公共卫生硕士(MPH)系列教材

总主编 姜庆五

医学数据分析

主编 赵耐青 尹 平

YIXUE SHUJU FENXI

复旦大学出版社
www.fudanpress.com.cn

公共卫生硕士(MPH)系列教材

总主编 姜庆五

医学数据分析

●主 编 赵耐青 尹 平

副主编 余红梅 凌 莉 秦国友

编 委（按姓氏拼音排序）

蒋红卫 华中科技大学

凌 莉 中山大学院

娄冬华 南京医科大学

秦国友 复旦大学

宋艳艳 上海交通大学

吴 骋 第二军医大学

薛付忠 山东大学

尹 平 华中科技大学

余红梅 山西医科大学

曾令霞 西安交通大学

赵耐青 复旦大学

朱彩蓉 四川大学

邹莉琳 同济大学

复旦大學出版社

图书在版编目(CIP)数据

医学数据分析/赵耐青,尹平主编. —上海:复旦大学出版社,2014.7
公共卫生硕士(MPH)系列教材
ISBN 978-7-309-10182-9

Ⅰ. 医… Ⅱ. ①赵…②尹… Ⅲ. 医学-数据-统计分析-研究生-教材 Ⅳ. R195.1

中国版本图书馆 CIP 数据核字(2013)第 275072 号

医学数据分析
赵耐青 尹 平 主编
责任编辑/傅淑娟

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
大丰市科星印刷有限责任公司

开本 787×960 1/16 印张 18.5 字数 297 千
2014 年 7 月第 1 版第 1 次印刷

ISBN 978-7-309-10182-9/R·1355
定价:45.00 元

主编简介

赵耐青，1954年1月出生。复旦大学公共卫生学院生物统计学教研室主任、教授、博士生导师。1983年毕业于复旦大学数学系理学学士，1996年毕业于澳大利亚Newcastle大学，获医学统计学硕士学位。上海市预防学会卫生统计学专业委员会主任委员，中国卫生信息学会卫生统计教育专业委员会副主任委员，现场统计研究会生物医学统计分会副理事长，中国医学数学会副理事长，中国卫生信息学会常务理事，中国卫生信息学会卫生统计理论与方法专业委员会常务理事。主要研究方向：预测、预报和预警方法及其应用，临床试验设计与统计学方法研究，医学研究中的统计方法学及其应用，时间序列分析和流行病学统计模型。主编和副主编10余本医学统计类教材，在国内外杂志发表200余篇论文。负责国家自然科学基金课题2项、科技部重大专项的子课题2项、其他国内外课题10余项，以及负责20余项临床试验的统计。

尹 平，1965年8月生，博士。华中科技大学公共卫生学院流行病与卫生统计学系副主任、教授、博士生导师。一直从事生物统计学的教学与研究工作，现为中国卫生信息学会卫生统计教育专业委员会副主任委员/统计理论与方法专业委员会常务委员、国际生物统计学会中国分会（IBS-CHINA）常务理事、武汉市预防医学会卫生统计专业委员会副主任委员。主编或副主编《医学统计学》、《SAS统计软件应用》等教材4部，主持国家及国际合作课题多项，发表学术论文60余篇，担任10多本专业杂志编委及审稿专家。

前　言

医学统计学理论是数据分析的统计方法学基础，但是仅仅学习和掌握统计方法学理论对于数据分析实践而言往往是不够的。不少医学统计学考试成绩很高的学生在数据分析实践中往往不知所措，特别在多因素统计分析中，同一样本资料，针对不同的研究问题和分析角度，往往得到的分析结果差异很大，对结果的解释和统计学结论也大相径庭。因此研究生的统计学学习不仅仅是学习统计学方法，更重要的是学习统计学分析的思路。要以研究问题作为分析目标，以研究设计和抽样方法或随机分组方法作为分析主线，围绕评价指标和研究因素的关系为核心，控制和评估其他因素对评价指标与研究因素的关联性的影响，根据资料分布类型选择合适的统计方法进行统计分析，科学地解释结果，依据分析结果所提供的证据，谨慎地给出统计学结论。

