主 编 李 卫

医疗器械临床试验统计方法

YILIAO QIXIE LINCHUANG SHIYAN TONGJI FANGFA

人民軍醫出版社
PEOPLE'S MILITARY MEDICAL PRESS

医疗器械临床试验统计方法

YILIAO QIXIE LINCHUANG SHIYAN
TONGJI FANGFA

ISBN 978-7-5091-5870-8
9 787509 158708 >

› 策划编辑　秦速励
› 封面设计　龙　岩

› 销售分类　基础医学

定价：55.00元

医疗器械临床试验统计方法

YILIAO QIXIE LINCHUANG SHIYAN TONGJI FANGFA

主　编　李　卫　中国医学科学院阜外心血管病医院

副主编　李　宁　美国约翰霍普金斯大学

周晓华　美国华盛顿大学

编　者　王　杨　中国医学科学院阜外心血管病医院

郭　晋　中国医学科学院阜外心血管病医院

陈　涛　中国医学科学院阜外心血管病医院

唐欣然　中国医学科学院阜外心血管病医院

姜宇莉　中国医学科学院阜外心血管病医院

吴振强　安徽医科大学流行病与卫生统计系

黄耀华　中国医学科学院阜外心血管病医院

王嘉琪　佳木斯医学院流行病与卫生统计系

胡　泊　河北联合大学流行病与卫生统计系

孙叶桓　安徽医科大学流行病与卫生统计系

贾　宣　中国医学科学阜外心血管病医院

成小如　中国医学科学院阜外心血管病医院

孙　毅　中国医学科学院阜外心血管病医院

人民軍醫出版社

PEOPLE'S MILITARY MEDICAL PRESS

北　京

图书在版编目(CIP)数据

医疗器械临床试验统计方法/李　卫主编．—北京：人民军医出版社，2012.9
ISBN 978-7-5091-5870-8

Ⅰ.①医…　Ⅱ.①李…　Ⅲ.①医疗器械—临床医学—试验—统计方法　Ⅳ.①R197.39—32

中国版本图书馆 CIP 数据核字(2012)第 181950 号

策划编辑：秦速励　**文字编辑**：　伦踪启　卢紫晔　**责任审读**：吴铁双
出版发行：人民军医出版社　　**经销**：新华书店
通信地址：北京市 100036 信箱 188 分箱　　**邮编**：100036
质量反馈电话：(010)51927290；(010)51927283
邮购电话：(010)51927252
策划编辑电话：(010)51927286
网址：www.pmmp.com.cn

印、装：北京华正印刷有限公司
开本：787mm×1092mm　1/16
印张：16　**字数**：382 千字
版、印次：2012 年 9 月第 1 版第 1 次印刷
印数：0001—3000
定价：55.00 元

作者简介

李卫，研究员，博士生导师。中国医学科学院阜外心血管病医院教授，国家心血管病中心医学研究统计中心主任，香港中文大学客座教授，中国医师协会循证医学分会理事，中国高血压联盟理事、中华医学会系列杂志审稿专家，中国新药杂志审稿专家、美国临床流行病学杂志审稿专家、美国统计学会及药物信息学会会员、美国临床试验学会会员、中国临床试验生物统计学组成员、中国卫生统计学会统计理论与方法专业委员会常务委员。国家食品药品监督管理局（SFDA）药物及医疗器械临床试验评审专家，国家科委及北京市科委评审专家。

长期致力于药物和医疗器械临床试验研究设计及统计方法学研究，并为政府、国内外企事业单位及大专院校从事临床研究相关人员，提供临床试验研究设计及统计方法学咨询服务。作为中国区负责人，负责国际多中心"前瞻性城乡流行病学（PURE）研究"项目；作为统计部分责任人，参加多项国家重大课题和数十项国内外多中心临床研究课题。在国内外 SCI 收录杂志上发表第一作者、通讯作者署名的英文文章数十篇，参与撰写论著6 部。

内容提要

本书介绍了医疗器械临床试验中涉及的试验设计方法、统计分析方法、计算原理以及软件实现过程与相关程序。书中除将复杂的统计计算原理结合医学语言详细地讲解之外，每种方法均给出医疗器械试验的实例及有针对性的分析。本书第1及第2章介绍了医疗器械临床试验的管理规范、方案制定原则，以及国家食品药品监督管理局要求的与统计相关的前期准备工作。第3章提纲挈领地介绍了器械临床试验的主要内容，包括数据管理、试验设计、统计分析，数据质控等。第4章详细介绍了多种样本量设计的方法，几乎涵盖了所有器械试验。第5章详细介绍了器械临床试验设计的原理。第6—17章，由简至繁地介绍了各种统计分析方法，每种方法均给出了相关的器械试验的实例，并附带了SAS的计算程序，便于读者理解操作。最后两章简要介绍了器械试验中可能涉及的统计分析方法。本书可作为基础医学、临床医学科研人员以及医疗器械生产厂家相关技术人员的工具书，亦可作为医学院校本科生、科研型研究生教学之用的参考书。

序　言

近年来，我国医疗器械和器材的研制和生产取得重大进展，随之而来的医疗器械临床试验领域也有了跨越式的发展，需要越来越多的高水平的专业人员和相关行业的从业人员，例如生物统计师、医生、护士、临床监察员等参加到这项工作中来。但是，由于其专业性极强，当前从事本领域研究的人才匮乏，人才的培养需要一个过程，然而至今尚没有一本适用的综合性书籍能够为这一领域的从业人员提供参考和指导，使新进入该行业的人员入门不易，在一定程度上制约了我国医疗器械临床试验的快速发展。本人作为我国第一批从事医疗器械临床试验的专业工作者，对此深感忧虑。

本书主编李卫教授是资深生物统计学专家，为最早介入我国医疗器械临床试验领域并参与指导原则制订、指导或负责临床试验研究设计及产品评审的专家之一。本书作者均为在医疗器械临床试验领域有丰富经验并工作在第一线的生物统计学专家或临床试验管理人员。作者们本着为工作在一线的研究人员或管理者提供有价值的参考资料，为行业培养新鲜血液为目的，结合多年的实际工作经验编写了这本书。

本书与以往的关于生物统计的书籍相比，有如下鲜明特色。

首先，具有明确的针对性。本书内容主要针对医疗器械临床试验。当前介绍临床试验的书很多，却很难找到一本涉及医疗器械临床试验方面的书籍。器械研究作为临床试验中很重要的一部分，医护人员、企业研究开发人员都需要掌握一定专门知识和技能，本书的实时出版正好满足了这一需要。

其次，具有广泛的实用性。书中的例子都是作者在实际工作中收集、整理的，可以为读者在实际工作中提供有实用价值的参考和借鉴，做到有的放矢。本书也涵盖了医疗器械临床试验的方方面面，包括最新的法规、流程、操作、试验设计、常用统计学方法及合理应用等，为读者展示了一幅医疗器械临床试验的全景图。

最后，有切实的可操作性。本书大部分章节都附有 SAS 程序。做到理论与实践相结合，让读者可以亲身体验其中的魅力，实时操作，使系统的理论变成鲜活的实例，亲身体验整个过程，体验解决问题的快乐。

本书内容丰富、充实，层次分明，深入浅出，理论与实践紧密结合，是一部难得的实用参考

书。本书可供从事医疗器械临床试验的各类研究人员、医生、生物统计学工作者、临床试验监察员、评审专家及制订和实施相关政策法规的管理人员借鉴参考，并可为有志于从事医疗器械临床试验研究的年轻科研工作者、本科生及研究生打开一扇知识的大门。

希望本书的出版能为进一步提高我国医疗器械临床试验的水平及规范化作出贡献！

于中国医学科学院　阜外心血管病医院

前　言

按照国家食品药品监督管理局5号令(二〇〇四年一月十七日)定义,医疗器械临床试验是指:获得医疗器械临床试验资格的医疗机构对申请注册的医疗器械在正常使用条件下按照规定在人体进行的试用或验证过程。其目的是评价受试产品是否具有预期的安全性和有效性。

由于临床试验通常是根据研究的目的,通过样本来研究器械对疾病及其预后等方面的作用,进而推论至整个的目标人群,指导真实的临床实践,因此,一个好的医疗器械临床试验应能够提供最客观的安全性和有效性评价,否则无法将来自样本的研究结论推广至总体。生物统计学是一门能够对试验相关因素做出合理的、有效的安排,并最大限度地控制试验误差,提高试验质量,并对试验结果进行科学合理分析的学科。

因此,生物统计学在医疗器械临床试验的全过程中有着不可缺少的重要作用,只有当研究结果既具有临床意义、又具有统计学意义时,才能够应用到真实的临床中,指导临床实践。

本书充分考虑了我国医疗器械临床研究的现状,参考了ICH E9文件以及美国食品药品监督管理局(FDA)的医疗器械现行统计学指导原则,以临床试验的基本要求和统计学原理为重点,从生物统计学角度为医疗器械临床试验过程提供了一个全面的实施办法,包含了对医疗器械临床试验的总体考虑以及试验设计时、试验实施过程中及试验结果分析和报告时的统计学问题,旨在为医疗器械注册申请人和临床试验的研究者在整个临床试验过程中如何进行设计、实施、质量控制、分析和安全性和有效性评价提供指导,以期保证医疗器械临床试验的科学性、严谨性和规范性,并力求符合中国国情,使之具有充分的可操作性。本书第一、二章介绍了医疗器械临床试验的管理规范、方案制定原则,以及国家食品药品监督管理局要求的与统计相关的前期准备工作。第三章提纲挈领的介绍了器械临床试验的主要内容,包括:数据管理,试验设计,统计分析,数据质控等。第四章详细地介绍了多种样本量设计的方法,几乎涵盖了所有医疗器械临床试验。第五章详细介绍了器械临床试验设计的原理。第六至十七章,由简至繁的介绍了各种统计分析方法,每种方法均给出了相关的器械试验的实例,并附带了相应的SAS计算程序,便于读者操作理解。最后两章简要介绍了器械试验中可能涉及的统计分析方法。

笔者长期致力于药物和医疗器械临床试验研究设计及统计方法学研究，并为政府、国内外企事业单位以及大专院校临床研究相关人员提供临床试验研究设计和统计方法学咨询服务。作为中方负责人，负责国际多中心“前瞻性城乡流行病学(PURE)研究”项目；作为统计部分负责任人，参加多项国家重大课题和数十项国内外临床试验研究课题。在国内外 SCI 收录杂志上发表第一作者、通讯作者署名的英文文章数十篇，参与撰写论著 6 部。每年培训政府、企事业单位临床研究相关人员、以及大专院校学生数千人次。如果读者有进一步的学习和需求，如器械临床试验研究设计、统计分析、科研课题合作研究以及临床研究相关人员培训，可以联系 liwei@mrbc-nccd.com。

在本书即将出版之际，笔者真挚的感谢为本书做出贡献的一直以来致力于器械临床试验工作的高等院校和科研院所的教授、医生、青年学者。感谢国家心血管病中心、中国医学科学院阜外心血管病医院原院所长、中国工程院院士高润霖教授为本书欣然作序；感谢美国约翰霍普金斯大学李宁教授、华盛顿大学周晓华教授为本书审稿所付出的辛勤劳动；感谢安徽医科大学孙业桓教授、河北联合大学胡泊副教授、佳木斯医学院王嘉琪老师对本书部分章节的撰写和校改；感谢北京协和医学院阜外心血管病医院博士生郭晋、陈涛以及国家心血管病中心医学研究统计中心统计部主管王杨以及他的团队成员唐欣然、姜宇莉、黄耀华和吴振强等人对本书的撰写和反复修改；感谢人民军医出版社秦速励编辑的大力支持与热忱帮助。正是由于全体编委会同仁的积极参与、不懈努力与真心奉献，才使这本书能够顺利出版。

由于笔者水平有限，书中难免出现个别不成熟的想法，甚至谬误，恳请广大读者不吝赐教，以便再版时修正。

主编　李　卫

国家心脏病中心医学研究统计中心
北京协和医学院阜外心血管病医院

目　录

第 1 章 医疗器械临床试验管理规范

医疗器械是从事临床医疗诊断与治疗的必要设备，是发展现代临床医学的重要基础，直接关系到人体健康。因此，医疗器械必须经过严格的安全性和有效性的评价方可上市，用于疾病的诊断与治疗。

医疗器械临床试验是指在由国家食品药品监督管理局(SFDA)认可具有临床试验资质的医疗机构、按照一定的期限和病例数量要求、对医疗器械的安全性和有效性进行评价的活动。临床试验的特征是根据有限的病人样本得出研究结果，对未来具有类似情况的病人总体作出统计学推断，以确认某一被试产品是否具有预期的安全性和有效性。由于临床试验直接涉及病人，在考虑临床试验的科学性的同时，还应更多地考虑到受试者的权益。因此，医疗器械临床试验既有法规方面的考虑，又有科学性和伦理学方面的考虑。

我国在2000年颁布的《医疗器械监督管理条例》，继而公布的《医疗器械注册管理办法》《医疗器械临床试验规定》(局令第5号)等法律法规明确规定了第二类、第三类医疗器械准入市场前应当通过临床试验，同时提出了医疗器械临床试验的基本要求、具体办法及实施细则，从而规范了医疗器械的临床试验。本章主要叙述国内外医疗器械临床试验管理和关于医疗器械注册时对临床试验资料的要求等规定，并重点就医疗器械临床试验设计中的统计学问题进行讨论，使得企事业单位法规注册部门及医疗器械临床试验工作者能基本了解医疗器械注册在临床试验管理方面的要求。

一、美国医疗器械临床试验管理

美国是国际上最早对医疗器械临床试验进行规范化的国家。美国食品药品监督管理局(FDA)美国食品、药品和化妆品法520(g)条和医疗器械安全法有“研究器械豁免”IDE(Investigational Device Exemption)法规，对医疗器械临床研究提出了要求。IDE是美国食品药品管理局(FDA)对医疗器械进行上市前审批(PMA)和510(K)审查过程中一个重要环节，它涵盖了医疗器械临床研究(Clinical Investigation Clinical Trial)的规定。IDE要求通过实施临床研究获得产品的安全性和有效性资料。

(一)临床研究法规

在IDE法规中对临床研究进行了规范，明确了哪些医疗器械需要进行临床研究，以及如何进行临床研究。在IDE法规要求下，制订了良好临床实践规范(Good Clinical Practice，GCP)，进一步明确了如何进行临床研究。随后，欧共体和日本也相应制订了与美国类似的GCP原则。IDE法规含有如下内容：①有重大风险及无重大风险的医疗器械临床研究的要求；②如何完成IDE申请；③保护受试者权益，建立伦理委员会(Ethics Committees，EC)或学术审查委员会(Institutional Review Boards，IRBs)；④规定申办者、监查员和研究者的职责。

在执行 IDE 法规时，还需要特别注意：在临床研究中需要有科学的证据支持器械的有效性；需要有很好的病历档案；需要有准备上市器械的使用说明，如手术方法及其经验。

(二)临床研究分类和要求

按 IDE 要求，任何一个医疗器械在进入临床研究前都要向 FDA 进行申请。也就是说在美国不经过适当的 FDA IDE 要求的程序及审批，就不能进行医疗器械的临床研究。图 1-1 是 IDE 对不同风险医疗器械进行临床研究的分类要求。表 1-1 是 IDE 对医疗器械临床研究的要求。

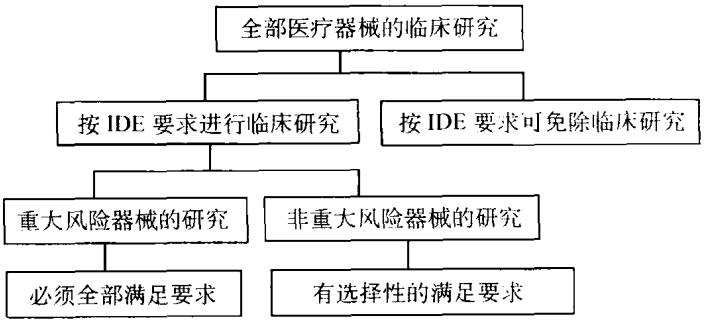

图 1-1　进行临床研究的医疗器械分类

表 1-1　IDE 对医疗器械临床研究的要求

非重大风险的医疗器械	重大风险的医疗器械
在器械的标签上必须标明名称、生产厂家地址、数量、适应证(相对适应证)、风险、不良反应、其他器械或设备的干扰和声明："警告：研究用器械，按美国法律只能用于研究"	在器械的标签上必须标明名称、生产厂家地址、数量、适应证(相对适应证)、风险、不良反应、其他器械或设备的干扰和声明："警告：研究用器械，按美国法律只能用于研究"
IRBs 审批	IRBs 审批
	FDA 审批
接到的同意文件	接到的同意文件
监查研究的单位	监查研究的单位
研究记录的保持和保留	研究记录的保持和保留
	临床研究报告
审计单位	审计单位
不能以任何方式促使器械进入市场或商业化	不能以任何方式促使器械进入市场或商业化
要严格按方案进行，不能非法延长研究时间	严格按方案进行，不能非法延长研究时间

(三)具有重大风险的医疗器械临床研究

具有重大风险的医疗器械除得到 IRBs 批准外，还必须向 FDA 提出申请，并得到 FDA 批准后才能进行临床研究。向 FDA 申请的资料应包括以下内容。

1. 申请者的名称和地址。
2. 申请前已完成的研究报告。
3. 已完成的研究方案或总结。
4. 对方法、设备、所用对照、操作过程、贮存、安装的详细描述。
5. 研究者的同意书。

6. 列出研究者的名字和地址。

7. 以一种声明或基本的综述表并带有签字的同意书作为证明。

8. 列出 IRBs 成员名单和他们的地址。

9. 已确认的临床研究单位。

10. 如果有的话，注明付费价格。

11. 说明产品不能出售或商业化。

12. 遵守美国环境保护法律方面的说明；器械标识。

13. 临床研究受试者的资料。

以上仅是基本要求，根据器械的复杂程度和具体情况可能还有其他要求。

在临床研究中，若临床研究方案发生重大变化，要向 FDA 申报，并要取得 FDA 同意。

在所有 510(k)审查中约有 20%的医疗器械需要提交临床研究资料。510(k)中的医疗器械临床研究资料只是为了证明其与另一已上市的器械具有相同的性能。FDA 不认为 510(k)中的临床研究资料可证明该器械的绝对安全和有效。

对于不同分类的医疗器械，临床研究资料在安全性和有效性方面应等同或优于另一同样用途的已上市器械的要求是不同的。例如，要制定外科用腹腔镜的等同性就不需或仅需少量的临床研究资料；相反，要证明置入式心脏起搏器的等同性就需要大量的临床研究资料。

对于 PMA(上市前审批)的医疗器械，必须要有临床研究资料，以证明该类器械的安全性和有效性。

二、我国医疗器械临床试验管理

1. 医疗器械临床试验分为医疗器械临床试用和医疗器械临床验证。

医疗器械临床试用是指通过临床试验来验证该医疗器械理论原理、基本结构、性能要素能否保证器械的安全性和有效性。适用于市场上尚未出现过、其安全性及有效性有待确认的医疗器械。

医疗器械临床验证是指通过临床试验来验证该医疗器械与已上市产品的主要结构、性能等要素是否实质性等同，是否具有同样的安全性和有效性。适用于同类产品已上市，其安全性及有效性需要进一步确认的医疗器械。

2. 第Ⅱ类、第Ⅲ类医疗器械产品机制已得到证实，可以通过临床试验证实其结构性能与已上市的医疗器械具有同样的安全性和有效性。

对于已上市的同类医疗器械出现不良事件或者疗效不明确的医疗器械，国家食品药品监督管理局可制订统一的临床试验方案，因此，属于该两种情况的医疗器械临床验证的方案须按规定在国家食品药品监督管理局进行备案。

3. 由国家食品药品监督管理局发布《国家药品临床研究基地目录》，作为承担医疗器械临床试验的医疗机构。

4. 对于高风险的置入器械、介入器械，机制、疗效不确切的治疗类器械和境内首次上市的器械均应进行临床试验，且临床试验方案必须符合国家食品药品监督管理部门的相关要求。对于市场上尚未出现过的第Ⅲ类置入体内的医疗器械，由于该类医疗器械临床风险较高，且风险还具有较大的不确定性，除了应进行严格、充分的动物实验(且结论为合格)以外，还应进行临床试验。

5. 对于借用中医理论制成的中医医疗器械，由于临床作用机制不够明晰，且缺乏有效的评价方法对临床效果进行定量的评价，容易给临床使用者带来不确切的临床效果的引导。故在《医疗器械临床试验规定》中认为此类借用中医理论制成的医疗器械的临床试验方案必须符合国家食品药品监督管理部门的相关要求。

根据《医疗器械监督管理条例》中的“国家对医疗器械实施再评价制度”的原则，对于已上市的同类医疗器械出现不良事件或者疗效不明确的医疗器械，国家食品药品监督管理局可制订统一的临床试验方案，并且开展此类医疗器械的临床试验的实施者、医疗机构及临床试验人员应当执行统一的临床试验方案。

（李　卫　贾　宣　成小如　李　宁　王　杨　郭　晋）

第2章 医疗器械临床试验资料与方案的实施

第一节 临床试验资料

我国的医疗器械法规详细规定了在中国境内进行临床试验的第二类、第三类医疗器械的临床试验资料提供的方式及内容。

一、医疗器械临床试验资料提供方式

医疗器械法规规定在医疗器械注册时应提交的临床试验资料。

申请第二类、第三类医疗器械注册，应当提交临床试验资料。

临床试验资料提供方式执行《医疗器械注册临床试验资料分项规定》。

《医疗器械注册临床试验资料分项规定》（以下简称《分项规定》）是根据境内外生产的产品、医疗器械类别、产品具体属性、企业具体条件及产品上市质量情况等多方面详细具体明确了医疗器械临床试验资料提供的不同要求。

1. 境内生产产品的临床试验资料

（1）医疗器械临床试验资料。

（2）本企业同类产品上市后的临床试验资料。

（3）已批准上市同类产品的临床试验资料。

2. 境外生产产品的临床试验资料

（1）在中国境内进行的临床试验资料。

（2）境外政府医疗器械主管部门批准该产品在本国（地区）上市时的临床试验资料，需经中国政府组织的专家组认可。

（3）境外政府医疗器械主管部门批准该产品在本国（地区）上市时的临床试验资料。

二、提供不同医疗器械临床试验资料的条件

提供不同医疗器械临床试验资料的条件可从下面几个方面考虑：

1. 医疗器械生产地点

（1）中国境内生产的医疗器械。

（2）中国境外生产的医疗器械。

2. 产品属性

（1）第三类医疗器械。

（2）置入型产品。

3. 产品特性

(1)属于采用超声、微波、激光、X线、伽马射线以及其他散射粒子作治疗源的治疗设备。

(2)不属于采用超声、微波、激光、X线、伽马射线以及其他散射粒子作治疗源的治疗设备。

(3)诊断器械。

4. 产品基本情况

(1)产品未进入市场。

(2)本企业无产品进入市场。

(3)本企业已有产品进入市场。

(4)本企业已有同型号产品进入市场。

(5)本企业已有同规格产品进入市场。

5. 企业质量体系

(1)质量体系已经中国政府认可,但不涵盖申请产品。

(2)质量体系已经中国政府认可,并涵盖申请产品。

6. 本企业其他产品上市情况

(1)本企业其他产品在中国销售有4年以上无投诉记录。

(2)本企业其他产品在中国销售有投诉记录。

7. 置入性产品

(1)境内生产　下列情况可以提供本企业同类产品上市时的临床试验资料。

①企业已有产品进入国内市场的。

②申请产品与已注册产品属同类产品,但不属于同一型号。

③申请产品与已注册产品属同型号产品,但不属于同一规格。

④经中国政府认可的质量体系涵盖所申请的产品(或型号),并在有效期内。

⑤本企业同类产品在国内销售有4年无投诉记录。

(2)境外生产:下列情况可以提供境外政府医疗器主管部门批准同类产品注册上市时的临床试验资料。

①企业已有产品进入中国市场的。

②申请产品与已注册产品属同类产品,但不属于同一型号。

③申请产品与已注册产品属同型号产品,但不属于同一规格。

④境外政府医疗器械主管部门已批准申请产品在本国(地区)上市。

⑤经中国政府认可的质量体系涵盖所申请的产品(或型号),并在有效期内。

⑥本企业同类产品在中国销售有4年无投诉的记录。

8. 其他第三类产品　下列情况可以提供本企业同类产品上市时的试验资料。

(1)企业已有产品进入中国市场的。

(2)申请产品与已注册产品属同类产品。

(3)本企业同类产品在中国销售有4年无投诉的记录。

三类产品见表2-1与表2-2。

表 2-1　境内产品(三类)

企业产品情况			产品属性							
			置入型产品			其他第三类产品				
			未批准上市	同类产品不同型号	同型号不同规格	未批准上市	属采用超声等作治疗源的治疗设备	不是用超声等作治疗源的治疗设备	诊断型产品	属采用超声等作治疗源的治疗设备的同类产品
企业无产品进入市场			A			A				
企业已有产品进入市场	体系不涵盖		A	A	A		A	A	A	
	体系涵盖	无投诉		B	B					B
		有投诉		A	A					A

表 2-2　境外产品(三类)

产品上市情况			无论何种情况	产品属性							
				置入性产品			其他第三类产品				
				未批准上市	同类产品不同型号	同型号不同规格	未批准上市	属采用超声等作治疗源的治疗设备	非采用超声等作治疗源的设备	诊断型产品	属采用超声等作治疗源的治疗设备同类产品
境外未批准在本国地区上市											
境外已批准在本国地区上市	企业无产品进入中国市场										
	企业已有产品进入中国市场	体系不涵盖		E	E	E		E	F	F	
		体系涵盖 无投诉		F	F	F					F
		体系涵盖 有投诉		E	E	E					E

9. 二类产品

(1)境内产品下列情况可以提交同类产品上市时的临床试验资料(中国政府已批准同类产

品在中国上市)。

(2)境外产品下列情况可以提交境外政府医疗器械主管部门批准同类产品注册上市时的临床试验资料(境外政府医疗器械主管部门已批准申请产品在本国(本地区)上市。

(3)执行国家、行业标准的检验、诊断类医疗器械不需要提供临床试验资料。

如果不符合上述情况的,一般都必须在中国境内进行临床试验。现将提供临床试验资料的9种方式如表2-2。

二类产品见表2-3。

表2-3 二类产品

生产地点	产品类别 / 产品情况	无论何种情况	申报产品第一次进入中国市场		
			检验、诊断类医疗企业		其他
境内产品	同类产品尚未批准上市	A	执行国家、行业标准	未执行国家、行业标准	
	已批准同类产品上市		不需要提供临床资料	C及(实质性等同)对比说明	C及(实质性等同)对比说明
境外产品	境外尚未批准在本国(本地区)上市	A			
	境外尚已批准在本国(本地区)上市		F		

A. 医疗器械临床试验资料;B. 本企业同类产品上市时的临床试验资料;C. 已批准上市同类产品的临床试验资料;D. 在中国境内进行的临床试验资料;E. 境外政府医疗器械主管部门批准该产品在本国(本地区)上市时的临床试验资料,经中国政府组织的专家组认可;F. 境外政府医疗器械主管部门批准该产品在本国(本地区)上市时的临床试验资料

三、实质性等同对比说明

在《分项规定》中提交同类产品的临床试验资料时需提交对比说明,对此说明其基本内容应是实质性等同,进行实质性等同对比时可以从下列两种情况考虑。

1. 与上市产品相比属于实质性等同的产品

(1)与已批准上市产品属于同类产品,其基本原理、主要功能、结构、材料、预期用途相同。

(2)已经批准上市的同类产品其临床适用范围正确有效,在使用过程中没有不良记录;没有与该类医疗器械固有特性(如医疗器械的设计或组织材料等)有关的不良事故记录。

2. 与上市产品相比有一定变化的产品

(1)产品与已经上市的同类产品不完全相同,其设计、组成、结构上有一定的变化,但是产品的安全性和有效性的特征指标通过临床试验及充分的技术试验已很好地建立,并获得了可靠的试验数据;按照医疗器械风险管理标准要求进行评价,认为这些变化在本质上不会形成潜在的、新的对人体的伤害,不会产生新的安全性及有效性问题,足以证明产品是安全有效的。

(2)医疗器械产品的任何变化不会导致其分类变化。

第二节　医疗器械临床试验实施

一、临床试验的前提条件

1. 该产品注册产品标准应执行相应的国家、行业标准，符合医疗器械基本安全有效要求。

2. 该产品应经检测符合注册产品标准。

3. 受试产品为首次用于置入人体的医疗器械，应当具有该产品的动物实验报告；其他需要动物实验确认产品对人体临床试验安全性的产品也应当提交动物实验报告。

二、临床试验受试者的权益保障

1. 临床试验受试原则

(1)受试者自愿参加临床试验。

(2)有权在临床试验的任何阶段退出。

(3)医疗器械临床试验负责人或其代委托人向受试者或其代理人详细说明受试者的知情权利、受试者的利益保护等内容。

2. 受试者的知情权利

(1)受试者充分了解医疗器械临床试验的内容。

(2)医疗器械临床试验负责人或其委托人应当向受试者详细说明医疗器械临床试验方案，特别是医疗器械临床试验目的、过程和期限，预期受试者可能的受益和可能产生的风险。

(3)医疗器械临床试验期间医疗机构有义务向受试者提供与该临床试验有关的信息资料。

3. 受试者的利益保护

(1)医疗器械临床试验不得向受试者收取费用。

(2)受试者个人资料保密。

(3)因受试产品原因造成受试者损害，实施者应当给予受试者相应的补偿；有关补偿事宜应当在医疗器械临床试验合同中载明。

(4)临床试验对受试者必须是利益大于风险。

(5)如果发现风险有可能超过利益或已经得出阳性结论和有利的结果时应当停止研究。

4. 病人知情同意书

(1)说明受试者获得病人知情同意书的基础。

(2)受试者在充分了解医疗器械临床试验内容基础上，获得病人知情同意。

(3)病人知情同意书的内容：

①必须向受试者或其代理人详细说明受试者自愿参加临床试验，有权在临床试验的任何阶段退出的受试原则，受试者获得知情的权利，受试者受到利益的保护。

②还应当包括以下内容：a. 医疗器械临床试验负责人签名及签名日期；b. 受试者或其法定代理人的签名及签名日期。

(4)病人知情同意书的修改。医疗机构在医疗器械临床试验中发现受试产品预期以外的临床影响，必须对《知情同意书》相关内容进行修改，并经受试者或其法定代理人重新签名确认。

三、临床试验方案

1. 临床试验方案制定的目的和意义　临床试验方案是指导参与临床试验所有研究者如何启动和实施临床试验的研究计划书，也是试验结束后进行资料统计分析的重要依据。医疗器械临床试验方案主要阐明临床试验目的、风险分析、总体设计、试验方法和步骤等内容。医疗器械临床试验在开始前应当制定试验方案，医疗器械临床试验必须按照该试验方案进行。

2. 临床试验方案制定的首要原则　医疗器械临床试验方案应当以最大限度地保障受试者权益、健康安全为首要原则。

3. 临床试验方案制定的格式　医疗器械临床试验方案按规定的格式设计制定。

4. 临床试验方案的制定者　医疗器械临床试验方案由负责临床试验的医疗机构和实施者制定。

5. 临床试验方案的认可　医疗器械临床试验方案报伦理委员会认可后实施。

6. 临床试验方案的修改　认可后的医疗器械临床试验方案若有修改必须经伦理委员会同意。

7. 临床试验方案的备案　下列医疗器械的临床试验方案应当向医疗器械审评机构备案。

(1)市场上尚未出现的第三类置入体内的医疗器械。

(2)借用中医理论制成的医疗器械。

8. 临床试验方案制定的要求

医疗器械临床试验方案应当针对具体受试产品的特性，确定临床试验例数、持续时间和临床评价标准，使试验结果既具有临床意义、又具有统计学意义。医疗器械临床验证方案应当证明受试产品与已上市产品的主要结构、性能等要素是否实质性等同，是否具有同样的安全性和有效性。

9. 临床试验方案的内容

(1)临床试验的题目。

(2)临床试验的目的、背景和内容。

(3)临床评价主要终点和次要终点。

(4)临床评价标准。

(5)临床试验的风险与受益分析。

(6)临床试验人员姓名、职务、职称和任职部门。

(7)总体设计，包括成功或失败的可能分析。

(8)临床试验持续时间及其确定理由。

(9)每病种临床试验例数及其确定理由(依据)。

(10)必要时对照组的设计。

(11)治疗性产品应当有明确的适应证或适用范围。

(12)临床性能的评价方法和统计分析方法。

(13)副作用预测及应当采取的措施。

(14)受试者《知情同意书》。

(15)各方职责。

10. 临床试验方案的签署　医疗机构与实施者签署双方同意的临床试验方案，并签订临床试验合同。

11. 特殊情况医疗器械临床试验方案的制定　对于特殊情况：

(1)已上市的同类医疗器械出现不良事件。

(2)疗效不明确的医疗器械。

国家食品药品监督管理局可制订统一的临床试验方案的规定。开展此类医疗器械的临床试验方案，实施者、医疗机构、临床试验人员应当执行统一的临床试验方案的规定。

12. 临床试验与临床试验方案的关系　医疗器械临床试验必须按照医疗器械临床试验方案进行。

医疗器械临床试验应当在两家以上(含两家)医疗机构进行。

四、临床试验的实施

1. 临床试验实施者　医疗器械临床试验的实施者为申请注册该医疗器械产品的单位。

2. 临床试验实施者职责

(1)依法选择医疗机构和数据管理及统计分析机构。

(2)向医疗机构提供《医疗器械临床试验须知》。

(3)与医疗机构和数据管理及统计分析机构共同设计、制定医疗器械临床试验方案、签署双方同意的医疗器械临床试验方案及合同。

(4)向医疗机构免费提供受试产品。

(5)对医疗器械临床试验人员进行培训。

(6)向医疗机构提供真实、可靠的试验数据并担保。

(7)如果发生严重不良反应应当如实、及时分别向受理该医疗器械注册申请的省、自治区、直辖市(食品)药品监督管理部门和国家食品药品监督管理局报告，同时向伦理委员会及进行该医疗器械临床试验的其他机构通报。

(8)实施者终止医疗器械临床试验前，应当通知医疗机构、伦理委员会和受理该医疗器械注册申请的省、自治区、直辖市(食品)药品监督管理部门和国家食品药品监督管理局，并说明理由。

(9)受试产品对受试者造成损害的，实施者应当按医疗器械临床试验合同给予受试者补偿。

3.《医疗器械临床试验须知》《医疗器械临床试验须知》应当包括如下内容。

(1)受试产品原理说明、适应证、功能、预期达到的使用目的、使用要求说明、安装要求说明。

(2)受试产品的技术指标。

(3)国务院食品药品监督管理部门会同国务院技术监督部门认可的检测机构出具的受试产品型式试验报告。

可能产生的风险，推荐的防范措施及紧急处理方法。

可能涉及的保密问题。

五、临床试验医疗机构及临床试验人员

1. 临床试验医疗机构及药品临床试验基地　国家食品药品监督管理局发布《国家药品临床研究基地目录》，规定了承担医疗器械临床试验的医疗机构。

2. 临床试验人员的资格和条件

(1)医疗器械临床试验人员应当具备的资格：负责医疗器械临床试验的医疗机构应当确定主持临床试验的专业技术人员作为临床试验负责人，临床试验负责人应当具备主治医师以上职称。

(2)临床试验人员应当具备条件：①具备承担该项临床试验的专业特长、资格和能力；②熟悉实施者所提供的与临床试验有关的资料与文献。

3. 临床试验机构和临床试验人员的职责

(1)应当熟悉实施者所提供的与临床试验有关的资料，并熟悉产品的使用。

(2)与实施者共同设计、制定临床试验方案，双方签署临床试验方案及合同。

(3)如实向受试者说明受试产品的详细情况，临床试验实施前必须给受试者充分的时间考虑是否参加临床试验。

(4)如实记录受试者的不良反应及不良事件，并分析原因，发生不良事件及严重不良反应，应当如实、及时分别向伦理委员会、受理该医疗器械注册申请的省、自治区、直辖市(食品)药品监督管理部门和国家食品药品监督管理局报告，发生严重不良反应应当在24h内报告。

(5)在发生不良反应时，临床试验人员应当及时作出临床判断，采取措施，保护受试者的利益；必要时，伦理委员会有权终止临床试验。

(6)临床试验终止的，应当通知受试者、实施者、伦理委员会和受理该医疗器械注册申请的省、自治区、直辖市(食品)药品监督管理部门和国家食品药品监督管理局，并说明理由。

(7)提出临床试验报告，并对报告的正确性及可靠性负责。

(8)对实施者提供的资料负有保密义务。

六、临床试验报告

1. 临床试验报告的出具　医疗器械临床试验完成后，由承担临床试验的医疗机构负责出具。

2. 临床试验报告的要求　临床试验报告应当包括下列内容。

(1)试验的病种、病例总数和病例的性别、年龄、分组分析、对照组的设置(必要时)。

(2)临床试验方法。

(3)所采用的数据管理及统计分析方法。

(4)临床评价方法及标准。

(5)临床试验结果。

(6)临床试验结论。

(7)临床试验中发现的不良事件和不良反应及其处理情况。

(8)临床试验效果分析。

(9)适应证、适用范围、禁忌证和注意事项。

(10)存在问题及改进意见。

临床试验报告的格式应符合规定的格式；应由临床试验人员签名并注明日期；由承担临床试验的医疗机构中的临床试验管理部门签署意见、注明日期、签章。

3. 临床试验资料的保存和管理

(1)医疗器械临床试验资料应当妥善保存和管理。

(2)医疗机构应当保存临床试验资料至试验终止后 5 年。

(3)实施者应当保存临床试验资料至最后生产的产品投入使用后 10 年。

（李　卫　孙　毅　成小如　贾　宣　王　杨　郭　晋）

第3章 医疗器械临床试验主要内容

按照国家食品药品监督管理局5号令(2004年1月17日)定义,医疗器械临床试验是指:获得医疗器械临床试验资格的医疗机构(以下称医疗机构)对申请注册的医疗器械在正常使用条件下按照规定在人体进行的试用或验证过程。其目的是评价受试产品是否具有预期的安全性和有效性。

由于临床试验通常是根据研究的目的,通过样本来研究器械对疾病及其预后等方面的作用,而一个好的医疗器械临床试验应能够提供最客观的安全性和有效性评价。因此,医疗器械临床试验设计必须应用统计学原理对试验相关的因素做出合理的、有效的安排,并最大限度地控制试验误差,提高试验质量,并对试验结果进行科学合理的分析,在保证试验结果科学、准确、可信的同时,尽可能做到高效、快速、经济。

因此,生物统计学在医疗器械临床试验的全过程中有着不可缺少的重要作用,只有当研究结果既具有临床意义,又具有统计学意义时,该器械才能获得批准。

生物统计学原则应贯穿在医疗器械临床试验的全过程(包括研究方案设计、试验实施、质量控制、数据管理及统计分析)。应由熟悉被试产品的、具有生物统计学资质的、专业的生物统计学人员,采用国内外公认的经典的统计分析方法和统计分析软件对医疗器械临床试验中数据进行统计分析。

本文充分考虑了我国医疗器械临床研究的现状,参考了ICH E9文件以及美国食品药品监督管理局(FDA)的医疗器械现行统计学指导原则,以临床试验的基本要求和统计学原理为重点,从生物统计学角度为医疗器械临床试验过程提供了一个全面的实施办法,包含了对医疗器械临床试验的总体考虑以及试验设计时、试验实施过程中及试验结果分析和报告时的统计学问题,旨在为医疗器械注册申请人和临床试验的研究者在整个临床试验过程中如何进行设计、实施、质量控制、分析和安全性、有效性的评价提供指导,以期保证医疗器械临床试验的科学性、严谨性和规范性,并力求符合中国国情,使之具有充分的可操作性。

一、临床试验全过程

医疗器械临床试验的主要目标是寻找是否存在其风险/效益比可接受的、使用安全有效的器械,同时还要确定该器械可能受益的特定人群及使用适应证(美国食品药品监督管理局定义)。为达到以上总体目标,需要设计具有特定目的的临床试验。

首先,应有一个详尽的临床试验研究方案,在每一个临床试验中,与该临床研究有关的所有设计、实施、质控及拟采用的统计分析方法等细节均应在试验开始前的临床试验方案中予以明确说明。

在该研究方案得到伦理委员会批准后,即进入具体实施阶段。为了保证数据的可溯源性,

对每一临床试验的所有受试者，均应建立原始观察记录（病历），和电子/纸质的病例报告表（case report rorm，CRF）。纸质病例报告表通常为一式两联（或三联）、无碳复写的病例报告表。在试验实施过程中，研究者（investigator）要及时、准确、无误、完整、清晰地将试验数据载入电子/纸质病例报告表中。

为了保证临床试验的质量，临床试验的实施者（厂家）应指派监查员（monitor）对临床试验的全过程进行监查和质控。

数据管理员应根据病例报告表和统计分析计划书要求建立数据库，并保证数据库运行的正确性。

统计分析人员应根据研究方案和病例报告表，撰写统计分析计划书。所有统计分析应建立在正确、完整的数据基础上，并采用国内外公认的经典统计分析方法和统计分析软件。最后，生物统计学专业人员根据统计结果写出统计分析报告，提供给主要研究者作为撰写临床试验总结报告的依据。

临床试验过程中的所有文件均应妥善保存以备核查。

对于某些高风险的、全新的医疗器械，需要首先进行小规模的探索性试验或预试验，这些试验也应有清晰和明确的研究目的。探索性试验有时需要更为灵活可变的方法进行设计并对数据进行探索性分析，以便根据逐渐积累的结果对后期的确证性试验设计提供相应的信息。虽然探索性试验对整个有效性的确证有所贡献，但不能作为产品证明安全性和有效性的正式依据。医疗器械的有效性和安全性只有通过探索性试验（如有）之后的确证性试验才能得到充分证实。一般来说，确证性试验是一种事先提出假设并对其进行检验的、具有良好对照的随机试验，以说明所研制的医疗器械对临床是有益的。

需要注意的是研究中所用的统计分析方法应与分析的目的相一致，而且要有全程证明文件。

二、多中心临床试验

多中心临床试验指由一个单位的主要研究者总负责、多个单位的研究者合作、按照同一个试验方案同时进行的临床试验。通常情况下多中心试验的每个研究单位由一名研究者负责。多中心试验可以在较短的时间内收集所需的病例数，且收集的病例范围广，临床试验的结果对将来的应用更具有代表性。

由于是多中心试验，统计分析时要将各中心的研究数据合并，以便达到所需要的样本量及研究的把握度，因此研究中心和研究者的选择对临床试验的成败是至关重要的。选择的中心必须有试验器械适用的充足的病人数，且各中心试验组和对照组的病例数的比例应与研究方案中规定的总样本的比例相同，以保证各中心的可比性。每一个中心必须具备研究方案中所描述的用于治疗病人的设施和手段，并且必须有具有研究资质的人员来实施该项临床试验。然而，应该注意，尽管使用了统一的研究方案，并且研究监查员尽了最大的努力，当合并各中心数据时，中心效应还是可能出现的。研究方案设计时应考虑如何排除由于中心效应所带来的潜在的偏性。

每个中心的主要研究者必须将合格的病人入选到试验中来，并且必须遵从方案所规定的标准操作规程（SOP）。如果研究者连续违背方案，则该中心数据不能被用于申办者器械的安全性及有效性评价。

临床试验基本上是一个基于人群的试验，因此不同于常规的医疗实践。应该注意，在很多研究中，均需要对中心进行意向性治疗（intention to treat，ITT）分析，在这种分析中，违背方案的病人数据将被作为无效数据。很明显，意向性治疗模型中违背方案的相对少量的病人就可以对最终结果产生巨大的、本质的影响。因此，由某一个研究者不遵从方案所入选的病人可以对试验结果的分析造成巨大的问题。

因此，保证研究者遵从研究方案是申办者的责任。试验前应对各中心所有参加临床试验的人员进行统一培训，试验实施过程中要有监查及质控措施。为了避免各中心疗效/安全性评价有较大差异，应采取相应的措施，如统一由中心实验室检验、阅片、评价等。候选的研究者无论什么原因显示出在试验的过程中有可能不能严格遵从方案时，申办者不应该让该研究者参加临床试验。

三、偏倚控制方法

偏倚又称偏性，是指在临床试验方案设计、实施及分析评价结果时，有关影响因素所致的系统误差，致使对器械疗效或安全性的评价偏离真值。偏倚干扰临床试验得出正确的结论，在临床试验的全过程中均须防范其发生。随机化和盲法是控制偏倚的重要措施。

（一）随机化

将治疗分配给病人时，应将选择偏倚降到最小。当具有一个或更多个重要基线特征的病人更频繁地出现在某一组时，选择偏倚就出现了。例如，如果某病发病率男性比女性高 2 倍，且某一组中的男性人数是女性的 2 倍，而另一组中女性人数是男性的 2 倍，那么，在不进行任何治疗的情况下就已经观察到两组发病率的差别了。此时如果对某一组分配治疗，治疗的疗效就会出现混杂，即由性别效果产生的无法区分的混杂。

因此，必须采取适当措施，使得各组间已知或可疑基线因素的不平衡达到最小。最好的控制选择偏倚的方法是随机。随机过程能够保证将病人分到治疗或对照组的机会是均等的。如果试验足够大，且具有有限的比较组，则随机将保证克服基线因素间的不平衡。随机还可以防止由于研究者有意或无意识的行为导致的组间不可比（如分配或选择最严重的病人到医生认为更有效的治疗组）。

随机化包括分组随机和试验顺序随机，与盲法合用，随机化有助于避免在受试者的选择和分组时因治疗分配的可预测性而导致的可能偏倚。

临床试验中可采用简单随机、分层随机、区组（block）随机或最小化随机等方法。有时，当试验样本量很小，但有很多组时，简单随机也许不能保证基线各组内预后因素（如疾病的严重程度）的均衡。在这种情况下，应采用将选择的诊断变量进行分组的分层随机方法。分层随机化有助于保持层内的组间均衡性，特别在多中心临床试验中，中心就是一个分层因素。另外为了使各层趋于均衡，按照基线资料中的重要预后因素等进行分层，对促使层内的均衡也很有意义。区组随机化有助于减少季节、疾病流行等因素对疗效的影响。需注意的是，区组太大可能造成组间基线不均衡，太小则可能使研究者猜测出实际分组。

当样本大小、分层因素及区组大小等因素决定后，生物统计学专业人员即可在计算机上使用统计软件产生随机分配表。临床试验的随机分配表就是用文件形式写出对受试者的治疗安排，即治疗的顺序表。随机分配表必须是可重复的，即当产生随机数表的初值、分层、区组决定后，能重新产生这组随机数。

一般来说，试验用器械应根据生物统计学专业人员产生的随机分配表进行编码，以达到随机化的要求，受试者应根据入组时间，严格按照试验用器械编号的顺序入组，不得随意变动，否则会破坏随机化效果。试验中所用的随机化的方法应在研究方案中说明，但容易使人预测分组的随机化的细节（如分段长度等）不应包含在试验方案中。

有时也可采用其他的分配治疗的方法，但除非使用了真正的随机模式，否则很难避免系统的或其他可能的偏倚。例如，按照某一系统模式将病人分到某一治疗组，假如每隔 4 个病人，这似乎是随机。然而，这样一种周期性的分配有时可能与病人看医生的周期一致，从而导致治疗组与对照组入选的不均衡，进而导致选择偏倚，因为治疗分配是可以预测的。

应该经常地监查治疗分配过程，以保证各组已知或可能影响终点的重要因素间的大致均衡。分层随机模式可自动地保证组间均衡，而其他随机方法（如最小化随机）需要监控和调整（因为较复杂，在此不赘述）。

（二）盲法

临床试验中可能出现的 3 个更严重的偏倚为研究者选择偏倚、效果评价时的可评价偏倚及安慰剂效应。当一个研究者有意识地或潜意识地喜欢某一组时，就会出现研究者的选择偏倚。例如，如果研究者知道哪一组是治疗组，他（她）就会更频繁地关注治疗组，从而使得治疗组与对照组被关注的程度有很大的不同，而两组之间的这种差异将严重地影响试验的结果（终点）。

可评价偏倚可以是研究者偏性中的一种，在这种偏倚中，评价疗效的人可以有意或无意地掩盖某一组的弱点，而倾向于另一组。在主观性研究或生活质量研究中，其终点就非常容易受这种偏倚的影响。

当病人处于一个非活性治疗模式，但他（她）相信自己正在进行有效的治疗并随后显示或报告症状有所改善时，安慰剂效应就出现了，这也是一种偏倚。

为了在临床试验的过程中防范这些潜在的偏性，必须使用盲法。

临床试验的盲法根据设盲的程度不同分为双盲、单盲、第三方盲法和非盲。设盲的程度取决于潜在偏倚的强度和严重性。单盲设计使病人不知道自己进入的是治疗组还是对照组。双盲设计使病人和研究者都不知道哪一组是治疗组。如条件许可，应尽可能采用双盲试验，尤其在试验的主要变量易受主观因素干扰时。如果双盲不可行，则应优先考虑单盲试验。在某些特殊情况下，由于一些原因而无法进行盲法试验时，可考虑进行非盲的临床试验（开放试验），但同时应采用第三方盲法评价治疗效果的研究设计（如中心实验室、中心阅片室、终点评价委员会等）。

设盲方法及理由，以及通过其他方法使偏倚达到最小的措施等，均应在试验方案中预先说明。

器械编盲与盲底保存应由不参与临床试验的人员负责。根据已产生的随机分配表对试验用器械进行分配编码的过程称为器械编盲。随机数、产生随机数的参数及试验用器械编码统称为双盲临床试验的盲底，用于编盲的随机数产生时间应尽量接近于器械分配的时间，编盲过程应有相应的监督措施和详细的编盲记录，完成编盲后的盲底应一式二份密封，交器械注册申办者保存，一份用于统计分析后试验接盲，另一份用于以后的监督核查。

不过，医疗器械由于其特殊性，很难进行单盲或双盲的临床试验，而取而代之的是第三方盲法评价（如中心实验室、中心阅片室、终点评价委员会等）。无论采用何种盲法设计，均应制

订相应的控制试验偏倚的措施，使已知的或可能的偏倚达到最小。例如，主要指标应尽可能客观、采用基于计算机系统的中央随机法（central randomization）入选受试者、参与疗效及安全性评价的研究者在试验过程中应尽量处于盲态等。双盲临床试验中，从随机数的产生、试验用器械的编码、受试者入组治疗、研究者记录试验结果和作出评价、监查员进行监查、数据管理直至统计分析，都必须保持盲态。由于违背盲法所引入的偏性在数据统计分析时是很难评价的，因此一定要在统计分析完成后再接盲。如果发生了任何非规定情况所致的盲底泄露，并影响了该试验结果的客观性，则该试验将被视作无效。

四、数据管理

数据的正确性、完整性和准确性对保证临床试验的质量，获得科学、有效、真实的研究结果极为重要，因此必须十分重视。在临床试验过程中应认真进行数据管理，以保证试验数据的正确和完整。研究者应根据受试者的原始观察记录，保证将数据准确、完整、清晰、及时地载入病例报告表。原则上，完成的病例报告表中不得有缺、漏项。

为了及时向研究者反馈数据管理信息、缩短数据管理时间、加快临床试验进程，数据管理可在临床试验启动后与临床试验同时进行。

（一）数据的收集和确认

获得和确认试验中所有测量变量的准确性的方法必须在试验开始前准备好，并要监督它的执行情况。每个研究点必须有足够的具有相关技术的人员以确保获得有效的数据。要特别注意每一细节，因为不可能回顾性地评价未在规定时间内取得的数据或不具有一定精度的数据。

这些方法必须包括数据测量、记录，转成电子媒介及确认的质控技术。试验前应对试验中观察的每一变量、条件或特征给予明确的定义。研究者应完全理解所有的定义术语，而且必须确保各研究者与各研究中心间的一致性。

试验术语的一致性是非常重要的，它能保证与文献中的其他试验或研究相比时具有可比性。

临床试验数据的收集和传送，可采用多种形式，目前较为常用的形式为电子或纸质病例报告表（case report form，CRF）。从试验数据的收集到数据库的完成，均应符合我国食品药品监督管理局（SFDA）《药物临床试验质量管理规范》（good clinical practice，GCP）的规定，尤其是及时的数据记录、错误修改、数据更正留痕等。这些步骤均是建立高质量数据库、完成试验方案并达到试验目的所必需的。

为了保证临床试验的质量，监查员要对临床试验的全过程进行监查和质控。

由于医疗器械的特殊性，很多临床试验无法进行双盲研究，因此，对疗效的科学评价就显得尤为重要。在条件许可的情况下，尽量建立独立于临床试验的第三方盲法评价（如终点评价委员会、中心实验室读片等），以便最大限度地获得真实可靠的疗效评价结果。

（二）缺失值及离群值

研究者应根据受试者的原始观察记录，保证将数据准确、完整、清晰、及时地载入病例报告表。由于缺失值是临床试验中的一个潜在的偏倚来源，因此，完成的病例报告表中原则上不应有缺失值，尤其是重要指标（如主要的疗效和安全性评价指标）必须填写清楚。对病例报告表中的基本数据，如性别、出生日期、入组日期和各种观察日期等不得缺失。试验中观察的阴性结果、测得的结果为零和未能测出者，均应有相应的符号表示，不能空缺，以便与缺失值区分。

离群值问题的处理，应当从医学和统计学专业两方面去判断，尤其应当从医学专业知识判断。对于离群值的处理方法，应在资料统计分析时进行灵敏度分析（包括和不包括离群值的两种结果比较），研究不同情况下的结果是否不一致以及产生不一致的直接原因等。

（三）数据质量控制

对于已完成的、经过监查员检查确认准确无误的病例报告表，应及时上传至临床试验数据管理中心进行数据管理，以便及时发现问题解决问题。对于纸质的病例报告表，应及时由监查员将一式两联（或三联）、无碳复写的病例报告表首页撕下送交临床试验数据管理中心进行数据管理，以便及时将文字信息转化为电子信息，及时发现问题解决问题。完成的病例报告表在研究者、监查员、数据管理员之间的传送应有专门的记录并妥善保存。根据病例报告表和统计分析要求，在第一份病例报告表送达数据管理中心以前，数据管理员应建立数据库，并在第一份 CRF 送达数据管理中心后对数据库进行调试，以保证数据库运行的正确和安全性。

数据管理员还应对每一份病例报告表进行初步审核（目视核查）。初步审核通过后，由两名计算机数据录入人员分别独立地将病例报告表输入数据库中（两遍录入），并用软件对两遍输入的结果进行比较（两遍比较）。如果两个数据库中数据不一致，需对照原始病例报告表查出原因，加以确认。

数据库中数据通过录入错误核查后，数据管理员还需编写计算机程序，对病例报告表中各指标的数值范围及其相互关系，进行范围和逻辑检查。

一旦发现源于病例报告表的任何错误，均要填写疑问表（Query Form），及时通知试验监查员，并通过监查员将疑问表转交研究者进行核实，并要求研究者在疑问表上作出回答。数据管理员根据核实的结果对数据库中数据进行修改。所有疑问表及相关文件应有详细记录并妥善保存。

最后，要再次对数据库中的指标（特别是主要疗效指标）进行抽样的人工检查并与病例报告表进行逐项核对。一般来说，关键性指标（如疗效及安全性指标）应 100% 核查，其余指标可按照一定比例抽查。

当所建立的数据库准确无误后，应由主要研究者、申办者、统计学家及相关人员对数据库进行锁定。此后，对数据库所作的任何改动只有在以上几方人员同意（以书面形式）的情况下才能进行。数据库锁定后需妥善保存备查，同时将数据库交统计学人员进行统计分析。

五、监查、干预与随访

（一）监查

试验研究过程中，由实施者指派临床监查员，定期对研究单位进行现场监督访问，以保证研究方案的所有内容都得到严格遵守，以及填写资料的正确无误。试验中心应客观、真实地记录和保留试验研究过程中所有的数据和方案执行、修改的情况。在招募病人阶段，应该尽可能保证各中心入选（排除）标准的一致性。

有时，在某些临床试验中，由于入选标准过于严格可能导致招募病人进度过慢，需要重新修改方案。研究方案中对于这部分的变化进行的修订应该考虑到任何统计学上的影响，如由于事件发生率变化引起的样本量调整，或是对于分析计划进行的修订。

（二）干预

干预的分配和应用都必须严格按照方案来进行。每一过程都必须有一个预先安排好的操

作规程。除了使用的器械不同，治疗组和对照组的操作规程应尽可能一样。

(三)随访

干预后的随访不是简单地安排时间约见受试者，应该有一个合理的安排以确保随访的受试者有较好的依从性。即使各组间在随访过程中仅存在中等程度的偏差，也可能导致数据分析时产生巨大的偏倚。

随访有两个重要特征：完整性和随访期。完整性的定义为入选试验的受试者完成全部随访的比例。非常重要的是这个比例应尽可能地接近100%，因为统计效能会随患者的失访而降低。比例<80%的随访通常被认为是质量很差的试验，且这些试验通常被认为是不完整的。同样重要的是各组及各研究中心间的随访比例应是相似的。不完整的随访是分析中主要考虑的问题。试验必须要有可行的程序来追踪那些失访的受试者。失访患者的估计是一个重要的分析问题，因为这些患者也许可以为临床试验提供最重要的信息，特别是如果这些患者的后果不好的情况下。因此，判定进入试验的所有病人(包括那些一次都没来随访的患者)的健康状况是非常重要的。

随访期是指在干预后研究个体被观察至被评估之间的时期。随访期的长短必须与安全性及有效性的要求一致，即它必须等于声称的发挥效力的时间，同时，随访期必须足够长，以便能够精确地估计已知的或可疑的不良事件发生率。各组间和各研究中心间的随访期也应该相同。

最终报告中应该记录所有参与该临床研究的病人以及设备，应该仔细记录被排除在最终分析之中的原因。同样，应该记录对于参与分析人群的所有病人及设备，所有重要变量在相关时刻的测量值；也应该简要报告其他关于参与了入选筛查，而未能参与随机化过程的病人情况。对于参加随机化分组的所有患者如果从治疗中退出，以及对于试验方案较严重违背对于分析主要变量所造成的影响，应该确定失访病人以及退出病人，并对其进行描述性分析，包括失访原因以及其与治疗及结果的关系。临床研究结论的报告是基于一份完整的统计学报告，其结论应该是基于对临床研究结果描述、解释以及分析。基于这一点，临床研究中的统计学专家应该是负责研究报告团队中的一员，并应该在临床研究报告上署名。

六、病历报告表

除了研究方案的撰写以外，临床研究者和生物统计学家在试验设计阶段共同从事的另一项重要工作是设计病历报告表。为了保证数据的可溯源性，对每一临床试验的所有受试者，均应建立原始观察记录(病历)和一式三联、无碳复写的病例报告表。在试验实施过程中，研究者要及时、准确无误、完整、清晰地将试验数据载入病例报告表中。

病历报告表中首先应由研究者(临床专家)设计并列出所有要验证研究假设所需要的全部疗效/安全性评价信息(包括病人的基线信息)，然后，应由生物统计学家从数据管理和统计分析角度进一步审阅CRF表格设计的合理性及逻辑性，一方面可以降低CRF填写人员的填写误差，另一方面根据统计方案，进一步确认CRF中确实收集了所有将来用于疗效和安全性评价的信息，易于将来的数据管理和统计分析。

七、试验目的与研究对象的选择

(一)试验目的

试验目的即所要研究的问题。一个有效且效率高的临床试验的设计肯定有一个清楚、准

确的目的。通常研究目的由一个研究问题所提出，包括器械的使用说明。如该器械是否和其他器械一样有效还是更有效？如果是这样，它是一样安全还是更安全？对这种器械的安全性和有效性的评价是否局限于某一特定的亚组人群？测量安全性和有效性的最好的临床方法是什么？

对上述这些问题及其类似问题的回答将使我们在设计临床试验时，对临床试验的观察指标进行标识。例如，如果一种新型器械被用来治疗一种进行性的不断恶化的眼科疾病，但针对这种疾病目前还存在另外一种已被认可的器械，在这种情况下怎样来界定疗效？这种新器械能否延缓或阻止病情恶化？如果可以，能否恢复以前失去的功能？能否缓解疼痛或不适感？是否与已被认可的器械做过比较？是否在某种用途上与已被认可的器械一样有效还是效果更好？是否有更少的不良反应？

我们可以看出，要回答这些问题不仅需要明确的研究目的，而且还要求申办者考虑一系列的其他问题，例如合适的终点或结局变量，对照人群的选择，可能用到的假设类型，以及其他一些问题等。

以上这些问题必须在方案完成前写好，因为我们必须决定所陈述的研究问题是否可以通过一个设计很好的临床试验来解决。即我们能否通过收集、分析和解释从临床试验中得到的数据来获得对所研究问题的准确的和客观的答案。

(二)研究对象的选择

研究人群是将来使用医疗器械的目标人群的一个代表人群。研究人群应在试验前由精通本研究领域的临床专家用严格、明确的入选(排除)标准进行定义，即由入选(排除)标准来定义研究人群的特征，并且通过这种方式确定今后医疗器械的用途。

可以将研究人群定义得小一点，以便组成人群的个体具有同质性，这样可以使临床试验有一个较小的样本量。也就是说，同质组中疗效的变异度通常比异质组低。而这种变异度的降低(假设所有其他重要因素不变)，将显著降低要观察出两组间有显著差异时所需要的样本量。它的缺点是由较窄范围的人群研究所得出的结论较难推广到一般人群上。因此，当临床试验开始前，申办者应与研究者认真讨论如何定义研究人群。

入选(排除)标准应包括诊断因素对结局变量的评价，因为这些变量中的一个或更多也许影响器械的有效性。例如，对某一特定疾病来说，性别也许是诊断因素。那么很自然的应评价性别在器械评价中的作用，然后确定入选(排除)标准、设计的其他方面以及相应分析方面的考虑等。还应考虑以下的问题：患者年龄、伴随疾病、治疗或一般情况(包括基线及以后随访)、疾病的严重程度以及其他因素等。

1. *病例组的选择*　病例应选自经金标准证实的患有目标疾病的病人，包括各种不同临床类型和病期的病例，如典型和不典型、病情严重程度不同的(轻、中、重)、不同病程阶段的(早、中、晚)，以及有无治疗史和有无并发症的，以便是诊断试验的结果能代表该病患者的总体。病例代表性的好坏，将直接影响诊断结果的推广价值。

2. *对照组的选择*　临床试验中对照组的设置应具备 3 个条件。

(1)专设：任何一个对照组都是为相应的试验组专门设立的，不得借用文献上的记载或以往研究的、其他研究的资料作为本研究的对照组。

(2)同步：对照组与试验组设立之后，在整个研究进程中始终处于同一空间和同一时间。

(3)对等：除研究因素外，对照组具备与试验组对等的一切因素。以上三个条件均是保证

对照组与试验组间的非实验因素的均衡一致，以便充分发挥对照组应有的作用。

八、预试验与可行性研究

如果一个申办者因为对器械在人群中的应用缺乏足够的经验而不能回答临床试验的关键问题，那么申办者应该设计一个小样本的人体试验来收集基本信息。这个小样本试验（通常称为预试验或可行性研究）的研究目的是确定器械的可能医疗用途、监查潜在的研究变量、检测试验流程以及决定潜在的反应变量的精确度。还可对可能导致偏倚的因素进行有限的评价。预试验的研究方案应报 SFDA。

预试验经常被用来做器械的检验，即申办者有一个关于器械用途的好的想法，需要一个小样本的试验来验证其理论或新技术，但预试验的范围不能太广。临床试验相关问题包括器械使用、病人处理和监控、数据的收集和确认以及医师的能力和所关心的问题等。应关注主要变量的测量，包括可能的结局变量和可能造成偏倚的影响变量。然而，需要长期终点的试验，通常不进行预试验。

预试验允许进行有限的假设检验，并且是进行探索性数据分析的理想场所，如寻找研究器械与结局变量之间有意义的关系，因为探索性的研究方法经常引发临床试验中可评价的研究问题。

九、研究终点和试验设计

评价申请注册器械的有效性和安全性应设立相应的临床试验观察指标（变量）。在有效性评价设计时一般应考虑两种变量：结果变量（或是终点）和影响变量。结果变量给出了临床研究目的中所提出问题的答案。并且应该对该设备声称的性能有直接影响。这些变量应该能够直接观测，尽可能客观。有最小的偏倚及误差，与接受设备干预病人的临床状况的生物效应有直接联系。

任何对于某种设备的临床研究应该有主要结果变量，有些时候存在次要结果变量。主要结果变量应该是能够提供直接与首要研究目的相关的，最可信的变量。通常仅有一个主要结果变量，该变量应该在试验方案中明确。在所有相同样本量的估计中，这是最恰当的变量。然而有时使用更多的主要变量可以更好地涵盖设备效应的范围。次要结果变量可能支持与主要目的相关的测量，也可能是与次要研究目的（如果存在）相关的测量。在试验方案中预先对其进行定义也十分重要。影响变量（混淆因素）或预后因素指的是，研究过程中任何可能干扰终点或治疗与结果之间关系的方面。因此，在设计某器械临床试验的过程中，应该小心地确定可能影响试验结果的影响因素。

（一）研究终点

观察指标是指能反映临床试验中器械疗效和安全性的观察项目。统计学中常将观察指标称为变量（variable）。观察指标必须在研究方案中有明确的定义和可靠的依据，不允许随意修改。观察指标分为测量指标和分类指标。

1. 主要终点（primary endpoint） 又叫主要指标，是能够为临床研究主要目的提供可靠证据的指标。主要终点常常是估计样本量及得到最终临床结论的依据。主要终点既可为疗效指标，也可为安全性评价指标。主要终点应为相关研究领域中已得到公认的准则或标准。可以采用前人在相关研究领域中已采用过的并已通过实践验证过的指标。应选择易于量化、客

观性强的指标作为主要终点。在研究方案中需确定有明确的定义，并说明选择的理由。在临床试验的疗效或安全性评价时，应以主要终点为依据。并用于试验样本含量的估计。

2. 次要终点(secondary endpoint)　又称为次要指标，是指为试验主要目的提供支持的附加指标，也可以是与试验次要目的有关的指标，在研究方案中也需明确说明与定义。

3. 联合终点(composite endpoint)　多个终点组合起来构成的终点称为联合终点，又称为复合指标。临床上经常采用的量表(rating scale)就是一种联合终点。当组成联合终点的某些独立终点具有临床意义时，也可以同时单独进行统计分析。

(二)试验设计

应特别强调的是，由某个研究者造成的背离研究方案将对试验的分析产生巨大的问题。申办者应保证研究者遵从研究方案。有迹象表明，在试验过程中不愿遵从方案的研究者，不论何种原因都不能参与临床试验。

1. 平行组设计　平行组设计也就是人们常说的头对头设计，即试验组和对照组同时开始、同时结束，是所有试验设计中最简单、也是最常见的一种试验设计类型。平行组设计中，各组受试者在试验中处于相同的条件，唯一的不同点是各组所使用的器械不同，如有的是试验器械，有的是对照器械，最后根据试验结果作出统计分析。

通常，根据试验方案的需要，可为试验组设置一个或多个对照组，试验器械也可按照若干种治疗强度设组。对照器械的选择应符合试验方案的要求。对照组可分为阳性或阴性对照。阳性对照一般采用按所选适应证的当前公认的有效器械，阴性对照一般采用疗效未被业界证实的安慰器械，但必须符合伦理学要求。

2. 交叉设计　是按事先设计好的试验次序，在各个时期对受试者逐一实施各种处理，以比较各处理组间的差异。交叉设计是将自身比较和组间比较设计思路综合应用的一种设计方法，可以很好地控制个体间的差异，同时减少受试者人数。

然而，实施交叉设计比实施平行设计更复杂，因此需要更密切的试验监查。每个试验阶段的治疗对后一阶段的延滞作用称为延滞效应。采用交叉设计时应避免延滞效应，即在每个试验阶段后需安排足够长的洗脱期或有效的洗脱手段，以消除第一阶段试验所产生的延滞效应。

最简单的交叉设计是2×2形式，对每个受试者安排两个试验阶段，分别接受两种器械治疗，而每一受试者在第一阶段接受何种器械是随机确定的，第二阶段必须接受与第一阶段不同的另一种试验用器械。因此，对于2×2交叉设计，每个受试者均需经历如下几个试验过程，即：准备阶段、第一试验阶段、洗脱期和第二试验阶段。在两个试验阶段分别观察两种试验用器械的疗效和安全性。

交叉设计的统计分析同样比平行设计更复杂，因为病人对任何试验阶段的疗效通常与对另一试验阶段的疗效相关联，因为同一个病人应用了不止一个试验阶段的治疗。因此，统计分析时需检验是否有延滞效应存在。

交叉设计常用于比较同一器械的两种或多种不同治疗强度的临床疗效。交叉设计应尽量避免受试者的失访。

3. 析因设计　可应用于医疗器械临床试验的第3种设计是析因设计。当一个医疗器械与一个治疗(如药物治疗)相比较时，经常使用这种方法。这种研究设计可回答如下问题：是否器械独自起作用？或是否器械与药物治疗相互影响，联合产生更强的作用？

本设计的不足之处是实施起来更复杂，因此申办者必须保证研究者严格按照研究方案实

施临床试验。

析因设计也许需要更大的样本量，但是由于这种设计类型基本上是两个临床试验合并成一个，因此效率更高。但是，如果试验的样本量是基于检验主效应计算的，则估计器械与药物的交互作用时会使检验效能降低。

还有一些其他的试验设计，如成组序贯设计、适应性设计等，使得评价起来更复杂。总之，所选择的设计应与申办者的目的相适应。研究目的也许导致复杂的研究，因此需要仔细地对研究进行设计、监查和评价。有时，可以通过限制试验范围的方法，来设计不太复杂的试验。但是应该认真地考虑能否这样做，因为这将导致对器械应用范围的限制。

4. *目标值法的单组设计*　对于某些器械，如果存在本研究领域临床认可的、国内/国外公认的疗效/安全性评价标准（如 FDA/SFDA 指导原则、ISO 标准、国标或部标等规范或指南），其中明确指出了该器械的主要疗效/安全性评价指标及其评价标准，那么可以以此评价标准为目标值计算临床试验样本量，并进行符合该目标值的单组试验。试验结果的评价，同样应采用主要疗效评价指标的 95%可信区间。这种与目标值进行比较的单组试验，称作 OPC（optimal performance criteria）研究。

（三）医疗器械临床试验比较类型

临床试验中比较的类型，按统计学中的假设检验可分为优效性检验、等效性检验和非劣效性检验。选择何种比较类型，应从临床实践角度出发考虑，并在制定研究方案时确定下来。

1. *优效性检验*　优效性检验的目的是显示试验器械的治疗效果优于对照器械，其零假设为 $H_0: \pi_T - \pi_C \leqslant \Delta$，备择假设为 $\pi_T - \pi_C > \Delta$。对照器械既可以是无任何疗效的安慰器械、也可以是疗效确证的阳性对照器械。采用何种器械作对照，应按照临床需要、并符合伦理学原则科学地设计。优效性检验的评价方法，类似于传统的假设检验。

2. *非劣效和等效性检验*　出于伦理学或者可操作性考虑，有时采用安慰器械作为对照器械往往在临床实践中不具有可行性。因此，在临床试验设计时，往往更多的采用疗效确实的阳性器械作为对照，即与已经被批准上市的、具有确切临床效果的同类器械进行对比，此时希望通过试验验证的结论为：试验器械的疗效与阳性对照器械实质等同，其假设检验为非劣效检验。即试验器械的疗效可以稍劣于对照器械，但此差距不能超过临床认可的非劣效界值。评价非劣效试验的结果时，传统假设检验的 P 值没有意义，应采用疗效差值的 95%可信区间进行评价。

等效性检验与非劣效检验类似，也是与已经被批准上市的、具有确切临床效果的同类器械进行对比。只是结果判断时，疗效差值的 95%可信区间必须落在临床事先给定的等效性界值区间之内。

非劣效界值和等效性界值均需在方案设计阶段由临床医生确定，且用于样本量计算。

等效性检验的目的是确认两种或多种器械的效果差别大小在临床上并无重要意义，即试验器械与对照器械在疗效上相当。而非劣效性检验目的是显示试验器械的治疗效果在临床上不劣于对照器械。在显示以上两种目的的试验设计中，对照器械的选择要非常慎重。所选择的对照器械应是已在临床广泛应用的、对相应适应证的疗效/安全性已被证实，使用它可以有把握地期望在对照试验中表现出相似的效果。理想情况下，对照组有曾与安慰器械对照的临床试验证据。

进行等效性检验或非劣效性检验时，需预先确定一个等效界值（上限和下限）或非劣效界

值(下限),这个界值不应超过临床上能接受的最大差别范围。等效界值或非劣效界值的确定需要由主要研究者(临床专家)从临床角度决定,但同时应满足统计学要求。

(四)样本含量

临床试验的目的是在目标人群的样本中收集有关医疗器械安全性和有效性的数据。然后用统计分析将结论推断到与试验人群具有相同特征的目标人群。而只有将研究问题翻译成具有人群特征的数学关系表达式,才能进行统计推断。我们称这个数学关系表达式为假设。对该假设所做的检验应该为该研究问题提供明确的答案。每个临床试验的样本量应符合统计学要求。

例如,如果研究问题是"对于某个疾病A,用试验器械治疗后,试验器械组主要后果的均值大于对照组的均值吗"?该问题产生两个假设。一个零假设:治疗组病人治疗后的均值等于(或差于)对照组的均值;另一个是备择(或研究)假设:治疗组病人治疗后的均值大于对照组的均值。当将从样本得出的结论推断到总体时,可能会犯两类决策错误。如果样本显示器械治疗组的均值大于对照组的均值(即拒绝无效假设),而人群中没有发现两组均值有差异时,就犯了Ⅰ类错误(也叫α错误)。另一方面,如果样本显示两组均值间无差异(即接受无效假设),而实际中,器械治疗组均值确实大于对照组时,就犯了Ⅱ类错误。犯Ⅱ类错误的概率也被称为β错误,统计效能就被定义为1−β。

在用于假设检验的样本量计算中均要用到上述两个错误概率,同时还应包含临床有意义的差值,即由临床专家确定的具有临床显著意义的结果变量间的差别。最常用的样本量计算公式包含处于分子位置的临床有意义差值的变异度的估计,和处于分母位置的临床有意义差值的估计。因此,对于一个已知的后果变量,变异度越大,所需要的样本量也就越大。类似地,当变异度已知时,要检测的临床差异越小,所需要的样本量就越大。

总而言之,样本的大小通常按照具体受试产品的特性、主要评价指标及其参数来确定。其具体计算方法以及计算过程中所需用到的统计量的估计值及其依据应在临床试验研究方案中给出,同时需要提供这些估计值的来源依据。在确证性试验中,样本量的确定主要依据已发表的文献资料或预试验的结果来估算。Ⅰ类错误常用5%,Ⅱ类错误应不>20%(检验效能不应低于80%)。

样本量的具体计算方法见第4章。

十、医疗器械临床试验的统计分析

当临床试验到达统计分析阶段时,除非在试验中发生了出乎预料的偏倚,否则应该预先在研究方案中把统计分析方法确定下来。大多数情况下,在试验实施过程中引入的任何大的偏倚,都无法通过统计分析的调整过程对其进行满意的调整。

临床试验中数据分析所采用的统计分析方法应是传统的、经典的统计分析方法,统计分析软件应是国内外公认的、经过国内外权威机构认证的、可再现全部编程过程的统计分析软件,如SAS®等。

(一)统计分析计划书

统计分析计划书由生物统计学家起草,并与主要研究者商定,其内容比试验方案中所规定的统计分析更为详细。

统计分析计划书上应列出研究背景、研究目的、统计分析集的选择、主要指标、次要指标、统

计分析方法、疗效及安全性评价方法等，并按预期的统计分析结果列出空白的统计分析表备用。

统计分析计划书应形成于试验方案和病例报告表完成之后。在临床试验实施过程中，可以进行修改、补充和完善。在盲态审核时（如适用）可再次进行修改完善。但是在第一次揭盲之前必须以文件形式予以确认，此后不能再变动。

（二）数据合并与转换

由于临床试验法规（5号令）要求，临床试验必须同时在两家或两家以上医院进行（多中心临床试验），因此，研究结束统计分析时，申办者除了出具每家医院的临床试验结果外，还应出具所有参加医院的总的临床试验结果报告。因此，必须将各个研究中心的数据进行合并，以便研究者撰写临床试验总报告。

数据合并前必须确认所有研究中心的临床操作过程都是按照研究方案中所描述的方式进行的，以及检查各个诊断因素之间是否均衡。有时，某个研究中心得到的数据会显著偏离其他中心的数据。申办者必须调查由研究中心造成的所有相关结果，并向审评单位报告这些情况，以决定为什么该研究中心会出现不同的结果。

数据变换是为了确保资料满足统计分析方法所基于的假设，变换方法的选择原则应是公认常用的。一些特定变量的常用变换方法已在某些特定的临床领域得到成功地应用。

统计分析之前对关键变量是否要进行变换，最好根据以前的研究中类似资料的性质，在试验设计时就作出决定。拟采用的变换（如对数、平方根等）及其原理需在试验方案中说明。

（三）统计分析人群（数据集）

用于统计的分析集需在试验方案的统计部分中明确定义，并在盲态审核时确认每位受试者所属的分析集。在定义分析数据集时，需遵循以下两个原则：①使偏倚达到最小；②控制Ⅰ类错误的增加。

应按照意向性分析（intention to treat，ITT）的基本原则，主要疗效评价指标的统计分析应包括所有随机的受试者。但是，在实际操作中往往很难做到，因此，常采用全分析集进行分析。全分析集（full analysis set，FAS）是指尽可能接近按意向性分析原则的理想的受试者集。该数据集是由所有随机的受试者中，以最小的和合理的方法剔除后得出的。在选择全分析集进行统计分析时，对主要疗效指标发生缺失时的估计，可利用最接近的一次观察值进行结转（last observation carry forward，LOCF）；对缺失数据进行截转还存在很多方法，截转时应选择对结果相对保守的方法，必要时需提供采用不同结转方法的灵敏性分析结果。

