一本临床医生看得懂、用得上的统计书

Step-by-step Examples of Data Analysis and Statistical Graphs in Clinical Researches

临床医学

研究中的统计分析和图形表达实例详解

EpiData+SPSS+GraphPad

主审◎马骏

主编◎周登远 崔壮 焦振山

军事医学科学出版社

Step-by-step Examples of Data Analysis and Statistical Graphs in Clinical Research

临床医学

研究中的统计分析和图形表达实例详解

临床医学研究中的统计分析和图形表达实例详解

（EpiData +SPSS +GraphPad）

主　审　马　骏（天津医科大学公共卫生学院）

主　编　周登远（天津市中医药研究院）

崔　壮（天津医科大学公共卫生学院）

焦振山（天津市中医药研究院）

副主编　贾真琳（天津医科大学公共卫生学院）

赵光宇（军事医学科学院微生物流行病研究所）

颜利求（河北医科大学沧州临床医学院）

雷铭德（天津医科大学第二医院泌尿外科研究所）

编　者　（以姓氏拼音排序）

李长平　（天津医科大学公共卫生学院）

刘学慧　（河北医科大学公共卫生学院）

于　婷　（军事医学科学院微生物流行病研究所）

郑　纺　（天津中医药大学）

张　磊　（武警医学院）

赵林胜　（天津市儿童医院）

军事医学科学出版社

·北　京·

内容提要

本书分为预备篇、统计分析篇和统计绘图篇。预备篇介绍了统计学基本知识、统计方法的选择和如何采用 EpiData 或 SPSS 建立数据文件；统计分析篇以 SPSS 18 中文版为介绍对象，实例解说了计量资料、计数资料、生存资料和诊断试验中的统计分析，涵盖了临床医学科研中 90% 以上的数据分析内容；统计绘图篇则在国内首次介绍如何采用 GraphPad Prism 绘制出版级统计图，以统计方法为框架，以实例详解的方式，简单明了。本书以“实用”为出发点，力图在最短时间最大限度地满足临床医学科研中的实际需求。本书见解独到、语言幽默、简单实用，适合临床医生、研究生和高年级本科生，同时对基础医学科研工作者也有极大的参考价值。

图书在版编目(CIP)数据

临床医学研究中的统计分析和图形表达实例详解/周登远，崔壮，焦振山主编.
-北京：军事医学科学出版社，2011.2
ISBN 978-7-80245-711-9

Ⅰ.①临… Ⅱ.①周… ②崔… ③焦… Ⅲ.①医学统计-统计分析-应用软件 Ⅳ.①R195.1-39

中国版本图书馆 CIP 数据核字(2011)第 018222 号

出　版：军事医学科学出版社
地　址：北京市海淀区太平路 27 号
邮　编：100850
联系电话：**发行部**：(010)66931051，66931049，63827166
编辑部：(010)66931039，66931127，66931038
86702759，86703183
传　真：(010)63801284
网　址：http://www.mmsp.cn
印　装：北京市顺义兴华印刷厂
发　行：新华书店

开　本：787mm×1092mm　1/16
印　张：17.5
字　数：427 千字
版　次：2011 年 5 月第 1 版
印　次：2011 年 5 月第 1 次
定　价：43.00 元

出版说明

“统计有用，统计难学”几乎成了所有临床研究人员共同的心声，大家对统计的理解还停留在那些晦涩难懂的数学公式，以及以编程为基础的SAS软件。虽然从本科开始就开设了统计学课程，但最后临床医生能独立完成数据分析和处理成了一种奢望，慢慢地甚至产生了统计恐惧症。该书的作者均是常年从事临床数据分析的一线统计专家，多年从事临床科研咨询、统计教学和稿件审阅等工作，结合工作，越发觉得需要编写一本临床医生看得懂、用得上的统计实用书籍，本书具有如下特点：