由于绝大多数研究生在本科期间在不同程度和不同层次上学习过医学统计学，但全国各个地区统计学教学范围有一定的差异，教学水平和统计学理论水平也是不尽相同。基于读者曾经学习过医学统计学或有其他统计学基础，本教材前 3 章内容主要以统计学基础复习为目的，相对简略地介绍相关统计学基础知识，完善读者在本科期间的统计学知识的薄弱点和强化统计学基本概念。第 4 章至第 14 章主要介绍多因素统计分析和其他常用统计分析。

本教材适用于公共卫生硕士（MPH）数据分析课程和多因素统计课程，也适用于八年制临床医学专业和其他医学专业的医学统计学研究生课程。本书可用作其他专业工作者的数据分析参考书。

本教材的编写得到了复旦大学公共卫生学院领导和复旦大学出版社的大力支持，在此表示衷心感谢。

本教材凝结着全国16所院校24位编委的智慧和心血，没有他们的辛勤劳动和无私奉献就没有这本教材。我谨在此向各位同仁致以崇高敬意和深深谢意！我们也特别珍惜在教材编写过程中结下的友谊！我还要感谢叶晨老师对稿件提出了许多修改意见，感谢我的研究生金欢、毕煜和李宝月进行整理稿件等烦琐事务。

由于本人能力所限，教材中不免存在不足之处，敬请广大师生提出宝贵意见。

赵耐青

2014年2月于上海

目　录

第一章　绪论 …… 1
第一节　数据分析中的若干基本概念 …… 1
第二节　常用多因素统计分析方法概述 …… 5
第三节　常用统计分析软件介绍 …… 6

第二章　定量资料的基本统计分析方法 …… 10
第一节　描述性统计分析 …… 10
第二节　样本资料平均水平的统计检验 …… 13
第三节　直线相关与直线回归 …… 31

第三章　定量资料的多因素分析(Ⅰ) …… 46
第一节　两因素方差分析 …… 46
第二节　多因素线性回归分析 …… 54
第三节　应用多因素回归控制混杂效应 …… 59
第四节　自变量的筛选 …… 62
第五节　多重线性相关和回归诊断分析简介 …… 66

第四章　定量资料的多因素分析(Ⅱ) …… 74
第一节　分类自变量的线性回归分析 …… 74
第二节　重复测量资料的 Mixed 模型 …… 80
第三节　离散型定量资料的 Poisson 回归和负二项回归 …… 85

第五章　分类资料的基本统计分析 …… 93
第一节　2×2 表格资料的统计分析 …… 93
第二节　多个 2×2 表格资料的统计分析 …… 101
第三节　2×C 表格有序分类资料的统计分析 …… 103

第四节 多个 2×C 表格有序分类资料的统计分析 …… 106
第五节 R×C 表格分类资料的统计分析 …… 109

第六章 Logistic 回归 …… 114
第一节 二分类反应变量的 Logistic 回归 …… 115
第二节 多分类反应变量的 Logistic 回归 …… 126
第三节 1∶M 条件 Logistic 回归 …… 133

第七章 对数线性模型在分类资料中的应用 …… 138
第一节 二维列联表的对数线性模型 …… 138
第二节 三维列联表的对数线性模型 …… 148
第三节 对数线性模型与 Logistic 模型的关系 …… 154

第八章 生存分析 …… 163
第一节 生存分析的基本概念 …… 163
第二节 生存率的估计与生存曲线 …… 168
第三节 生存曲线的 Log-rank 检验 …… 173
第四节 Cox 比例风险模型 …… 175
第五节 应用实例 …… 181

第九章 判别分析 …… 190
第一节 二类判别分析 …… 190
第二节 多类判别分析 …… 192
第三节 逐步判别 …… 195
第四节 应用实例 …… 198