受试者的“符合方案集”（per protocol，PP），亦称为“可评价病例”样本。它是全分析集的一个子集，这些受试者应完成全部试验，并且其主要疗效评价指标没有严重违背研究方案。将受试者排除在符合方案集之外的理由应在统计分析报告和临床试验报告中写明。

在确证性试验中，对器械进行有效性评价时，宜同时用全分析集和符合方案集进行统计分析。当以上两种数据集的分析结论一致时，可以增强试验结果的可信度。当不一致时，应对其差异进行细致的讨论和解释。如果从符合方案集中排除受试者的比例太大，则对试验的总有效性会产生疑问。

全分析集分析法所得到的疗效评价是保守的估计，但更能反映以后临床实践中的真实情况。应用符合方案集可以显示试验器械按规定的方案使用的效果。但可能高估以后临床实践中的疗效。对于优效性研究更倾向参考全分析集的结果，而对于非劣效或等效研究，则更倾向参考符合方案集的结果。

对安全性评价的数据集选择应在方案中明确定义，通常安全性数据集应包括所有随机化后至少有一次安全性评价的所有受试者。

(四)具体分析方法

统计分析应建立在正确、完整的数据基础上，采用的统计模型应根据研究目的、试验方案和观察指标选择，一般可概括为以下几个方面。

1. 临床试验基本情况描述　在临床统计报告的开始，通常用语言或表格对受试者的参与情况做出总结。包括各研究组别入选的受试者数、全分析集(FAS)对应的受试者例数及占总入选人数的百分比、符合方案集(PPS)对应的受试者例数及百分比、、安全集(SS)对应的受试者数及百分比；整个试验过程中，因违背研究方案而被剔除、或者没能完成研究中途失访的受试者，也应分别描述其例数及百分比，并且列明剔除或失访的具体原因。

2. 基线分析　不论临床试验是否采用随机的方法，基线观测值均应基于治疗前所有患者的值。

基线数据的评价有助于标识治疗组间必须均衡的因素，如病人目前的病情、伴随用药、治疗、年龄、性别、社会经济状况、既往病史以及其他可能影响主要终点的因素。对基线数据的评价允许选择及采用使潜在偏倚最小化的方法。例如，对那些已知会影响主要终点的因素，可在分配治疗时采用分层或均衡分配的方法。

如果在临床试验过程中发现影响主要终点的因素在两组间不均衡，则在数据分析过程中就应该使用调整或标化的方法来使各组间的不均衡达到最小。目前国际上较流行采用倾向性评分(propensity score)方法进行调整。

描述参加试验者的人口统计学资料(如性别、年龄等)及伴随疾病等，同时，做组间均衡性检验(如为平行组设计)。

3. 期中分析　是指正式完成临床试验前，按事先制订的分析计划，比较处理组间的有效性和安全性所做的分析。由于期中分析的结果会对后续试验的结果产生影响，因此，一个临床试验的期中分析次数应严格控制。同时，任何设计不良的期中分析都可能使研究结果有误，所得结论缺乏可靠性，因此应避免设计不良的期中分析。如进行了计划外的期中分析，在研究报告中应解释其必要性，提供可能导致的偏倚的严重程度以及对结果解释的影响。最后，期中分析应考虑Ⅰ型误差膨胀的问题。

4. 验证假设　在开始一个详细的统计分析之前，有必要对预计的统计分析中使用的假设进行验证。这些假设包括用于假设检验或估计的概率分布的主要特征；诊断因素在各中心及各组间分布的相似性；以及验证变量间可疑的关联(相关或不相关)。

对统计检验中要用到的分布和方差假设进行验证是非常重要的。只有当所有假设被验证时，此种统计检验才能被应用。例如，假如假设服从正态(高斯)分布，那么应对数据用适当的统计方法进行检验，以保证数据没有显著地偏离正态分布。如果数据显著地偏离了正态分布，那么就要使用其他的更合适的检验方法，比如非参数方法(自由分布)。

评价各因素在各研究中心以及各对照组间是否平衡也是非常必要的。任何观测到的不平衡都必须进行校正，以保证最终进行比较的样本组间具有可比性。如果需要调整的变量数不多，并且要调整的变量与因变量高度相关，那么协方差分析将是一种非常有效的调整工具。但是，如果需要调整的变量数很多，要想把所有的变量都调整得很好，将是一件很困难的事情。因此，应该非常严谨地设计和实施临床试验，因为希尔(Hill，1967)说过："如果考虑不周详，怀

着各种侥幸心理开始试验，寄希望于最终的统计分析可以解决各种问题，这种做法必将导致灾难性的后果。”

如果统计分析时假设变量是独立的，但实际上变量间是相关的或不独立的，则可导致假设检验的重大错误。

5. *参数估计、置信区间估计与假设检验* 参数估计、置信区间估计和假设检验是对主要指标及次要指标进行评价和估计的必不可少的手段。基本上，所有的比较分析都要进行假设检验。统计分析报告中应明确地陈述要检验的假设、待估计的处理效应、选择的统计检验方法以及所涉及的统计模型等。处理效应的估计应同时给出可信区间，并说明计算方法。假设检验应明确说明所采用的是单侧还是双侧，如果采用单侧检验，应说明理由。

在某些情况下，可应用可行的（历史）数据对疾病的进展或其他特征建立数学模型。临床试验中收集的数据可以被用来验证模型，即通过将模型中的设定特征与调查中得到的结果进行比较来验证。这种类型的比较可用来构成模型特征的假设检验。

6. *协变量分析* 除器械本身的作用以外，常常还有其他一些因素可影响器械的有效性评价，如受试者的基线情况、不同研究中心之间的差异（中心效应）等。这些因素在统计学上可作为协变量处理。在试验前应深思熟虑地识别可能对主要疗效指标有重要影响的协变量及如何进行分析以提高估计的精度，补偿处理组间由于协变量不均衡所产生的影响。

在多中心临床试验中，如果中心间处理效应是齐性的，则在模型中常规地包含交互作用项将会降低主效应检验的效能。因此对主要指标的分析如采用一个考虑到中心间差异的统计模型来研究处理的主效应时，不应包含中心与处理的交互作用项。如中心间处理效应是非齐性的，则处理效应的解释是复杂的。

7. *疗效评价* 对主要指标及次要指标进行评价和估计的必不可少的手段是进行参数估计、可信区间和假设检验。基本上，所有的比较分析都要进行假设检验。统计分析报告中应明确地陈述要检验的假设、待估计的处理效应、选择的统计检验方法以及所涉及的统计模型等。处理效应的估计应同时给出可信区间，并说明计算方法。假设检验应明确说明所采用的是单侧还是双侧，如果采用单侧检验，应说明理由。

非劣效试验应该使用待估计参数的点估计及其95%可信区间来评价。

除器械本身的作用以外，常常还有其他一些因素可影响器械的有效性评价，如受试者的基线情况、不同研究中心之间的差异（中心效应）等。这些因素在统计学上可作为协变量处理。一般来说，对于连续型主要疗效评价指标（如冠脉支架手术中的晚期管腔丢失等），常使用调整中心和基线效应的协方差分析；对于离散型主要疗效评价指标（如成功、失败等），常使用多元Logistic回归分析方法，此时，中心和基线效应作为协变量放在模型中进行调整。也可使用CMH χ^2的方法进行调整。对于生存分析资料，还可以使用Cox比例风险回归模型进行调整。

在某些医疗器械临床试验中，有时由于伦理原因或可行性原因，很难实行随机双盲临床试验，因此可能会碰到基线不均衡的研究，此时可使用目前国外流行的倾向性得分法（propensity score，PS）进行调整。

因此，在试验前应深思熟虑地识别可能对主要疗效指标有重要影响的协变量及如何进行分析以提高估计的精度，补偿处理组间由于协变量不均衡所产生的影响。

8. *安全性评价* 临床试验中，安全性评价是非常重要的一个方面。在临床试验的早期（如适用），这一评价主要是探索性的，且只对不良反应明显的表现敏感，在后期，器械的安全性

评价一般通过较大的样本来更加全面的了解。后期的对照试验，是一个重要的以无偏的方式探索任何新的、潜在的器械不良反应的方法。

为了说明在安全性方面与其他器械比较的优效性或等效性，可设计某些试验，这种评价需要相应的确证性试验的支持，这与相应的有效性评价的要求是相同的。

器械安全性评价的常用统计指标为不良事件发生率和不良反应发生率。对于试验时间较长、有较大的退出治疗比例或死亡比例时，需用生存分析计算累计不良事件发生率。用于评价器械安全性的方法以及度量准则依赖于非临床研究和早期临床研究的信息、器械使用方法、受试者类型以及试验的持续时间等。而构成安全性评价的资料则主要来源于临床不良事件、实验室检查（包括临床化学、血液学）及生命指征等。

从受试者中收集的安全性变量应尽可能全面，包括受试者出现的所有不良事件的类型、发生时间、严重程度、处理措施、持续的时间、转归及是否认为与试验器械有关等。

所有的安全性指标在评价中都需十分重视，其主要分析方法需在研究方案中指明。所有的不良事件均需报告，无论是否认为与试验器械有关。在评价中，研究人群的所有可用资料均需说明。实验室应提供检查指标的度量单位以及参考值范围等。

在大多数的临床试验中，对安全性的评价常采用描述性统计分析方法进行。

（五）统计分析报告

为了给主要研究者撰写临床试验总结报告提供素材，在统计分析结束后，生物统计学专业人员应写出统计分析报告。

统计分析报告由文字、统计图和统计表格组成，一般包括如下内容：临床试验的目的、研究设计、随机方法、对照组的选择、是否双盲、主要（次要）疗效评价指标的定义、统计分析数据集的定义等。其次对统计分析报告中涉及的统计模型，应准确而完整地予以描述，如选用的统计分析软件、统计描述的内容、对检验水准的规定、以及进行假设检验和建立可信区间的统计学方法。如果统计分析过程中进行了数据转换，应同时提供选择数据转换的基本原理。统计分析结论应用精确的统计学术语予以阐述。最后，按照统计分析计划书设计的统计分析表格详细描述统计分析结果。

对器械进行有效性评价时，应给出每个观察时间点的描述性统计分析结果。列出检验统计量、P 值。例如，两个样本的 t 检验的结果中应包括每个样本的数量、均值、标准差、中位数，最小值、最大值、两样本比较的 t 值和 P 值。用方差分析进行主要指标有效性分析时，至少应包括各中心的均值和标准差，考虑各种治疗器械、各中心和基线值的协方差分析表。对于交叉设计资料的分析，应包括治疗器械顺序资料、每个阶段开始时的基线值、洗脱期及洗脱期长度、每个阶段中的脱落情况、还有用于分析治疗、阶段、治疗与阶段的交互作用方差分析表。器械的安全性评价，主要以描述性统计分析为主，包括使用器械情况（使用器械持续时间等）、不良事件发生率及不良事件的具体描述（包括不良事件的类型、严重程度、发生及持续时间、与试验器械的关系等）；实验室检验结果在试验前后的变化情况；发生的异常改变及其与试验用器械的关系及随访结果等。

总之，统计分析报告是非常重要的，是主要研究者撰写临床试验总结报告的基础，也是器械上市报批审评的基础。

（李　卫　王　杨　郭　晋　成小如　孙　毅　贾　宣）

第4章 医疗器械临床研究试验设计

本章将列举一些医疗器械临床试验中涉及的试验设计(experimental design)方法，对于各类试验设计对应的统计分析方法以及SAS软件实现将在后续章节介绍。

一、平行组设计

当仅考察一个因素，该因素分成若干个水平，将受试对象随机的分配到各个水平组中进行试验，这种试验设计类型称为平行组设计(parallel group design)。平行组设计中，各组受试者在试验中处于相同的条件，唯一的不同点是各组所使用的器械不同，如有的是试验器械，有的是对照器械，最后根据试验结果作出统计分析。平行组设计是所有试验设计中最简单、也是最常见的一种试验设计类型。

通常，根据试验方案的需要，可为试验组设置一个或多个对照组，试验器械也可按照若干种治疗强度设组。对照器械的选择应符合试验方案的要求。对照组可分为阳性或阴性对照。阳性对照一般采用按所选适应证的当前公认的有效器械，阴性对照一般采用疗效未被业界证实的安慰器械，但必须符合伦理学要求。

【例4-1】 本临床试验采用多中心、单盲、随机对照方法，验证西罗莫司(雷帕霉素)可降解涂层钴铬合金冠状动脉药物洗脱支架系统的有效性和安全性，评价该支架的输送系统。与药物洗脱支架进行对照。按照入选和排除标准选取合适患者参加本次验证，对所有入选患者在270d(±30d)时进行造影随访，以标准的定量冠状动脉造影(QCA)评测获得的晚期管腔丢失为主要疗效指标以评价试验产品的有效性。以试验随访过程中的主要心脏不良事件(MACE)为主要安全性指标以评价试验产品的安全性。

【例4-1分析】 本研究是一个平行组设计，试验器械为“NOYA雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统”，对照器械“Firebird2药物洗脱支架”，主要终点为定量冠状动脉造影(QCA)，各中心内比较，可使用t检验，如不满足参数检验条件，可使用相应的非参数方法；中心间比较为在检验各中心方差齐性后，使用调整中心效应和基线效应的协方差分析。两组心脏不良事件(MACE)的比较可采用定性资料统计分析方法，CMH(Cochran Mantel-Haenszel)χ^2检验。

二、非劣效临床试验设计

以安慰剂作为对照的随机双盲试验在临床试验中通常被视为金标准。但随着医疗技术的发展，越来越多有效产品的出现，开发疗效有突破的新产品变得十分困难。尤其在医疗器械临床试验中，当现有的治疗手段可以为患者提供很大的生存获益时(例如，延长生存时间或防止不可逆的损伤)，单纯地使用安慰剂、空白对照或者低剂量的对照既不符合伦理要求，也无法比

较新产品与已上市的产品之间的优劣。基于这个原因，现在国内外的医疗器械临床试验大都采用了非劣效试验设计。但遗憾的是目前一部分采用非劣效设计的试验往往存在着适用条件不符合，阳性对照选择错误、界值确定不合理等诸多问题，笔者参考了众多文献并结合实际工作中的一些经验，以期为国内正确设计非劣效试验提供参考。

(一)优效性试验与非劣效性试验的比较

优效性试验是指主要研究目的是显示所研究的产品疗效优于对照产品的疗效(阳性或安慰剂对照)的试验。现在常用的 4 种用来验证有效性的平行对照试验类型，其中有 3 种是优效性试验——安慰剂对照，空白对照和剂量组间对照。第 4 种是非劣效试验，采用阳性药物对照组与试验组比较。如果试验目的是为了验证试验组比阳性药物组更有效，那么可以采用优效性设计。但目前更普遍的情况是用来验证试验组与阳性对照组之间的差别不大，在一定程度上可以认为他们的疗效没有本质差异。这里差别不大的概念是指试验组比对照组疗效差，但差的程度非常小，小到可以认为与阳性对照组相比，试验组依然是有效的。如何设计和解释这样的试验，使它能达到上面所说的要求是一个很艰巨的挑战。

优效性试验与非劣效性试验最大的区别在于，一个设计合理执行良好的优效性试验可以在不做进一步假设的前提下，完整地解释显著性检验的结果；也就是说，试验结果本身就可以说明问题，不需要参考其他信息。而非劣效性试验是建立在一些非本试验的信息上的，即阳性对照组的预期疗效。如果在非劣效性试验中阳性对照组完全无效(完全没有预期的效果)，那么即使试验组与阳性对照组只有很小的疗效差异也是毫无意义的，没有任何证据可以表明试验组是有效的。非劣效试验很大程度上依赖于外部信息，它需要参考历史对照试验的数据。

(二)非劣效界值的选择

非劣效试验旨在验证阳性对照组与试验组之间疗效的差异(对照组优于试验组的部分)小于预先制定的非劣效界值。在非劣效试验中，非劣效介值不能大于对照组预期的效应，这种基于对照组效应设定的界值通常称 M1。M1 不是从非劣效试验中得到的，设定时应参考对照组以前做过的一个设计良好的安慰剂对照试验的结果或综合多个试验结果进行 Meta 分析。

为了方便表达，我们将高优指标定义为数值越大疗效越好，低优指标定义为数值越小疗效越好，比如有效率是高优指标，死亡率是低优指标。若记非劣效界值 $\Delta>0$。

非劣效临床试验的检验假设：

1. 高优指标

$H_0: C-T\geqslant\Delta, \Delta>0$

$H_1: C-T<\Delta$

或

$H_0: \ln(C/T)\geqslant\Delta, \Delta>0$

$H_1: \ln(C/T)<\Delta$

2. 低优指标

$H_0: T-C\geqslant\Delta, \Delta>0$

$H_1: T-C<\Delta$

或

$H_0: \ln(T/C)\geqslant\Delta, \Delta>0$

$H_0: \ln(T/C)<\Delta$

对高优指标，取$(C-P)$区间下限作为阳性对照的疗效估计，记为 M(可以认为本次试验中的阳性对照的疗效有 97.5%以上的可能大于 M)。若取 $M_1<M$，令 $\Delta=M_1$，如果拒绝 H_0，则可间接推论出试验组疗效优于安慰剂，即$C-T<\Delta\Leftrightarrow T-P>C-P-\Delta>0$；若取$M_2=(1-f)M_1$，$0<f<1$，令 $\Delta=M_2$，如果拒绝 H_0，则可推论出试验组非劣效于阳性对照，且至少保持了阳性对照疗效 M 的 f，即 $C-T<\Delta\Leftrightarrow T-P>C-P-(1-f)M_1\Leftrightarrow T-P>f(C-P)$ 。

对于低优指标，构建(P—C)区间估计后，取区间下限作为阳性对照的疗效估计，记为 M。若取 $M_1<M$，令 $\Delta=M_1$，如果拒绝 H_0，则可间接推论出试验组疗效优于安慰剂，即 $T-C<\Delta\Leftrightarrow P-T>P-C-\Delta>0$。若取 $M_2=(1-F)M_1$，$0<f<1$，令 $\Delta=M_2$，如果拒绝 H_0，则可推论出试验组非劣效于阳性对照，且至少保持了阳性对照疗效 M 的 f，即 $T-C<\Delta\Leftrightarrow P-T>P-C-(1-f)M_1\Leftrightarrow P-T>f(P-C)$ 。

以下就高优指标给予详细说明：

在一个安慰剂对照试验中，无效假设(H_0)是指试验组(C)的效应小于等于安慰剂(P)的效应；备择假设是指试验组的疗效大于安慰剂的疗效。

H_0：$C\leqslant P$；　$C-P\leqslant 0$

H_1：$C>P$；　$C-P>0$

多数情况下，统计学上是通过 $C-P$ 的双侧 95%可信区间的下限>0 来确定治疗结果的有效性(相当于单侧 97.5%可信区间的下限)。这表明了试验组的疗效是>0 的。见图 4-1。

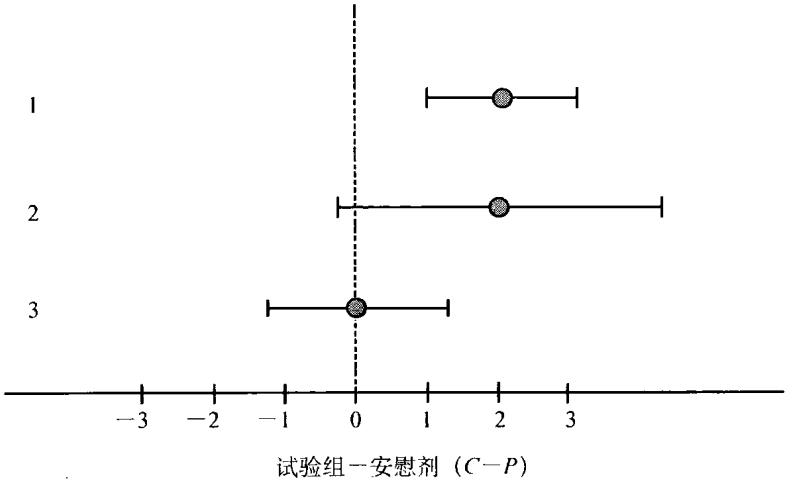

图 4-1　安慰剂对照优效性试验的 3 种结果

1. 点估计的结果为 2；95%CI 下限为 1；结论：试验组是有效的，有效的额度至少为 1。2. 点估计的结果为 2；95%CI 下限<0(可能由于样本量太小)。结论：不能证明试验组有效；点估计的结果为 0；95%CI 下限远<0；结论：没有任何迹象显示试验组有效

$C-P$ 的 95%CI 下限值 >0 时，拒绝 H_0，说明试验组的疗效优于安慰剂。取$(C-P)$区间下限作为阳性对照的疗效估计，记为 M(可以认为本次试验中的阳性对照的疗效有 97.5%以上的可能>M)。为保守起见，建议 M_1 要小于置信区间下限值。取 $M_1<M$，令 $\Delta=M_1$。

在一个非劣效试验中，无效假设是指对照组(C)与试验组(T)的疗效差$(C-T)$大于非劣效界值 M_1，其中 M_1 代表阳性对照组与安慰剂间的疗效差即阳性对照组的绝对疗效。

$H_0: C-T \geqslant M_1$　（试验组比对照组的差，且差距$\geqslant M_1$）

$Ha: C-T < M_1$（试验组比对照组的差，差距$< M_1$）

当 C-T 的双侧可信区间上限$< M_1$ 时，非劣效成立。如果 M_1 的值可以完全反映非劣效试验中阳性对照组的效应，那么想要得到非劣效结果就意味着试验组的效应必须>0（图 4-2）。

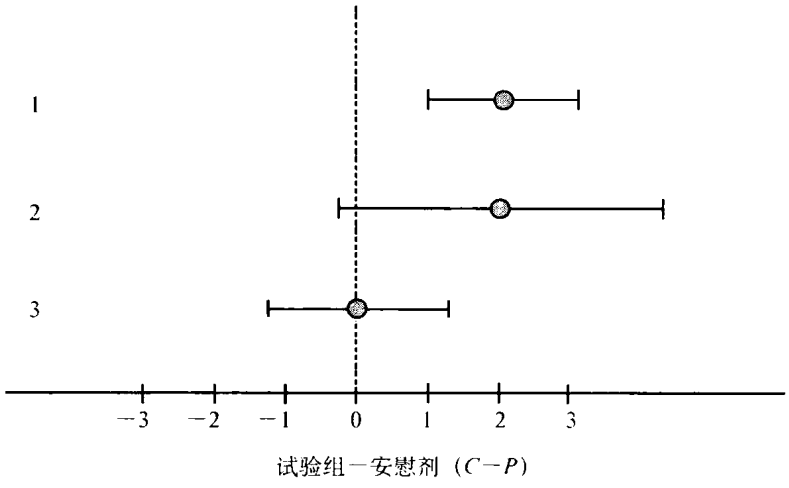

图 4-2　非劣效试验结果——C-T 的值和 95%CI($M_1=2$)为啥是对照组-试验组，这样理解起来有点困难

1. $C-T$ 的点估计值为 0，暗示效应可能相等。$C-T$ 95%CI 的上限为 1，$< M_1$，证明它是非劣效的；2. $C-T$ 点估计值>0；$C-T$ 95%CI 的上限为>2，即$> M_1$；这种情况不是非劣效；3. $C-T$ 的点估计值为 0，暗示效应可能相等。但 $C-T$ 95%CI 的上限为>2，即$> M_1$，这种情况也不是非劣效；4. 点估计的结果偏向 $T(<0)$；证明它是非劣效，但不能证明是优效；5. 点估计的结果偏向 $T(<0)$；证明它是非劣效，且同时也是优效的；6. $C > T$ 且有统计学意义。但由于 $C-T$ 95%CI 上限$< 2(M_1)$，在非劣效界值为 M_1 的情况下，它依然是非劣效的

如果对照组比试验组的疗效好，且差值$> M_1$，那么该试验药就被认为没有疗效。因此，我们用对照组与试验组的差值$(C-T) < M_1$ 来说明药物的有效性，可以用 $C-T$ 的 95%置信区间的上限$< M_1$ 来判断。

M_1 的确定是其中最关键的问题，它不是通过非劣效性试验得出来的（没有设置相应的安慰剂对照）。它的估计需要综合考虑既有阳性对照试验的结果以及考虑历史试验和现试验之间条件环境上的差异。

选定非劣效界值的第二步是设定临床认可的较对照组差的最大值。虽然 M_1 的非劣效性检验能够保证试验组比安慰剂好，试验组的效果>0。但在很多实例中，这并不能保证试验产品的有效性具有临床意义。应用非劣效性检验设计的原因在于需要考虑对照组的有效性，如果试验组的效果只有对照组效果的很小一部分，那么试验组就变得没有价值了。因此在非劣效研究中，需要确定这个界限，一般选择临床认可的最大差值作为非劣效性检验的界值，记为 M_2，$M_2 = f \times M_1 (0 < f < 1)$。$M_2$ 是临床可以接受的试验组较对照组疗效差的最大界值，它应该$< M_1$，因为当 $M_2 > M_1$ 时，表示试验组将失去对照组的所有效果，也就是说，以对照组为参照试验组是完全无效的。f 越小，说明试验组的疗效越接近阳性对照组。但 f 过小，可能导致样本量过大而无法实施试验。f 的意义就是要求试验组至少能保持对照组疗效的比例。所

以要衡量临床意义与样本量大小，选择一个合适的值，一般定在0.5～0.8。

试验组与对照组的非劣效判断，可采用可信区间法；若上限$<M_2$，可认为试验组与对照组相比非劣效。见图4-3。

H_0：$C-T \geq M_2$ （试验组比对照组的差，且差距$\geq M_2$）

Ha：$C-T < M_2$（试验组比对照组的差，差距$<M_2$）

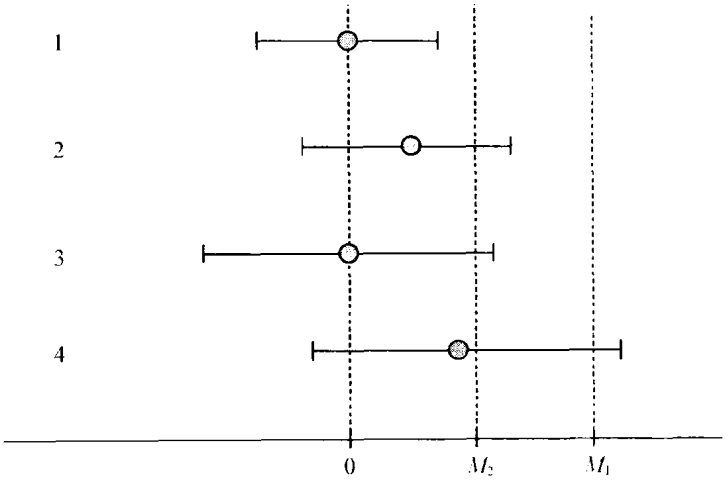

图4-3 阳性对照组与试验组疗效差异（点估计，95%CI）

1. $C-T$点估计值 = 0，双侧95%CI上限$<M_2$，提示试验组有效且试验药物非劣于阳性对照药；2. $C-T$点估计值 > 0，表明阳性对照组更好，95%CI上限$<M_1$但$>M_2$，显示试验组虽然有效，但对阳性药物疗效的丧失不可接受；3. $C-T$点估计值 > 0，95%CI上限$<M_1$但稍$>M_2$，显示试验组有效，但非劣效不成立；4. $C-T$点估计值 > 0，表明阳性对照组更好，95%CI上限$>M_1$，没有任何证据提示试验药有效

如果试验产品具有一些重要的优势（例如在安全性或次要指标方面），则临床认可的界值可以比较宽松。在非劣效界值判定过程中，多步骤和界值假定都是潜在的不可靠因素，会对试验结果和结论造成影响。

从定义上讲，非劣效研究设计提供了两种比较：①试验药与阳性对照药的直接比较；②试验药与安慰剂组的间接对照。非劣效试验中阳性对照药估计的疗效是根据以往试验中的疗效得出，如果以前有过对阳性对照药的疗效估计或者建立了循证医学证据HESDE（historical evidence of sensitivity to drug effect - ICH E-10），考虑到历史试验研究与现行研究的差异，我们必须探讨这个疗效值在新的研究中是否依然适用，是作为恒定值还是以某种方式进行调整后才能用。我们必须确保非劣效试验的研究质量，因为低质量的研究会降低对照药的疗效，会对对照药的药效估计产生不良影响。

（三）非劣效研究的样本量估计

对于非劣效试验来说，样本量估计方法很重要。当试验药是真的非劣效时，使试验有把握度来判断非劣效界值是否已经被减掉通常不是一项轻松的工作。在方案设计阶段，非劣效界值的大小已经被界定好；样本量的大小必须满足疗效差$<M_2$的要求。被减去的界值是样本量

估计中一个重要的环节，但是我们无法知道治疗效果估计的变异度。有时候在新的研究中事件发生率有可能会降低，假如非劣效界值（C－T 的 95%置信区间上限）为增加的风险不大于 25%，那么当阳性对照组的事件发生率由 4%上升为 8%时这个要求将变得更加容易达到，而且在这两种情况下阳性对照组都是 20%优效于安慰剂组。这样的问题在优效性试验的样本量计算时同样存在。

临床试验中所需的样本量应足够大，以确保对所提出的问题给予一个可靠的回答。样本量的大小通常根据试验的主要指标来确定。一般只有一个主要疗效指标。Ⅰ类错误概率常用单侧 0.025，Ⅱ类错误概率应不大于 0.2。

率差非劣效样本量计算公式：

$$n_c \frac{(Z_\alpha+Z_\beta)^2}{(\delta-\Delta)^2}\left[\frac{T(1-T)}{K}+C(1-C)\right] \tag{6.1}$$

其中 T 为试验组率的估算值、C 为阳性对照组率的估算值，$\delta=C-T\geqslant 0$，$\Delta>0$ 为非劣效界值，K 为分组比例，n_c 为对照组样本量，kn_c 为试验组样本的样本量。

率比非劣效样本量计算公式：

$$n_c=\frac{(Z_\alpha+Z\beta)^2}{(\delta-\Delta)^2}\left(\frac{1}{KT(1-T)}+\frac{1}{C(1-C)}\right) \tag{6.2}$$

其中 $\delta=In(C/T)\geqslant 0$，$T$ 为试验组率的估算值、C 为阳性对照组率的估算值，K 为分组比例

均数之差非劣效样本量计算公式：

$$n_c=\frac{(Z_\alpha+Z_\beta)^2}{(\delta-\Delta)^2}\sigma^2\left[1+\frac{1}{K}\right] \tag{6.3}$$

其中 σ^2 为合并方差的估算值，$\delta=C-T\geqslant 0$，T 为试验组均数的估算值、C 为阳性对照组均数的估算值，$\Delta>0$ 为非劣效界值，K 为分组比例

低优指标时(6.1)(6.3)$\delta=C-T\geqslant 0$；(6.3)式中 $\delta=In(T-C)\geqslant 0$

（四）非劣效研究中的潜在偏倚

通常临床上随机优效性试验的结果分析要满足意向性治疗（ITT）的原则，就是说，所有的随机化病人都是根据随机结果来进行相应的治疗。这样分析的目的是避免引起像病人转换治疗方式，选择性偏倚，和脱落/撤回等对观察治疗效果造成影响的偏倚。通常这被认为是保守的策略。在优效性试验中我们也会偏向于进行意向性治疗分析来防止各式各样的偏倚。在非劣效试验中，许多对优效性试验致命的问题，比如不依从性，主要终点指标的错误分类，或者是更常见的度量问题（像噪声的度量），或者是在试验组中大量病例的脱落会使结果不断偏倚直至没有差异，而这个“相对的没有差异”正是非劣效试验想要的结果。这种情况下，试验的正确性将受到质疑，可能会将实际上不存在的非劣效结果当成非劣效。这一点会对试验结果造成很大的影响，所以务必要注意，非劣效试验对试验质量有比较高的要求。虽然在非劣效研究中“被当成试验组”分析经常是主要的分析方法，也有需要信息审查的可能性。因此，对于非劣效研究来说，意向性治疗和“被当成试验组”分析都很重要。这两种分析方法的差异有待进一步检验。最好的建议是，在进行非劣效研究过程时，在试验的设计阶段注意这些潜在的问题，并通过监察来把这些问题最小化，因为它们会严重影响非劣效研究的正确性。

其他任何发生于研究中的偏倚也要在非劣效研究中予以考虑，尤其是在不设盲的研究中怎

样最好地确保终点指标的无偏估计，无偏地纳入患者，以及其他各类潜在偏倚都应该特别注意。

（五）非劣效研究中的非劣效及优效性检验

一般来说，当只有一个终点指标和一个试验组时，非劣效设计可以用来检验优效性，且不必校正Ⅰ类错误，也就是说，检验非劣效性时通过预定的界值得出的95%或者更高的置信区间的值，可以用来检验优效性。我们也可以把它当成是一种两步分析法，在非劣效过程中采用95%置信区间（如果试验药确实是优效的那么总是会成功）分析，然后再接着做优效性检验。这种连续的检验在非劣效和优效试验中的一类错误率都控制在5%以下。与此相反，在一个失败的优效性研究之后进行一个非劣效的研究，通常会得到一个错误的结果，这样的研究通常也会被认为是一个失败的研究。因此，成功的非劣效研究允许再进行优效性研究。一个失败的优效性研究是不允许再进行非劣效研究的。

当非劣效试验有多个终点或者多个试验组时，有效的统计决策将变得非常复杂。先不考虑是非劣效检验还是优效性检验，如果在每个终点水平或者每一组都采用相同的95%置信区间来检验非劣效和优效性，就会使总的一类错误增加，导致在多重比较分析中得出一个或多个错误的结论。因此，对所有在检验中涉及的比较时所犯总体一类错误的估计以及对其做出适当的统计学调整就变得至关重要。

非劣效设计中的有些结果的解释要比优效性检验中的更为灵活，尤其像设计和执行上的问题比如说依从性问题，或者错误分类/测量误差等那些会严重影响优效性研究的问题，显然可以考虑采用非劣效设计。

【例4-2】 对于随机对照部分，根据一个注册研究结果显示，某药物洗脱支架平均置入9个月时，支架内晚期管腔丢失（late loss）为（0.21±0.39）mm；另一项注册研究报道了对112处靶病变平均244d造影随访结果，显示该支架的支架内晚期管腔丢失为（0.12±0.34）mm；此外一项入选超过500例病人的平行对照研究显示，该支架的支架内晚期管腔丢失为（0.18±0.07）mm；综合以上研究报道，将本研究中对照组支架的管腔丢失估计为（0.21±0.35）mm，临床认可的非劣效界值为0.12mm；当统计学显著性水平取（双侧）5%，把握度80%时，

每组的 $n=\frac{2\times 7.9\times 0.35^2}{0.12^2}=134.4\approx 135$，[注：$(Z_{0.05}+Z_{0.2})^2=7.9$]

考虑30%的脱落率后 $n=135/0.7=192.9\approx 193$，两组样本量 $N=193\times 2=386$ 例。

三、交叉设计

交叉设计是按事先设计好的试验次序，在各个时期对受试者逐一实施各种处理，以比较各处理组间的差异。交叉设计是将自身比较和组间比较设计思路综合应用的一种设计方法，可以很好地控制个体间的差异，同时减少受试者人数。然而，实施交叉设计比实施平行设计更复杂，因此需要更密切的试验监查。每个试验阶段的治疗对后一阶段的延滞作用称为延滞效应。采用交叉设计时应避免延滞效应，即：在每个试验阶段后需安排足够长的洗脱期或有效的洗脱手段，以消除第一阶段试验所产生的延滞效应。

交叉设计的统计分析同样比平行设计更复杂，因为病人对任何试验阶段的疗效通常与对另一试验阶段的疗效相关联，因为同一个病人应用了不止一个试验阶段的治疗。因此，统计分析时需检验是否有延滞效应存在。交叉设计常用于比较同一器械的两种或多种不同治疗强度的临床疗效。交叉设计应尽量避免受试者的失访。

(一)二阶段交叉设计(2×2 交叉设计)

1. 定义　最简单的交叉设计是二阶段交叉设计,对每个受试者安排两个试验阶段,分别接受两种器械治疗,而每一受试者在第一阶段接受何种器械是随机确定的,第二阶段必须接受与第一阶段不同的另一种试验用器械。因此,对于二阶段交叉设计,每个受试者均需经历如下几个试验过程,即准备阶段、第一试验阶段、洗脱期和第二试验阶段。在两个试验阶段分别观察两种试验用器械的疗效和安全性。二阶段交叉设计,又被称为一次交叉设计或 2×2 交叉设计。

2. 特点　该设计可以考察一个具有两水平的试验因素和两个区组因素(即个体差异、测定顺序)对观测结果的影响;实验因素的两个水平施加的先后顺序对两个试验分组的影响是动态平衡的;对于每一受试者而言,均有一个"洗脱期";在此设计中,试验因素和顺序(或阶段)因素均取 2 水平,而受试者应取偶数个,以便配对或均分成样本含量相等的两个组。

3. 设计　先将 $2n$ 个受试对象完全随机地均分成两组,然后,随机地决定其中一组受试对象接受两种处理的先后顺序,另一组接受处理的顺序正好相反。当试验中仅涉及一个具有两水平的试验因素,而且,根据专业需要,希望试验因素的两个水平要先后作用于每一个受试对象,此时,宜选用交叉设计。

【例 4-3】　用 A、B 两种闪烁液分别测定血浆中的 3H-cGMP。第一阶段 1,3,4,7,9 号用 A 液测定,2,5,6,8,10 号用 B 液测定;第二阶段 1,3,4,7,9 号用 B 液测定,2,5,6,8,10 号用 A 液测定。试比较 A,B 两种闪烁液测定结果之间的差别有无统计学意义(表 4-1)。

表 4-1　考察 A,B 两种闪烁液测定血浆中^{3}H-cGMP 的交叉试验结果

受试者	闪烁液种类(血浆中的^{3}H-cGMP)	
测定顺序	1	2
1	A(760)	B(770)
2	B(860)	A(855)
3	A(568)	B(602)
4	A(780)	B(800)
5	B(960)	A(958)
6	B(940)	A(952)
7	A(635)	B(650)
8	B(440)	A(450)
9	A(528)	B(530)
10	B(800)	A(803)

【例 4-3 分析】　本例中的一个试验因素是"闪烁液",有 A,B 两个水平;两个区组因素分别是测定顺序和受试者号;观测的定量指标为"血浆中的 3H-cGMP"。由于两种处理在同一对受试对象之间施加的顺序是交叉开的,是由 5 个 2×2 拉丁方设计纵向排列的结果,故为二阶段交叉设计。

(二)三阶段交叉设计

1. 定义　当试验中涉及一个具有两水平的试验因素,设其水平为 A,B。根据需要,希望

该试验因素的两个水平要在3个时期作用于同一个受试者，其顺序要么是ABA，要么是BAB。该试验设计类型为三阶段交叉设计或二次交叉设计。

2. 特点　该设计可以考察一个具有两水平的试验因素和两个区组因素（即个体差异、测定顺序）对观测结果的影响；试验因素的两个水平施加的次数在两个试验分组中达到动态平衡；对于每一受试者而言，均有两个"洗脱期"；较二阶段交叉设计而言，携带效应可能更加突出。

应尽可能避免使用三阶段交叉设计，因为对每一位受试者来说，使用器械的种类是不平衡的，即对于一个受试者，使用两次A器械治疗而使用一次B器械治疗，或使用两次B器械治疗而使用一次A器械治疗，当然可以进行四阶段交叉设计，即每位受试者接受处理的顺序均为ABAB或BABA，以便ABAB与BABA这两种平衡组合在同一对受试者或两组受试者中交叉实施，从试验设计的均衡原则角度看，其效果更好一些。不论三阶段交叉设计还是四阶段交叉设计，携带效应非常突出，因此这种试验设计类型在医疗器械临床试验中是不多见的。

3. 设计　先将$2n$个受试对象完全随机地均分成两组，然后，随机地决定其中一组受试对象在3个时期接受两种处理的先后顺序（如BAB），另一组接受处理的顺序正好相反（ABA）。

（三）3×3交叉设计

1. 定义　将3种处理（不同的器械或者药物）分3个时期先后给予同一个受试者，观察受试者接受每种处理后的反应。为了减少总是A处理、B处理、C处理这样一种固定顺序所带来的误差（称为顺序误差），需要将A，B，C处理施加的顺序随机化，共有3！=6种排列方式，即ABC，ACB，BAC，BCA，CAB，CBA，故至少要将受试者分为6个组，每组中至少要有一位受试者，若各组中有多位受试者，最好数目相等为宜。

2. 特点　该设计可以考察一个具有3个水平的试验因素及两个区组因素（即个体差异、测定顺序）对观测结果的影响；试验因素的3个水平施加的顺序随机化排列，在6种全排列组中达到动态平衡；对于每一受试者而言，均有两个"洗脱期"；较二阶段交叉设计而言，携带效应可能更加突出。

3. 设计　受试对象的个数为6的倍数，将它们完全随机地均分入6个处理组，这6个处理组中的受试对象在3个试验时期接受处理因素的3个水平A，B，C的顺序依次为：ABC，ACB，BAC，BCA，CAB，CBA。当试验中涉及一个具有三水平的试验因素，根据需要，要求试验因素的3个水平要在3个试验时期作用于每个受试对象时，可选用此设计。

【例4-4】　假定某临床研究为了研究3种理疗仪A，B，C对关节疼痛的治疗效果，特别希望考察3种理疗仪先后用于同一位患者身上会产生怎样的疗效，进行了如下的临床预试验研究。试验设计方案如下：从符合入选标准的关节疼痛患者中随机地选取12名，再将他们随机地均分成6个组，每组患者接受理疗仪治疗的顺序依次为ABC，ACB，BAC，BCA，CAB，CBA，每次治疗后患者依据评分标准进行评分，设计与资料见表4-2。

【例4-4分析】　本试验涉及一个3水平的试验因素（即理疗仪的种类），其3个水平分别为A，B和C理疗仪；两个区组因素分别是：顺序因素和患者个体。根据对受试对象分组的方法可知，表4-2为"3×3交叉设计"。

表 4-2　A,B,C 3 种理疗仪用于每一关节痛患者的疗效评分结果

受试者编号	组别	疗效评分结果		
		时期:1	2	3
1	1	A 171	B 146	C 164
2	1	A 145	B 125	C 130
3	2	A 192	C 150	B 160
4	2	A 194	C 208	B 160
5	3	B 184	A 192	C 176
6	3	B 140	A 150	C 150
7	4	B 136	C 132	A 138
8	4	B 145	C 154	A 166
9	5	C 206	A 220	B 210
10	5	C 160	A 180	B 145
11	6	C 190	B 145	A 160
12	6	C 180	B 180	A 208

四、析因设计

可应用于医疗器械临床试验的第 3 种设计是析因设计。当一个医疗器械与一个治疗(如药物治疗)相比较时,经常使用这种方法。这种研究设计可回答如下问题:是否器械独自起作用?或是否器械与药物治疗相互影响,联合产生更强的作用?

本设计的不足之处是实施起来更复杂,因此申办者必须保证研究者严格按照研究方案实施临床试验。析因设计也许需要更大的样本量,但是由于这种设计类型基本上是两个临床试验合并成一个,因此效率更高。但是,如果试验的样本量是基于检验主效应计算的,则估计器械与药物的交互作用时会使检验效能降低。

1. 定义　如果临床试验所涉及的试验因素的个数 $m \geq 2$,当各因素在试验中同时实施且所处的地位基本平等,因素之间存在一级(即 2 个因素之间)、二级(即 3 个因素之间)乃至更复杂的交互作用且需要加以考察时,所采用的一种多因素试验设计类型叫做析因设计。

2. 特点　此设计具有下列 4 个突出特点:其一,实验中涉及 m 个实验因素($m \geq 2$);其二,所有 m 个试验因素的水平都互相搭配到,构成 s 个试验条件(s 为 m 个因素的水平数之积);其三,做试验时,每次都涉及全部因素,即因素是同时施加的;其四,进行统计分析时,将全部因素视为对终点的影响是同等重要的,具体体现在分析每一项(包括主效应和交互效应)时所用的误差项是相同的,它被称为模型的误差项。

由于上述的 4 个特点,决定了析因设计有以下突出的优点和缺点:优点是它可以用来分析全部主效应(即各个单因素的作用)和因素之间的各级交互作用(即任何两个因素之间的交互作用、任何 3 个因素之间的交互作用)的大小;缺点是设计较为复杂,需要更大的样本量,研究者常无法承受。

3. 设计　利用横向与纵向两个方向来排列全部试验因素及其水平,使试验因素之间的全部水平组合都能以纵横交叉的形式呈现出来。设试验因素的个数为 m,当 $m \geq 2$ 且试验因素之间的各级交互作用不可忽视,可以选用此设计。一般来说,医疗器械临床试验该试验设计类

型不十分常用。

【例 4-5】 某临床试验研究人工水球对于肥胖患者的治疗效果，试验组使用新的人工水球(A)，对照组使用常规人工水球(B)；同时希望观察到不同性别人群的减肥情况，并考察其与试验因素(不同人工水球)的交互作用，选择析因设计，资料如表 4-3。

表 4-3 两种人工水球及性别对减肥结果的影响

人工水球	性别	编号	体重减轻量(kg)	皮下脂肪厚度减小量(mm)	BMI 变化(cm)
A	男	1	15	2.5	4.7
		2	10	1.8	3.5
		3	7.8	1.3	2.3
		⋮	⋮	⋮	⋮
		20	3.4	1.0	1.2
	女	1	6.8	1.5	1.7
		2	7	1.0	2
		3	1.9	0.5	1.2
		⋮	⋮	⋮	⋮
		20	3.9	1.4	1.7
B	男	1	10	2.0	4
		2	19	3.1	6.5
		3	15	3.0	5.7
		⋮	⋮	⋮	⋮
		20	14	2.1	6.1
	女	1	13	2.5	5.7
		2	15.5	2.7	6.5
		3	18	3.1	5.2
		⋮	⋮	⋮	⋮
		20	8.5	2.4	3.4

【例 4-5 分析】 本研究除了考察两种人工水球对减肥患者疗效的差别，同时认为性别的不同对于减肥同样是有影响的，并需要考察两者的交互作用，每个组合下的样本量是充足的(需要经过样本含量计算)，因此该临床试验设计可以认为是析因设计。

五、目标值法的单组设计

对于某些器械，如果存在本研究领域临床认可的、国内/国外公认的疗效/安全性评价标准(如：FDA/SFDA 指导原则、ISO 标准、国标或部标等规范或指南)，其中明确指出了该器械的主要疗效/安全性评价指标及其评价标准，那么可以以此评价标准为目标值计算临床试验样本量，并进行符合该目标值的单组试验。试验结果的评价，同样应采用主要疗效评价指标的 95%可信区间。这种与目标值进行比较的单组试验，称作 OPC(optimal performance criteria)研究。

六、重复测量设计(随访研究)

受试对象接受不同处理后，其观测指标的数值会随着时间的推移发生动态变化，为了比较

准确地描述和分析这种变化，需要在不同时间点上从每位受试对象身上观测指标的数值，这种考察和评价处理因素和时间因素对观测指标影响的试验设计方法就是重复测量设计。在医疗器械临床试验中，当器械用于受试对象后，经常会对其进行重复观测（随访），此时的研究设计类型属于重复测量设计。在重复测量试验设计中，主要终点既可能为定量指标也可能为定性指标。

1. 定义　具有重复测量的设计，即在给予某种处理后，在几个不同的时间点上从同一个受试对象上重复获得指标的观察值；有时是从同一个个体的不同部位（或组织）上重复获得指标的观测值。

2. 特点　此设计具有一个明显的特点，即在不同条件下，从同一个受试对象身上观测到 k 个数据（$k \geq 2$）；测自同一受试者的 k 个数据之间具有不相等的相关性，即间隔越近相关性越强，反之亦然。

3. 设计　先考虑试验分组因素，即根据某个或某些试验因素将受试对象完全随机地分成几个独立的组。例如，有 3 种不同的治疗器械治疗某病的患者，则属于具有 3 水平的试验分组因素；若有 3 种治疗器械，并伴随用药，药物可取两种剂量，每位患者只能用一种特定剂量，则共有 6 个试验分组，它们由治疗器械的种类与伴随用药的剂量两因素组合而成。然后，考虑在重复测量方向上有几个因素。例如，考察患者使用治疗器械并服药后在 5 个不同时间点（手术后、1 个月、3 个月、6 个月、1 年）进行随访，这个研究即为 3 因素的重复测量试验。

【例 4-6】 本临床试验采用多中心、单盲、随机对照方法，验证雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统的有效性和安全性，评价该支架的输送系统。与药物洗脱支架进行对照。按照入选和排除标准选取合适患者参加本次验证，对所有入选患者在支架置入后 30d，90d，180d，270d，365d 时进行临床随访；对所有入选患者在 270d（±30d）时进行造影随访，以标准的定量冠状动脉造影（QCA）评测获得的晚期管腔丢失为主要疗效指标以评价试验产品的有效性。以试验随访过程中的主要心脏不良事件（MACE）为主要安全性指标以评价试验产品的安全性。QCA 情况中参照血管直径（mm）如表 4-4。

表 4-4　2 组受试者的参照血管直径（mm）

受试对象使用器械情况	受试对象编号	参照血管直径（mm）		
		时间：术前	术后	270d
雷帕霉素支架	1	2.89	3.03	3.02
	⋮	⋮	⋮	⋮
	166	2.85	3.00	2.94
药物洗脱支架	167	2.81	2.91	2.90
	⋮	⋮	⋮	⋮
	347	2.84	2.99	2.94

【例 4-6 分析】 这个研究受试对象使用器械的情况属于平行组设计，但是对于某些观测指标，如“参照血管直径（mm）”涉及随访观察，在样本失访率不高的前提下，这部分研究应视为重复测量设计，应该使用重复测量设计下的一些统计分析方法，如将术前参照血管直径作为协变量的协方差分析等。

【例 4-7】 目的：在传统抗血小板和抗凝治疗基础上，加用血小板糖蛋白Ⅱb/Ⅲa受体拮抗药替罗非班，能否改善ST段抬高型急性心肌梗死(STEMI)患者冠状动脉介入治疗(PCI)术后的心肌组织灌注水平。方法：通过冠状动脉造影(CAG)确诊为STEMI的患者144例，87例患者接受传统抗血小板和抗凝治疗作为对照组，57例患者在CAG术后在对照组治疗的基础上加用替罗非班作为治疗组，治疗组患者在接受替罗非班静脉负荷量后即刻开始PCI，对照组患者在CAG术后亦立即接受PCI治疗。应用TIMI分级和TIMI心肌灌注分级(TMP)评价术前术后心肌灌注变化，并分析治疗前后患者心电图ST段偏移总和比值(sumSTR)的变化。试分析两组患者治疗前后TIMI分级人数有无变化，数据如表4-5所示。

表4-5 两组患者心肌灌注术前术后TIMI血流分级比较

分组	分级	例数	
	时间：	术前	术后
治疗组	TIMI0～2级	53	5
	3级	4	52
对照组	TIMI0～2级	82	16
	3级	5	71

【例4-7分析】 该资料有3个定性变量，其中时间变量涉及重复测量，属于重复测量的研究，但是列表应按表4-6。

表4-6 两组患者心肌灌注术前术后TIMI血流分级比较

分组	疗效		例数
	时间：	治疗前	治疗后
治疗组		0	1	48
		0	0	5
		1	0	0
		1	1	4
对照组		0	1	66
		0	0	16
		1	0	0
		1	1	5

表内“0”代表“TIMI0～2级”，“1”代表“3级”。

还有一些其他的试验设计，如成组序贯设计、适应性设计等，使得评价起来更复杂。总之，所选择的设计应与申办者的目的相适应。研究目的也许导致复杂的研究，因此需要仔细地对研究进行设计、监查和评价。有时，可以通过限制试验范围的方法，来设计不太复杂的试验。但是应该认真地考虑能否这样做，因为这将导致对器械应用范围的限制。

（李 卫 郭 晋 姜宇莉）

第5章 样本含量的估计

一、检验效能与把握度

临床试验的目的是在目标人群的样本中收集有关医疗器械安全性和有效性的证据，然后用统计分析将试验结论推断到与试验人群具有相同特征的目标人群，这个过程叫做统计推断。而只有将待研究问题翻译成具有人群特征的数学关系表达式，才能进行统计推断。我们称这个数学关系表达式为研究假设。研究假设分为零假设（或无效假设）和备择假设。

例如，如果研究问题是“对于某个疾病，用试验器械治疗后，治疗组疗效优于对照组吗”？针对该问题的两个假设是：

零假设 H_0：治疗组疗效比对照组疗效差。

备择假设 H_1：治疗组疗效不比对照组疗效差。

我们的目的就是要否定零假设，接受备择假设（即治疗组疗效优于对照组疗效），并将从样本得出的结论推断到总体。

在上述统计推断过程中，可能会犯两类决策错误。如果样本显示器械治疗组的疗效大于对照组疗效（即：拒绝零假设，接受备择假设），而临床实践中并没有发现两组疗效有差异时，就犯了Ⅰ类错误（也称为 α 错误，或假阳性错误）。另一方面，如果样本显示两组间疗效无差异（即：没有足够证据否定零假设），而临床实践中，器械组疗效确实优于对照组时，就犯了Ⅱ类错误（也称为 β 错误，或假阴性错误）。我们通常将 α 称作显著性水平，把 $1-\beta$ 定义为检验效能，或把握度。

一般而言，临床试验中对Ⅰ类错误 α 和Ⅱ类错误 β 的大小是有明确规定的。通常情况下，双侧显著性水平 α 不得超过 5%（或单侧显著性水平不得超过 2.5%），β 应不大于 20%（把握度不低于 80%）。其实际意义为：在将由试验样本得出的结论推断到总体时，结论出错的可能性不超过 5%，同时有 80%的把握得出与临床试验完全相同的结论，也就是说，如果治疗组与对照组间确实存在疗效差异的话，那么就有 80%的把握能够将此种差异检测出来。由此可见，把握度在临床试验中起着举足轻重的作用，它是保证我们获得阳性结果的前提条件，在条件允许的情况下，把握度越大越好。

总而言之，样本的大小通常按照受试产品具体的特性、主要评价指标及其参数来确定。计算过程中所需用到的统计量的估计值以及该估计值的来源依据应写在相应的临床试验研究方案中。一般来说，统计量值应主要参照已公开发表的国内外相关文献资料或待测产品预试验的结果来估算。

样本量通常是基于病人数的，有些临床试验在某一个病人身上重复进行多次试验，由于来自相同病人的数据原则上是高度相关的，因此，统计分析时应随机抽取同一病人的一次测量值

进行疗效评价，除非有明确的数据支持，说明每个病人各次测量值之间不相关。

在非劣效试验中，由于我们的目的是要验证两个器械实质等同，因此，计算样本量时应假设被试器械与对照器械来自同一总体，他们应该具有相同的总体率（成功率、事件发生率等）或总体均数，或者，按照非劣效定义，被试器械的总体率或总体均数稍劣于对照器械。否则，可能造成低估样本量，以至于没有足够的研究把握度获得预期的阳性结果。

如果一个申办者因为对器械在人群中的应用缺乏足够的经验而不能回答临床试验的关键问题，那么申办者应该设计一个小样本的人体试验来收集有关器械的基本信息。这个小样本试验（通常称为预试验或可行性研究）的研究目的是确定器械的可能医疗用途、监查潜在的研究变量、检测试验流程以及决定潜在的主要终点的精确度。还可对可能导致偏倚的因素进行有限的评价。

预试验或探索性试验也应有清晰和明确的研究目标。探索性试验有时需要更为灵活可变的方法进行设计并对数据进行探索性分析，以便根据逐渐积累的结果对后期的确证性试验设计提供相应的信息。虽然探索性试验对整个有效性的验证有所贡献，但不能作为证明产品安全性和有效性的正式依据。医疗器械的有效性和安全性只有通过探索性试验（如有）之后的验证性试验才能得到充分证实。验证性试验是一种事先提出假设并对其进行检验的具有良好对照的随机试验，以说明所研制的医疗器械对临床是有益的。

二、优效性试验样本含量估计

（一）定性指标

需要预先指定的参数为：

π_C：对照组总体率

π_T：试验组总体率

P_C：对照组样本率

P_T：试验组样本率

n_T：试验组样本数量

n_C：对照组样本数量

N：总的样本量，$N = n_T + n_C$

Δ：优效性界值

D：希望检测的两组率之差，$D = P_T - P_C$；如果指标为低优指标，$D = P_C - P_T$

α：第Ⅰ类错误

β：第Ⅱ类错误

检验假设为：

H_0：$\pi_T - \pi_C \leqslant \Delta$，两组器械疗效差等于或低于优效性界值

H_1：$\pi_T - \pi_C > \Delta$，两组器械疗效差大于优效性界值

则每组样本量为：

$$n_T = n_C = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 [P_C(1-P_C) + P_T(1-P_T)]}{(D-\Delta)^2}$$

其中：$Z_{1-\alpha}$，$Z_{1-\beta}$ 为标准正态分布的分位数，可在相关统计书中查到。也可用 f（α，β）值表近似计算（表 5-1）。设 $f(\alpha,\beta)=(Z_{1-\alpha/2}+Z_{1-\beta})^2$，则可利用下表查 $(Z_{1-\alpha/2}+Z_{1-\beta})^2$。如：$\alpha=0.05$，$\beta=0.20$ 查得 $f(\alpha,\beta)=7.9$

表 5-1　$f(\alpha,\beta)$ 值表

第Ⅰ类错误概率 α（双侧）	$f(\alpha,\beta)$值			
	β（第Ⅱ类错误概率）	0.05	0.1	0.15	0.2
0.10		10.8	8.6	4.3	6.2
0.05		13.0	10.5	5.8	7.9
0.02		15.8	13.0	7.6	10.0
0.01		17.8	14.9	9.1	11.7

【例 5-1】　在心脏换瓣临床试验中，对照瓣膜 3 个月血栓栓塞发生率为 10%，估计被试瓣膜 3 个月血栓栓塞发生率降至 2%。临床认为，被试瓣膜相比对照瓣膜，血栓发生率之差超过 3%才有临床意义。设计一个检验水平 α 为 0.05；检验效能 $1-\beta=0.80$。则：至少需要多少样本，可正确检测出试验组优于对照组？

解：两组事件发生率之差为：$D=0.10-0.02=0.08$，$\Delta=0.03$

用查表的方法，$\alpha=0.05$，$\beta=0.10$ 时，可查表中 $\alpha=0.05$，$\beta=0.10$ 所对应的 f（α，β）值，得 $f(0.05,0.20)=7.9$，因此：

$$n_T=n_C=\frac{[0.02(1-0.02)+0.10(1-0.10)]\times 7.9}{[0.08-0.03]^2}=346$$

则：每组至少需要 346 例，两组共需要 692 例。

（二）定量指标

需要预先指定的参数为：

μ_C：对照组总体均数

μ_T：试验组总体均数

x_C：对照组样本均数

x_T：试验组样本均数

n_T：试验组样本数量

n_C：对照组样本数量

N：总的样本量，$N=n_T+n_C$

Δ：优效性界值

D：希望检测的两组率之差，$D=\mu_T-\mu_C$；如果指标为低优指标，$D=\mu_C-\mu_T$

σ：标准差

α：第Ⅰ类错误

β：第Ⅱ类错误

检验假设为：

H_0：$\mu_T-\mu_C\leqslant\Delta$，两组器械疗效差等于或低于优效性界值

H_1：$\mu_T-\mu_C>\Delta$，两组器械疗效差大于优效性界值

则每组样本量为：

$$n=\frac{2(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma^2}{(D-\Delta)^2}$$

【例 5-2】 一种新的美容产品，其疗效可平均保持 6 个月（$\mu_T=6$），对照组的疗效可平均保持 4 个月（$\mu_C=4$），假设两组标准差σ相同为 3.5。临床认为，新美容产品相比旧美容产品，疗效保持时间之差高于 15d（$\Delta=0.5$），即可认为具有临床意义。当Ⅰ类错误 $\alpha=0.05$，Ⅱ类错误 $\beta=0.2$ 时，至少需要多少样本量？

解：由 $\alpha=0.05$ 和 $\beta=0.2$ 查表得 $f(\alpha,\beta)=f(0.05,0.20)=7.9$

代入公式得：

$$n_T=n_C=\frac{2\times3.5^2\times7.9}{[(6-4)-0.5]^2}=86$$

则：每组至少需 86 例，两组共需要 172 例。

三、等效性试验含量估计

有些器械临床试验希望被试产品既不比对照好、也不比对照差，此时可用等效性设计。例如希望一种价格较廉的新产品与现有的已上市产品相比，其性能既不比对照好、也不比对照差；这时可以采用等效性设计方法。以下是计算等效性试验样本量的公式。

（一）定性指标

首先确定下面一些参数：

π_C：对照组总体率

π_T：试验组总体率

P_C：对照组样本率

P_T：试验组样本率

n_T：试验组样本数量

n_C：对照组样本数量

N：总的样本量，$N=n_T+n_C$

Δ：等效性界值，即如果两个总体率的差别不超过 Δ 时，这种率的差别是没有临床意义的。这里，Δ 取值为正值

D：希望检测的两组率之差，$D=P_T-P_C$

α：第Ⅰ类错误

β：第Ⅱ类错误

检验假设为：

$$H_0：|\pi_T-\pi_C|\leqslant\Delta,$$

$$H_1：|\pi_T-\pi_C|>\Delta$$

原假设意为：试验器械疗效劣于对照器械，其差值$\leqslant-\Delta$；或试验器械疗效优于对照器械，其差值$\geqslant\Delta$。

备择假设意为：试验器械与对照器械疗效之差不超过 Δ。

每组所需例数为：

$$n_T = n_C = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 [P_C(1-P_C) + P_T(1-P_T)]}{(\Delta - |D|)^2}$$

其中：$Z_{1-\alpha}$，$Z_{1-\beta}$ 为标准正态分布的分位数，可由相关统计书中查到。也可用 $f(\alpha,\beta)$ 值表近似计算

【例 5-3】 某试验器械疗效为 90%，对照器械疗效为 88%。新器械 B 与 A 器械最多相差 10%可被接受。假设犯第Ⅰ类错误的概率为 0.05（$\alpha = 0.05$），为了有 80%的把握得到两器械疗效完全相同的结果（第Ⅱ类错误 $\beta = 0.2$），需要多大样本量？

解：$P_T = 0.90$，$P_C = 0.88$；$\Delta = 10\%$；$\alpha = 0.05$；$\beta = 0.2$

查 $f(\alpha,\beta)$ 值表得 $f(0.05, 0.2) = 7.9$

代入公式得：

$$n_T = n_C = \frac{7.9 \times [0.88(1-0.88) + 0.90(1-0.90)]}{(0.1-0.02)^2} = 241.4 \approx 242$$

则：每组所需例数为 242 例，两组共需要 484 例。

（二）定量指标

需要预先指定的参数为：

μ_C：对照组总体均数

μ_T：试验组总体均数

X_C：对照组样本均数

X_T：试验组样本均数

n_T：试验组样本数量

n_C：对照组样本数量

N：总的样本量，$N = n_T + n_C$

D：希望检测的两组率之差，$D = \mu_T - \mu_C$

σ：标准差

α：第Ⅰ类错误

β：第Ⅱ类错误

检验假设为：

$$H_0 : |\mu_T - \mu_C| \leqslant \Delta,$$

$$H_1 : |\mu_T - \mu_C| > \Delta$$

原假设意为：试验器械疗效劣于对照器械，其差值$\leqslant -\Delta$；或试验器械疗效优于对照器械，其差值$\geqslant \Delta$。

备择假设意为：试验器械与对照器械疗效之差不超过 Δ。

则每组样本量为：

公式如下：

$$n_T = n_C = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{(\Delta - |D|)^2}$$

【例 5-4】 某人工关节试验，对照关节可承受压力约为 97MPa，被试关节可承受压力约为 90MPa。希望被试关节与对照组相比可承受压力之差（Δ）$\leqslant$17MPa。标准变异（σ）为

21MPa。假设犯第Ⅰ类错误的概率为0.05($\alpha=0.05$),为了有80%的把握得到两器械疗效相当的结果(第Ⅱ类错误$\beta=0.2$),试估计其样本量。

解:$\sigma=$ 21MPa;$\Delta=17$MPa;$\alpha=0.05$;$\beta=0.2$

查$f(\alpha,\beta)$值表得$f(0.05,0.2)=7.9$

代入公式得:

$$n_T=n_C=\frac{2\times 7.9\times 21^2}{(17-|97-90|)^2}=69.7\approx 70$$

则:每一组例数为70例,两组共需要140例。

四、非劣效试验样本含量估计

(一)定性指标

首先确定下面一些参数:

π_C:对照组总体率

π_T:试验组总体率

P_C:对照组样本率

P_T:试验组样本率

n_T:试验组样本数量

n_C:对照组样本数量

N:总的样本量,$N=n_T+n_C$

D:希望检测的两组率之差,$D=P_T-P_C$;如果指标为低优指标,$D=P_C-P_T$

Δ:非劣效界值,即:如果两个产品的疗效差别不超过Δ时,说明试验组疗效不比对照组差(即:实质等同),也就是说,临床认为这种差别没有临床意义,可以忽略不计

非劣效试验中,界值Δ取负值

α:第Ⅰ类错误

β:第Ⅱ类错误

假设检验的公式:

$$H_0:\pi_T-\pi_C\leqslant\Delta$$

$$H_1:\pi_T-\pi_C>\Delta$$

原假设意为:试验组与对照组差值的绝对值大于非劣效界值的绝对值。

备择假设意为:尽管试验器械疗效稍劣于对照器械,但是,其差值的绝对值小于非劣效界值的绝对值。

每组所需例数为:

$$n_T=n_C=\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2[P_C(1-P_C)+P_T(1-P_T)]}{(D-\Delta)^2}$$

其中:$Z_{1-\alpha}$,$Z_{1-\beta}$为标准正态分布的分位数,可由相关统计书中查到。也可用$f(\alpha,\beta)$值表近似计算。

如果无法得到对照组和试验组的率P_C,P_T,可以近似认为两个率相等,即$D=0$,样本量计算公式中分母,即为界值的平方。

【例5-5】 对一器械临床试验,对照器械疗效为90%,试验器械疗效87%,在试验中,如果

试验器械的疗效比对照器械最多差10%可以接受，假设犯第Ⅰ类错误的概率为0.05（$\alpha=0.05$），把握度80%（第Ⅱ类错误$\beta=0.2$），需要多大样本量得到试验器械非劣于对照器械的结论？

解：$P_T=0.87$，$P_C=0.90$；$\Delta=-10\%$；$\alpha=0.05$；$\beta=0.2$

查$f(\alpha,\beta)$值表得$f(0.05,0.2)=7.9$

代入公式得：

$$N=\frac{7.9\times[0.87(1-0.87)+0.90(1-0.90)]}{[(0.87-0.90)-(-0.10)]^2}=327.3\approx328$$

则：每组所需例数为328例，两组共需要656例。

（二）定量指标

需要预先指定的参数为：

μ_C：对照组总体均数

μ_T：试验组总体均数

$\bar{x}_C$：对照组样本均数

$\bar{x}_T$：试验组样本均数

n_T：试验组样本数量

n_C：对照组样本数量

N：总的样本量，$N=n_T+n_C$

D：希望检测的两组率之差，$D=\mu_T-\mu_C$；如果指标为低优指标，$D=\mu_C-\mu_T$

Δ：非劣效界值，即：如果两个产品的疗效差别不超过Δ时，说明试验组疗效不比对照组差（即：实质等同），也就是说，临床认为这种差别没有临床意义，可以忽略不计。非劣效试验中，界值取负值

σ：标准差

α：第Ⅰ类错误

β：第Ⅱ类错误

假设检验的公式：

$$H_0:\mu_T-\mu_C\leqslant\Delta$$

$$H_1:\mu_T-\mu_C>\Delta$$

原假设意为：试验组与对照组差值的绝对值大于非劣效界值的绝对值。

备择假设意为：试验器械疗效劣于对照器械，其差值的绝对值小于非劣效界值的绝对值。

则每组样本量为：

$$n_T=n_C=\frac{2(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma^2}{(D-\Delta)^2}$$

如果无法得到对照组和试验组的率P_C，P_T，可以近似认为两个率相等，即$D=0$，样本量计算公式中分母，即为界值的平方。

【例5-6】　某药物洗脱冠状动脉支架产品，预与其已上市的一代产品进行比较，主要终点设为术后9个月时冠状动脉造影得到的管腔丢失。据文献报道：其一代产品9个月时的管腔丢失均值为0.61mm，试验组支架产品0.63mm（低优指标），标准差为0.5mm，则：当临床认可的非劣效界值为0.2mm，统计学显著性水平为0.05，把握度为80%时，需要多大样本？

解：本研究类型为非劣效设计，非劣效界值 $=-0.2$mm。应用定量指标的非劣效试验公 $\alpha=0.05$，$\beta=0.10$ 时，所对应的 $f(\alpha,\beta)$值为 7.9，因此：

$$n_T=n_C=\frac{2\times 7.9\times 0.5^2}{[(0.61-0.63)-(-0.2)]^2}=121.9\approx 122$$

以上结果说明，试验组和对照组每组至少需要 122 例受试者参加此项临床试验。由于造影随访的失访率很高，若考虑 30%的病人失访，则试验组和对照组至少应分别入选 159 例受试者，两组合计共需入选 318 例患者。

五、单组目标值法试验样本含量估计

(一)定性指标

首先确定下面一些参数：

P_T：预期的产品性能指标(如：有效率、成功率、事件发生率等)

P_0：目标值

α：第Ⅰ类错误

β：第Ⅱ类错误

则每组所需例数为：

$$n=\frac{\left[Z_{1-\alpha/2}\sqrt{P_0(1-P_0)}+Z_{1-\beta}\sqrt{P_T(1-P_T)}\right]^2}{(P_T-P_0)^2}$$

其中：$Z_{1-\alpha/2}$，$Z_{1-\beta}$ 为标准正态分布的分位数

【例 5-7】 某射频消融导管临床验证研究，按照美国 FDA 射频消融导管临床试验指导原则(见 Cardiac Ablation Catheters Generic Arrhythmia Indications for Use; Guidance for Industry)，其中明确规定了该类产品的主要疗效评价指标为：即刻手术成功率和术后 3～6 个月随访时的成功率，且即刻手术成功率必须达到 95%，其 95%可信区间下限必须＞85%；术后 3～6 个月随访时的成功率必须达到 90%，其 95%可信区间下限必须＞80%；当显著性水平取 0.05，检验效能为 80%时，应用定性指标的单组试验公式：

(1)以即刻手术成功率计算：

$$n=\frac{\left[Z_{1-\alpha/2}\sqrt{p_0(1-p_0)}+Z_{1-\beta}\sqrt{p_T(1-p_T)}\right]^2}{(p_T-p_0)^2}$$

$$=\frac{\left[1.96\sqrt{0.85(1-0.85)}+1.28\sqrt{0.95(1-0.95)}\right]^2}{(0.95-0.85)^2}=79$$

(2)以术后 3 个月随访成功率计算：

$$n=\frac{\left[Z_{1-\alpha/2}\sqrt{p_0(1-p_0)}+\mu_{1-\beta}\sqrt{p_T(1-p_T)}\right]^2}{(p_T-p_0)^2}$$

$$=\frac{\left[1.96\sqrt{0.80(1-0.80)}+1.28\sqrt{0.90(1-0.90)}\right]^2}{(0.90-0.80)^2}=108$$

因此，综合以上结果，对于射频消融导管的临床验证研究，至少应入选 108 例受试者。若考虑 10%的病人脱落率，则至少应入选 130 例患者。

(二)定量指标

需要预先指定的参数为：

μ_T：试验组观测指标均数

μ_0：目标值

σ：标准差

α：第Ⅰ类错误

β：第Ⅱ类错误

则每组样本量为：

$$n=\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma^2}{(\mu_T-\mu_0)^2}$$

六、诊断试验样本含量估计

样本量的大小应该根据产品特性，在符合统计学原则的基础上，同时符合 2007 年 4 月 SFDA 规定的一般要求和特殊要求。

统计学要求：

（一）根据抽样调查的样本量公式计算

公式：

$$n=Z^2_{1-\alpha/2}P(1-P)/\Delta^2$$

Δ 为允许误差，一般取 P 的 95%CI 宽度的一半。可 0.05～0.10，P 为灵敏度或特异度（灵敏度计算病例组样本量，特异度计算对照组样本量）。对于诊断试验定量指标样本含量的计算，可参考非劣效试验的样本含量计算，此处从略。

对于抽样调查法，只保证参数估计的精度（可信区间的宽度一定），而未考虑是否能够达到最终的标准。或者说对于抽样调查设计，实际并没有提出明确的统计假设检验，所以根本不存在把握度的概念。

适用条件：主要适用于参数估计，在满足研究者期望精度水平的前提下，证明诊断试验的灵敏度和特异度。

（二）根据单组目标值法样本量公式计算

单组目标值公式 $n=\dfrac{\mu_\alpha\times\sqrt{P_0\times(1-P_0)}+\mu_\beta\times\sqrt{P\times(1-P)}}{P-P_0}$

其中，P_0：为临床能够接受灵敏度（或特异度）的最低标准；P：为预期灵敏度（或特异度）；μ_α 和 μ_β：显著性水平和把握度对应的正态分布函数的分位数

可以给出明确的假设检验，保证把握度。

适用条件：主要适用于已统计检验为目的的情况，研究者想检验某种诊断方法的前提下证明诊断试验的灵敏度和特异度是否不低于某个临床认可的标准。相对于抽样调查方法更为保守。

【例 5-8】 B 型超声波对胆石症诊断的预期灵敏度为 80%，特异度为 60%，在允许误差为 8%，显著性水平 α 为 0.05 时，要多大样本才能具有统计学意义？

查表：$Z_{0.05}=1.96$，$\Delta=0.08$，因此：

$$n_{有病组}=1.96^2\times0.8(1-0.8)/0.08^2=96.1=97$$

$$n_{无病组}=1.96^2\times0.6(1-0.6)/0.08^2=144.1=145$$

共需要至少 242 例受试对象。

1. SFDA 对诊断试剂的一般要求　三类：总样本数至少为 1000 例；二类：总样本数至少为

200 例；一类：不需进行临床研究 。

2. SFDA 对诊断试剂的特殊要求

(1)国家法定用于血源筛查的项目，及预期用途为血源筛查的诊断试剂：总样本数至少为 10 000 例。

(2)采用体外核酸扩增(PCR)方法、用于病原体检测的诊断试剂：总样本数至少为 500 例。

(3)与麻醉药品、精神药品、医疗用毒性药品检测相关的诊断试剂：总样本数至少为 500 例。

(4)采用放射性核素标记的体外诊断试剂：临床研究总样本数至少为 500 例。

(5)新诊断试剂产品(未在国内批准注册的产品、被测物相同但分析敏感度指标不在国家已批准注册产品范围内，且具有新的临床诊断意义的产品)，其临床研究样本量要求同第三类产品。

3. 需要注意的是

(1)估计的灵敏度和特异度应当基于目标人群中的样本。

(2)如果换一组人群，或同一组人群不同时间，估计的灵敏度和特异度可能完全不同。

(4)应根据产品临床使用目的，及该产品相关疾病的患病率确定临床研究的样本量。

(5)在符合指导原则有关最低样本量要求的前提下，还应符合统计学要求。

(6)罕见病、特殊病种及特殊情况可酌减样本量，但应说明理由，并满足评价的需要。

七、配对或交叉设计资料假设检验时检验效能的计算

预先指定的统计参数：

μ_C ：对照组观测指标均数

μ_T ：试验组观测指标均数

σ：标准差

α ：第Ⅰ类错误

β ：第Ⅱ类错误

$$n=\frac{\sigma^2\ (u_{\alpha/2}+u_{\beta})^2}{(\mu_T-\mu_C)^2}$$

八、多组平行组设计定量资料统计分析时样本含量估计

按下式计算：

$$n=\psi^2\left(\sum_{i=1}^{k}S_i^{\ 2}/k\right)/\left[\sum_{i=1}^{k}(\overline{X}_i-\overline{X})^2/(k-1)\right]$$

式中 n 为各样本所需样本含量，$\overline{X}_i$ 和 S_i 分别为第 i 个样本的均数和标准差的初估值。$\overline{X}=\sum_{i=1}^{k}\overline{X}_i/k$ ，k 为组数。ψ 值查本书附录统计用表中 ψ 值表获得，先以 α ，β ，$\upsilon_1=k-1$ ，$\upsilon_2=\infty$ 查得 ψ 值，代入公式中，求得 $n_{(1)}$ ，第二次由 $\upsilon_1=k-1$ ，$\upsilon_2=k(n_{(1)}-1)$ 查 ψ 值表，代入式中求 $n_{(2)}$ ，仿此进行，直至前后两次求得结果趋于稳定为止，即为所求样本含量

九、多组平行组设计定性资料(多个样本率)比较时样本含量估计

按下式进行计算：

$$n=\frac{\lambda}{2\left(\arcsin\sqrt{p_{\max}}-\arcsin\sqrt{p_{\min}}\right)^{2}}$$

式中 n 为每个样本所需观察例数，$p_{\max}$ 和 $p_{\min}$ 为所有总体率估计值中的最大率和最小率，当仅知最大率和最小率差值 p_d 时，则取 $p_{\max}=0.5+p_d/2$，$p_{\min}=0.5-p_d/2$，λ 是以 α，β，自由度 $\upsilon=k-1$，查本书附录统计用表中 λ 值表而得，k 为组数

（李　卫　郭　晋　王　杨　吴振强）

第6章 非诊断试验平行组设计定性资料的统计分析

医疗器械临床试验中经常涉及大量定性资料，在平行组设计中，主要终点变量可能为二值变量，多值有序变量等定性变量，对于这类资料的统计分析包括统计描述、参数估计和假设检验，本章将对平行组设计定性资料的统计描述、率的参数估计及有关的假设检验方法进行介绍。