一、选择EpiData 3.1录入数据，SPSS 18中文版统计分析，GraphPad Prism 5绘制图形

人生的智慧在于选择，作为一名非统计专业人员，选择软件需要兼顾易用性和权威性，而各医科大学在临床研究生中教授SAS的使用，如同教一个不会开车的人参加F1方程赛，其结果可想而知了。而同样权威的SPSS 18中文版，以全中文的图形界面，使独立完成统计分析变得如此简单。而采用GraphPad Prism通过简单几步即可完成出版级的统计图绘制，EpiData对于大型问卷调查录入具有无法比拟的优势。综合发挥各软件的优势，让它们成为你科研生涯的忠实伴侣。

二、以数据类型为框架，让你明明白白选择统计方法

传统的统计学教科书是以统计理论为框架，从分布开始谈起，不断深入；而统计软件教程则往往以软件界面为框架，逐个菜单栏进行介绍，但这两种框架在实际运用中会产生一些问题，一个研究中必定存在这些数据资料“能进行哪些统计分析”，以及如果要回答研究者提出的研究假设“需要采用何种统计分析”，所以“能做什么”和“想做什么”成为研究过程中最关心的问题。本书以数据资料（能做什么）为框架，在各种资料中，又以研究假设（想做什么）作为分类标准来介绍统计方法，详尽介绍了计量资料、计数资料、生存资料和诊断试验这四类临床科研中最常用数据资料的统计方法，而撇弃了医学临床科研中很少使用的聚类分析、时间序列分析等方法。中国有句古话“有所为，有所不为”，让一个临床医生掌握经济预测方法和市场调查分析意义并不很大。本书还单独列出“第二章　统计方法的选择”，以帮助大家更好更迅速地选用恰当的统计方法。

三、各节单独成章，以线性流程的讲解方式让你迅速掌握统计应用

书中大部分例子取自方积乾教授主编的卫生部统编教材《卫生统计学》(第6版)，熟悉的例题，通过SPSS软件再次重现。每节均由方法原理－分析示例－研究假设－数据录入－操作流程－结果解释－注意事项这样7个部分组成，你只要按书中提示一步一步完成即可掌握该统计方法。

四、以统计方法为框架介绍统计图绘制

以往书籍介绍统计图形均就图论图，逐一介绍直方图、条图、饼图和线图等的绘制，但在实际应用中对数据采用何种统计图形表达却是个难题，该书则以统计方法为框架，将统计分析和统计图形联系起来，让大家迅速选用合适的统计图形。

五、各节单独成章，以线性流程的讲解方式让你迅速掌握图形绘制

本书在国内首次对GraphPad Prism软件应用进行系统介绍，所有章节均由示例－该示例的图形表达－图形类型的选择－数据录入－坐标轴和图形调整－文字修饰6个部分组成，让你掌握发表级图形的绘制。

本书以“一切为了应用”为出发点，深刻把握临床医生在科研中的实际需求，帮助大家迅速有效地掌握统计知识，增强科研核心竞争力。

本书数据文件 www.mmsp.cn

作者信箱：周登远 zhoudengyuan@hotmail.com；崔壮 nicholastricle@hotmail.com

作者博客：www.zhoudenyuan.com

周登远　崔壮　焦振山

目 录

上篇　预备篇

中篇　统计分析篇(SPSS 18 中文版)

下篇　统计绘图篇(GraphPad Prism 5)

上篇

预备篇

>> 第一章　统计学的基本概念

一、样本与总体

1. 总体(population)　是根据观察目的而确定的同质观察单位的全体,即同质的所有观察单位某种变量值的集合。

2. 样本(sample)　样本是总体中随机抽取的部分观察单位的实测值的集合。

科学研究一般是通过样本来推断总体特征,其做法是从研究总体中抽取少量有代表性的个体,称为抽样(sampling),对这些个体组成的样本进行深入的观察与测量,获取数据(data),利用统计知识,透过样本数据对研究总体的规律进行推断(inference)。

二、变异与同质

1. 变异(variation)　同质个体同指标之间的差异称为变异。

2. 同质(homogeneity)　指研究事物现象存在的共性。它是统计研究的基础,是资料整理和分析的前提。

三、变量的分类

变量(variable):总体中个体特征总是通过一个或多个变量来描述,变异性的存在决定了我们要处理的是变量。本书把变量分为定性(qualitative)和定量(quantitative)两种。