第十章 主成分分析 …… 210
第一节 主成分分析的原理 …… 210
第二节 主成分分析的方法与步骤 …… 211
第三节 应用实例 …… 212

第十一章 因子分析 …… 217
第一节 探索性因子分析 …… 218

第二节　确定性因子分析 …… 226
第三节　应用实例 …… 228

第十二章　诊断试验 …… 233
第一节　试验设计中的基本概念 …… 233
第二节　常用诊断试验的评价指标 …… 234
第三节　ROC 曲线的应用 …… 238

第十三章　综合评价和综合分析方法 …… 249
第一节　综合评价与综合分析的基本概念与步骤 …… 249
第二节　层次分析法 …… 257
第三节　Meta 分析 …… 262
第四节　应用实例 …… 271

第十四章　群体评价指标的统计推断方法 …… 278
第一节　群体评价指标的统计推断问题 …… 278
第二节　Bootstrap 实现 …… 279

参考文献 …… 288

第一章 绪 论

传统的卫生统计学教学观念比较强调统计方法学的讲授，相对忽略数据分析技能的教学，有些人甚至把卫生统计学方法与数据分析技能的内容视为一谈。事实上，许多刚刚学习完卫生统计学的学生在面对实际问题的数据分析时往往感到束手无策，不知从何下手。国内许多刚刚从数理统计专业毕业的本科学生在面对实际问题的数据分析时也往往感到非常困惑。在一些国际知名大学中，一般都会开设数据分析技能的课程，帮助学生从掌握统计分析方法过渡到掌握数据分析技能。由此可见，数据分析技能与统计分析方法既有密切的联系，也有互相不能代替的方面。本教材编写的目的就是要帮助广大公共卫生硕士(MPH)学生和医务工作者能够较迅速地掌握数据分析的基本技能，拓展基本的多因素统计分析方法和提高数据综合分析技能。

第一节 数据分析中的若干基本概念

一、确认性研究与探索性研究的要求差异

一般而言，医学研究可以粗略地分为确认性研究、探索性研究和其他类型的研究。确认性研究的研究目的就是验证某一研究假设是否成立，如Ⅲ期临床试验就是一个确认性研究。探索性研究一般是某一类研究问题的初期研究或者为确认性研究设计做前期工作的一类研究，如Ⅱ期临床试验就是一个探索性研究，一般就是为Ⅲ期临床试验摸索最佳剂量和初步评价安全性的研究。

由于确认性研究所得出的结论往往被行内专业人士和官方机构采纳并且作为肯定性的结论应用于实际工作中，因此确认性研究需要严格控制统计学中的第一类错误发生的概率，尽可能避免出现假阳性的错误结论。例如：在Ⅲ期临床试验中，一个实际无疗效的药却获得优于对照药的结论，这就是典型的假阳性的错误结论。为了尽量减少出现假阳性的结果和结论，确认性研究在数据分析中往往要做出许多限制，使所下的结论为假阳性的概率 $\leqslant \alpha$。因此，确认性研究一般需要在研究设计中严格定义主要评价指标。在大多数情况下，确认性研究只定义一个主要评价指标。如果确实需要定义多个主要