一、描述性统计

（一）相对数的种类与计算

描述定性资料的内部构成情况或相对比值或某现象的发生强度，有相对数，它包括比和率。相对数看起来很简单，但在使用中稍不留神，很容易出错。

相对数可分为结构相对数和强度相对数。相对数也叫相对指标，它是两个有联系的指标之比，按性质和用途可分为比（常分为构成比、相对比、动态数列的定基比和环比）和率。“比与率”有时较难分清，因此，人们在使用中经常混淆。

它们的共同点在于：求率与比时所用公式的基本形式是完全相同的，都是由两个绝对数之商乘以100%（或1000‰等）；它们的不同点在于：率反映某种事物或现象发生的强度，而比则反映“部分与整体”或“某一部分与另一部分”之间的关系。

1. *率* 率是强度相对数，表示在一定范围和时间内，某现象的发生次数 k 与该现象可能发生的总次数 n 之比，说明该现象发生的强度。按式(6-1)计算。

$$率=\frac{k}{n}\times 100\%（或\ 1000‰）\tag{6-1}$$

2. *构成比* 它表示仅具有属性 i 的那一部分个体数目 n_i 占全部个体总数 n 的比重。按式(6-2)计算。

$$构成比=\frac{n_i}{n}\times 100\%（或\ 1000‰等）\tag{6-2}$$

3. *相对比* 它是两个有关指标之比，说明两者的对比水平。按式(6-3)计算。

$$相对比=\frac{甲指标}{乙指标}\times 100\%\tag{6-3}$$

动态数列的定基比与环比：动态数列是一系列按时间顺序排列起来的统计指标（包括绝对数、相对数和平均数），用以说明事物在时间上的变化和发展趋势。常用动态数列数据计算的统计指标有定基比与环比，根据不同的算法，这两种比可用来反映随时间的推移，事物的发展速度（%）或增长速度（%）。

用来反映发展速度时，定基比与环比分别为：

定基比：各时间点上的统计指标都以第1个时间点上的统计指标为分母求得。

环比：各时间点上的统计指标都以它前面的那个时间点上的统计指标为分母求得。

用来反映增长速度时，定基比与环比分别为：

定基比：各时间点上定基比发展速度减 1。

环比：各时间点上环比发展速度减 1。

（二）相对数的应用

在医疗器械临床试验中，相对数应用是比较广泛的，使用相对数应注意以下几点。

1. 分母不宜过小，一般＜20 例，是不易计算“率”的，应用绝对数表达比较好。

2. 当相对数的分母不可比时，仅凭两个彼此独立的相对数来说话，也是没有说服力的。

3. 构成比只能说明事物所占的比重或分布，不能说明事物发生的强度或频率，不能用构成比替代率。

二、率的标准误与区间估计

（一）率的标准误

样本率与总体率之间存在着抽样误差。率的抽样误差指样本率与样本率之间，样本率与总体率之间的差别。

率的抽样误差可用率的标准误来表示，计算公式为：

$$S_p=\sqrt{\frac{p(1-p)}{n}} \tag{6-4}$$

式中，S_p 为率的标准误，p 为样本率（表示实际事物发生的率），$1-p$ 为某事物实际不发生的率，也可以用 q 表示，n 为样本观察例数

率的标准误的大小可表示抽样误差的大小，说明样本率对总体率的代表性。率的误差小，说明抽样误差小，表示样本率与总体率比较接近，用样本率代表总体率的可靠性就大；反之，率的标准误大，说明抽样误差大，表示样本率与总体率相差较远，用样本率代替总体率的可靠性就小。

（二）总体率的区间估计

率的抽样研究的目的，就是通过样本来估计总体的水平。由于样本率正好等于总体率的情况是极少的，所以用样本率对总体率的估计往往是一个区间，这个区间是指可能包括总体率的一个区间，称为总体率的置信区间。

常用的总体率置信区间的估计方法有两种：

1. 正态近似法

当 n 足够大，且 p 和 $1-p$ 均不接近于零时，例如 $n>1000$，且 p 或 $1-p\geqslant 1\%$ 时，或 np 与 $n(1-p)$ 均>5 时，p 的抽样分布逼近正态分布，可按式(6-5)求总体率的 $100(1-\alpha)\%$ 置信限，并由两个限值构成置信区间。

$$p\pm u_\alpha S_p \tag{6-5}$$

式中，u_α 为正态离差值，对应于 95% 和 99% 的 u_α 为 1.96 和 2.58

2. 查表法　当样本例数 n 很小时，特别是 p 接近于 0 或 1 时，用正态近似法的准确性较差，可查百分率的置信区间表，得到总体率的置信区间。

三、定性资料假设检验

假设检验，又称显著性检验，通常是先对总体的参数或分布做出某种假设，然后用适当

的方法根据样本对总体提供的信息，推断此假设应当被拒绝或接受。其结果将有助于研究者作出决策，采取措施。例如，两种医疗器械的治疗有效率之间存在一定的差别，能否断定其中一种器械的有效率就一定高于另一种器械呢？严格地说，仅凭两组试验数据是无法做出令人信服的推断结论的。因为两种医疗器械表现出来的差别，可能有两种原因所致：一是单纯由抽样误差所致（即两种器械的有效率实际是相同的）；二是除抽样误差外，两种器械的有效率确实有所不同。判断其不同有两种可能的方法，其一，将试验一批一批地重复做，每一批的样本都比较大，而且重复试验的批次又比较多，如果95%的批次都显示甲器械的治愈率优于乙器械，根据常理，我们可以接受“平均来说，甲器械的治愈率高于乙器械的治愈率”的结论；其二，运用随机变量的概率分布规律，将一批试验的结果放入某个特定的分布规律之中去，相当于把相同的试验重复了无穷多批，从而借助随机变量的概率分布规律，也能得出具有一定置信度（如95%或99%）的统计推断结论来，实现这一推理过程的统计学方法，即假设检验。很显然，第一种方法费时费力，工作量巨大难以完成；而第二种方法具有较强的可操作性，易于为人们所接受。

根据上述假设检验的基本概念和思想的介绍，可总结出定性资料假设检验的一般步骤：

1. *建立假设，确定检验水准（显著性水平）* 假设检验中有原假设（无效假设）H_0和备择假设H_1两种假设。原假设：根据检验结果准备予以拒绝或接受的假设；备择假设：与原假设不相容（对立）的假设。在建立原假设H_0和备择假设H_1的同时，要设定检验水准（显著性水平）α，即拒绝原假设可能犯错误的概率，它也是假设检验最后一步计算出特定随机事件发生概率（P值）之后，作出肯定或否定原假设的判定标准。$P \leqslant \alpha$时，拒绝H_0，接受H_1。

对于定性资料，例如：对两种器械的治疗有效率相等（未知）的假设检验：

H_0：$\pi_1 = \pi_2$（两种器械的治疗有效率相等）

H_1：$\pi_1 \neq \pi_2$（两种器械的治疗有效率不相等）

双侧$\alpha = 0.05$（根据具体情况，有时也取$\alpha = 0.01$）。

2. *选择适当的假设检验方法，计算相应的检验统计量* 应根据资料类型、临床试验设计类型、统计分析目的和各种假设检验方法的应用条件选择恰当的检验方法。对于定性资料，根据设计类型和结果变量的性质可选择χ^2检验、秩和检验等。在选定方法后，在H_0成立的条件下可计算相应的检验统计量，即将从样本资料获得的信息代入相应的计算公式。

3. *确定P值，作出结论* 根据计算出的检验统计量的值，查相应的界值表即可获得P值，也可通过统计软件直接计算出检验统计量在特定区域内取值的概率。

将P值与事先规定的检验水准α进行比较，即可作出结论，当$P \leqslant \alpha$时，拒绝H_0，接受H_1，认为对比的两个统计量所代表的两个总体相应量（如两个总体率）之间的差别有统计学意义；反之，$P > \alpha$时，不拒绝H_0，认为对比的两个统计量所代表的两个总体相应量（如两个总体率）之间的差别无统计学意义。

一般来说，推断的结论应包含统计结论和专业结论两部分。上述表述仅为统计结论，专业结论是根据统计结论结合实际问题中是否不同以及差异的方向作出推断。如从样本资料提供的信息得知，A器械的有效率大于B器械的有效率，采用正确的假设检验后得到$P \leqslant \alpha$时，可认为：在总体上，A器械的有效率高于B器械。

四、平行组设计 2×2 列联表的假设检验(两组率的比较)

(一)基本概念

两组平行组(试验器械，对照器械)的研究设计，指标为率(有效率、治愈率、缓解率、不良事件率)，可以表达为列联表的形式，见表 6-1。

表 6-1　同时观测 n 个受试者两个属性的分组结果

试验分组 A	例数		
指标 B	B_1(发生)	B_2(没发生)	合计
A_1	a	b	$n_1=a+b$
A_2	c	d	$n_2=c+d$
合计	$m_1=a+c$	$m_2=b+d$	$n=a+b+c+d$

这种研究设计的目的是通过两组样本的频数分布情况来推测两组样本对应的总体频数分布是否相同，表中各格数字 a,b,c,d 为一次抽样得到的实际频数，也称为观测频数。通过观察频数来推断两组总体频数分布情况时肯定存在抽样误差，所以当两样本频率不相同时可能有两种原因：

1. 抽样误差所致。

2. 这两个样本频率对应的总体概率(以下简称总体率)本来就不同。

设第一行上指标 B_1 发生的概率为 π_1，设第二行上指标 B_1 发生的概率为 π_2，则平行组设计 2×2 列联表资料的假设检验的零假设和备择假设分别为：

H_0：$\pi_1=\pi_2$(两种器械的治疗有效率相等)

H_1：$\pi_1\neq\pi_2$(两种器械的治疗有效率不相等)

若上述的第一种原因成立，则 a 格中理论频数为 $(m_1/n)\times n_1$，即得到 a 格的理论频数，同理可计算出各个格子上的理论频数，如表 6-2。

表 6-2　各个格子的理论频数

试验分组 A	例数		
指标 B_i	B_1	B_2
A_1	$n_1\times m_1/n$	$n_1\times m_2/n$
A_2	$n_2\times m_1/n$	$n_2\times m_2/n$

如果零假设为真，则由式(6-6)定义的检验统计量不会超过某临界值；如果计算出的检验统计量越过该临界值，则拒绝接受零假设。此处的假设检验方法称为 Pearson χ^2 检验，其计算公式如下：

$$\chi^2=\sum\frac{(O-T)^2}{T} \tag{6-6}$$

上式中 O 代表实际频数，T 代表理论频数，该式也可用于 R×C 列联表资料的独立性检验。

在 2×2 表资料中，据表(6-6)和式(6-7)可知：

$T_a=\frac{(a+b)(a+c)}{n}$，$T_b=\frac{(a+b)(b+d)}{n}$，$T_c=\frac{(c+d)(a+c)}{n}$，$T_d=\frac{(c+d)(b+b)}{n}$

$T_a\cdots T_d$ 分别代表与观察频数 a,b,c,d 对应的理论频数。

所以，此处式(6-6)可转化为计算 2×2 列联表资料检验统计量的专用公式：

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{6-7}$$

由式(6-6)和式(6-7)计算出的 χ^2 值近似服从自由度为 υ 的 χ^2 分布，自由度 υ 按下式计算：

$$\upsilon=(R-1)(C-1) \tag{6-8}$$

上式中 R 为行数，C 为列数。四格表资料中 R 和 C 均为 2，所以 $\upsilon=1$，此时，与 $\alpha=0.05$ 和 $\alpha=0.01$ 对应的两个临界值分别为 $\chi^2_{(1)0.05}=3.841$ 和 $\chi^2_{(1)0.01}=6.635$。

按式(6-6)或式(6-7)计算出检验统计量的值，根据 χ^2 分布规律、自由度及检验水准可以得到相应的临界值，如果现有检验统计量的值等于临界值，则 $P=\alpha$；如果现有统计量大于检验临界值，则 $P<\alpha$；如果现有统计量小于检验临界值，则 $P>\alpha$。得到 P 值即可作出相应的统计推断。

需要注意的是，运用式(6-6)或式(6-7)计算检验统计量 χ^2 值的前提条件是：$n\geqslant 40$，无<5的理论频数；当表格中 $n\geqslant 40$，但至少有一个理论频数为 $1<T\leqslant 5$ 时，需改用下面的连续校正 χ^2 检验公式计算检验统计量 χ^2 值：

$$\chi^2_C=\frac{n(|ad-bc|-0.5n)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{6-9}$$

当表格中 $n<40$ 或至少有一个理论频数为 $T\leqslant 1$ 时，需要用 Fisher R A 借助超几何分布提出的直接计算概率的方法，此方法被称为 Fisher 精确检验法。此方法不属于 χ^2 检验的范畴，但可以作为四格表 χ^2 检验应用的补充。四格表资料 Fisher 精确检验法计算公式如下：

$$p_a=\frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ b!\ c!\ d!\ n!}，P=\sum_a p_a \tag{6-10}$$

上式中 p_a 为获得某个四格表的概率，对于 Fisher 精确检验法，P 值就是“出现目前状况和更极端状况的概率”，其计算方法就是将小于或等于“样本观察值概率 p_a”的所有可能结局的概率求和。之所以用 p_a 表示，因为它是观察到的 2×2 列联表中与“a”取值直接有关系的概率。Fisher 精确检验法可作为四格表资料检验的通用方法。

(二)假设检验的步骤

1. *建立检验假设，确定检验水准* 对表 6-1 数据进行统计分析，首先应建立检验假设。

H_0：两组(A_1 与 A_2)在终点事件 B 的两个水平(发生、不发生)上总体分布相同，都服从某一理论分布，即 $\pi_1=\pi_2=\pi$

H_1：两组(A_1 与 A_2)在终点事件 B 的两个水平上总体分布不同，即 $\pi_1\neq\pi_2$

确定检验水准为 α。

2. *选择检验方法和计算检验统计量* 根据资料情况，可以选择 Pearson χ^2 检验或连续校正的 χ^2 检验，检验统计量为 χ^2，或者用 Fisher 精确检验法直接计算概率。

3. *根据检验统计量的值确定 P 值，并作出统计推断和专业结论* 根据式(6-6)或式(6-7)或式(6-10)可计算出检验统计量 χ^2 值，然后与相应的检验临界值比较即可得到 P 值，或采用 Fisher 精确检验法直接得到 P 值，根据 P 值作出统计推断和专业结论。

【例 6-1】 某临床试验欲比较人工胃内水球与常规胃内水球对肥胖的疗效，将病情相近的 200 名患者随机均分成两组，分别用两种水球进行治疗。结果见表 6-3，试分析两种水球的治疗效果之间的差别是否具有统计学意义。

表 6-3　两种胃内水球治疗肥胖的治疗效果比较

分组	例数		
治疗效果	有效	无效	合计
人工胃内水球	89	11	100
常规水球	75	25	100
合计	164	36	200

本数据不代表真实疗效，且数据需要经过样本含量计算，此处仅为说明统计运算使用

分析与计算：表 6-3 是关于两种药物治疗效果的评价，可视为横断面研究设计 2×2 表定性资料。因总频数 $n=200>40$，且无<5 的理论频数，故可用式(6-7)，即一般 χ^2 检验公式计算。

(1)建立检验假设，确定检验水准：表 6-3 是 2×2 列联表，首先建立检验假设。

H_0：两药物的总体治疗愈率相同，即 $\pi_1=\pi_2$

H_1：两药物的总体治疗愈率不同，即 $\pi_1\neq\pi_2$

检验水准 $\alpha=0.05$。

(2)选择检验方法和计算检验统计量：采用一般 χ^2 检验方法，检验统计量为 χ^2，可手工按式(6-6)或式(6-7)计算，此处用 SAS 统计软件实现统计分析。SAS 程序(程序名 CT6-1)如下：

```
Data CT6-1;
Do a=1 TO 2;Do b=1 TO 2;
Input f @@;
Output; End; End;
Cards;
89  11
75  25;
Run;
Ods html;
Proc freq data=CT6-1;
Weight f;
Tables a * b / chisq;
Run;
Ods html close;
```

SAS 程序说明：数据步，建立数据集名为 CT6-1，建立数值型变量 a,b,f，分别读入行号、列号、每格实际频数；过程步，调用 freq 过程，指定频数变量为 f，用 tables a * b 语句表示二维列联表资料，加参数 chisq 进行一般 χ^2 检验

SAS 主要输出结果如下：

第一部分：卡方检验。

a * b 表的统计量

统计量	自由度	值	概率
卡方	1	6.639 6	0.010 0
似然比卡方	1	6.787 3	0.009 2
连续校正卡方	1	5.724 9	0.016 7
Mantel-Haenszel 卡方	1	6.606 4	0.010 2
Phi 系数		0.182 2	
列联系数		0.179 3	
Cramer V 统计量		0.182 2	

第二部分：Fisher 精确检验。

本例资料符合一般 χ^2 检验条件，第一部分 Chi-Square 结果，检验统计量 χ^2 值为 6.639 6，$P=0.010\ 0$。

Fisher 精确检验

单元格 (1,1)	频数 (F)89
左侧 $Pr <= F$	0.997 4
右侧 $Pr >= F$	0.008 0
表概率 (P)	0.005 3
双侧 $Pr <= P$	0.015 9

(3)根据检验统计量确定 P 值，并作出统计推断和专业推断：在 SAS 分析结果中给出了 P 值，所以可以直接作出统计推断，$P=0.01<0.05$，拒绝 H_0，可以认为两种胃内水球对于治疗肥胖效果是不同的，新的人工胃内水球的有效率比常规水球高。

对于多个平行组研究设计率的比较同 2×2 列联表资料的方法是基本相同的，此处从略。

五、R×C 列联表的假设检验

所谓结果变量为有序变量的单向有序 R×C 列联表是指表中仅结果变量的取值为有序的，而原因变量是无序的，如某资料中原因是“治疗与对照”，结果是“治愈、显效、好转、无效”或某指标的取值为“－，＋，＋＋，＋＋＋”等。结果变量为有序变量的单向有序的 R×C 表资料的统计分析方法可选用秩和检验、Ridit 分析以及有序变量的 logistic 回归分析。

在器械临床试验中，终点的水平可能不仅仅为“有效、无效”这样的二值变量，而是有序变量，平行组也可能设计为一个对照组多个试验组，或多个对照组一个试验组，因此资料类型即为结果为有序的 $R\times C$ 列联表资料。

【例 6-2】 某临床试验比较两种新器械与一种常规器械(A，B，C)对于关节痛的治疗效果。将 162 例关节痛患者分为 3 组，A 组 56 使用常规器械；B 组 43 例使用试验器械 B；C 组 63 例使用试验器械 C。治疗效果见表 6-4。对 3 组优劣进行比较。

表 6-4 3 组器械治疗效果比较

分组	例数				
治疗效果：	治愈	显著改善	改善	无效	合计
A 组	15	19	19	3	56
B 组	7	10	18	8	43
C 组	11	21	24	7	63
合计	33	50	61	18	162

(一)建立检验假设

H_0：3 组之间疗效相同

H_1：3 组之间疗效不同或不全相同

$\alpha=0.05$

(二)分析

上表中的原因变量为分组(名义变量)，结果变量为治疗效果，其取值为“治愈、显著改善、改善、无效”，为有序变量。因此应按结果变量为有序变量的单向有序的 R×C 表资料进行分析，可选用秩和检验、Ridit 分析。因一般 χ^2 检验与变量的有序性没有联系，用一般的 χ^2 检验进行分析，得到的结论是 3 组的频数分布是否相同，而不能得出 3 组疗效之间的差别是否有统

计学意义的结论。

此处使用秩和检验进行分析，此法的前提是假定各总体是连续的和相同的，利用多个样本的秩和来推断它们所代表的总体之分布位置是否差异。

(三)SAS 程序与结果解释：

SAS 程序(CT6-2)如下：

```
DATA CT6-2;
DO a=1 TO 3;
DO b=1 TO 4;
INPUTf@@; OUTPUT;
END; END; CARDS;
15 19 19 3
7 10 18 8
11 21 24 7;
Run;
ods html;
PROC NPAR1WAY WILCOXON data=CT6-2;
Freq f;
CLASS a;VAR b;
RUN;
ods html close;
```

所得的计算结果及其解释如下：

SAS 系统

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable b

Classified by Variable a

a	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	56	3 993.0	4 564.00	270.492 043	71.303 571
2	43	3 984.0	3 504.50	251.139 634	92.651 163
3	63	5 226.0	5 134.50	277.265 236	82.952 381

Average scores were used for ties.

以上是 3 个组的打分结果，其中各组的平均秩和为 A 组为 71.303 571，B 组为 92.651 163，C 组为 82.952 381。

以上是对 3 个组的治疗效果进行 Kruskal-Wallis 检验，因 $H_c \approx \chi^2 = 5.660\ 1$，$P > 0.05$，所以 3 个组治疗效果之间的差别无统计学意义。

Kruskal-Wallis Test	
Chi-Square	5.660 1
DF	2
Pr > Chi-Square	0.059 0

专业结论：因治疗效果由“治愈”“显著改善”“改善”“无效”的顺序排列，且程序中“DO”语句对治疗效果设定的水平由“1”到“3”，说明平均秩和越小治疗效果越好。但是由于 $\chi^2 = 5.660\ 1$，$P > 0.05$，治疗效果间差别无统计学意义，因此可以下结论：3 个组间的治疗效果基本相同。

(郭　晋)

第7章 非诊断试验平行组设计定量资料统计分析

平行组设计是医疗器械临床试验中最常用的一种试验设计类型，本章将对平行组设计定性资料的统计描述、及有关的假设检验方法进行介绍。主要终点变量可能为二值变量，多值有序变量等定性变量，对于这类资料的统计分析包括统计描述、参数估计和假设检验。

一、描述性统计

(一)集中趋势的描述

为了用简捷的方式表达一组性质相同的许多定量资料的平均水平，最常用的方法就是求其平均值。它说明一组观察值的集中趋势、中心位置或平均水平。通常我们计算的总和除以例数得到的是算术平均值，除此之外，还有几何平均值、调和平均值、中位数和众数。由于一组性质相同的定量数据的表现不同(在统计学上称为“分布”)，因此我们需要选用不同的平均指标来表达它们。

1. 算术平均值的定义与计算　算术平均值：n 个性质相同的定量数据之和除以 n 所得的结果叫做算术平均值。算术平均值见式(7-1)。

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{7-1}$$

式(7-1)中的 x_i 为第 i 个观测值，n 为样本大小(或样本含量，即全部数据的个数)，Σ 为求和符号，求和范围为 $i=1$ 到 $i=n$。利用频数分布表计算算术平均值见式(7-2)

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k}x_j f_j \tag{7-2}$$

式(7-2)中，k 为频数分布表的组数，n 为样本大小，x_j 为第 j 组数据的组中值，f_j 为第 j 组的频数

算术平均值适用于一组性质相同的、单峰的、且近似服从对称分布的。如图 7-1。

2. 几何平均值的定义与计算　几何平均值：对 n 个性质相同的定量数据分别取对数变换后，按算术平均值计算，然后再求其反对数所得的结果，叫做几何平均值。

利用全部原始数据计算几何平均值见式(7-3)。

$$G=\lg^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\lg x_i\right) \tag{7-3}$$

利用频数分布表计算几何平均值见式(7-4)。

$$G=\lg^{-1}\left(\frac{1}{n}\sum_{j=1}^{k}f_j\lg x_j\right) \tag{7-4}$$

式(7-3)中，“lg”代表取常用对数，“$\lg^{-1}$”代表取常用对数的反对数，也可用自然对数“ln”代替；

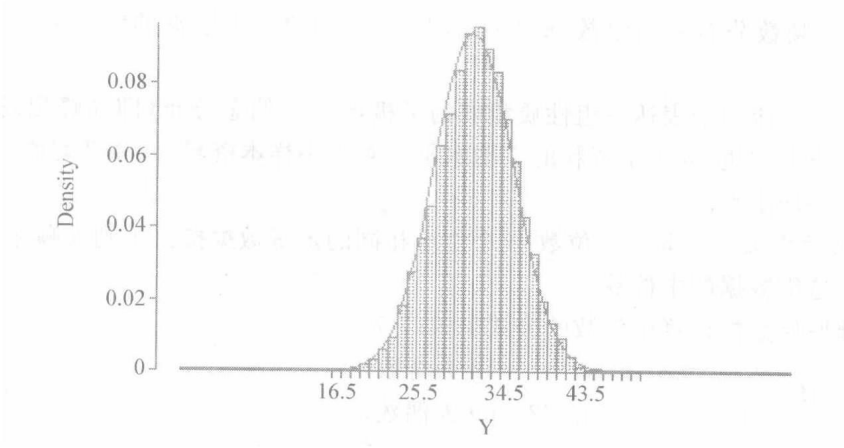

图 7-1　适合选用算术平均值的大样本定量资料的频数分布

式(7-4)中，k 为频数分布表的组数，n 为样本大小，x_j 为第 j 组数据的组中值，f_j 为第 j 组的频数

几何平均值适用于一组性质相同的、单峰的、且服从正偏态分布的(最好是服从对数正态分布的，即数据取对数变换后服从正态分布)定量资料，如图 7-2 所示。

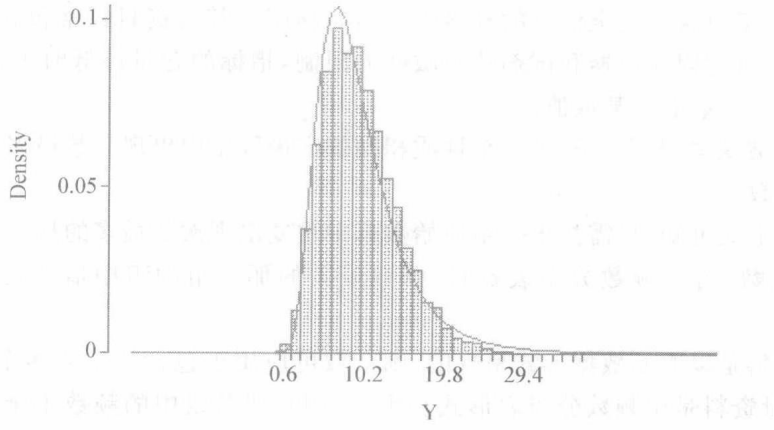

图 7-2　适合选用几何平均值的大样本定量资料的频数分布

3. 调和平均值的定义与计算　调和平均值：对 n 个性质相同的定量数据分别取倒数变换后，按算术平均值计算，然后再求其倒数所得的结果，叫做调和平均值。

利用全部原始数据计算调和平均值见式(7-5)。

$$H=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \tag{7-5}$$

式(7-5)中 $1/x_i$ 就是求 x_i 的倒数。利用频数分布表计算调和平均值见式(7-6)

$$H=\frac{n}{\sum_{j=1}^{k}\frac{f_j}{x_j}} \tag{7-6}$$

式(7-6)中，k 为频数分布表的组数，n 为样本大小，x_j 为第 j 组数据的组中值，f_j 为第 j 组的频数

调和平均值可应用于表达一组性质相同的呈极严重正偏态分布(即高峰出现在全部数据取值范围的中心点左边)的定量资料的平均水平。对于小样本资料，调和平均值常用于求类似"速度"数据的平均水平。

4. 中位数的定义与计算　中位数：n 个性质相同的定量数据按由小到大顺序排列后，居中的数据就叫做这组数据的中位数。

利用全部原始数据计算中位数的公式见式(7-7)。

$$M=\begin{cases} x_{(n+1)/2} & (n\text{ 为奇数}) \\ (x_{n/2}+x_{(n/2+1)})/2 & (n\text{ 为偶数}) \end{cases} \tag{7-7}$$

式(7-7)中的 $n/2$，$n/2+1$、$(n+1)/2$ 为数据由小到大排列后的位次

利用频数分布表计算中位数的公式见式(7-8)。

$$M=L_m+\frac{i_m}{f_m}\left(\frac{n}{2}-C\right) \tag{7-8}$$

式(7-8)中的 L_m，i_m，f_m 分别代表中位数所在组的下限、组距、频数，n 为总频数，C 为中位数所在组之前的所有组(不含中位数所在组)内的频数之和

中位数可以应用于任何定量资料，通常用于不适合用几何平均值和调和平均值的偏态资料中，尤其适用于包含不完全信息的资料中。例如临床上随访资料经常包含一些中途失访患者的某些数据；有时因受仪器和试剂的灵敏度的限制，指标的含量过低时无法准确测得，只知道一组数中有几个数低于某数值。

5. 众数的定义与计算　众数：n 个性质相同的定量数据中出现次数最多的那个数，就叫做这组数据的众数。

由众数的定义可知，只需找出一组原始数据中重复出现次数最多的那个数据，它就是这组定量资料的众数；若是频数分布表资料，频数最大的那一组的组中值就是这组定量资料的众数。

若定量资料是以原始数据形式呈现的，则众数可应用于包含两个或多个相同数据的定量资料中；若定量资料是用频数分布表形式呈现的，则只要各组中的频数不全是 1，就可以应用众数。在医学上，众数常用来表示某病或某次食物中毒的潜伏期，在实际的医疗器械临床研究中，该指标运用的较少。

(二)离散趋势的描述

仅用平均指标来描述一组定量资料的全貌是不完善的，因为平均指标仅能反映一组定量资料的平均水平，而无法反映其离散程度(即各观测值偏离平均值的程度)的大小。例如，下面的两组数的算术平均值是完全相等的，但它们的离散程度却相差很大。

第一组：23，25，27，29，31　$\bar{x}=27$，$S=3.162$

第二组：1，8，27，43，56　$\bar{x}=27$，$S=23.098$

上面这两组数据的算术平均值虽然都是 27，但第一组数据彼此之间离开平均值 27 的距离都比较小，而第二组数据彼此之间离开平均值 27 的距离都比较大。度量一组定量资料中的每一个离开其算术平均值的离散程度大小的最常用的变异指标叫做标准差(就是上面用 S 表示的数)。除了标准差外，还有方差、标准误、变异系数、四分位数间距、极差等。

1. 极差　一组性质相同的定量数据中最大值与最小值之差，叫做极差。公式见式(7-9)。

$$R=\mathrm{Max}(x_i)-\mathrm{Min}(x_i),i=1,2\cdots n \tag{7-9}$$

极差计算反应的数据离散趋势的信息量较少，因而较少使用。

2. 方差　一组性质相同的定量数据中的每一个与其样本算术平均值的差量的平方和除以数据个数与 1 的差量，所得的结果叫做样本方差。用原始数据计算见式(7-10)，用频数分布表计算见式(7-11)。

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] \tag{7-10}$$

$$S^2=\frac{1}{n-1}\sum_{j=1}^{k}f_j(x_j-\bar{x})^2=\frac{1}{n-1}\left[\sum_{j=1}^{k}f_jx_j^2-\frac{1}{n}\left(\sum_{j=1}^{k}f_jx_j\right)^2\right] \tag{7-11}$$

式(7-10)和(7-11)中，k 为频数分布表的组数，n 为样本大小，x_j 为第 j 组数据的组中值，f_j 为第 j 组的频数

一组性质相同的定量数据中的每一个与其总体算术平均值的差量的平方和除以数据个数，所得的结果叫做总体方差。用原始数据计算见式(7-12)，用频数分布表计算见式(7-13)。

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2=\frac{1}{N}\left[\sum_{i=1}^{N}x_i^2-(2\mu)\sum_{i=1}^{N}x_i+N\mu^2\right] \tag{7-12}$$

$$\sigma^2=\frac{1}{N}\sum_{j=1}^{k}f_j(x_j-\mu)^2=\frac{1}{N}\left[\sum_{j=1}^{k}f_jx_j^2-(2\mu)\sum_{j=1}^{k}x_j+N\mu^2\right] \tag{7-13}$$

式(7-12)和(7-13)中，k 为频数分布表的组数，N 为样本大小，x_j 为第 j 组数据的组中值，f_j 为第 j 组的频数，μ 为总体算术平均值

3. 标准差　方差的算术平方根叫做标准差，可分为样本方差和总体方差，分别见式(7-14)和式(7-15)。

$$\text{样本标准差 } S=\sqrt{\text{样本方差}}=\sqrt{S^2} \tag{7-14}$$

$$\text{总体标准差 } \sigma=\sqrt{\text{总体方差}}=\sqrt{\sigma^2} \tag{7-15}$$

式(7-14)中 S^2 的计算见式(7-10)和式(7-11)；式(7-15)中 σ^2 的计算见式(7-12)和式(7-13)

在可以计算标准差和方差的资料中，为了反映资料的离散度大小时，通常只用标准差，而不用方差，仅在对定量资料作统计分析(如方差分析)时要用到方差(称作均方)。

4. 变异系数　标准差与算术平均值之比值(通常以百分数形式给出)，叫做变异系数，记作 CV。其计算公式见式(7-16)。

$$CV=\frac{S}{\bar{x}}\times 100\% \tag{7-16}$$

标准差与变异系数的用法比较：当比较两组或多组定量资料的离散度大小时，在下面两种情形下不适合使用标准差，而必须使用变异系数。

情形一，当各组定量资料的单位(或叫量纲)不同时。

情形二，当各组定量资料的算术平均值相差悬殊时。

5. 标准误　为了明确地给出标准误的定义，需要知道两个重要的名词。它们分别是“统计量”和“参数”。

参数：反映总体中数据分布特征的量，称为参数。如总体平均值 μ、总体标准差 σ 等。

统计量：由样本数据确定的，不含任何未知参数的统计指标，叫做统计量。如样本平均值

$\bar{x}$，样本标准差 S，样本率，样本变异系数等。

标准误：统计量的标准差，叫做标准误。样本平均值的标准误记作 $S_{\bar{x}}$，样本率的标准误记作 S_p，变异系数的标准误记作 S_{CV}。它们的计算公式分别见式(7-17)至式(7-20)。

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \tag{7-17}$$

$$S_s = \sqrt{\frac{\sigma}{2n}} \tag{7-18}$$

$$S_P = \sqrt{\frac{P(1-P)}{n}} \tag{7-19}$$

$$S_{CV} = \sqrt{\frac{CV^2(1+2CV^2)}{2n}} \tag{7-20}$$

式(7-18)是在大样本(n 至少要>30，最好>100)条件下样本标准差的标准误，其中 σ 为总体标准差

标准差与标准误的用法比较：标准差适合用来反映一组性质相同的定量数据离开其算术平均值的波动大小，它反映了在相同条件下试验的重现性好坏(即精密度的高低)；而平均值的标准误则更适合用来反映在相同条件下试验的准确度的高低，它暗含对总体算术平均值 μ 数值大小的推测。因此，在表达几组类似实验结果离散度大小时，建议使用标准差，而不要盲目使用平均值的标准误。因为对同一组定量资料而言，标准误小于标准差，容易掩盖实际存在的较大的离散度，使那些原本不适合用近似正态分布法表达和处理的资料，都错误地运用了这些方法。

6. *分位数间距* 先用分位数法求出两个关于中位数对称的同类分位数，如第一四分位数(Q_1)与第三四分位数(Q_3)，令 $Q_R = Q_3 - Q_1$，则 Q_R 就叫做四分位数间距，这个数值的大小，标志着一组呈偏态分布定量资料居中的 50% 数值的离散度的大小。那么，我们可以用($P_{97.5} - P_{2.5}$)来反映一组呈偏态分布定量资料居中的 95% 数值的离散度的大小。

极差与四分位数间距的用法比较：四分位数间距反映了一组性质相同的定量数据中居中的 50% 的数据所在的范围，它比极差更有参考价值。

标准差与四分位数间距的用法比较：在偏态分布资料中，一般不适合使用标准差，此时可用四分位数间距取代标准差。

二、假设检验

参数估计和假设检验是统计推断的两个组成部分，它们都是利用样本对总体进行某种推断，然而推断的角度不同。参数估计是用样本统计量估计总体参数的方法，总体参数在估计前是未知的；假设检验的方法是首先假设样本对应的总体参数(或分布)与某个已知总体参数(或分布)相同，然后根据样本统计量的抽样分布规律，推断样本信息是否支持这种假设，并对原假设作出接受或拒绝的决定，以便在实践中采取相应的对策。

(一)假设检验的步骤

下面继续用例 2-1 说明假设检验的步骤。

1. *建立检验假设，确定检验性水准* 根据研究目的和专业知识，检验假设可分为单侧检验和双侧检验。

双侧检验：零假设 H_0：$\mu=\mu_0$，备择假设 H_1：$\mu\neq\mu_0$

单侧检验：零假设 H_0：$\mu=\mu_0$，备择假设 H_1：$\mu<\mu_0$，或者，

零假设 H_0：$\mu=\mu_0$，备择假设 H_1：$\mu>\mu_0$

检验水准 α 可定为 0.05，0.01 等。

如果研究目的是推断两总体均数是否相等（或者是否不相等），用双侧检验；如果在专业上认为 $\mu<\mu_0$（或 $\mu>\mu_0$）的情况不可能出现时，用单侧检验。在专业依据不明确的情况下，一般认为双侧检验比较稳妥，故双侧检验比较常用。

2. *选择检验方法和计算统计统计量*　根据研究目的和设计类型、资料的性质和定量资料所具备的前提条件等选择检验方法，计算检验统计量的值。不同的检验方法有其相应的检验统计量（即计算公式）和前提条件，例如，完全随机设计（两组平行组设计）的两样本均数比较，可根据定量资料所具备的前提条件是否满足选用 t 检验、F 检验、Z 检验（或叫 U 检验）或秩和检验等，相应的检验统计量分别为 t 值、F 值、Z 值（或 U 值）或 H_C 值。

3. *根据检验统计量确定 P 值，并作出统计推断和专业推断*　t 值是符合 t 分布的随机变量（即检验统计量），可以根据 t 分布规律及确定的检验水准确定 P 值。具体确定 P 值的方法有两种：

（1）手工查表得出 P 值。可以在“t 临界值表”中，通过自由度、单侧或是双侧检验、检验水准等信息，查得相应的检验统计量临界值。如果样本统计量绝对值 $\geqslant$ 临界值，则 $P\leqslant\alpha$，在 α 水准上，拒绝 H_0，接受 H_1；如果样本统计量绝对值<临界值，则 $P>\alpha$，在 α 水准不拒绝 H_0。

（2）由统计软件（SAS，SPSS 等）直接给出 P 值。若 $P\leqslant\alpha$，则在 $\alpha=0.05$ 水准上，拒绝 H_0，接受 H_1；若 $P>\alpha$，则不拒绝 H_0。

（二）把握度

拒绝不正确的 H_0 的概率，在统计学中称为检验功效，即把握度，记为 $1-\beta$。把握度的意义是：当两个总体参数间存在差异时，所使用的统计检验能够发现这种差异的概率，一般情况下要求把握度应在 0.8 以上。

在假设检验时，应兼顾犯Ⅰ类错误和犯Ⅱ类错误的概率。如果犯Ⅰ类错误的概率定得很小，势必增加犯Ⅱ类错误的概率，从而降低检验功效；反之，如果把犯Ⅱ类错误的概率定得很小，提高检验功效，则犯Ⅰ类错误的概率增大。为了同时减小犯Ⅰ类和Ⅱ类错误的概率，只有通过增加样本含量，减少抽样误差来实现。

三、两组平行组设计定量资料的假设检验

当两组平行组的定量资料满足参数检验的前提条件（独立性、正态性、等方差）时，选用 t 检验；否则可寻找到合适的变量变换方法使变换后的定量资料满足参数检验的前提条件，对变换后的数据采用 t 检验；或直接选用非参数的方法，如秩和检验。

若观测的定量指标只有 1 个，这种资料为平行组设计一元定量资料；若观测的定量指标有 $m(m>1)$ 个，即为平行组设计多元定量资料，可以选择多元统计分析方法。

（一）t 检验的适用条件

t 检验以 t 分布为理论基础，主要用于两组定量资料的总体均数比较，是定量资料分析中最常用的假设检验方法。应用 t 检验对定量资料进行统计推断之前应注意其使用条件，概括起来就是“独立性、正态性和等方差”：

1. 独立性　即各个观测值之间相互独立，比如以时间为一个重复测量因素的重复测量设计下收集的定量资料，在各时间点上测自同一个受试对象的数据之间具有一定联系，不符合独立性要求。

2. 正态性　即待分析的各组定量资料服从（或近似服从）正态分布，或者通过数据转换使之符合正态分布。

3. 等方差　例如，进行定量资料两总体均数比较时，要求两组定量资料所对应的两总体方差相等。

（二）一元参数检验（t 检验）与非参数检验（秩和检验）

【例 7-1】　本临床试验采用多中心、单盲、随机对照方法，验证雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统（试验组）的有效性和安全性，评价该支架的输送系统。与药物洗脱支架（对照组）进行对照，每组 20 例。终点为术后病变节段内直径狭窄程度（%），比较置入支架的两组病人术后病变节段内直径狭窄程度有无差别。数据如下：

试验组：19.3，25.8，28.3，16.0，17.9，11.2，27.3，17.3，41.6，17.3，21.6，11.8，35.8，15.5，10.3，16.3，11.0，10.8，23.8，31.1。

对照组：35.3，36.0，32.9，7.1，19.9，26.2，13.3，5.7，4.5，15.6，27.5，28.4，21.5，20.6，5.0，34.9，27.8，16.5，7.6，40.9。

对于这类资料，可以首先检查数据是否满足参数检验的前提条件，即是否服从正态分布，及是否等方差，如果满足，可以运用 t 检验，不满足选择秩和检验。

t 检验的步骤：

第一步：建立假设

H_0：$\mu_1 = \mu_2$，H_1：$\mu_1 \neq \mu_2$，$\alpha = 0.05$

其检验统计量为：

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S_{\bar{x}_1 - \bar{x}_2}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 + n_2)}{n_1 n_2 (n_1 + n_2 - 2)}(SS_1 + SS_2)}} \tag{7-21}$$

式（7-21）定义的检验统计量 t 服从自由度 $\nu = n_1 + n_2 - 2$ 的 t 分布，当 $t \geqslant t_{\alpha(n_1+n_2-2)}$ 时，就有 $P \leqslant \alpha$；反之，当 $t < t_{\alpha(n_1+n_2-2)}$ 时，就有 $P > \alpha$

第二步，计算检验统计量 t 值

第三步，确定概率作出统计推断。

秩和检验的步骤：

第一步：建立假设

H_0：两组患者退热时间总体分布相同，H_1：两组患者退热时间总体分布不同，$\alpha = 0.05$

第二步，排秩次：将两个组退热时间进行混合排序，以顺序号作为它们的秩次，遇到相同秩次取它们的平均数，但必须清楚每个数据所属的组别。

第三步，求秩和，确定检验统计量。具体略。

第四步，确定概率作出统计推断。

具体 SAS 程序（程序名为 CT7-1）如下：

程序	说明	程序	说明
DATACT7-1;	建立数据集	PROC SORT; BY g;	以 g 作为分组变量排序
INPUT g$;		RUN;	
DO i=1 to20;		Ods html;	
INPUT x @@;		PROC UNIVARIATENORMAL;	调用单变量过程
OUTPUT;		VAR x; BY g;	
END;CARDS;		RUN;	
A		PROC TTEST COCHRAN;	进行 t 检验
29.3 25.8 16.3 16.0 17.9		CLASS g;	
15.2 27.3 17.3 41.6 17.3		VAR x; RUN;	
11.6 11.8 35.8 15.5 16.3		PROC NPAR1WAY WILCOXON;	进行秩和检验
16.3 11.0 10.8 23.8 21.1		CLASS g; VAR x;	
B		RUN;	
35.3 36.0 32.9 7.1 19.9		Ods html close;	
26.2 13.3 5.7 4.5 15.6			
27.5 28.4 21.5 20.6 5.0			
34.9 27.8 16.5 7.6 40.9;			
RUN;			

主要结果及结果解释：

g=A

矩			
N	20	权重总和	20
均值	20.5	观测总和	410
标准差	8.828 959 53	方差	77.950 52 63
偏度	0.889 866 83	峰度	0.205 967 53
未校平方和	9 886.06	校正平方和	1 481.06
变异系数	43.068 09 52	标准误差均值	1.974 215 37

这是试验组的一般统计量计算指标。

正态性检验				
检验	统计量		P 值	
Shapiro-Wilk	W	0.915 017	$Pr < W$	0.079 5
Kolmogorov-Smirnov	D	0.165 807	$Pr > D$	>0.150 0
Cramer-von Mises	W-Sq	0.089 283	$Pr > W-Sq$	0.148 0
Anderson-Darling	A-Sq	0.559 637	$Pr > A-Sq$	0.133 7

这是试验组数据正态性检验结果，W 检验显示，$P=0.0795>0.05$，数据服从正态分布

g=B

矩			
N	20	权重总和	20
均值	21.36	观测总和	427.2
标准差	11.638 068 2	方差	135.444 632
偏度	−0.039 421 3	峰度	−1.244 936 8
未校平方和	11 698.44	校正平方和	2 573.448
变异系数	54.485 338	标准误差均值	2.602 351 16

这是对照组的一般统计量计算指标。

正态性检验				
检验	统计量		*P* 值	
Shapiro-Wilk	W	0.938 638	*Pr* < W	0.225 9
Kolmogorov-Smirnov	D	0.131 462	*Pr* > D	>0.150 0
Cramer-von Mises	W-Sq	0.049 707	*Pr* > W-Sq	>0.250 0
Anderson-Darling	A-Sq	0.385 463	*Pr* > A-Sq	>0.250 0

这是对照组数据正态性检验结果，W 检验显示，$P=0.2259>0.05$，数据服从正态分布。

Equality of Variances					
Variable	Method	Num DF	Den *DF*	F Value	*Pr* > F
x	Folded F	19	19	1.74	0.237 6

以上是两组方差齐性检验的结果，$P=0.237\ 6>0.05$，两组总体方差相等。

T-Tests					
Variable	Method	Variances	*DF*	*t* Value	*Pr* > |*t*|
x	Pooled	Equal	38	−0.26	0.793 8
x	Satterthwaite	Unequal	35.4	−0.26	0.793 9
x	Cochran	Unequal	19	−0.26	0.795 2

这是两组 *t* 检验的结果，$t=-0.26$，$P=0.793\ 8>0.05$，说明两组人群术后病变节段内直径狭窄程度(%)是没有统计学差异的。

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x Classified by Variable g					
g	*N*	Sum of Scores	Expected Under H_0	Std Dev Under H_0	Mean Score
A	20	399.0	410.0	36.966 721	19.950
B	20	421.0	410.0	36.966 721	21.050
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	399.000 0
Normal Approximation	
Z	−0.284 0
One-Sided $Pr < Z$	0.388 2
Two-Sided $Pr > \|Z\|$	0.776 4
t Approximation	
One-Sided $Pr < Z$	0.388 9
Two-Sided $Pr > \|Z\|$	0.777 9
Z includes a continuity correction of 0.5	

这是两组秩和检验的结果，如果数据不满足正态性或等方差，可以查验这部分结果，$P=0.776\ 4>0.05$，两组之间差异无统计学意义。

结论：NOYA 雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统与对照器械相比，患者术后病变节段内直径狭窄程度(%)是没有差别的。

(三)两组平行组设计多元统计分析(T^2检验)

对于平行组设计，如果终点是多个，而临床认为终点之间存在着关联，即观测的定量指标有 $m(m>1)$个，此时可以选择多元统计分析方法，如，T^2检验。

【例 7-2】　本临床试验采用多中心、单盲、随机对照方法，验证雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统(试验组)的有效性和安全性，评价该支架的输送系统。与药物洗脱支架(对照组)进行对照，每组 20 例。主要指标为术后参照血管直径(mm)、支架内直径狭窄程度(%)、支架内最小管腔直径(mm)，比较置入支架的两组病人术后 3 项指标有无差别。数据如表 7-1 所示。

表 7-1　试验组与对照组术后情况

分组	参照血管直径(mm)	支架内直径狭窄程度(%)	支架内最小管腔直径(mm)
试验组	2.91	22.45	2.44
	3.13	8.51	2.47
	⋮	⋮	⋮
	3.22	1.05	2.24
对照组	2.63	15.16	2.85
	1.37	12.82	1.93
	⋮	⋮	⋮
	3.38	5.46	2.22

样本含量在实际计算中需要计算，具体数据见 SAS 程序。

T^2检验的具体步骤为：

第一步，建立检验假设。$H_0:\mu_A=\mu_B$，$H_1:\mu_A\neq\mu_B$，$\alpha=0.05$

第二步，按下式计算检验统计量 Hotelling T^2的值。

$$T^2=\frac{n_An_B}{n_A+n_B}[\overline{X}_A-\overline{X}_B]'S^{-1}[\overline{X}_A-\overline{X}_B]$$

其中 $\overline{X}_A$，$\overline{X}_B$ 分别代表各组的样本均值向量，S 为样本协方差矩阵的逆矩阵。n_A，n_B 分别为各组样本数

F 与 Hotelling T^2 有如下关系：

$$F=\frac{n_A+n_B-m-1}{(n_A+n_B-2)m}T^2,\nu_1=m,\nu_2=n_A+n_B-m-1$$

m 指反应变量，即观测的定量指标的个数。

第三步，根据相应的检验统计量 F 和自由度查临界值表确定 p 值，下结论。

具体 SAS 程序(程序名为 CT7-2)如下：

程序	程序
DATA CT7-2; Do group=0 to 1;C Do i= 1 to 20; INPUT X_1-X_3@@; OUTPUT; END;END; CARDS; 2.91 22.45 2.44 3.13 8.51 2.47 3.64 5.29 2.49 3.61 24.89 2.84 3.06 19.30 2.65 2.53 6.60 2.46 2.65 8.43 2.74 4.12 13.98 2.90 2.46 7.06 1.90 2.67 16.12 3.08 2.51 5.97 2.64 2.41 6.30 2.20 2.41 20.12 2.80 2.85 20.40 2.83 2.76 4.85 2.64 2.39 16.34 2.13 2.00 36.54 2.86 3.60 8.27 3.13 2.61 27.95 2.06 3.22 1.05 2.24 2.63 15.16 2.85 1.37 12.82 1.93	2.80 0.49 2.87 3.08 16.36 2.30 2.98 4.42 2.69 3.04 16.87 2.64 3.27 2.50 2.87 2.69 15.28 2.90 2.92 14.17 3.40 3.28 3.14 2.93 3.24 3.15 2.56 3.27 11.14 1.88 4.00 3.61 2.60 3.18 21.21 2.68 3.16 3.97 2.31 3.24 25.29 2.41 2.50 18.78 3.08 2.67 6.86 3.02 3.47 28.52 2.77 3.38 5.46 2.22; RUN; ODS HTML; PROC GLM data= CT7-2; CLASSgroup; MODEL X_1-X_3=group / SS3 NOUNI; MANOVA H=group; MEANSgroup; RUN;QUIT; Ods html close;

主要结果及结果解释：

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall group Effect
H = Type Ⅲ SSCP Matrix for group
E = Error SSCP Matrix
S=1 M=0.5 N=17

Statistic	Value	*F* Value	Num DF	Den DF	*Pr* > F
Wilks' Lambda	0.958 874 29	0.51	3	36	0.674 8
Pillai's Trace	0.041 125 71	0.51	3	36	0.674 8
Hotelling-Lawley Trace	0.042 889 58	0.51	3	36	0.674 8
Roy's Greatest Root	0.042 889 58	0.51	3	36	0.674 8

Level of group	N	X_1		X_2		X_3	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
0	20	2.877 000 00	0.535 823 03	14.021 000 0	9.338 928 65	2.575 000 00	0.340 162 50
1	20	3.008 500 00	0.516 305 45	11.460 000 0	8.224 351 52	2.645 500 00	0.382 587 80

三元方差分析的结果为 Wilks′λ=0.958 9，由其转换而来的近似 $F=0.51$，$P=0.674\ 8$。说明试验组与对照组人群在术后参照血管直径(mm)、支架内直径狭窄程度(%)、支架内最小管腔直径(mm)3 个终点上总的来说并无明显差异。

四、多个平行组设计定量资料的假设检验

通常，根据试验方案的需要，可为试验组设置一个或多个对照组，试验器械也可按照若干种治疗强度设组，此时即为多个平行组设计，选择的统计分析方法也不尽相同。

(一)方差分析

方差分析(anlysis of variance，常缩写成 ANOVA)又称 F 检验，其理论基础是 F 分布，目的是推断两组或多组定量资料的总体均数是否相同，检验两个或多个样本均数的差异是否具有统计学意义，一般用于：①多个平行组设计下的多个均数之间的比较；②分析两个或多个因素对定量观测结果影响时各因素的主效应及其交互作用效应是否具有统计学意义；③简单回归方程分析中某些方面的假设检验；④多重线性回归分析中某些方面的假设检验；⑤两个或多个总体方差齐性检验等。

应用方差分析前提条件和单因素两水平设计定量资料的 t 检验相同，即“独立性、正态性和等方差”。

方差分析的基本步骤：

H_0：各组样本均数对应的总体均数相等；

H_1：各组样本均数对应总体均数不等或不全相等。

检验水准为 α。

若不拒绝 H_0 时，可认为各样本均数间的差异是由于抽样误差所致，而不是由于处理因素的作用所致。理论上，此时的组间变异与组内变异应相等，两者的比值即检验统计量的取值为 1；由于存在抽样误差，两者往往不恰好相等，但相差不会太大，检验统计量的值应接近于 1。

若拒绝 H_0，接受 H_1 时，可认为各组均数间的差异，不仅是由抽样误差所致，还有处理因

素的作用。此时的组间变异远>组内变异，两者的比值即检验统计量的取值远>1。在实际应用中，当检验统计量的值超过其相应的检验临界值时，拒绝 H_0，接受 H_1，即意味着各组均数间的差异，不仅是由抽样误差所致，更主要是由处理因素引起的。

组间变异和组内变异与自由度有关，所以不能直接比较离均差平方和。为减小自由度的影响，将各部分的离均差平方和除以各自的自由度，得到相应的平均变异指标—均方。组间及组内均方计算公式为：

$$MS_{组间}=\frac{SS_{组间}}{\upsilon_{组间}}$$

$$MS_{组内}=\frac{SS_{组内}}{\upsilon_{组内}}$$

将组内均方除以组间均方，即可得到方差分析的检验统计量 F：

$$F=\frac{MS_{组间}}{MS_{组内}}，\nu_{总}=\nu_{组间}+\nu_{组内} \tag{7-22}$$

可以证明，当各组样本来自同一总体时，F 是服从分子自由度为 $\upsilon_{间组}$，分母自由度为 $\upsilon_{组内}$ 的 F 分布，当 $F=F_{\alpha(\upsilon_1,\upsilon_2)}$ 时，则 $P=\alpha$；$F>F_{\alpha(\upsilon_1,\upsilon_2)}$ 时，则 $P<\alpha$；$F<F_{\alpha(\upsilon_1,\upsilon_2)}$ 时，则 $P>\alpha$，ν_1、ν_2 分别为组间自由度、组内自由度。

(二)一元参数检验(方差分析)与非参数检验(秩和检验)

对多个平行组设计设计定量资料进行方差分析时，对定量资料的要求与两个平行组设计设计定量资料 t 检验的前提条件是一样的，即定量资料应满足独立性、正态性、等方差。若定量资料不满足参数检验的前提条件时，则可选用秩和检验进行分析。

【例 7-3】 本临床试验采用多中心、单盲、随机对照方法，验证雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统(试验组)的有效性和安全性，评价该支架的输送系统。与市场常规支架系统进行对照，设计两个对照组(A,B)每组 20 例。终点为术后病变节段内直径狭窄程度(%)，比较置入支架的 3 组病人术后病变节段内直径狭窄程度有无差别。数据如下：

试验组：19.3，25.8，28.3，16.0，17.9，11.2，27.3，17.3，41.6，17.3，21.6，11.8，35.8，15.5，10.3，16.3，11.0，10.8，23.8，31.1。

对照 A 组：35.3，36.0，32.9，7.1，19.9，26.2，13.3，5.7，4.5，15.6，27.5，28.4，21.5，20.6，5.0，34.9，27.8，16.5，7.6，40.9。

对照 B 组：18.7，11.2，27.3，17.3，9.4，17.3，14.6，11.8，25.8，12.6，27.5，22.3，29.5，28.6，15.2，34.9，21.4，21.5，27.4，7.8。

方差分析的主要步骤：

第一步，给出检验假设及检验水准：

$H_0:\mu_1=\mu_2=\mu_3$；$H_1:\mu_1,\mu_2,\mu_3$不等或不全相等；$\alpha=0.05$。

第二步，计算检验统计量 F 值：按式(7-23)计算。

$$F=\frac{MS_b}{MS_e}，\ \nu_b=k-1,\nu_e=N-k,\nu_t=N-1 \tag{7-23}$$

式中，MS_b 与 MS_e 的计算步骤如下：

$$C(校正数)=\frac{(\sum\sum X_{ij})^2}{N}=全部数据和的平方除以总例数$$

SS_t(总离均差平方和)= $\sum\sum X_{ij}^2 - C$ =各数据平方求和减去 C

T_j(第 j 组数据之和)= $\sum_{i=1}^{n_j} X_{ij}$

SS_b(组间离均差平方和)= $\sum_{j=1}^{k} \frac{T_j^2}{n_j} - C$ =各组数据和的平方除以各组例数后求和再减去 C

$SS_e = SS_t - SS_b$

MS(组间均方)= $\frac{SS_b}{df_b}$；MS(组内均方)= $\frac{SS_e}{df_e}$

查方差分析用的 F 界值表，得 $F_{\alpha(df_b,df_e)}$，若 $F \geqslant F_{\alpha(df_b,df_t)}$，则 $P<\alpha$，反之，则有 $P>\alpha$。

第三步，确定 P 值并作出统计推断。最后，作出统计推断，并结合临床专业知识给出专业结论。

秩和检验主要步骤：

第一步，给出检验假设及检验水准：

H_0：3 组定量资料的总体分布相同；H_1：3 组定量资料的总体分布不同或不完全相同；$\alpha = 0.05$。

第二步，编秩：将 3 组原始数据由小到大统一排序。编秩时如遇同组相同数据则顺列秩次；如遇不同组相同数据则取平均秩次。各数据所属组号必须清楚，以便分别求秩和。

第三步，求秩和：将各组秩次相加求出 R_i，下标 i 表示组序。

第四步，计算检验统计量 H 值：如果没有相同秩次时，秩次服从均匀分布，按式(7-24)计算。

$$H = \frac{12}{N(N+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(N+1) \tag{7-24}$$

式中，n_i：各组测定值个数；$N = \sum n_i$：各组测定值个数之和

如果有相同秩次时，上述以均匀分布为基础导出的公式(7-24)，需要进行校正，校正因子为 C：

$$C = 1 - \frac{\sum_{j=1}^{l}(t_j^3 - t)}{(N^3 - N)} \tag{7-25}$$

校正的 Kruskal-Wallis 检验的检验统计量为：

$$H_c = H/C \tag{7-26}$$

需要指出的是用 SAS 输出的结果为校正后的结果。

第五步，确定 P 值并作出统计推断。若组数 $k=3$，且每组例数 $\leqslant 5$，可查 H 界值表，得出 P 值。若最小样本的例数>5，则 H 近似服从 $\upsilon = k-1$ 的 χ^2 分布。最后，作出统计推断，并结合专业知识，给出专业结论。

具体 SAS 程序(程序名为 CT7-3 如下：

程序	说明
DATA CT7-3;	以下是建立数据集
INPUT GROUP $ n;	
DO i=1 to n;	
INPUT x @@; OUTPUT;	
END;	
CARDS;	输入数据
GROUP120	
19.3 25.8 28.3 16.0 17.9 11.2 27.3 17.3 41.6	
17.3	
21.6 11.8 35.8 15.5 10.3 16.3 11.0 10.8 23.8	
31.1	
GROUP220	
35.3 36.0 32.9 7.1 19.9 26.2 13.3 5.7 4.5 15.6	
27.5	
28.4 21.5 20.6 5.0 34.9 27.8 16.5 7.6 40.9	
GROUP320	
18.7 11.2 27.3 17.3 9.4 17.3 14.6 11.8 25.8 12.6	
27.5 22.3 29.5 28.6 15.2 34.9 21.4 21.5 27.4 7.8;	
RUN;	
ODS HTML;	
PROC SORT data= CT7-3; BY GROUP; RUN;	
PROC UNIVARIATE NORMAL data= CT7-3;	
VAR x; BY GROUP;	对各组进行正态性检验
RUN;	
PROC GLM DATA=CT7-3;	
CLASS GROUP;MODEL X=GROUP / SS3;	调用 GLM 过程
MEANS GROUP / HOVTEST SNK ;	进行方差分析
RUN;	进行方差齐性检验和两两比较
PROC NPAR1WAY WILCOXON data= CT7-3;	
CLASS GROUP; VAR x;	调用非参数检验过程
RUN;	
PROC RANK DATA= CT7-3 OUT=BBB;	
VAR X; RANKS RX;	调用 RANK 过程对 3 组数据排序
RUN;	
PROC ANOVA DATA=BBB;	排秩后生成新变量 RX
CLASS GROUP; MODEL RX=GROUP;	对排秩后的数据调用 ANOVA 过程进行两两比较
MEANS GROUP/SNK;	
RUN;	
ODS HTML CLOSE;	

主要输出结果及结果解释：

GROUP=GROUP1

正态性检验

检验	统计量		P 值	
Shapiro-Wilk	W	0.915 017	Pr < W	0.079 5
Kolmogorov-Smirnov	D	0.165 807	Pr >D	>0.150 0
Cramer-von Mises	W-Sq	0.089 283	Pr > W-Sq	0.148 0
Anderson-Darling	A-Sq	0.559 637	Pr > A-Sq	0.133 7

GROUP=GROUP2

正态性检验

检验	统计量		P 值	
Shapiro-Wilk	W	0.938 638	Pr<W	0.225 9
Kolmogorov-Smirnov	D	0.131 462	Pr>D	>0.150 0
Cramer-von Mises	W-Sq	0.049 707	Pr > W-Sq	>0.250 0
Anderson-Darling	A-Sq	0.385 463	Pr > A-Sq	>0.250 0

GROUP=GROUP3

正态性检验

检验	统计量		P 值	
Shapiro-Wilk	W	0.961 563	Pr<W	0.575 5
Kolmogorov-Smirnov	D	0.125 379	Pr>D	>0.150 0
Cramer-von Mises	W-Sq	0.048 012	Pr > W-Sq	>0.250 0
Anderson-Darling	A-Sq	0.309 181	Pr > A-Sq	>0.250 0

以上是 3 组正态性检验的结果，由结果可知，3 组数据均服从正态分布，3 组概率分别为 $P=0.079\ 5$，$P=0.225\ 9$，$P=0.575\ 5$。

Levene's Test for Homogeneity of x Variance

ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
GROUP	2	57 115.8	28 557.9	3.15	0.050 3
Error	57	516 236	9 056.8		

这是方差齐性检验的结果，由结果可知，3 组总体方差相等，$P=0.050\ 3>0.05$。

The GLM Procedure

Dependent Variable: x

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16.471 000	8.235 500	0.09	0.913 4
Error	57	5 177.057 500	90.825 570		
Corrected Total	59	51 93.528 500			

这是方差分析的结果，$F=0.09$，$P=0.9134$，说明3组之间差别无统计学意义。

Student-Newman-Keuls Test for x

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	GROUP
A	21.360	20	GROUP2
A			
A	20.500	20	GROUP1
A			
A	20.105	20	GROUP3

这是3组之间两两比较的结果，可知，各组平均值之间差别均无统计学意义。

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable x Classified by Variable GROUP

GROUP	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
GROUP1	20	596.00	610.0	63.755 359	29.800 0
GROUP2	20	634.50	610.0	63.755 359	31.725 0
GROUP3	20	599.50	610.0	63.755 359	29.975 0

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	0.1487
DF	2
$Pr >$ Chi-Square	0.9284

这是3组非参数秩和检验的结果，如果数据不满足正态性，或总体方差不相等，可以查看此部分结果，$P=0.9284>0.05$。

Student-Newman-Keuls Test for RX

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	GROUP
A	31.725	20	GROUP2
A			
A	29.975	20	GROUP3
A			
A	29.800	20	GROUP1

这是对 3 组排秩后，进行两两比较的结果，如果 3 组不服从参数检验前提条件时查看此部分结果，结果可知，3 组之间两两比较差别无统计学意义。

结论：NOYA 雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统（试验组）与 A，B 两个对照组人群术后病变节段内直径狭窄程度没有差别。

（三）多元参数检验（多元方差分析）

多个平行组设计中，如果终点是多个，临床认为多个终点之间存在关联，此时应该使用多元统计分析方法考察多个组在多个终点是否存在差别。多元方差分析与一元方差分析是类似的。多元方差分析的主要思想是对方差-协方差矩阵的分解。自由度的分解与单因素分析时是一致的（表 7-2）。

多组均值向量比较的检验统计量 Wilks 's Λ：

$$\Lambda=\frac{|E|}{|E+H|} \tag{7-27}$$

表示组内变异（误差）在总变异中的比例。

表 7-2　多元方差分析的方差分解表

方差来源	DF	离均差平方和矩阵
组间	$G-1$	$E-\sum_{g=1}^{G} n_g(X_g-\bar{X}_{..})(X_g-\bar{X}_{..})'$
组内	$\sum_{g=1}^{G} n_g-G$	$H=\sum_{g=1}^{G}\sum_{j=1}^{n_g}(X_{gj}-\bar{X}_g)(X_{gj}-X_g)'=\sum_{g=1}^{G}(n_g-1)S_g$
合计	$\sum_{g=1}^{G} n_g-1$	$H+E$

说明：其中，n_g 表示第 g 组中的样本量，X_{gj} 表示第 g 组中第 j 个个体的观察向量，$\bar{X}_g$ 表示第 g 组中的均数向量，$X_{..}$ 表示个体总均数向量，H（组间）相当于 ANOVA 中的 SS_B（组间），E（组内）相当于 SS_W（组内），(…)'(…) 表示列向量与行向量的乘积，结果是一个 p×p 维矩阵。在 SAS 输出结果中，Λ 可自动转换为检验统计量 F

【例 7-4】　本临床试验采用多中心、单盲、随机对照方法，验证雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统（试验组）的有效性和安全性，评价该支架的输送系统。与两组对照组（A，B）进行比较，每组 20 例。主要指标为术后参照血管直径（mm）、支架内直径狭窄程度（%）、支架内最小管腔直径（mm），比较置入支架的两组病人术后 3 项指标有无差别。数据如表 7-3 所示。

表 7-3　试验组与对照组术后情况

分组	参照血管直径（mm）	支架内直径狭窄程度（%）	支架内最小管腔直径（mm）
试验组	2.91	22.45	2.44
	3.13	8.51	2.47
	⋮	⋮	⋮
	3.22	1.05	2.24

（续 表）

分组	参照血管直径(mm)	支架内直径狭窄程度(%)	支架内最小管腔直径(mm)
对照组 A	2.63	15.16	2.85
	1.37	12.82	1.93
	⋮	⋮	⋮
	3.38	5.46	2.22
对照组 B	3.66	12.33	2.99
	2.50	5.97	2.59
	⋮	⋮	⋮
	2.92	15.22	3.44

样本含量在实际计算中需要计算，具体数据见 SAS 程序

第一步，建立检验假设。

$H_0:\mu_A=\mu_B=\mu_C$，$H_1:\mu_A,\mu_B,\mu_C$ 不等或不全相等，$\alpha=0.05$

第二步，按式(7-27)计算检验统计量 Λ 的值，并将 Λ 的值转化为 F 值。

第三步，确定 p 值，下结论。

具体 SAS 程序(程序名为 CT7-4)如下：

程序	说明
DATA CT7-4;	建立数据集
Do A=1 to 3;	
Do i= 1 to 20;	
INPUT X1-X3 @@;	
OUTPUT; END;END;	
CARDS;	
/*输入表 7-3 数据*/;	输入数据
RUN;	
Odshtml;	
PROC GLM data=CT7-4;	
CLASSA; MODEL X1-X3=A/NOUNI;	调用 GLM 过程 NOUNI 表示只做多元方差分析不做一元方差分析
CONTRAST 'A1 vs A2' A1 -1 0;	C ONTRAST 表示对 A 因素的 3 个水平，任何两个水平之间的比较时，相应的两个水平中 1 个用“1”表示，另一个用“-1”
CONTRAST 'A1 vs A3' A1 0 -1;	
CONTRAST 'A2 vs A3' A0 1 -1;	
MANOVA H=A; MEANS A;	
RUN;	
Odshtml close;	

主要输出结果及结果解释如下：

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall A Effect

H = Type Ⅲ SSCP Matrix for A

E = Error SSCP Matrix

S=2 M=0 N=26.5

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.934 157 22	0.64	6	110	0.701 9
Pillai's Trace	0.066 641 01	0.64	6	112	0.6953
Hotelling-Lawley Trace	0.069 629 13	0.63	6	71.583	0.703 6
Roy's Greatest Root	0.053 723 83	1.00	3	56	0.398 4

NOTE: F Statistic forRoy's Greatest Root is an upper bound

NOTE: F Statistic for Wilks' Lambda is exact

这是多元方差分析的结果，Wilks's Lambda=0.066 6，与之对应的近似 *F* 值为 0.64，*P*=0.69 53>0.05，说明 3 组在结果 3 个指标上的差别没有统计学意义。

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda(A_1 vs A_2)	0.969 972 12	0.57	3	55	0.638 7
Wilks' Lambda(A_1 vs A_3)	0.979 708 76	0.38	3	55	0.768 0
Wilks' Lambda(A_2 vs A_3)	0.949 870 83	0.97	3	55	0.414 7

以上是 3 组间两两比较多元方差分析的结果，*P* 均>0.05，说明三组两两比较结果差别无统计学意义。

Level of A	N	X_1		X_2		X_3	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1	20	2.877 000 00	0.535 823 03	14.021 000 0	9.338 928 65	2.575 000 00	0.340 162 50
2	20	3.008 500 00	0.516 305 45	11.460 000 0	8.224 351 52	2.645 500 00	0.382 587 80
3	20	2.706 000 00	0.711 864 86	13.490 500 0	8.606 248 33	2.613 000 00	0.343 190 91

以上是不同组别临床终点(参照血管直径、支架内直径狭窄程度、支架内最小管腔直径)的均值向量。

结论：NOYA 雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统与两组对照组的 3 个临床终点(参照血管直径、支架内直径狭窄程度、支架内最小管腔直径)指标没有差别。

(李　卫　郭　晋　王　杨)

第 8 章 非诊断试验中的多重比较

在临床试验中，通常根据试验方案的需要，可为试验组设置一个或多个对照组，试验器械也可按照若干种治疗强度设组，此时就会出现多个平行组，对于这类数据进行组与组之间的统计分析统称为多重比较或者两两比较，本章就两两比较的方法作出简要的介绍，并结合实例给出相应的 SAS 程序和输出结果的具体解释。

在多个平行组设计中，对组与组之间进行两两比较时，如果仍然使用 t 检验或者两个独立样本比较的 Wilcoxon 秩和检验，就会增加犯Ⅰ类错误的概率。其原因在于，虽然每进行一次比较犯Ⅰ类错误的概率依旧是事先所确定的显著性水平，通常为 $\alpha=0.05$，但是比较的次数却大大增加了。假设有一个对照组，3 个试验组，也就是有 4 个样本均数，则进行任意两组间相互比较的总次数为 $C_4^2=6$ 次，这时完成全部 45 次比较后，所犯的Ⅰ类错误的概率为 $1-(1-0.05)^6=0.265$。显然，在这种情况下选择 t 检验进行多个水平之间的两两比较是不恰当的，需要选择其他的校正的统计分析方法。

与两两比较有关的一些基本的概念，这里做一简单介绍。每做一次比较，犯Ⅰ类错误的概率称为比较误差率；做完全部比较所犯Ⅰ类错误的概率称为实验误差率；完全无效假设是指所有组资料所对应的总体均数都相等，也就是进行方差分析时建立的原假设；部分无效假设则是指部分组所对应的总体均数相等；在任何完全或部分无效假设下，做完全部比较所犯Ⅰ类错误概率的最大值称为最大实验误差率。

一、多个平均值两两比较方法

多重比较从功能上讲主要分为两类，第一类即任何两个均数之间都要进行比较；第二类为多个试验组均数之间不比较，但需要将各试验组分别与同一个对照组均数比较。

设某试验因素有 k 个水平，也就是有 k 个均数，第 i 组和第 j 组的总体均数分别为 μ_i 和 μ_j，样本均数分别为 $\bar{X}_i$ 和 $\bar{X}_j$，样本例数分别为 n_i 和 n_j，$MS_{误差}$ 为某种特定试验设计定量资料方差分析中的误差均方，误差项的自由度为 $\nu_{误差}$。

（一）所有处理组与对照组均数的比较

在很多研究中，分析的目的是比较多个处理组与一个对照组均数之间差异有无统计学意义，在试验因素的 k 个水平中，有 $k-1$ 个处理组和一个对照组，此时只需要进行 $k-1$ 次检验即可。常用的方法为 Dunnett's t 检验法，它是 t 检验法的一种修正。

假设 μ_0 为对照组对应的总体均数，$\bar{X}_0$ 与 n_0 分别为对照组的样本均数和例数，μ_i 为第 i 个处理组的总体均数，$i=1,2,\cdots,k-1$，则原假设和备择假设分别为：

$H_0:\mu_i=\mu_0$

$H_1:\mu_i\neq\mu_0$

检验统计量为：

$$t=\frac{\bar{X}_i-\bar{X}_0}{S_{X_i\ X_0}},\nu=\nu_{误差} \tag{8-1}$$

$$S_{X_i\ X_0}=\sqrt{MS_{误差}\left(\frac{1}{n_i}+\frac{1}{n_0}\right)} \tag{8-2}$$

计算出检验统计量后，将其与 Dunnett 检验的临界值 $d_{\alpha/2,(k-1,\nu)}$ 比较，临界值可以从 Dunnett 检验专用的界值表中查得，由检验的显著性水平 α，处理组个数 $k-1$ 和自由度 ν 确定。需要注意的是，该界值表有单侧和双侧之分。对于双侧检验，当显著性水平 $\alpha=0.05$ 时，若有 $|t|\geqslant d_{0.05/2,(k-1,\nu)}$，则 $P\leqslant 0.05$。如果在专业上有明确的证据，已知处理组均数不会大于或小于对照组均数时，可以进行单侧检验。SAS 软件可以进行 Dunnett 法的单侧和双侧检验。

上述假设检验可以用置信区间的形式表达，两总体均数差值 $\mu_i-\mu_0$ 的双侧 $100(1-\alpha)\%$ 置信区间为：

$$\bar{X}_i-\bar{X}_0-d_{\alpha/2,(k-1,\nu)}S_{X_i\ X_0}\leqslant\mu_i-\mu_0\leqslant\bar{X}_i-\bar{X}_0+d_{\alpha/2,(k-1,\nu)}S_{X_i-X_0} \tag{8-3}$$

同样，对于单侧检验，可以计算相应的单侧 $100(1-\alpha)\%$ 置信区间，置信区间的结论与假设检验是一致的。

Dunnett 检验控制最大试验误差率在不超过事先给定的 α 水平上。

(二)所有组之间的两两比较

在进行多个总体均数两两之间的全面比较时，需要进行的检验的次数为 $\binom{k}{2}=\frac{k\times(k-1)}{2}$。检验的原假设和备择假设分别为：

$H_0:\mu_i=\mu_j$

$H_1:\mu_i\neq\mu_j$

所有组之间的两两比较的方法较多，下面分别给出简要介绍。

1. LSD 法　LSD 法也称最小显著差异法(least significant difference)，适用于一对或几对在专业上有特殊意义的总体均数间的比较。当某种特定试验设计定量资料方差分析的结果表明某因素各水平组的均数之间的差异有统计学意义时，使用 LSD 法作进一步的分析。检验统计量的计算公式为：

$$t=\frac{\bar{X}_i-\bar{X}_j}{S_{X_i\ X_j}},\nu=\nu_{误差} \tag{8-4}$$

$$S_{X_i\ X_j}=\sqrt{MS_{误差}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{8-5}$$

如果设计是平衡的，即各组样本例数相等，此时 $n_i=n_j=n$，式(8-5)可以简化为 $\sqrt{\frac{2MS_{误差}}{n}}$。

$$\mathrm{LSD}=t_{\alpha,\nu}\sqrt{MS_{误差}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{8-6}$$

称为最小显著性差异。使用 LSD 法时，只要简单地将观察到的每一对样本平均值的差值与对应的 LSD 值比较就可以了。就双侧检验而言，如果 $|\bar{X}_i-\bar{X}_j|\geqslant$LSD，则 $P\leqslant\alpha$，就说明被比较的两组总体均数之间的差异具有统计学意义。

用这种方法做比较的次数越多，试验误差率就越大，即做完全部比较所犯Ⅰ类错误的概率就越大。

2. Bonferroni 法　Bonferroni 提出，如果在 α' 水准上进行 c 次假设检验，当原假设为真时，至少有一次拒绝原假设的累积Ⅰ类错误概率，即做完全部比较所犯Ⅰ类错误的概率 α 不超过 $c\times\alpha'$，也就是有不等式 $\alpha<c\times\alpha'$。因此可以重新选择Ⅰ类错误概率水准 α'，以便使试验误差率 α 控制在规定的水平之内。取 $\alpha'=\alpha/c$，当因素有 k 个水平时，$c=k(k-1)/2$，则有 $\alpha'=2\alpha/[k(k-1)]$。假设事先规定 $\alpha=0.05$，另外 $k=3$，则比较的总次数 $c=[3\times(3-1)]2=3$，每次进行比较的检验水准 $\alpha'=0.05/3=0.016$。

用 Bonferroni 法进行多个平均值之间两两比较的检验统计量计算公式同式(8-4)，就前一段中的例子来说，每次进行比较的临界值为 $t_{0.016/2,\nu}$，对于双侧检验，若检验统计量 $|t|\geqslant t_{0.016/2,\nu}$，则 $P\leqslant 0.016$，两个均数的差别有统计学意义。

当比较次数不多时，Bonferroni 法的效果较好；但当比较次数较多(例如在 10 次以上)时，由于其检验水准选择过低，结论偏于保守。

3. Sidak 法　与 Bonferroni 法类似，根据 Sidak 的不等式同样可以对 t 检验法进行校正，规定每次进行比较的检验水准 $\alpha'=1-(1-\alpha)^{1/c}$。仍沿用前例，$\alpha'=1-(1-0.05)^{1/3}=0.017$，对于双侧检验，若检验统计量 $|t|\geqslant t_{0.017/2,\nu}$，则 $P\leqslant 0.017$，两个均数的差别有统计学意义。