1. 定性变量又分为分类变量和有序变量(等级变量)

(1)分类变量(categorical variable):又称名义变量(nominative variable),例如职业是一个分类变量,其可能的"取值"不是数字,而是工、农、商、学、兵等,这些成为分类变量的水平(level),为便于输入计算机,一般采用代码(code)1、2、3、4、5 等来表示各水平。最简单也是最常用的变量为二分类变量(binary variable),如性别(男、女)、疾病(有、无)和结局(生、死)等。

(2)有序变量(ordinal variable):指分类变量种种可能的"取值"中自然地存在着次序。例如,问卷调查中常问对某件事情的满意程度,给出了 5 个答案:极不满意、有点满意、中度满意、很满意、极满意。有些临床体检或实验室检验常用 -、±、+、++ 和 +++ 来表示测量结果。

2. 定量变量又分为离散变量和连续变量

(1)离散变量(discrete variable):离散型变量只能取整数值,如一个月中手术病人数,一年里的新生儿数。

(2)连续变量(continuous variable):连续型变量可以取实数轴上的任何数值。有些变量的数值由测量而得到,它们大多属于连续变量,如血压、身高、体重等。而有一些测量值,如红细胞数,虽然以"个"为单位时只能取整数值,但当数值很大而以"千"或"万"为单位时,又可以

取小数值，所以通常把这些变量也称为连续变量。

有时为了数据分析的方便，人们将一种类型的变量转化为另一种类型，但变量只能由“高级”向“低级”转化（定量变量—有序变量—分类变量—二分类变量）。

四、频率与概率

（一）频率（frequency）

指在相同条件下，进行有限 n 次重复试验，某随机事件 A 发生的次数与 n 次试验的比值，频率是个变数，随样本变化而改变。

（二）概率（probability）

是描述随机事件 A 发生的可能性大小的度量，概率是一个定值。假设在相同的条件下，独立进行 n 次重复试验，随着 n 逐渐增加，频率摆动的幅度越来越小，则该事件 A 为随机事件，其频率可作为概率的估计值。

五、误差的分类

误差（error）可以分为随机误差和非随机误差。

1. 随机误差又分为抽样误差和随机测量误差

（1）抽样误差：由于产生的根本原因是生物个体的变异性，故抽样误差的分布具有规律性。

（2）随机测量误差：对同一观察单位某项指标在同一条件下进行反复测量所产生的误差。

2. 非随机误差又分为系统误差和过失误差

（1）系统误差：可产生于设计人员、调查者或调查对象，也可由于考虑不当、汇总计算有误等造成，一般带有倾向性，其产生原因复杂，贯穿于研究全过程并对研究结果有影响，很难用统计方法评价它的影响。

（2）过失误差：是错误，一般应杜绝。

六、统计分析的流程

（一）根据临床实践，提出研究问题，进行科研设计

在医学实践过程中提出科研问题，然后围绕提出的研究问题，制订研究方案，统计分析人员应当从设计阶段就参与研究项目，而不是临床医生获得数据之后，才想到统计分析。医学研究一般有干预性研究（intervention study）和观察性研究（observational study）两种。医学干预性研究是人们通过规定对象的准入条件（entry criteria）、随机化、重复、匹配（match）以及盲法（blinding）等措施来控制主要的混杂因素。公共卫生方面的研究大多属于观察性研究，这类研究不可能人为地控制许多混杂因素，人们能做的主要是观察已经或将要发生的事情。对于混杂因素的处理办法是精心设计抽样方法、无误地记录可能有用的信息。

（二）进行科学研究，分析清楚资料的性质，并分解出观察与变量

研究方案出来后，需要严格按照研究方案进行，资料大体上可以分为计量资料和计数资料，计量资料指测定每个观察单位的某项指标量的大小所得的资料；而将观察单位按照某种属性或类别分组计数，所得各组观察单位数称为计数资料。分清楚资料类型后，需要将资料分解成观察与变量。变量在临床上称为指标，是指具有相同属性的测量值的集合；而观察是指对同