评价指标，则需要在研究设计中事先定义多个主要评价指标与研究结论之间的关系，并且据此定义统计检验水平 α 与这些指标的关系。例如，在某个Ⅲ期临床试验中，定义了两个主要评价指标，并且规定如果两个评价指标的统计检验结果 P 必须均小于等于 α（称两个评价指标为“且”的关系）才可以得出试验药优于对照药的结论，该研究要求出现假阳性结论的概率小于等于0.05，则可以取 $\alpha = 0.05$。如果在上述临床试验中，规定两个主要评价指标中至少有一个指标的统计分析检验结果 $P \leqslant \alpha$（称两个评价指标为“或”的关系），就可以得出试验药优于对照药的结论，并且同样要求出现假阳性结论的概率小于等于0.05，则可以取 $\alpha = 0.05/2 = 0.025$。研究设计还可以针对两个评价指标的重要性不同，分别定义两个评价指标的统计检验水平。如定义第一个评价指标的统计检验水平为 $\alpha_1 < 0.05$（称该指标消耗检验水准为 α_1），第二个评价指标的统计检验水平为 $\alpha_2 = 0.05 - \alpha_1$（称该指标消耗检验水准为 α_2），则只要第一个评价指标的统计检验 $P \leqslant \alpha_1$ 或第二个评价指标的统计检验 $P \leqslant \alpha_2$，就可以得出试验药优于对照药的结论。在多个评价指标的关系为“或”的情况下，要求各个评价指标所消耗的检验水准 α_i 之和小于等于0.05。除了主要评价指标外还可以定义次要评价指标，但验证该研究假设是否成立的结论是要依据该研究所定义的主要评价指标统计结果得到的，但如果次要评价指标统计结果与主要评价指标统计结果所得到的结论是相冲突的，则研究者对下该结论将要十分谨慎，并要有充分的理由解释主要评价指标的统计结果与次要评价指标的结果为何是冲突的，且要充分解释这些评价指标的合理性和科学性。绝大多数确认性研究都是在探索性研究结果的基础上定义一个主要评价指标，然后通过确认性研究进行验证，当然也有少量的确认性研究的背景必须用两个或多个主要评价指标，如果该研究确实需要多个评价指标，则需在研究设计中定义多个主要评价指标之间的关系。对于探索性研究不仅应该完成研究设计中所指定的统计分析计划，而且应该积极探索研究计划以外的可能有用信息，为进一步研究及其研究设计提供尽可能丰富的信息，所以探索性研究在研究设计中可以设定多个主要疗效指标，也不需要根据主要疗效指标的个数调整检验水平 α。

确认性研究设计中，要制订非常详细的统计分析计划，其中包括对数据集的定义，哪些指标选用何种统计分析方法等。

二、统计学结论的表述

在假设检验中，当 $P \leqslant \alpha$，可以拒绝 H_0，推断 H_1 成立，但当 $P > \alpha$ 时，不能

推断 H_0。例如：两个样本均数比较检验时，如果满足成组 t 检验条件，则可以用成组 t 检验进行统计分析，$\alpha = 0.05$，当统计检验结果为 $P > 0.05$，则只能称两组差异无统计学显著意义，但不能称两个总体均数相等。

三、单侧假设检验的选择

在假设检验中，存在双侧检验和单侧检验。一般而言，选择单侧检验需要有足够的理由。例如：在成组 t 检验中，有足够的其他证据表明某个总体均数不会低于另一个总体均数（$\mu_1 \geqslant \mu_2$），则可以考虑采用单侧统计检验；又例如：检验某个水源中的水含细菌数的平均数 μ 是否少于某个合格标准 μ_0。对于该水源的抽样数据，如果采用双侧检验（$H_0: \mu = \mu_0$　$H_1: \mu \neq \mu_0$），则对于背景而言是没有意义的，因此需要采用符合背景的单侧检验 $H_0: \mu \leqslant \mu_0$（水源中的水含细菌数没有超过标准，合格），$H_1: \mu > \mu_0$（水源中的水含细菌数超标，不合格）。一般而言，在没有足够证据的情况下，都采用双侧假设检验。即使某种情况需要采用单侧假设检验，检验显著性水平往往选择较低的水平（如：$\alpha = 0.025$）。

四、变量和资料

变量亦称观察指标。变量的取值表示观察值（或测量值）或对应的观察结果，亦称资料。例如：身高测量值；尿常规检查结果可以为－、＋、＋＋；血型为 A、B、AB、O。由于医学观察结果往往在观察前是未知的，而且即使在相同条件下，重复观察结果往往也是不同的（如观察血压等），故称这种观察结果是随机的，可以称这种观察为随机试验，并且称这类观察指标为随机变量，简称为变量。根据变量取值的特征不同，变量可以分为连续型变量和离散型变量。