4. Scheffé 法　它是由 Scheffé 于 1953 年和 1959 年提出的控制试验误差率的方法，其检验统计量为 F，计算公式为：

$$F=\frac{t^2}{k-1} \tag{8-7}$$

其中 t 的计算见式(5-4)。当 $F\geqslant F_{\alpha(k-1,\nu)}$ 时，认为两总体均数的差异有统计学意义。

Scheffé 检验的结果与多组平行组设计定量资料方差分析的结果是相容的，即若方差分析的结果有统计学意义，则用此法至少能发现一次比较的结果有统计学意义；反之，若方差分析的结果为无统计学意义，则用此法也找不出任何两个均数之间差别有统计学意义(然而，大部分多重比较方法可能会发现某些对比组之间差别有统计学意义)。

如果比较的次数明显地大于均数的个数时，Scheffé 法的检验功效可能高于 Bonferroni 法和 Sidak 法。

5. Tukey-Kramer 法　Tukey-Kramer 法也称为 Tukey 法或诚实显著性差异(Honestly significant difference，HSD)检验，它由 Tukey 提出，建立在学生化极差统计量的基础上。当各组样本含量相等时，按下式计算 HSD 值：

$$\mathrm{HSD}=q_{\alpha(k,\nu)}\sqrt{\frac{MS_{\text{误差}}}{n}} \tag{8-8}$$

其中 $q_{\alpha(k,\nu)}$ 为临界值，可由 Student-Newman-Keuls 检验用的临界值表查得，n 为每组的样本含量。当各组样本含量不等时，该法的计算公式为：

$$\mathrm{HSD}=q_{\alpha(k,\nu)}\sqrt{\frac{MS_{\text{误差}}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{8-9}$$

如果两组平均数的差值 $|\overline{X}_i-\overline{X}_j|\geqslant\mathrm{HSD}$，则 $P\leqslant\alpha$，称被比较的两个总体均数之间差别有统计学意义。

对 Tukey-Kramer 法控制试验误差率没有一般的证明，但 Dunnett(1980)用 Monte Carlo 法研究此法非常好。该法的检验功效高于 Bonferroni 法、Sidak 法或 Scheffé 法，但比 Duncan 法或 Student-Newman-Keuls 法低。

6. GT2 法　GT2 法由 Hochberg(1974)提出，类似于 Tukey 法，它用学生化最大模代替学生化极差，并运用 Sidak 的未校正的 t 不等式。如果 $|t|\geqslant m_{\alpha(c,\nu)}$，则有 $P\leqslant\alpha$，两个均数的差别有统计学意义。$m_{\alpha(c,\nu)}$ 是具有自由度为 ν 的 c 个独立正态随机变量的学生化最大模分布的 α 水平对应的临界值，$c=k(k-1)/2$。

一般认为，该法的检验功效低于 Tukey-Kramer 法。

7. Gabriel 法　该法由 Gabriel(1978)提出，用于样本含量不等的情形。此法建立在学生化最大模的基础之上，检验统计量为：

$$t=\frac{\bar{X}_i-\bar{X}_j}{\sqrt{MS_{误差}}\left(\frac{1}{\sqrt{2n_i}}+\frac{1}{\sqrt{2n_j}}\right)} \tag{8-10}$$

当 $|t|\geqslant m_{\alpha(k,\nu)}$ 时，则有 $P\leqslant\alpha$，两个均数的差别有统计学意义。这里临界值 $m_{\alpha(k,\nu)}$ 中用 k 取代 GT2 法中的 c。

样本含量相等时，Gabriel 法与 GT2 法是等价的；样本含量不等时，Gabriel 法比 GT2 法具有更高的检验功效，但当样本含量相差悬殊时，此法可能变得不精确。

8. 多级检验(Multiple-Stage Test)　多级检验方法可以分为步长增加法和步长减少法，其中步长减少法应用较为广泛，SAS/STAT 中采用的也是此法，这里主要就步长减少法的基本步骤及各种方法作一简要介绍。

沿用前述中的假设，有 k 个均数需要进行比较，则步长减少法的具体步骤为：

首先将 k 个样本均数 $\bar{X}_1$，$\bar{X}_2\cdots\bar{X}_k$，按照由小到大的顺序重新排列为 $\bar{X}_{[1]}\leqslant\bar{X}_{[2]}\leqslant\cdots\leqslant\bar{X}_{[k]}$，其中带有括号的下标[1]，[2]…[k]表示重新排序后的顺序标号。

第二步是比较取值最小的均数 $\bar{X}_{[1]}$ 与最大的均数 $\bar{X}_{[k]}$，此时是跨度(即处理组数)为 k 的两个总体均数之间的比较。若两者之间差别无统计学意义，则意味着其他任何两个总体均数之间的差别也都无统计学意义，应该停止一切比较；反之，则继续进行下面的步骤。

第三步，分别比较 $\bar{X}_{[1]}$ 与 $\bar{X}_{[k-1]}$，$\bar{X}_{[2]}$ 与 $\bar{X}_{[k]}$，此时是跨度为 $k-1$ 的两个总体均数之间的比较。沿用第二步中的思路，一直进行下去，如果每一步都有不满足停止比较的对比组，最后应达到跨度为 2 的所有需要比较的相邻两总体均数间都做完比较时为止。

多级检验法在做每一级比较时，通过控制 $\gamma_p(p=k,k-1\cdots2)$ 的水平来实现其最终要控制的某种误差率。γ_p 在特定的分布中所起的作用相当于 t 分布和 F 分布中概率 α，即 γ_p 也是一种显著性水平，它与对比的两个均数之间的跨度(即处理组数 p)有关。

多级检验法中最著名的是 Duncan 法和 SNK 法，由于它们都使用学生化极差统计量，因此也常常被称为多重极差检验法。

(1)Duncan 法：虽然 Duncan 法是目前使用的多级检验法中历史较为久远的一种，但是它也被称为新多级极差检验。它所控制的是比较误差率，也就是每做一次比较犯 I 类错误的概率，其

$$\gamma_p=1-(1-\alpha)^{p-1} \tag{8-11}$$

当各组样本含量相等时，按下式计算检验统计量：

$$q=\frac{\overline{X}_i-\overline{X}_j}{\sqrt{\frac{MS_{误差}}{n}}} \tag{8-12}$$

其中 n 为每组的样本数。当各组样本含量不等时，用作比较的两组样本例数的调和平均数 n_h 代替 n，与式(5-9)类似：

$$n_h=\frac{2}{\frac{1}{n_i}+\frac{1}{n_j}} \tag{8-13}$$

q 统计量称为学生化极差统计量，与式(5-4)中统计量 t 的关系为 $q=\sqrt{2}\,t$。将统计量与临界值比较，如果 $q\geqslant q_{\gamma_p(p,\nu)}$，则有 $P\leqslant\gamma_p$，两总体均数之间的差别有统计学意义。$q_{\gamma_p(p,\nu)}$ 为自由度为 ν 时学生化极差统计量的上 α 百分位数，p 为所比较的两个均数包含的处理组数。$q_{\gamma_p(p,\nu)}$ 与 $\sqrt{\frac{MS_{误差}}{n}}$ 的乘积称为最小显著性极差。当样本均数按照由小到大的顺序排列以后，如果其中任意两个均数 $\overline{X}_i$ 和 $\overline{X}_j$ 之间还有 $p-2(p\geqslant 2)$ 个均数，那么这两个均值的差称为 p 级极差。做任何两组均数间的比较时，也可以将各级极差与最小显著性极差比较即可。

Duncan 法功效较高，也就是说，当真正存在差异时，这一方法能把均数间的差异有效检测出来。如果不考虑较高的Ⅰ型误差率，则此法优于 Tukey 法，正因为如此，Duncan 多重极差检验法应用较为普遍。

(2)SNK 法：该法即 Student-Newman-Keuls 法，又称 q 检验。它和 Duncan 法相似，不同之处在于临界值的计算稍有不同，其对应的 $\gamma_p=\alpha$。检验统计量的计算与 Duncan 法完全相同，算得学生化极差统计量后，如果 $q\geqslant q_{\alpha(p,\nu)}$，则 $P\leqslant\alpha$，两总体均数之间的差别有统计学意义。

SNK 法比 Duncan 法保守，它的Ⅰ类错误概率较小。特别是，对所有涉及相同均数个数的检验，比较的Ⅰ类错误概率都是 α。因此，由于 α 一般说来较小，SNK 法的功效通常比 Duncan 检验法为低。

另外两种多级检验法不像 Duncan 法和 SNK 法那样出名，但它们却是到 20 世纪 70 年代为止文献中介绍的最有效的步长减少的多级检验法，它们是 REGWQ 法和 REGWF 法，由 Ryan(1959,1960)、Einot 和 Gabriel(1975)、Welsch(1977)研究出来，其中 γ_p 由下式定义：

$$\gamma_p=\begin{cases}1-(1-\alpha)^{p/k}, if\ p<k-1\\ \alpha, if\ p\geqslant k-1\end{cases} \tag{8-14}$$

(3)REGWQ 法：REGWQ 法检验统计量的计算与式(5-12)相同，各组样本含量不同时也是以两组样本例数的调和平均数 n_h 代替 n，不同之处在于 γ_p 的定义，也就是临界值的计算不同。

(4)REGWF 法：REGWF 法的检验统计量为如下。

$$F=\frac{n_h\left(\sum_{u=i}^{j}\overline{X}_u^2-\left(\sum_{u=i}^{j}\overline{X}_u\right)^2/k\right)}{(p-1)MS_{误差}} \tag{8-15}$$

其中 n_h 为两组样本例数的调和平均数，p 为所比较的两个均数包含的处理组数，求和范围从 $u=i$ 到 $u=j$。当 $F\geqslant F_{\gamma_p(p,\nu)}$ 时，有 $P\leqslant\gamma_p$，两总体均数之间的差别有统计学意义

(三)两两比较的 Bayes 方法

多个平均值之间的两两比较也可以采用 Bayes 方法，也就是 Waller 法，由 Waller 和 Duncan(1969)采用。它不是控制犯 Ⅰ 类错误的概率，而是在附加损失条件下使 Bayes 风险达到最小值。该法的假定条件是：各组总体均数具有未知方差的先验正态分布，均数的方差的对数具有先验均匀分布。检验的原假设和备择假设分别为：

$H_0:\mu_i-\mu_j\leqslant 0$；

$H_1:\mu_i-\mu_j>0$。

对于任一对(i,j)，让 d_0 表示有利于 H_0 的一个决策，用 d_1 表示有利于 H_1 的一个决策，取 $\delta=\mu_i-\mu_j$。与欲比较的均数对子(i,j)的决策相对应的损失函数为：

$$L(d_0\mid\delta)=\begin{cases}0,\ \text{当}\ \delta\leqslant 0\\ \delta,\ \text{当}\ \delta>0\end{cases}\tag{8-16}$$

$$L(d_1\mid\delta)=\begin{cases}K\delta,\ \text{当}\ \delta\leqslant 0\\ 0,\ \text{当}\ \delta>0\end{cases}\tag{8-17}$$

这里 K 代表一个常数，在 SAS 中用户可以指定，注意它不是所比较的均数个数，要与上文中的符号 k 区分开。

总的损失等于各对均数比较的损失之和。对于对子(i,j)，如果满足

$$\bar{X}_i-\bar{X}_j\geqslant t_B\sqrt{\frac{2MS_{误差}}{n}}\tag{8-18}$$

则拒绝 H_0，说明两总体均数之间的差别有统计学意义。这里 t_B 是 Bayesian t 值，它依赖于常数 K、单因素多水平设计定量资料方差分析的检验统计量 F 及其自由度。t_B 的值是 F 的单调减函数，故当 F 统计量增加时，Waller-Duncan 检验，即 Waller 法变得不精确。

(四)析因设计方差分析中的多重比较

在多因素析因设计中，在方差分析的基础上，也可以进一步对各因素不同水平进行两两比较。当交互作用无统计学意义时，可以直接对处理因素各水平的平均值进行比较；当交互作用有统计学意义时，直接对各因素主效应诸均数比较时，可能被交互作用所掩盖，在这种情况下，可将一个因素或多个因素的组合固定在一定水平上，然后对需要比较的那个因素进行均数间的两两比较。本节以两因素析因设计为例，结合 Tukey 法说明具体的比较过程。

设有两个处理因素 A 和 B，水平数分别为 a 及 b，则总的组合数共有 $a\times b$ 种。用 $i(i=1,2\cdots a)$表示因素 A 的水平号，$j(j=1,2\cdots b)$表示因素 B 的水平号；$\bar{X}_i$ 表示因素 A 第 i 个水平的均数，$\bar{X}_j$ 表示因素 B 第 j 个水平的均数。

1. *两因素交互作用无统计学意义时的比较方法*　在交互作用无统计学意义时，首先计算每一因素各水平均数的方差，处理因素 A 第 i 个水平均数的方差为：

$$S^2(\bar{X}_i)=\frac{MS_{误差}}{b^2}\left(\frac{1}{n_{i1}}+\frac{1}{n_{i2}}+\cdots+\frac{1}{n_{ib}}\right)\tag{8-19}$$

处理因素 B 第 j 个水平均数的方差为：

$$S^2(\bar{X}_j)=\frac{MS_{误差}}{a^2}\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}+\cdots+\frac{1}{n_{aj}}\right)\tag{8-20}$$

式(8-19)中 n_{ib} 是因素 A 第 i 个水平与因素 B 的第 b 个水平组合下的样本例数，式(9-20)中 n_{aj} 是因素 B 第 j 个水平与因素 A 第 a 个水平组合下的样本例数

然后计算同一因素在两个不同水平 l,m 下的平均值之差的方差：

$$S^2(\bar{X}_l - \bar{X}_m) = \frac{1}{2}[S^2(\bar{X}_l) + S^2(\bar{X}_m)] \tag{8-21}$$

检验统计量为：

$$t = \frac{\bar{X}_l - \bar{X}_m}{\sqrt{S^2(X_l - X_m)}} \tag{8-22}$$

当 $|t| \geqslant q_{\alpha(k,\nu)}$ 时，则有 $P \leqslant \alpha$，说明同一因素两水平均数之间的差异有统计学意义。当对因素 A 的不同水平的平均值进行比较时，计算临界值时的组数 $k=a$；当对因素 B 的不同水平的平均值进行比较时，计算临界值时的组数 $k=b$。自由度 ν 为两因素析因设计定量资料方差分析中误差的自由度 $\nu_{误差}$。

2. *两因素间交互作用具有统计学意义时两两比较的方法*　当两因素间交互作用具有统计学意义时，需要将其中一个因素固定在一个水平上，然后对另外一个因素的全部水平下的平均值进行两两比较。首先计算两因素各水平相互组合的格子(i,j)中的平均值 $\bar{X}_{ij}$ 的方差：

$$S^2(\bar{X}_{ij}) = \frac{MS_{误差}}{n_{ij}} \tag{8-23}$$

式中 n_{ij} 为因素 A 第 i 个水平与因素 B 第 j 个水平组合下的样本含量

当对因素 B 的两个不同水平 l,m 下的均数进行比较时，固定因素 A 在第 i 个水平上，计算两水平均数之差的方差为：

$$S^2(\bar{X}_{il} - \bar{X}_{im}) = \frac{1}{2}[S^2(\bar{X}_{il}) + S^2(\bar{X}_{im})] \tag{8-24}$$

检验统计量为：

$$t = \frac{\bar{X}_{il} - \bar{X}_{im}}{\sqrt{S^2(X_{il} - X_{im})}} \tag{8-25}$$

当 $|t| \geqslant q_{\alpha(k,\nu)}$ 时，在因素 A 固定在第 i 个水平上时，因素 B 两水平均数之间的差异有统计学意义。计算临界值时的参数 $k=a\times b$，自由度 ν 为两因素析因设计定量资料方差分析中误差的自由度 $\nu_{误差}$。

【例 8-1】　某器械临床试验，研究 X 线机的读片清晰度，使用国内新型生产的两种 X 线机与国外新型生产的一类 X 线机作为试验组(A,B,C)，对照组选择市场已有器械。按完全随机设计的方法将病例随机分为四组，分别接受不同的照射，指标为读片清晰度得分。测量结果见表 8-1。分析不同的 X 线机读片清晰度的差别有无统计学意义，如差别有统计学意义，进一步分析各试验组与对照组平均值之间的差别是否具有统计学意义。

表 8-1　3 种新型 X 线机与对照的读片清晰程度结果

对照组	试验组 A	试验组 B	试验组 C
3	3	5	3
4	2	8	2
4	2	4	3
4	1	5	4

（续　表）

对照组	试验组 A	试验组 B	试验组 C
3	4	5	5
5	3	7	3
6	2	6	6
4	1	7	4
5	2	5	3
4	1	8	3

本例数据经检验符合正态性和方差齐性，首先进行多个平行组设计方差分析，在此基础上进一步对 3 个试验组和对照组进行比较，使用 SAS 中的 GLM 过程，即一般线性模型过程。

本例用 group 表示不同试验组。Means 语句之后的选项指定分别使用 LSD 法、Bonferroni 法、Scheffé 法、Tukey-Kramer 法、SNK 法和 Bayes 方法进行两两比较。

用 GLM 过程进行试验组与对照组均数比较的程序（程序名 CT8-1）如下：

程　序	说　明
DATA CT8-1; DO GROUP=1 TO 4; INPUT X@@; OUTPUT; END; CARDS; 3 3 5 3 4 2 8 2 4 2 4 3 4 1 5 4 3 4 5 5 5 3 7 3 6 2 6 6 4 1 7 4 5 2 5 3 4 1 8 3; RUN; Ods html;	建立数据集
PROC SORT DATA= CT8-1; BY GROUP; RUN;	对数据集排序
PROC UNIVARIATE DATA= CT8-1 NORMAL; VAR X; BY GROUP; RUN;	进行正态性检验
PROC GLM DATA= CT8-1; CLASS GROUP; MODEL X=GROUP; MEANS GROUP/HOVTEST DUNNETT(1) ALPHA=0.01; RUN; Ods html close;	选项 dunnett 规定进行 Dunnett 检验，指定比较的对照组为第 1 组，采用双侧检验 选项 alpha 规定检验的显著性水准为 0.01

主要输出结果及结果解释：

The GLM Procedure

Dependent Variable: *X*

Source	*DF*	Sum of Squares	Mean Square	*F* Value	*Pr* > *F*
Model	3	58.305 555 56	19.435 185 19	15.29	<.000 1
Error	32	40.666 666 67	1.270 833 33		
Corrected Total	35	98.972 222 22			

以上是方差分析的结果，$F=15.29$，$P<0.000\ 1$，说明四组之间结果的差别具有统计学意义

The GLM Procedure

Dunnett's *t* Tests for *x*

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.01
Error Degrees of Freedom	32
Error Mean Square	1.270 833
Critical Value of Dunnett's *t*	3.137 85
Minimum Significant Difference	1.667 5

Comparisons significant at the 0.01 level

are indicated by * * *

GROUP Comparison	Difference Between Means	Simultaneous 99% Confidence Limits		
3-1	1.555 6	−0.112 0	3.223 1	
4-1	−0.555 6	−2.223 1	1.112 0	
2-1	−2.000 0	−3.667 5	−0.332 5	* * *

程序及输出结果的解释与结论：本例用 group 表示处理因素，由 1 到 4 分别代表对照组和 3 个不同的试验器械组。选项 dunnett 指定进行双侧检验，也可以用选项 dunnettl（当试验组均数只会小于对照组均数时用）或 dunnettu（当试验组均数只会大于对照组均数时用）说明进行单侧检验。

Dunnett 检验的输出结果主要包括两部分，第一部分首先说明该检验控制试验误差率，给出一些统计量的值，包括显著性水准 $\alpha=0.01$，误差的自由度 $\nu_{误差}=32$，误差均方 $MS_{误差}=1.27$，检验的临界值以及最小显著性差异值；第二部分为各试验组与对照组比较的情况，用“* * *”号说明所比较的两组之间差异有统计学意义，可以看到，3 个试验组只有第 2 个试验组与对照组之间的差异具有统计学意义，同时也给出两个均数的差值以及差值的 99%置信区间。这里将各差值的绝对值与最小显著性差值进行比较也可以看到，第二个试验组与对照组均数差值的绝对值大于最小显著性差值 1.667 5，差异有统计学意义。此外，根据置信区间也能够得到相同的结论。

结论：3 个试验器械 X 线机，第 2 个与对照组读片清晰度存在差别，其他类与对照器械相

比无差别。

【例 8-2】 使用【例 8-1】 数据，比较 4 组(3 个试验组，1 个对照器械组)平均值之间差异有无统计学意义。

本例数据经检验符合正态性和方差齐性，首先 4 个平行组之间的方差分析，在此基础上对所有处理组进行两两比较，仍然使用 SAS 中的 GLM 过程。

用 GLM 过程进行所有处理组两两比较的程序(程序名 CT8-2)如下：

程　　序	说　　明
DATA CT8-2;	建立数据集
DO GROUP=1 TO 4;	
INPUT X@@;	
OUTPUT; END;	
CARDS;	输入数据
3 3 5 3	
4 2 8 2	
4 2 4 3	
4 1 5 4	
3 4 5 5	
5 3 7 3	
6 2 6 6	
4 1 7 4	
5 2 5 3	
4 1 8 3;	
RUN;	
Ods html;	
PROC SORT DATA= CT8-2;	
BY GROUP;	对数据集排序
RUN;	
PROC UNIVARIATE DATA= CT8-2 NORMAL;	
VAR X;	进行正态性检验
BY GROUP;	
RUN;	
PROC GLM DATA= CT8-2;	
CLASS GROUP;	
MODEL X=GROUP;	
MEANS GROUP/HOVTEST LSD BON SCHEFFE	
TUKEY SNK WALLER;	选项 LSD,BON,SCHEFFE,TUKEY,SNK,
RUN;	WALLER 规定进行相应两两比较
Ods html close;	

主要输出结果为：

The GLM Procedure

Waller-Duncan K-ratio *t* Test for *X*

Note: This test minimizes the Bayes risk under additive loss and certain other assumptions.

Kratio	100
Error Degrees of Freedom	36
Error Mean Square	1.302 778
F Value	19.98
Critical Value of *t*	1.849 42
Minimum Significant Difference	0.944

Means with the same letter

are not significantly different.

Waller Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

以上为 Bayes 方法，即 Waller-Duncan 检验的结果。一开始说明它是在附加损失条件下使 Bayes 风险达到最小值，接着给出 Kratio 值、误差自由度、误差均方、*F* 值、检验的临界值和最小显著性差异，其中 Kratio 值是 Waller-Duncan 检验两类误差的重要比例，Kratio 的合理值为 50，100，500，大致相当于两水平情况下的 α 为 0.1，0.05 和 0.01，默认值是 100，此处是默认值。结果中指出不同组别间用相同的字母标示差异无统计学意义，提供各组的均数，并按照均数由大到小的顺序对各组进行排列，本例第一组和第 4 组（对照组与第 3 试验组）之间的差别无统计学意义，第 1 组、第二组和第 3 组（对照组，第 1 试验组，第 2 试验组）两两之间的差别均有统计学意义。

The GLM Procedure

t* Tests (LSD) for *X

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	1.302 778
Critical Value of *t*	2.028 09
Least Significant Difference	1.035 2

Means with the same letter are not significantly different.

t Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

以上给出的是 LSD 检验的结果，说明它只控制比较误差率，而不是试验误差率。结果部分大体与前面类似，第 1 组和第 4 组（对照组与第 3 试验组）之间的差别无统计学意义，第 1 组、第 2 组和第 3 组（对照组，第 1 试验组，第 2 试验组）两两之间的差别均有统计学意义。

The GLM Procedure

Student-Newman-Keuls Test for *X*

Note: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	1.302 778

Number of Means	2	3	4
Critical Range	1.035 242 7	1.247 682 9	1.374 748 6

Means with the same letter are not significantly different.

SNK Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

第 3 个结果使用的是 SNK 法，它控制的是在完全无效假设下的试验误差率。使用 SNK 法有多个临界值，本例中存在 4 个平行组，所以存在 3 个临界值，结果中列出了包括不同组数时对应的临界值，检验的结论与 Waller-Duncan 检验相同。

The GLM Procedure

Tukey's Studentized Range (HSD) Test for *X*

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type Ⅱ error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	1.302 778
Critical Value of Studentized Range	3.808 80
Minimum Significant Difference	1.374 7

Means with the same letter are not significantly different.

Tukey Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

Bonferroni (Dunn) *t* Tests for *X*

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type Ⅱ error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	1.302 778
Critical Value of t	2.791 97
Minimum Significant Difference	1.425 2

Means with the same letter are not significantly different

Bon Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

以上结果分别列出 Tukey-Kramer 法与 Bonferroni 法的结果，它们都是控制试验误差率，但是与 REGWQ 法相比，它们通常有一个更高的Ⅱ类错误的概率，检验的结论与 Waller-Dun-

can 法及 SNK 法相同。

The GLM Procedure

Scheffe's Test for *X*

Note: This test controls the Type Ⅰ experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	1.302 778
Critical Value of *F*	2.866 27
Minimum Significant Difference	1.496 8

Means with the same letter are not significantly different

Scheffe Grouping	Mean	*N*	GROUP
A	6.000 0	10	3
B	4.200 0	10	1
B			
B	3.600 0	10	4
C	2.100 0	10	2

最后给出的是 Scheffé 检验的结果，此法要求比较严格，临界值普遍较大。在大部分多重比较方法可能会发现有统计学意义的差别的对比组时，Scheffé 法有可能找不出任何两个均数之间差别有统计学意义。在本例中，结果与前述方法一致。

结论：综合上述几种多重检验方法的结果，对照组与第 3 个试验组之间没有差别，第 1 试验组、第 2 试验组与对照组两两之间均存在差别。

二、多个平均秩的两两比较方法

在多个平行组设计中，当资料不满足参数检验的前提条件时，可以使用 Kruskal-Wallis 检验对多个独立样本进行比较，在检验结果为拒绝 H_0，接受 H_1 时，需要进一步做多个平均秩的两两比较。检验的原假设和备择假设分别为：

H_0：第 i 组和第 j 组某指标总体分布相同；

H_1：第 i 组和第 j 组某指标总体分布不同；

多个平均秩两两比较的方法之一是使用平行组设计中两独立样本比较的 Mann-Whitney U 检验，此时需要对检验水准进行调整，如果检验的次数为 c，要保证总的犯Ⅰ类错误的概率不超过 α，则每次进行检验的显著性水平 $\alpha' = \alpha/c$。

另外一种方法是由 Dunn(1964)提出的，其检验统计量为：

$$z = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{8-26}$$

其中 $\bar{R}_i$ 和 $\bar{R}_j$ 分别为第 i 组和第 j 组的平均秩，N 为总例数，n_i 与 n_j 分别为第 i 组和第 j 组的

例数

此处检验水准也需要调整，为了保证总的犯Ⅰ类错误的概率不超过 α，当进行所有组之间的两两比较时，每次检验的显著性水平 $\alpha' = \alpha/[k(k-1)]$，$k$ 为实验因素的水平数；当多个处理组与一个试验组比较时，$\alpha' = \alpha[2(k-1)]$。

如果 $z \geqslant z_{\alpha'}$，则有 $P \leqslant \alpha'$，两组总体平均秩的差异有统计学意义。其中 $z_{\alpha'}$ 为标准正态分布的分位数。也可以计算出 $z_{\alpha'}\sqrt{\frac{N(N+1)}{12}(\frac{1}{n_i}+\frac{1}{n_j})}$ 的值，然后将两组平均秩之差的绝对值与其进行比较得出结论。

在样本量较大时也可以使用临界值 $q_{\alpha(k,\infty)}$，这里自由度恒取 $\nu=\infty$，检验统计量 $q=\sqrt{2}\,z$。

Nemenyi(1963)建议的方法是使用 χ^2 统计量，与式(5-26)中统计量的关系为 $\chi^2=z^2$，比较的临界值为 $\chi^2_{\alpha,\nu}$，自由度 $\nu=k-1$。与前面提到的 q 检验相比，Nemenyi 法要更为保守。

当数据中存在相同秩时，需要对 χ^2 值进行校正，校正系数为：

$$C=1-\sum(t_j^3-t_j)/(N^3-N) \tag{8-27}$$

式中 t_j 为第 j 个相同秩的个数。校正后的计算公式为：

$$\chi^2_C=\chi^2/C \tag{8-28}$$

【例 8-3】 使用【例 8-1】 数据，假定 4 组平行组数据不服从参数检验的前提条件(正态性、等方差)比较 4 组(3 个试验组，1 个对照器械组)平均秩之间差异有无统计学意义。

当数据经检验不满足正态性和等方差，所以首先使用 Kruskal-Wallis 检验，检验 4 组之间的差异有无统计学意义，在此基础上进一步对 4 组平均秩进行两两比较。关于 Kruskal-Wallis 检验的 SAS 程序及输出结果具体可参见其他章节，此处不再赘述，仅给出两两比较的具体内容。

用 GLM 过程进行多个水平两两比较的程序(程序名 CT8-3)如下：

程　序	说　明
DATA CT8-3;	建立数据集
DOgroup=1 to 4;	
INPUT X @@;	
OUTPUT; END;	
CARDS;	输入数据
3 3 5 3	
4 2 8 2	
4 2 4 3	
4 1 5 4	
3 4 5 5	
5 3 7 3	
6 2 6 6	
4 1 7 4	
5 2 5 3	
4 1 8 3;	
RUN;	
Ods html;	
PROC RANK data=CT8-3 OUT=CT8-3a;	使用 rank 过程对原始数据排秩

（续　表）

程　　序	说　　明
VAR X; RANKS R; RUN; PROC GLM data=CT8-3a; CLASS GROUP; MODEL R= GROUP; MEANS GROUP/TUKEY; RUN; Ods html close;	使用 ranks 语句定义排秩后生成的新变量为 r 选项 tukey 规定使用 Tukey 法

主要输出结果及结果解释：

Tukey's Studentized Range (HSD) Test for *R*

Note: This test controls the Type Ⅰ experimentwise error rate, but it generally has a higher Type Ⅱ error rate than REGWQ

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	51.011 11
Critical Value of Studentized Range	3.808 80
Minimum Significant Difference	8.602 4

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	group
A	33.100	10	3
B	23.100	10	1
B			
B	18.000	10	4
C	7.800	10	2

程序及输出结果的解释与结论：由于 SAS 的 Nparlway 过程并不提供多组平均秩之间的两两比较结果，所以先使用 Rank 过程对原始数据排秩，再用 GLM 过程进行多组平均秩的两两比较。

输出结果中，省略了方差分析的其他结果，仅给出使用 Tukey 法进行两两比较的输出内容。首先给出检验水准、误差自由度、误差均方、检验的临界值和最小显著性差异；第二部分给出 3 组的平均秩及比较结果。

结论：对照组与第 3 个试验组之间没有差别，第 1 试验组、第 2 试验组与对照组两两之间均存在差别。

三、定性资料的多重比较

对于结局变量为定型变量的多个平行组设计，两两比较的方法有很多种，这里着重介绍R×2、R×C列联表资料两两比较的一些方法，并给出一些SAS程序供读者参考。

(一)R×2列联表资料的两两比较

R×2列联表资料的两两比较也称率的多重比较，可分为各试验组与对照组的比较和全部组间的两两比较。前者可分析各试验组与对照组之间频数分布的差异有无统计学意义，后者可分析任意两组间频数分布的差异有无统计学意义。若直接采用多次2×2列联表资料的χ^2检验或Fisher精确检验来进行R×2列联表资料的两两比较，将会明显增大犯假阳性错误的概率。因此，统计学界提出了一些校正的方法，目前已有30种之多，下面简要介绍一些常用的两两比较方法。

1. 调整显著性水准

(1)各试验组与对照组之间的比较：一般采用Brunden法，它是由M. N. Brunden于1972年提出的，其显著性水准的计算公式为：

$$\alpha' = \frac{\alpha}{2(k-1)} \tag{8-29}$$

式中α'为调整后的显著性水准，并以它作为评价各试验组与对照组之间的差异是否有统计学意义的共同显著性水准。α为研究者规定的一般显著性水准，常为0.05或0.01。k为样本率的个数，即列联表资料的行数。如对于一个5×2的列联表资料，它有4个试验组和1个对照组，在统计分析后认为5个组间的频数分布不完全相同时，欲进一步研究哪些试验组与对照组的频数分布不同，需进行4次试验组与对照组的比较。所以，$k=5$，而需要比较的次数为$k-1=4$，调整后的显著性水准就变为$\alpha/[2\times(5-1)]$，即$\alpha/8$。需要说明的是，由该方法估计的检验水准α'可以使研究者犯假阳性错误的概率大大降低，但另一方面，该检验水准也比较保守

(2)各组间的两两比较：调整各组间两两比较的显著性水准应用较为广泛的有两种方法，分别是Bonferroni和Sidak方法。

Bonferroni方法是通过将两两比较所获得的P值乘以一定的倍数(两两比较的次数)来实现对两两比较原始P值的校正。设校正后的P值为P'，某一分表假设检验获得的P值仍记为P，两者之间的关系可表示如下：

$$P' = \begin{cases} nP & \text{如果 } nP<1 \\ 1 & \text{如果 } nP\geqslant 1 \end{cases} \tag{8-30}$$

其中n为两两比较的次数，其计算公式为：

$$n = \frac{k(k-1)}{2} \tag{8-31}$$

上面是修正两两比较假设检验所得的P值。当然，更为常用的做法是通过上述方法来修正临界水准。设α'为调整后的显著性水准，并以它作为评价各组间两两比较的共同显著性水准。α为研究者规定的一般显著性水准，则α'与α的关系可表示如下：

$$\alpha' = \frac{\alpha}{\frac{k(k-1)}{2}} = \frac{2\alpha}{k(k-1)} \tag{8-32}$$

Sidak 方法通过另一种方法对两两比较所获得的 P 值进行修正。设校正后的 P 值为 P'，某一分表假设检验获得的 P 值仍记为 P，两者之间的关系可表示如下：

$$P' = 1 - (1 - P)^n \tag{8-33}$$

其中 n 为两两比较的次数，其计算公式同(8-31)

与 Bonferroni 方法类似，这也是修正两两比较假设检验所得的 P 值，更常用的做法则是修正临界水准。设 α' 为调整后的显著性水准，并以它作为评价各组间的两两比较的共同显著性水准。α 为研究者规定的一般显著性水准，则 α' 与 α 的关系可表示如下：

$$\alpha' = 1 - (1 - \alpha)^{1/n} \tag{8-34}$$

Sidak 方法与 Bonferroni 方法类似，区别在于它们修正 P 值或显著性水准 α 的计算方法略有差异。前者通过计算多次两两比较后犯假阳性错误概率的理论值 $1-(1-\alpha')^n$，并使 $1-(1-\alpha')^n$ 控制在原本设想的假阳性错误概率 α 之内来实现的，后者是通过控制多次两两比较的假阳性错误概率之和在原本设想的假阳性错误概率 α 之内来实现的。理论上来讲，前者更为精确，可从理论上进行推导获得，后者更大程度上是一种经验上的简便算法。两者的差异不是很大，读者可根据实际需要选用，若上述两种方法的假设检验结果存在矛盾，建议读者多采用一些两两比较的方法进行检验并结合临床专业情况来决定取何种结果。

2. *基于重复抽样的 Bootstrap 方法或 Permutation 方法*　SAS 的 Multtest 过程(即多重检验过程)，提供了两种方法来进行各试验组与对照组的比较或各组间的两两比较，即 Bootstrap(自助法)和 Permutation(置换法)方法，这两种方法均基于重复抽样。Bootstrap 和 Permutation 分别用有放回和无放回的方法对数据进行重抽样，来逼近所有检验中最小 P 值的分布，然后用该分布对单个原始 P 值进行修正。

【例 8-4】　某临床试验研究使用 3 种理疗仪对治疗膝关节疼痛的治疗效果，同时设立常规器械作为对照组，治疗结果见表 8-2。现已经过统计分析得出各组间改善率不完全相同，需进一步研究哪些试验组与对照组之间改善率的差异存在统计学意义。

表 8-2　3 种试验器械与对照药器械疗效的比较

组　别	例数(人)		
疗效：	改善	无效	合计
对照组	16	14	30
试验器械 1	19	11	30
试验器械 2	13	17	30
试验器械 3	27	3	30
合计	75	45	120

SAS 程序及说明：SAS 程序见 CT8-4。

```
%macro rc(dataset,a);                    /*①*/
proc sql;                                /*②*/
create table pTest1(compare char(16));
create table pTest2(p num);
quit;

%do i=2 %to &a;

proc sql;                                /*③*/
create table a1&i as
select * from &dataset where a=1 or a=&i;
quit;

ods listing close;                       /*④*/
proc freq;
weight f;
tables a*b/exact;output out=p1&i(keep=XP2_FISH)
fisher;
run;
ods listing;

proc sql;                                /*⑤*/
insert into pTest1 values("ROW 1 VS ROW &i");

insert into pTest2 select * from p1&i;
quit;
%end;

data a;                                  /*⑥*/
merge pTest1 pTest2;
run;

data outcome;                            /*⑦*/
set a;
pj=round(p*2*(&a-1),0.001);
if pj>1 then pj=1;
run;

ods html;
proc print;                              /*⑧*/
run;
ods html close;

%mend rc;
data CT8-4;                              /*⑨*/
do a=1 to 4;
do b=1 to 2;
input f @@;
output;
end; end;
cards;
16  14
19  11
13  17
27  3;
run;

%rc(CT8-4,4);                            /*⑩*/
```

程序中第①步为创建两两比较的宏，包含两个宏参数：数据集名称 dataset 和列联表行数 a(即组数)，在调用这个宏时需要给出这两个宏参数的值。第②步为创建两个表，分别包含对比组名称和对比结果 P 值(见第⑤步)，用来帮助输出最后两两比较的结果。第③步为将原 R×C 表分割成由每一个试验组分别与对照组构成的 2×2 表。第④步为对每个分割表进行 Fisher 精确检验，并将相应的 P 值输出。因为此宏对每一个 2×2 列联表都执行相同的统计检验，所以建议读者仅当所有分割出来的 2×2 列联表都满足一般 χ^2 检验的应用条件时，才能用 χ^2 检验替代本步内的 Fisher 精确检验。否则，都采用 Fisher 精确检验即可。此步中的首行"ods listing close;"及尾行"ods listing;"调用了 SAS 输出传送系统(output delivery system)，目的是在结果中不输出对每个分割表的统计分析结果。若读者认为需要每个分割表的统计分析结果，可将此两句同时删除。第⑤步为将对比组名称插入 pTest1 表中，将对比结果 P 值插入 pTest2 表中。此步骤与第二步相呼应。第⑥步为将 pTest1 和 pTest2 两个表合并，使最后输出的统计分析结果简单清晰。第⑦步为采用 Brunden 法对每一个分割表处理所获得的 Fisher 精确检验的 P 值进行修正。第⑧步为将两两比较的结果输出，同时调用 SAS 输出传送系统将结果输出为 html 格式。第⑨步为创建数据集。第⑩步为宏 rc 输入两个宏参数的值，数据集名称为 CT8-4，列联表行数为 4，即调用宏。

输出结果：

Obs	compare	*p*	*pj*
1	ROW 1 VS ROW 2	0.600 95	1.000
2	ROW 1 VS ROW 3	0.605 81	1.000
3	ROW 1 VS ROW 4	0.003 42	0.021

输出结果的解释：以上是 4×2 列联表各试验组与对照组两两比较的结果，第 2 列“compare”给出了进行比较的组别，第 3 列“*P*”为对分割表进行 Fisher 精确检验的双侧尾端概率值，第 4 列“*PJ*”是采用 Brunden 法校正后的 P 值，读者在下统计结论时应以此列为准。从上面的结果中可以看出，若以 0.05 水平为基准，第 4 组（第 3 试验组）与第 1 组（对照组）之间改善率的差别具有统计学意义。

R×2 列联表资料中，各试验组与对照组之间的两两比较也可以通过 SAS 的 Multtest 过程来实现，SAS 程序及结果如下。

SAS 程序及说明：SAS 程序见 CT8-5。

```
data CT8-5;                                    /*①*/
do j=1 to 4;
do i=0 to 1;
input f@@;
output; end; end;
cards;
16  14
19  11
13  17
27  3;
run;
proc multtest data=CT8-5 order=data out=p
permutation nsample=1000 seed=999999;          /*②*/
test fisher(i);
class j;
freq f;
contrast '12' -1 0 0 ;
contrast '13' 0 -1 0 ;
contrast '14' 0 0 -1 ;
run;

ods html;
proc print data=p;                             /*③*/
run;
ods html close;
```

第①步为创建数据集，需要注意的是结果变量的取值必须为 0 或 1，这是接下来要调用的 Multtest 过程规定的，但其赋值顺序（既 0,1 或 1,0）不影响最终的结果。第②步调用 Multtest 过程，采用 permutation 方法（或采用 bootstrap 方法，两者的区别在于重抽样是否为有放回抽样）来修正 *P* 值，需要设定样本量 nsample 和随机种子数 seed，nsample 表示采用 permutation 或 bootstrap 时所使用的重抽样次数，缺省值为 100，seed 为规定一个正整数作为重抽样的随机数发生器的初始种子，缺省值为计算机时钟值。Test 语句后规定检验统计量，做 R×2 列联表的多重比较在这里只能选用 Fisher 精确检验，因为 Multtest 过程不提供 χ^2 检验。Fisher(*)中，“*”代表结果变量，class 后接原因变量。contrast 语句后规定需要进行比较的组，只能为－1,0,1 3 个取值，其中取值为 1 的组被合并，取值为－1 的组被合并，两合并组进行比较，取值为 0 的组不参与比较。本例中，统计分析的目的是进行两两比较，所以每次只有一个取值为 1 的组和一个取值为－1 的组之间的比较，不会合并组然后进行比较。第③步将

两两比较的结果输出。

注意：对于这种一个对照组与多个试验组进行两两比较的情况，采用 Multtest 过程，可不必写出 contrast 语句，SAS 自动生成所有试验组与对照组的对比，但必须在 SAS 数据步中将对照组的数据首先录入，因为 SAS 默认 class 后的变量的第一水平为对照组。此外，对于所有组间的两两比较则不能缺省 contrast 语句，必须将需要对比的所有组一一列出。

输出结果：

Obs	_test_	_var_	_contrast_	_xval_	_mval_	_yval_	_nval_	raw_P	perm_P	sim_se
1	FISHER	i	12	14	30	11	30	0.600 95	0.836	0.011 709
2	FISHER	i	13	14	30	17	30	0.605 81	0.836	0.011 709
3	FISHER	i	14	14	30	3	30	0.003 42	0.006	0.002 442

输出结果的解释：以上是采用 SAS 的 Multtest 过程获得的 4×2 列联表各试验组与对照组两两比较的结果。其中，第 9 列“raw_P”是对分割表进行 Fisher 精确检验的双侧尾端概率值，第 10 列“perm_P”是采用 permutation 方法对数据进行重抽样后估计的 P 值，读者在下统计结论时应以此列为准。从上面的结果中可以看出，若以 0.05 水平为基准，则第一组与第四组（即对照组与试验第 3 组）之间改善率的差异有统计学意义。

【例 8-5】 使用【例 8-4】 的数据，现已经过统计分析得出各组间改善率不完全相同，对四组（3 个试验组，1 个对照）之间进行两两比较，分析每两组改善率的差异是否存在统计学意义。

SAS 程序及说明：SAS 程序见 CT8-6。

```
%macro rc(dataset,a);                                    /*①*/
proc sql;                                                /*②*/
create table pTest1(compare
char(16));
create table pTest2(p num);
quit;

%do i=1 %to &a-1;
%do j=&i+1 %to &a;

proc sql;                                                /*③*/
create table a&i&j as
select * from &dataset where a=&i or
a=&j;
quit;

ods listing close;                                       /*④*/
```

```
data outcome;                                            /*⑦*/
set a;
pj=round(p*&a*(&a-1)/2,0.001);
if pj>1 then pj=1;
run;

ods html;
proc print;                                              /*⑧*/
run;
ods html close;
%mend rc;
data CT8-6;                                              /*⑨*/
do a=1 to 4;
do b=1 to 2;

input f @@;
output;
```

（续　表）

```
proc freq;
weight f;
tables a * b/exact;output
out=p&i&j(keep=XP2_FISH) fisher;
run;
ods listing;

proc sql;                                    /*⑤*/
insert into pTest1 values("ROW &i VS ROW &j");
insert into pTest2 select * from
p&i&j;
quit;
%end;
%end;

data a;                                      /*⑥*/
merge pTest1 pTest2;
run;
end; end;
cards;
16  14
19  11
13  17
27  3;
run;

%rc(CT8-6,4);                                /*⑩*/
```

此程序与例 8-4 中第一个 SAS 程序基本相同，区别只有两处。第一处为第②步与第③步之间的 do 循环语句，例 17-2 第一个 SAS 程序中此位置的 do 循环语句是为了将 R×2 列联表分割成由每一个试验组分别与对照组组合而成的多个 2×2 列联表，而本例的 do 循环语句是为了将 R×2 列联表分割成任意两个组都分别进行组合而成的多个 2×2 列联表。另外第⑦步中修正 P 值的方法也不相同，例 8-4 采用 Brunden 法修正 P 值，而本例采用 Bonferroni 法修正 P 值。其他步骤的 SAS 程序实现目的与例 8-4 中第一个 SAS 程序相同。

输出结果：

Obs	compare	P	Pj
1	ROW 1 VS ROW 2	0.600 95	1.000
2	ROW 1 VS ROW 3	0.605 81	1.000
3	ROW 1 VS ROW 4	0.003 42	0.021
4	ROW 2 VS ROW 3	0.195 36	1.000
5	ROW 2 VS ROW 4	0.030 34	0.182
6	ROW 3 VS ROW 4	0.000 25	0.002

输出结果的解释：以上是 4×2 列联表各组两两比较的结果，第 2 列“compare”给出了进行比较的组别，第 3 列“P”为对分割表进行 Fisher 精确检验的双侧尾端概率值，第 4 列“Pj”是采用 Bonferroni 法校正后的 P 值，读者在下统计结论时应以此列为准。从上面的结果中可以看出，第 1 组与第 4 组（对照组与试验 3 组）、第 3 组与第 4 组（试验 2 组与试验 3 组）改善率差别具有统计学意义。

R×2 列联表资料中，各组之间的两两比较也可以通过 SAS 的 Multtest 过程来实现，SAS 程序及结果如下。

SAS 程序及说明：SAS 程序见 CT8-7。

```
data CT8-7;                                          /*①*/
do j=1 to 4;
do i=0 to 1;
input F@@;
output;
end;end;
cards;
16  14
19  11
13  17
27  3;
run;
proc multtest data= CT8-7
order=data out=p permutation nsample=1000
seed=764511;                                         /*②*/
test fisher(i);
class j;
freq F;
contrast '12' 1 -1 0 0 ;
  contrast '13' 1 0 -1 0 ;
  contrast '14' 1 0 0 -1 ;
contrast '23' 0 1 -1 0 ;
  contrast '24' 0 1 0 -1 ;
  contrast '34' 0 0 1 -1 ;
run;

ods html;
proc print data=p;                                   /*③*/
run;
ods html close;
```

输出结果：

Obs	_test_	_var_	_contrast_	_xval_	_mval_	_yval_	_nval_	raw_P	perm_P	sim_se
1	FISHER	i	12	14	30	11	30	0.600 95	0.932	0.007 961
2	FISHER	i	13	14	30	17	30	0.605 81	0.932	0.007 961
3	FISHER	i	14	14	30	3	30	0.003 42	0.008	0.002 817
4	FISHER	i	23	11	30	17	30	0.195 36	0.451	0.015 735
5	FISHER	i	24	11	30	3	30	0.030 34	0.083	0.008 724
6	FISHER	i	34	17	30	3	30	0.000 25	0.000	0.000 000

输出结果的解释：以上是采用 SAS 的 Multtest 过程获得的 9×2 列联表各组两两比较的结果。其中，第 8 列“raw_P”是对分割表进行 Fisher 精确检验的双侧尾端概率值，第 9 列“perm_P”是采用 Permutation 方法对数据进行重抽样后估计的 P 值，下统计结论时应以此列为准。第 1 组与第 4 组（对照组与试验 3 组）、第 3 组与第 4 组（试验 2 组与试验 3 组）改善率差别具有统计学意义。

（二）R×C 列联表资料的两两比较

与 R×2 列联表两两比较不同，R×C 列联表两两比较目前研究并不多。我们以 Bonferroni 法为基础对 R×C 列联表两两比较所得的 P 值进行校正，仅供读者参考。

【例 8-6】 某临床试验比较两种新器械与一种常规器械（A，B，C）对于关节痛的治疗效果。将 162 例关节痛患者分为 3 组，A 组 56 使用常规器械；B 组 43 例使用试验器械 B；C 组 63 例使用试验器械 C。治疗效果见表 8-3。对 3 组优劣进行两两比较。

表 8-3　3 组器械治疗效果比较

分组	例数				
	治疗效果：	治愈	显著改善	改善	无效	合计
A 组		15	19	19	3	56
B 组		7	10	18	8	43
C 组		24	21	11	7	63
合计		46	50	48	18	162

SAS 程序及说明：SAS 程序见 CT8-8

```
%macro rc(dataset,a);                    /*①  */
proc sql;                                /*②  */
create table pTest1(compare
char(16));
create table pTest2(p num);
quit;

%do i=1 %to &a-1;
%do j=&i+1 %to &a;

proc sql;                                /*③  */
create table a&i&j as
select * from &dataset where a=&i or
a=&j;
quit;

ods listing close;                       /*④  */
proc nparlway wilcoxon;
class a;
var b;
output out=p&i&j(keep=P2_WIL);
run;
ods listing;

proc sql;                                /*⑤  */
insert into pTest1 values("ROW &i VS ROW
&j");
insert into pTest2 select * from
P&i&j;
quit;
%end;
%end;
```

```
data a;                                  /*⑥  */
merge pTest1 pTest2;
run;

data outcome;                            /*⑦  */
set a;
pj=round(p*&a*(&a-1)/2,0.001);
if pj>1 then pj=1;
run;

ods html;
proc print;                              /*⑧  */
run;
ods html close;
%mend rc;

data CT8-8;                              /*⑨  */
do a=1 to 3;
do b=1 to 4;
input f @@;
do k=1 to f;
output;
end; end; end;
cards;
15 19 19 3
7 10 18 8
24 21 11  7
;
run;

%rc(CT8-8,3);                            /*⑩  */
```

本资料为结果变量为有序变量的单向有序的 R×C 表资料，分割而成的多个 2×4 列联表仍为结果变量为有序变量的二维列联表资料。研究 3 组差异是否有统计学意义，需采用 Wilcoxon 秩和检验。故在第④步中对程序略作调整，采用 Wilcoxon 秩和检验。

输出结果：

Obs	compare	P	Pj
1	ROW 1 VSROW 2	0.020 98	0.063
2	ROW 1 VSROW 3	0.243 51	0.731
3	ROW 2 VSROW 3	0.002 16	0.006

输出结果的解释：上面是对原 3×4 列联表资料进行两两比较的结果。根据校正后的 P 值(即 Pj)可知，若以 0.05 水平为基准，第 2 组与第 3 组之间疗效上的差异有统计学意义。

（郭　晋　李　卫　姜宇莉　唐欣然　黄耀华）

第 9 章 非诊断试验交叉设计统计分析

在医疗器械临床试验中有时会用到交叉试验(cross-over trial)设计，相关统计分析方法也与其他设计略有差别。由于定量资料和定性资料的统计分析方法不同，本章主要介绍定量资料的分析方法及相关软件实现。

一、交叉试验设计概述

交叉试验是按事先设计好的处理顺序(sequence)，在试验对象上按各个时期(period)逐一依次实施各项处理(treatment)，以比较这些处理的作用。这是一种自身比较的试验方法。在每一个试验对象上所经历的试验过程大致为：设试验对象按时期依次安排使用不同仪器(图 9-1，图中A，B)。

图 9-1 交叉试验过程

时期 1. 准备阶段(run in)：指试验对象经过一段时间不加任何处理的观察，确认已进入自然状态，可以进入试验；时期 2. 处理：按事先安排好的次序在各个相应时期中仅使用一种器械来治疗；3. 洗脱期(wash out)：指不加任何处理的观察，确认前一器械的治疗效果已经消失，试验对象又回到了自然状态以保证后一时期的治疗效果不受前一时期器械治疗的影响

(一)交叉设计的优点和缺点

1. 优点

(1)交叉设计是一种自身比较的方法：该方法是将个体的差异从处理比较中分离出来，能同时研究时期效应、治疗效应、延迟效应，当有基础数据时，还可以扣除其影响作协方差分析，所以效率较高，临床试验工作者乐于采用。

(2)每个试验对象安排了多个时期，可实施多个治疗，因此降低了试验的样本量。

2. 缺点

(1)需在同一试验对象上作多种处理，因此每个处理时间不能过长。若过长，则造成试验周期太长，何况还要增加一定的洗脱期，由于周期过长，可能造成无法坚持到底而中断了试验。

(2)安排洗脱期：为了清除前一时期的处理作用，必须安排洗脱期，否则，无法排除上一治疗的延迟效应。

(3)变化：当试验过程中试验对象的状态发生了根本的变化，如治愈、死亡等，后一时期的处理无法施加。

(4)退出试验:试验对象一旦在某一时期退出试验,就造成了数据的缺失,增加了统计分析之困难。

(二)交叉试验统计分析的一般内容

1. 不考虑时期作用时,检验处理作用。

2. 扣除时期作用后,检验处理作用。

3. 检验时期作用。

4. 检验延滞作用(carry over)。所谓的延滞作用的检验是指检验前一时期的处理是否会延续到后一个时期的处理上,尽管试验中已经安排了洗脱期。

5. 扣除协变量影响后的检验。

二、2×2 交叉试验计量资料的统计分析

2×2 交叉试验指的是两个时期,两个处理的交叉试验设计,在临床仪器试验常用。当交叉试验观察结果为计量资料时,其分析方法一般有参数统计分析方法及非参数统计分析方法,前者需假定数据呈正态分布。对于交叉试验计量资料,统计学家更愿意选择正态分布的假设,其理由为:①许多临床试验中资料常是计量连续的,其中一部分为正态分布,另有一些通过变量置换也可能使其接近正态分布;②基于正态分布的统计方法,常是稳健的(robustness),适用范围广,选择余地大;③使用正态分布方法作统计分析时,可达到一定的精度;④基于正态分布的假设,结果的解释比较容易;⑤虽然有些资料,如生存资料在临床试验中占一定的地位,它是非正态的,但由于在交叉试验中极少遇到,这反过来提高了正态分布在交叉试验中的用途。

例如:将 A、B 两种仪器先后施于同一批试验对象,随机地使半数对象先接受 A,后接受 B;另一半对象先接受 B,后接受 A;两种仪器在全部试验过程中"交叉"进行称为交叉试验。由于 A 和 B 处于先后两个试验阶段的机会是相等的,因此平衡了试验顺序的影响,而且能把处理方法之间的差别与时间先后之间的差别分开来分析(表 9-1)。

表 9-1　2×2 交叉试验设计试验结果的数值符号

次序	样本量	和值		差值	
		$\bar{T}_i$	S_{T_i}	$\bar{D}_i$	S_{D_i}
AB	n_1	$\bar{T}_1$	S_{T1}	$\bar{D}_1$	S_{D1}
BA	n_2	$\bar{T}_2$	S_{T2}	$\bar{D}_2$	S_{D2}

$\bar{T}_i$ 为和的均值;S_{T_i} 为和的标准差;$\bar{T}_1$、$\bar{T}_2$、S_{T_1}、S_{T_2} 表示次序 AB,BA 两组试验对象两阶段观察值和的平均值和标准差。$\bar{D}_i$ 为差的均值;S_{D_i} 为差的标准差;$\bar{D}_1$、$\bar{D}_2$、S_{D_1}、S_{D_2} 示次序 AB、BA 两组试验对象两阶段实验观察值之差的差数平均数及标准差

1. 处理效应

(1)处理效应的差别检验:假定处理无剩余效应,要判断两种处理的直接效应有无差别,即是否 $\tau_1=\tau_2$(τ 表示总体均数),可用 t 检验:

$$\tau=\frac{|\bar{D}_1-\bar{D}_2|}{S_D}\sqrt{\frac{n_1n_2}{n_1+n_2}}, \qquad (9\text{-}1)$$

$$\upsilon=n_2+n_2-2,$$

$$S_D{}^2=\frac{(n_1-1)S_{D_1}{}^2+(n_2-1)5_{D_2}{}^2}{n_1+n_2-2}$$

SAS 程序-1

```
data crossover_2by2_1;
input sq$ p1 p2 @@;
d=p1-p2;
output;
cards;
AB 数据
BA 数据;
proc ttest ci=equal;
var d;
class sq;
run;
quit;
```

(2)处理效应差值的可信区间：$(\tau_1-\tau_2)$的 $100(1-\alpha)\%$可信区间可用下式计算：

$$\frac{\overline{D}_1-\overline{D}_2}{2}\pm t_{\alpha(n_1+n_2-2)}\times\frac{S_D}{\sqrt{\frac{1}{n_1}+\frac{2}{n_2}}}$$

2. *时期效应*　各种处理随着不同的阶段其效应可能有所不同，这种因研究阶段不同而产生的差异为时期效应。检验两种处理时期效应有无不同，可用如下 t 检验：

$$t=\frac{|\overline{D}_1+\overline{D}_2|}{S_D}\sqrt{\frac{n_1n_2}{n_1+n_2}} \tag{9-2}$$

$$\upsilon=n_1+n_2-2,$$

$$S_D{}^2=\frac{(n_1-n_2)S_{D_1}{}^2+(n_1-1)S_{D_2}{}^2}{n_1+n_2-2};$$

SAS 程序-2

```
data crossover_2by2_2;
input sq$ p1 p2 @@;
d=p1-P2;
output;
cards;
AB 数据
BA 数据;
data crossover_temp;
set crossover_2by2_2;
if sq='ab' then cd=d;
else cd=-d;
run;
proc ttest alpha=0.05 ci=equal;
var cd;
class sq;
run;
```

3. *剩余效应*　剩余效应又称延续效应(carry-over effect)，指交叉设计中前一阶段仪器作用的效果至后一阶段时仍然存在的作用。一项符合要求的交叉设计试验各处理应当无剩余效应或剩余效应相同。判断两种处理剩余效应有无不同，可用下式检验：

$$t=\frac{|\overline{T}_1-\overline{T}_2|}{S_T}\sqrt{\frac{n_1n_2}{n_1+n_2}}, \tag{9-3}$$

$$\upsilon=n_2+n_2-2;$$

$$S_T^2=\frac{(n_1-2)S_{T_1}^2+(n_2-1)S_{T_2}^2}{n_1+n_2-2};$$

SAS 程序-3	
data crossover_2by2_3; input sq $ p1 p2 @@; T=p1+p2; output; cards; AB 数据 BA 数据;	proc ttest alpha=0.05 ci=equal; var T; class sq; run;

对于剩余效应的检验，有学者建议取显著水准 $\alpha=0.10$ 或以上，以便较敏感地发现剩余效应不等的情形。当两种仪器作用结果剩余效应无显著性差别时，可用上述方法；当对时期效应、剩余效应的检验均无差别时，两处理的效应比较还可以进一步简化为配对 t 检验。注意，剩余效应偶显著性差异时，只能用第一阶段的结果进行比较，第二阶段的结果将归于无用。

【例 9-1】 为检查 A、B 两台代谢测定仪器测定耗氧量得结果是否相同，以条件近似的 14 名健康人测试，将 14 名健康人随机分成两组，一组人先用 A 仪器测试，再用 B 仪器测试，次序为 AB；另一组则相反，次序为 BA。现将两组健康人测试结果列在表 9-2。

表 9-2 14 名健康人先后用两台测定仪器测定的耗氧量

受试者	阶段		和值	差值
按 AB 次序	Ⅰ	Ⅱ		
	1 237	1 256	2 493	−19
	1 179	1 275	2 454	−96
	1 000	981	1 981	19
	1 295	1 387	268 2	−92
	1 218	1 187	2 405	31
	1 138	1 175	2 313	−37
	971	1 012	1 983	−41
按 BA 次序				
	1 387	1 348	2 735	39
	1 025	1 022	2 047	3
	1 225	1 226	2 451	−1
	1 050	1 026	2 076	24
	1 050	1 031	2 081	19
	1 387	1 298	2 685	89
	1 150	1 108	2 258	42

SAS 计算程序：

数据录入 sq，$p1$，$p2$ 分别为次序、时期 1，时期 2，过程步里面首先使用 t 检验分别计算处理效应、时期效应和剩余效应，然后根据效应结果判断最后的结果显示。

```
data crossover_2by2;
input sq $ p1 p2 @@;
T=p1+p2;
D=p1-p2;
output;
cards;
ab 1237 1256
ab 1179 1275
ab 1000 981
ab 1295 1387
ab 1218 1187
ab 1138 1175
ab 971 1012
ba 1387 1348
ba 1025 1022
ba 1225 1226
ba 1050 1026
ba 1050 1031
ba 1387 1298
ba 1150 1108;
data crossover_temp;
set crossover_2by2;
if sq='ab' then cd=d;
else cd=-d;
run;
%macro tjy(datain,varx,group);/* t 检验宏命令
*/
proc ttest data=&datain ci=equal alpha=0.05 ;
class &group;
var &varx;
run;
%mend tjy;
%macro result(a);/* 判断选择何种效应,并显示
结果 */
data temp;
set dataout3;/* 检查剩余效应 */
%if Probt>&a %then  %do;
data temp;
set dataout2;/* 检查时期效应 */
%if probt>&a %then %do;
title1 '剩余效应,时期效应均无显著性差别';
title2 '最后显示结果采用"配对 t 检验"';
data m1; /* 调整数据集格式 */
set crossover_temp;
if sq='ab' then d1=d;
else delete;
data m2;
set crossover_temp;
if sq='ba' then d2=d;else delete;
data mm;
merge  m1 m2;
keep d1 d2;
run;
proc ttest data=mm; /* 进行配对 t 检验 */
paired d1*d2;
run;
%end;
%else  %do;
title1 '剩余效应无显著性差别';
title2 '时期效应';
%tjy(crossover_temp,cd,sq);
title2 '处理效应';
%tjy(crossover_temp,d,sq);
%end;  %end;
%else  %do;
title1 '剩余效应有显著性差别';
title2 '使用第一阶段的结果进行检验';
%tjy(crossover_temp,p1,sq);
%end;
%mend;
ods output ttests=dataout1 ;/* 将处理效应检验结
果输出为数据集 */
title '处理效应';
%tjy(crossover_temp,d,sq);
ods output close;
ods output ttests=dataout2;/* 将时期效应检验结
果输出为数据集 */
title '时期效应';
%tjy(crossover_temp,cd,sq);
ods output close;
ods output ttests=dataout3;/* 将剩余效应检验结
果输出为数据集 */
title '剩余效应';
%tjy(crossover_temp,T,sq);
ods output close;
%result(0.1);
Run;
Quit;
```

SAS 分析结果：

处理效应

The TTEST Procedure

Statistics

Variable	sq	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CLStd Dev	Std Err	Minimum	Maximum
D	ab	7	−79.07	−33.57	11.931	31.704	49.2	108.34	18.596	−96	31
D	ba	7	2.587 6	30.714	58.841	19.597	30.412	66.97	11.495	−1	89
D	Diff (1-2)		−111.9	−64.29	−16.65	29.328	40.899	67.514	21.862		

T-Tests

Variable	Method	Variances	*DF*	*t* Value	*Pr* > |*t*|
D	Pooled	Equal	12	−2.94	0.012 4
D	Satterthwaite	Unequal	10	−2.94	0.014 8

Equality of Variances

Variable	Method	Num *DF*	Den *DF*	*F* Value	*Pr* > *F*
D	Folded F	6	6	2.62	0.266 7

时期效应

The TTEST Procedure

Statistics

Variable	sq	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Err	Std Err	Minimum	Maximum
cd	ab	7	−79.07	−33.57	11.931	31.704	49.2	108.34	18.596	−96	31
cd	ba	7	−58.84	−30.71	−2.588	19.597	30.412	66.97	11.495	−89	1
cd	Diff (1−2)		−50.49	−2.857	44.775	29.328	40.899	67.514	21.862		

T-Tests

Variable	Method	Variances	*DF*	*t* Value	*Pr* > |*t*|
cd	Pooled	Equal	12	−0.13	0.898 2
cd	Satterthwaite	Unequal	10	−0.13	0.898 6

Equality of Variances

Variable	Method	Num *DF*	Den *DF*	*F* Value	*Pr* > F
cd	Folded F	6	6	2.62	0.266 7

剩余效应

The TTEST Procedure

Statistics

Variable	sq	N	Lower		Upper	Lower		Upper		Minimum	Maximum
			CL Mean	Mean	CL Mean	CL Std Dev	Std Dev	CL Std Dev	Std Err		
T	ab	7	2 087.2	2 330.1	2 573.1	169.26	262.67	578.42	99.28	198 1	2 682
T	ba	7	2 062	2 333.3	2 604.6	189.05	293.37	646.03	110.88	2 04 7	2 735
T	Diff (1-2)		−327.4	−3.143	321.14	199.67	278.44	459.64	148.84		

T-Tests

Variable	Method	Variances	DF	*t* Value	$Pr > \|t\|$
T	Pooled	Equal	12	−0.02	0.983 5
T	Satterthwaite	Unequal	11.9	−0.02	0.983 5

Equality of Variances

Variable	Method	Num DF	Den DF	*F* Value	$Pr > F$
T	Folded F	6	6	1.25	0.795 2

以上的部分为处理效应、时期效应、剩余效应的结果；其中第一部分为 *t* 检验统计描述的部分包括“Mean”平均值，“Std Dev”标准差，“Std Err”标准误等信息；第二部分“T-Tests”为 *t* 检验结果中，其中包括方差齐同时使用 Pooled 方法，方差不齐时使用 Satterthwaite 方法；第三部分为“equality of variances”方差齐性检验。

剩余效应、时期效应均无显著性差别

最后显示结果采用“配对 *t* 检验”

The TTEST Procedure

Statistics

Difference	*N*	Lower		Upper	Lower		Upper		Minimum	Maximum
		CL Mean	Mean	CL Mean	CL Std Dev	Std Dev	CLStd Dev	Std Err		
d1-d2	7	−119	−64.29	−9.569	38.124	59.163	130.28	22.361	−126	20

T-Tests

Difference	*DF*	*t* Value	$Pr > \|t\|$
d1-d2	6	−2.87	0.028 2

最后一部分为程序总体判断的结果，本次检验剩余效应和时期效应的检验结果均无显著性差异，因此采用配对 *t* 检验。结果中的第一部分为配对差值的统计学描述信息，包括“Mean”均数、“Std Dev”标准差、“Std Err”标准误；第二部分为配对 *t* 检验的结果“*t* Value”。

采用方差分析方法对例题进行分析。

SAS 程序如下：

sq $，pa，pd，tr $，oc 分别录入次序、病人号、时期、处理、反应值，过程步为方差分析。

```
data co_2_2;
input sq $ pa pd tr $ oc;
cards;
ab 11 a 1237
ab 21 a 1179
ab 31 a 1000
ab 41 a 1295
ab 51 a 1218
ab 61 a 1138
ab 71 a 971
ba 81 b 1387
ba 91 b 1025
ba 10 1 b 1225
ba 11 1 b 1050
ba 12 1 b 1050
ba 13 1 b 1387
ba 14 1 b 1150
ab 12 b 1256
ab 22 b 1275
ab 32 b 981
ab 42 b 1387
ab 52 b 1187
ab 62 b 1175
ab 72 b 1012
ba 82 a 1348
ba 92 a 1022
ba 102 a 1226
ba 112 a 1026
ba 122 a 1031
ba 132 a 1298
ba 142 a 1108;
proc anova;
class pd pa tr sq;
model oc=pd pa tr sq;
means tr;
run;
quit;
```

SAS 分析结果如下：

SAS 系统

The ANOVA Procedure

Class Level Information

Class	Levels	Values
pd	2	1 2
pa	14	1 2 3 4 5 6 7 8 9 10 11 12 13 14
tr	2	a b
sq	2	ab ba

Number of Observations Read 28

Number of Observations Used 28

SAS 系统

The ANOVA Procedure

Dependent Variable: oc Source	*DF*	Sum of Squares	Mean Square	*F* Value	*Pr* > *F*
Model	16	472 470.142 9	29 529.383 9	32.42	<0.000 1
Error	11	10 019.285 7	910.844 2		
Corrected Total	27	482 489.428 6			

R-Square	Coeff Var	Root MSE	oc Mean
0.979 234	2.588 670	30.180 19	1 165.857

Source	*DF*	Anova SS	Mean Square	*F* Value	*Pr* > *F*
pd	1	14.285 7	14.285 7	0.02	0.902 6
pa	13	465 206.428 6	35 785.109 9	39.29	<0.000 1
tr	1	7 232.142 9	7 232.142 9	7.94	0.016 7
sq	1	17.285 7	17.285 7	0.02	0.892 9

SAS 系统

The ANOVA Procedure

Level of tr	N	oc	
		Mean	Std Dev
a	14	1 149.785 71	125.057 635
b	14	1 181.928 57	144.633 470

方差分析结果，不同的时期检验结果 $F=0.02$，$P=0.902\ 6>0.1$；不同次序检验结果 $F=0.02$，$P=0.892\ 9>0.1$；不同的处理检验结果 $F=7.94$，$P=0.016\ 7<0.05$。

（王嘉琪　郭　晋）

第10章 非诊断试验析因设计统计分析

当一种医疗器械与一种治疗(如药物治疗)相比较时,研究者除了需要考察器械或药物的单独治疗效果,往往需要继续考察器械与药物治疗相互影响,联合作用是否产生更强的作用(即器械治疗与药物治疗存在交互作用)。此时需要一种更为复杂的试验设计类型——析因设计,相对应的统计分析方法也较单因素设计复杂一些。本章将介绍析因设计定量资料的假设检验方法。由于相对应的非参数检验方法仍然不够成熟,且该类型设计在医疗器械临床试验中应用相对较少,因此本章仅简要介绍析因设计一元参数检验(即一元方差分析)和多元参数检验(即多元方差分析)。

一、析因设计定量资料的一元方差分析

析因设计是一种比较常见的多因素试验设计,在试验研究中应用得比较多。一般来说,如果在试验中涉及的试验因素不超过4个或5个,每个因素的水平数也比较少时,可以考虑用析因设计。析因设计也称全因子试验设计,它要求将全部试验因素的水平进行全面组合,每种组合叫做一个试验点,在各试验点上要求至少做2次以上独立的重复试验,它可以对各因素及因素之间的各级交互作用的效应估计得比较准确。在医疗器械临床试验中,具体的样本含量需要在试验前设计之初根据各个统计参数确定(见样本含量估计一章)。

本设计的不足之处是实施起来更复杂,因此申办者必须保证研究者严格按照研究方案实施临床试验。析因设计会需要更大的样本量,但是由于这种设计类型基本上是两个临床试验合并成一个,因此效率更高。但是,如果试验的样本量是基于检验主效应计算的,则估计器械与药物的交互作用时会使检验效能降低。

当试验按照析因设计来安排,统计指标只有一个而且是定量的,所得到的试验结果就叫做析因设计一元定量资料。若资料满足独立性、正态性和方差齐性,则可以采用析因设计一元定量资料的方差分析来处理。

【例10-1】 某临床试验研究人工胃内水球与某种减肥药对于某些特殊的肥胖患者的治疗效果,试验三组分别为:单独使用人工胃内水球、单独使用减肥药、使用人工胃内水球联用减肥药,对照组为空白对照,所有组受试者均辅助适当体育锻炼,随访期6个月;希望观察使用胃内水球、减肥药的治疗效果,以及使用胃内水球与使用减肥药的交互效应,选择析因设计数据,如表10-1。

分析:针对析因设计,在满足参数检验的前提条件下,统计分析方法选择方差分析,本研究假定数据正态性、方差齐性检验均满足。

表 10-1　胃内水球与减肥药对减肥效果的影响

胃内水球的使用	减肥药的使用	体重减轻量(kg)
否	否	2.8 2.1 2.6 2.1 1.3 1.9 2.3 2.4 1.8 1.9
		2.3 1.0 2.1 2.0 1.0 1.7 1.8 3.4 1.7 0.7
	是	3.7 1.0 1.9 4.1 3.8 4.5 6.5 4.9 3.7 1.9
		5.0 3.9 3.5 2.5 6.3 3.3 6.6 3.7 3.2 3.1
是	否	10.4 10.7 7.3 8.3 8.1 8.0 6.6 6.5 3.4 6.0
		3.9 10.7 12.1 8.6 3.8 6.4 7.9 7.4 6.1 5.9
	是	12.7 13.4 8.7 11.4 16.2 17.5 8.9 11.2 12.7 13.0
		11.7 14.2 6.4 12.0 6.1 8.8 12.1 11.5 10.5 14.8

在实际试验设计时，样本量需要严格计算

所需的 SAS 程序(名为 CT10-1)如下：

程序	说明

DATA CT10-1;
doA=0,1; do B=0,1;
do i=1 to20;
inpu tweight @@; output;
end;end;end;
cards;
	建立数据集
	
	输入数据

2.8 2.1 2.6 2.1 1.3 1.9 2.3 2.4 1.8 1.9 2.3 1.0 2.1 2.0 1.0 1.7 1.8 3.4 1.7 0.7 3.7 1.0 1.9 4.1 3.8 4.5 6.5 4.9 3.7 1.9 5.0 3.9 3.5 2.5 6.3 3.3 6.6 3.7 3.2 3.1 10.4 10.7 7.3 8.3 8.1 8.0 6.6 6.5 3.4 6.0 3.9 10.7 12.1 8.6 3.8 6.4 7.9 7.4 6.1 5.9 12.7 13.4 8.7 11.4 16.2 17.5 8.9 11.2 12.7 13.0 11.7 14.2 6.4 12.0 6.1 8.8 12.1 11.5 10.5 14.8; run; Ods html; PROC GLM data= CT10-1; classA B; model weight=A B A * B / ss3; lsmeansA * B / tdiff pdiff; run; Ods html close; QUIT;	 调用一般线性过程进行分析 "lsmeans"进行两两比较 "tdiff,pdiff"意为输出两两比较的统计量和概率值

此程序的主要输出结果及其解释如下：

The GLM Procedure

Dependent Variable: weight

Source	*DF*	Sum of Squares	Mean Square	F Value	*Pr* > *F*
Model	3	1 103.878 375	367.959 458	87.62	<0.000 1
Error	76	319.146 500	4.199 296		
Corrected Total	79	1 423.024 875			

R-Square	Coeff Var	Root MSE	weight Mean
0.775 727	32.925 78	2.049 218	6.223 750

Source	*DF*	Type Ⅲ SS	Mean Square	*F* Value	*Pr* > *F*
A	1	883.785 125 0	883.785 125 0	210.46	<0.000 1
B	1	191.890 125 0	191.890 125 0	45.70	<0.000 1
A * B	1	28.203 125 0	28.203 125 0	6.72	0.011 5

以上是两因素析因设计定量资料的双向方差分析结果，由最后3行可知因素A(是否使用胃内水球)和B(是否使用减肥药)对指标weight(体重减轻量)影响均有统计学意义($F=210.46$，$P<0.000\ 1$；$F=45.70$，$P<0.000\ 1$)；两因素之间的交互作用(A * B)也有统计学意义($F=6.72$，$P<0.000\ 1$)。

The GLM Procedure

Least Squares Means

A	B	weight LSMEAN	LSMEAN Number
0	0	1.945 000 0	1
0	1	3.855 000 0	2
1	0	7.405 000 0	3
1	1	11.690 000 0	4

Least Squares Means for Effect A * B

***t* for H_0: LSMean(*i*)=LSMean(*j*) / *Pr* > |*t*|**

Dependent Variable: weight

i/j	1	2	3	4
1		−2.947 44 0.004 3	−8.425 67 <0.000 1	−15.038 1 <0.000 1
2	2.947 441 0.004 3		−5.478 23 <0.000 1	−12.090 7 <0.000 1
3	8.425 669 <0.000 1	5.178 228 <0.000 1		−6.612 45 <0.000 1
4	15.038 12 <0.000 1	12.090 68 <0.000 1	6.612 453 <0.000 1	

To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used

以上是4个均值之间两两比较的结果，横向与纵向的编号都是1～4号，横向与纵向交叉处就是相应的两个均值比较的结果，上行数值代表t统计量的数值，下行数值代表相应的概率P值。此处"t检验"的实质仍是方差分析，t值平方就是方差分析中的F值。

结论：使用胃内水球和减肥药均可减轻肥胖患者的体重，且两者的影响具有相互联系。根据具体数据可以发现，使用胃内水球并适当使用减肥药对于该肥胖人群的减肥的效果是最好的。

二、析因设计定量资料的多元方差分析

对于析因设计定量资料，如果有两个或两个以上终点，且临床认为几个终点之间存在相互关系，对资料进行统计分析时需要同时予以考虑，则此时的统计分析方法可选择多元方差分析。

【例 10-2】　沿用【例 10-1】数据，3 个临床终点分别为体重减轻量(kg)、皮下脂肪厚度减小量(mm)、BMI 变化，数据如表 10-2 所示。试验欲考察使用胃内水球和使用减肥药 3 个临床终点的共同影响，假定数据满足参数检验的前提条件，即正态性、等方差。

表 10-2　胃内水球与减肥药对减肥效果(3 个临床终点)的影响

胃内水球	减肥药	编号	体重减轻量(kg)	皮下脂肪厚度减小量(mm)	BMI 变化
否	否	1	2.8	0.3	0.9
		2	2.1	0.2	0.7
		3	2.6	0.3	0.8
		⋮	⋮	⋮	⋮
		20	0.7	0.1	0.2
	是	1	3.7	0.7	1.2
		2	1.0	0.1	0.3
		3	1.9	0.3	0.6
		⋮	⋮	⋮	⋮
		20	3.1	0.4	1.2
是	否	1	10.4	1.7	3.2
		2	10.7	1.9	3.3
		3	7.3	1.2	2.4
		⋮	⋮	⋮	⋮
		20	5.9	1	1.9
	是	1	12.7	2	3.9
		2	13.4	2.4	4.1
		3	8.7	1.4	2.4
		⋮	⋮	⋮	⋮
		20	14.8	2.2	4.3

