应用一般的统计方法进行数据分析,意味着我们假定没有群间变异,这将导致一个 P 值偏小的分析性偏性,得到一个虚假的统计显著性结果。在流行病学文献中,Gounfield(1978)就这一问题提出了一个著名的观点:"将适合于分析个体随机试验的统计方法应用于分析群随机试验实际上是在玩弄一种自我欺骗的游戏,这样的分析结果不应被接受"。相似的观点也经常出现于教育学文献中(Zucker,1990)。但是,如果研究者希望在数据分析时得到一个干预有效的结论,或者是对群随机试验的统计方法了解甚少,则难以抵制这种自我欺骗的诱惑。在此,我们想要强调的是:将适合于分析个体随机试验的统计方法应用于分析群随机试验是十分错误的。事实上,这也是在文献上许多试验被报告在统计上差别有显著性意义,而实际上这种差别并没有统计学意义的一个主要原因。显然我们也可以说,使用估计随机个体试验样本含量的公式去估计群随机试验所需的样本含量也是一种自我欺骗,因为前者可导致Ⅰ型错误(type Ⅰ error)的显著增大,而后者则可显著增大Ⅱ型错误(type Ⅱ error)。

不幸的是,许多群随机试验在设计和分析阶段并没有考虑群间变异。例如:Simpson 等(1995)对 21 个有关初级预防保健的群随机试验进行综述后发现,在样本含量或统计效能计算时考虑群间变异的仅仅只有 4 个试验(19%),在结果分析时考虑群间变异的也只有 12 个试验(57%)。Donner 等(1990) 的研究与 Simpson 等(1995)的研究非常相似,Donner 等在对 16 个群随机试验进行综述时发现,仅仅有 3 个试验在估计样本含量时考虑了群间变异(19%),8 个试验在结果分析时考虑了群间变异(50%)。其他一些关于群随机试验方法学的综述性论文(Butler & Bachmann,1996;Rooney & Murray,1996)也显示了相似的令人沮丧的结果。

有些研究者认为,群间变异在理论上固然重要,但在实际工作中却无关紧要,这可能是他们在实验设计和结果分析时没有考虑群间变异的原因。例如,在一个以家庭为单元将儿童随机分配到给予维生素 A 补充的试验组和不给予维生素 A 补充的对照组的研究中,研究者未考虑儿童死亡率在家庭间的差别是根据"这一预期的差别可以忽略不计"。我们认为,除非有强有力的证据,否则在实际工作中随意做出这样的假定是很危险的。对这一问题,我们将在第三节中进行更深入的讨论。

虽然采用群随机化主要影响样本含量估计和统计分析方法,然而它也影响临床试验的执行和研究结果的解释。这些问题包括分层(stratification)的作用,对象的盲法(subject blindness),知情同意(informed consent),以及随访和失访等一些特殊问题,对此,我们将在第三章第五节中予以深入讨论。

第三节 定量群效应

为了比较两组的均数(mean),将每群具有 m 个个体的 k 个群随机分配到试验

组和对照组，假定应变量 Y 服从正态分布且未知的总体方差 σ^2 相等，把试验组样本均数 $\bar{Y}_1$ 和对照组样本均数 $\bar{Y}_2$ 作为其总体均数 μ_1 和 μ_2 的点估计值。从著名的群抽样方法(Moser & Kalton,1972)可知，其样本均数的方差是：

$$Var(\bar{Y}_i) = \frac{\sigma^2}{km}[1 + (m-1)\rho] \quad i = 1,2 \tag{1.1}$$

这里，ρ 是定义在第一节中的群内相关系数。若 $\rho=0$，则方程(1.1)简化为个体抽样的样本均数方差公式。若用 P 代表事件的发生率，用 $P(1-P)$ 代替 σ^2，公式 1.1 也就是群抽样中样本率方差的计算公式。

根据公式 1.1 可知，要获得与随机分配 km 个个体到不存在群效应的各个组相同的统计效能，在计算样本含量时应在原来的公式上乘方差膨胀因子(variance inflation factor, IF)，$IF=1+(m-1)\rho$。这个公式也经常出现在抽样调查的文献中，在这些文献中方差膨胀因子被称为“设计效应”(design effect)(Kish,1965)。在方差膨胀因子中，群内相关系数 ρ 是一个重要参数，若 $\rho=0$，则表示群内各个体间在统计上相互独立；若 $\rho=1$，则表示群内各个体完全相关，或者说一个群内各个体的反应是完全相同的，也就是说整个群所提供的信息量与这个群内某个个体所提供的信息量相同，这被称为“有效群样本含量”(effective cluster size)为 1。有效群样本含量可用公式 $m/[1+(m-1)\rho]=m/IF$ 来表示，当各群样本含量不等时，常用群样本含量的平均数 $\bar{m}$ 代替 IF 中的 m；用 $\bar{m}$ 代替 m 计算出的 $Var(\bar{Y}_i)$ 偏小，但当各群样本含量相差不大时，这种偏小是微不足道的。

如上节所述，参数 ρ 可被理解为在同一群内的任两个观察对象间的相关系数，假如我们进一步假定群内相关系数大于等于零，那么也可以说 ρ 等于群间方差占总方差的比重，即 $\rho=\sigma_A^2/\sigma^2=\sigma_A^2/(\sigma_A^2+\sigma_W^2)$，式中 σ^2 代表总方差，σ_A^2 代表总方差的群间变异组分，σ_W^2 代表总方差的群内组分。第 i 组均数的方差则可写成 $Var(\bar{Y}_i)=(\sigma_A^2+\sigma_W^2/m)/k$。请注意，我们也经常写 $\sigma_W^2=\sigma^2(1-\rho)$，这个式子说明，通过乘因子 $1-\rho$，群内方差 σ_W^2 将随着群内相关系数 ρ 的增加而减小，群内方差的减小则意味着同一群内各个体间相关性的增加。

设 k 为每组的群数，m 为每群的个体数，MSC 为群间均方误差(mean square error)，MSW 为群内均方误差。使用一般的方差分析(analysis of variance)公式可计算出 ρ 的估计值为(Snedecor & Cochran,1989)：

$$\hat{\rho} = \frac{MSC - MSW}{MSC + (m-1)MSW} = S_A^2/(S_A^2 + S_W^2) \tag{1.2}$$

式 1.2 中，$S_A^2=(MSC-MSW)/m$ 和 $S_W^2=MSW$ 分别是 σ_A^2 和 σ_W^2 的样本估计值。

当各群的样本含量不等时，若计算 $\hat{\rho}$，则须用每群调整个体数 m_0 代替公式