五、连续型变量

连续型变量是可以在一个区间中任意取值的，即在忽略测量精度的情况下，连续型变量在理论上可以取到区间中的任意一个值，并且通常含有测量单位。观察连续型变量所得到的数据资料称为计量资料。例如，身高变量就是连续型变量，身高资料为计量资料。

六、离散型变量

变量的取值范围是有限个值或者为一个数列。离散型变量的取值情况

可以分为具有分类性质的资料和不具有分类性质的资料，表示分类情况的离散型变量亦称分类变量。观察分类变量所得到的资料称为分类资料。分类资料可以分为二分类资料和多分类资料，而多分类资料又分成无序分类资料和有序分类资料(图 1-1)。

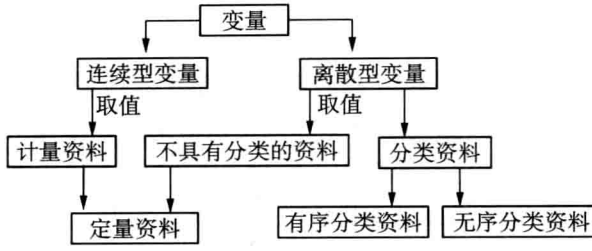

图 1-1 变量与资料的分类关系

七、二分类资料

如症状指标分为感染或未感染，疗效指标为有效或无效；又如性别变量为男或女等。观察可能的结果只有两个，记录这种观察结果的资料称为二分类资料。通常用变量取值为 0 和 1 对应两种可能的结果，所以这种变量也称为 0-1 变量，相应其取值的资料称为 0-1 资料。虽然二分类资料也可以从背景上分为有序情况和无序情况，但无论有序情况和无序情况，在统计学分析时所用的统计分析方法是相同的，所以二分类资料一般不区分有序和无序的情况，并且都可以归类为无序分类资料。

八、无序多分类资料

如血型可以分为 A、B、AB 和 O 型；又如慢性气管炎可以分为单纯型、喘息型、单纯型合并肺气肿和喘息型合并肺气肿。观察可能的结果只有若干个，并且这种观察结果在背景意义上没有程度或等级的含义，通常可用变量取值为 1，2，…，m 表示 m 个无序分类的属性或类别，故这种变量的取值没有大小的背景意义，仅是指示不同类别的作用，所以这种分类变量是无序的，这类资料称为无序多分类资料。无序分类资料可以先按类汇总，分别统计每一类的个体数，并将按类汇总的统计结果编制成表格形式的资料，这种汇总后的资料又可称为计数资料。

九、有序多分类资料

如病情指标分为无症状、轻度、中度和重度；又如疗效可以分为无效、进

步、有效和痊愈。对于每个个体而言，可能的观察结果只是若干个中的一个，若干个观察结果在研究背景意义上含有程度或等级上的差别，通常用变量取值为 1，2，…，m 对应表示相应 m 个程度或等级，所以这种分类变量称为有序变量，而这种分类变量值的集合构成有序分类资料。有序分类资料可以按类汇总，统计每个等级的个体数，并将按类汇总的统计结果编制成表格形式的资料，这种汇总后的资料又可称为等级资料。

十、不具有分类性质的离散型资料

有些观察指标，例如白细胞计数，其取值虽然是离散的，但不具有分类的性质，因此通常把这类观察指标的资料按较为特殊的计量资料处理。

在许多研究中，往往根据研究需要对变量进行变换。例如，研究对象的年龄是一个连续型变量，对年龄进行分组，定义一个年龄组变量 ageg：用 ageg＝1 表示年龄≤45 岁；ageg＝2 表示年龄＞45 岁并且年龄≤60 岁；ageg＝3 表示年龄＞60 岁。显然 ageg 是离散型变量，其取值为有序分类资料。如果定义 ageg＝0 表示年龄≤45 岁；ageg＝1 表示年龄＞45 岁，则 ageg 为二分类资料。

所以，上述资料的类型并不是一成不变的，可以根据研究目的的需要进行转化。一般而言，定量资料可以转换为有序分类资料，有序分类资料可以转换为二分类资料；反之，二分类资料不能转换为有序分类资料，有序分类资料也不能转化为定量资料。