分析：该试验涉及 3 个指标：体重减轻量(kg)、皮下脂肪厚度减小量(mm)、BMI 变化，这 3 个结果变量临床上认为明显存在相互联系或影响，因此对资料进行统计分析时宜将 3 个结果变量同时予以考虑，进行多元统计分析。由于假定资料满足参数检验的前提条件，采用三元方差分析处理。

由于多元方差分析的计算比较复杂，故不介绍手工算法，直接借助 SAS 软件来实现统计计算。

用“A”表示“是否使用胃内水球”，其两个水平值“1”和“2”分别表示“不使用”和“使用”；用“B”表示“是否使用减肥药”，其两个水平值“1”和“2”分别表示“不使用”和“使用”；用“Y1”“Y2”和“Y3”分别表示“体重减轻量(kg)”“皮下脂肪厚度减小量(mm)”“BMI 变化”3 个结果变量。

实现本资料统计分析的 SAS 程序(名为 CT10-2)如下：

程序			程序		
DATA CT10-2;			8.3	1.1	2.5
DO A=1 TO 2;			8.1	1.8	2.7
DO B=1 TO 2;			8.0	2	2.7
DO Repeat=1 TO 20;			6.6	1.4	1.9
INPUT Y1-Y3; OUTPUT;			6.5	1.7	2
END; END; END;			3.4	0.8	1.1
CARDS;			6.0	0.7	2.1
2.8	0.3	0.9	3.9	0.6	1.3
2.1	0.2	0.7	10.7	2.1	3.2
2.6	0.3	0.8	12.1	1.7	4
2.1	2.5	0.6	8.6	1.5	2.7
1.3	0.2	0.3	3.8	0.7	1.3
1.9	0.1	0.6	6.4	0.8	2
2.3	0.3	0.7	7.9	1.7	2.4
2.3	0.3	0.7	7.4	1.8	2.3
1.0	0	0.4	6.1	1.2	1.9
2.1	0.3	0.7	5.9	1.0	1.9
2.0	0.3	0.6	12.7	2.0	3.9
1.0	0.1	0.4	13.4	2.4	4.1
1.7	0.1	0.5	8.7	1.4	2.4
1.8	0.1	0.5	11.4	1.7	3.6
0.7	0	0.1	16.2	1.8	5.3
2.4	0.2	0.8	17.5	2.6	5.8
1.8	0.1	0.5	8.9	2.7	3.1
1.9	0.1	0.5	11.2	2.8	3.9
3.4	0.3	1.2	12.7	2.7	4
1.7	0.1	0.2	13.0	2.4	4.2
3.7	0.7	1.2	11.7	2.6	3.9
1.0	0.1	0.3	14.2	2.7	4.7
1.9	0.3	0.6	6.4	1.7	2.1
4.1	0.5	1.4	12.0	2.7	3.9
3.8	0.7	1.1	6.1	1.7	1.9
4.5	0.9	1.6	8.8	1.4	2.7
6.5	1.1	2.1	12.1	1.7	4
4.9	1	1.5	11.5	1.8	3.5
3.7	0.6	1.3	10.5	2.1	3.6
1.9	0.3	0.7	14.8	2.2	4.3;
5.0	1	1.6	RUN;		
3.9	0.3	1.4	ODS HTML;		
3.5	0.3	1.4	PROC GLM data= CT10-2;		
2.5	0.3	0.8	CLASS A B;		
6.3	0.7	1.9	MODEL Y1-Y3=A B A*B/ SS3;		
3.3	0.8	1.2	MANOVA H=A B A*B;		
6.6	1	1.9	LSMEANS A * B / SLICE= A;		
3.7	0.4	1.3	LSMEANS A * B / SLICE= B;		
3.2	0.3	1.1	LSMEANS A * B / TDIFF PDIFF;		
3.1	0.4	1.2	RUN;		
10.4	1.7	3.2	ODS HTML close;		
10.7	1.9	3.3			
7.3	1.2	2.4			

主要输出结果及其解释：

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall A Effect

***H* = Type Ⅲ SSCP Matrix for A**

***E* = Error SSCP Matrix**

***S*=1 *M*=0.5 *N*=36**

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.232 397 40	81.47	3	74	<0.000 1
Pillai's Trace	0.767 602 60	81.47	3	74	<0.000 1
Hotelling-Lawley Trace	3.302 974 13	81.47	3	74	<0.000 1
Roy's Greatest Root	3.302 974 13	81.47	3	74	<0.000 1

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall B Effect

***H* = Type Ⅲ SSCP Matrix for B**

***E* = Error SSCP Matrix**

***S*=1 *M*=0.5 *N*=36**

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.582 617 08	17.67	3	74	<0.000 1
Pillai's Trace	0.417 382 92	17.67	3	74	<0.000 1
Hotelling-Lawley Trace	0.716 393 22	17.67	3	74	<0.000 1
Roy's Greatest Root	0.716 393 22	17.67	3	74	<0.000 1

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall A * B Effect

***H* = Type Ⅲ SSCP Matrix for A * B**

***E* = Error SSCP Matrix**

***S*=1 *M*=0.5 *N*=36**

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.894 545 04	2.91	3	74	0.040 2
Pillai's Trace	0.105 454 96	2.91	3	74	0.040 2
Hotelling-Lawley Trace	0.117 886 70	2.91	3	74	0.040 2
Roy's Greatest Root	0.117 886 70	2.91	3	74	0.040 2

这里给出的是多元方差分析的结果：A 因素(是否使用胃内水球)对 Y1，Y2 和 Y3 整体的影响有统计学意义(Wilks' λ =0.232，F=0.26，P<0.000 1)；B(是否使用减肥药)对 Y1、Y2 和 Y3 整体的影响有统计学意义(Wilks' λ =0.583，F=17.67，P<0.000 1)；交互作用项 A * B 对 Y1，Y2 和 Y3 整体的影响有统计学意义(Wilks' λ =0.895，F=2.91，P<0.000 1)。

Least Squares Means

A	B	Y1 LSMEAN	Y2 LSMEAN	Y3 LSMEAN
1	1	1.945 000 0	0.295 000 00	0.585 000 00
1	2	3.855 000 0	0.585 000 00	1.280 000 00
2	1	7.405 000 0	1.370 000 00	2.345 000 00
2	2	11.690 000 0	2.155 000 00	3.745 000 00

这是两因素各水平组合下3项指标的校正均数。

两两比较的结果此处从略。

结论：使用胃内水球和减肥药对体重减轻量(kg)、皮下脂肪厚度减小量(mm)、BMI变化3个指标总体而言具有影响，且两者之间具有相互联系，使用胃内水球并适当使用减肥药对于该肥胖人群的减肥的效果是最好的。

（李　卫　郭　晋　王　扬）

第11章 非诊断试验重复测量设计统计分析

受试对象接受不同处理后，其观测指标的数值会随着时间的推移发生动态变化，为了比较准确地描述和分析这种变化，需要在不同时间点上从每位受试对象身上观测指标的数值，这种考察和评价处理因素和时间因素对观测指标影响的试验设计方法就是重复测量设计。在医疗器械临床试验中，当器械用于受试对象后，经常会对其进行重复观测（随访），此时的研究设计类型属于重复测量设计。在重复测量试验设计中，主要终点既可能为定量指标也可能为定性指标。

一、重复测量设计定量资料的方差分析

按重复测量设计方案实施的试验所得到的试验资料，如果只有一个定量的观测指标，或者虽然有多个定量的观测指标但各个定量观测指标之间在专业上并不存在相互联系或影响，则分析资料时，在资料满足特定参数检验的前提条件下，应采用重复测量设计定量资料的一元方差分析处理。对于重复测量多个临床终点的多元方差分析，方法更为复杂，在医疗器械临床试验中运用的很少，一般对多个指标仍然按多个一元方差分析处理，因此，本章对这部分内容不作介绍，有兴趣读者可参阅相关统计文献。

对于重复测量设计一元定量资料，本质上只有一个指标，但是在重复测量变量各个水平下（各个时间点）的值由于是对同一例受试者进行的连续观测，因此一般具有很强的相关性，此时如果仍然视其为单变量的多个水平，需进行球性检验（sphericity test，Anderson，1958），因为单变量分析方法对协方差结构有着严格的要求，在 SAS 中，只要在 REPEATED 语句中加上选择项/PRINTE，便可实现此检验。如果将资料视为多元资料，可以运用多元方差分析，或者运用混合效应线性模型来处理，它们都允许资料存在相关性，以及协方差矩阵存在多样性。本节将介绍一元方差分析（GLM 过程）和混合效应线性过程（MIXED 过程）两种方法的 SAS 应用。

【例 11-1】 本临床试验采用多中心、单盲、随机对照方法，验证 NOYA 雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统的有效性和安全性，评价该支架的输送系统。与上海微创公司的 Firebird2 药物洗脱支架进行对照。按照入选和排除标准选取合适患者参加本次验证，对所有入选患者在支架置入后 30d、90d、180d、270d、365d 时进行临床随访；对所有入选患者在 180d、270d（±30d）时进行造影随访，以标准的定量冠状动脉造影（QCA）评测获得的晚期管腔丢失为主要疗效指标以评价试验产品的有效性。以试验随访过程中的主要心脏不良事件（MACE）为主要安全性指标以评价试验产品的安全性。QCA 情况中参照血管直径（mm）见表 11-1。

表 11-1　2 组受试者的参照血管直径(mm)

受试对象使用器械情况	受试对象编号	参照血管直径(mm)		
		时间:术后	术后 180d	270d
NOYA 支架	1	3.11	2.96	2.42
	⋮	⋮	⋮	⋮
	15	3.22	3.55	2.48
Firebird2 药物洗脱支架	16	1.90	2.73	3.22
	⋮	⋮	⋮	⋮
	30	3.01	2.82	2.68

分析:受试对象使用器械的情况属于平行组设计,但是对于某些观测指标,如"参照血管直径(mm)"涉及随访观察,在样本失访率不高的前提下,这部分研究应视为重复测量设计,应该使用重复测量设计下的一些统计分析方法,即重复测量方差分析。

利用 GLM 过程分析,程序(程序名为 CT11-1)如下:

程序	说明
DATA CT11-1;	建立数据集
input group $ IDshuhou shu180 shu270;	输入数据
cards;	
1 1 3.11 2.96 2.42	
1 2 4.29 2.14 2.98	
1 3 2.05 2.72 3.31	
1 4 2.82 2.73 2.51	
1 5 2.81 3.26 2.50	
1 6 2.89 2.63 2.86	
1 7 2.38 3.11 2.85	
1 8 2.33 3.32 3.51	
1 9 2.22 3.75 2.31	
1 10 3.01 3.05 2.82	
1 11 3.11 2.99 2.00	
1 12 2.91 2.75 2.95	
1 13 3.08 2.96 2.49	
1 14 2.41 3.19 3.30	
1 15 3.32 3.55 2.48	
2 16 1.90 2.73 3.22	
2 17 2.56 2.32 3.77	
2 18 2.65 3.41 3.78	
2 19 3.26 3.73 3.10	
2 20 1.48 2.34 2.83	
2 21 2.44 3.45 2.55	
2 22 2.10 1.98 2.45	
2 23 3.14 2.89 3.37	
2 24 2.90 3.04 3.35	
2 25 2.76 1.90 2.64	
2 26 2.28 3.52 3.26	

（续　表）

程序	说明
2　27　2.80　2.27　2.61 2　28　3.04　2.64　3.06 2　29　2.78　2.45　3.21 2　30　3.01　2.82　2.68 run; Ods html; PROC GLM data=CT11-1; class group; model shuhou shu180 shu270=group/nouni; repeated time 3 /printe; run; Ods html close;	调用一般线性过程进行分析 “Printe”意为进行球性检验

主要分析结果：

Sphericty Tests

Variables	*DF*	Mauchly's Criterion	Chi-Square	*Pr* > ChiSq
Transformed Variates	2	0.857 514 2	4.150 374 4	0.125 5
Orthogonal Components	2	0.949 044 2	1.412 097 1	0.493 6

这是球性检验的结果，“Orthogonal Components”输出的是正交对比的 Mauchly 球性检验统计量为 0.949，$\chi^2=1.412$，$P=0.493\ 6>0.05$，说明复合球性检验的假设。

The GLM Procedure

Repeated Measures Analysis of Variance

Tests of Hypotheses for Between Subjects Effects

Source	*DF*	Type Ⅲ SS	Mean Square	*F* Value	*Pr* > *F*
group	1	0.079 210 00	0.079 210 00	0.36	0.555 4
Error	28	6.224 688 89	0.222 310 32		

The GLM Procedure

Repeated Measures Analysis of Variance

Univariate Tests of Hypotheses for Within Subject Effects

Source	*DF*	Type Ⅲ SS	Mean Square	*F* Value	*Pr* > *F*	Adj *Pr* > *F*	
						G - G	H - F
time	2	0.571 015 56	0.285 507 78	1.22	0.303 3	0.302 3	0.303 3
time * group	2	1.501 526 67	0.750 763 33	3.20	0.048 1	0.050 8	0.048 1
Error(time)	56	13.119 057 78	0.234 268 89				

Greenhouse-Geisser Epsilon	0.951 5
Huynh-Feldt Epsilon	1.055 5

这里输出了最终的分析的结果，关于主效应“group”$F=0.36$，$P=0.555\ 4>0.05$，两组的总体参照血管直径(mm)平均值差别无统计学意义，故不能拒绝无组间差异的无效假设。关于主效应“time”$F=1.22$，$P=0.303\ 3$，不同时间点参照血管直径(mm)平均值是差别无统计学意义。交互项 time * group $F=3.20$，$P=0.048\ 1<0.05$，可以认为不同的组(试验组、对照组)在不同的时间点上参照血管直径(mm)是有差别的。

如果数据不服从球性假设，需要使用 H-F 值(Huynh-Feldt Epsilon)对其自由度进行校正。此时，亦可使用 Mixed 过程进行分析。

结论：试验组与对照组参照血管直径(mm)没有差别，各个时间点的参照血管直径(mm)无差别，但是二者具有一定的交互作用，具体可以进行两两比较。

在通常分析中，如果交互项有统计学意义，应该先考察交互项，它表示不同组别在不同时间点的指标均值具有差别，而后可以考虑进行两两比较的分析。当交互项不显著时才对主效应(即分组)进行分析。

利用 MIXED 过程分析，程序如下，程序名为 CT11-2。

程序	说明
DATA CT11-2;	建立数据集
Do group=1 to 2;	
Do number=1 to 15;	
Do time= 1 to 3;	
Input blood @@; output;end;end;end;	
cards;	输入数据
3.11 2.96 2.42	
4.29 2.14 2.98	
2.05 2.72 3.31	
2.82 2.73 2.51	
2.81 3.26 2.50	
2.89 2.63 2.86	
2.38 3.11 2.85	
2.33 3.32 3.51	
2.22 3.75 2.31	
3.01 3.05 2.82	
3.11 2.99 2.00	
2.91 2.75 2.95	
3.08 2.96 2.49	
2.41 3.19 3.30	
3.32 3.55 2.48	
1.90 2.73 3.22	
2.56 2.32 3.77	
2.65 3.41 3.78	
3.26 3.73 3.10	
1.48 2.34 2.83	
2.44 3.45 2.55	
2.10 1.98 2.45	
3.14 2.89 3.37	
2.90 3.04 3.35	

（续　表）

程序	说明
2.76　1.90　2.64 2.28　3.52　3.26 2.80　2.27　2.61 3.04　2.64　3.06 2.78　2.45　3.21 3.01　2.82　2.68; run;	

%macro mixed(model);
PROC MIXED data= CT11-2;
class group number time;
model blood=group time group * time;
repeated / type=&model sub=number(group);
run; quit;
%mend;
	调用混合效应线性模型过程进行分析
	
	程序利用宏，写了 5 个过程步，其本质相同，只是“TYPE=”后的选择项不同，它标志着估计时间点之间的协方差矩阵的类型不同。TYPE=VC 称作方差分量型模型、TYPE=CS 称作复合对称型模型、TYPE=UN 称作无结构型模型、TYPE=AR(1)称作一阶自回归型模型、TYPE=SP(POW)叫作空间幂型模型

%mixed(vc); %mixed(cs); %mixed(un); %mixed(AR(1)); %mixed(SP(pow)(Time));	

主要分析结果：见表 11-2。

表 11-2　5 种模型的 Fit Statistics

协方差结构模型	−2 Res Log Likelihood	AIC (smaller is better)	AICC (smaller is better)	BIC (smaller is better)
VC	131.3	133.3	133.3	134.7
CS	131.3	135.3	135.4	138.1
UN	129.3	141.3	142.3	149.7
AR(1)	131.3	135.3	135.4	138.1
SP(POW)	131.3	135.3	135.4	138.1

5 个模型待估计的协方差结构中的参数个数分别是 1,2,6,2,2，对上面的结果进行选择，其中的要求是 AIC,AICC,BIC 都是越小越好，模型中所用的参数越少越好，另外，用似然比的数值，还可以对上面的模型进行更量化的比较，具体方法可以参见有关专著。直观上比较，可以认为上面应该选择 VC 这个模型，因为其 AIC,AICC,BIC 值都是最小的，且参数的个数也是最少的，因此应按上面由“TYPE=VC”计算的结果来统计和专业结论为宜，将程序修改如下：

程序名为 CT11-3

程序	说明
DATA CT11-3;	建立数据集
Do group=1 to 2;	
Do number=1 to 15;	输入数据
Do time= 1 to 3;	
Input blood @@; output;end;end;end;	
cards;	
/*输入程序 CT11-2 的数据*/;	
run;	
Ods html;	
PROC MIXED data= CT11-3;	
class group number time;	调用混合效应线性模型进行分析
model blood=group time group*time;	
repeated / type=un sub=number(group);	
lsmeans group*time /cl diff pdiff;	
run; quit;	对其进行两两比较
Ods html close;	

主要输出结果：

The Mixed Procedure

Null Model Likelihood Ratio Test

DF	Chi-Square	*Pr* > ChiSq
5	2.03	0.845 1

Type 3 Tests of Fixed Effects

Effect	Num *DF*	Den *DF*	*F* Value	*Pr* > *F*
group	1	28	0.36	0.555 4
time	2	28	1.04	0.366 2
group*time	2	28	3.98	0.030 2

这里输出了最终的分析的结果，关于主效应“group”$F=0.36$，$P=0.555\ 4>0.05$，两组的总体参照血管直径(mm)平均值差别无统计学意义，故不能拒绝无组间差异的无效假设。关于主效应“time”$F=1.04$，$P=0.366\ 2$，不同时间点参照血管直径(mm)平均值是差别无统计学意义。交互项 time*group $F=3.98$，$P=0.030\ 2<0.05$，可以认为不同的组(试验组、对照组)在不同的时间点上参照血管直径(mm)是有差别的。

Differences of Least Squares Means

Effect	group	time	_group	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
group * time	1	1	1	2	−0.158 0	0.191 8	28	−0.82	0.417 1	0.05	−0.550 9	0.234 9
group * time	1	1	1	3	0.096 67	0.179 5	28	0.54	0.594 5	0.05	−0.271 0	0.464 3
group * time	1	1	2	1	0.242 7	0.190 7	28	1.27	0.213 8	0.05	−0.148 1	0.633 4
group * time	1	1	2	2	0.083 33	0.185 2	28	0.45	0.656 2	0.05	−0.296 0	0.462 7
group * time	1	1	2	3	−0.209 3	0.173 1	28	−1.21	0.236 6	0.05	−0.563 9	0.145 2
group * time	1	2	1	3	0.254 7	0.157 2	28	1.62	0.116 3	0.05	−0.067 25	0.576 6
group * time	1	2	2	1	0.400 7	0.185 2	28	2.16	0.039 2	0.05	0.021 32	0.780 0
group * time	1	2	2	2	0.241 3	0.179 5	28	1.34	0.189 5	0.05	−0.126 3	0.608 9
group * time	1	2	2	3	−0.051 33	0.166 9	28	−0.31	0.760 7	0.05	−0.393 3	0.290 6
group * time	1	3	2	1	0.146 0	0.173 1	28	0.84	0.406 0	0.05	−0.208 5	0.500 5
group * time	1	3	2	2	−0.013 33	0.166 9	28	−0.08	0.936 9	0.05	−0.355 3	0.328 6
group * time	1	3	2	3	−0.306 0	0.153 4	28	−2.00	0.055 8	0.05	−0.620 2	0.008 170
group * time	2	1	2	2	−0.159 3	0.191 8	28	−0.83	0.413 2	0.05	−0.552 2	0.233 6
group * time	2	1	2	3	−0.452 0	0.179 5	28	−2.52	0.017 8	0.05	−0.819 7	−0.084 32
group * time	2	2	2	3	−0.292 7	0.157 2	28	−1.86	0.073 1	0.05	−0.614 6	0.029 25

这是两两比较的结果，“$Pr > |t|$”一列即为两两比较的概率，“group＝2，time＝1，group ＝2，time＝3”一行的概率为 0.017 8，说明对照组支架在术后和术后 270d 的参照血管直径(mm)是有差别的。

结论：试验组和对照组在各个时间点参照血管直径(mm)上无差别，但是对照组在术后和术后 270d 参照血管直径(mm)存在一定的差别。

二、重复测量设计定量资料的一元协方差分析

在重复测量设计中，有时除观测临床终点变量外，还会存在协变量，它们的存在会影响终点变量的计算，此时如果将其作为一般的重复测量资料进行方差分析，结果有可能是不准确的。我们需要对这种数据运用协方差分析。例如，试验中对每一个受试对象在不同的时间点上某一指标进行了重复测量，而且在“术前”观测了 1 次，在“术后”观测了 $k(k \geqslant 2)$ 次，则应以术前点观测的结果作为“协变量”来处理比较恰当，因为“术前”和“术后 6 个月”“术后 12 个月”“术后 24 个月”的观察，除了“时间”不同，是否进行了手术也是不同的；换言之，它们不是一个变量的 4 个水平，“术后 6 个月”“术后 12 个月”“术后 24 个月”的观察值可以作为一个变量(即术后不同时间)的 3 个水平，而“术前”观察值则应作为协变量，由于每个受试对象“术前”某指标的观察值就是不同的，它可能对结果产生影响，因此我们需要利用协方差分析消除“因基础值不同”而对术后观测结果所造成的影响，更好地阐明不同的受试者在术后不同时间点上观测指标的动态变化情况。

【例 11-2】沿用【例 11-1】 的数据，对于受试者术前的参照血管直径(mm)也给予了测量，数据如下，利用协方差分析，评价试验组与对照组受试者参照血管直径(mm)随时间变化是否存在差异(表 11-3)。

表 11-3　2 组受试者的参照血管直径(mm)

受试对象使用器械情况	受试对象编号	参照血管直径(mm)			
		时间：术前	术后	术后 180d	270d
NOYA 支架	1	1	3.11	2.96	2.42
	⋮	⋮	⋮	⋮	⋮
	15	15	3.22	3.55	2.48
Firebird2 药物洗脱支架	16	16	1.90	2.73	3.22
	⋮	⋮	⋮	⋮	⋮
	30	30	3.01	2.82	2.68

程序名为 CT11-4

程序	说明
DATA CT11-4; input group $ ID shuqian shuhou shu180 shu270 @@; cards;	建立数据集 输入数据
1 1 3.10 3.11 2.96 2.42	
1 2 3.26 4.29 2.14 2.98	
1 3 2.76 2.05 2.72 3.31	
1 4 2.14 2.82 2.73 2.51	
1 5 3.64 2.81 3.26 2.50	
1 6 3.20 2.89 2.63 2.86	
1 7 2.64 2.38 3.11 2.85	
1 8 3.00 2.33 3.32 3.51	
1 9 2.46 2.22 3.75 2.31	
1 10 3.13 3.01 3.05 2.82	
1 11 3.03 3.11 2.99 2.00	
1 12 2.89 2.91 2.75 2.95	
1 13 2.61 3.08 2.96 2.49	
1 14 2.82 2.41 3.19 3.30	
1 15 2.67 3.32 3.55 2.48	
2 16 2.74 1.90 2.73 3.22	
2 17 3.10 2.56 2.32 3.77	
2 18 2.54 2.65 3.41 3.78	
2 19 3.35 3.26 3.73 3.10	
2 20 2.02 1.48 2.34 2.83	
2 21 3.15 2.44 3.45 2.55	
2 22 3.28 2.10 1.98 2.45	
2 23 2.38 3.14 2.89 3.37	
2 24 2.77 2.90 3.04 3.35	
2 25 2.43 2.76 1.90 2.64	
2 26 2.47 2.28 3.52 3.26	
2 27 2.95 2.80 2.27 2.61	
2 28 3.33 3.04 2.64 3.06	
2 29 1.82 2.78 2.45 3.21	
2 30 2.80 3.01 2.82 2.68	

（续　表）

程序	说明
run; Ods html; PROC GLM data=CT11-4; class group; model shuhou shu180 shu270=group shuqian/nouni E X XPX; MANOVA H=group shuqian/printe printh; Lsmeans group; run; Ods html close;	调用一般线性过程进行分析 "MANOVA"意为将"术前"作为协变量行多元方差分析 "Lsmeans group"是输入校正"术前"指标值后，术后几次观测的均值

主要分析结果：

The GLM Procedure

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall group Effect
H = Type Ⅲ SSCP Matrix for group
E = Error SSCP Matrix
$S=1$ $M=0.5$ $N=11.5$

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.799 305 00	2.09	3	25	0.126 7
Pillai's Trace	0.200 695 00	2.09	3	25	0.126 7
Hotelling-Lawley Trace	0.251 086 88	2.09	3	25	0.126 7
Roy's Greatest Root	0.251 086 88	2.09	3	25	0.126 7

这是对两组随时间变化参照血管直径(mm)是否存在差别的假设检验结果，Wilks' Lambda=0.799，$P=0.126\ 7>0.05$，因此可以说明，对照组和试验组受试者术后 3 个时间点参照血管直径(mm)无差别。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall shuqian Effect
H = Type Ⅲ SSCP Matrix for shuqian
E = Error SSCP Matrix
$S=1$ $M=0.5$ $N=11.5$

Statistic	Value	*F* Value	Num *DF*	Den *DF*	*Pr* > *F*
Wilks' Lambda	0.907 339 74	0.85	3	25	0.479 2
Pillai's Trace	0.092 660 26	0.85	3	25	0.479 2
Hotelling-Lawley Trace	0.102 123 01	0.85	3	25	0.479 2
Roy's Greatest Root	0.102 123 01	0.85	3	25	0.479 2

这是"术前"对"术后"参照血管直径(mm)是否存在影响的假设检验结果，Wilks' Lambda=0.90，$P=0.479\ 2>0.05$，因此可以说明，"术前"的指标值对于"术后"的观测结果无影响。

The GLM Procedure

Least Squares Means

group	shuhou LSMEAN	shu180 LSMEAN	shu270 LSMEAN
1	2.822 027 66	3.004 782 05	2.757 092 16
2	2.633 972 34	2.768 551 28	3.054 241 18

这是校正“术前”参照血管直径(mm)值后，“术后”几次观测的均值。

三、具有重复测量变量的高维列联表的统计分析

所谓具有重复测量变量的高维列联表资料是指表中所涉及的定性变量的个数≥3，其中一个变量涉及重复测量。对于此类列联表统计分析可使用加权最小二乘法分析。

【例 11-3】 在传统抗血小板和抗凝治疗基础上，加用血小板糖蛋白Ⅱb/Ⅲa受体拮抗药替罗非班，能否改善ST段抬高型急性心肌梗死(STEMI)患者冠状动脉介入治疗(PCI)术后的心肌组织灌注水平。方法：通过冠状动脉造影(CAG)确诊为STEMI的患者144例，87例患者接受传统抗血小板和抗凝治疗作为对照组，57例患者在CAG术后在对照组治疗的基础上加用替罗非班作为治疗组，治疗组患者在接受替罗非班静脉负荷量后即刻开始PCI，对照组患者在CAG术后亦立即接受PCI治疗。应用TIMI分级和TIMI心肌灌注分级(TMP)评价术前术后心肌灌注变化，并分析治疗前后患者心电图ST段偏移总和比值(sumSTR)的变化。试分析2组患者治疗前后TIMI分级人数有无变化，数据如表11-4所示。

表 11-4 2组患者心肌灌注术前术后TIMI血流分级比较

分组	分级	例数	
	时间：	术前	术后
治疗组	TIMI 0～2级	53	5
	3级	4	52
对照组	TIMI 0～2级	82	16
	3级	5	71

分析：该资料有3个定性变量，其中时间变量涉及重复测量，可使用最小二乘法分析。但是上表是无法了解受试者治疗前后TIMI分级的变化信息的，向软件录入时，应按表11-5录入数据。

表 11-5 2组患者心肌灌注术前术后TIMI血流分级比较

分组	疗效		例数
	时间：	治疗前	治疗后
治疗组		0	1	48
		0	0	5
		1	0	0
		1	1	4

（续　表）

分组	疗效		例数
	时间：	治疗前	治疗后
对照组		0	1	66
		0	0	16
		1	0	0
		1	1	5

表内“0”代表“TIMI 0～2 级”“1”代表“3 级”

SAS 程序（名为 CT11-5）程序如下：

```
DATA CT11-5;
INPUT g t1 t2 f;
CARDS;
1  0  1  48
1  0  0  5
1  1  0  0
1  1  1  4
2  0  1  66
2  0  0  16
2  1  0  0
2  1  1  5;
run;
```

```
ods html;
PROC CATMOD data=CT11-5;
WEIGHT f;
RESPONSE MARGINALS;
MODEL t1 * t2=g|_RESPONSE_ /pred=freq
cov;
repeated t;
run;
ods html close;
```

输出结果如下：

Response Functions and Covariance Matrix

Sample Function Response Covariance Matrix

Number	Function	1	2
1	0.929 82	0.001 144 8	0.000 108
2	0.087 72	0.000 108	0.001 403 9
1	0.942 53	0.000 622 6	0.000 121 5
2	0.183 91	0.000 121 5	0.001 725 1

Analysis of Variance

Source	*DF*	Chi-Square	*Pr* > ChiSq
Intercept	1	858.32	<0.000 1
g	1	2.21	0.136 8
t	1	577.43	<0.000 1
g * t	1	1.57	0.210 1
Residual	0	.	.

Analysis of Weighted Least Squares Estimates

Effect	Parameter	Estimate	Standard Error	Chi-Square	$Pr > ChiSq$
Intercept	1	0.536 0	0.018 3	858.32	<0.000 1
g	2	−0.027 2	0.018 3	2.21	0.136 8
t	3	0.400 2	0.016 7	577.43	<0.000 1
g * t	4	0.020 9	0.016 7	1.57	0.210 1

Predicted Values for Response Functions

Function Number		Observed		Predicted		Residual
		Function	Standard Error	Function	Standard Error	
1	1	0.929 825	0.033 834	0.929 825	0.033 834	0
	2	0.087 719	0.037 469	0.087 719	0.037 469	0
2	1	0.942 529	0.024 952	0.942 529	0.024 952	0
	2	0.183 908	0.041 535	0.183 908	0.041 535	0

结论：由输出结果可以得到，分组的不同疗效差别是没有有统计学意义的，$\chi^2=2.21$，$P=0.136\ 8>0.05$。治疗前、治疗后疗效差别是有统计学意义的，$\chi^2=577.43$，$P<0.000\ 1$。

具有重复测量的高维列联表资料的整理是比较复杂的，只按表中整理数据是极难分析的，本例只有两个时间点，并且分级也只有两个水平，如果有多个时间点、分级多个水平，则应该规范的将数据整理表的形式，否则数据信息是表达不完整的，也无法录入软件进行分析。

（郭　晋　黄耀华　李　卫）

第12章 非诊断试验单组目标值设计的评价方法

随机对照试验(RCT)通过随机分组和设立对照来减少试验可能发生的偏倚,提高试验质量。由于随机对照临床试验的要求严格、科学性强,研究证据可靠,现已被广泛推崇。然而,由于医疗器械具有一些特殊性,往往并不适用随机对照试验,比如,所研究的器械无法进行盲法操作,受试者不能依从随机化分组方案等。

近年来,非随机对照设计越来越受到重视,其中,目标值法在医疗器械非随机对照临床试验中应用特别多。在美国FDA对医疗器械审评中,目标值法在某些医疗器械临床试验中的应用已被认可,并将其作为RCT不适合时的替代方法之一。本章将具体介绍单组目标值法的应用步骤、试验假设、样本量的确定及目标值可信区间的估计。

一、方法介绍

美国FDA对目标值(objective performance criteria, OPC)的定义为:从大量历史数据库(如,文献资料或历史记录)的数据中得到的一系列可被广泛认可的性能标准,这些标准可以作为说明某些器械的安全性或有效性的替代指标或临床终点。通常目标值由靶值(target)和单侧可信区间界限(通常为95%单侧可信区间界限)组成,其中,单侧可信区间界限最为重要,近些年来,有时仅用单侧可信区间界限表示。

对于某些器械,如果存在本研究领域临床认可的、国内外公认的该器械的主要疗效/安全性评价指标及其评价标准,那么可以此评价标准作为目标值,并根据该目标值计算临床试验所需样本量,进行符合该目标值的单组试验。

试验结果的评价,同样应采用主要评价指标的点估计值和单侧可信区间,并将他们与相应目标值进行比较。一般要求点估计值及其单侧可信区间界限均分别不低于(或不高于)OPC的靶值和单侧可信区间界限。但通常,更侧重单侧可信区间界限间的比较,如果单侧可信区间界限不低于(或不高于)OPC的单侧可信区间界限,即可认为医疗器械的有效性(或安全性)达标。

二、应用步骤

应用目标值法进行临床试验的基本步骤为:

1. 根据临床专业知识确定研究器械的主要疗效/安全性评价指标。

2. 根据历史数据或审评机构,明确本研究领域临床认可的、国内/国外公认的疗效/安全性评价标准(如FDA/SFDA指导原则、ISO标准、国标或部标等规范或指南),并以此作为主要评价指标的目标值。

3. 根据目标值确定该临床试验所需的样本量,即受试者人数。

4. 所有受试者接受该医疗器械的治疗或检查，观察其有效性和安全性指标。

5. 估计主要评价指标的单侧95%可信区间，并与目标值进行比较，得出结论。

三、检验假设与统计分析

应用目标值法的医疗器械临床试验是将主要指标的单侧95%可信区间与目标值进行比较的单组试验，其分析的主要评价指标往往为二分类变量（例如，验证射频消融导管的即刻手术成功率>85%），此时，可将该类临床试验视为单样本率与已知总体率的比较研究，故其检验假设为：

$$H_0：P \geq P_0，H_1：P < P_0 \quad 或 \quad H_0：P \leq P_0，H_1：P > P_0$$

其中，P 为预期的主要终点事件发生率，P_0 为OPC单侧95%可信区间界限

当主要评价指标为连续变量时，可将该临床试验视为单样本均数与已知总体均数的比较研究，故其检验假设为：

$$H_0：\mu \geq \mu_0，H_1：\mu < \mu_0 \quad 或 \quad H_0：\mu \leq \mu_0，H_1：\mu > \mu_0$$

其中，μ 为预期的主要评价指标均值，μ_0 为OPC单侧95%可信区间界限

当主要评价指标为高优指标，即 P/μ 的取值越大越好时，则需提前明确主要指标OPC的单侧可信区间下限（例如，射频消融导管的即刻手术成功率至少为85%），检验假设为 $H_0：P \leq P_0$，$H_1：P > P_0$ 或 $H_0：\mu \leq \mu_0$，$H_1：\mu > \mu_0$。通过统计分析得到主要指标的单侧可信区间，若其可信区间下限大于OPC的可信区间下限，即可认为医疗器械的有效性/安全性达标。

当主要评价指标为低优指标，即 P/μ 的取值越小越好时，则需提前明确主要指标OPC的单侧可信区间上限（例如，不良事件的发生率至多为10%），检验假设为 $H_0：P \geq P_0$，$H_1：P < P_0$ 或 $H_0：\mu \geq \mu_0$，$H_1：\mu < \mu_0$。通过统计分析得到主要指标的单侧可信区间，若其可信区间上限小于OPC的可信区间上限，即可认为医疗器械的有效性/安全性达标。

1. *样本量的确定* 样本量估计是临床试验设计中极为重要的环节，充足的样本量才能保证试验有足够的把握度发现实际存在的差异。然而，当样本量过大，会使过多的受试者暴露于危险因素之下，违背伦理要求，也会造成不必要的人力、物力和财力的浪费。因此，准确的估计样本量是临床试验得以有效进行的保证。以下分别介绍主要评价指标为定性指标和定量指标的样本量计算方法。

（1）定性指标：当 P 或 P_0 接近0.5时，Dixon，Massey（1983）提出，应根据正态近似法计算单组率的样本量，计算公式如下：

$$N = \frac{\left[Z_{1-\alpha}\sqrt{P_0(1-P_0)} + Z_{1-\beta}\sqrt{P(1-P)}\right]^2}{(P_0 - P)^2} \qquad (12\text{-}1)$$

式中，α 为检验水准，常取双侧0.05或单侧0.025；$1\text{-}\beta$ 为检验效能，一般取值80%；P_0 为OPC的95%可信区间界限值；P 为预期主要终点的发生率；$Z_{1-\alpha}$ 和 $Z_{1-\beta}$ 为标准正态分布的分位数。值得注意的是，当数据来自人数有限的总体时，需对样本量的估计值进行调整，即 $n' = nN/(n+N)$，其中 n' 为调整后的样本量估计值，n 为通过公式求得的样本量估计值，N 为总体的人数

当 P 或 P_0 接近0或1时，Dixon，Massey（1983）和Chernick，Liu（2002）提出，应基于累积二项分布计算检验效能，再由预期的检验效能反推所需样本量。在计算样本量时，先设定样本量初始值，然后根据检验水准 α 和二项分布 $B(n, P_0)$ 得到拒绝域，再由拒绝域和二项分布B

(n, P)求得检验效能，迭代样本量直到所得的检验效能满足条件为止，此时的样本量，即研究所需的样本量。P_0为 OPC 单侧 95%可信区间界限值；P 为预期主要终点的发生率。

对于检验水准为 α 的单侧检验，H_0：$P \geqslant P_0$，H_1：$P < P_0$，其拒绝域为$(0, k)$，其中 k 为 r 的最大值，r 满足下式，k，r 均为非负整数。

$$\sum_{i=0}^{r} \frac{n!}{i!\ (n-i)!} P_0^i\ (1-P_0)^{n-i} \leqslant \alpha \tag{12-2}$$

相应的检验效能 $1-\beta$ 计算公式如下：

$$1-\beta = \sum_{i=0}^{k} \frac{n!}{i!\ (n-i)!} P^i\ (1-P)^{n-i} \tag{12-3}$$

对于检验水准为 α 的单侧检验，H_0：$P \leqslant P_0$，H_1：$P > P_0$，其拒绝域为$(k, +\infty)$，其中 k 为 r 的最小值，r 满足下式：

$$\sum_{i=0}^{r} \frac{n!}{i!\ (n-i)!} P_0^i\ (1-P_0)^{n-i} \geqslant 1-\alpha \tag{12-4}$$

相应的检验效能 $1-\beta$ 计算公式如下：

$$1-\beta = 1 - \sum_{i=0}^{k} \frac{n!}{i!\ (n-i)!} P^i\ (1-P)^{n-i} \tag{12-5}$$

(2)定量指标：当主要指标为连续变量时，目标值的假设检验可视为单样本 t 检验。O'Brien 和 Muller (1993)给出的单样本 t 检验的样本量估计是建立在自由度为 n-1，非中心参数为 $\sqrt{n}\left(\frac{\mu-\mu_0}{\sigma}\right)$ 的非中心 t 分布基础上。其检验效能的计算公式为：

$$1-\beta = 1 - \text{Prob}\ t\left[t_{1-\alpha,\ n-1}, n-1, \sqrt{n}\left(\frac{|\mu-\mu_0|}{\sigma}\right)\right] \tag{12-6}$$

式中，μ 为预期的主要评价指标均值；μ_0 为 OPC 单侧 95%可信区间界限；σ 为预期的总体标准差。在计算样本量时，一般先设定样本量初始值，然后迭代样本量直到所得的检验效能满足条件为止。此时的样本量，即研究所需的样本量

对于部分定性指标和定量指标的样本量估计，需要不断迭代来计算，如果手工计算会非常复杂。随着计算机技术的发展，可应用样本量计算的专业软件，如 nQuery，PASS 等，进行估计，也可用 SAS 软件编程来估算样本量。

2. *可信区间的估计*　区间估计是按预先给定的概率(1-α)所确定的包含未知总体参数的一个范围，该范围称为参数的可信区间或置信区间(confidence interval，CI)，其含义为：如果能进行重复抽样试验，平均有 1-α(如 95%)的可信区间包含了总体参数；预先给定的概率 1-α 称为可信度或置信度(confidence level)，常取双侧 0.05 或单侧 0.025。可信区间通常由两个可信界限(confidence limit，CL)构成，其中较小的值是可信下限(lower limit，L)，较大的值是可信上限(upper limit，U)，一般表示为(L，U)。

(1)定性指标：对于定性资料，可根据 $np(1-p)$ 的大小，采用近似正态法或精确二项分布法来估计可信区间，其中，n 为病例数，P 为试验得到的主要终点发生率。

当 $nP(1-P)>5$ 时，采用正态近似法，根据 u 分布可得主要终点发生率的单侧 1-α 可信区间为：

$$\mu > L = P - u_{\alpha}\sqrt{\frac{P(1-P)}{n}} \text{ 或 } \mu < U = P + u_{\alpha}\sqrt{\frac{P(1-P)}{n}} \qquad (12\text{-}7)$$

当 $nP(1\text{-}P)<5$ 时，采用精确二项分布法，若 $P=0$，则率的可信区间下限为 0，上限的计算公式为 $U=1-\sqrt[n]{\alpha}$；若 $P>0$，则由二项分布与 F 分布的关系，可得率的单侧可信区间为：

$$\mu > L = x/[x + (n-x+1)F(2(n-x+1), 2x, 1-\alpha/2)] \qquad (12\text{-}8)$$

$$\text{或 } \mu < U = \frac{(x+1)F[2(x+1), 2(n-x)]}{(n-x) + (x+1)F[2(x+1), 2(n-x), 1-\alpha]} \qquad (12\text{-}9)$$

式中 x 为事件发生数，$F(df1, df2, P)$ 为 F 分布的分位数，在 SAS 软件中可用 FINV(P, ndf, ddf) 函数计算该分位数

(2)定量指标：对于定量资料，可根据样本含量的大小，采用 t 分布或 u 分布两种方法估计可信区间。样本量较小时，根据 t 分布的原理可得主要指标均数的单侧 1-α 可信区间为：

$$\mu > L = \bar{x} - t_{\alpha,n-1}S_{\bar{x}} \text{ 或 } \mu < U = \bar{x} + t_{\alpha,n-1}S_{\bar{x}} \qquad (12\text{-}10)$$

当样本量足够大(如 $n>60$)时，根据 u 分布的原理可得主要指标均数的单侧 1-α 可信区间为：

$$\mu > L = \bar{x} - u_{\alpha}S_{\bar{x}} \text{ 或 } \mu < U = \bar{x} + u_{\alpha}S_{\bar{x}} \qquad (12\text{-}11)$$

其中，$\bar{x}$ 为样本均数，$S_{\bar{x}}$ 为样本均数的标准误

四、实例分析

某临床试验欲验证某射频消融导管的有效性和安全性，由于射频消融导管技术成熟风险可控，在临床的应用较为广泛，其预期疗效相对明确，所以采用单组目标值设计。

美国 FDA 有相关产品的临床试验指导原则(见 Cardiac Ablation Catheters Generic Arrhythmia Indications for Use; Guidance for Industry)，其中明确规定了该类产品的主要疗效评价指标为：即刻手术成功率和术后 3～6 个月随访时的成功率。且规定射频消融导管的评价标准为：即刻手术成功率的 95%可信区间下限必须>85%；术后 3～6 个月随访时的成功率的 95%可信区间下限必须>80%。

以指导原则规定的成功率的 95%可信区间下限作为目标值，如果通过临床验证试验，证明试验产品的成功率的 95%可信区间下限达到并超过上述目标值，则可以认为该射频消融导管能够满足临床应用的要求。接下来以主要评价指标之一的即刻手术成功率为例，给出样本量和可信区间估计的过程。

1. 建立检验假设，确定检验水准

$$H_0: P \leqslant P_0,\ H_1: P > P_0$$

其中，P 为试验预期射频消融导管的即刻手术成功率(预期能达到 95%)，P_0 为 OPC95%可信区间下限值(规定为 85%)。检验水准 α 取单侧 0.025。

2. 确定试验所需样本量　验预期射频消融导管的即刻手术成功率为 95%，OPC 的可信区间下限为 85%。由于 P 和 P_0 较接近 1，当检验水准取单侧 0.025，检验效能取 80%时，根据累积二项分布法得出，本试验至少需要入选受试者 79 例；考虑研究过程中最大可能出现 20%的脱落率，故本研究预计入选患者 99 例。根据计算公式，用 SAS 统计软件编程实现样本量的

估计，SAS 程序如下：

```
%macro Binomial(a,P0,P,n);
data Binomial;
  a=&a;P0=&P0;P=&P;n=&n;
  if P<P0 then do;
    k=n;
    do until(CDF('BINOM',k,P0,n)<=a);
      k=k-1;
    end;
    power=CDF('BINOM',k,p,n) ;
  end;
  if P>P0 then do;
    k=0;
    do
until[(1-CDF('BINOM',k,P0,n)]<=a);
      k=k+1;
    end;
power=1-CDF('BINOM',k,p,n) ;
  end;
  format power 4.2;
run;
proc print data=Binomial label;
  var a P0 p power n;
  label a='检验水准'
    P0='目标值'
    P='预期主要终点发生率'
    power='检验效能'
    n='样本量';
quit;
%mend Binomial;
%Binomial(0.025,0.85,0.95,79);
```

SAS 程序说明：创建名为 Binomial 的宏程序，其参数的意义分别为：a 为检验水准；P_0 为目标值；P 为试验预期主要终点发生率；power 为检验效能，即把握度；n 为样本量。

SAS 输出结果如下：

检验水准	目标值	预期主要终点发生率	检验效能	样本量
0.025	0.85	0.95	0.80	79

3. *估计成功率的可信区间*　本研究选取 100 名受试者进行临床试验，有 97 名受试者即刻手术取得成功。由于 $nP(1-P)=2.91<5$，应采用精确二项分布法计算可信区间下限为，成功率的单侧可信区间下限计算公式如下：

$$\mu > L = x/[x+(n-x+1)F(2(n-x+1),2x,1-\alpha/2)]$$
$$= 97/[97+4F(8,194,0.987\,5)] = 0.91 = 91\%$$

4. *结果解释*　统计分析结果显示，该射频消融导管的即刻手术成功率的单侧 95%可信区间下限值为 91%，达到并超过目标值 85%，可以认为该射频消融导管能够满足临床应用的要求。

五、单组目标设计方法的适用条件

1. *低风险、成熟的产品*　目标值法由于没有对照组，因此无法与市场同类产品对比，不能得到临床实践的广泛认可，因此仅适用于低风险产品，手术成功率必须高于 90%，在试验进行前与进行过程中，需要反复与 SFDA 及临床专家、统计专家进行论证。

2. *目标值的确诊需要获得充分的支持*　临床试验应用单组目标值法的关键是确定目标值。目标值应由大样本数据汇总得到，有足够多的文献数据的支持，且单侧可信区间界限值应被临床业界广泛认可和接受，必须为国内/国外公认的疗效/安全性评价标准（如：FDA/SFDA 指导原则、ISO 标准、国标或部标等规范或指南）。

3. 保证试验样本量充足　在试验设计阶段，要准确地估计样本量，保证具有充足的检验效能。此外，与随机对照试验相比，目标值法的优点是需要的样本量少。虽然目标值法使得医疗器械临床试验简单易行，但试验前期工作需要搜集大量的同类产品的临床试验结果资料，来确定可被广泛接受的目标值。

4. 充分了解试验的影响因素　目标值法不设立对照组，无法调整非处理因素对结果的影响，所以在解释上带来一定困难。研究结果的解释必须限定在某些特定条件之下。如受试者的入选标准和排除标准必须明确，如果年龄、性别、疾病严重程度发生改变，临床试验结果也会随之改变；医院的医疗及护理水平的高低也会对试验结果造成一定影响。这不仅要求临床试验的质量控制要严格，也要求确定目标值时要充分考虑到各种因素的影响。

（李　卫　唐欣然　郭　晋）

第13章 诊断试验的统计分析

一、诊断试验的常用统计指标

诊断(diagnosis)是指医生通过观察和思维，对就诊者所患疾病的症状、体征和其他表现进行分析从而识别、判断疾病的过程。为了给病人作出诊断所应用的各种实验室检查、医疗仪器检查及其他方法检查，称为诊断试验(diagnostic test)。

广义的诊断试验包括各种实验室检查、病史、体检的临床资料，各种影像诊断和仪器诊断等。目前新的诊断试验日益增多，研究和评价这些新的诊断试验，掌握现有诊断试验的特性和临床价值，有助于临床医生和研究者正确选用各种诊断试验，科学地解释诊断试验的各种结果，提高诊断水平，为准确、合理地防治疾病提供依据，避免单凭经验造成的错误。诊断试验的结果通常可整理成表 13-1 所示的形式。

表 13-1 诊断试验结果的表达形式

诊断试验结果	人数		
	金标准诊断结果：阳性(患者)	阴性(非患者)	合计
阳性(患者)	a(真阳性)	b(假阳性)	$a+b$
阴性(非患者)	c(假阴性)	d(真阴性)	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d$

诊断试验的常用统计指标比较多，下面分类进行介绍。

(一)评价诊断试验真实性的指标

真实性(validity)是指测量值与实际值相符合的程度，亦称效度，又称准确性(accuracy)。评价诊断试验真实性的统计指标如下：

1. 金标准(gold standard)或标准诊断　评价任何一种诊断试验在临床上的价值，都是与金标准相比较而言的。所谓金标准，是医学家们公认的、具有诊断某一疾病价值的最佳方法，如国际、国内公认的诊断标准、方法、病理学检查、手术所见及特殊检查所获结论等。若不能得到金标准，则应该用目前公认较可靠的“经典方法”代替。

研究对象应是公认标准诊断确诊的病人或非病人，根据试验诊断与标准诊断(金标准)结果列出四格表，计算其他一系列指标来评价诊断性试验的临床应用价值。如研究者无明确标准诊断(金标准)，将影响对此项诊断试验的评判。

2. 灵敏度(sensitivity，se)　即真阳性率，指标准诊断确诊阳性组中，诊断试验阳性结果人数与患者组总人数的比值，即：

$$Se=\frac{\text{真阳性人数}}{\text{患者组总人数}}=\frac{a}{a+c}\times 100\% \tag{13-1}$$

灵敏度反映了诊断试验将实际有病的人正确地判为患者的能力。

3. 假阴性率(false negative rate,FNR) 即漏诊率,指标准诊断确诊阳性组中,诊断试验阴性结果人数与患者组总人数的比值,即:

$$FNR=\frac{\text{假阴性人数}}{\text{患者组总人数}}=\frac{c}{a+c}\times 100\%=1-\text{灵敏度} \tag{13-2}$$

假阴性率(即漏诊率)反映了诊断试验将实际有病的人错判为非患者的比例。

4. 特异度(specificity,Sp) 即真阴性率,指标准诊断确诊阴性组中,诊断试验阴性结果人数与非患者组总人数的比值,即:

$$Sp=\frac{\text{真阴性人数}}{\text{非患者组总人数}}=\frac{d}{b+d}\times 100\% \tag{13-3}$$

特异度反映了诊断试验将实际无病的人正确地判定为非患者的能力。

5. 假阳性率(false positive rate,FPR) 即误诊率,指标准诊断确诊阴性组中,诊断试验阳性结果人数与非患者组总人数的比值,即:

$$FPR=\frac{\text{假阳性人数}}{\text{非患者组总人数}}=\frac{b}{b+d}\times 100\%=1-\text{特异度} \tag{13-4}$$

假阳性率(即误诊率)反映了诊断试验将实际无病的人错误地判断为患者的比例。

灵敏度只与患者组有关,与非患者组无关。灵敏度越高,漏诊机会越小。当疾病漏诊可能造成严重后果时,如癌症的诊断,必须选用灵敏度高的试验,而灵敏度高常伴之高误诊率,故在普查粗筛患者时灵敏度高的试验可减少漏诊,对筛查结果阳性者再进一步检查确诊。一项灵敏度高的试验,结果为阴性时,对于排除此病是有意义的。如结核菌素试验结果为阴性,一般可认为没有感染结核。

特异度只与非患者组有关,与患者组无关。特异度越高,误诊机会越小。如在普查粗筛患者的基础上需进一步确诊某病或挑选科研病例时,则应选用特异度高的诊断方法。某一特异度高的诊断试验结果为阳性时,说明患该病的概率较高,是确诊的有力证据。但某一诊断试验的特异度越高,伴随而来的是漏诊率也越高。例如用检测血清 SM 抗体诊断系统性红斑狼疮(SLE),SM 抗原的特异性很强,查出 SM 抗体者,确诊红斑狼疮无疑,但在确实患红斑狼疮的病人中,约有 70%将因无 SM 抗体而漏诊。

灵敏度和特异度是评价诊断性试验的两个重要指标,它们是诊断性试验本身所固有的,比较稳定,受患病率影响较小。

6. 约登指数(Yonden index) 又称正确指数,是评价真实性的综合指标,表示筛检/诊断试验发现真正的患者与非患者的总能力。

$$\text{约登指数}=\text{灵敏度}+\text{特异度}-1 \tag{13-5}$$

7. 似然比(likelihood ratio,LR) 指患者中得出某一试验结果的概率与非患者中得出这一试验结果的概率的比值。是同时反映灵敏度和特异度的复合指标,分为阳性似然比($+LR$)和阴性似然比($-LR$)。

$$+LR=\frac{\text{真阳性率}}{\text{假阳性率}}=\frac{\text{灵敏度}}{1-\text{特异度}} \tag{13-6}$$

$$-LR=\frac{\text{假阴性率}}{\text{真阴性率}}=\frac{1-\text{灵敏度}}{\text{特异度}} \tag{13-7}$$

阳性似然比表示诊断试验正确判断阳性可能性是错误判断阳性可能性的倍数，显然此数值越大越好，说明诊断与客观相符的程度越高。

阴性似然比表示诊断试验错判阴性可能性是正确判断阴性可能性的倍数，当然此数值越小越好，说明试验错判的可能性越小。

可根据受试者工作特性曲线（receiver operating characteristic curve，简称 ROC 曲线）确定试验阳性结果的截断值（与测得的观察值的分布有关）。ROC 曲线是用真阳性率（即灵敏度）和假阳性率（即误诊率）作图得出的曲线，可以表示灵敏度和特异度之间的关系。其斜率是阳性似然比。

（二）评价诊断试验可靠性的指标

可靠性（reliability）又称信度，重复性、精确性。指筛检/诊断试验在相同条件下重复测量同一受试者时，所获结果的一致性。评价诊断试验可靠性的指标如下。

1. 定性资料用诊断符合率（diagnose accordance rate，DAR）和 Kappa 值来评价。

$$e=\frac{\text{真阳性人数}+\text{真阴性人数}}{\text{受试总人数}}=\frac{a+d}{a+b+c+d}\times 100\% \tag{13-8}$$

诊断符合率反映了诊断试验结果与金标准诊断结果的符合程度。

$$\text{Kappa}=\frac{P_o-P_e}{1-P_e} \tag{13-9}$$

式中，P_o 为两检测法一致率（即诊断符合率 e），P_e 为机遇一致率

2. 定量资料组内相关系数或 Bland-Altman 方法来评价。

（三）评价诊断试验的其他指标

1. 预测值　预测值（predictive value）分为阳性预测值和阴性预测值。阳性预测值反映的是试验阳性者属于真病例的可能性。阴性预测值反映的是试验阴性者属于非患者的可能性。计算公式为：

$$\text{阳性预测值}=\frac{a}{a+b}\times 100\% \tag{13-10}$$

$$\text{阴性预测值}=\frac{d}{c+d}\times 100\% \tag{13-11}$$

灵敏度越高，阴性预测值越高；特异度越高，阳性预测值越高。预测值与受检人群目标疾病患病率（prevalence）密切相关。

$$\text{阳性预测值}=\frac{\text{灵敏度}\times\text{患病率}}{\text{灵敏度}\times\text{患病率}+(1-\text{患病率})(1-\text{特异度})} \tag{13-12}$$

$$\text{阴性预测值}=\frac{\text{特异度}\times(1-\text{患病率})}{\text{特异度}\times(1-\text{患病率})+(1-\text{灵敏度})\times\text{患病率}} \tag{13-13}$$

2. 信息量　信息量（information content，I）有专门的计算公式，但较为复杂。它是由阳性预测值、阴性预测值及就诊患病率三者所确定的一项综合指标。信息量表示就诊者经诊断后所减少的其患病与否的不肯定程度。因此，信息量越大，诊断试验的价值越高。基于医学诊断大多面向的是来医院就诊的人群，因而可用就诊患病率替代人群患病率。

3. 平均信息量　鉴于信息量与受试人群患病率有关，故提出了一种独立于患病率的信息指标——平均信息量（mean information content），作为诊断试验在所有可能患病率下所提供信息大小的一种平均量度，与患病率无关。平均信息量越大，表明试验用于诊断越有效，即试

验的鉴别诊断价值越高。

二、关于定性指标的一致性评价

试验中如有金标准，在与金标准比较时，应报告灵敏性和特异性、阳性似然比和阴性似然比、阳性预测值和阴性预测值、其双侧95%置信区间。与非金标准比较时，应报告阳性一致性百分比、阴性一致性百分比、总体一致性百分比，仅仅使用“敏感性”和“特异性”描述新试验与非金标准的比较结果是不恰当的。当计算诊断的准确性或一致性时，应抛弃模棱两可的诊断结果。对于一致性的计算，应进行假设检验，使用 Kappa 检验的方法以及加权 Kappa 检验的方法予以计算。并给出 Kappa 值。

(一)配对 2×2 表资料的一致性评价

配对研究设计 2×2 表数据是指接受了两种不同的处理方法得到的数据，每种处理方法的结果都可分为“阳性”和“阴性”两种。其中一种方法应为金标准，需检查另一种方法与其是否一致。格式如表 13-2。

表 13-2　配对研究设计 2×2

处理方法 A 所得结果	例数		
方法 B 所得结果：	+	−	合计
+	a	b	$n_1=a+b$
−	c	d	$n_2=c+d$
合计	$m_1=a+c$	$m_2=b+d$	$n=a+b+c+d$

这类资料分析方法有两种：其一，检验两种处理方法检测结果不一致的部分差别是否具有统计学意义，可以用 McNemarχ^2 检验；其二，检验两种处理方法检测结果是否具有一致性，可以用 Kappa 检验，即一致性检验法。

McNemar χ^2 检验计算公式如下：

若 $b+c \geqslant 40$ 时可应用未校正的公式：

$$\chi^2=\frac{(b-c)^2}{b+c},\nu=1 \tag{13-14}$$

若 $b+c<40$ 时应用连续性校正公式：

$$\chi^2=\frac{(|b-c|-1)^2}{b+c},\nu=1 \tag{13-15}$$

McNemar χ^2 检验统计量近似服从自由度为 1 的 χ^2 分布；Kappa 检验在 SAS 统计软件中直接给出检验统计量 Kappa 值和 P 值，并对总体 Kappa 值与“0”之间的差别是否具有统计学意义进行假设检验，借助 P 值即可进行相关统计推断。

【例 13-1】 对于甲状旁腺功能亢进的研究，一种诊断方法是正电子放射断层照相(PET)(A 法)，另一种方法为双阶段^{99m}Tc-sestamibi-SPEC(B 法)，试比较两种方法检测结果之间不一致部分的差别是否具有统计学意义；并检验其一致性是否具有统计学意义(表 13-3)。

表 13-3　两种方法对甲状旁腺功能亢进检测结果比较

A	例数		
B 方法检测结果：	+	−	合计
+	37	33	70
−	19	121	140
合计	56	154	210

比较两种检测方法不一致部分的频数之间的差别是否具有统计学意义，应选用配对设计定性资料的χ^2检验，即 McNemar χ^2检验；同时选择 kappa 检验检测两种方法的一致性。

用 McNemar χ^2检验分析资料，SAS 程序如下（程序名 CT13-1）：

程序	说明
DATA CT13-1; Do a=1 TO 2; Do b=1 TO 2; Input f @@; Output; End; End; Cards; 37　33 19　121 Run; Ods html; Proc freq; Weight F; Tables a * b; test agree; Run; Ods html close;	建立数据集 输入数据 “Test agree”进行 McNemar χ^2检验和 kappa 检验 如果只进行一致性检验，可使用“Test　Kappa”语句进行 Kappa 检验

SAS 输出结果及其解释如下：

McNemar 检验	
统计量（S）	3.769 2
自由度	1
$Pr > S$	0.052 2

以上是 McNemar Test 结果，检验统计量为 3.769 2，$P=0.052\ 2>0.05$，可认为两种方法在检测甲状旁腺功能亢进方面不一致部分的差别无统计学意义。

简单 Kappa 系数	
Kappa	0.413 5
渐近标准误差	0.067 3
95%置信下限	0.281 7
95%置信上限	0.545 4

H_0 检验：Kappa = 0	
H_0 下的渐近标准误差	0.068 1
Z	6.068 8
单侧 $Pr> Z$	<0.000 1
双侧 $Pr>\|Z\|$	<0.000 1

以上是 Kappa 检验结果，Kappa 值为 0.413 5，此值比较小，假设检验结果显示$P<0.000\ 1$，具有统计学意义。观察的一致率=158/210=75.23%，如果临床认可，即可以认为两种方法具有较高的一致性。

(二)R×C 表资料的一致性评价

与配对 2×2 表结构类似，R×C 表同样为两种检测方法得到的数据，在表中表现为行变量与列变量，只是每个变量的水平数不是 2，而是多个等级，例如“优、良、可、差”或“轻、中、重”等。两个变量的水平数应相等。我们需要分析得到两种方法的检测结果之间是否具有一致性，可利用 Kappa 检验。

一致性检验(Kappa 检验)计算公式如下：

$$Pa=\frac{\text{实际观察的一致数}}{\text{总观察例数}} \tag{13-16}$$

$$Pe=\frac{\text{期望一致数}}{\text{总观察例数}} \tag{13-17}$$

$$K=\text{Kappa}=\frac{(P_a-P_e)}{(1-P_e)} \tag{13-18}$$

$$Sk=\frac{\sqrt{P_e+P_e^2-\frac{1}{n^3}\sum R_iC_j(R_i+C_j)}}{(1-P_e)\sqrt{n}} \tag{13-19}$$

$$U=\frac{K}{S_k}\sim N(0,1) \tag{13-20}$$

【例 13-2】 某临床试验，检验其生产的 X 线机(A 组)与市场使用的 X 线机对疾病诊断的一致性，数据如表 13-4 所示，检验其一致性是否具有统计学意义。

表 13-4 两种方法对疾病检测结果比较

A 器械	例数			合计
B 器械：	严重	中度	无	
严重	37	29	3	69
中度	21	99	9	129
无	2	14	27	43
合计	60	142	39	241

建立检验假设：

H_0：Kappa=0(两诊断结果不一致)

H_1：Kappa≠0(两诊断结果一致)

$\alpha=0.05$

SAS 程序 CT13-2：

程序	说明
DATA CT13-2;	建立数据集
DO A=1 TO 3;	
DO B=1 TO 3;	
INPUT F@@;	
OUTPUT;	
END; END;	输入数据
CARDS;	
37　29　3	
21　99　9	
2　14　27;	
ods html;	
PROC FREQ data=CT13-2;	
WEIGHT F;	
TABLES A * B;	
Test Kappa wtkap;	"Kappa"语句进行 Kappa 检验
RUN;	
ods html close;	

程序运行的主要结果及其解释如下：

对称性检验	
统计量（S）	2.567 0
自由度	3
$Pr > S$	0.463 3

对称性检验的结果为：$S=2.567\ 0$，$P=0.463\ 3>0.05$，说明此方表的频数满足对称性假设，即此表中的各频数是关于主对角线对称的。

简单 Kappa 系数	
Kappa	0.446 2
渐近标准误差	0.051 9
95%置信下限	0.344 6
95%置信上限	0.547 9

H_0 检验：Kappa = 0	
H_0 下的渐近标准误差	0.047 5
Z	9.398 4
单侧 $Pr> Z$	<0.000 1
双侧 $Pr>\|Z\|$	<0.000 1

Kappa 系数部分给出了简单统计量的值、渐近标准误、总体 Kappa 值的 95%置信区间。最后给出了一致性检验的结果：Kappa=0.446 2，$P<0.000\ 1$，表明两种 X 线机的诊断结果之间的一致性具有统计学意义。

加权的 Kappa 系数	
加权的 Kappa	0.489 9
渐近标准误差	0.050 4
95%置信下限	0.391 3
95%置信上限	0.588 6

H_0 检验：加权的 Kappa = 0	
H_0 下的渐近标准误差	0.047 6
Z	10.295 5
单侧 $Pr>Z$	<0.000 1
双侧 $Pr>\|Z\|$	<0.000 1

如果表格不满足对称性假设，即此表中的各频数并不关于主对角线对称。此部分 Kappa 系数部分给出了加权 Kappa 统计量的值，此时采用加权 Kappa 检验更为合适（充分利用了非对角线上的信息），Weighted kappa=0.489 9，$P<0.000\ 1$，表明两种 X 线机测定结果之间的一致性具有统计学意义。

结论：经 Kappa 检验（即一致性检验），两个诊断结果是基本一致的，若临床上认为观察一致率 $P=(37+99+27)/241=67.6\%$ 符合要求，即两种 X 线机的检测结果是一致的。

三、定量指标一致性的计算

（一）配对 t 检验

配对 t 检验主要检验的是两诊断器械的系统误差是否有差别，即对两测量结果的系统误差敏感，但不能兼顾随机误差。

计算公式：

$$t=\frac{|\bar{d}-0|}{s_d},\ s_d=\frac{s}{\sqrt{n}},\ df=n-1 \tag{13-21}$$

（二）Pearson 相关系数

Pearson 相关系数是用于表示两定量资料线性相关关系的密切程度，而非一致性。仅当斜率=1，截距=0 时才为一致。样本量少会导致低估相关系数。对相关系数的检验即使 $P<0.05$，不能下结论。

计算公式：

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \tag{13-22}$$

对 r 进行假设检验：$t_r=\dfrac{|r-0|}{S_r}=\dfrac{|r|}{\sqrt{(1-r^2)/(n-2)}}$，$df=n-2$　(13-23)

(三)组内相关系数(Intra-class Correlation Coeficients,ICC)

检测测量方法间变异占总变异的比例，对系统误差和随机误差均敏感。ICC 值较高表明作为两诊断器械测量差别的系统误差与随机误差均小，数据一致性良好。判断标准：ICC＞0.7，样本量少会导致 ICC 被低估。

(四)Bland-Altman 法

该法可较好地评价定量数据两种测量结果的一致性，同时控制系统误差和随机误差。首先应绘制两诊断器械测量值差值对应于均值的散点图(Bland-Altman 图)，设 $D=X_M-X_N$，$A=(X_M+X_N)/2$，即绘制 D 与 A 的散点图，观察 D 与 A 的关系。

1. D 与 A 独立，D 及 A 相关关系无统计学意义。计算一致性限度(limits of agreement)作为评价一致性的指标。即差值 D 的 95%参考值范围 D±1.96S。此时，一致性限度既可衡量系统误差又能估计随机误差的大小。

“D±1.96S”必须被包含在临床认可的界值之内。

界值的确定：有临床意义的一致性界限(由临床医生决定)。

2. D 与 A 不独立，D 及 A 相关关系有统计学意义。进行回归分析 $D=\alpha+\beta A+\varepsilon$，检查 α，β 与 0 之间差别是否有统计学意义。当 $\alpha=\beta=0$，说明两种诊断器械的诊断结果具有一致性。

附 Bland-Altman 点图(图 13-1)：

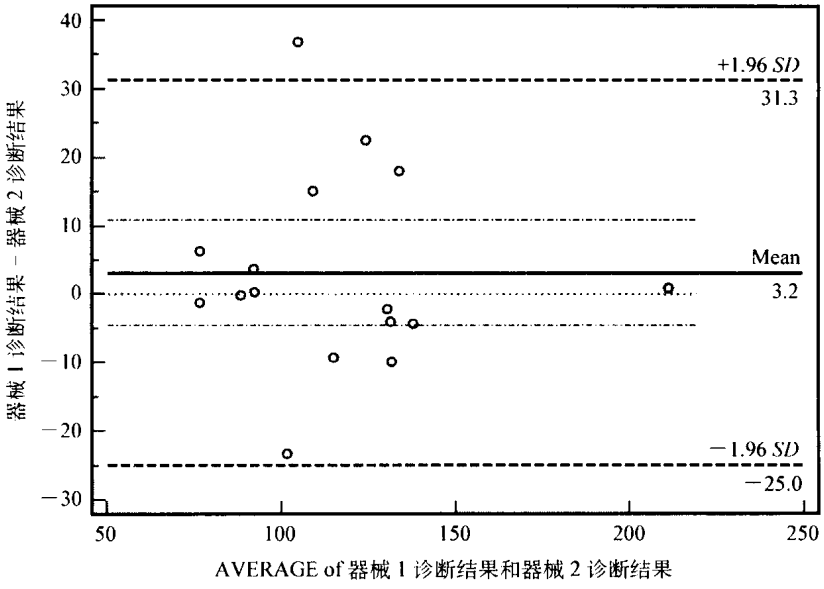

图 13-1　两诊断器械诊断结果 Bland-Altman 点图

(五)ATE / LER 区域(FDA 可豁免诊断器械一致性评价方法)

首先建议两个区域，LER(限制误差区域)和 ATE(容许误差区域)，见图 13-2。

对于 ATE 区域，观察值所落在 ATE 区域(包括低、中、高不同程度)的数量应在接近 95%，对于整个测量范围，在单侧 95%可信区间下落在较低区域中的数量应超过 92%。

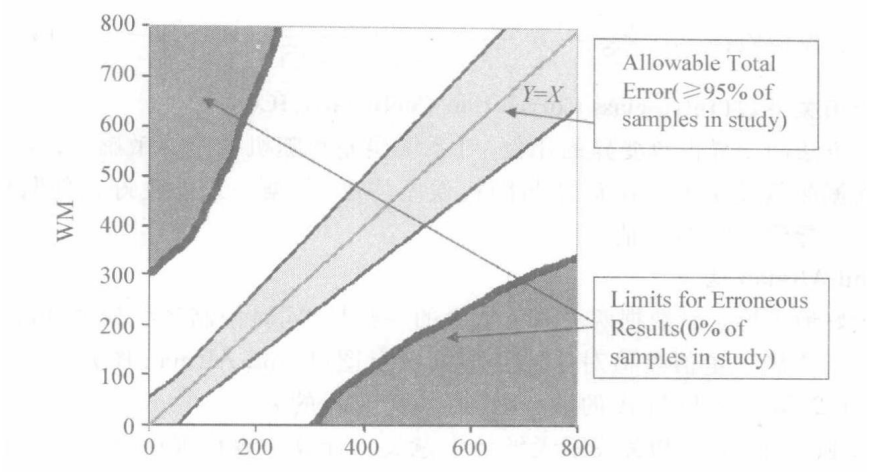

图 13-2　LER 区域和 ATE 区域

对于 LER 区域，需要报告落下 LER 区域下不同程度的值。对于整个测量范围，在单侧 95%可信区间下落在较高区域中的数量应少于 1%。

同时满足上述条件，即可认为一致性较好。

(六)最小二乘回归估计

依据最小二乘法，即残差平方和最小原理，以诊断器械 2 诊断结果 y 对诊断器械 1(金标准)诊断结果 x 拟和直线回归方程，通过对斜率与截距的检测，考察诊断器械的系统误差。最小二乘估计是假定诊断结果 x(金标准)没有随机测定误差，另一诊断结果 y 在给定水平上具有正态分布，且标准差在整个测定范围上是固定的。

经假设检验，斜率 b 接近 1，截距 a 接近 0，说明两种诊断器械的一致性比较好。最小二乘回归分析需要绘制散布图(图 13-3)。

计算公式：

回归线估计：$\hat{y} = a + bx$　　(13-24)

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{l_{xy}}{l_{xy}}　　(13\text{-}25)$$

$$a = \bar{y} - b\,\bar{x}　　(13\text{-}26)$$

对 b 进行假设检验：$t_b = \frac{|b-0|}{S_b} = \frac{|b|}{S_{y.x}/\sqrt{l_{xx}}}$　　(13-27)

对 a 进行假设检验：$t_a = \frac{|a-0|}{S_a} = \frac{|a|}{S_{y.x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}}}$　　(13-28)

其中，　$S_{y.x} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$　　(13-29)

$$l_{xx} = \sum (x_i - \bar{x})^2　　(13\text{-}30)$$

附散布图，见图 13-3。

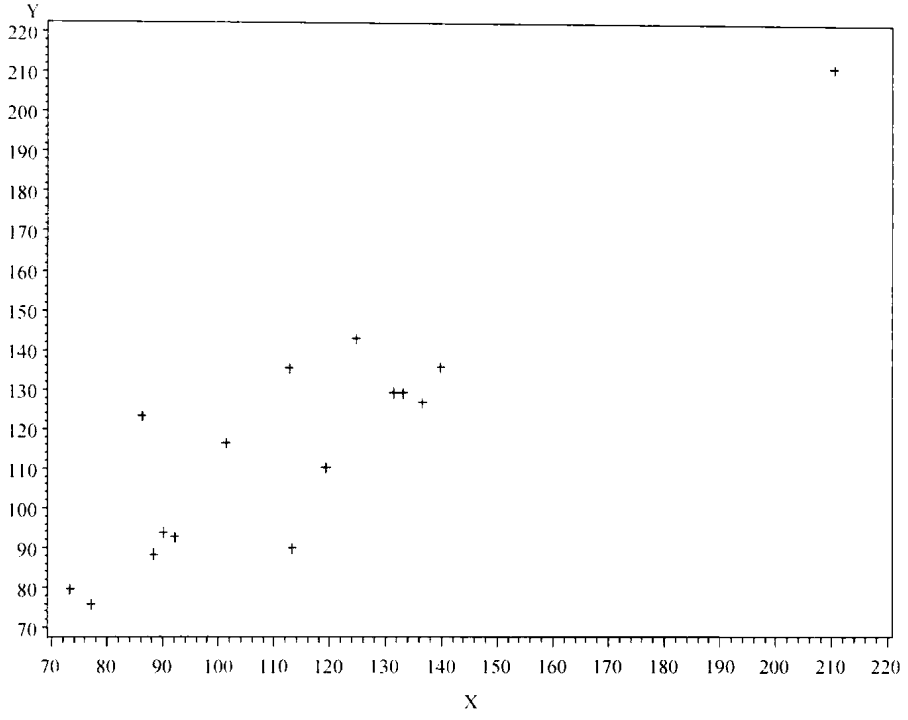

图 13-3　两诊断器械诊断结果散布图

(七)Deming 回归估计

在医疗器械临床检验中，对于两种诊断器械的一致性进行比较，传统的统计方法可以使用最小二乘法，即线性回归分析。这种模型要求一种诊断器械的测量是“金标准”，即，方程中自变量无明显的系统误差和随机误差，为固定变量；而响应变量为随机变量。然后，在有些情况下，两种诊断方法都存在着误差，即，新方法和标准的方法均为随机变量，这种情况是不满足最小二乘估计的假定条件的。我们在分析时，考虑使用 Deming 回归方法。

Deming 回归估计考虑到了两种诊断器械测量的随机测定误差，但是它假定分析标准差的比值是固定的。Deming 回归分析需要对每种诊断器械对于每个个体进行双份测定。经假设检验，斜率 b 接近 1，截距 a 接近 0，说明两种诊断器械的一致性比较好。

Deming 回归估计的计算公式为：

$$\hat{y} = a + bx \tag{13-31}$$

$$b = [(\hat{\lambda} q - u) + \sqrt{(u - \hat{\lambda} q)^2 + 4\hat{\lambda} p^2}]/2\hat{\lambda} p \tag{13-32-1}$$

$$a = y - b\bar{x} \tag{13-32-2}$$

$$\text{其中，} \hat{\lambda} = \frac{(1/k)\sum (x_{1i} - x_{2i})^2}{(1/k)\sum (y_{1i} - y_{2i})^2} \quad \text{，} k \text{ 代表 } k \text{ 对观测} \tag{13-33}$$

$$u = \sum (x_i - \bar{x})^2 \tag{13-34}$$

$$q = \sum (y_i - \bar{y})^2 \tag{13-35}$$

$$p = \sum (x_i - \bar{x})(y_i - \bar{y}) \tag{13-36}$$

其中，u，q 为 x，y 变量的离均差平方和；由于每个个体使用 x 器械诊断了 2 次，因此将 2 次值取均值作为该个体的测量值，y 器械同样计算，并由此计算两个离均差平方和 u，q，p 为离均差交叉乘积和。

Deming 回归与最小二乘估计最大的区别，在于它们对回归系数 b 估计的方法不同。Deming 回归考虑到了 x，y 的残差波动，使得二者残差和达到最小，这点是和最小二乘回归不同的，因此 Deming 回归进行定量指标一致性检验更为合理。