一观察对象的不同属性的集合，就构成了一条观察；观察与变量结合起来就能准确地描述二维空间的物体特征。区分资料为计量资料还是计数资料，然后将资料分解为观察与变量，这是资料分析的基础。

（三）结合以上两点，罗列出能够回答该问题的可选统计方法

根据研究目的和资料性质，选用相应的统计学方法，如计量资料中的 t 检验、方差分析、线性回归分析等，计数资料中的卡方检验、Logistic 回归分析等。不同的资料和研究目的有多种可供选择的统计方法；而多种不同的统计方法可以对应多种资料类型，回答多种问题，如秩和检验能处理不符合方差分析条件的计量资料，也可以分析等级资料。

（四）选用统计软件，尝试着进行相关的统计分析

大家经常认为，只要形成了数据表格，选对了统计方法，用软件一操作，就万事大吉了。其实，事情没有这么简单，统计分析是一个反复的过程，是一个系统工程，需要进行预分析、正式分析等，如一份计量资料，我们首先考虑进行方差分析，但是分析过程中发现方差不齐，我们可以改做秩和检验。如在 Logistic 回归过程中，可以选用全部进入法和逐步回归法，也可以两者均尝试一下，然后比较两种方法所得出结果的差异，再根据专业知识和分析目的，作出判断，可见统计分析不是一锤定音、一成不变的过程，而是不断尝试、不断思考、不断判断的过程。

（五）评估统计结果，结合专业来回答提出的研究问题

从统计结论到专业结论，大家都需要特别慎重，不可妄加推断，任意发挥。

>> 第二章 统计方法的选择

统计方法的选择既是一个科学问题，也是一个艺术问题，同样的数据可以采用不同的统计方法，而不同的数据也可以采用同一种统计方法，因此我们需要在统计方法的选择中把握其根本：数据资料的性质决定“能做什么”，而研究设计或研究目的则决定了“想做什么”。本书将数据按性质分为计量资料和计数资料，对于两种资料又根据不同的研究设计或研究目的，讨论其具体的统计方法。而生存分析和诊断试验两章，则属于特定的分析方法，其数据特征和分析目的有其固定的特征，所以单独列出进行讨论。

需要特别说明的是，计量资料和计数资料只是对资料的一种通俗叫法，并不是计量资料中没有定性变量，也不是计数资料中没有定量变量。我们测量和感兴趣的指标为定量变量（如每组病例的血压值）或定性变量（如每组中治愈的人数），我们就称之为计量资料或计数资料。下面就计量资料、计数资料、生存资料和诊断试验的统计方法的选择进行讲解。

第一节 计量资料的统计方法选择

计量资料的统计分析按照研究设计或研究目的的不同，可以分为三类：

成组设计：其目的在于比较各组所代表的总体之间均数或中位数的差别，包括 t 检验、方差分析和秩和检验三种类型。

相关分析：其目的在于研究两个变量之间联系的密切程度，又可分为线性相关和秩相关。

因果联系：其目的在于探讨自变量和因变量之间的因果关系，因变量又称为结果变量，通常为身高、血压等连续变量；自变量又称解释变量，自变量可以为连续变量、等级变量和分类变量，其分析方法称为线性回归分析。

一、成组设计

1. 简单成组设计的统计方法选择

单组或两组计量资料的统计方法

设计名称	前提条件是否满足及假设检验方法的选择	
	满足	不满足
单组设计	单样本 t 检验	单样本秩和检验
配对设计	配对 t 检验	配对样本秩和检验
成组设计	成组 t 检验	两组独立样本秩和检验

上表统计方法是成组设计中最简单的统计方法，也是应用非常广泛的统计方法。t 检验中要求数据来自正态总体，如果这一前提条件不满足，则采用对应的非参数检验（秩和检验）。后面介绍的方差分析和秩和检验可以视为该表的扩展。