由于变量类型与数据类型都是对应的，所以在实际应用中往往不严格区分变量与资料之间的差异，如有时称计量资料为连续型变量资料或连续型资料等。

第二节 常用多因素统计分析方法概述

在医学研究中，绝大多数的观察指标往往受多个因素的影响，并且各种因素也可能相互影响，因此单因素统计分析的结果很可能受到其他因素影响而产生偏差，造成结论不可靠或错误，所以在许多人群的观察性研究或临床研究中往往需要用到多因素统计分析，使其分析结果更可靠，结论更可信。从预测和分类角度分析，多因素统计的预测效果比单因素统计更优，同样多因素分类统计的效果也比单因素分类统计更佳。

最常用的多因素统计分析包括：多因素方差分析，协方差分析，多因素线性回归分析，Logistic 回归分析，Cox 模型，Poisson 回归分析，主成分分析，聚

类分析，判别分析和典型相关等。

多因素回归分析主要是刻画连续型因变量的总体均数与各个自变量构成的线性对应和变化关系，特别多因素方差分析和协方差分析也可以用多因素线性回归分析实现。

Logistic 模型主要刻画二分类或多分类因变量的概率在 Logit 变换下与各个自变量之间的线性对应和变化关系。Logistic 模型分为条件 Logistic 回归模型和非条件 Logistic 回归模型，条件 Logistic 回归模型适用于配对病例对照研究的样本资料，非条件 Logistic 回归模型适用于独立样本资料，非条件 Logistic 模型包含二分类 Logistic 回归模型、有序 Logistic 回归模型和无序 Logistic 回归模型，其中二分类 Logistic 回归模型是最常用的回归分析模型。

Cox 模型属于比例风险模型，它主要是分析各个影响因素变化所对应风险函数比（Hazard ratio，HR），并且可以证明对于满足比例风险模型条件前提下，对于风险函数（Hazard function）与生存率是一一对应的。因此，分析两组对象的生存率是否相同等价于分析两组对象的风险函数之比 HR 是否等于 1，因此 Cox 模型的统计分析主要关注点是风险函数比 HR 是否为 1 的统计推断。

Poisson 回归是要求资料满足：固定自变量取值情况下，用因变量 Y 刻画某一类事件的发生数服从 Poisson 分布，并且 Poisson 回归主要在资料满足 Poisson 回归条件下，刻画因变量总体均数的对数与各个自变量之间的线性对应和变化关系。

主成分分析的目的是把 m 个相关性较高或很高的观察指标构建成 $p(m \geqslant p)$ 个不相关的新指标，称这 p 个新指标为主成分，并且这 p 个主成分是由这 m 个观察指标的线性表达式构成的。原观察指标在回归模型中作为自变量因相关而容易造成多元共线，而采用主成分在回归模型中作为自变量不会造成多元共线。当 p 远小于 m 时，可以把 p 个主成分视为原 m 个观察指标的综合指标应用在实际研究中。

读者应根据自己的研究问题选择合适的统计分析方法进行统计分析，而不能根据自己所掌握的统计方法生搬硬套去进行统计分析。

第三节 常用统计分析软件介绍

由于计算机技术的高速发展，目前绝大多数情况下数据分析都是借助统计软件实现，比较常用的统计软件是 SPSS，Stata，SAS 和 R，以下将简单介

绍这4个统计软件的特点。

SPSS软件的研发是以社会科学研究为主要背景的，该软件的最大特色是菜单功能很强，操作比较方便，运算速度快，但统计分析方法不太全，该软件往往没有提供深入统计分析的功能，因此对于较为粗犷的统计分析是比较适合使用的，而对统计分析输出结果的再处理和再分析很困难。