附 Deming 回归图，见图 13-4。

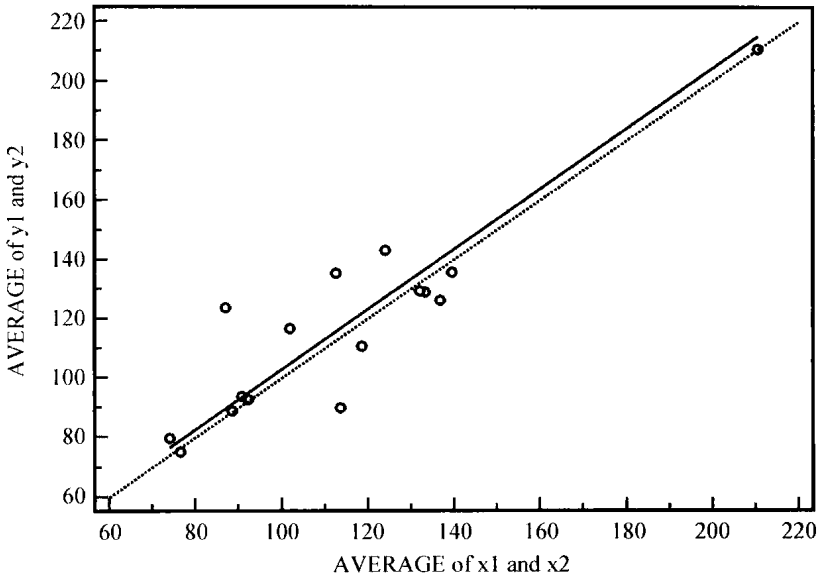

图 13-4　两诊断器械双份测定结果 Deming 回归散布图及拟合直线

（八）Passing-Bablok 回归估计

如异常值较多，可选用 Passing-Bablok 回归，即任取的两点确定直线，屡次反复，得到多条直线并计算斜率值，而后对多个斜率值取中位数并对其调整。该方法假定两诊断器械测定结果随机测定误差遵循同一种分布，标准差比值恒定，样本分布任意。

附 Passing-Bablok 回归图，见图 13-5。

四、诊断试验 ROC 分析

（一）ROC 分析简介

ROC 是受试者工作特征（receiver operating characteristic，ROC）或相对工作特征（relative operating characteristic）的缩写。ROC 分析 20 世纪 50 年代起源于统计决策理论。后来应用于雷达信号观察能力的评价。20 世纪 60 年代中期大量成功地用于实验心理学和心理物理学研究。1971 年，Lusted 描述了如何将心理物理学上常用的受试者工作特征曲线方法，即 ROC 曲线方法用于医学决策，该方法通过包括所有决策值的 ROC 曲线，克服了诊断试验中仅用灵敏度与特异度及其相关指标的缺陷，也是描述诊断试验固有准确度的一种方法。自此以后，该方法变成了非常有价值的描述与比较诊断试验的工具。

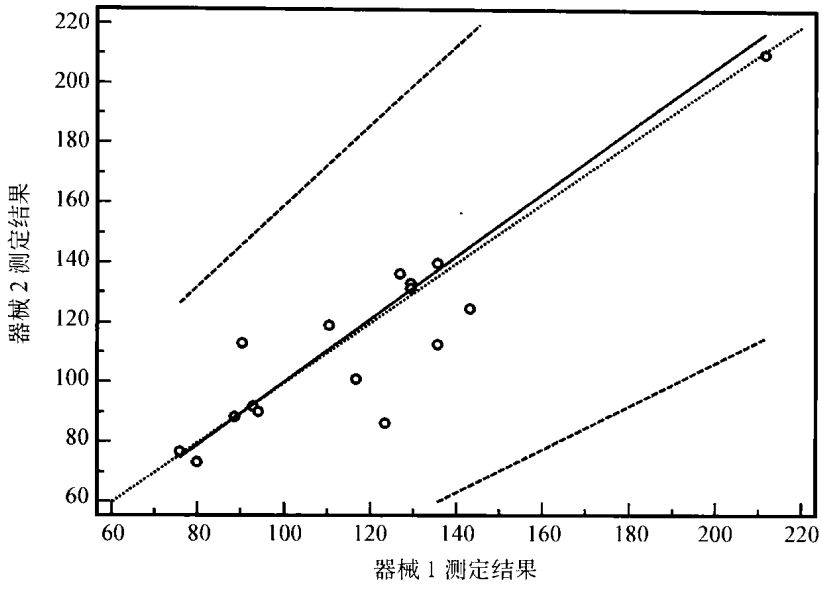

图 13-5　两诊断器械测定结果 Passing-Bablok 回归散布图及拟合直线

(二)诊断试验准确性评价指标的局限性

对于二分类总体,如对照与病例(或无病与有病、正常与异常等),诊断试验结果分别写成阴性与阳性,其资料可列成四格表形式。这时可计算出诊断符合率、灵敏度、特异度等指标。这几个指标均可不同程度反映诊断的准确性。诊断符合率是患者被正确诊断为阳性与非患者被正确诊断为阴性的例数之和占总例数的百分比,它很大程度上依赖患病率,比如患病率为5%时,完全无价值地诊断所有样本为阴性,也可有 95%的诊断符合率。这是其一。其二,诊断符合率没有揭示假阴性和假阳性错误诊断的频率。相同的诊断符合率,假阴性和假阳性的情况可能很不相同。其三,诊断符合率受诊断阈值的限制。更好的方法是计算灵敏度和特异度,它们的值越高,诊断性能越好。灵敏度是患者被正确诊断为阳性的比例,也叫真阳性率(true positive rate,简称 TPR)。特异度是非患者被正确诊断为阴性的比例,也叫真阴性率。(1-特异度)为假阳性率(false positive rate,简称 FPR)。应用灵敏度和特异度这对指标时最明显的问题是:比较两种诊断方法时,可能出现一种诊断方法的灵敏度高,而另一种诊断方法的特异度高,从而难以判断哪一种诊断方法更好的情况。此时,可将灵敏度和特异度相结合,改变诊断阈值,获得多对灵敏度值和(1-特异度)值,即 TPR 值和 FPR 值,绘制 ROC 曲线,做 ROC 分析来解决。

(三)ROC 曲线的构建

ROC 曲线的横轴(x 轴)表示诊断试验的假阳性率(FPR,即误诊率),纵轴(y 轴)表示诊断试验的灵敏度(Se),由不同决策界值产生图中的不同点。采用线段连接图中所有可能界值产生的点,即形成经验 ROC 曲线(empirical ROC curve)。随着灵敏度的增加,FPR 也增加,ROC 曲线正好反映了这种增加的数量大小。

图 13-6 与图 13-7 分别显示了心脏瓣膜影像数据与乳腺 X 线照相数据的 ROC 曲线。图中每个小圆圈表示不同决策界值相应的点(FPR , Se)。

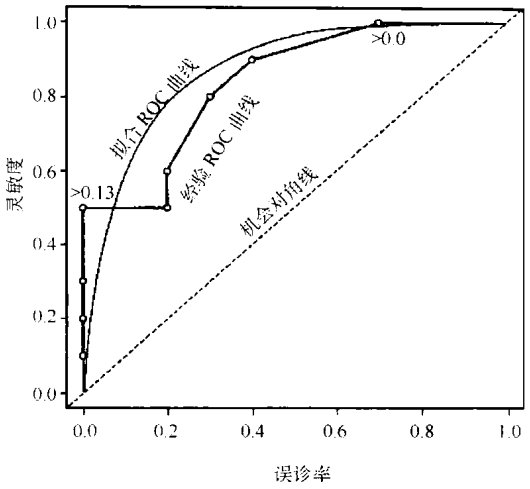

图 13-6 心脏瓣膜影像数据的经验与拟合 ROC 曲线

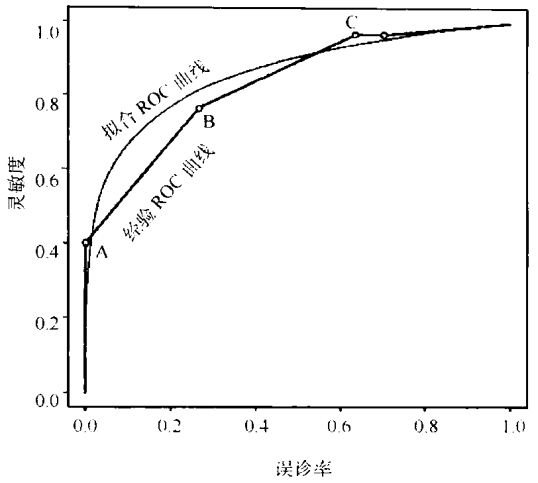

图 13-7 乳腺 X 线照相数据的经验与拟合 ROC 曲线

对患者样本试验结果可方便地拟合统计学模型，图 13-6 和图 13-7 分别显示了心脏瓣膜影像数据与乳腺 X 线照相数据的模型拟合 ROC 曲线(fitted ROC curves)，有时也称为光滑 ROC 曲线，采用的统计学模型是双正态(binormal distribution，即病例组与对照组的试验结果均服从正态分布)模型。诊断医学中通常采用这种模型拟合 ROC 曲线，这种曲线可以用两个参数来表示，一个参数记为 a，是病例与对照试验结果的标准化差值；另一个参数记为 b，是对照与病例试验结果的标准差之比。

ROC 曲线可以全面反映诊断试验的固有准确度。ROC 曲线对诊断的准确性提供了一个直观的视觉印象，描述了相反两种状态间诊断系统的判别能力。曲线上的每一点代表了随着病例诊断阈值(或置信阈)变化的灵敏度与特异度的折中。

(四)ROC 分析的准确性评价指标

一般用“ROC 曲线下面积”反映诊断系统的准确性。理论上，这一指标取值范围为 0.5 至 1.0，完全无价值的诊断 ROC 曲线下面积为 0.5；完善的诊断 ROC 曲线下面积为 1.0。该指标及其标准误的计算目前有非参数、半参数和参数方法。其中得到广泛应用的方法有：Wilcoxon 非参数法和最大似然估计“双正态”参数法。参数法与非参数法估计 ROC 曲线下面积均适用于诊断结果为连续性定量资料或等级资料的诊断试验的评价。

非参数法是根据实验结果直接计算绘制 ROC 曲线所需的工作点，由此绘制的 ROC 曲线为经验 ROC 曲线，曲线下面积与患者和非患者实验结果秩和检验的 Mann-Whitney 统计量相等，但其结果常小于真实的面积值。

参数法是根据试验结果拟合双正态模型，利用最大似然法估计其 2 个参数 a 和 b，由 2 个参数可得到光滑的 ROC 曲线及曲线下面积的估计值。参数法的应用条件为：患者与非患者的试验结果服从双正态分布，但这是指 ROC 曲线的函数形式，而不是指试验结果的基本分布，因为变量变换几乎可以使任何试验结果转换为正态分布，而且在样本量较大、相同值较少时参数法与非参数法估计的 ROC 曲线下面积常常近似相等。双正态模型参数法估计 ROC

曲线下面积的公式为：

$$A=\Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{13-37}$$

$$a=\frac{\bar{x}-\bar{y}}{s_x} \tag{13-38}$$

$$b=\frac{s_y}{s_x} \tag{13-39}$$

式中，A 为 ROC 曲线下面积，a 和 b 分别为双正态模型的 2 个参数，Φ 表示标准正态分布函数，$\bar{x}$ ，$\bar{y}$ 分别为患者组和非患者组检测结果的均数(假定患者组高于非患者组，$\mu_x>\mu_y$)，s_x 和 s_y 分别为患者组和非患者组检测结果的标准差

当 2 条 ROC 曲线无交叉时相关 ROC 曲线下面积的比较可根据曲线下面积、面积的方差及面积间协方差由以下公式计算检验统计量从而得出结论：

$$Z=\frac{A_1-A_2}{\sqrt{Var(A_1)+Var(A_2)-2Cov(A_1,A_2)}} \tag{13-40}$$

式中的 Z 值服从或近似服从标准正态分布，查正态分布表可得相应的 P 值，A_1，A_2 分别为两诊断试验 ROC 曲线下面积，$Var(A_1)$ ，$Var(A_2)$ 分别为两曲线下面积的方差，$Cov(A_1,A_2)$ 为两曲线下面积的协方差，当两诊断试验独立时，此协方差项为 0

两面积的等效性检验根据上述公式，以对照试验 ROC 曲线下面积的 5%为参照，面积的差值与之相比进行统计学检验。当 2 条 ROC 曲线交叉时，两诊断试验的比较应比较部分 ROC 曲线下的面积或固定假阳性率时的灵敏度。

(五)其他与 ROC 分析有关的问题

1. ROC 工作点的置信区间　对于每个 ROC 工作点，可根据二项分布标准误计算公式计算 FPR 和 TPR 的标准误。

$$S_p=\sqrt{\frac{p(1-p)}{n}} \tag{13-41}$$

对于 FPR ，$p=FPR$ ，$n=$非患者组总例数；对于 TPR ，$p=TPR$ ，$n=$患者组总例数。

对于资料中的工作点(FPR ，TPR)=(0.239,0.578)，FPR 和 TPR 的标准误分别为 $S_{FPR}=\sqrt{\frac{0.239(1-0.239)}{234}}$ 和 $S_{TPR}=\sqrt{\frac{0.578(1-0.578)}{166}}$ ，即 0.028 和 0.038；FPR 和 TPR 的 95%的置信区间分别为 $0.239\pm(1.96\times0.028)$ 和 $0.578\pm(1.96\times0.038)$，即0.184～0.294 和 0.504～0.652。

2.“金标准”　对于诊断方法的准确性评价，首先应知道受试者(人、动物或影像等)的真实情况，即哪些是对照组，哪些是病例组。划分它们的标准就是“金标准”。虽然“金标准”不需要十全十美，但是它们应比待评价的诊断方法更可靠，且与待评价的诊断方法无关。

3. 最佳工作点的选择　一般选择阳性似然比或约登指数最大者为最佳工作点。

也有人采用$(C/B)[(1-P)/P]$计算最佳工作点的斜率，表达式中 C，B 和 P 分别表示花费、收益和患病率。在假定对患者组实施治疗，而对非患者组不治疗的前提下，这一表达式表示治疗疾病的花费(cost)和收益(benefit)之比与(1－患病率)和患病率之比的乘积。在 ROC 曲线上，从(FPR ，TPR)=(0,0)到(1,1)，工作点斜率从大到小单调改变。从表达式可以看

出，如果治疗花费多，收益少，或患病率低，则斜率大，最佳工作点接近(0,0)，确保了假阳性率的减少。如果疾病治疗花费少，收益多，或患病率高，则斜率小，最佳工作点接近(1,1)，确保了假阴性率的减少。

（李　卫　李　宁　周晓华　郭　晋）

第14章 直线相关与回归分析

事物之间是互相联系且有内在规律的。对于变量之间的关系，有的可以用函数关系表达，即自变量取某一值时，有一因变量与之完全对应。相关分析(correlation analysis)的任务就是要说明客观事物或现象间的数量关系的密切程度并用适当的统计指标表示出来。而回归分析(regression analysis)的任务则是把客观事物或现象间的数量关系用一定的函数形式表示出来。本章仅对简单相关与直线回归分析进行简要的介绍。

一、线性相关分析的计算

(一)定量资料的 Pearson 直线相关分析

对于两个变量均为连续性变量，可求解 Peason 相关系数，从而判断其相关性，这里要求两个变量近似服从正态分布。

1. *直线相关系数 r 的计算* 通常用 Pearson 乘积矩相关系数(correlation coefficient)来定量地描述线性相关的程度。总体相关系数，习惯上记为 ρ，若 $\rho\neq0$，则称 x 和 y 呈线性相关关系，简称相关；若 $\rho=0$，则称 x 与 y 不呈线性相关关系。进行直线相关分析的两个变量之间无自变量和因变量之分，分析的目的是研究在专业上有一定联系的两个定量变量呈直线关系的密切程度和方向，所用的统计量称为样本相关系数 r，其计算公式见式为：

$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}}\quad(14-1)$$

其中：$l_{XX}=\sum(X-\bar{X})^2=\sum X^2-\frac{(\sum X)^2}{n}$ 表示 X 的离均差平方和

$l_{YY}=\sum(Y-\bar{Y})^2=\sum Y^2-\frac{(\sum Y)^2}{n}$ 表示 Y 的离均差平方和

$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y})=\sum XY-\frac{(\sum X)(\sum Y)}{n}$ 表示 X 与 Y 的离均差积和

相关系数是一个无量纲的统计指标，其取值范围为 $-1\leqslant r\leqslant1$，同样，$-1\leqslant\rho\leqslant$。若 $|r|$ 越接近于0，表明 X 与 Y 呈直线关系的密切程度越低，若 $|r|$ 越接近于1，表明 X 与 Y 呈直线关系的密切程度越高。

2. *直线相关系数 r 的假设检验* 相关系数的大小受数据的对子数和随机误差的影响，当 r 所代表的总体相关系数 $\rho=0$ 时，$|r|$ 可能明显>0，为了尽可能排除抽样误差的影响，较客观地反映出两个变量之间呈直线关系的密切程度，须进行假设检验。

其假设为：$H_0:\rho=0$；$H_1:\rho\neq0$，$\alpha=0.05$

$$t_r=\frac{|r-0|}{S_r}=\frac{|r|}{\sqrt{(1-r^2)/(n-2)}},df=n-2,\nu=n-2 \qquad (14\text{-}2)$$

求出统计量 t_r 的值后，查 t 界值表，下结论的方法与均数比较时所用的 t 检验相同。

若 $r>0$，且检验结果为 $P<0.05$，则认为两个定量变量之间呈正相关关系；若 $r<0$，且检验结果为 $P<0.05$，则可认为两个定量变量之间呈负相关关系。当 r 值接近于零，且 $H_0(\rho=0)$被接受时，认为两变量之间不呈直线关系，但此时不能排除两变量之间可能存在某种曲线关系。

3. *总体相关系数 ρ 的置信区间* 样本相关系数 r 值的分布并非正态，尤其当总体相关系数 ρ 的绝对值较大时，其随机样本 r 值分布之偏态特征更为明显。计算总体相关系数的置信区间时需先对 r 按下式作 z 变换：

$$z=\tanh^{-1}r \quad 或 \quad z=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (14\text{-}3)$$

$$r=\tanh z \quad 或 \quad r=\frac{e^{2z}-1}{e^{2z}+1} \qquad (14\text{-}4)$$

式中，tanh：双曲正切函数；$\tanh^{-1}$：反正切双曲函数

按正态近似原理，z 的 $1-\alpha$ 置信区间按式(14-5)计算。

$$(z-u_\alpha/\sqrt{n-3},z+u_\alpha/\sqrt{n-3}),缩写为\ z\pm u_\alpha/\sqrt{n-3} \qquad (14\text{-}5)$$

最后通过对 z 的 $1-\alpha$ 置信区间按式(14-4)变换为 r 值得置信区间。

(二)定性资料的 Spearman 秩相关分析

1. *秩相关系数 r_s 的计算* 秩相关系数(rank correlation coefficient)又称等级相关系数。基本思想是，对于不符合正态分布的资料，不用原始数据计算相关系数，而是将原始观察值由小到大编秩，然后根据秩次来计算秩相关系数。

设有 n 例观察对象，对每一例观察对象同时取得两个测定值(X_i,Y_i)，分别按 X_i、$Y_i(i=1,2\cdots n)$的值由小到大编秩为 $1,2,3\cdots n$。用 R_{Xi} 表示 X_i 的秩次，R_{Yi} 表示 Y_i 的秩次。因为 n 是固定的，所以总秩相等，即 $\sum R_{X_i}=\sum R_{Y_i}=n(n+1)/2$，$\sum(R_{X_i}-\bar{R}_X)^2=\sum(R_{Y_i}-\bar{R}_Y)^2=(n^3-n)/12$，以及平均秩次 $\bar{R}_X=\bar{R}_Y=(n+1)/2$。但 X_i 的秩顺序不一定与 Y_i 的秩顺序相同，故所对应的 R_{Xi} 与 R_{Yi} 不一定相等。

计算秩相关系数 r_s 的公式为：

$$r_s=\frac{\sum(R_{X_i}-\bar{R}_X)(R_{Y_i}-\bar{R}_Y)}{\sqrt{\sum(R_{X_i}-\bar{R}_X)^2\sum(R_{Y_i}-\bar{R}_Y)^2}} \qquad (14\text{-}6)$$

令同一观察对象的两个秩次差为：

$$d_i=R_{X_i}-R_{Y_i} \quad (i=1,2,3\ldots n) \qquad (14\text{-}7)$$

由式 14-6 及式 14-7 得到计算秩相关系数的简化公式为：

$$r_s=1-\frac{6\sum d^2}{n^3-n} \qquad (14\text{-}8)$$

式中 n 为观察例数，r_s 的取值为 $|r_s|\leqslant 1$，它的解释与直线相关系数 r 一致

2. *秩相关系数的假设检验* r_s 是样本相关系数，对总体秩相关系数 ρ_s 是否为 0 做假设检验，根据样本含量 n 的大小，假设检验的方法有两种：

(1)查表法：当 $n \leqslant 50$ 时，查秩相关系数界值表进行假设检验。

(2)计算法：当 $n > 50$ 时，按式 14-9 计算统计量 t 值：

$$t = \frac{|r_s|}{\sqrt{(1-r_s^2)/(n-2)}}, \quad \nu = n-2 \tag{14-9}$$

根据 t 分布作出推断。Spearman 等级相关系数的 t 检验与直线相关系数的 t 检验是类似的。

二、简单线性回归分析的计算

(一)截距 a 和斜率 b 的计算

进行直线回归分析的两个变量之间一般有自变量和因变量之分，即使在专业上无法区分时，常把容易测量的变量看作自变量，另一个较难测量的变量看作因变量，分析的目的是建立两定量变量之间的回归方程，检验该方程是否成立，并结合专业知识说明该方程是否值得应用以及如何应用。经典的回归分析模型，要求资料符合下列条件。①线性(linear)：即 X 和 Y 之间的关系为线性关系；②独立(independent)：即 n 个个体的观察资料间必须是独立的；③正态(normal)：即给定 X 后，Y 为正态分布，且均数就是回归线上对应于 X 值的点；④等方差(equal variance)：即不同 X 值对应的 Y 分布具有相同的方差，换句话说 Y 的方差与 X 无关。设总体的线性模型为：$Y=\alpha+\beta X+\varepsilon$，$\varepsilon$ 为随机误差。

样本直线回归方程的一般表达式：

$$\hat{Y} = a + bX \tag{14-10}$$

由于变量 X 与变量 Y 的关系具有非确定性，故以 $\hat{Y}$ 表示回归方程所求得的估计值，它是当 X 固定时，Y 的总体均数 μY，X 的估计值。式中的 a，b 是决定回归直线的两个参数，分别为 α，β 的估计值。a 为回归直线在 Y 轴上的截距(intercept)；b 为回归系数(regression coefficient)，即回归直线的斜率(slope)。根据最小平方法(或最小二乘法)原理，可导出计算 a，b 的公式。见式 14-11 和式 14-12。

$$b = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sum (X-\bar{X})^2} = \frac{l_{XY}}{l_{XX}} \tag{14-11}$$

$$a = \bar{Y} - b\bar{X} \tag{14-12}$$

如果根据专业知识需求过定点 (x_0, y_0) 的直线回归方程，式中 $\bar{x}$ 取 x_0，$\bar{y}$ 取 y_0；其特殊情况就是求过原点(0,0)的直线回归方程，式中 $\bar{x}$ 取 0、$\bar{y}$ 取 0。

(二)截距 a 和斜率 b 的假设检验

与需要对相关系数 r 进行假设检验的理由相同，对斜率和截距也需做检验。

对 β(总体斜率)做检验的假设和方法如下：

H_0：$\beta=0$；H_1：$\beta\neq 0$；$\alpha=0.05$。

$$t_b = \frac{|b-0|}{S_b} = \frac{|b|}{S_{y.x}/\sqrt{l_{xx}}} \tag{14-13}$$

$$S_{y.x} = \sqrt{\frac{\sum (y_i-\hat{y})^2}{n-2}} \tag{14-14}$$

$$l_{xx} = \sum (x_i-\bar{x})^2 \tag{14-15}$$

式(14-13)所对应的自由度 df=n-2,S_b为 b 的标准误。

上述式(14-14)中 $S_{y.x}$称为剩余标准差,是排除了 x 的影响后,单独 y 方面的变异大小,常用它作为预报精确度的标志。因为它的单位与 y 一致,最容易在实际中进行比较和检验,所以,一个回归能否对解决实际问题有所帮助,只要比较 $S_{y.x}$与允许的偏差就行,故它是检验一个回归是否有效的极其重要的标志。

与对斜率检验等价的还有一种常用的方法:即对回归方程是否具有统计学意义作方差分析。其基本思想是:计算出 y 的总离均差平方和 SS_T,由回归所能解释的离均差平方和 SS_R,它们的差值就是回归所无法解释的量,称为误差,记为 SS_E,然后,用回归的均方除以误差的均方,构造出 F 统计量,进而根据 F 分布推断出所求的直线回归方程是否有统计学意义。关于 SS_T,SS_R,SS_E的计算公式见式(14-16～14-18)。

$$SS_T = \sum (y_i - \bar{y})^2 \tag{14-16}$$

$$SS_R = \sum (\hat{y}_i - \bar{y})^2 \tag{14-17}$$

$$SS_E = SS_T - SS_R \tag{14-18}$$

$df_T = n-1$,$df_R = 1$,$df_E = n-2$。

有了上述统计量,就可引入一个与相关系数 r 有关的统计量——决定系数 r^2。可以证明:$SS_R = r^2 SS_T$,于是,得:

$$r^2 = \frac{SS_R}{SS_T} \tag{14-19}$$

这说明决定系数 r^2就是回归的离均差平方和占 y 的总离均差平方和的百分比,它即建立了相关与回归之间的联系,又通过具体的数量大小反映了回归的贡献大小,这是回归分析中一个十分有用的统计量。有时,由于样本含量 n 很大,即使算得的相关系数 r 的绝对值较小,也能得出检验结果具有统计学意义,此时,计算一下决定系数 r^2,若发现 r^2的值远远<0.5,说明由 x 的变化所引起 y 的变化部分不足 50%,此时若求出 y 依 x 变化的直线回归方程,就没有多大的实用价值了。

对 α(总体截距)做检验的假设和方法如下。

H_0:$\alpha=0$;H_1:$\alpha\neq 0$;α(显著性水平)=0.05。

$$t_a = \frac{|a-0|}{S_a} = \frac{|a|}{S_{y.x}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}} \tag{14-20}$$

式(14-20)中的 S_a为 α 的标准误,$S_{y.x}$,l_{xx}与式(14-14～14-15)中相应符号的含义相同。

(三)总体截距 a 和斜率 b 的置信区间

总体中的截距 α,斜率 β 的 100(1−α)%置信区间为:

$$a - t_{\alpha(n-2)} S_a \leqslant \alpha \leqslant a + t_{\alpha(n-2)} S_a \tag{14-21}$$

$$b - t_{\alpha(n-2)} S_b \leqslant \beta \leqslant b + t_{\alpha(n-2)} S_b \tag{14-22}$$

(四)简单线性回归分析中其他有关的区间估计问题

若记 $\mu_{y|x=x_0}$ 为给定 $x=x_0$ 条件下 y 的总体均数,则它的 100(1−α)%置信区间可按式(14-23)和式(14-24)计算。

$$\hat{y} - t_{\alpha(n-2)} S_{\hat{y}} \leqslant \mu_{y|x=x_0} \leqslant \hat{y} + t_{\alpha(n-2)} S_{\hat{y}} \tag{14-23}$$

$$S_{\hat{y}} = S_{y.x}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{14-24}$$

在给定 $x=x_0$ 条件下，y 的个体值的近似 $100(1-\alpha)\%$ 容许区间可按式(14-25)和式(14-26)计算，这就解决了对因变量 y 进行预报的问题。

$$(\hat{y}-t_{\alpha(n-2)}S_y, \hat{y}+t_{\alpha(n-2)}S_y) \tag{14-25}$$

$$S_y = S_{y.x}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{14-26}$$

三、相关 SAS 语句与程序

CORR 过程(相关过程)存在于 SAS 软件 base 模块，用于计算变量间的相关系数，它可以计算 Pearson 积矩相关系数、Spearman 秩相关系数、Kendall 的 tau-b(τ_b)统计量、Hoeffding 的相关性度量 D 以及 Pearson，Spearman，以及 Kendall 偏相关系数。另外，它还可以计算用于估计可靠性的 Cronbach 系数 α 。

CORR 过程语句格式如下：

```
PROC CORR < options >;
BY variables ;
FREQ variable ;
PARTIAL variables ;
VAR variables ;
WEIGHT variable ;
WITH variables ;
RUN;
```

(1)PROC CORR 语句：PROC CORR 语句调用 CORR 过程(相关分析过程)，只有 PROC CORR 语句为 CORR 过程所必需，其他语句均为可选语句(表 14-1)。

表 14-1　PROC CORR 语句选项

功能	选项
指定数据集类型	
指定即将要分析的数据集	DATA=
输出含有 Hoeffding's D 统计量的数据集	OUTH=
输出含有 Kendall 相关统计量的数据集	OUTK=
输出含有 Pearson 相关统计量的数据集	OUTP=
输出含有 Spearman 相关统计量的数据集	OUTS=
调节统计分析类型	
将分析中 weight 变量取值为非正数的观测排除	EXCLNPWGT
将分析中含有缺失值的观测排除	NOMISS
计算并输出 Hoeffding's D 统计量	HOEFFDING
计算并输出 Kendall's tau-b 系数	KENDALL
计算并输出 Pearson 积矩相关系数	PEARSON

（续表）

功能	选项
计算并输出 Spearman 秩相关系数	SPEARMAN
使用 Fisher z 变换计算 Pearson 相关系数	FISHER PEARSON
使用 Fisher z 变换计算 Spearman 秩相关系数	FISHER SPEARMAN
控制 Pearson 相关系数统计量	
计算并输出 Cronbach 系数 α	ALPHA
计算并输出协方差矩阵	COV
计算并输出校正的离均差平方和及离均差积和矩阵	CSSCP
依据 Fisher z 变换计算校正统计量	FISHER
设置变量奇异性标准	SINGULAR=
计算并输出离均差平方和及离均差积和矩阵	SSCP
指定方差计算的分母	VARDEF=
控制打印输出	
将相关系数以绝对值大小降序排列输出	BEST=
不输出 Pearson 相关系数	NOCORR
所有结果均不输出	NOPRINT
不输出相关系数对应的 P 值	NOPROB
不输出描述性统计量	NOSIMPLE
将所有相关系数排序输出	RANK

(2)BY 语句：指定分组变量。同 PROC CORR 一起使用能够获得用 BY 变量定义的分组观测的独立分析结果。

(3)FREQ 语句：指定作为观测频数的变量。

(4)PARTIAL 语句：对指定的变量计算偏相关系数或偏统计量，可计算 Pearson 偏相关、Spearman 偏秩序相关、Kendall 偏 tau-b，使用该语句可指定一个或多个变量名称。当语句中设置了 HOEFFDING 选项时，partial 语句不起作用。

(5)VAR 语句：指定待分析变量，即指定要计算相关系数的变量。

(6)WEIGHT 语句：计算加权的乘积矩相关系数，用该语句指定权数变量名称。该语句仅用于 Pearson 相关，对于选项 SPEARMAN，KENDALL，HOEFFDING 均无效。

(7)WITH 语句：得到变量间特殊组合的相关，该语句与 VAR 语句共同使用。当有 WITH 语句存在时，VAR 变量之间不进行相关分析，而是在每个 var 变量和每一个 with 变量之间进行相关分析。用 VAR 语句列出的变量放在输出相关阵的上方，而用 WITH 语句列出的变量竖在相关阵的左边。

(8)相关 SAS 程序 CT14-1

程序	说明
DATA CT14-1;	建立临时数据集 CT14-1
inputx y @@;	以下输入两个变量
cards;	
/* 输入 x,y 变量值 */;	
PROC GPLOT data=CT14-1;	
ploty* x ='s';	绘制散点图
run;	
PROC CORR data=CT14-1;	
var x y;	进行相关分析
run;	
PROC REG data=14-1;	
model y=x;	直线回归分析
run;	

（郭　晋）

第15章 多重线性回归分析

在医疗器械临床研究中，器械置入或用于受试对象，器械的疗效除了与受试人群使用的器械种类这一影响因素有关之外，可能还受到受试者年龄、性别、遗传因素、饮食、吸烟、饮酒、肥胖等多个因素的影响，尽管 RCT 研究尽可能使两组受试对象的这类信息均衡，然而仍然有可能某些基线信息存在差别。从流行病学的角度，这类因素称之为混杂因素，我们应排除此类因素的干扰，准确判断不同种类的器械在受试人群中的疗效是否存在差别，此时我们需要一种回归分析的方法能够建立起一个因变量与多个自变量之间的关系，它既区别于单因素的差异性分析，又和简单相关回归分析（研究一个自变量与一个因变量的关系）不同。如果因变量为连续性变量，可运用多重线性回归分析；如果因变量为离散型变量（定性变量），可利用 logistic 回归分析。本章将介绍多重线性回归分析的概念，假设检验，SAS 运用等相关内容。

一、多重线性回归模型概念

设因变量为 y，自变量为 x_1、x_2…x_p，建立起一个因变量与多个自变量的关系，即多重线性回归的模型：

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

成为 p 重线性回归。$\hat{y}$ 为 y 的预测值（predicted value），表示各个自变量赋值后 y 的估计值；α 为截距；β_i 为 y 的偏回归系数（partial regression coefficient），它表示在其他自变量固定不变的情况下，x_i 没改变一个测量单位时所引起的因变量 y 的平均改变量。

线性回归方程也可以表示为：

$\hat{y} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon$

方程中 ε 之前的部分为自变量决定的部分，即预测值，ε 称为残差，是不能由现有自变量决定的部分。

线性回归方程可以矩阵表示：

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad B = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

回归方程的矩阵形式为：

$$Y = XB + E = \hat{Y} + E$$

$$\hat{Y} + XB$$

数据进行多重线性回归分析，需要满足独立性、正态性、方差齐性，此外自变量与因变量关系需呈线性关系。

二、回归系数的估计与假设检验

(一)回归系数的计算

拟求解回归系数 β_1，$\beta_2 \cdots \beta_p$ 的估计值 b_0，$b_1 \cdots b_p$，在多重线性回归分析中，对回归系数的估计仍然使用最小二乘法(least square，LS)。即求得残差平方和(sum of squares for residual)。

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-b_2x_{i2}-\cdots-b_px_{ip})^2$$

达到最小。回归系数需满足：

$$\frac{\partial Q}{\partial b_0}=0,\ \frac{\partial Q}{\partial b_1}=0\cdots\frac{\partial Q}{\partial b_p}=0$$

该方程组也称为正规方程组，可化为：

$$\begin{cases} nb_0+\sum x_{i1}b_1+\sum x_{i2}b_2+\cdots+\sum x_{ip}b_p=\sum y_i \\ \sum x_{i1}b_0+\sum x_{i1}^2b_1+\sum x_{i1}x_{i2}b_2+\cdots+\sum x_{i1}x_{ip}b_p=\sum x_{i1}y_i \\ \sum x_{i2}b_0+\sum x_{i1}x_{i2}b_1+\sum x_{i2}^2b_2+\cdots+\sum x_{i2}x_{ip}b_p=\sum x_{i2}y_i \\ \cdots\cdots \\ \sum x_{ip}b_0+\sum x_{i1}x_{ip}b_1+\sum x_{i2}x_{ip}b_2+\cdots+\sum x_{ip}^2b_p=\sum x_{ip}y_i \end{cases}$$

正规方程左端的系数用矩阵表示为：

$$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}=XX'$$

右端的参数项可以写为：

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}=YX'$$

故可表示为：

$$XX'B=X'Y$$

其解可以表示为：

$$B=(X'X)^{-1}X'Y$$

其中 $(X'X)^{-1}$ 表示系数矩阵 $X'X$ 的逆矩阵。回归系数的求解非常复杂，可利用软件计算。

(二)回归方程的假设检验

可利用方差分析方法来检验因变量 y 与 p 个自变量之间是否存在线性回归关系。因变量总的离均差平方和 $SS_{总}$ 可分解为回归平方和 $SS_{回}$ 与残差平方和 $SS_{残}$ 两部分，他们的计算方法与自由度分别为：

$SS_{总}=\sum_{i=1}^{n}(y_i-\bar{y})^2$，自由度 $\nu_{总}=n-1$

$SS_{回}=\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$，自由度 $\nu_{回}=p$

$SS_{残}=\sum_{i=1}^{n}(y-\hat{y})^2$，$SS_{总}-SS_{回}$，自由度 $\nu_{残}=n-p-1$

检验假设为：

H_0：总体中各偏回归系数均为 0。

H_1：总体中偏回归系数不为 0 或不全为 0。

统计量 F 的计算为：

$$F=\frac{SS_{回}/p}{SS_{残}/(n-p-1)}=\frac{MS_{回}}{MS_{残}}，\nu_{回}=p，\nu_{残}=n-p-1$$

$MS_{回}$ 与 $MS_{残}$ 分别为回归均方与残差均方。

(三)回归系数的假设检验

在对方程进行假设检验之后，仍有必要对每个自变量进行检验，检验每个偏回归系数是否为 0。

检验假设为：

H_0：$\beta_j=0$；H_1：$\beta_j\neq0$

运用 t 检验，统计量为：

$$t_j=\frac{b_j}{s_{b_j}}，(j=1,2,\cdots,p)，\nu=n-p-1$$

其中，s_{b_j} 为回归系数 b_j 的标准误

$$s_{b_j}=s_{y\cdot12\cdots p}\sqrt{c_{ii}}$$

c_{ii} 是矩阵 $(X'X)^{-1}$ 对角线上对应于 x_i 的元素

三、回归变量的选择

(一)回归变量的选择标准

1. 复相关系数 R　复相关系数 R(multiple correlation coefficient)反映模型的拟合优度，其值越大越好。R^2 为决定系数(determination coefficient)表示回归 SS 占总的 SS 的比重。

$$R^2=\frac{SS_{回}}{SS_{总}}=1-\frac{SS_{残}}{SS_{总}}$$

复相关系数 R 系数的特点是随着方程中纳入的自变量个数的增加，复相关系数总是增加的，即使增加的变量无统计学意义，这显然不是足够准确的。

2. 修正复相关系数　修正复相关系数 R_{adj}(adjusted multiple correlation coefficient)的定义如下：

$$R^2_{adj}=R^2-\frac{p(1-R^2)}{n-p-1}$$

R^2 为决定系数，n 为样本含量，p 为自变量个数，由上式可知，R_{adj} 不一定随着决定系数 R^2 的值增大而增大。

3. C_p 准则　C_p 统计量的定义是：

$$C_p=(n-p-1)\left(\frac{MS_{残,p}}{MS_{残,全部}}-1\right)+(p+1)$$

式中，$MS_{残,p}$ 为 p 个自变量残差平方和，$MS_{残,全部}$ 为全部自变量作回归的残差均方，p 为包括常数项在内的自变量个数。在实际计算中，可以选择 C_P 值小的模型作为最佳回归模型

4. AIC 准则　AIC 准则(Akaike's information criterion，AIC)是日本学者赤池于 1973 年提出的，广泛应用于多重回归、广义线性回归中自变量的筛选，以及非线性模型的比较。它的定义为：

(1)当模型是用最小二乘法估计时：

$$AIC = n\ln((n-p)/n \times s^2_{y x_1 x_2 \cdots x_p}) + 2p$$

(2)当模型或方程是用极大似然法估计时：

$$AIC = -2\ln(L) + 2p$$

式中，p 为模型中参数的个数，L 是模型的极大似然函数，n 为样本量

AIC 有两部分组成，前一部分反映了回归方程的拟合精度，值越小越好；后一部分反映了回归中变量的个数的多少，即参数越少越好。因此，AIC 值越小，模型可作为最佳模型。

(二)自变量的筛选

线性独立的若干个自变量对因变量的贡献会不尽相同，有些对因变量的影响可能很小。因此，需要对自变量进行筛选。在 SAS 中共有 9 种筛选方法，分别为全模型法(NONE)、向前选择法(FORWARD)、向后选择法(BACKWARD)、逐步选择法(STEPWISE)、基于最大 R^2 增量法(MAXR)、基于最小 R^2 增量法(MINR)、基于 R^2 数值大小的选择变量法(RSQUARE)、基于校正 R^2 数值大小的选择变量法(ADJRSQ)、基于 Mallow 's C_P 统计量数值大小的选择变量法(CP)。这些筛选变量的方法的基本思想如下。

1. *向前选择法*(FORWARD)　事先给定一个入选标准，即 I 类错误的概率 α。当 P 小于 SLENTRY(REG 过程中规定的选变量进入方程的检验水平 α 则该变量入选，否则不能入选；当回归方程中变量少时某变量不符合入选标准，但随着回归方程中变量逐次增多时，该变量就可能符合入选标准；这样直到没有变量可入选为止。SLENTRY 缺省值定为 0.5，亦可定为 0.2 到 0.4，如果自变量很多，此值还应取得更小一些，如让 SLENTRY=0.05。

向前选择法的局限性：SLENTRY 取值小时，可能没有一个变量能入选；SLENTRY 大时，开始选入的变量后来在新条件下不再进行检验，因而不能剔除后来变得无统计学意义的变量。

2. *向后消去法*(BACKWARD)　从模型语句中所包含的全部自变量开始，计算留在回归方程中的各个自变量所产生的 F 统计量和 P 值，当 P 小于 SLSTAY(程序中规定的从方程中剔除变量的检验水准)则将此变量保留在方程中，否则，从最大的 P 值所对应的自变量开始逐一剔除，直到回归方程中没有变量可以被剔除时为止。SLSTAY 缺省值为 0.10，欲使保留在方程中的变量都在 $\alpha=0.05$ 水平上有统计学意义时，应让 SLSTAY=0.05。

程序能运行时，因要求所选自变量的子集矩阵满秩，所以当观测点少、且自变量过多时程序会自动从中选择出观测点数减 1 个变量。

向后消去法的局限性：SLSTAY 大时，任何一个自变量都不能被剔除；SLSTAY 小时，开始被剔除的自变量后来在新条件下即使变得对因变量有较大的贡献了，也不能再次被入选回归方程并参与检验。

3. *逐步筛选法*(STEPWISE)　此法是向前选择法和向后消去法的结合。回归方程中的变量从无到有像向前选择法那样，根据 F 统计量按 SLENTRY 水平决定该自变量是否入选；当回归方程选入自变量后，又像向后消去法那样，根据 F 统计量按 SLSTAY 水平剔除无统计

学意义的各自变量，依次类推。这样直到没有自变量可入选，也没有自变量可被剔除或入选的自变量就是刚被剔除的自变量，则停止逐步筛选过程。

逐步筛选法比向前选择法和向后消去法都能更好地选出变量构造模型，但也有它的局限性：其一，当有 m 个变量入选后，选第 $m+1$ 个变量时，对它来说，前 m 个变量不一定是最佳组合；其二，选入或剔除自变量仅以 F 值作标准，完全没考虑其他标准。

4. 最大 R^2 增量法（MAXR） 首先找到具有最大决定系数 R^2 的单变量回归模型，其次引入产生最大 R^2 增量的另一变量。然后对于该两变量的回归方程，用其他变量逐次替换，每一次替换都计算其 R^2，如果换后的回归方程能产生最大 R^2 增量，即为两变量最优回归方程，如此再找下去，直到入选自变量数太多，使设计矩阵不再满秩时为止。

其实它也是一种逐步筛选法，只是筛选变量所用的准则不同，不是用 F 值，而是用决定系数 R^2 判定自变量是否入选。因它不受 SLENTRY 和 SLSTAY 的限制，总能从变量中找到相对最大者，这就克服了用本节筛选法 1～3 法时的一种局限性：即找不到任何自变量可进入回归方程的情况。

本法与本节第 3 种方法都是逐步筛选变量法，每一步选进或剔除自变量都是只限于一个，因而二者局限性也相似：第一，当有 m 个变量入选后，选第 $m+1$ 个变量时，对它来说，前 m 个变量不一定是最佳组合；第二，选入或剔除自变量仅以 R^2 值作标准，完全没考虑其他标准。

5. 最小 R^2 增量法（MINR） 首先找到具有最小决定系数 R^2 的单变量回归方程，然后从其余自变量中选出一个自变量，使它构成的回归方程比其他自变量所产生的 R^2 增量最小，不断用新变量来替换老变量，依次类推，这样就会顺次列出全部单变量回归方程，依次为：（R^2 最小者、R^2 增量最小者、R^2 增量次小者…R^2 增量最大者）所对应的含一个自变量的回归方程，最后一个为单变量最佳回归方程；两变量最小 R^2 增量的筛选类似本节第 4 种方法，但引入的是产生最小 R^2 增量的另一变量。对该两变量的回归方程，再用其他变量替换，换成产生最小 R^2 增量者，直至 R^2 不能再增加，即为两变量最优回归方程。依次类推，继续找含 3 个或更多自变量的最优回归方程，等等，变量有进有出。

它与本节第 4 种方法选的结果不一定相同，但它在寻找最优回归方程过程中所考虑的中间回归方程要比本节第 4 种方法多。

本法的局限性与本节第 3 及 4 种方法相似：第一，当有 m 个变量入选后，选第 $m+1$ 个变量时，每次只有 1 个自变量进或出，各自变量间有复杂关系时，就有可能找不到最佳组合；第二，选入自变量或替换自变量仅以 R^2 值作标准，完全没有考虑其他标准。

在 SAS 编程法中，通过在模型语句中增加适当的选择项，可以有 8 种筛选变量的方法，其关键词分别为：STEPWISE（逐步回归法）、BACKWARD（向后剔除变量法）、FORWARD（向前选择变量法）、MAXR（基于最大 R^2 增量法）、MINR（基于最小 R^2 增量法）、RSQUARE（基于 R^2 数值大小的选择变量法）、ADJRSQ（基于校正 R^2 数值大小的选择变量法）、C_P（基于 Mallow C_P 统计量数值大小的选择变量法）；而在非编程模块 SAS/ASSIST 和 SAS/AA 中，也提供了这 8 种筛选变量的方法。用前 5 种方法筛选变量后，一般都会给出回归方程中参数的估计值，但用后 3 种方法筛选变量后，一般只给出仅含 1 个自变量、仅含 2 个自变量…仅含 m 个自变量的“最优”变量组合，故后 3 种筛选变量的方法可统称为求“最优回归子集”的方法。此时，欲得到回归参数的估计值，需给定变量的组合，按不筛选变量法直接拟合多重线性回归方程。

四、回归诊断

当自变量存在多重共线性时，即当一个(或几个)自变量可以由另外的自变量线性表达时，由数理统计知识可以知道在运用最小二乘法对回归系数进行估计时，必然使回归系数的估计变得不确定，变量间即使不是完全的共线性，但有近似的共线性时(如相关系数接近于 1)，将使最小二乘法原理失效，使得回归方程中参数变得不确定，因此对回归系数的估计、预测值的精度产生很大的影响。当变量间有共线性关系时，需要对自变量之间是否存在较强的共线性关系进行研究，称为共线性诊断(the diagnosis of collinearity)。在使用 SAS 软件的 REG 过程时，通过在模型语句中增加选择项“COLLIN”和“COLLINOINT”来实现共线性诊断。当截距项的假设检验结果无统计学意义时，可用“COLLIN(未对截距项进行校正)”输出的诊断结果为判断有无共线性的依据；当截距项的假设检验结果有统计学意义时，应看由“COLLINOINT(对截距项进行了校正)”输出的诊断结果。在非编程模块 SAS/ASSIST 和 SAS/INSIGHT 中，也有相应的选择项来实现这一目的。

若个别观测点与多数观测点偏离很远或因过失误差(如抄写或输入错误所致)，它们也会对回归方程的质量产生极坏的影响。对这两方面的问题进行监测和分析的方法，称为异常点诊断(the diagnosis of outlier)。

下面就 SAS 系统的 REG 过程运行后不同输出结果，仅从回归诊断方面理解和分析说明如下。

1. *用条件数和方差分量来进行共线性诊断*　各入选变量的共线性诊断借助 SAS 的模型语句的选择项 COLLIN 或 COLLINOINT 来完成。二者都给出信息矩阵的特征根和条件数(condition number)，还给出各变量的方差在各主成分上的分解(decomposition)所得的分量，以百分数的形式给出，每个入选变量上的方差分量之和为 1。COLLIN 和 COLLINOINT 的区别在于后者对模型中截距项作了校正。当截距项无统计学意义时，看由 COLLIN 输出的结果；反之，应看由 COLLINOINT 输出的结果。

条件数：先求出信息矩阵 $X'X$ 的各特征根，条件指数(condition indices)定义为：最大特征根与每个特征根比值的平方根，其中最大条件指数 k 称为矩阵 $X'X$ 的条件数。最大的条件数大，说明设计矩阵有较强的共线性，使结果不稳定，甚至使离开试验点的各估计值或预测值毫无意义。直观上，条件数度量了信息矩阵 $X'X$ 的特征根散布程度，可用来判断多重共线性是否存在以及多重共线性严重程度。在应用经验中，若 $0 \leqslant k < 10$，则认为没有多重共线性；$10 \leqslant k \leqslant 30$，则认为存在中等程度或较强的多重共线性；$k > 30$，则认为存在严重的多重共线性。

方差分量：强的多重共线性同时还会表现在变量的方差分量上：对最大的条件数同时有 2 个以上自变量的方差分量超过 50%，就意味这些自变量间有一定程度的相关性。

2. *用方差膨胀因子来进行共线性诊断*　容许度(tolerance，在 Model 语句中的选择项为 TOL)：对一个入选变量而言，该统计量等于 $1-R^2$，这里 R^2 是把该自变量当做因变量对回归方程中所有其余自变量的决定系数，R^2 大(趋于 1)，则 $1-R^2=$TOL 小(趋于 0)，容许度小，表明该自变量与其他自变量之间的关系密切。

方差膨胀因子(variance inflation factor，VIF)：VIF＝1/TOL，对于不好的试验设计，VIF 的取值可能趋于无限大。VIF 达到什么数值就可认为自变量间存在共线性？尚无明确的临界值。有人根据经验得出：VIF＞5 或 10 时，就有严重的多重共线性存在。

3. 用学生化残差对观测点中的强影响点进行诊断　对因变量的预测值影响特别大，甚至容易导致相反结论的观测点，被称为强影响点(influence case)或称为异常点(outlier)。有若干个统计量(如：Cook 's D 距离统计量、h_i统计量、STUDENT 统计量、RSTUDENT 统计量等，这些统计量(其定义较复杂，此处从略)可用于诊断哪些点对因变量的预测值影响大，其中最便于判断的是学生化残差 STUDENT 统计量。当该统计量的绝对值>2 时，所对应的观测点可能是异常点，此时，需认真核对原始数据。若属抄写或输入数据时人为造成的错误，应当予以纠正；若属非过失误差所致，可将异常点剔除前后各作出一个最好的回归方程，并对所得到的结果进行分析和讨论。如果有可能，最好在此点上补做试验，以便进一步确认可疑的"异常点"是否确属异常点。

4. 用残差分解法对观测的独立性进行诊断　如果各观测不独立，将会对参数的估计有很大的影响，因此，如果对实验设计没能很好的把握，有可能会对同一个样品进行重复的观测，此时应对其进行假设检验。其原理是将残差进行分解。假设有 n_i个重复观测 $Y_{i1}\cdots\cdots Y_{im}$，那么对于因变量就有一个平均值 $\bar{Y}_i$ 和一个预测值 $\hat{Y}_i$，对它们的残差进行分解：

$$\sum_i \sum_{j=1}^{ni} (Y_{ij} - \hat{Y}_i)^2 = \sum_i \sum_{j=1}^{ni} (Y_{ij} - \bar{Y}_i)^2 + \sum_i n_i (\bar{Y}_{ij} - \hat{Y}_i)^2$$

式的左边为总的方差，右边的第一项为由因变量的平均值造成的误差，第二项为偏差造成的误差，如果这些观测对此回归方程是适当的，那么这两个方差的差别是很小的。否则，第二项误差会比第一项大很多，这时所进行的假设检验就会有统计学意义。在 SAS 9.2 当中进行此项假设检验时应在模型语句中加入 lackfit 选项。而在 SAS 9.2 以前的版本中则没有此选项。

【例 15-1】 本临床试验采用多中心、单盲、随机对照方法，验证 NOYA 雷帕霉素可降解涂层钴铬合金冠状动脉药物洗脱支架系统(试验组)的有效性和安全性，评价该支架的输送系统。与上海微创公司的 Firebird2 药物洗脱支架(对照组)进行对照，每组 30 例。主要指标为术后病变阶段内直径狭窄程度(%)，假定两组人群的基线信息、疾病史可能不相同，并且可能对直径狭窄程度产生影响，试建立回归方程，比较两组人群的直径狭窄程度的影响因素。x_1：受试组别(0：对照组；1：试验组)，x_2：年龄，x_3：bmi，x_4：糖尿病史(0：无，1：有)，x_5：高血压史(0：无，1：有)。数据如表 15-1 所示。

表 15-1　60 例受试对象基线信息、疾病史及术后病变阶段内直径狭窄程度(%)

id	x_1	x_2	x_3	x_4	x_5	y	id	x_1	x_2	x_3	x_4	x_5	y
1	0	61	20.6	0	1	24.24	31	1	47	26.6	1	0	11.53
2	0	49	19.8	0	0	30.41	32	1	46	24.4	0	0	16.69
3	0	48	27.0	0	1	23.89	33	1	78	28.5	0	1	16.10
4	0	47	29.1	0	0	22.64	34	1	47	25.7	0	1	12.81
5	0	53	23.6	1	1	25.44	35	1	67	27.8	0	1	14.48
6	0	39	25.7	1	0	26.30	36	1	56	21.2	1	0	20.44
7	0	52	30.0	0	0	24.62	37	1	75	27.3	0	0	34.67
8	0	61	23.2	0	1	19.51	38	1	47	23.9	0	1	17.33
9	0	67	23.7	0	0	22.29	39	1	57	20.7	0	1	6.61
10	0	51	28.7	0	1	31.67	40	1	66	20.6	0	1	20.13

（续　表）

id	x_1	x_2	x_3	x_4	x_5	y	id	x_1	x_2	x_3	x_4	x_5	y
11	0	42	19.9	0	1	26.40	41	1	57	23.0	1	0	14.54
12	0	57	27.7	0	0	27.00	42	1	64	20.2	0	0	15.15
13	0	66	28.4	0	1	25.11	43	1	59	20.7	0	1	15.40
14	0	64	21.0	1	1	20.25	44	1	39	17.6	0	1	17.88
15	0	44	24.7	0	0	17.94	45	1	40	21.5	0	0	9.06
16	0	51	28.1	1	0	18.60	46	1	63	22.5	1	0	6.61
17	0	67	19.1	0	1	18.12	47	1	77	24.2	1	1	20.24
18	0	61	24.9	0	0	16.48	48	1	62	19.7	0	0	5.00
19	0	38	22.7	1	1	24.97	49	1	58	24.2	0	0	19.32
20	0	69	29.2	0	1	32.49	50	1	70	28.9	0	1	21.48
21	0	63	26.1	0	0	26.81	51	1	50	23.2	0	1	18.02
22	0	71	23.9	0	0	29.53	52	1	56	24.6	0	1	11.16
23	0	59	22.2	1	1	24.49	53	1	70	26.7	1	0	19.62
24	0	48	26.8	0	0	20.87	54	1	37	22.0	0	0	18.77
25	0	53	33.1	0	1	21.82	55	1	49	25.4	0	0	23.30
26	0	77	27.8	1	1	17.71	56	1	51	25.1	1	1	18.96
27	0	59	29.7	0	0	36.21	57	1	60	29.9	0	0	9.15
28	0	70	24.1	1	1	23.93	58	1	66	26.5	1	0	14.06
29	0	67	25.8	0	0	28.11	59	1	55	26.1	0	0	11.85
30	0	67	26.2	1	0	28.80	60	1	56	25.2	1	1	17.14

SAS 程序如 CT15-1：

```
DATA CT15-1;
input id x1-x5 @@ y;
cards;
1  0  61  20.6  0  1  24.24  31  1  47  26.6  1  0  11.53
2  0  49  19.8  0  0  30.41  32  1  46  24.4  0  0  16.69
3  0  48  27.0  0  1  23.89  33  1  78  28.5  0  1  16.10
4  0  47  29.1  0  0  22.64  34  1  47  25.7  0  1  12.81
5  0  53  23.6  1  1  25.44  35  1  67  27.8  0  1  14.48
6  0  39  25.7  1  0  26.30  36  1  56  21.2  1  0  20.44
7  0  52  30.0  0  0  24.62  37  1  75  27.3  0  0  34.67
8  0  61  23.2  0  1  19.51  38  1  47  23.9  0  1  17.33
9  0  67  23.7  0  0  22.29  39  1  57  20.7  0  1  6.61
10 0  51  28.7  0  1  31.67  40  1  66  20.6  0  1  20.13
11 0  42  19.9  0  1  26.40  41  1  57  23.0  1  0  14.54
12 0  57  27.7  0  0  27.00  42  1  64  20.2  0  0  15.15
13 0  66  28.4  0  1  25.11  43  1  59  20.7  0  1  15.40
14 0  64  21.0  1  1  20.25  44  1  39  17.6  0  1  17.88
15 0  44  24.7  0  0  17.94  45  1  40  21.5  0  0  9.06
16 0  51  28.1  1  0  18.60  46  1  63  22.5  1  0  6.61
17 0  67  19.1  0  1  18.12  47  1  77  24.2  1  1  20.24
18 0  61  24.9  0  0  16.48  48  1  62  19.7  0  0  5.00
```

（续 表）

19	0	38	22.7	1	1	24.97	49	1	58	24.2	0	0	19.32
20	0	69	29.2	0	1	32.49	50	1	70	28.9	0	1	21.48
21	0	63	26.1	0	0	26.81	51	1	50	23.2	0	1	18.02
22	0	71	23.9	0	0	29.53	52	1	56	24.6	0	1	11.16
23	0	59	22.2	1	1	24.49	53	1	70	26.7	1	0	19.62
24	0	48	26.8	0	0	20.87	54	1	37	22.0	0	0	18.77
25	0	53	33.1	0	1	21.82	55	1	49	25.4	0	0	23.30
26	0	77	27.8	1	1	17.71	56	1	51	25.1	1	1	18.96
27	0	59	29.7	0	0	36.21	57	1	60	29.9	0	0	9.15
28	0	70	24.1	1	1	23.93	58	1	66	26.5	1	0	14.06
29	0	67	25.8	0	0	28.11	59	1	55	26.1	0	0	11.85
30	0	67	26.2	1	0	28.80	60	1	56	25.2	1	1	17.14;

```
run;
ods html;
ods graphics on;
PROC REG data=CT15-1 outest=aa  lineprinter;
m1: model y=x1-x5/COLLIN  COLLINOINT r aic bic;
m2: model y=x1-x5/SELECTION=STEPWISE SLE=0.1 SLS=0.05;
m3: model y=x1-x5/SELECTION=CP best=5;
m4: model y=x1-x5/SELECTION=ADJRSQ best=5;
m5: model y=x1-x5/SELECTION=RSQUARE best=5;
run;
ods graphics off;
proc print data=aa;run;
ods html close;
```

程序过程的第一行是说明对数据集 CT15-1 进行分析，并且将 model 语句中的有关模型统计量写入到数据集“aa”中。随后有五个 model 语句，并分别取名为 m1-m5。m1 首先对数据进行共线性诊断。SELECTION＝选项是指明自变量筛选的方法，m2 运用了“逐步选择法”，并设置了相关的参数，“SLE＝0.1”表示自变量进入模型的水准为 0.1，“SLS＝0.05”表示模型中剔除自变量的水准为 0.05；m2 运用了“基于 Mallow CP 统计量数值大小的选择变量法”筛选变量；m3 运用了“基于校正 R^2 数值大小的选择变量法”筛选变量；m4 运用了“基于 R^2 数值大小的选择变量法”筛选变量。“best＝5”是指明在所有的模型组合中选取 5 个最优的模型。当 model 语句中有选项“B”时，则除了显示最优的变量组合外，还将其参数的估计值列出。

主要结果如下：

The REG Procedure

Model: m1

Dependent Variable: *y*

Parameter Estimates

Variable	*DF*	Parameter Estimate	Standard Error	*t* Value	*Pr* > |*t*|
Intercept	1	14.987 43	6.444 10	2.33	0.023 8
x_1	1	−8.296 66	1.435 45	−5.78	<0.000 1

（续　表）

Parameter Estimates

Variable	*DF*	Parameter Estimate	Standard Error	*t* Value	$Pr > \|t\|$
x_2	1	0.051 90	0.069 37	0.75	0.457 6
x_3	1	0.271 62	0.225 92	1.20	0.234 5
x_4	1	−0.874 11	1.539 95	−0.57	0.572 6
x_5	1	−0.108 05	1.427 82	−0.08	0.940 0

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	x_1	x_2	x_3	x_4	x_5
1	4.426 85	1.000 00	0.000 575 72	0.014 46	0.001 45	0.000 735 49	0.014 46	0.015 01
2	0.659 83	2.590 18	0.000 162 42	0.044 20	0.000 307 91	0.000 204 40	0.919 82	0.027 43
3	0.525 33	2.902 90	0.000 004 08	0.423 30	9.635 945E-7	5.590 938E-7	0.001 45	0.506 16
4	0.359 24	3.510 41	0.003 21	0.448 77	0.008 04	0.006 17	0.053 82	0.412 28
5	0.021 33	14.404 98	0.046 36	0.003 26	0.944 12	0.176 29	0.007 16	0.014 46
6	0.007 41	24.434 49	0.949 69	0.066 01	0.046 09	0.816 60	0.003 30	0.024 66

Collinearity Diagnostics (intercept adjusted)

Number	Eigenvalue	Condition Index	Proportion of Variation				
			x_1	x_2	x_3	x_4	x_5
1	1.266 53	1.000 00	0.199 51	0.162 76	0.364 14	0.002 33	0.002 05
2	1.142 39	1.052 93	0.027 41	0.235 04	0.036 69	0.225 46	0.316 14
3	0.997 75	1.126 67	0.220 26	0.037 60	0.014 59	0.241 77	0.454 44
4	0.944 24	1.158 15	0.311 80	0.202 77	0.024 52	0.479 09	0.003 94
5	0.649 08	1.396 87	0.241 02	0.361 83	0.560 06	0.051 35	0.223 43

以上是对截距项未校正和校正之后的共线性回归诊断结果：因本例的截距项有统计学意义，故用校正之后的共线性回归诊断结果。由以上可以看出最大条件数为 1.396 87（condition Index）<10，说明此 3 个自变量的复共线性对参数估计的影响很小。可以进行参数估计。

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook D
1	24.240 0	23.640 8	1.697 3	0.599 2	5.160	0.116	| | |	0.000
2	11.530 0	15.481 1	1.896 7	−3.951 1	5.090	−0.776	| * | |	0.014
3	30.410 0	22.908 7	1.877 5	7.501 3	5.097	1.472	| | * * |	0.049

（续 表）

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook D
4	16.690 0	15.705 8	1.437 8	0.984 2	5.238	0.188	| | |	0.000
5	23.890 0	24.704 4	1.571 2	−0.814 4	5.200	−0.157	| | |	0.000
6	16.100 0	18.372 3	2.053 0	−2.272 3	5.029	−0.452	| |	0.006
7	22.640 0	25.330 9	1.652 9	−2.690 9	5.174	−0.520	| * | |	0.005
8	12.810 0	16.002 7	1.636 9	−3.1927	5.179	−0.616	| * | |	0.006
9	25.440 0	23.166 3	1.687 2	2.273 7	5.163	0.440	| | |	0.003
10	14.480 0	17.611 2	1.657 3	−3.131 2	5.173	−0.605	| * | |	0.006
11	26.300 0	23.118 1	2.087 1	3.181 9	5.015	0.634	| | * |	0.012
12	20.440 0	14.481 5	1.750 7	5.958 5	5.142	1.159	| | * * |	0.026
13	24.620 0	25.834 9	1.640 1	−1.214 9	5.178	−0.235	| | |	0.001
14	34.670 0	17.998 7	1.849 0	16.671 3	5.107	3.264	| | * * * * * * |	0.233
15	19.510 0	24.347 0	1.405 5	−4.837 0	5.247	−0.922	| * | |	0.010
16	17.330 0	15.513 8	1.538 0	1.816 2	5.210	0.349	| | |	0.002
17	22.290 0	24.902 3	1.644 3	−2.612 3	5.177	−0.505	| * | |	0.004
18	6.610 0	15.163 7	1.497 2	−8.553 7	5.221	−1.638	| * * * | |	0.037
19	31.670 0	25.321 8	1.655 8	6.348 2	5.173	1.227	| | * * |	0.026
20	20.130 0	15.603 6	1.660 7	4.526 4	5.172	0.875	| | * |	0.013
21	26.400 0	22.464 5	1.939 7	3.935 5	5.074	0.776	| | * |	0.015
22	14.540 0	15.022 3	1.635 8	−0.482 3	5.180	−0.093 1	| | |	0.000
23	27.000 0	25.469 7	1.353 3	1.530 3	5.261	0.291	| | |	0.001
24	15.150 0	15.499 2	1.772 1	−0.349 2	5.135	−0.068 0	| | |	0.000
25	25.110 0	26.018 9	1.549 6	−0.908 9	5.206	−0.175	| | |	0.000
26	15.400 0	15.267 5	1.510 2	0.132 5	5.218	0.025 4	| | |	0.000
27	20.250 0	23.031 0	1.877 7	−2.781 0	5.097	−0.546	| * | |	0.007
28	17.880 0	13.387 4	2.170 3	4.492 6	4.979	0.902	| | * |	0.026
29	17.940 0	23.980 1	1.519 4	−6.040 1	5.215	−1.158	| * * | |	0.019
30	9.060 0	14.606 7	1.728 7	−5.546 7	5.149	−1.077	| * * | |	0.022
31	18.600 0	24.392 8	1.814 8	−5.792 8	5.120	−1.131	| * * | |	0.027
32	6.610 0	15.197 9	1.714 9	−8.587 9	5.154	−1.666	| * * * | |	0.051
33	18.120 0	23.544 8	2.075 0	−5.424 8	5.020	−1.081	| * * | |	0.033
34	20.240 0	16.278 3	1.993 6	3.961 7	5.053	0.784	| | * |	0.016
35	16.480 0	24.916 8	1.377 6	−8.436 8	5.254	−1.606	| * * * | |	0.030
36	5.000 0	15.259 6	1.789 4	−10.259 6	5.129	−2.000	| * * * * | |	0.081
37	24.970 0	22.143 3	2.199 9	2.826 7	4.966	0.569	| | * |	0.011
38	19.320 0	16.274 3	1.282 9	3.045 7	5.278	0.577	| | * |	0.003
39	32.490 0	26.391 9	1.691 3	6.098 1	5.162	1.181	| | * * |	0.025

（续　表）

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook D
40	21.480 0	18.065 7	1.843 5	3.414 3	5.109	0.668	| |* |	0.010
41	26.810 0	25.346 5	1.388 1	1.463 5	5.251	0.279	| | |	0.001
42	18.020 0	15.479 4	1.438 3	2.540 6	5.238	0.485	| | |	0.003
43	29.530 0	25.164 2	1.792 3	4.365 8	5.128	0.851	| |* |	0.015
44	11.160 0	16.171 1	1.354 5	−5.011 1	5.260	−0.953	| *| |	0.010
45	24.490 0	23.097 5	1.713 7	1.392 5	5.154	0.270	| | |	0.001
46	19.620 0	16.702 1	1.842 9	2.917 9	5.110	0.571	| |* |	0.007
47	20.870 0	24.758 1	1.428 0	−3.888 1	5.241	−0.742	| *| |	0.007
48	18.770 0	14.586 8	1.829 2	4.183 2	5.115	0.818	| |* |	0.014
49	21.820 0	26.620 8	2.306 5	−4.800 8	4.918	−0.976	| *| |	0.035
50	23.300 0	16.133 1	1.380 8	7.166 9	5.253	1.364	| |* * |	0.021
51	17.710 0	25.552 8	2.028 2	−7.842 8	5.039	−1.556	| * * *| |	0.065
52	18.960 0	15.173 3	1.811 2	3.786 7	5.121	0.739	| |* |	0.011
53	36.210 0	26.116 7	1.543 2	10.093 3	5.208	1.938	| |* * * |	0.055
54	9.150 0	17.926 3	1.731 0	−8.776 3	5.149	−1.705	| * * *| |	0.055
55	23.930 0	24.184 5	1.776 3	−0.254 5	5.133	−0.049 6	| | |	0.000
56	14.060 0	16.440 1	1.742 0	−2.380 1	5.145	−0.463	| | |	0.004
57	28.110 0	25.472 7	1.517 5	2.637 3	5.216	0.506	| |* |	0.004
58	11.850 0	16.634 7	1.319 3	−4.784 7	5.269	−0.908	| *| |	0.009
59	28.800 0	24.707 2	1.747 1	4.092 8	5.143	0.796	| |* |	0.012
60	17.140 0	15.459 9	1.711 4	1.680 1	5.155	0.326	| | |	0.002

以上是对 60 名受试对象的原始数据进行的残差分析，可以看出第 14 受试对象的残差较大为 16.67，可以认为该例的数据为异常点，这个点的存在将会使得模型参数的估计值发生较大的偏差，应将其从数据集中剔除，重新进行模型拟合。剔除观测不必在数据集中将其删除，而是运用 reweight 和 refit 语句即可。

Model: m2

Dependent Variable: *y*

Stepwise Selection: Step 1

Variable x1 Entered: R-Square = 0.400 4 and C(p) = 0.805 2

Analysis of Variance					
Source	*DF*	Sum of Squares	Mean Square	*F* Value	*Pr* > *F*
Model	1	1 119.312 04	1 119.312 04	38.73	<0.000 1
Error	58	1 676.046 62	28.897 36		
Corrected Total	59	2 795.358 66			

Variable	Parameter Estimate	Standard Error	Type Ⅱ SS	*F* Value	*Pr* > *F*
Intercept	24.555 00	0.981 45	18 088	625.95	<0.000 1
x_1	−8.638 33	1.387 98	1 119.312 04	38.73	<0.000 1

Bounds on condition number: 1, 1

模型 m2 进行了逐步回归分析，只进行了一步，只有 1 个自变量 X_1 保留在模型内。经检验，其 $F=38.73$，$P<0.000\ 1$，按 $\alpha=0.05$ 水准，说明 X_1 对因变量的影响具有统计学意义。

Model: m3

Dependent Variable: *y*

C(p) Selection Method

Number in Model	C(p)	R-Square	Variables in Model
2	0.803 3	0.421 5	x_1 x_3
1	0.805 2	0.400 4	x_1
2	1.923 3	0.409 7	x_1 x_2
3	2.328 2	0.426 6	x_1 x_2 x_3
2	2.502 2	0.403 6	x_1 x_4

以上是模型 m3 的运行结果，它采用的是 CP 法进行自变量的筛选。CP 值越小，说明模型拟合越好，从以上结果可以看出最优的模型中自变量组合为 x_1，x_3；其次为 x_1。

Model: m4

Dependent Variable: *y*

Adjusted R-Square Selection Method

Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.401 3	0.421 5	x_1 x_3
3	0.395 8	0.426 6	x_1 x_2 x_3
3	0.393 3	0.424 1	x_1 x_3 x_4
3	0.390 6	0.421 6	x_1 x_3 x_5
1	0.390 1	0.400 4	x_1

以上是模型 m4 的运行结果，它采用的是校正 R^2 选择变量法进行自变量的筛选。可以看出最优的模型中自变量组合为 x_1，x_3；其次为 x_1，x_2，x_3。

Model: m5

Dependent Variable: *y*

R-Square Selection Method

Number in Model	R-Square	Variables in Model
1	0.400 4	x_1
1	0.071 9	x_3

（续　表）

Number in Model	R-Square	Variables in Model
1	0.008 5	x_2
1	0.003 2	x_4
1	0.000 0	x_5
2	0.421 5	x_1 x_3
2	0.409 7	x_1 x_2
2	0.403 6	x_1 x_4
2	0.400 7	x_1 x_5
2	0.074 0	x_3 x_4
3	0.426 6	x_1 x_2 x_3
3	0.424 1	x_1 x_3 x_4
3	0.421 6	x_1 x_3 x_5
3	0.414 1	x_1 x_2 x_4
3	0.410 5	x_1 x_2 x_5
4	0.430 0	x_1 x_2 x_3 x_4
4	0.426 6	x_1 x_2 x_3 x_5
4	0.424 1	x_1 x_3 x_4 x_5
4	0.414 8	x_1 x_2 x_4 x_5
4	0.077 4	x_2 x_3 x_4 x_5
5	0.430 0	x_1 x_2 x_3 x_4 x_5

以上是模型 m5 的运行结果，它采用的是 R^2 数值大小的选择变量法进行自变量的筛选。R^2 越大说明模型越好，但是全模型时的 R^2 最大，不符合实际情况。从上述结果中可以看出，随自变量个数的增加，R^2 是不断增大的，但是在自变量个数为 2 时 R^2 值为 0.421 5，已接近了最大的 R^2 值 0.927，因此，最佳模型的自变量组合为 x_1，x_3。

根据上述的分析，重新拟合方程，程序如下：

```
proc reg data=CT15-1 outest=aa;
model y=x1 x3/stb R AIC BIC;
reweight obs. =14;
refit;
plot r. * p. =" * ";
Run;
proc print data=aa; run;
```

结果如下：

Analysis of Variance					
Source	*DF*	Sum of Squares	Mean Square	*F* Value	*Pr* > *F*
Model	2	1 298.795 26	649.397 63	28.31	<0.000 1
Error	56	1 284.686 95	22.940 84		
Corrected Total	58	2 583.482 22			

Root MSE	4.789 66	R-Square	0.502 7
Dependent Mean	19.991 19	Adj R-Sq	0.485 0
Coeff Var	23.958 86		

Parameter Estimates

Variable	*DF*	Parameter Estimate	Standard Error	*t* Value	$Pr > \|t\|$	Standardized Estimate
Intercept	1	19.123 23	5.033 60	3.80	0.000 4	0
x_1	1	−8.984 65	1.277 06	−7.04	<.000 1	−0.678 78
x_3	1	0.213 62	0.194 96	1.10	0.277 9	0.105 72

以上是回归分析的最后结果，给出对整个回归模型的方差分析结果、参数估计值及其假设检验结果、标准化回归系数。由标准化回归系数可以看出，仅有 x_1 对因变量 Y 的影响最大，AIC=188，BIC=190（AIC，BIC 是通过"proc print data=aa；run；"来实现的），较之前的模型拟合 AIC，BIC 值低很多。可以看出重新拟合的模型优于先前的模型。最后的多重线性回归方程为：$\hat{Y} = -0.679x_1 + 0.106x_3$。其中 $R^2 = 0.502\ 7$，说明自变量仅能解释方程的 50.2%的信息，因此此方程的实际意义并不大。

这是残差分析图（图 15-1），由于 14 例观测为异常值，因此在拟合方程中删除，此处残差图并未显示。

结论：根据现有资料，建模的意义并不大，但是可知 x_1 变量（使用器械分组不同）对结果是有意义的，说明两组人群（对照组、试验组）术后病变阶段内直径狭窄程度是不同的。

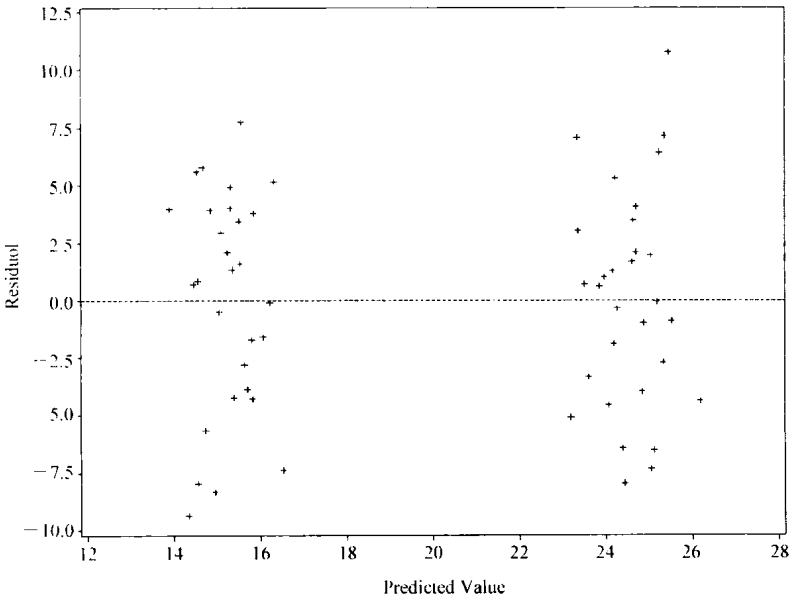

图 15-1　残差分析图

（郭　晋　李　卫）

第16章 回归分析

在医疗器械临床试验中，有时除了分组因素不同（试验组，对照组）可能导致受试对象的疗效具有差别以外，可能还受到受试者年龄、性别、遗传因素、饮食、吸烟、饮酒、肥胖等多个因素的影响，尽管 RCT 研究尽可能使两组受试对象的这类信息均衡，然而仍然有可能某些基线信息存在差别。此时我们仍需要建立回归模型对其进行分析，排除混杂因素的干扰，得到主要试验因素（使用器械的不同）对其疗效的影响。与 15 章相比，Logistic 回归分析适用于临床终点（因变量）是分类变量的情况。

Logistic 回归按照因变量的类型可分为：因变量为二分类的 Logistic 回归、因变量为多值有序变量的 Logistic 回归、因变量为多值无序变量的 Logistic 回归；Logistic 回归按研究设计类型可分为非条件 Logistic 回归（平行组设计）和条件 Logistic 回归（匹配设计）。在临床试验中，二分类的 Logistic 回归、多值有序变量的 Logistic 回归应用相对较多，本章将重点介绍这两类 Logistic 回归模型的理论，SAS 编程方法，以及结果解释。

一、二值变量的 Logistic 回归分析

（一）模型结构

上一章节介绍的多重线性回归模型，因变量与自变量的线性组合即，$\hat{y}=\alpha+\beta_1 x_1+\beta_2 x_2\cdots\beta_p x_p$，考察多个自变量对结果的影响。对于结果变量如果为二分类变量的数据，设结果变量 $Y=1$ 的概率为 P，则有 $Y=0$ 的概率为 $1-P$，P 的取值应在 0～1，而线性模型的应变量取值是任意的，因此利用线性模型去估计结果变量为二分类变量的数据显然是不合适的。