2. 复杂成组设计的统计方法选择　在谈到复杂成组设计之前，有三个概念必须弄清楚：

（1）因素（factor）：指对测量结果可能有影响的变量，一般来说，因素不止一个水平，而分析的目的就是比较同一因素内各水平之间测量结果是否相同。在方差分析中，一般有一个或多个因素。

（2）水平（level）：因素的不同取值称为水平，例如，因素“性别”有男、女两个水平，需要注意的是有时水平是人为划分的，比如身高被分为高、中、低三个水平。

（3）交互作用（interaction）：如果一个因素的效应大小在另一个因素不同水平下明显不同，则称为两因素间存在交互作用。当存在交互作用时，单纯研究某个因素的作用是没有意义的，必须在另一个因素的不同水平下研究该因素的作用大小。有时两因素之间的交互作用无法测量，如随机区组设计的方差分析。

根据因素、水平和交互作用的不同，可以将方差分析分为如下几类：

方差分析类型

方差分析类型	因素	水平	交互作用
完全随机设计	1	>2	无
随机区组设计	2	>2	无
析因设计	2	≥2	有

（1）完全随机设计可以视为成组 t 检验的扩展，两者均为一个因素，但是成组 t 检验中水平数为2，而完全随机设计方差分析中水平数大于2。由于只有一个因素，故不存在交互作用。

（2）随机区组设计存在两个因素，但两个因素的地位并不相同，如考察不同饲料剂量对大白鼠体重的影响，饲料是研究因素，而大白鼠窝别为区组因素，区组因素是为了消除混杂因素而引入的，从研究设计上就要求饲料和窝别之间不存在交互作用，否则该设计不合理。

（3）析因设计存在两个因素，且两个因素地位相同，如考察缝合方法和缝合时间对于大白兔神经损伤后愈合的影响（测量指标为轴突通过率，为计量资料），分析时就应当考虑缝合方法和缝合时间是否存在交互作用。

复杂成组设计的统计方法选择

设计名称	前提条件是否满足及假设检验方法的选择	
	满足	不满足
完全随机设计	完全随机设计的方差分析	Kruskal-Wallis 秩和检验
随机区组设计	随机区组设计的方差分析	Friedman 秩和检验
析因设计	析因设计的方差分析	非参数较少，进行数据变换
重复测量设计	重复测量设计的方差分析	

此处需要注意方差分析的适用条件：

(1)独立性：要求各样本为相互独立的随机样本，才能保证变异的可加性(可分解性)。

(2)正态性：即所有观察值均从正态总体中抽样得出。

(3)方差齐：指假设总的模型无意义时方差齐。

以上条件中，独立性要求最严，其次为正态性和方差齐性。在重复测量设计中，由于各次测量违反了独立性原则，所以采用特殊的重复测量的方差分析。

二、相关分析

(1)线性相关：两个随机变量之间的联系，即适用于二元正态分布的资料，常用 Pearson 相关系数表示。

(2)秩相关：对于不服从双变量正态分布的资料，还有总体分布未知的资料和原始数据用等级表示的资料，常用 Spearman 秩相关系数表示。

三、因果联系

(1)简单线性回归：因变量(结果变量)为连续变量，自变量(解释变量)也只有一个连续变量。

(2)多重线性回归：因变量(结果变量)为连续变量，自变量(解释变量)有多个变量，可以为连续变量、等级变量和分类变量。统计软件只能处理连续变量，分类变量可以转换为哑变量处理，等级变量则可以按连续变量或哑变量处理。

第二节　计数资料的统计方法选择

R × C 表：包括四格表和列联表，是计数资料中最常见的一种类型。

因果联系：其目的在于探讨自变量和因变量之间的因果关系，因变量又称为结果变量，通常为二分类变量、多分类变量和等级变量，本书只谈到最常用的二分类变量；自变量又称解释变量，可以有多个自变量，自变量可以为连续变量、等级变量和分类变量，其分析方法称为 Logistic 回归。

一、R × C 表

四格表统计分析

项目	统计方法
一般四格表	χ^2 检验、Fisher 精确检验
配对四格表	McNemar 检验、Kappa 检验