Stata统计软件中相当部分的模块是以医学研究为背景研发的，该软件在统计学上是比较严谨的，一般不提供那些存在较大争议的统计方法的分析模块，该软件提供临床诊断试验、流行病学、生存分析等与医学关系非常密切的专用模块、缺失数据和开口资料统计分析的模块、样本量估计等模块。因此，可以认为该软件提供了较为完备的统计分析模块，速度相对较快。正是由于统计分析方法较全，其菜单往往不能包容这么多统计分析方法，即其菜单功能相对较弱，其统计分析输出结果的再处理和再分析功能也相对较弱。

SAS统计软件提供了非常强大的统计分析模块，一般可以认为SAS提供绝大多数的统计分析方法的分析模块，可以借助该软件的ODS功能，通过编程把统计分析结果直接输出成按照研究者所设计格式的Word文档和Word表格，因此在一些正规的研究中(如临床试验等)一般都采用SAS软件进行统计分析。同样一些国家级重要项目也一般采用SAS软件进行分析，往往借助预先编制的宏(统计模块的整合成一个新的模块称为宏)，把一个项目的统计分析报告在SAS软件输出成一个Word文档。SAS最大的缺陷是任何统计分析都是用通过编程来调用相关的模块实现的，所以对临床医生和其他医学研究者而言，SAS软件不是一个最好的选择，SAS软件应该是统计工作者的专业软件。

R是一个免费统计软件包，它是一些统计工作者和统计爱好者自发研究上传的一个软件包，在这个软件包中包含几乎所有的统计分析方法的模块，其输出结果再处理和再分析功能强大，通常为统计方法研究者所使用。该软件没有提供操作菜单，任何统计分析都需要编制程序完成，更大的担忧是R的统计模块包是没有经过专业认证的，所以分析结果可能有误，也无人对其输出结果承担责任。

对于研究生学习医学统计分析而言，强调记忆一些统计方法的公式已经没有很大的实际意义，但是学习操作统计软件并不能代替学习统计方法，更不能取代医学研究中的统计设计、整体统计分析思路和统计分析计划。我们一般强调：学生应该牢固掌握统计学基本概念，理解各个统计量的统计学意义，掌握统计分析的基本理念和分析思路，正确和合理地解释分析结果以及

谨慎地依据统计分析结果下统计学结论，同时也要至少熟练掌握一个统计软件的基本操作。

小 结

本章主要涉及的概念为：变量和资料类型，其中包括连续型变量和离散型变量，离散型变量所对应的资料基于其背景意义分为分类资料和不具有分类性质的资料（归为定量资料），分类资料又可以分为无序多分类资料和有序多分类资料，二分类资料归为无序多分类资料。

一般情况下，统计检验都是采用双侧检验，并且取 $\alpha = 0.05$；特殊情况下，可以采用单侧检验，并且取 $\alpha = 0.025$。

常用的多因素回归模型为多因素线性模型，多因素 Logistic 模型和 Cox 模型。如果因变量是连续型变量，可以考虑用多因素线性模型；如果因变量是分类变量，可以考虑用 Logistic 模型；如果因变量是刻画生存时间（更一般的情况为因变量是刻画某事件经过多少时间后发生的），则可以考虑采用 Cox 模型，但均需要考虑资料是否符合这些模型的要求。

思考与练习

一、是非题

1. 家庭中子女数是离散型的定量变量。

2. 学校对某个课程进行 1 次考试，可以理解为对学生掌握该课程知识的一次随机抽样。

3. 取 $\alpha = 0.05$，则选择单侧检验的第一类错误的概率高于选择双侧检验的第一类错误的概率。

二、选择题

1. 下列属于连续型变量的是（　　）。

A. 血压　　B. 职业　　C. 性别　　D. 民族

2. 某高校欲了解大学新生心理健康状况，随机选取了 1 000 例大学新生调查，这 1 000 例大学生新生调查问卷是（　　）。

A. 一份随机样本　　B. 研究总体　　C. 目标总体　　D. 个体

3. 某研究用 X 表示儿童在一年中患感冒的次数，共收集了 1 000 人，请问：儿童在一年中患感冒次数的资料属于（　　）。