1970 年，Cox 提出 Logit 变换（Logit transformation）。称发病的概率 P 与未发病的概率 $1-P$ 之比为“优势（odds）”。对 P 作 Logit 变换，Logit(P)定义为优势之对数（log odds），即：

$$\text{Logit}(P)=\ln\left(\frac{P}{1-P}\right)$$

因此，多重 Logistic 回归方程的线性表达式定义为：

$$\text{Logit}(P)=\alpha+\beta_1 x_1+\beta_2 x_2\cdots\beta_i x_i$$

方程等号左边，是 $P/(1-P)$的对数，取值范围在 $-\infty$～$+\infty$，却限定了 P 的取值在 0～1。Logit(P)与各因素间呈线性关系，β_i为 Logistic 回归偏回归系数，表示在其他变量都固定的情况下，x_i对 Y 即 Logit(P)影响的大小。取 β_i的反对数可得优势比 $OR=\exp(\beta_i)$，表示其他变量不变的情况下，x_i每变化一个单位时所引起的优势比的自然对数改变量。当某种疾病的发病率或死亡率很低时，$OR\approx RR$（即相对危险度）。

如果将 Logit(P)视为因变量，Logistic 回归与多重线性回归形式上是一样的，但是 Logistic 回归适用的误差分布是二项分布，回归系数的估计是利用极大似然估计，模型的检验运用

的是 Wald 检验或似然比检验，这些都是与多重现性回归分析不一致的。

对于 Logistic 回归方程的拟合一般利用 SAS 中 Logistic 过程完成，一些较复杂的 Logistic 回归分析则需要通过调用 SAS 中的 CATMOD 过程或 PHREG 过程完成。本章仅介绍调用 SAS 中的 Logistic 过程完成一般的 Logistic 回归分析。

(二)Logistic 回归方程的参数估计及假设检验

1. *Logistic 回归参数的估计* Logistic 回归系数的估计通常采用极大似然法(maximum likelihood，ML)。

当各事件为独立发生时，则 n 个观察对象所构成的似然函数 L 是每一观察对象的似然函数贡献量的乘积，即似然函数。

$$L=\prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i},\ i=1,2\cdots n$$

式中，i 为观察对象编号，$\prod_{i=1}^{n}$ 为个体 1 到个体 n 的连乘积。Y_i 为因变量，其取值为 0 或 1。p 为预报概率

将以上似然函数 L 两边取自然对数有：

$$\ln L=\sum_{i=1}^{n}[Y_i\ln p_i+(1-Y_i)\ln(1-p_i)]$$

称为对数似然函数。

与 L 比较，$\ln L$ 的表达式相对简单。用 $\ln L$ 取一阶导数求解参数，相对于参数 β_j，令 $\ln L$ 的一阶导数为 0，即 $\frac{\partial \ln L}{\partial \beta_j}=0$，采用迭代方法解此方程，可得到 β_j 的估计值 b_j，其中 j 是每个参数的编号，$i=1,2\cdots p$。

2. *回归系数的假设检验* 由于 Logistic 回归是一种概率模型，回归系数估计常用的方法是最大似然法。最大似然估计是总体参数的渐近无偏和有效的点估计。Logistic 回归系数的最大似然估计近似服从正态分布，所以可以直接对回归系数进行假设检验。常用的回归系数的假设检验方法有。

(1)Wald 检验：对于规模很大的样本，检验其总体系数是否为 0 可以采用 Z 统计量：

$$Z=\frac{\hat{\beta}_k}{SE_{\hat{\beta}_k}}$$

其中，$SE_{\hat{\beta}_k}$ 为 $\hat{\beta}_k$ 的标准误。当不拒绝 H_0 时(即 $H_0:\beta_k=0$)，Z 为标准正态分布

(2)似然比检验(likelihood ratio，LR)：统计学已经证明，在大样本时，如果两个模型之间有嵌套关系，那么两个模型之间的对数似然值乘以 -2 的结果(简称为 $-2LL$)之差近似服从 x^2 分布。这一检验统计量称为似然比统计量。

回归系数的置信区间：对于选定的 α，参数 β_k 的 $(1-\alpha)100\%$ 的置信区间为：

$$\hat{\beta}\pm Z_{\alpha/2}\times SE_{\hat{\beta}_k}$$

其中 $\pm Z_{\alpha/2}$ 为与正态曲线下中心部分为 $(1-\alpha)100\%$ 概率所围成的左、右两个边界的横坐标；$SE_{\hat{\beta}_k}$ 为回归系数估计值 $\hat{\beta}$ 的标准误；$(\hat{\beta}-Z_{\alpha/2}\times SE_{\hat{\beta}_k})$ 和 $(\hat{\beta}+Z_{\alpha/2}\times SE_{\hat{\beta}_k})$ 两值便分别是置信区间的下限和上限。

优势比(OR)的置信区间为：

$$\left(e^{\hat{\beta}-Z_{\alpha/2}\times SE_{\hat{\beta}_k}}, e^{\hat{\beta}+Z_{\alpha/2}\times SE_{\hat{\beta}_k}}\right)$$

(三)变量筛选方法

拟合多重 Logistic 回归方程与拟合多重线性回归方程一样，也须对自变量进行筛选，只保留对回归方程具有统计学意义的自变量。在 SAS 中，筛选变量的方法主要有前进(FORWARD)法、后退(BACKWARD)法、逐步(STEPWISE)法、最优回归子集(SCORE)法。最优回归子集法基于似然得分统计量，这种方法分别对包含 1 个、2 个、3 个变量，直至包含所有自变量的模型，输出指定个数的最佳模型。其中 SCORE 法常与 BEST=n 选项一同使用，BEST 用于显示具有最优回归子集的自变量的个数。

1. *模型 x^2 统计*　所谓 x^2 模型检验定义为零假设模型与所设模型之间在 $-2LL$ 上的差距。模型 x^2 的自由度为零假设模型自由度与所设模型自由度之间的差，等于回归系数个数。模型检验是关于自变量是否与所研究事件的对数发生比呈线性相关的检验。

2. *拟合优度检验*　如果模型的预测值能够与对应的观测值有较高的一致性，就认为这一模型对数据的拟合是合适的。从这一点上说，模型的适当性指的是拟合优度。实际上我们在评价模型时需要测量的是模型预测值与实际值之间的偏差。换句话说，检验的是模型预测的“劣度”，而不是优度，即拟合不佳检验。

(1)Pearson x^2：用来比较模型预测的和观测的所关心的某事件发生和不发生的频数，检验关于模型的假设是否成立。事件发生的预测值为这一模式中所有案例的预测概率 p_j 之和；事件不发生的预测值为 $1-P_j$ 的和。将观测频数和预测频数带入标准 x^2 统计量计算公式，有：

$$\chi^2=\sum_{j=1}^{J}\frac{(O_j-E_j)^2}{E_j}$$

其中 $j=1,2\cdots J$，且 J 是变量组合类型的种类数目。O_j 和 E_j 分别为第 j 类协变类型中的观测频数和预测频数。

(2)偏差：观测值与预测值的比较还可以根据对数似然函数表示。所谓“似然”即在一定参数估计条件下得到该观测结果的概率。以 $\dot{L}_s$ 作为设定模型所估计的最大似然值，它概括了样本数据由这一模型所拟合的程度。但是，由于这一统计量不能独立于样本规模，因此不能仅根据它的值来估计模型的拟合优度。对于同一组数据必须有一个基准模型作为比较所设模型拟合优度的标准。一个基准模型是饱和模型，标注为 $\dot{L}_f$。通过比较 $\dot{L}_s$ 与 $\dot{L}_f$，便可以估计所拟合模型代表数据的充分程度。

在似然函数的基础上，比较 $\dot{L}_s$ 与 $\dot{L}_f$ 实际上就是比较预测值和观测值。通常采用 -2 乘以设定模型与饱和模型之间最大似然值之比的对数。

$$D=-2\ln\left(\frac{\dot{L}_s}{\dot{L}_f}\right)=-2(\ln\dot{L}_s-\ln\dot{L}_f)$$

当样本足够大时，它服从 x^2 分布，其自由度等于所拟合模型中变量组合方式的个数减去系数个数所得之差。当 $\dot{L}_s$ 与 $\dot{L}_f$ 值近似时，D 值就会很小，表示模型拟合很好。D 统计量和 Pearson x^2 有着同样的渐进 x^2 分布，然而，这两个统计量的值一般有所不同。模型的最大似然估计是所设立的模型取得的似然函数作为拟合优度指标的。偏差通过这些估计便取得其最小值。就这点而言，当用最大似然值来拟合 Logistic 回归模型时，偏差比 x^2 更适用于测量拟合优度。

当模型中有连续自变量时或变量组合方式的数目太多时，它们又不能适当的用于同一目的。但是，不管组合方式的数量如何，偏差指标在比较嵌套模型（即一个模型在原有模型变量的基础上再加入或减少某些变量）时仍是重要的指标。两个有嵌套关系的模型之间的差别可以用来判断特定变量的重要性，但 Pearson x^2 却不能实现这一目的。

（3）Hosmer-Lemeshow 拟合优度指标：当自变量数量增加时，尤其连续自变量纳入模型之后，变量组合方式的数量便会很大，于是许多组合方式下只有很少的观测案例。结果指标 D 和 Pearson x^2 不再适用于评价拟合优度。此时可采用 Hosmer-Lemeshow 指标来度量模型的拟合优度。

Hosmer-Lemeshow 指标（记为 HL）是一种类似于 Pearsonx^2 统计量的指标。它可以从观测频数和预测频数构成 2×G 交互表中求得。公式如下：

$$HL=\sum_{g=1}^{G}\frac{(y_g-n_g\dot{p}_g)}{n_g\dot{p}_g(1-\dot{p}_g)}$$

其中 G 代表分组数，且 G≤10；n_g 为第 g 组中的发生数；y_g 为第 g 组事件的观测数量；$\dot{p}_g$ 为第 g 组的预测事件概率；$n_g\dot{p}_g$ 为事件的预测数，实际上它等于第 g 组的预测概率之和。HL 是广为接受的拟合优度指标

（4）信息测量指标

①AIC 准则（Akaike information criterion，AIC），它的定义为：

$$AIC=(\frac{-2\dot{LL}_s+2(K+S)}{n})$$

其中 K 为模型中自变量的数目；S 为反应变量类别总数减 1；n 为观测例数；$\dot{LL}_s$ 为所拟合模型的最大似然估计值的自然对数，其值越大表示拟合效果越好。$-2\dot{LL}_s$ 的值域为 0 至 $+\infty$，其值越小说明拟合越好。较小的 AIC 值表示模型拟合得较好

②贝叶斯信息标准（Bayesian information criterion，BIC）。有两种不同的类型。第一种指标定义为：

$$BIC=-2\dot{LL}_s-d.f._s\times\ln(n)$$

其中 $-2\dot{LL}_s$ 是 −2 乘以所拟合模型的对数似然值；$d.f._s$ 为模型的自由度，它等于样本例数与模型中待估计参数的数目之差（即 $d.f._s$ =n−K−1）；ln(n) 为样本例数的自然对数

另一种 BIC 指标用来对所拟合模型与零假设对应的模型（即只包含常数项的模型）进行比较。其统计公式为：

$$BIC'=-G_s+d.f._s\times\ln(n)$$

其中 $d.f._s$ 为自变量数目（即 K），而 G_s 为 −2 乘以所拟合模型与零假设对应的模型之间的最大似然比之差的对数，即：$G_s=-2\dot{LL}_0-(-2\dot{LL}_s)=2\dot{LL}_s-2\dot{LL}_0$

对于零假设对应的模型而言，G_0 和 $d.f._0$ 的值都是 0，所以 BIC_0 的值也是 0。因此，$BIC_S>0$ 表示所拟合模型比零假设对应的模型差，而 $BIC_S<0$ 表示所拟合模型比零假设对应的模型要好。具有最小的 BIC 或 BIC' 值的模型最好。

另外，SAS 系统提供了各种方法用来修正过度分散，其中包括适应于二项分布数据的 Williams 方法。拟合模型的充分性可以用各种拟合优度检验来评估，其中包括用于二值响应数据的 Hosmer Lemeshow 检验。

【例 16-1】 纳米银祛痘凝胶为某公司研制开发的治疗粉刺、寻常性痤疮、脂溢性皮炎等表浅炎症性皮肤病的治疗的医疗器械，为考察其疗效和安全性，某院于 2005 年 7 月至 9 月采用该产品对寻常性痤疮(包括粉刺)40 例，脂溢性皮炎 20 例受试者进行治疗，观察治疗前后症状和体征的改善情况。所有病例 60 例均来自皮肤科门诊，符合入选标准，年龄 15～45 岁。将患者随机分为两组对照组为普通祛痘凝胶，试验组为纳米银祛痘凝胶，考虑到多个因素对于疗效的影响，建立 Logistic 回归模型评价两种凝胶类型对炎症性皮肤病的治疗效果。

x_1：受试组别(0：对照组；1：试验组)，x_2：性别(1：男性；2：女性)。x_3：年龄，x_4：皮肤病种类(1：寻常性痤疮；2：脂溢性皮炎)，x_5：皮炎症状程度(0：轻度，1：中度)，y：术后疗效(0：无效；1：有效)数据如表 16-1 所示。

表 16-1　60 例患者基线信息、皮肤病种类、炎症程度以及术后疗效

id	x_1	x_2	x_3	x_4	x_5	y	id	x_1	x_2	x_3	x_4	x_5	y
1	0	1	30	1	1	0	31	1	1	17	1	0	1
2	0	2	19	1	0	1	32	1	1	16	1	0	1
3	0	1	18	2	1	1	33	1	1	44	1	1	0
4	0	2	17	2	0	0	34	1	1	17	2	1	1
5	0	1	23	1	1	0	35	1	1	37	1	1	1
6	0	1	19	1	0	1	36	1	1	26	1	0	1
7	0	1	22	1	0	0	37	1	2	45	2	0	1
8	0	2	31	2	1	1	38	1	2	17	1	1	1
9	0	2	37	1	0	1	39	1	1	27	1	1	1
10	0	1	21	1	1	0	40	1	1	36	1	1	0
11	0	1	19	1	1	0	41	1	1	27	2	0	1
12	0	1	27	2	0	1	42	1	2	34	1	0	1
13	0	1	36	1	1	1	43	1	1	29	1	1	0
14	0	2	34	1	1	0	44	1	1	19	1	1	1
15	0	2	15	2	0	1	45	1	1	15	2	0	1
16	0	1	21	2	0	1	46	1	1	33	2	0	0
17	0	1	37	1	1	0	47	1	2	45	2	1	1
18	0	1	31	1	0	0	48	1	1	32	1	0	1
19	0	1	18	2	1	1	49	1	1	28	2	0	1
20	0	1	39	1	1	0	50	1	2	40	1	1	1
21	0	2	33	1	0	1	51	1	1	20	1	1	1
22	0	1	41	1	0	1	52	1	1	26	1	1	1
23	0	1	29	1	1	0	53	1	1	40	1	0	0
24	0	1	18	2	0	0	54	1	2	17	1	0	1
25	0	1	23	1	1	1	55	1	1	19	1	0	1
26	0	2	45	2	1	0	56	1	1	21	2	1	1
27	0	1	29	2	0	1	57	1	2	30	2	0	1
28	0	1	40	1	1	0	58	1	2	36	2	0	0
29	0	2	37	1	0	1	59	1	1	25	1	0	1
30	0	2	37	1	0	1	60	1	1	26	1	1	1

SAS 程序如 CT16-1：

程序	说明
DATACT16-1； Input id x_1-x_5@@ y； cards； / * 录入表 16-1 数据 * /；	建立数据集 输入数据
run； ods html； proc logistic data=CT16-1 descending； model Y=x_1-x_5 / selection=stepwise sle=0.10 sls=0.05 rsq cl stb scale=none aggregate； run； ods html close；	 "descending"选择项是要求 SAS 程序按照降序输出结果，即输出 Y 取值较大时(Y=1)的参数估计值 "stepwise"选择逐步回归 "rsq cl"选择项是要求输出 R-square 值和回归系数的 95%置信区间；"stb"选择项是要求输出标准化回归系数；"scale=none aggregate"选择项是要求对模型进行拟合优度检验

当数据中的自变量有多分类变量(本例中自变量均为二分类变量或连续性变量)时，需要对变量赋哑变量，较为简便的做法是在"model"语句之前加"class x1(param=ref ref='1');" "x_1"即为多分类变量，"ref='1'"意为变量的各水平均与赋值为"1"的水平相比求得 OR 值。

主要结果与结果解释：

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	78.382	75.480
SC	80.476	79.668
−2 Log L	76.382	71.480

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	*DF*	*Pr* > ChiSq
Likelihood Ratio	4.902 2	1	0.026 8
Score	4.800 0	1	0.028 5
Wald	4.585 4	1	0.032 2

此处检验回归系数是否来自"回归系数均为 0 总体"进行检验，在"Model Fit Statistics"部分输出了 3 种不同的回归系数总体是否为 0 的检验统计量，"AIC(Akaike Information Criterion)"检验法，"SC(Schwarte Coriterion)"检验法和"−2Log L(Likelihood Ratio)"检验法，前二者均用于比较同一数据下自变量个数不同的模型，值越小提示回归模型越合适。经过似然比检验(Likelihood Ratio)、得分检验(Score)、Waldx^2 卡方检验 P 均<0.05，说明模型整体有意义。

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	*DF*	Value/*DF*	*Pr* > ChiSq
Deviance	68.707 0	53	1.296 4	0.072 2
Pearson	57.991 0	53	1.094 2	0.296 4

此处对模型拟合数据的效果进行检验，SAS 提供了 2 种方法的拟合优度检验(Deviance and Pearson Goodness-of-Fit Statistics)P 值分别为 0.072 2 和 0.296 4，均<0.05，提示模型对资料总体上拟合效果较好。

Analysis of Maximum Likelihood Estimates						
Parameter	*DF*	Estimate	Standard Error	Wald Chi-Square	*Pr* > ChiSq	Standardized Estimate
Intercept	1	0.133 5	0.366 0	0.133 1	0.715 2	
x_1	1	1.252 8	0.585 0	4.585 4	0.032 2	0.348 3

此处显示，截距项无统计学意义，$P=0.715\ 2$，需要在程序中“model $Y=x_1$-x_5/”加“noint”一项，意为去掉截距项，重新拟合模型。结果如下：

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	*DF*	*Pr* > ChiSq
Likelihood Ratio	11.564 7	1	0.000 7
Score	10.800 0	1	0.001 0
Wald	9.224 5	1	0.002 4

R-Square	0.175 3	Max-rescaled R-Square	0.233 7

R-Square 与线性回归中的决定系数 R^2 类似，说明尽管模型有统计学意义，但是自变量仅能解释因变量的 23.37%。

nalysis of Maximum Likelihood Estimates						
Parameter	*DF*	Estimate	Standard Error	Wald Chi-Square	*Pr* > ChiSq	Standardized Estimate
x_1	1	1.386 3	0.456 4	9.224 5	0.002 4	0.385 4

去掉截距项重新拟合模型的结果显示，仅有变量 x_1 对 y 的影响有统计学意义，$P=0.002\ 4$，说明两组(试验组，对照组)疗效是有差别的。

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
x_1	4.000	1.635	9.785

Wald Confidence Interval for Parameters			
Parameter	Estimate	95% Confidence Limits	
x_1	1.386 3	0.491 7	2.280 9

这是 x_1 的 *OR* 值以及 *OR* 值的置信区间，回归系数的置信区间估计。

拟合 Logistic 回归模型为：

$$\ln\left(\frac{P}{1-P}\right)=0.385\ 4X_1$$

可以转换为：

$$P(Y=1)=\frac{e^{0.385\ 4X_1}}{1+e^{0.385\ 4X_1}}$$

结论：试验组纳米银祛痘凝胶与对照组产品的疗效是不同的。

二、有序变量的多重 Logistic 回归分析

当因变量为多值有序变量时，需要运用有序的 Logistic 回归分析。

设结果变量 y 为 K 个等级的有序变量，K 个等级分别用 $1,2\cdots K$ 表示。$X^T=(X_1,X_2\cdots X_p)$ 为自变量。记等级为 $j(j=1,2\cdots K)$ 的概率为：$P(Y=j\mid X)$，则等级$\leqslant j(j=1,2\cdots K)$ 的概率为：

$$P(Y\leqslant j\mid X)=P(Y=1\mid X)+\cdots P(Y=j\mid X)$$

称为等级$\leqslant j$ 的累计概率(cumulative probability)，$P(Y>j\mid X)=1-P(Y\leqslant j\mid X)$。作 Logit 变换：

$$\mathrm{Logit}P_j=\mathrm{Logit}[P(Y>j\mid X)]=\ln\frac{P(Y>j\mid X)}{1-P(Y>j\mid X)}$$

有序分类结果的 Logistic 回归定义为：

$$\mathrm{Logit}P_j=\mathrm{Logit}[P(Y>j\mid X)]=-\alpha_j+\sum_{i=1}^{P}\beta_iX_i$$

$$j=1,2\cdots K-1$$

等价于：

$$P(Y\leqslant j\mid X)=\frac{1}{1+\exp(-\alpha_j+\sum_{i=1}^{P}\beta_iX_i)}$$

实际上式将 K 个等级人为的分为两类：$\{1\cdots j\}$ 和 $\{j+1\cdots K\}$，在这两类的基础上定义 Logit 表示属于后 $K-j$ 个等级的累积概率与前 j 个等级的累积概率的优势之对数，故该概率称为累积优势模型(cumulative odds model)。

有序结果的累积优势模型有$(K-1)+P$ 个参数，α_j 和 β_j 为待估参数($j=1,2\cdots K-1,i=1,2\cdots P$)，对任一 j，Logit P 是自变量的线性函数。α_j 是解释变量均为 0 时，在某一固定 j 下的两类不同概率之比的对数值，由于回归系数 β_j 与 j 无关，故有：

$$\alpha_1<\alpha_2<\cdots<\alpha_k$$

根据有序结果的 Logistic 回归，可得每类结果的概率：

$$P(Y=j \mid X)=P(Y \leqslant j \mid X)-P(Y \leqslant j-1 \mid X)=P(\alpha_{j-1}<\sum_i \beta_i X_i+u \leqslant \alpha_j)$$

$$=\frac{1}{1+e^{-\alpha_j+\sum_i \beta_i X_i}}-\frac{1}{1+e^{-\alpha_{j-1}+\sum_i \beta_i X_i}}$$

$$j=1,2\cdots K$$

这里，α_0 定义为 $-\infty$，α_K 定义为 $+\infty$。

当其他变量不变时，X_i 的两个不同取值水平为 a，b，其优势比(odds ratio)为：

$$OR=\exp[\beta_i(b-a)]$$

可见 OR 值与 α_i 无关，回归系数 β_i 表示 X_i 每改变一个单位，Y 值提高一个及一个以上等级之优势比的对数值。若 X_i 为 0～1 变量，则 e^{β_i} 即为该变量的 OR 值。

累积优势模型中，假设自变量的回归系数 β_i 与 j 无关。如在两种治疗方案(分别记为 $x=0,1$)的评价中，结果变量依次为：无效，有效，显效，治愈四个等级(分别记为 $Y=0,1,2,3$)。按有序分类将其分为两类，有 3 种分法：

第一种：{无效}，{有效，显效，治愈}

第二种：{无效，有效}，{显效，治愈}

第三种：{无效，有效，显效}，{治愈}

按照累积优势模型的假定，3 种模型建立的 Logistic 回归，回归系数应相等，即无论哪种方法，治疗方案的效应是相同的。

【例 16-2】沿用【例 16-1】数据，将疗效分为 4 个等级，分别为无效、有效、显效、治愈。建立有序结果变量的 Logistic 回归模型，评价各因素对术后疗效的影响。x_1：受试组别(0：对照组；1：试验组)，x_2：性别(1：男性；2：女性)。x_3：年龄，x_4：皮肤病种类(1：寻常性痤疮；2：脂溢性皮炎)，x_5：皮炎症状程度(0：轻度，1：中度)，y：术后疗效(0：无效；1：有效；2：显效；3：治愈)数据如表 16-2 所示。

表 16-2　60 例患者基线信息、皮肤病种类、炎症程度以及术后疗效

id	x_1	x_2	x_3	x_4	x_5	y	id	x_1	x_2	x_3	x_4	x_5	y
1	0	1	30	1	1	0	31	1	1	17	1	0	1
2	0	2	19	1	0	2	32	1	1	16	1	0	3
3	0	1	18	2	1	2	33	1	1	44	1	1	0
4	0	2	17	2	0	0	34	1	1	17	2	1	1
5	0	1	23	1	1	0	35	1	1	37	1	1	1
6	0	1	19	1	0	1	36	1	1	26	1	0	3
7	0	1	22	1	0	0	37	1	2	45	2	0	1
8	0	2	31	2	1	1	38	1	2	17	1	1	1
9	0	2	37	1	0	1	39	1	1	27	1	1	2
10	0	1	21	1	1	0	40	1	1	36	1	1	0
11	0	1	19	1	1	0	41	1	1	27	2	0	1
12	0	1	27	2	0	1	42	1	2	34	1	0	2
13	0	1	36	1	1	1	43	1	1	29	1	1	0

（续 表）

id	x_1	x_2	x_3	x_4	x_5	y	id	x_1	x_2	x_3	x_4	x_5	
14	0	2	34	1	1	0	44	1	1	19	1	1	1
15	0	2	15	2	0	1	45	1	1	15	2	0	2
16	0	1	21	2	0	2	46	1	1	33	2	0	0
17	0	1	37	1	1	0	47	1	2	45	2	1	1
18	0	1	31	1	0	0	48	1	1	32	1	0	3
19	0	1	18	2	1	1	49	1	1	28	2	0	3
20	0	1	39	1	1	0	50	1	2	40	1	1	1
21	0	2	33	1	0	1	51	1	1	20	1	1	2
22	0	1	41	1	0	3	52	1	1	26	1	1	1
23	0	1	29	1	1	0	53	1	1	40	1	0	0
24	0	1	18	2	0	2	54	1	2	17	1	0	3
25	0	1	23	1	1	0	55	1	1	19	1	0	3
26	0	2	45	2	1	1	56	1	1	21	2	1	2
27	0	1	29	2	0	0	57	1	2	30	2	0	1
28	0	1	40	1	1	1	58	1	2	36	2	0	2
29	0	2	37	1	0	1	59	1	1	25	1	0	2
30	0	2	37	1	0	2	60	1	1	26	1	1	1

SAS 程序如 CT16-2：

程序	说明
DataCT16-2;	建立数据集
Input id x_1-x_5@@y;	
cards;	
/* 录入表 16-2 数据 */;	输入数据
run;	
ods html ;	
proc logistic data=CT16-2 descending;	“descending”选择项是要求 SAS 程序按照降序输出结果，即输出 Y 取值较大时(Y=1)的参数估计值
model Y=x_1-x_5/	
selection=stepwise sle=0.10 sls=0.05	“stepwise”选择逐步回归
rsq cl stb;	“rsq cl”选择项是要求输出 R-square 值和回归系数的 95%置信区间；“stb”选择项是要求输出标准化回归系数
run;	
ods html close;	

主要输出结果及结果解释：

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	162.155	148.140
SC	168.438	158.612
−2 Log L	156.155	138.140

R-Square	0.259 4	Max-rescaled R-Square	0.280 1

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	*DF*	*Pr* > ChiSq
Likelihood Ratio	18.014 6	2	0.000 1
Score	15.229 6	2	0.000 5
Wald	16.409 4	2	0.000 3

经过似然比检验(Likelihood Ratio)、得分检验(Score)、Wald x^2 卡方检验 P 均<0.05，说明模型整体有意义。但是 R-Square 值只有 28.01%，说明自变量仅能解释因变量的 28.01%。

Analysis of Maximum Likelihood Estimates						
Parameter		*DF*	Estimate	Standard Error	Wald Chi-Square	*Pr* > ChiSq	Standardized Estimate
Intercept	3	1	−2.333 2	0.550 9	17.938 8	<0.000 1	
Intercept	2	1	−0.812 8	0.453 8	3.208 3	0.073 3	
Intercept	1	1	1.183 0	0.471 6	6.291 7	0.012 1	
x_1		1	1.467 4	0.512 5	8.198 4	0.004 2	0.407 9
x_5		1	−1.664 3	0.522 9	10.129 9	0.001 5	−0.462 4

变量 x_1、x_5 对 y 的影响有统计学意义，P=0.004 2，P=0.004 2，说明两组(试验组，对照组)疗效是有差别的，且炎症的程度对疗效也是有影响的。

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
x_1	4.338	1.589	11.844
x_5	0.189	0.068	0.528

Wald Confidence Interval for Parameters				
Parameter		Estimate	95% Confidence Limits	
Intercept	3	−2.333 2	−3.412 9	−1.253 5
Intercept	2	−0.812 8	−1.702 1	0.076 6
Intercept	1	1.183 0	0.258 6	2.107 4
x_1		1.467 4	0.462 9	2.471 8
x_5		−1.664 3	−2.689 3	−0.639 4

这是 x_1，x_5 的 *OR* 值以及 *OR* 值的置信区间，回归系数的置信区间估计。

结果中输出了3个截距项(Intercept1至Intercept 3),说明拟合了3个模型,分别对应"无效"与"有效,显效,治愈"("P_1"与"$1-P_1$")"无效,有效"与"显效和治愈"("P_2"与"$1-P_2$")、"有效,有效,显效"与"治愈"("P_3"与"$1-P_3$")的概率。其中截距项2(Intercept2)经检验$P>0.05$,可以认为"无效和有效"与"显效和治愈"之间的差异无统计学意义。该模型的2个有意义的Logistic模型为:

$$P_1=\frac{e^{1.183+1.467x_1-1.6643x_5}}{1+e^{1.183+1.467x_1-1.6643x_5}} \qquad p_3=\frac{e^{2.3332+1.467x_1-1.6643x_5}}{1+e^{2.3332+1.467x_1-1.6643x_5}}$$

可以将x_1,x_5不同水平的值代入模型,计算得到不同疗效的概率。

结论:试验组和对照组的疗效是有差别的,且不同的炎症分级对疗效也有影响。

(郭 晋 李 卫 胡 泊)

第17章 生存分析

在医疗器械临床研究中，治疗器械的效果如何，有时需要长期随访观察。例如对体内置入的支架系统的疗效评估。评价治疗效果不仅要看临床终点，即器械治疗是否有效，还需要考察出现临床终点的时间。在这种情况下，原有的疗效指标如有效率、治愈率等难以适用。生存分析(survival analysis)是将终点事件的出现与否和达到终点所经历的时间结合起来的一种统计分析方法。

本章主要介绍生存率估计的 Kaplan-Meier 法和寿命表法、生存曲线比较的 log-rank 检验、Cox 比例风险回归模型、参数回归模型以及 SAS 软件中的 LIFETEST，PHREG 和 LIFEREG 过程及其应用。

一、基本概念与统计描述

(一)基本概念

1. *生存时间*(survival time)　一般可以认为是患某种病的病人从确诊到死亡所经历的时间。准确的定义应该为从规定的观察起点到某一给定终点事件出现的时间。终点事件可以是某种疾病的发生、某种处理(治疗)的反应、病情的复发或死亡等。由于涉及此类的随访资料以生存问题最多，因此该类型资料又称为生存资料(survival data)，有关的统计分析也统称为生存分析。

2. *完全数据*　指从观察的起点到临床终点发生所经历的时间。以观察终点死亡为例，即对某些观察对象从某种病确诊到死亡所经历的时间，称为完全数据(complete data)。完全数据所提供的关于生存时间的信息是完整的。

3. *截尾数据*　随访研究中，观察期内由于某种原因对某些观察对象未能观察到临床终点，因为无法得知确切的生存时间，称为该类型数据为截尾数据(censored data)，又称为删失数据。

产生截尾数据的原因大致有①失访：如失去联系，随访无回信，搬迁，电话无应答等。②退出：退出研究，如病人死于非此研究的其他疾病，交通意外死亡，或临时改变想法，退出研究。③终止：研究期限已到，病人仍未发生终点事件而终止观察。截尾数据常在其右上角标记“+”，表示真实的生存时间未知，只知道比观察到的截尾时间要长。

含有截尾数据是生存资料的主要特点。另外，生存时间的分布也与常见资料的统计分布(常假定符合正态分布)有明显不同，如呈指数分布、Weibull 分布、对数正态分布、对数 Logistic 分布或 Gamma 分布，因此，需要有能分析这类数据的特殊统计分析方法。

(二)统计描述

1. *死亡概率、死亡率*

(1)死亡概率(mortality probability)：记为 q，是指以后一个时段内死亡的可能性大小。

年死亡概率的计算公式为：

$$q=\frac{\text{年内死亡数}}{\text{年初观察例数}}$$

若年内有截尾，则分母用校正的人口数，如：

$$\text{校正人口数}=\text{年初人口数}-(\text{截尾例数}/2)$$

(2)死亡率(mortality rate，death rate)：记为 M，表示在某单位事件里的平均死亡强度的频率，描述过去已经发生的情况。年死亡率的计算公式为：

$$M=\frac{\text{年内死亡数}}{\text{年平均人口数}}\times 1\,000‰$$

年平均人口数通常用年中人口数代替，为：

$$\text{年平均人口数}=(\text{年初人口数}+\text{年底人口数})/2$$

2. *生存概率与生存率*

(1)生存概率(probability of survival)：记为 P，是死亡概率的对立，表示某单位时段开始时存活的个体，到该时段结束时仍存活的可能性。如 1 年生存概率 P 表示该年年初尚存人口存活满 1 年的可能性。

$$P=1-q=\frac{\text{某年活满一年人数}}{\text{某年年初人口数}}$$

(2)生存率(survival rate)：记为 $S(t)$，指观察对象经历 t 个单位时段后仍存活的可能性。生存率常随时间逐渐下降，又称生存函数(survival function)。资料中无截尾数据时计算生存率的公式如下：

$$S(t)=P(T>t)=\frac{t\text{ 时刻未仍存活的例数}}{\text{观察总例数}}$$

若含有截尾数据，须分时段计算。假定观察对象在各个时段的生存事件独立，应用概率乘法定理，则宜采用如下的公式计算：

$$S(t_k)=P(T\geqslant t_k)=P_1\times P_2\times\cdots\times P_k=S(t_{k-1})\cdot P_k$$

式中 P_i（$i=1,2\cdots k$）为各分时段的生存概率。

生存率的标准误：其中，Greenwood 法比较常用。公式为：

$$SE[S(t_k)]=S(t_k)\sqrt{\sum_{j=1}^{k}\frac{q_j}{P_j n_j}}$$

3. *风险函数*(hazard function)　又称为危险率函数、条件死亡率、死亡力、瞬时死亡率等，表示个体在生存过程中每单位时间死亡的危险度，用 $h(t)$ 表示。

$$h(t)=\frac{\text{死于区间}(t,t+\Delta t)\text{ 的病人数}}{\text{在 }t\text{ 时刻尚存的病人数}\times\Delta t}$$

4. *生存曲线*(survival curve)　是以观察(随访)时间为横轴，以生存率为纵轴，将各个时间点所对应的生存率连接在一起的曲线图。生存曲线是一条下降的曲线，分析时应注意曲线的高度和下降的坡度。平缓的生存曲线表示高生存率或较长生存期；陡峭的生存曲线表示低生存率或较短生存期。

5. *半数生存期与四分位数间距*(median survival time)　又称中位生存期，表示恰有 50% 的个体尚存活的时间。中位生存期越长，表示疾病的预后越好；反之，中位生存期越短，预后越差。估计中位生存期常用图解法或线性内插法。

四分位数间距：记为 Q，表示半数病人生存期的分布范围。

二、生存率和生存曲线估计的非参数方法

非参数法估计生存率有 Kaplan-Meier 法和寿命表法(life table method，actuarial method)，前者适用于小样本或大样本未分组资料，后者适用于观察例数较多的分组资料。

(一)Kaplan-Meier 法

Kaplan-Meier 法(以下简称 KM 法)由 Kaplan 和 Meier 于 1958 年首先提出，又称乘积限法(product-limit method)。该法利用概率乘法定理计算生存率。

【例 17-1】 观察 10 例急性心肌梗死患者置入支架治疗发生 MACE 事件的时间(月)，试估计 MACE 事件发生率。

发生 MACE 事件时间：14　18　20^{+}　22　24　24^{+}　25　27^{+}　28　38

该研究未对观察对象施加干预，属于调查研究或观察性研究。设计阶段应明确规定观察起点、终点事件、随访终止时间、截尾。本例中，终点是发生再狭窄，对于“生存率”的估计实际是对未发生 MACE 事件的率的估计(表 17-1)。

表 17-1　生存率计算

序号(j)	未发生 MACE 时间(月)	t 时刻 MACE 事件数(d)	t 时刻截尾数(c)	t 时刻期初例数(n_0)	MACE 事件概率 $q=d/n_0$	生存概率 $q=1-p$	生存率 $S(t)$	标准误
1	14	1	0	10	1/10	9/10	9/10=0.9	0.094 9
2	18	1	0	9	1/9	8/9	(9/10)(8/9)=0.8	0.126 5
3	20	0	1	8	0/8	8/8	(9/10)(8/9)=0.8	0
4	22	1	0	7	1/7	6/7	(9/10)(8/9)(6/7)=0.685 7	0.151 5
5	24	1	0	6	1/6	5/6	0.685 7(5/6)=0.571 4	0.163 8
6	24	0	1	5	0/5	5/5	0.571 4	0
7	25	1	0	4	1/4	3/4	0.428 6	0.174 3
8	27	0	1	3	0/3	3/3	0.428 6	0
9	28	1	0	2	1/2	1/2	0.214 3	0.174 8
10	38	1	0	1	1/1	0/1	0	0

Greenwood 生存率标准误近似计算公式：

$$SE[S(t_k)]=S(t_k)\sqrt{\sum_{j=1}^{k}\frac{q_j}{p_j n_j}}$$

式中 j 要求为完全数据的顺序号。假定生存率近似服从正态分布，则总体生存率的(1-α)置信区间为：

$$S(t_k)\pm u_{\alpha/2}\cdot SE[S(t_k)]$$

程序名为 CT17-1：

```
DATA CT17-1;
input t status @@;
cards;
14  1  18  1  20  0  22  1
24  1  24  0  25  1  27  0
28  1  38  1;
Run;
Ods html;
PROC LIFETESTdata=CT17-1
method=pl plots=(s);
time t*status(0);
run;
ods html close;
```

数据步中变量 t，status 分别表示生存时间和生存结局，其中 status=1 表示某终点事件发生，如死亡；status=0 表示截尾。

LIFETEST 过程可对生存资料进行 KM 法或寿命表法估计，并对两组或多组生存率作组间比较，包括 log-rank 检验、Wilcoxon 检验和似然比检验，其中似然比检验基于指数分布。

PROC LIFETEST 语句中的选项：

(1)METHOD=PL|KM|LT|LIFE|ACT，指定生存率估计方法。PL 和 KM 表示乘积限法或 KM 法(缺省值)；LT，LIFE 和 ACT 表示寿命表法。当使用寿命表法时，SAS 可自动生成生存时间的分组区间，也可用 WIDTH= 人为指定区间宽度，或用 INTERVALS=(a TO b BY c)，a，b，c 分别表示区间的初值、终值和区间的宽度。

(2)PLOTS=( )，要求绘图。括号内可填写的内容有：S，LS，LLS，H，若同时填多项，各项之间用逗号隔开，H 选项仅在使用寿命表法时有效。各选项的具体含义如下：SURVIVAL(或 S)表示生存率 $S(t)$ 对生存时间 t 的图形，即生存曲线；LOGSURV(或 LS)表示 $-\log S(t)$ 对 t 的图形，用于判断生存时间是否服从指数分布。若呈指数分布，则图形应呈过原点的直线；LOGLOGS(或 LLS)表示 $\log[-\log S(t)]$ 对 $\log t$ 的图形，用于判断生存时间是否服从 Weibull 分布。若呈 Weibull 分布，则图形应呈直线；HAZARD(或 H)表示累积风险 $h(t)$ 对 t 的图形。只有当使用寿命表法时，才出现 HAZARD 图形。

TIME 语句为 LIFETEST 过程的必须语句，设置生存时间变量和生存结局变量，括号内为截尾变量的标示值。

输出结果如下：

The LIFETEST Procedure

Product-Limit Survival Estimates

t		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000 0		1.000 0	0	0	0	10
14.000 0		0.900 0	0.100 0	0.094 9	1	9
18.000 0		0.800 0	0.200 0	0.126 5	2	8
20.000 0	*	.	.	.	2	7
22.000 0		0.685 7	0.314 3	0.151 5	3	6
24.000 0		0.571 4	0.428 6	0.163 8	4	5
24.000 0	*	.	.	.	4	4
25.000 0		0.428 6	0.571 4	0.174 3	5	3
27.000 0	*	.	.	.	5	2
28.000 0		0.214 3	0.785 7	0.174 8	6	1
38.000 0		0	1.000 0	0	7	0

以上输出是生存率的 KM 法估计结果。统计量包括各生存时间点对应的生存率(survival)、死亡率(failure)(即发生 MACE 事件)、生存率标准误(survival standard error)、累积死亡数(number failed)和期初例数(number left)。

Summary Statistics for Time Variable *t*

Quartile Estimates			
Percent	Point Estimate	95% Confidence Interval	
		Lower	Upper
75	28.000 0	25.000 0	38.000 0
50	25.000 0	22.000 0	38.000 0
25	22.000 0	14.000 0	28.000 0

Mean	Standard Error
26.171 1	2.890 7

Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
10	7	3	30.00

结果表明,14 个月不发生 MACE 事件率 90%(即 MACE 事件发生率 10%),18 个月不发生 MACE 事件率 80%(即 MACE 事件发生率 10%),其他依此类推。注意:①生存时间一栏标有"*"者为截尾生存时间,其生存率和生存率标准误用"."表示,实际上该截尾生存时间的生存率和生存率标准误与前一个完全生存时间对应数值相同。如 20 个月对应不发生 MACE 事件率 80%。②两个完全生存时间之间任意时间的生存率均等于前一个完全生存时间的生存率。如 16 个月不发生 MACE 事件率 90%。

第二部分输出生存时间的分位数,包括 75%、50%和 25%分位数的点估计及 95%置信区间。50%分位数即中位生存期为 25 个月,95%置信区间 22～38(月)。

最后一部分概括截尾和非截尾例数。本例总例数 10,其中发生 MACE 事件 7 例,截尾 3 例,截尾百分比 30%。

Kaplan-Meier 生存曲线为阶梯形曲线,见图 17-1。

(二)寿命表法

寿命表法除需计算期初有效例数 n_i 外,其余同 KM 法。假定截尾可发生在各区间内任一时间,按截尾者平均每人观察了该区间宽度的 50%,则期初有效例数应为期初观察例数 $n_i' - c_i/2$,即 $n_i = n_i' - c_i/2$,其中 c_i 为第 i 个区间的宽度。

【例 17-2】 观察 94 例急性心肌梗死患者置入支架治疗发生 MACE 事件的时间(月),试估计 MACE 事件发生率(表 17-2)。

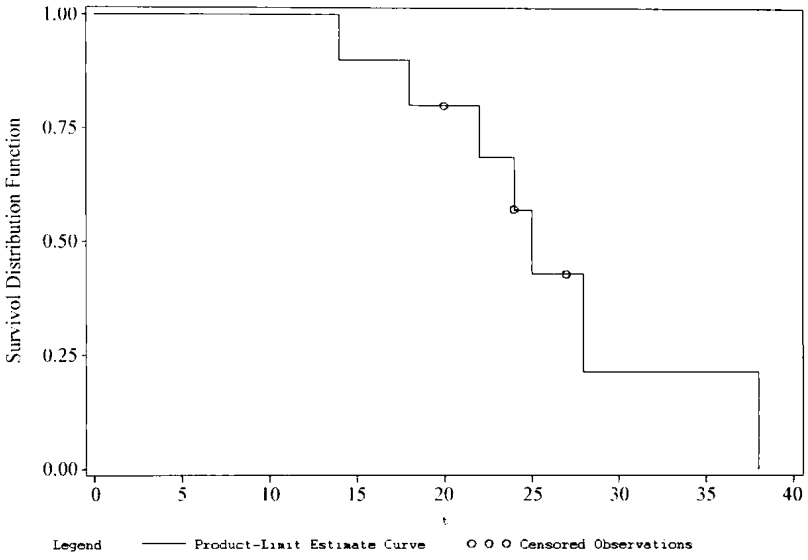

图 17-1　Kaplan-Meier 法绘制生存曲线

表 17-2　MACE 事件发生情况

时间	发生 MACE 事件人数	截尾人数
0～1 年	2	0
1～2 年	8	2
2～3 年	24	3
3～4 年	25	4
4 年	3	23

程序名为:CT17-2

```
DATA CT17-2;
INPUT t status number @@;
CARDS;
0 1 2    0 0 0
1 1 8    1 0 2
2 1 24   2 0 3
3 1 25   0 4
4 1 3    4 0 23;
run;
ods html;
proc lifetest data=CT17-2
method=lt plots=(s) width=1;
time t * status(0);
freq number;
run;
ods html close;
```

数据步中变量 t、status 和 number 分别表示每个时间区间的中点、生存结局(发生 MACE 事件=1;截尾=0)及其频数。过程步中指定生存率的估计方法为寿命表法,时间区间宽度为 1。同时加 FREQ 语句指定频数变量。时间的选择为生存时间的上限,即"0,1,2,3,4"。

SAS 输出结果如下:

The LIFETEST Procedure

Life Table Survival Estimates

Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	Evaluated at the Midpoint of the Interval			
Lower	Upper											PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	1	2	0	94.0	0.021 3	0.014 9	1.000 0	0	0	3.428 2	0.174 5	0.021 3	0.014 9	0.021 505	0.015 206
1	2	8	2	91.0	0.087 9	0.029 7	0.978 7	0.021 3	0.014 9	2.464 2	0.173 6	0.086 0	0.029 1	0.091 954	0.032 476
2	3	24	3	80.5	0.298 1	0.051 0	0.892 7	0.107 3	0.032 1	1.609 7	0.168 3	0.266 1	0.046 5	0.350 365	0.070 412
3	4	25	4	53.0	0.471 7	0.068 6	0.626 5	0.373 5	0.050 8	.	.	0.295 5	0.049 2	0.617 284	0.117 429
4	.	3	23	14.5	0.206 9	0.106 4	0.331 0	0.669 0	0.050 6	.	.	.	.	.	.

Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
94	62	32	34.04

SAS 输出统计量包括各生存时间区间(interval)的死亡数(number failed)、截尾数(number censored)、期初有效例数(effective sample size)、条件死亡概率(conditional probability of failure)、条件死亡概率标准误(conditional probability standard error)、区间左端点处生存率(survival)(即不发生 MACE 事件率)、死亡率(failure)(即 MACE 事件发生率)、生存率标准误(survival standard error)、中位剩余寿命(median residual lifetime)、中位剩余寿命标准误(median standard error)及各时间区间中点处的概率密度函数(PDF)、概率密度函数标准误(PDF standard error)、风险函数(hazard)和风险函数标准误(hazard standard error)。

结果表明,置入支架患者 1 年未发生 MACE 事件率 97.87%,MACE 事件发生率 2.13%,其他依此类推。

三、生存曲线(图 17-2)比较

Log-rank 检验基本思想是当 H_0(各组各时点生存率均相等)成立时,根据 t_i 时点的死亡率,可计算出各组的理论死亡数,则 χ^2 统计量计算公式为:

$$\chi^2 = \frac{[\Sigma w_i(d_{gi} - T_{gi})]^2}{V_g} \qquad \nu = g - 1$$

式中 d_{gi} 和 T_{gi} 分别表示各组在时间点 t_i 上的实际死亡数和理论死亡数,V_g 为第 g 组理论死亡数 T_g 的方差估计值

$$V_g = \sum w_i^2 \frac{n_{gi}}{n_i}(1 - \frac{n_{gi}}{n_i})(\frac{n_i - d_i}{n_i - 1})d_i$$

式中 w_i 为权重,对 log-rank 检验,$w_i = 1$。当比较两总体生存曲线呈比例时,检验效能最大;$w_i = n_i$ 则对应 Gehan 检验(1965)或 Wilcoxon 检验,该检验给实际死亡数与理论死亡数的早期差别更大的权重

【例 17-3】 观察 20 例急性心肌梗死患者置入支架治疗发生 MACE 事件的时间(月),将 20 例患者随机分为两组,应用两种不同的支架 A、B,比较 MACE 事件发生率是否存在差别。

应用 A 支架发生 MACE 事件时间:14 18 20$^+$ 22 24 24$^+$ 25 27$^+$ 28 38

应用 B 支架发生 MACE 事件时间:18$^+$ 20 22 27 28 28$^+$ 30 32$^+$ 34 40

log-rank 检验和 Wilcoxon 检验是生存率比较的非参数方法,由于能对各组的生存曲线做整体比较,实际工作中应用较多。其中,log-rank 检验给组间死亡的远期差别更大的权重,即对远期差异敏感;Wilcoxon 检验给组间死亡的近期差别更大的权重,即对近期差异敏感。本例选用 log-rank 检验。

程序名为:CT17-3

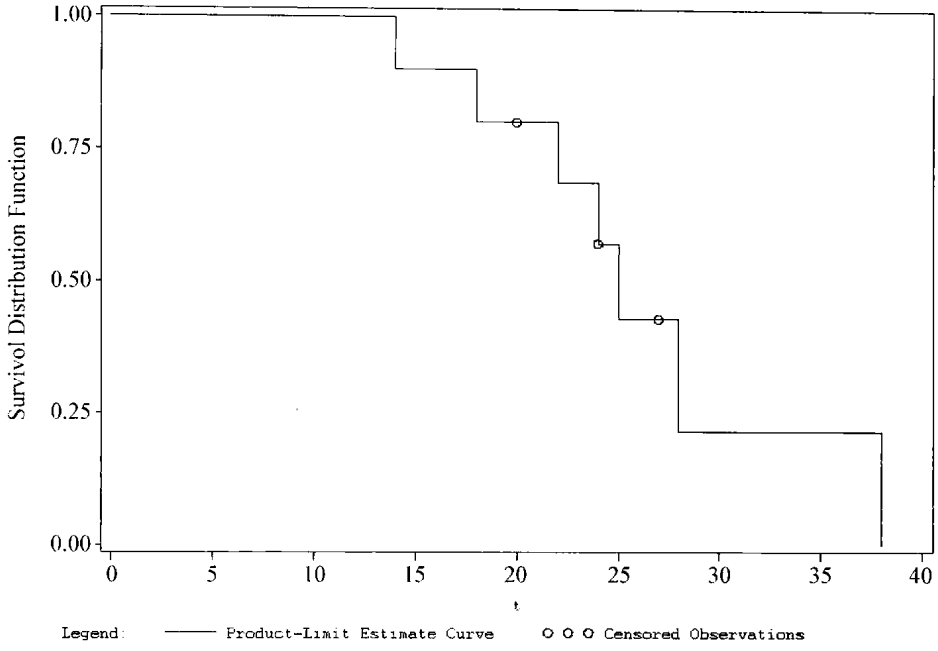

图 17-2　寿命表法绘制生存曲线

```
DATA CT17-3;
input group $ n;
do i=1 to n;
input t status @@;
output;end;
cards;
A 10
14 1 18 1 20 0 22 1 24 1 24 0 25 1 27 0
28 1 38 1
B 10
18 0 20 1 22 1 27 1 28 1 28 0 30 1 32 0
34 1 40 1;
run;
PROC LIFETESTdata=CT17-3 method=pl plots
=(s);
time t*status(0);
strata group;
run;
```

输出结果及结果解释：

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for t over Strata

Rank Statistics		
group	Log-Rank	Wilcoxon
A	1.963 9	27.000
B	−1.963 9	−27.000

Covariance Matrix for the Log-Rank Statistics		
group	A	B
A	2.854 24	−2.854 24
B	−2.854 24	2.854 24

Covariance Matrix for the Wilcoxon Statistics		
group	A	B
A	484.571	−484.571
B	−484.571	484.571

Test of Equality over Strata			
Test	Chi-Square	*DF*	*Pr* >Chi-Square
Log-Rank	1.351 3	1	0.245 1
Wilcoxon	1.504 4	1	0.220 0
−2Log(LR)	0.079 3	1	0.778 3

生存曲线组间比较结果(图 17-3),包括 log-rank 检验和 Wilcoxon 检验的秩统计量(rank statistics)、协方差矩阵(covariance matrix)、χ^2 统计量和 P 值。

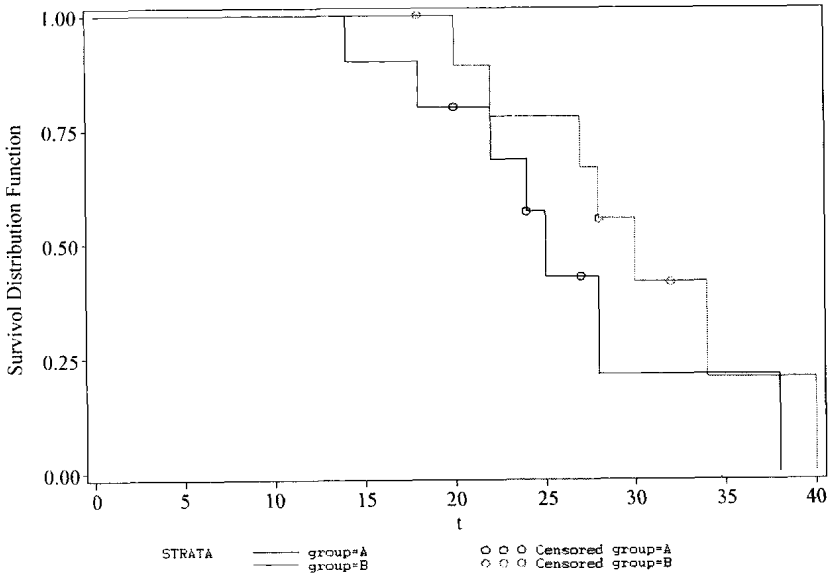

图 17-3 两种支架 MACE 事件发生情况绘制生存曲线

log-rank 检验结果 Chi-Square 为 1.35,自由度为 1,P 值为 0.245 1。尚不能认为两条生存曲线不同。

四、Cox 回归模型

目前对生存资料的多因素分析最常用的方法是 Cox 比例风险回归模型(proportional hazards regression model),简称 Cox 模型。该模型是一种多因素的生存分析方法,它可同时分析众多因素对生存期的影响,分析带截尾生存时间的资料,且不要求估计资料的生存分布类型。由于上述优良性质,该模型自英国统计学家 D. R. Cox 于 1972 年提出以来,在医学随访研究中得到非常广泛的应用。

假定 X_1, $X_2\cdots X_p$ 为协变量或影响因素;$h(t)$ 为具有协变量 X_1, $X_2\cdots X_p$ 的个体在 t 时刻的风险函数或瞬时死亡率,表示生存时间已达 t 的人在 t 时刻的瞬时死亡率;$h_0(t)$ 为 t 的未知函数,即 $X_1=X_2=\cdots=X_p=0$ 时 t 时刻的风险函数,称为基准风险函数(baseline hazard function)。变量 X_1 的作用是使个体的风险函数由 $h_0(t)$ 增至 $h_0(t)\exp(\beta_1)$;变量 X_2 的作用是使个体的风险函数由 $h_0(t)$ 增至 $h_0(t)\exp(\beta_2)$;X_1, $X_2\cdots X_pp$ 个变量共同影响下的风险函数为 $h(t)=h_0(t)\cdot\exp(\beta_1X_1)\cdot\exp(\beta_2X_2)\cdots\exp(\beta_pX_p)$,使风险函数由 $h_0(t)$ 变为 $h_0(t)$ 的 $\exp(\beta_1X_1)\cdot\exp(\beta_2X_2)\cdots\exp(\beta_pX_p)$ 倍,故 Cox 模型是一种乘法模型。

基本 Cox 模型表达式为

$$h(t)=h_0(t)\exp(\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)$$

β_1, $\beta_2\cdots\beta_p$ 为各协变量所对应的回归系数,需由样本资料作出估计

任意两个个体风险函数之比,即风险比(hazard ratio,HR)

$$HR=\frac{h_i(t)}{h_j(t)}=\frac{h_0(t)\exp(\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_pX_{ip})}{h_0(t)\exp(\beta_1X_{j1}+\beta_2X_{j2}+\cdots+\beta_pX_{jp})}$$
$$=\exp[\beta_1(X_{i1}-X_{j1})+\beta_2(X_{i2}-X_{j2})+\cdots+\beta_p(X_{ip}-X_{jp})]$$

$i\neq j$, $i,j=1,2\cdots n$

该比值保持一个恒定的比例,与时间 t 无关,称为比例风险(proportional hazards)假定,简称 PH 假定,即模型中协变量的效应不随时间改变而改变。

对上式两边取对数

$$\ln(HR)=\ln\left[\frac{h_i(t)}{h_j(t)}\right]=\beta_1(X_{i1}-X_{j1})+\beta_2(X_{i2}-X_{j2})+\cdots+\beta_p(X_{ip}-X_{jp})$$

式中,左边为风险比的自然对数,右边为协变量变化量与相应回归系数的线性组合。故 β_j ($j=1,2\cdots p$)的统计学意义是,在其他变量相同条件下,变量 X_j 每变化一个单位所引起的风险比的自然对数,或使风险函数成为原来数值的 $\exp(\beta_j)$ 倍

当 $\beta_j>0$ 时,$\exp(\beta_j)$ 或 $HR>1$,说明 X_j 增加时,风险函数增加,即 X_j 为危险因子;当 $\beta_j<0$ 时,$\exp(\beta_j)$ 或 $HR<1$,说明 X_j 增加时,风险函数下降,即 X_j 为保护因子;当 $\beta_j=0$ 时,$\exp(\beta_j)$ 或 $HR=1$,说明 X_j 增加时,风险函数不变,即 X_j 是与危险无关的因子。

参数估计:

$$L(\beta)=\prod_{i=1}^{D}\frac{\exp\left[\sum_{k=1}^{p}\beta^TZ_{(i)k}\right]}{\sum_{j\in R_i}\exp\left[\sum_{k=1}^{p}\beta^TZ_{(j)k}\right]}$$

其中,D 为死亡(终点事件)发生的个体,Z 为个体各个变量实际取值,R_i 为观测时间不小

于 t_i 的个体的集合

似然方程组为：

$$\frac{\partial \ln L(\beta)}{\partial \beta_k}=0\ ,\quad k=1,\cdots,p$$

所得之解即为极大似然估计。

在建立 Cox 回归方程时，风险比例可能会随时间变化而变化，即有些危险因素作用的强度随时间而变化，这样的资料是不适合前面所讲的一般的 Cox 回归模型的。对 Cox 模型的简单形式的直接扩展是引入时间相依协变量(time-dependent covariates)。此时的模型变为：

$$h(t\mid(t))=h_0(t)\exp\Big[\sum_{k=1}^{p}\beta_k Z_k(t)\Big]$$

参数估计：

$$L(\beta)=\prod_{i=1}^{D}\frac{\exp\Big[\sum_{k=1}^{p}\beta^T Z_{(i)k}(t)\Big]}{\sum_{j\in R_i}\exp\Big[\sum_{k=1}^{p}\beta^T Z_{(j)k}(t)\Big]}$$

回归系数的检验方法有 3 种。①Score 检验：常用于模型中新变量的引入；② Wald 检验：常用于模型中不重要变量的剔除；③似然比检验：常用于模型中不重要变量的剔除和新变量的引入。以上 3 种检验方法均为 χ^2 检验，自由度为模型中待检验的参数个数。

【例 17-4】 某研究欲考察某类型心脏病患者置入起搏器的预后生存时间和结局情况，以及两种起搏器是否存在差别，统计接受治疗的 60 例患者预后生存情况，x_1：起搏器种类(0：A 类；1：B 类)，x_2：年龄，x_3：bmi，x_4：糖尿病史(0：无，1：有)，x_5：高血压史(0：无，1：有)。t：生存时间(月)；status：生存结局，死亡＝1，截尾＝0。数据如表 17-3 所示。

表 17-3　60 例受试对象基线信息、疾病史及术后生存时间及结局 (%)

id	x_1	x_2	x_3	x_4	x_5	t	status	id	x_1	x_2	x_3	x_4	x_5	t	status
1	0	81	20.6	0	1	61	1	31	1	67	26.6	1	0	74	1
2	0	69	19.8	0	0	87	0	32	1	86	24.4	0	0	24	0
3	0	68	27.0	0	1	72	0	33	1	78	28.5	0	1	44	1
4	0	80	29.1	0	0	60	1	34	1	67	25.7	0	1	72	1
5	0	63	23.6	1	1	27	1	35	1	77	27.8	0	1	83	1
6	0	79	25.7	1	0	67	1	36	1	86	21.2	1	0	13	0
7	0	82	30.0	0	0	71	1	37	1	75	27.3	0	0	41	0
8	0	67	23.2	0	1	90	1	38	1	87	23.9	0	1	52	1
9	0	64	23.7	0	0	40	0	39	1	73	20.7	0	1	49	1
10	0	81	28.7	0	1	45	1	40	1	66	20.6	0	1	81	1
11	0	62	19.9	0	1	80	0	41	1	69	23.0	1	0	72	1
12	0	77	27.7	0	0	50	1	42	1	64	20.2	0	0	91	0
13	0	86	28.4	0	1	39	1	43	1	84	20.7	0	1	36	1
14	0	64	21.0	1	1	34	1	44	1	65	17.6	0	1	82	1
15	0	84	24.7	0	0	67	1	45	1	60	21.5	0	0	97	0
16	0	71	28.1	1	0	43	1	46	1	63	22.5	1	0	48	1

（续　表）

id	x_1	x_2	x_3	x_4	x_5	t	status	id	x_1	x_2	x_3	x_4	x_5	t	status
17	0	67	19.1	0	1	57	1	47	1	77	24.2	1	1	21	1
18	0	81	24.9	0	0	61	1	48	1	82	19.7	0	0	37	0
19	0	78	22.7	1	1	32	1	49	1	68	24.2	0	0	69	0
20	0	69	29.2	0	1	67	1	50	1	70	28.9	0	1	71	1
21	0	63	26.1	0	0	90	0	51	1	75	23.2	0	1	42	1
22	0	71	23.9	0	0	87	1	52	1	66	24.6	0	1	92	1
23	0	79	22.2	1	1	22	1	53	1	80	26.7	1	0	68	1
24	0	68	26.8	0	0	87	0	54	1	87	22.0	0	0	38	0
25	0	63	33.1	0	1	78	1	55	1	79	25.4	0	0	90	1
26	0	77	27.8	1	1	27	1	56	1	71	25.1	1	1	40	1
27	0	69	29.7	0	0	72	0	57	1	80	29.9	0	1	32	1
28	0	80	24.1	1	1	26	1	58	1	66	26.5	1	0	72	1
29	0	87	25.8	0	0	16	0	59	1	81	26.1	1	1	19	1
30	0	67	26.2	1	0	91	1	60	1	66	25.2	1	1	51	1

研究者欲分析影响心脏病患者生存时间长短的因素，包括置入起搏器的种类、年龄、bmi 指数、糖尿病史、高血压史，并根据影响因素进行不同时间点上生存率的预测。

Cox 模型属比例风险模型簇，其基本假定之一是比例风险假定。只有满足该假定前提下，基于此模型的分析预测才是可靠有效的。检查某协变量是否满足 PH 假定，最简单的方法是观察按该变量分组的Kaplan-Meier生存曲线，若生存曲线交叉，提示不满足PH假定。第二种方法是绘制按该变量分组的 $\ln[-\ln\hat{S}(t)]$ 对生存时间 t 的图，曲线应大致平行或等距。如各协变量均满足或近似满足 PH 假定，可直接应用基本 Cox 模型。

表 17-3 数据中，年龄为连续性变量，将年龄转化为两分类变量（<80 岁，即低龄老年人和≥80 岁，高龄组），bmi 指数分为（<=25 和>25）、使用起搏器种类、糖尿病史、高血压病史个变量的生存曲线见图 17-4。图中 X_1，X_3-1 两个变量绘制生存曲线是交叉的，其余 3 个变量满足 PH 假定，本例中仍假定数据满足 PH 假定，使用 COX 回归分析模型。

程序名为 CT17-4：

PHREG 过程是实现 Cox 模型的标准过程，其中 MODEL 语句是必需语句。MODEL 语句左边为生存时间和生存结局变量（括号内为截尾值），右边为协变量。该语句比较重要的选项有：

（1）TIES= DISCRETE | EXACT | BRESLOW | EFRON，指定重合生存时间或称结点（ties）的处理方法。DISCRETE 和 EXACT 为精确法，DISCRETE 法假定事件确实发生在相同的时间，而 EXACT 法假定结点来自连续的无结点资料。BRESLOW 法（1974，缺省值）和 EFRON 法（1977）是精确法的近似。关于四种方法的选择，没有结点时，4 种方法结果相同；结点比例不是很大时，4 种方法结果相近；结点比例很大时，两种近似结果有偏性，考虑计算耗时，可选 EFRON 近似法。

（2）SELECTION= FORWARD | BACKWARD | STEPWISE | NONE | SCORE，指定变量筛选方法，分别表示前进法、后退法、逐步法、全回归模型（缺省值）和最优子集法。

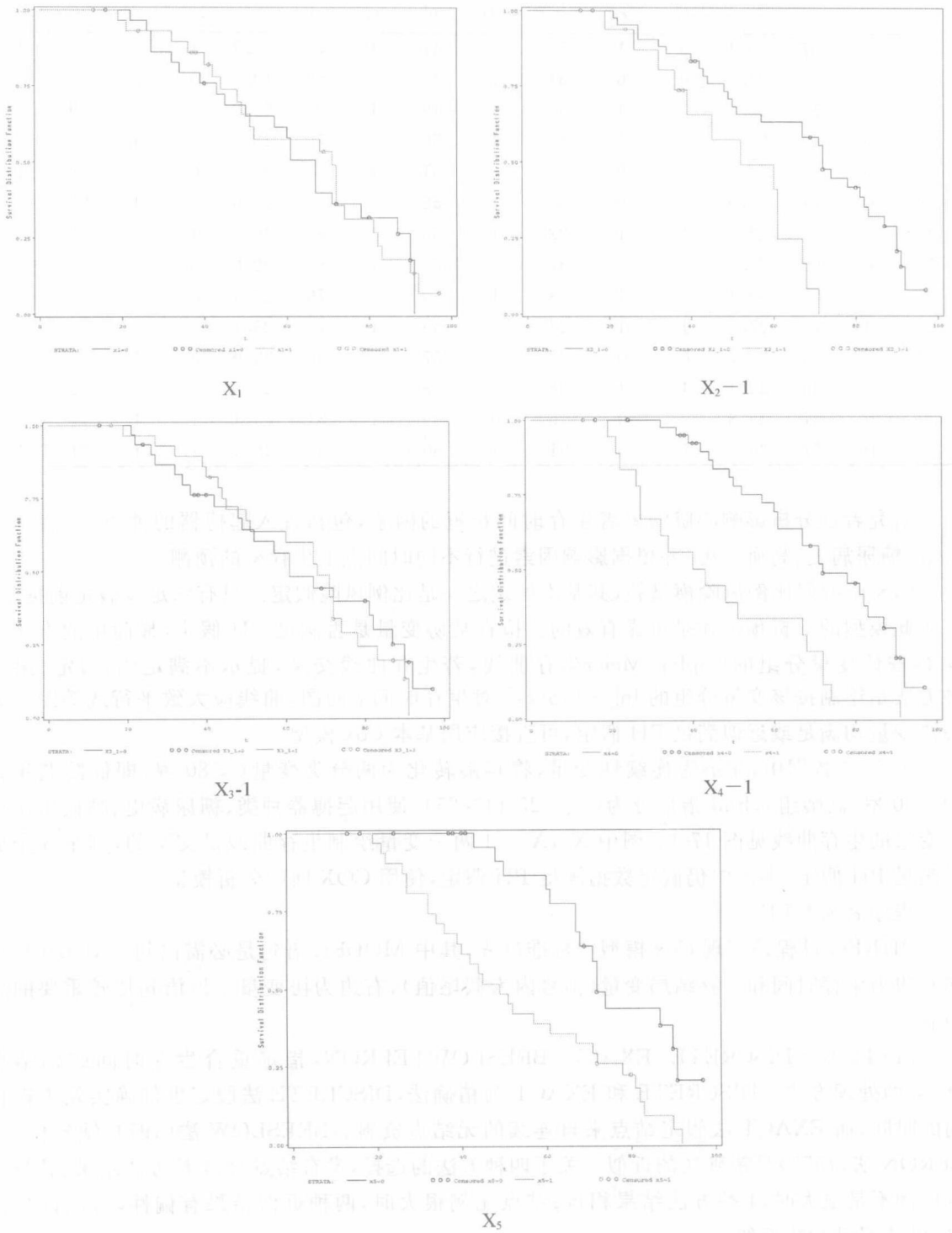

图 17-4　5 个变量的生存曲线

（X_1 起搏器种类；X_2-1 年龄；X_3-1bmi；X_4 糖尿病史；X_5 高血压病史）

(3)SLE=和 SLS= 分别指定引入和剔除变量的显著性水平 α。缺省值为 $\alpha=0.05$。

(4)RL 要求输出风险比 HR 的 95%置信区间。

OUTPUT 语句创建一个新的 SAS 数据集，含有为每一个观测计算的一些统计量，SAS 为每一个统计量定义一个关键字，如生存率和预后指数分别用 SURVIVAL 和 XBETA 表示。选项 ORDER=DATA 规定输出的数据集中的观测顺序与输入数据集中的顺序一致；METHOD=PL|CH|EMP 规定用于计算生存率的方法，PL 表示生存率的乘积限法(缺省值)，CH 和 EMP 表示生存率的经验累积风险函数估计法。

程 序	说 明
DATA CT17-4; INPUT id x_1-x_5 t status @@; X_2-1='';if x_2<80 then x_2-1=0; else x_2-1=1; X_3-1='';if x_3<=25 then x_3-1=0; else x_3-1=1; CARDS; /*输入表 17-4 的数据*/; Run; Ods html;	以下是建立数据集 t 为生存时间，status 为生存结局 将连续变量离散化 输入数据

PROC LIFETEST method=km plot=(s,ls,lls) graphics ;
time t * status(0); strata x_1; run;
PROC LIFETEST method=km plot=(s,ls,lls) graphics ;
time t * status(0); strata x_2-1; run;
PROC LIFETEST method=km plot=(s,ls,lls) graphics ;
time t * status(0); strata x_3-1; run;
PROC LIFETEST method=km plot=(s,ls,lls) graphics ;
time t * status(0); strata x_4; run;
PROC LIFETEST method=km plot=(s,ls,lls) graphics ;
time t * status(0); strata x_5; run;
	验证等比例条件是否成立，由于 x_2，x_3 为连续型变量，因此将其离散化后再绘制 Kaplan-Meier 曲线图

PROC PHREG; MODEL t * status(0)=x_1 x_2 x_3 x_4 x_5/SELECTION=STEPWISE SLE=0.05 SLS=0.05 RL; OUTPUT OUT=report SURVIVAL=s XBETA=pi/ORDER=DATA METHOD=PL; Run; PROC PRINT DATA=report; RUN; Ods html close;	调用 phreg 过程进行 Cox 回归分析 SELECTION=STEPWISE，指定变量筛选方法为逐步法，缺省则为全回归模型。“SLE=”和“SLS=”分别指定引入和剔除变量的显著性水平 α。缺省值为 $\alpha=0.05$。RL 要求输出风险比 HR 的 95%置信区间

PHREG 过程的其他语句有：①STRATA 语句用于建立分层 Cox 模型。格式为，STRATA 分层变量。注意分层变量不宜太多，否则每层观察单位数太少，会影响分析结果。②BASELINE语句用于输出对原有数据集以外某新数据集中观测的生存预测。格式为，BASELINE OUT=输出数据集名 COVARIATES=新数据集名，关键字=变量名/选项。新数据集中的变量必须与最后 Cox 模型中的变量相对应。常用关键字有 SURVIVAL(生存率)和 XBETA(预后指数)等。该语句中“/”后的选项有 METHOD=PL|CH|EMP，意义同 OUTPUT 语句。

主要分析结果及解释：

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	286.553	223.367
AIC	286.553	229.367
SBC	286.553	234.719

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	*DF*	*Pr* > ChiSq
Likelihood Ratio	63.186 6	3	<0.000 1
Score	56.927 3	3	<0.000 1
Wald	46.807 4	3	<0.000 1

Analysis of Maximum Likelihood Estimates								
Variable	*DF*	Parameter Estimate	Standard Error	Chi-Square	*Pr* > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
x_2	1	0.137 27	0.025 79	28.334 3	<0.000 1	1.147	1.091	1.207
x_4	1	2.423 91	0.435 79	30.936 7	<0.000 1	11.290	4.806	26.524
x_5	1	2.180 77	0.438 98	24.679 7	<0.000 1	8.853	3.745	20.929

Summary of Stepwise Selection						
Step	Variable		Number In	Score Chi-Square	Wald Chi-Square	*Pr* > ChiSq
	Entered	Removed				
1	x_2		1	18.934 4	.	<0.000 1
2	x_4		2	18.464 1	.	<0.000 1
3	x_5		3	30.355 2	.	<0.000 1

Obs	*t*	status	x_1	x_2	x_3	x_4	x_5	pi	s
1	61	1	0	81	20.6	0	1	13.299 3	0.146 69
2	74	1	1	67	26.6	1	0	11.620 7	0.357 91
3	87	0	0	69	19.8	0	0	9.471 3	0.798 49
4	24	0	1	86	24.4	0	0	11.804 8	0.988 58
5	72	0	0	68	27.0	0	1	11.514 8	0.437 80
6	44	1	1	78	28.5	0	1	12.887 5	0.638 88
7	60	1	0	80	29.1	0	0	10.981 3	0.855 94
8	72	1	1	67	25.7	0	1	11.377 6	0.486 73
9	27	1	0	63	23.6	1	1	13.252 4	0.878 74
10	83	1	1	77	27.8	0	1	12.750 2	0.006 38

逐步法第一步、第二步和第三步分别引入变量 x_2，x_4 和 x_5，SAS 输出每一步的模型拟合统计量（−2 LOG L，AIC 和 SBC）及模型检验结果（似然比检验、Score 检验和 Wald 检验），之后是模型的最大似然估计，包括参数估计（parameter estimate）、估计值标准误（standard Error）、Waldχ^2（Chi-Square）、P 值（$Pr>$ChiSq）、风险比 HR 及 HR95%置信区间（hazard ratio confidence limits）。

Cox 模型结果显示：年龄、糖尿病史、高血压病史均为置入起搏器的心脏病患者发生死亡的危险因素。3 个变量的回归系数均为正值，提示年龄＞80 岁、有糖尿病史和高血压病史的患者死亡的概率增高。糖尿病史，高血压病史不变的情形下，年龄每增加 1 岁，死亡风险增加 0.147 倍；年龄与高血压病史不改变情形下，有糖尿病史的患者死亡风险是无糖尿病史患者死亡风险的 4.806 倍；年龄与年龄病史不改变情形下，有高血压病史的患者死亡风险是无高血压病史患者死亡风险的 8.85 倍，增加了 7.85 倍。由 Cox 回归分析结果，得出风险函数的表达式为：

$$h(t)=h_0(t)\exp(0.137\,27\times x_2+2.423\,91\times x_4+2.180\,77\times x_5)$$

此表达式右边指数部分取值越大，则风险函数 $h(t)$ 越大，预后越差，故称为预后指数（prognostic index，PI）。

生存率可由下式估计

$$S(t)=[S_0(t)]^{\exp(\sum\hat{\beta}_iX_i)}$$

式中 $\hat{S}_0(t)$ 为基准生存率，可采用 Breslow 估计，

$$S_0(t)=\prod_{t_{(j)}\le t}\left[\exp\frac{-d_j}{\sum_{i\in R_j}\exp(\sum\hat{\beta}_iX_i)}\right]$$

式中 $\prod$ 为连乘积符号，$t_{(j)}$ 为排序后的完全生存时间，d_j 为 t_j 处死亡数，R_j 为 t_j 处风险集

输出结果显示了前 10 例患者的预后指数 PI 及其所对应生存时间的生存率。如第 1 例患者 PI=13.29，61 个月生存率 14.67%。

五、参数回归

生存时间数据分析的一个重要内容是模型拟合或分布拟合，描述生存时间分布的模型通常有指数分布、Weibull 分布、对数正态分布、Gamma 分布等，常见生存时间分布的概率密度函数 $f(t)$、生存函数 $S(t)$ 和风险函数 $h(t)$ 见表 17-4。实际对生存数据作分布拟合时，可用上述模型分别进行拟合，根据拟合优度检验的结果选择适当的模型。但是，对于一批生存数据，事先不知道生存时间分布的总体趋势，也不好判断应该用什么样的模型最合适，这时许多研究者一般直接采用非参数方法或半参数法。

但是，如果一批数据确实符合某特定的参数模型，由于非参数方法的精度一般低于参数方法，因此，按照非参数方法进行的分析就不能有效地利用和阐述样本数据所包含的信息，同时它对样本量的要求也高于参数方法。

（一）指数模型

指数分布是一种纯随机死亡模型，在任何时间上的风险函数为一常数，即风险函数的大小不受生存时间长短的影响，以独特的“无记忆性”而闻名。λ 为指数分布的风险率，称为刻度参数或尺度参数，其大小决定了生存时间的长短。风险率越大，生存率下降越快，生存时间越短；

风险率越小，生存时间越长。

表 17-4　常见生存时间分布的概率密度函数 $f(t)$、生存函数 $S(t)$ 和风险函数 $h(t)$

分布	$f(t)$	$S(t)$	$h(t)$
指数分布	$\lambda\exp(-\lambda t)$	$\exp(-\lambda t)$	λ
Weibull 分布	$\lambda\gamma(t)^{\gamma-1}\exp[-\lambda(t)^{\gamma}]$	$\exp[-\lambda(t)^{\gamma}]$	$\lambda\gamma(t)^{\gamma-1}$
Gamma 分布	$\frac{\lambda^{\gamma}t^{\gamma-1}\exp(-\lambda t)}{\Gamma(\gamma)}$	$1-\mathrm{I}(\lambda t,\gamma)$	$\frac{f(t)}{S(t)}$
对数正态分布	$\frac{\exp[-\frac{1}{2}(\frac{\ln t-\mu}{\sigma})^{2}]}{t(2\pi)^{1/2}\sigma}$	$1-\Phi[\frac{\ln t-\mu}{\sigma}]$	$\frac{f(t)}{S(t)}$
对数 Logistic 分布	$\frac{\gamma t^{\gamma-1}\lambda}{[1+\lambda t^{\gamma}]^{2}}$	$\frac{1}{1+\lambda t^{\gamma}}$	$\frac{\gamma t^{\gamma-1}\lambda}{1+\lambda t^{\gamma}}$
广义 Gamma 分布	$\frac{\gamma\lambda^{k}t^{\gamma k-1}\exp(-t\lambda^{\gamma})}{\Gamma(k)}$	$1-\mathrm{I}[\lambda t^{\gamma},k]$	$\frac{f(t)}{S(t)}$

指数回归是在指数分布的基础上引入预后因素后的模型。设 X_1，$X_2\cdots X_p$ 为影响因素，如果生存时间服从指数分布，则参数 λ 与各因素间的关系可用如下回归方程表示：

$$\ln\lambda=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$$

风险函数为：$h(t)=\lambda=\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_P)$

相应 t 时刻的生存率为：$S(t)=\exp[-t*\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p)]$

(二)Weibull 模型

Weibull 分布也是生存分析的理论基础，由瑞典科学家 Waloddi Weibull 提出。Weibull 分布是指数分布的一种推广形式，它不像指数分布假定危险率是常数，因而有更广的应用性。

λ 和 γ 为两个参数。λ 称为尺度参数，它决定分布的分散度；γ 为形状参数，它决定该分布的形态。$\gamma>1$ 时风险函数随时间单调递增；$\gamma<1$ 时风险函数随时间单调递减；显然，当 $\gamma=1$ 时，风险不随时间变化，Weibull 分布退化为指数分布，所以指数分布是 Weibull 分布在 $\gamma=1$ 时的特例。

(三)Gamma 模型

生存分析讨论两类不同的 Gamma 模型：标准 Gamma 模型(2 参数)和广义 Gamma 模型(3 参数)。标准 Gamma 分布的特性取决于两个参数 γ 和 λ，γ 为形状参数，λ 为尺度参数。当 $0<\gamma<1$ 时，若时间从 0 增加到无穷时，风险函数从无穷单调地减小到 λ，表现为负老化；当 $\gamma>1$ 时，若时间从 0 增加到无穷时，风险函数从 0 增加到 λ，表现为正老化；当 $\gamma=1$ 时，风险等于常数 λ，即指数分布情形。

广义 Gamma 模型比我们之前考虑的其他模型多一个参数，它的风险函数可呈现更多的形状。特别地，它可以是 U 形或浴盆形的风险函数，在这样的函数中风险先下降，下降到最小值后又升高。众所周知，人类在整个生命周期中的死亡危险性就属于这种形状。

一般地，似然比统计量用于比较嵌套模型。如果限制模型 B 中的参数可得到模型 A，那么模型 A 嵌套于模型 B。比如，指数模型同时嵌套于 Weibull 模型和标准 Gamma 模型。当 Weibull 模型的 $\gamma=1$ 时，或当标准 Gamma 模型的形状参数和尺度参数都=1 时，便得到指数模型。如果模型 A 嵌套于模型 B，可以通过取两模型对数似然值的正差值的 2 倍来评价 A 模型的拟合优度。

广义 Gamma 分布是一个相当灵活的三参数分布族，指数模型（$\lambda=\gamma=1$）、Weibull 模型（$\gamma=1$）和标准 Gamma 模型（$\lambda=\gamma$）都是广义 Gamma 模型的特例。可据此进行参数回归模型的拟合优度检验。

【例 17-5】　在 17 年里追踪调查了 149 位糖尿病患者，数据见表 17-5。变量及其赋值如下，试进行患者生存时间的影响因素分析并进行生存预测。

结局（status，1 表示死亡，0 表示截尾）；生存时间（t，年）；随访开始时年龄（age1，岁）；体重指数（BMI）；诊断出糖尿病时的年龄（Age0，岁）；吸烟状况（smk，0 表示不吸烟；1 表示曾吸烟；2 表示吸烟）；收缩压（SBP，mmHg）；舒张压（DBP，mmHg）；心电图读数（ECG，0 表示正常；1 表示可疑；2 表示异常）；病人是否有冠心病（CHD，0 表示无；1 表示有）。

表 17-5　149 位糖尿病患者生存资料

Id	status	t	Age1	BMI	Age0	smk	SBP	DBP	ECG	CHD
1	0	12.4	44	34.2	41	0	132	96	0	0
2	0	12.4	49	32.6	48	2	130	72	0	0
3	0	9.6	49	22.0	35	2	108	58	0	1
4	0	7.2	47	37.9	45	0	128	76	1	1
5	0	14.1	43	42.2	42	2	142	80	0	0
6	0	14.1	47	48.1	44	0	156	94	0	0
7	0	12.4	50	36.5	48	0	140	86	1	1
8	0	14.2	36	38.5	33	2	144	88	0	0
9	0	12.4	50	41.5	47	1	134	78	0	1
10	0	14.5	49	34.1	45	0	102	68	0	0
11	0	12.4	50	39.5	48	2	142	84	0	0
12	0	10.8	54	42.9	43	0	128	74	0	0
13	1	10.9	42	29.8	36	2	156	86	0	0
14	0	10.3	44	48.2	43	2	102	58	0	0
15	1	13.6	40	27.5	26	2	146	98	0	0
16	0	11.9	48	25.3	48	0	120	68	1	1
17	0	12.5	50	31.6	44	1	142	76	0	0
18	0	5.9	47	26.3	38	1	144	82	0	0
19	0	12.4	38	32.4	36	2	150	98	1	1
20	0	14.1	35	47.0	33	1	134	78	0	0
21	1	9.8	51	26.5	47	2	130	76	0	0
22	0	7.2	40	43.9	34	0	122	92	0	0
23	0	3.5	54	32.3	52	1	132	80	0	0
24	0	0.0	53	34.5	47	2	150	88	2	1
25	1	12.1	45	18.9	40	1	134	98	0	0
26	0	1.9	41	32.0	31	1	142	90	1	1
27	0	8.6	34	48.9	30	2	124	66	0	0
28	0	14.0	38	23.7	28	0	102	60	0	0
29	0	14.3	43	24.8	43	0	134	80	0	0
30	0	12.4	45	26.6	41	2	118	66	1	1
31	0	12.4	40	39.2	35	2	192	108	0	0

（续 表）

Id	status	t	Agel	BMI	Age0	smk	SBP	DBP	ECG	CHD
32	0	14.4	44	32.7	36	2	122	78	0	0
33	0	14.2	48	48.5	43	1	122	92	0	0
34	0	14.5	51	31.2	49	2	112	74	0	0
35	0	12.4	36	24.2	30	2	142	90	0	0
36	0	14.3	52	31.6	48	1	152	96	0	0
37	1	13.7	41	30.7	39	2	112	74	0	0
38	0	13.4	49	28.0	35	2	118	84	0	0
39	0	12.5	44	32.0	29	0	152	88	0	0
40	0	14.4	37	32.7	36	2	136	88	0	0
41	0	12.6	51	24.2	42	2	134	90	0	0
42	0	13.8	47	18.7	42	0	130	78	1	1
43	0	14.0	45	25.6	36	0	108	72	0	0
44	0	6.8	38	22.8	27	2	126	66	1	1
45	0	12.4	35	30.1	33	0	132	78	0	0
46	0	12.9	50	27.7	49	1	144	88	0	0
47	0	8.9	53	27.6	49	2	126	68	0	0
48	0	12.4	48	28.1	47	1	128	70	0	0
49	0	14.5	40	31.7	37	2	132	82	0	0
50	0	13.0	43	26.1	42	2	128	80	0	0
51	0	13.4	54	30.8	54	1	142	80	1	1
52	0	10.6	52	36.9	50	1	132	80	1	1
53	0	13.9	69	24.2	63	1	148	78	0	0
54	0	16.9	38	27.5	26	2	170	100	0	0
55	0	3.6	50	27.3	44	1	140	90	0	0
56	0	10.2	64	30.1	58	0	138	76	1	1
57	0	15.7	44	36.1	41	0	112	78	0	0
58	0	12.0	38	43.1	39	2	140	78	0	0
59	1	6.7	62	34.6	58	0	138	78	2	1
60	0	11.6	47	39.0	45	0	130	82	0	0
61	1	2.0	78	28.7	77	0	178	86	1	1
62	0	10.2	49	28.2	43	2	158	80	0	0
63	0	3.6	63	25.1	46	1	168	88	2	1
64	0	15.4	71	26.0	59	0	146	88	0	0
65	0	11.3	51	32.0	49	2	128	76	0	0
66	0	10.3	59	28.1	57	1	132	76	0	1
67	0	5.8	50	26.1	49	1	154	80	0	0
68	1	8.0	66	45.3	49	0	154	92	0	0
69	0	14.6	42	30.0	41	1	122	80	0	0
70	0	11.4	40	35.7	36	2	144	76	1	1
71	0	7.2	67	28.1	61	0	178	96	0	0
72	0	5.5	86	32.9	61	0	162	60	0	0
73	0	11.1	52	37.6	46	1	142	80	0	0

（续　表）

Id	status	t	Agel	BMI	Age0	smk	SBP	DBP	ECG	CHD
74	0	16.5	42	43.4	37	0	120	76	0	0
75	0	10.9	60	25.4	60	0	124	64	0	0
76	0	2.5	75	49.7	57	1	174	82	1	1
77	1	10.8	81	35.2	81	0	142	88	0	0
78	0	4.7	60	37.3	39	0	160	78	0	0
79	1	5.5	60	26.0	42	0	122	68	2	1
80	0	4.5	63	21.8	60	2	162	98	0	1
81	0	9.0	62	18.2	43	0	132	72	1	1
82	0	6.8	57	34.1	41	2	116	60	2	1
83	1	3.6	71	25.6	54	1	152	84	2	1
84	0	12.1	58	35.1	45	0	144	68	1	1
85	0	8.1	42	32.5	28	1	98	68	2	1
86	0	11.1	45	44.1	40	0	138	76	0	1
87	1	7.0	66	29.7	59	1	138	78	0	0
88	0	1.5	61	29.2	54	0	184	80	1	1
89	0	11.7	48	25.2	30	2	158	98	0	0
90	0	0.3	82	25.3	50	0	176	96	0	1
91	0	13.6	35	25.8	34	1	118	72	0	0
92	0	15.0	57	48.7	57	2	172	98	0	0
93	0	11.2	56	39.5	55	1	182	100	0	1
94	0	3.0	49	32.9	48	0	144	90	1	1
95	0	13.7	50	37.1	50	0	142	80	0	0
96	0	10.2	53	35.3	53	2	154	76	0	0
97	0	12.4	71	29.3	70	0	122	60	0	0
98	0	1.1	55	22.1	33	2	222	102	1	1
99	0	16.3	69	23.6	43	0	150	80	0	1
100	0	6.7	59	26.1	55	2	142	66	0	0
101	0	15.4	47	32.5	45	2	128	82	0	0
102	1	7.6	75	29.8	67	0	122	76	2	1
103	1	3.6	80	24.4	80	1	162	88	1	1
104	0	11.5	57	26.3	54	0	172	82	1	1
105	0	13.5	52	30.8	46	2	132	70	0	1
106	0	10.6	48	29.4	46	0	112	68	0	0
107	1	6.5	57	29.1	47	1	138	92	1	1
108	1	14.3	58	30.1	56	0	128	74	0	0
109	0	11.6	51	31.0	37	2	132	78	0	0
110	0	15.4	33	34.0	33	2	120	78	0	0
111	0	11.0	36	38.1	33	1	122	70	0	0
112	1	11.0	52	37.0	46	0	140	98	0	0
113	1	4.8	64	31.2	57	2	172	88	2	1
114	0	14.8	31	38.8	29	1	136	76	0	0
115	0	1.8	69	22.3	56	0	152	74	2	1

（续 表）

Id	status	t	Agel	BMI	Age0	smk	SBP	DBP	ECG	CHD
116	0	15.8	59	25.0	58	0	126	80	0	0
117	0	14.1	38	31.3	38	2	104	58	0	0
118	0	4.6	49	59.7	49	1	142	82	0	0
119	0	15.5	49	34.0	41	0	128	76	0	0
120	0	7.2	68	29.4	66	1	122	58	2	1
121	0	14.5	40	43.2	41	1	122	70	0	0
122	0	10.5	36	35.1	32	2	122	68	0	0
123	0	14.3	60	37.0	54	0	122	70	0	0
124	1	2.2	74	27.1	54	1	168	84	1	1
125	0	5.0	61	27.6	51	0	162	82	0	0
126	0	12.4	54	25.2	51	0	116	76	0	0
127	0	1.1	35	25.8	34	2	126	82	0	0
128	0	15.4	46	32.2	42	2	180	98	0	0
129	0	14.3	40	41.6	41	2	132	98	0	0
130	0	15.6	53	39.8	52	0	150	88	0	0
131	1	12.5	66	26.6	54	1	106	70	0	1
132	0	12.3	61	48.3	55	0	154	88	0	0
133	0	14.8	41	27.7	38	1	122	76	0	0
134	0	10.2	64	26.6	51	2	130	68	0	0
135	0	12.3	41	25.0	38	2	120	58	0	0
136	0	10.3	46	54.3	45	1	144	86	0	0
137	0	8.5	80	29.4	79	1	134	60	0	1
138	0	10.2	63	48.1	60	1	148	80	1	1
139	1	10.0	72	27.3	68	1	170	78	2	1
140	0	7.3	41	36.9	33	0	160	92	1	1
141	1	15.3	52	40.2	36	0	154	96	0	0
142	0	14.0	53	32.7	48	2	124	76	1	1
143	0	15.8	61	48.2	57	1	130	70	0	0
144	0	11.4	53	41.4	47	1	156	78	0	0
145	1	5.5	75	35.8	66	0	162	78	0	0
146	0	11.0	40	34.0	38	2	132	76	0	0
147	0	7.3	61	19.9	37	0	120	60	1	1
148	1	10.6	62	30.6	49	0	160	86	1	1
149	0	10.5	49	30.8	47	1	146	86	0	0

资料来源：ET Lee(陈家鼎，戴中维译)生存数据分析的统计方法．中国统计出版社，1998：72-76

考虑到例 17-5 中收缩压和舒张压两个变量有一定的相关性，数据分析时取平均血压(MBP)，即令 MBP=SBP＊(1/3)+DBP＊(2/3)。程序名为 CT17-5。

程 序	说 明
DATA CT17-5;	建立数据集
input id status t agel bmi age0 smk sbp dbp ecg chd;	
mbp=sbp * (1/3)+dbp * (2/3);	
cards;	
/ * 输入表 17-4 的数据 * /;	取平均血压 输入数据
Run;	
ods html;	
PROC LIFEREG;	
class chd smk ecg;	
model t * status(0)=agel bmi age0 mbp chd smk ecg/ dist=exponential;	
PROC LIFEREG;	
class chd smk ecg;	拟合指数模型
model t * status(0)=agel bmi age0 mbp chd smk ecg/ dist=weibull;	
PROC LIFEREG;	
class chd smk ecg;	拟合 Weibull 模型
model t * status(0)=agel bmi age0 mbp chd smk ecg/ dist=gamma;	
run;	
PROC LIFEREG;	拟合广义 Gamma 模型
class ecg;	
model t * status(0)=agel mbp ecg/ dist=gamma;	
output out=a p=median std=s;	
run;	
PROC PRINT data=a;	保留有统计学意义的变量，重新进行分析
var t status -prob- median s;	
run;	
ods html close;	

PROC LIFEREG 过程对生存数据拟合参数模型，其最大特点在于可以处理右截尾、左截尾或区间截尾数据，同时含有丰富的生存分布形式，特别是其中的广义 Gamma 分布可以进行许多其他概率分布的似然比拟合优度检验。

CLASS 语句用于说明分类变量。

MODEL 语句指出哪些变量用于该模型的回归部分以及模型的误差项或随机项的分布是什么。MODEL 语句可用的选项：

(1)DISTRIBUTION|DIST|D=distribution-type(分布的类型)，说明生存时间的分布类型。exponential，weibull，Gamma，normal，Inormal，Logistic，Logistic 指定指数分布、Weibull 分布、Gamma 分布、正态分布、对数正态分布、Logistic 分布和对数 Logistic 分布。

(2)NOLOG 要求不对反应变量进行对数变换，缺省时 LIFEREG 过程对反应变量进行对数变换。

(3)SCALE=value(值)，要求尺度参数以这个值作为初始值。

(4)NOSCALE 要求尺度参数固定。

(5)SHAPE1= value(值),要求形状参数用规定的 value 值为初始值。

(6)NOSHAPE1 要求第一个形状参数 SHAPE1 保持固定。

OUTPUT 语句创建一个新 SAS 数据集,它包含模型拟合之后计算的统计量。

OUTPUT <OUT=SAS-data-set> keyword=name

OUT=SAS-data-set(SAS 数据集),命名输出数据集。keyword=name(关键词=名字),规定在 OUTPUT 数据集中包含的统计量(如下),并给出包含这些统计量的新名字。

①CONTROL 在输入数据集中命名用于控制分位数估计的变量。

②PREDICTED|P,命名存放分位数估计结果的变量。缺省时计算第 50 百分位数即中位生存时间。

③QUANTILES|Q,给出所要求计算的分位数列表。

④STD_ERR|STD,命名存放分位数标准差估计结果的变量。

⑤XBETA 命名存放 $x'b$ 计算结果的变量。

LIFEREG 过程也可以得到原始数据集外其他协变量值所对应的预测值。模型拟合前,将这些协变量值附加在原数据集后,生存时间设置为缺失值。这样这些观测不用于模型拟合,但可生成它们的预测值。如果只需要几个观察值(真实值或假想值)的预测,则生成一个变量如 USE,若需要预测,则该变量=1;否则,该变量=0。OUTPUT 语句中包括 CONTROL=USE。

主要分析结果及解释:

以下是程序 CT17-5 输出的主要结果及其解释。

指数模型输出结果:

Model Information	
Data Set	WORK.SASTJFX27-1
Dependent Variable	Log(t)
Censoring Variable	status
Censoring Value(s)	0
Number of Observations	148
Noncensored Values	24
Right Censored Values	124
Left Censored Values	0
Interval Censored Values	0
Zero or Negative Response	1
Name of Distribution	Exponential
Log Likelihood	−56.822 577 2

Analysis of Parameter Estimates

Parameter		*DF*	Estimate	Standard Error	95% Confidence Limits		Chi-Square	*Pr* > ChiSq
Intercept		1	10.191 8	2.385 7	5.516 0	14.867 7	18.25	<0.000 1
age1		1	−0.092 6	0.034 9	−0.161 0	−0.024 1	7.02	0.008 0
bmi		1	0.031 5	0.033 1	−0.033 1	0.096 4	0.91	0.341 0
age0		1	0.023 2	0.030 9	−0.037 3	0.083 7	0.57	0.451 9
mbp		1	−0.039 5	0.018 1	−0.074 9	−0.004 0	4.77	0.029 0
chd	0	1	−0.983 8	1.080 3	−3.101 1	1.133 5	0.83	0.362 4
chd	1	0	0.000 0	.	.	.	.	.
smk	0	1	0.066 9	0.639 2	−1.186 0	1.319 8	0.01	0.916 7
smk	1	1	0.041 7	0.655 3	−1.242 6	1.325 9	0.00	0.949 3
smk	2	0	0.000 0	.	.	.	.	.
ecg	0	1	2.106 7	1.097 0	−0.043 3	4.256 7	3.69	0.054 8

（续　表）

Analysis of Parameter Estimates								
Parameter		*DF*	Estimate	Standard Error	95% Confidence Limits		Chi-Square	*Pr* > ChiSq
ecg	1	1	0.813 4	0.641 0	−0.442 9	2.069 7	1.61	0.204 4
ecg	2	0	0.000 0	.	.	.	.	.
Scale		0	1.000 0	0.000 0	1.000 0	1.000 0		
Weibull Shape		0	1.000 0	0.000 0	1.000 0	1.000 0		

Weibull 模型输出结果：

Data Set	WORK. SASTJFX27-1
Dependent Variable	Log(t)
Censoring Variable	status
Censoring Value(s)	0
Number of Observations	148
Noncensored Values	24
Right Censored Values	124
Left Censored Values	0
Interval Censored Values	0
Zero or Negative Response	1
Name of Distribution	Weibull
Log Likelihood	−42.646 785 78

Analysis of Parameter Estimates								
Parameter		*DF*	Estimate	Standard Error	95% Confidence Limits		Chi-Square	*Pr* > ChiSq
Intercept		1	5.803 3	1.067 9	3.710 3	7.896 4	29.53	<0.000 1
age1		1	0.055 4	0.015 1	−0.085 0	−0.025 8	13.45	0.000 2
bmi		1	0.007 5	0.011 9	−0.015 8	0.030 8	0.40	0.528 9
age0		1	0.022 5	0.012 8	−0.002 7	0.047 6	3.07	0.079 6
mbp		1	−0.015 9	0.007 5	−0.030 6	−0.001 3	4.54	0.033 0
chd	0	1	−0.580 4	0.401 6	−1.367 4	0.206 7	2.09	0.148 4
chd	1	0	0.000 0	.	.	.	.	.
smk	0	1	0.166 1	0.239 6	−0.303 5	0.635 7	0.48	0.488 1
smk	1	1	0.049 9	0.248 1	−0.436 3	0.536 1	0.04	0.840 5
smk	2	0	0.000 0	.	.	.	.	.
ecg	0	1	1.224 0	0.426 9	0.387 4	2.060 7	8.22	0.004 1
ecg	1	1	0.290 8	0.254 3	−0.207 6	0.789 2	1.31	0.252 8
ecg	2	0	0.000 0	.	.	.	.	.
Scale		1	0.349 0	0.055 5	0.255 5	0.476 7		
Weibull Shape		1	2.865 3	0.455 9	2.097 7	3.913 9		

广义 Gamma 模型输出结果：

Data Set	WORK. SASTJFX27-1
Dependent Variable	Log(t)
Censoring Variable	status
Censoring Value(s)	0
Number of Observations	148
Noncensored Values	24
Right Censored Values	124
Left Censored Values	0
Interval Censored Values	0
Zero or Negative Response	1
Name of Distribution	Gamma
Log Likelihood	−36.856 591 16

Analysis of Parameter Estimates

Parameter		*DF*	Estimate	Standard Error	95% Confidence Limits		Chi-Square	*Pr* > ChiSq
Intercept		1	1.707 8	0.825 3	3.090 3	6.325 4	32.54	<0.000 1
age1		1	−0.049 5	0.011 4	−0.071 9	−0.027 1	18.73	<0.000 1
bmi		1	0.026 5	0.011 4	0.004 0	0.048 9	5.35	0.020 7
age0		1	0.008 7	0.009 5	−0.009 9	0.027 3	0.84	0.360 3
mbp		1	−0.013 5	0.006 6	−0.026 4	−0.000 6	4.21	0.040 3
chd	0	1	−0.473 4	0.270 7	−1.004 0	0.057 1	3.06	0.080 3
chd	1	0	0.000 0	.	.	.	.	.
smk	0	1	0.302 0	0.172 2	−0.035 4	0.639 4	3.08	0.079 4
smk	1	1	0.318 4	0.209 6	−0.092 5	0.729 3	2.31	0.128 8
smk	2	0	0.000 0	.	.	.	.	.
ecg	0	1	0.855 4	0.299 1	0.269 1	1.441 7	8.18	0.004 2
ecg	1	1	0.021 1	0.323 9	−0.613 7	0.656 0	0.00	0.947 9
ecg	2	0	0.000 0	.	.	.	.	.
Scale		1	0.416 5	0.151 4	0.204 3	0.849 4		
Shape		1	−2.752 8	1.661 1	−6.008 5	0.502 9		

在 PROC LIFEREG 中没有直接拟合标准 Gamma 模型的方法，但 PROC LIFEREG 可以将尺度参数和形状参数设定为特定值。若拟合标准 Gamma 模型，可试用许多不同的值（比如用直线搜索法），直到找到一个能使对数似然值达到最大的共同的尺度参数和形状参数。本例不再尝试。

现比对 3 个模型的拟合效果，可采用似然比检验，似然比统计量的公式为：

$$\chi^2_v = -2\log L_q - (-2\log L_{q+v})$$

式中 χ^2_v 服从自由度为 v 的 χ^2 分布，$-2\log L_q$ 和 $-2\log L_{q+v}$ 分别为含 q 和 $q+v$ 个参数的模型的对数似然函数值。因 Weibull 分布包含两个参数，指数分布包含 1 个参数，广义 Gamma 分布包含 3 个参数，标准 Gamma 分布包含 2 个参数。所以，各种分布拟合效果有无差异的假设检验结果可汇总如下，见表 17-6。

表 17-6　糖尿病资料似然比拟合优度检验

比较模型			自由度	χ^2 统计量	P 值
指数模型	VS	Weibull 模型	1	28.351	$P<0.05$
指数模型	VS	广义 Gamma 模型	2	40.390	$P<0.05$
Weibull 模型	VS	广义 Gamma 模型	1	12.039	$P<0.05$

可见，各分布的拟合效果之间均有统计学差异。故这里应选用－2logL 值最小的分布，即广义 Gamma 分布（－2logL 值为 73.255）。查看广义 Gamma 模型的输出结果，只有 Agel，BMI 和 ECG3 个变量有统计学差异。故仅保留这 3 个变量，重新采用广义 Gamma 模型来进行分析。所得结果如下：

Analysis of Parameter Estimates

Parameter		*DF*	Estimate	Standard Error	95% Confidence Limits		Chi-Square	*Pr* > ChiSq
Intercept		1	5.893 7	0.806 0	4.314 0	7.473 4	53.17	<0.000 1
agel		1	−0.034 1	0.006 5	−0.046 8	−0.021 4	27.73	<0.000 1
mbp		1	−0.017 7	0.006 3	−0.030 1	−0.005 3	7.88	0.005 0
ecg	0	1	0.757 1	0.251 8	0.263 5	1.250 6	9.04	0.002 6
ecg	1	1	0.376 0	0.282 1	−0.176 9	0.929 0	1.78	0.182 6
ecg	2	0	0.000 0	.	.	.	.	.
Scale		1	0.586 5	0.109 4	0.406 9	0.845 3		
Shape		1	−0.588 0	0.909 3	−2.370 1	1.194 1		

广义 Gamma 回归模型结果表明，随访开始时年龄和平均血压的回归系数均为负值（−0.0341 和−0.0177），说明这 2 个变量取值越大，生存时间越短，死亡风险越大；心电图读数的 0 水平与 2 水平比较时，ECG0 对应的 P 值为 0.002 6，且其回归系数为正值（0.757 1），说明心电图异常者生存时间短于正常者。

以下是基于广义 Gamma 模型预测各观测中位生存时间的结果。

Obs	*t*	status	_PROB_	median	s
1	12.4	0	0.5	28.722 7	6.442 5
2	12.4	0	0.5	32.542 2	7.191 4
3	9.6	0	0.5	43.717 8	11.853 0
4	7.2	0	0.5	22.971 3	6.181 8
5	14.1	0	0.5	33.840 5	7.548 9
⋮	⋮	⋮	⋮	⋮	⋮
148	10.6	1	0.5	10.136 3	2.305 8
149	10.5	0	0.5	25.097 0	5.348 0

前 5 位患者的输出结果表明，第 1 号患者随访开始时年龄（age1）44 岁，收缩压（SBP）

17.6kPa(132mmHg),舒张压(DBP)12.8kPa(96mmHg),心电图读数正常.预测此类患者的中位生存时间为28.7年。第2号患者随访开始时年龄(age1)49岁.收缩压(SBP)17.3kPa(130mmHg),舒张压(DBP)9.6kPa(72mmHg),心电图读数正常,预测此类患者的中位生存时间为32.5年。其他依此类推。

(郭 晋 李 卫 王 杨 陈 涛 唐欣然)

第18章 贝叶斯统计分析简介

2010年12月，美国食品与药品监督管理局（FDA）器械与辐射防护中心（CDRH）向产业和食品药品管理从业人员发布了《医疗器械临床试验的贝叶斯统计学评价指导原则》，其中讨论了贝叶斯统计学方法用于医疗器械临床疗效评价的关键问题，并提出了相应的指导原则。目前已有许多以贝叶斯分析为主的上市许可申请和多个临床试验申请获得批准。贝叶斯统计分析方法应用于医疗器械评价已成为当前的一种趋势。

一、贝叶斯统计分析

在贝叶斯统计分析中，任一未知参数β都可看作是一个随机变量，$P(\beta)$是在获取数据之前对β的认知，称作β的先验分布。假定$U=(u_1,u_2\cdots u_{n-1},u_n)$　是来自某总体的样本，该总体的密度函数为$P(U\mid\beta)$，$\theta=(\beta_1,\beta_2\cdots\beta_{n-1},\beta_n)$。$\beta$的条件概率分布：

$$P(\beta\mid U)=\frac{P(U\mid\beta)P(\beta)}{P(U)}$$

称作β的后验分布，是贝叶斯统计对参数β或其函数进行任何推断的基础。

二、贝叶斯统计和传统的统计分析的区别

传统的统计分析方法仅仅在研究设计阶段运用到相关的研究提供位置参数的先验信息和样本信息；在统计分析阶段这些信息仅仅作为传统分析结果的一个补充。然而，贝叶斯统计分析方法则是综合未知参数的先验信息和样本信息（随着观察值不断地获取，通过贝叶斯定理，可以不断地更新相关的信息，即：今天的后验即时明天的先验），进而，求出的后验分布来推断位置参数。

通过对先验信息的利用，贝叶斯统计可以为最后的器械审评提供更多的信息；若先验信息质量很高，则可以显著的减少样本量，缩短试验周期；同时贝叶斯方法可以随着试验的进行，灵活进行试验中期分析和处理中途变更（包括停止一个不合理的治疗组、修改随机计划），其中修改随机计划适用于人群敏感的医疗器械疗效评价（如：试验组收集不到病例）；贝叶斯的方法在控制治疗过程中的随机比例变化时尤其灵活；另外，当传统的统计分析只能进行近似分析时，通过贝叶斯分析可以获得精确地结果，并且贝叶斯统计分析在缺失值的处理方面更加灵活。

由于贝叶斯分析具有上述特点，并且医疗器械的更新换代更多的是在原来基础上改进，提供了很多高质量的先验信息；因此，贝叶斯统计分析方法非常符合医疗器械的临床疗效评价。

三、贝叶斯统计的临床试验设计

1. 良好的试验设计　与传统的统计分析方法相同，贝叶斯统计同样需要一个良好的试验

设计。试验设计的主要内容包括：试验目的、待评估的终点指标、试验所需条件、试验人群、统计分析计划等。

减小偏倚：器械疗效评价中可以用随机的方法确定病人接受哪一组的治疗，从而最小化偏倚带来的影响；通过对医生设盲可以最小化有意或无意的选择偏倚；对病人的设盲可以最小化安慰剂效应。我们建议在试验开始前统计分析的类型就要确定（贝叶斯统计分析或传统的统计分析）。若我们已经收集到所需数据，然后再根据不同方法的统计结果选择统计分析方法是有问题的。

终点指标：器械评价的临终点应该是与临床相关的、可以直接观察的、对病人是非常重要的、与器械安全性和有效性相关的指标。例如：临床终点可以是一个能在试验中观察到的，可以评价重要结局平均变化的指标（死亡率、发病率、生活质量等）。

2. *对照组的选择*　为了更好地评估临床试验的结果，我们推荐设置一个对照组作为参考。对照的设置主要有平行对照，自身对照和历史对照。其中自身对照和历史对照与平行对照相比存在很多潜在的偏倚，因此我们推荐使用平行对照。其中在贝叶斯统计方法中，采用阳性对照（平行对照的一种）证明新器械的非劣效性已得到广泛的应用。

3. *关于试验终点的初始信息——先验分布*　先验信息是指在研究进行之前根据以往研究所得到的关于终点指标的相关信息，我们用先验分布来描述这些先验信息的不确定性。新的医疗器械研究的先验信息可能来自于新器械自身的信息、对照器械的信息或两者兼之。先验信息的获得主要来源于已经完成的国外开展的试验研究、类似产品的临床资料或早期试点研究资料。当有高质量的先验信息可以使用时，我们可以根据这些信息确定出先验分布；当我们对所研究的终点指标一无所知时，也可以指定一个无信息的先验分布（通常为均匀分布）。

4. *确定样本含量*　临床试验中所需的样本量是指能获得预期结果（观察到期望差异）时所需的最小样本量。贝叶斯统计分析中样本含量一般由下面几个因素决定：样本的变异程度、先验信息的质量、分析时用的数学模型、分析模型中参数的分布、判断标准。

传统的统计分析中，样本量要在试验设计中提前确定，然而，在贝叶斯统计分析中不事先确定要本含量，而是指定一个特定的停止试验的标准：获得关于试验终点指标足够的信息量或达到一个事先指定的足够高的概率。在进行贝叶斯统计分析时，任何一点均能算出达到预期停止标准尚需要的额外观察数。例如，在一个以器械成功率为终点心脏支架非劣效性试验中，在其他条件不变时，当阳性对照组和试验组的成功率均为 0.85 时，需样本量为每组 40 人；当阳性对照组和试验组的成功率分别达到 0.90 和 0.95 时，需样本量为每组 25 人。也就是说，随着试验的进行，样本量将会被不断地调整。

5. *运用多层模型从其他研究中借用力量*　在医疗器械疗效评价中，这种借用的力量可以理解为样本量，其中借的程度依赖于两个研究的接近程度。如果研究的相似性很高（具体表现在个体的可交换性），我们当前的研究则可以借到很多的力量（节省很多样本量）；但是当研究相似性较低时，当前研究借用到的力量就会很少。其中可交换性是指两个个体能够提供等价的统计信息。

四、贝叶斯统计的临床试验分析

贝叶斯统计的临床试验分析主要包括：后验分布、假设检验、区间估计、预测概率和中期分析等几个方面。

1. 后验分布　后验分布包含了所有的先验信息，通过似然函数把试验中所得的结果结合起来。贝叶斯分析中，所有的结论都是基于后验分布。

2. 假设检验　统计推断包括假设检验和区间估计。贝叶斯分析的假设检验是基于后验分布计算出的特定假设（原假设或备择假设）是真的概率。若原假设为真的概率大于备择假设，则接受原假设，反之亦然；若两者的概率相近时，建议进一步收集样本信息，然后再作结论。

3. 区间估计　贝叶斯区间估计是基于后验分布求得一个区间（置信区间），使得参数落入区间的概率为所给定的概率。需要注意的是，贝叶斯的可信区间仅仅基于先验信息和当前数据，并不涉及重复抽样，因此在结果解释上与传统统计分析置信区间的意义不同。

4. 预测概率　预测概率是指特定结局在未来发生的概率，主要应用在以下几个方面：确定停止试验的时间、预测病人的临床结局、缺失值的预测填充和模型检验。可以在制定临床试验时，将预测概率作为停止试验的标准。

5. 中期分析　贝叶斯统计分析主要有后验概率和决策分析两种方法做临床疗效评价的中期分析。

（1）后验概率：若中期分析时假设的后验概率足够大，则可以提前终止试验，即：相同的贝叶斯假设检验在试验过程中重复进行。

（2）预测分布：若预测的成功（观察到预期的差异）概率足够高，则试验可以提前终止；若预测的成功概率非常低，则试验也可以提前终止。

医疗器械的开发是一个改进和创新的过程，在这个过程中我们能够积累丰富的经验（先验信息），并且贝叶斯方法可以充分利用这些先验信息，因此贝叶斯统计分析在医疗器械疗效评价的临床试验中应用前景很广泛。但在应用贝叶斯统计分析时也存在一些困难和挑战。如：与传统的统计分析的试验设计相比，贝叶斯统计分析的临床试验设计要更为复杂，要求更为严格；试验设计包括的先验信息和模型的选择必须要经过专家的反复的讨论和协商；贝叶斯统计分析常常涉及到高微积分的求解，因此需要配套的统计和计算机技术。

（吴振强　孙叶桓　李　卫　郭　晋）

附　表

附表 1　标准正态分布曲线下的面积[φ(*u*)值]

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247

（续 表）

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

附表 2　t 分布临界值

自由度 ν	概率(P)								
	单侧：0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	双侧：0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	1.0000	3.0777	6.3138	12.7062	31.821	63.657	127.32	318.309	636.619
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248	14.089	22.3271	31.5991
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145	12.9240
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732	8.6103
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934	6.8688
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.2076	5.9588
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995	4.0293	4.7853	5.4079
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0413
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.7809
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4370
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1405
15	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862	4.0150
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458	3.9651
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453	3.1534	3.5518	3.8495
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352	3.5272	3.8193
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787	3.0669	3.4350	3.7066
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564	3.0380	3.3962	3.6594
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460

（续 表）

自由度 ν	概率(P)								
	单侧:0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	双侧:0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440	3.0221	3.3749	3.6335
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385	3.0149	3.3653	3.6218
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333	3.0082	3.3563	3.6109
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284	3.0020	3.3479	3.6007
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238	2.9960	3.3400	3.5911
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195	2.9905	3.3326	3.5821
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154	2.9852	3.3256	3.5737
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116	2.9803	3.3190	3.5657
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079	2.9756	3.3128	3.5581
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.5510
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012	2.9670	3.3013	3.5442
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981	2.9630	3.2960	3.5377
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951	2.9592	3.2909	3.5316
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923	2.9555	3.2861	3.5258
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896	2.9521	3.2815	3.5203
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870	2.9488	3.2771	3.5150
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846	2.9456	3.2729	3.5099
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822	2.9426	3.2689	3.5051
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800	2.9397	3.2651	3.5004
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778	2.9370	3.2614	3.4960
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146	3.2317	3.4602
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479	2.8987	3.2108	3.4350
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387	2.8870	3.1953	3.4163
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316	2.8779	3.1833	3.4019
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737	3.3905
200	0.6757	1.2858	1.6525	1.9719	2.3451	2.6006	2.8385	3.1315	3.3398
300	0.6753	1.2844	1.6499	1.9679	2.3388	2.5923	2.8279	3.1176	3.3233
500	0.6750	1.2832	1.6479	1.9647	2.3338	2.5857	2.8195	3.1066	3.3101
1000	0.6747	1.2824	1.6464	1.9623	2.3301	2.5808	2.8133	3.0984	3.3003
∞	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758	2.8070	3.0902	3.2905

附录3 x^2 分布临界值

自由度 ν	概率 (P)						
	0.995	0.990	0.975	0.950	0.900	0.750	0.500
1	.	.	.	.	0.02	0.10	0.45
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33

（续 表）

自由度 ν	概率（P）						
	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.31
21	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.19	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.9	106.6	112.3	116.3	124.8
90	98.65	107.6	113.2	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

附表 4　*F* 分布临界值(方差齐性检验用,双侧概率为 0.05)

分母的自由度 ν_2	分子的自由度(ν_1)						
	1	2	3	4	5	6	7
1	647.8	799.5	864.2	899.6	921.9	937.1	948.2
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29

（续 表）

分母的自由度 ν_2	分子的自由度(ν_1)						
	8	9	10	11	12	13	14
1	956.7	963.3	968.6	973.0	976.7	979.8	982.5
2	39.37	39.39	39.40	39.41	39.41	39.42	39.43
3	14.54	14.47	14.42	14.37	14.34	14.30	14.28
4	8.98	8.90	8.84	8.79	8.75	8.71	8.68
5	6.76	6.68	6.62	6.57	6.52	6.49	6.46
6	5.60	5.52	5.46	5.41	5.37	5.33	5.30
7	4.90	4.82	4.76	4.71	4.67	4.63	4.6
8	4.43	4.36	4.30	4.24	4.20	4.16	4.13
9	4.10	4.03	3.96	3.91	3.87	3.83	3.80
10	3.85	3.78	3.72	3.66	3.62	3.58	3.55
11	3.66	3.59	3.53	3.47	3.43	3.39	3.36
12	3.51	3.44	3.37	3.32	3.28	3.24	3.21
13	3.39	3.31	3.25	3.20	3.15	3.12	3.08
14	3.29	3.21	3.15	3.09	3.05	3.01	2.98
15	3.20	3.12	3.06	3.01	2.96	2.92	2.89
16	3.12	3.05	2.99	2.93	2.89	2.85	2.82
17	3.06	2.98	2.92	2.87	2.82	2.79	2.75
18	3.01	2.93	2.87	2.81	2.77	2.73	2.70
19	2.96	2.88	2.82	2.76	2.72	2.68	2.65
20	2.91	2.84	2.77	2.72	2.68	2.64	2.60
21	2.87	2.80	2.73	2.68	2.64	2.60	2.56
22	2.84	2.76	2.70	2.65	2.60	2.56	2.53
23	2.81	2.73	2.67	2.62	2.57	2.53	2.50
24	2.78	2.70	2.64	2.59	2.54	2.50	2.47
25	2.75	2.68	2.61	2.56	2.51	2.48	2.44
26	2.73	2.65	2.59	2.54	2.49	2.45	2.42
27	2.71	2.63	2.57	2.51	2.47	2.43	2.39
28	2.69	2.61	2.55	2.49	2.45	2.41	2.37
29	2.67	2.59	2.53	2.48	2.43	2.39	2.36
30	2.65	2.57	2.51	2.46	2.41	2.37	2.34
40	2.53	2.45	2.39	2.33	2.29	2.25	2.21
60	2.41	2.33	2.27	2.22	2.17	2.13	2.09
100	2.32	2.24	2.18	2.12	2.08	2.04	2.00
120	2.30	2.22	2.16	2.10	2.05	2.01	1.98
∞	2.19	2.11	2.05	1.99	1.94	1.90	1.87

（续 表）

分母的自由度 ν_2	分子的自由度 ν_1						
	15	20	30	40	60	100	∞
1	984.9	993.1	1001	1006	1010	1013	1018
2	39.43	39.45	39.46	39.47	39.48	39.49	39.50
3	14.25	14.17	14.08	14.04	13.99	13.96	13.90
4	8.66	8.56	8.46	8.41	8.36	8.32	8.26
5	6.43	6.33	6.23	6.18	6.12	6.08	6.02
6	5.27	5.17	5.07	5.01	4.96	4.92	4.85
7	4.57	4.47	4.36	4.31	4.25	4.21	4.14
8	4.10	4.00	3.89	3.84	3.78	3.74	3.67
9	3.77	3.67	3.56	3.51	3.45	3.40	3.33
10	3.52	3.42	3.31	3.26	3.20	3.15	3.08
11	3.33	3.23	3.12	3.06	3.00	2.96	2.88
12	3.18	3.07	2.96	2.91	2.85	2.80	2.72
13	3.05	2.95	2.84	2.78	2.72	2.67	2.60
14	2.95	2.84	2.73	2.67	2.61	2.56	2.49
15	2.86	2.76	2.64	2.59	2.52	2.47	2.40
16	2.79	2.68	2.57	2.51	2.45	2.40	2.32
17	2.72	2.62	2.50	2.44	2.38	2.33	2.25
18	2.67	2.56	2.44	2.38	2.32	2.27	2.19
19	2.62	2.51	2.39	2.33	2.27	2.22	2.13
20	2.57	2.46	2.35	2.29	2.22	2.17	2.09
21	2.53	2.42	2.31	2.25	2.18	2.13	2.04
22	2.50	2.39	2.27	2.21	2.14	2.09	2.00
23	2.47	2.36	2.24	2.18	2.11	2.06	1.97
24	2.44	2.33	2.21	2.15	2.08	2.02	1.94
25	2.41	2.30	2.18	2.12	2.05	2.00	1.91
26	2.39	2.28	2.16	2.09	2.03	1.97	1.88
27	2.36	2.25	2.13	2.07	2.00	1.94	1.85
28	2.34	2.23	2.11	2.05	1.98	1.92	1.83
29	2.32	2.21	2.09	2.03	1.96	1.90	1.81
30	2.31	2.20	2.07	2.01	1.94	1.88	1.79
40	2.18	2.07	1.94	1.88	1.80	1.74	1.64
60	2.06	1.94	1.82	1.74	1.67	1.60	1.48
100	1.97	1.85	1.71	1.64	1.56	1.48	1.35
120	1.94	1.82	1.69	1.61	1.53	1.45	1.31
∞	1.83	1.71	1.57	1.48	1.39	1.30	1.00

附表 5 F 分布临界值(方差分析用,$\alpha=0.05$)

ν_2	ν_1						
	1	2	3	4	5	6	7
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09
∞	3.84	3.00	2.66	2.37	2.21	2.10	2.01

（续 表）

ν_2	ν_1						
	8	9	10	11	12	13	14
1	238.9	240.5	241.9	243.0	243.9	244.7	245.4
2	19.37	19.38	19.40	19.40	19.41	19.42	19.42
3	8.85	8.81	8.79	8.76	8.74	8.73	8.71
4	6.04	6.00	5.96	5.94	5.91	5.89	5.87
5	4.82	4.77	4.74	4.70	4.68	4.66	4.64
6	4.15	4.10	4.06	4.03	4.00	3.98	3.96
7	3.73	3.68	3.64	3.60	3.57	3.55	3.53
8	3.44	3.39	3.35	3.31	3.28	3.26	3.24
9	3.23	3.18	3.14	3.10	3.07	3.05	3.03
10	3.07	3.02	2.98	2.94	2.91	2.89	2.86
11	2.95	2.90	2.85	2.82	2.79	2.76	2.74
12	2.85	2.80	2.75	2.72	2.69	2.66	2.64
13	2.77	2.71	2.67	2.63	2.60	2.58	2.55
14	2.70	2.65	2.60	2.57	2.53	2.51	2.48
15	2.64	2.59	2.54	2.51	2.48	2.45	2.42
16	2.59	2.54	2.49	2.46	2.42	2.40	2.37
17	2.55	2.49	2.45	2.41	2.38	2.35	2.33
18	2.51	2.46	2.41	2.37	2.34	2.31	2.29
19	2.48	2.42	2.38	2.34	2.31	2.28	2.26
20	2.45	2.39	2.35	2.31	2.28	2.25	2.22
21	2.42	2.37	2.32	2.28	2.25	2.22	2.20
22	2.40	2.34	2.30	2.26	2.23	2.20	2.17
23	2.37	2.32	2.27	2.24	2.20	2.18	2.15
24	2.36	2.30	2.25	2.22	2.18	2.15	2.13
25	2.34	2.28	2.24	2.20	2.16	2.14	2.11
26	2.32	2.27	2.22	2.18	2.15	2.12	2.09
27	2.31	2.25	2.20	2.17	2.13	2.10	2.08
28	2.29	2.24	2.19	2.15	2.12	2.09	2.06
29	2.28	2.22	2.18	2.14	2.10	2.08	2.05
30	2.27	2.21	2.16	2.13	2.09	2.06	2.04
40	2.18	2.12	2.08	2.04	2.00	1.97	1.95
60	2.10	2.04	1.99	1.95	1.92	1.89	1.86
100	2.03	1.97	1.93	1.89	1.85	1.82	1.79
120	2.02	1.96	1.91	1.87	1.83	1.80	1.78
∞	1.94	1.88	1.83	1.79	1.75	1.72	1.69

（续 表）

ν_2	ν_1						
	20	40	50	100	200	500	∞
1	248.0	251.1	251.8	253.0	253.7	254.1	254.3
2	19.45	19.47	19.48	19.49	19.49	19.49	19.50
3	8.66	8.59	8.58	8.55	8.54	8.53	8.53
4	5.80	5.72	5.70	5.66	5.65	5.64	5.63
5	4.56	4.46	4.44	4.41	4.39	4.37	4.36
6	3.87	3.77	3.75	3.71	3.69	3.68	3.67
7	3.44	3.34	3.32	3.27	3.25	3.24	3.23
8	3.15	3.04	3.02	2.97	2.95	2.94	2.93
9	2.94	2.83	2.80	2.76	2.73	2.72	2.71
10	2.77	2.66	2.64	2.59	2.56	2.55	2.54
11	2.65	2.53	2.51	2.46	2.43	2.42	2.40
12	2.54	2.43	2.40	2.35	2.32	2.31	2.30
13	2.46	2.34	2.31	2.26	2.23	2.22	2.21
14	2.39	2.27	2.24	2.19	2.16	2.14	2.13
15	2.33	2.20	2.18	2.12	2.10	2.08	2.07
16	2.28	2.15	2.12	2.07	2.04	2.02	2.01
17	2.23	2.10	2.08	2.02	1.99	1.97	1.96
18	2.19	2.06	2.04	1.98	1.95	1.93	1.92
19	2.16	2.03	2.00	1.94	1.91	1.89	1.88
20	2.12	1.99	1.97	1.91	1.88	1.86	1.84
21	2.10	1.96	1.94	1.88	1.84	1.83	1.81
22	2.07	1.94	1.91	1.85	1.82	1.80	1.78
23	2.05	1.91	1.88	1.82	1.79	1.77	1.76
24	2.03	1.89	1.86	1.80	1.77	1.75	1.73
25	2.01	1.87	1.84	1.78	1.75	1.73	1.71
26	1.99	1.85	1.82	1.76	1.73	1.71	1.69
27	1.97	1.84	1.81	1.74	1.71	1.69	1.67
28	1.96	1.82	1.79	1.73	1.69	1.67	1.65
29	1.94	1.81	1.77	1.71	1.67	1.65	1.64
30	1.93	1.79	1.76	1.70	1.66	1.64	1.62
40	1.84	1.69	1.66	1.59	1.55	1.53	1.51
60	1.75	1.59	1.56	1.48	1.44	1.41	1.39
100	1.68	1.52	1.48	1.39	1.34	1.31	1.28
120	1.66	1.50	1.46	1.37	1.32	1.28	1.25
∞	1.57	1.39	1.35	1.24	1.17	1.11	1.00

附表 6　ψ 值(多个样本均数比较时所需样本例数的估计用 $\alpha=0.05,\beta=0.1$)

υ_2	υ_1																
	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
2	6.80	6.71	6.68	6.67	6.66	6.65	6.65	6.65	6.64	6.64	6.64	6.63	6.63	6.63	6.63	6.63	6.62
3	5.01	4.63	4.47	4.39	4.34	4.30	4.27	4.25	4.23	4.22	4.18	4.16	4.14	4.13	4.12	4.11	4.09
4	4.40	3.90	3.69	3.58	3.50	3.45	3.41	3.38	3.36	3.34	3.28	3.25	3.22	3.20	3.19	3.17	3.15
5	4.09	3.54	3.30	3.17	3.08	3.02	2.97	2.94	2.91	2.89	2.81	2.78	2.74	2.72	2.70	2.68	2.66
6	3.91	3.32	3.07	2.92	2.83	2.76	2.71	2.67	2.64	2.61	2.53	2.49	2.44	2.42	2.40	2.37	2.35
7	3.80	3.18	2.91	2.76	2.66	2.58	2.53	2.49	2.45	2.42	2.33	2.29	2.24	2.21	2.19	2.16	2.18
8	3.71	3.08	2.81	2.64	2.51	2.46	2.40	2.35	2.32	2.29	2.19	2.14	2.09	2.06	2.03	2.00	1.97
9	3.65	3.01	2.72	2.56	2.44	2.36	2.30	2.26	2.22	2.19	2.09	2.03	1.97	1.94	1.91	1.88	1.85
10	3.60	2.95	2.66	2.49	2.37	2.29	2.23	2.18	2.14	2.11	2.00	1.94	1.88	1.85	1.82	1.78	1.75
11	3.57	2.91	2.61	2.44	2.32	2.23	2.17	2.12	2.08	2.04	1.93	1.87	1.81	1.78	1.74	1.70	1.67
12	3.54	2.87	2.57	2.39	2.27	2.19	2.12	2.07	2.02	1.99	1.88	1.81	1.75	1.71	1.68	1.64	1.60
13	3.51	2.84	2.54	2.36	2.23	2.15	2.08	2.02	1.98	1.95	1.83	1.76	1.69	1.66	1.62	1.58	1.54
14	3.49	2.81	2.51	2.33	2.20	2.11	2.04	1.99	1.94	1.91	1.79	1.72	1.65	1.61	1.57	1.53	1.49
15	3.47	2.79	2.48	2.30	2.17	2.08	2.01	1.96	1.91	1.87	1.75	1.68	1.61	1.57	1.53	1.49	1.44
16	3.46	2.77	2.46	2.28	2.15	2.06	1.99	1.93	1.88	1.85	1.72	1.65	1.58	1.54	1.49	1.45	1.40
17	3.44	2.76	2.44	2.26	2.13	2.04	1.96	1.91	1.86	1.82	1.69	1.62	1.55	1.50	1.46	1.41	1.36
18	3.43	2.74	2.43	2.24	2.11	2.02	1.94	1.89	1.84	1.80	1.67	1.60	1.52	1.48	1.43	1.38	1.33
19	3.42	2.73	2.41	2.22	2.09	2.00	1.93	1.87	1.82	1.78	1.65	1.58	1.49	1.45	1.40	1.35	1.30
20	3.41	2.72	2.40	2.21	2.08	1.98	1.91	1.85	1.80	1.76	1.63	1.55	1.47	1.43	1.38	1.33	1.27
21	3.40	2.71	2.39	2.20	2.07	1.97	1.90	1.84	1.79	1.75	1.61	1.54	1.45	1.41	1.36	1.30	1.25
22	3.39	2.70	2.38	2.19	2.05	1.96	1.88	1.82	1.77	1.73	1.60	1.52	1.43	1.39	1.34	1.28	1.22
23	3.39	2.69	2.37	2.18	2.04	1.95	1.87	1.81	1.76	1.72	1.58	1.50	1.42	1.37	1.32	1.26	1.20
24	3.38	2.68	2.36	2.17	2.03	1.94	1.86	1.80	1.75	1.71	1.57	1.49	1.40	1.35	1.30	1.24	1.18
25	3.37	2.68	2.35	2.16	2.02	1.93	1.85	1.79	1.74	1.70	1.56	1.48	1.39	1.34	1.28	1.23	1.16
26	3.37	2.67	2.35	2.15	2.02	1.92	1.84	1.78	1.73	1.69	1.54	1.46	1.37	1.32	1.27	1.21	1.15
27	3.36	2.66	2.34	2.14	2.01	1.91	1.83	1.77	1.72	1.68	1.53	1.45	1.36	1.31	1.26	1.20	1.13
28	3.36	2.66	2.33	2.14	2.00	1.90	1.82	1.76	1.71	1.67	1.52	1.44	1.35	1.30	1.24	1.18	1.11
29	3.36	2.65	2.33	2.13	1.99	1.89	1.82	1.75	1.70	1.66	1.51	1.43	1.34	1.29	1.23	1.17	1.10
30	3.35	2.65	2.32	2.12	1.99	1.89	1.81	1.75	1.70	1.65	1.51	1.42	1.33	1.28	1.22	1.16	1.08
31	3.35	2.64	2.32	2.12	1.98	1.88	1.80	1.74	1.69	1.64	1.50	1.41	1.32	1.27	1.21	1.14	1.07
32	3.34	2.64	2.31	2.11	1.98	1.88	1.80	1.73	1.68	1.64	1.49	1.41	1.31	1.26	1.20	1.13	1.06
33	3.34	2.63	2.31	2.11	1.97	1.87	1.79	1.73	1.68	1.63	1.48	1.40	1.30	1.25	1.19	1.12	1.05
34	3.34	2.63	2.30	2.10	1.97	1.87	1.79	1.72	1.67	1.63	1.48	1.39	1.29	1.24	1.18	1.11	1.04
35	3.34	2.63	2.30	2.10	1.96	1.86	1.78	1.72	1.66	1.62	1.47	1.38	1.29	1.23	1.17	1.10	1.02
36	3.33	2.62	2.30	2.10	1.96	1.86	1.78	1.71	1.66	1.62	1.47	1.38	1.28	1.22	1.16	1.09	1.01
37	3.33	2.62	2.29	2.09	1.95	1.85	1.77	1.71	1.65	1.61	1.46	1.37	1.27	1.22	1.15	1.08	1.09

（续 表）

v_2	v_1																
	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
38	3.33	2.62	2.29	2.09	1.95	1.85	1.77	1.70	1.65	1.61	1.45	1.37	1.27	1.21	1.15	1.08	0.99
39	3.33	2.62	2.29	2.09	1.95	1.84	1.76	1.70	1.65	1.60	1.45	1.36	1.26	1.20	1.14	1.07	0.99
40	3.32	2.61	2.28	2.08	1.94	1.84	1.76	1.70	1.64	1.60	1.44	1.36	1.25	1.20	1.13	1.06	0.98
41	3.32	2.61	2.28	2.08	1.94	1.84	1.76	1.69	1.64	1.59	1.44	1.35	1.25	1.19	1.13	1.05	0.97
42	3.32	2.61	2.28	2.08	1.94	1.83	1.75	1.69	1.63	1.59	1.44	1.35	1.24	1.18	1.12	1.05	0.96
43	3.32	2.61	2.28	2.07	1.93	1.83	1.75	1.69	1.63	1.59	1.43	1.34	1.24	1.18	1.11	1.04	0.95
44	3.32	2.60	2.27	2.07	1.93	1.83	1.75	1.68	1.63	1.58	1.43	1.34	1.23	1.17	1.11	1.03	0.94
45	3.31	2.60	2.27	2.07	1.93	1.83	1.74	1.68	1.62	1.58	1.42	1.33	1.23	1.17	1.10	1.03	0.94
46	3.31	2.60	2.27	2.07	1.93	1.82	1.74	1.68	1.62	1.58	1.42	1.33	1.22	1.16	1.10	1.02	0.93
47	3.31	2.60	2.27	2.06	1.92	1.82	1.74	1.67	1.62	1.57	1.42	1.33	1.22	1.16	1.09	1.02	0.92
48	3.31	2.60	2.26	2.06	1.92	1.82	1.74	1.67	1.62	1.57	1.41	1.32	1.22	1.15	1.09	1.01	0.92
49	3.31	2.59	2.26	2.06	1.92	1.82	1.73	1.67	1.61	1.57	1.41	1.32	1.21	1.15	1.08	1.00	0.91
50	3.31	2.59	2.26	2.06	1.92	1.81	1.73	1.67	1.61	1.56	1.41	1.31	1.21	1.15	1.08	1.00	0.90
60	3.30	2.58	2.25	2.04	1.90	1.79	1.71	1.64	1.59	1.54	1.38	1.29	1.18	1.11	1.04	0.95	0.85
80	3.28	2.56	2.23	2.02	1.88	1.77	1.69	1.62	1.56	1.51	1.35	1.25	1.14	1.07	0.99	0.90	0.77
120	3.27	2.55	2.21	2.00	1.86	1.75	1.66	1.59	1.54	1.49	1.32	1.22	1.09	1.02	0.94	0.83	0.68
240	3.26	2.53	2.19	1.98	1.84	1.73	1.64	1.57	1.51	1.46	1.29	1.18	1.05	0.97	0.88	0.76	0.56
∞	3.24	2.52	2.17	1.96	1.81	1.70	1.62	1.54	1.48	1.43	1.25	1.14	1.01	0.92	0.82	0.65	0.00

附表 7　λ值(多个样本率比较时所需样本例数的估计用 α=0.05)

υ	β								
	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
1	0.43	1.24	2.06	2.91	3.84	4.90	6.17	7.85	10.51
2	0.62	1.73	2.78	3.83	4.96	6.21	7.70	9.63	12.65
3	0.78	2.10	3.30	4.50	5.76	7.15	8.79	10.90	14.17
4	0.91	2.40	3.74	5.05	6.42	7.92	9.68	11.94	15.41
5	1.03	2.67	4.12	5.53	6.99	8.59	10.45	12.83	16.47
6	1.13	2.91	4.46	5.96	7.50	9.19	11.14	13.62	17.42
7	1.23	3.13	4.77	6.35	7.97	9.73	11.77	14.35	18.28
8	1.32	3.33	5.06	6.71	8.40	10.24	12.35	15.02	19.08
9	1.40	3.53	5.33	7.05	8.81	10.71	12.89	15.65	19.83
10	1.49	3.71	5.59	7.37	9.19	11.15	13.40	16.24	20.53
11	1.56	3.88	5.83	7.68	9.56	11.57	13.89	16.80	21.20
12	1.64	4.05	6.06	7.97	9.90	11.98	14.35	17.34	21.83
13	1.71	4.20	6.29	8.25	10.23	12.36	14.80	17.85	22.44
14	1.77	4.36	6.50	8.52	10.55	12.73	15.22	18.34	23.02
15	1.84	4.50	6.71	8.78	10.86	13.09	15.63	18.81	23.58
16	1.90	4.65	6.91	9.03	11.16	13.43	16.03	19.27	24.13
17	1.97	4.78	7.10	9.27	11.45	13.77	16.41	19.71	24.65
18	2.03	4.92	7.29	9.50	11.73	14.09	16.78	20.14	25.16
19	2.08	5.05	7.47	9.73	12.00	14.41	17.14	20.56	25.65
20	2.14	5.18	7.65	9.96	12.26	14.71	17.50	20.96	26.13
21	2.20	5.30	7.83	10.17	12.52	15.01	17.84	21.36	26.60
22	2.25	5.42	8.00	10.38	12.77	15.30	18.17	21.74	27.06
23	2.30	5.54	8.16	10.59	13.02	15.59	18.50	22.12	27.50
24	2.36	5.66	8.33	10.79	13.26	15.87	18.82	22.49	27.94
25	2.41	5.77	8.48	10.99	13.49	16.14	19.13	22.85	28.37
26	2.46	5.88	8.64	11.19	13.72	16.41	19.44	23.20	28.78
27	2.51	5.99	8.79	11.38	13.95	16.67	19.74	23.55	29.19
28	2.56	6.10	8.94	11.57	14.17	16.93	20.04	23.89	29.60
29	2.60	6.20	9.09	11.75	14.39	17.18	20.33	24.22	29.99
30	2.65	6.31	9.24	11.93	14.60	17.43	20.61	24.55	30.38
31	2.69	6.41	9.38	12.11	14.82	17.67	20.89	24.87	30.76
32	2.74	6.51	9.52	12.28	15.02	17.91	21.17	25.19	31.13
33	2.78	6.61	9.66	12.45	15.23	18.15	21.44	25.50	31.50
34	2.83	6.70	9.79	12.62	15.43	18.38	21.70	25.80	31.87
35	2.87	6.80	9.93	12.79	15.63	18.61	21.97	26.11	32.23
36	2.91	6.89	10.06	12.96	15.82	18.84	22.23	26.41	32.58

（续 表）

ν	β								
	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
37	2.96	6.99	10.19	13.12	16.01	19.06	22.48	26.70	32.93
38	3.00	7.08	10.32	13.28	16.20	19.28	22.73	26.99	33.27
39	3.04	7.17	10.45	13.44	16.39	19.50	22.98	27.27	33.61
40	3.08	7.26	10.57	13.59	16.58	19.71	23.23	27.56	33.94
50	3.46	8.10	11.75	15.06	18.31	21.72	25.53	30.20	37.07
60	3.80	8.86	12.81	16.38	19.88	23.53	27.61	32.59	39.89
70	4.12	9.56	13.79	17.60	21.32	25.20	29.52	34.79	42.48
80	4.41	10.21	14.70	18.74	22.67	26.75	31.29	36.83	44.89
90	4.69	10.83	15.56	19.80	23.93	28.21	32.96	38.74	47.16
100	4.95	11.41	16.37	20.81	25.12	29.59	34.54	40.56	49.29
110	5.20	11.96	17.14	21.77	26.25	30.90	36.04	42.28	51.33
120	5.44	12.49	17.88	22.68	27.34	32.15	37.47	43.92	53.27

附表 8　相关系数 r 临界值

n	P:0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	0.707	0.951	0.988	0.997	1.000	1.000	1.000	1.000	1.000
2	0.500	0.800	0.900	0.950	0.980	0.990	0.995	0.998	0.999
3	0.404	0.687	0.805	0.878	0.934	0.959	0.974	0.986	0.991
4	0.347	0.608	0.729	0.811	0.882	0.917	0.942	0.963	0.974
5	0.309	0.551	0.669	0.754	0.833	0.875	0.906	0.935	0.951
6	0.281	0.507	0.621	0.707	0.789	0.834	0.870	0.905	0.925
7	0.260	0.472	0.582	0.666	0.750	0.798	0.836	0.875	0.898
8	0.242	0.443	0.549	0.632	0.715	0.765	0.805	0.847	0.872
9	0.228	0.419	0.521	0.602	0.685	0.735	0.776	0.820	0.847
10	0.216	0.398	0.497	0.576	0.658	0.708	0.750	0.795	0.823
11	0.206	0.380	0.476	0.553	0.634	0.684	0.726	0.772	0.801
12	0.197	0.365	0.458	0.532	0.612	0.661	0.703	0.750	0.780
13	0.189	0.351	0.441	0.514	0.592	0.641	0.683	0.730	0.760
14	0.182	0.338	0.426	0.497	0.574	0.623	0.664	0.711	0.742
15	0.176	0.327	0.412	0.482	0.558	0.606	0.647	0.694	0.725
16	0.170	0.317	0.400	0.468	0.543	0.590	0.631	0.678	0.708
17	0.165	0.308	0.389	0.456	0.529	0.575	0.616	0.662	0.693
18	0.160	0.299	0.378	0.444	0.516	0.561	0.602	0.648	0.679
19	0.156	0.291	0.369	0.433	0.503	0.549	0.589	0.635	0.665
20	0.152	0.284	0.360	0.423	0.492	0.537	0.576	0.622	0.652
21	0.148	0.277	0.352	0.413	0.482	0.526	0.565	0.610	0.640
22	0.145	0.271	0.344	0.404	0.472	0.515	0.554	0.599	0.629
23	0.141	0.265	0.337	0.396	0.462	0.505	0.543	0.588	0.618
24	0.138	0.260	0.330	0.388	0.453	0.496	0.534	0.578	0.607
25	0.136	0.255	0.323	0.381	0.445	0.487	0.524	0.568	0.597
26	0.133	0.250	0.317	0.374	0.437	0.479	0.515	0.559	0.588
27	0.130	0.245	0.311	0.367	0.430	0.471	0.507	0.550	0.579
28	0.128	0.241	0.306	0.361	0.423	0.463	0.499	0.541	0.570
29	0.126	0.237	0.301	0.355	0.416	0.456	0.491	0.533	0.562
30	0.124	0.233	0.296	0.349	0.409	0.449	0.484	0.526	0.554
31	0.122	0.229	0.291	0.344	0.403	0.442	0.477	0.518	0.547
32	0.120	0.225	0.287	0.339	0.397	0.436	0.470	0.511	0.539
33	0.118	0.222	0.283	0.334	0.392	0.430	0.464	0.504	0.532
34	0.116	0.219	0.279	0.329	0.386	0.424	0.458	0.498	0.525
35	0.114	0.216	0.275	0.325	0.381	0.418	0.452	0.492	0.519
36	0.113	0.213	0.271	0.320	0.376	0.413	0.446	0.486	0.513
37	0.111	0.210	0.267	0.316	0.371	0.408	0.441	0.480	0.507

（续 表）

υ	P:0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
38	0.110	0.207	0.264	0.312	0.367	0.403	0.435	0.474	0.501
39	0.108	0.204	0.260	0.308	0.362	0.398	0.430	0.469	0.495
40	0.107	0.202	0.257	0.304	0.358	0.393	0.425	0.463	0.490
41	0.106	0.199	0.254	0.301	0.354	0.389	0.420	0.458	0.484
42	0.104	0.197	0.251	0.297	0.350	0.384	0.416	0.453	0.479
43	0.103	0.195	0.248	0.294	0.346	0.380	0.411	0.449	0.474
44	0.102	0.192	0.246	0.291	0.342	0.376	0.407	0.444	0.469
45	0.101	0.190	0.243	0.288	0.338	0.372	0.403	0.439	0.465
46	0.100	0.188	0.240	0.285	0.335	0.368	0.399	0.435	0.460
47	0.099	0.186	0.238	0.282	0.331	0.365	0.395	0.431	0.456
48	0.098	0.184	0.235	0.279	0.328	0.361	0.391	0.427	0.451
49	0.097	0.182	0.233	0.276	0.325	0.358	0.387	0.423	0.447
50	0.096	0.181	0.231	0.273	0.322	0.354	0.384	0.419	0.443
60	0.087	0.165	0.211	0.250	0.295	0.325	0.352	0.385	0.408
70	0.081	0.153	0.195	0.232	0.274	0.302	0.327	0.358	0.380
80	0.076	0.143	0.183	0.217	0.257	0.283	0.307	0.336	0.357
90	0.071	0.135	0.173	0.205	0.242	0.267	0.290	0.318	0.338
100	0.068	0.128	0.164	0.195	0.230	0.254	0.276	0.303	0.321
200	0.048	0.091	0.116	0.138	0.164	0.181	0.197	0.216	0.230
300	0.039	0.074	0.095	0.113	0.134	0.148	0.161	0.177	0.188
500	0.030	0.057	0.073	0.088	0.104	0.115	0.125	0.138	0.146
1000	0.021	0.041	0.052	0.062	0.073	0.081	0.089	0.098	0.104
∞	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

$\nu=n-2$ 为自由度

附表 9　Spearman 秩相关系数($\rho_s=0$ 的界值表)

n	概率 P								
	单侧:0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	双侧:0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
4	0.600	1.000	1.000						
5	0.500	0.800	0.900	1.000	1.000				
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.217	0.406	0.503	0.587	0.678	0.727	0.769	0.818	0.846
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.791	0.824
14	0.200	0.367	0.464	0.538	0.626	0.679	0.723	0.771	0.802
15	0.189	0.354	0.446	0.521	0.604	0.654	0.700	0.750	0.779
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.729	0.762
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.382	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519

（续 表）

n	概率 P								
	单侧：0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	双侧：0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507
41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456

参 考 文 献

[1] 医疗器械临床试验规定．国家食品药品监督管理局 2004 年 4 月实施

[2] 国家食品药品监督管理局器械司、国家食品药品监督管理局培训中心．医疗器械监管专业技术基础．

[3] 黄嘉华．医疗器械注册与管理．北京：科学出版社

[4] 药物临床试验质量管理规范．国家食品药品监督管理局，2003

[5] 国家食品药品监督管理局．化学药品和生物制品临床试验的生物统计学指导原则(第二稿)：2003 年 12 月

[6] STATISTICAL GUIDANCE for CLINICAL TRIALS of NON-DIAGNOSTIC MEDICAL DEVICES. FDA. 1996

[7] 苏炳华．新药临床试验统计分析新进展．上海：上海科学技术文献出版社，2000

[8] 刘玉秀，洪立基．新药临床研究设计与统计分析．南京：南京大学出版社，1999

[9] 金丕焕．医用统计方法．上海：复旦大学出版社，2003

[10] 姚晨．医疗器械临床试验设计的统计学基本原理．首都医药，2007，3：11

[11] 杨晓芳，奚廷斐．医疗器械临床试验概述．中国医疗器械信息，2006，12(7)：47

[12] Guidance for Industry and FDA Staff: Recommendations for Clinical Laboratory Improvement Amendments of 1988 (CLIA) Waiver Applications for Manufacturers of In Vitro Diagnostic Devices.

[13] 李镒冲，李晓松．两种测量方法定量测量结果的一致性评价．现代预防医学，2007，31(17)：3253

[14] 王治国．临床检验方法比较研究中回归分析方法的介绍．陕西医学检验，1996，11(2)：18

[15] 王建华．流行病学．北京：人民卫生出版社，2008：85-98

[16] 宇传华译．诊断医学统计学．北京：人民卫生出版社，2005：137-154

[17] 胡良平．口腔医学科研设计与统计分析．北京：人民军医出版社，2007：272-275

[18] 陈卉．Bland-Altman 分析在临床测量方法一致性评价中的应用．中国卫生统计，2001，21(3)：308

[19] Jason C. Hsu. Multiple Comparisons Theory and Methods. London: Chapman & Hall, 1996

[20] Yosef Hochberg, Ajit C. Tamhane. Multiple Comparison Procedures. New York: John Wiley & Sons, Inc., 1987

[21] SAS Institute Inc. 2004. SAS/STAT 9.1 User's Guide. Cary, NC: SAS Institute Inc

[22] 郭祖超．医用数理统计方法．3 版．北京：人民卫生出版社，1988，10

[23] 方积乾．生物医学研究的统计方法．北京：高等教育出版社，2007

[24] 胡良平．现代统计学与 SAS 应用．北京：军事医学科学出版社，2000

[25] Douglas C. Montgomery 著．汪仁官，陈荣昭译．实验设计与分析．北京：中国统计出版社，1998

[26] 孙振球．医学统计学．北京：人民卫生出版社，2002

[27] 王万中，茆诗松．试验的设计与分析．上海：华东师范大学出版社，1997

[28] 余松林．医学统计学．北京：人民卫生出版社，2002

[29] 高惠璇等编译．SAS 系统 SAS/STAT 软件使用手册．北京：中国统计出版社，1997

[30] 徐勇勇．医学统计学．北京：高等教育出版社，2004

[31] 杨树勤．中国医学百科全书．医学统计分册．上海：上海科学技术出版社，1985

[32] SAS Institute Inc. 2008. SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc

[33] 陈希儒，王松桂．近代回归分析——原理方法及应用．合肥：安徽教育出版社，1987

[34] Food and Drug Administration Center for Devices and Radiological Health. Clinical Study Designs for Catheter Ablation Devices for Treatment of Atria Flutter. Guidance for Industry and FDA Staff, 2008.

http://www.fda.gov/cdrb./ode/guidance/1678.html

[35] Food andDrug Administration Center for Devices and Radiological Health. The Least Burdensome Provisions of the FDA Modernization Act of 1997: Concept and Principles; Final Guidance for FDA and Industry. http://www.fda.gov/cdrb./ode/guidance/1332.html

[36] summary of Safety and Effectiveness Data (SSED)http://www.fda.gov/cdrb./pdf5/p050029b.pdf

[37] 吕德良，李雪迎，朱赛楠，等．目标值法在医疗器械非随机对照临床试验中的应用．中国卫生统计，2009，(3)

[38] Dixon, WJ, Massey, FJ. Introduction to Statistical Analysis. 4th Edition McGraw-Hill (1983)

[39] Chernick, M. R, Liu, CY. "The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods."The American Statistician,2002(56):149

[40] OBrien RG, Muller KE. Applied Analysis of Variance in Behavioral Science. New York: Marcel Dekker, 1993, 297

[41] Jovanovic B, Zalenski R. Safety Evaluation and Confidence Intervals When the Number of Observed Events is Small or Zero. Annals of Emergency Medicine, 1997,30(3):301

[42] 陈家鼎，戴中维译．生存数据分析的统计方法．北京：中国统计出版社，1998：77-321

[43] 陈峰．医用多元统计分析方法．北京：中国统计出版社，2000：132-159

[44] Nora Mutalima, Elizabeth Molyneux, Harold Jaffe, et al. Associations between Burkitt 朵 Lymphoma among children in Malawi and infection with HIV, EBV and Malaria: results from a case-control study. PLoS ONE . 2008,3(6):e2505

[45] 胡良平．Windows SAS 6.12 & 8.0 实用统计分析教程．北京：军事医学科学出版社，2001：336-362

[46] 张家放．医用多元统计方法．武汉：华中科技大学出版社，2002：127-157

[47] 金丕焕，苏炳华，贺佳．医用 SAS 统计分析．上海：复旦大学出版社，2000：80-105

[48] 方积乾．卫生统计学．5 版．北京：人民卫生出版社，2007：356-367

[49] 方积乾．生物医学研究的统计方法．北京：高等教育出版社，2007：347-371

[50] 胡克震，马德锡．医学随访统计方法．北京：科学技术文献出版社，1993：71-100

[51] 茆诗松，濮晓龙，刘忠译．寿命数据中的统计模型和方法．北京：中国统计出版社，1998：278-407

[52] 高惠璇编译．SAS 系统：SAS/STAT 软件使用手册．北京：中国统计出版社，1997：772-833

[53] Paul D. Allision. Survival analysis using the SAS system: A Practical Guide. SAS Publishing, 2001: 29-184

[54] Alan B. Cantor. SAS survival analysis techniques for medical research. SAS Publishing, 2003: 153-186

[55] 余松林．临床随访资料的统计分析方法．北京：人民卫生出版社，1991：1-37

[56] 周惠馨，李卫，王杨，等．倾向得分法在非随机临床研究中的应用．中国循环杂志，2008，23(4)：289

[57] 汪涛，山口拓洋，大桥靖雄，等．倾向指数方法的蒙特卡罗研究．中华流行病学杂志，2005，26(6)：458

[58] 王永吉，蔡宏伟，夏结来，等．倾向指数的基本概念和研究步骤．中华流行病学杂志，2010，31(3)：347

[59] 王永吉，蔡宏伟，夏结来，等．倾向指数常用研究方法．中华流行病学杂志，2010，31(5)：584

[60] 王永吉，蔡宏伟，夏结来，等．应用中的关键问题．中华流行病学杂志，2010，31(7)：823

[61] 李智文，张乐，刘建蒙，等．倾向评分配比在流行病学设计中的应用．中华流行病学杂志，2009，30(5)：514

[62] Guiping Yang, Stephen Stemkowski, William Saunders. A Review of Propensity Score Application in Healthcare Outcome and Epidemiology. 2007

[63] U. S. Department of Health and Human Services, Food and Drug Administration Center for Drug Evaluation and Research (CDER) , Center for Biologics Evaluation and Research (CBER). Guidance for Industry Non-Inferiority Clinical Trials, 2010

[64] 刘玉秀，姚晨，陈峰，等．临床非劣效性/等效性评价的统计学方法．中国临床药理学与治疗学，2000，5(4)

[65] 刘玉秀，姚晨，陈峰，等．非劣性/等效性试验的样本含量估计及统计推断，中国新药杂志，2003，12(5))：371

[66] International Conference on Harmonization: Statistical Principles for Clinical Trials (ICH E-9), Food and Drug Administration, DHHS, 1998