2011年度国家出版基金资助项目

中华医学统计百科全书

徐天和 / 总主编

描述性统计分册

田考聪 / 主　编

中国统计出版社
China Statistics Press

2011年度国家出版基金资助项目

中华医学统计百科全书

徐天和 / 总主编

描述性统计分册

田考聪 / 主　编

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

中华医学统计百科全书．描述性统计分册/田考聪主编．—北京：中国统计出版社，2011.12

ISBN 978－7－5037－6468－4

Ⅰ.①中… Ⅱ.①田… Ⅲ.①医学统计－中国－百科全书 Ⅳ.①R195.1－61

中国版本图书馆 CIP 数据核字(2012)第 000699 号

描述性统计分册

作　　者/田考聪
责任编辑/梁　超
装帧设计/杨　超　李雪燕
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/787×1092 mm　1/16
字　　数/360 千字
印　　张/16.5
版　　别/2012 年 4 月第 1 版
版　　次/2012 年 4 月第 1 次印刷
书　　号/ISBN 978－7－5037－6468－4/R·11
定　　价/38.00 元

《中华医学统计百科全书》
专家指导委员会

《描述性统计分册》
编　委　会

序　言

国家统计局局长 马建堂

随着时代前进和科学技术的进步，我国的统计科学和医学统计工作的发展进入了一个崭新的阶段。统计科学既是认识社会现象与自然现象数量特征的手段，又是获取信息和进行科学研究的重要工具，历来为人们所重视。自20世纪20年代起，统计学理论与方法日益广泛地被应用于医学领域。近些年来，随着基因组学、蛋白质组学、药物开发、公共卫生、计算机和信息等学科的迅猛发展，统计学与医学学科的交叉融合不断深入，统计科学在医学领域中的应用与发展提高到了一个新水平。

医学统计是统计科学的重要分支，也是国民经济和社会发展统计的重要组成部分，它关系到人民健康水平的提高和国家的长足发展。医学是强国健民学科，医学研究的对象是人及人群的健康，具有复杂性、特殊性及变异性等特点，这无疑需要全面系统的统计分析方法的支持与帮助。随着统计科学的迅猛发展，一些新的统计方法如遗传统计、多水平模型、结构方程模型、健康量表等不断涌现。一方面这些新的统计方法和理论亟需在医学科学领域内推广应用，为医学发展提供支持和帮助，另一方面，医学科研工作者为了科学研究工作的需要也迫切要求了解和掌握一些最新的、全面系统的统计方法和理论。因此，对当代医学科学研究中的统计分析方法进行全面系统的研究与介绍，是十分重要的一件事情，《中华医学统计百科全书》正是在这样的背景下编纂而成的，它满足了当前医学科学发展的需要，不失为一部好的大型医学统计参考书。

《中华医学统计百科全书》自2009年1月开始编写，由国内外著名医学院校的统计学教授和专家担任主编和编委，可谓编写力量强大，在编写过程中，他们本着精益求精的精神，精雕细琢，采百家之所长，融国内外华人统计学专家之所成。历时三年，终成其册。本套书内容浩繁，共八个分册，包含描述性统计分册、单变量推断统计分册、多元统计分册、非参数统计分册、管理与健康统计分册、医学研究与临床统计设计分册、健康测量分册和遗传统计分册。各

分册在内容上相互衔接并互为补充，贯穿“从简单到复杂”，“从一般、传统到先进、前沿”的循序渐进的编纂思路，一改目前医学统计著述中普遍存在的方法之间或评价指标之间缺乏相互联系、过于分散和单一的状况，使医学统计理论与方法更加具备了系统性、完整性与时代前沿性。本套书结构严谨，层次分明，科学性强，既突破了传统的辞典式编撰方法，又吸取了辞典的某些特点，在实用性、知识性、可读性、可查性等方面均具独到之处。

《中华医学统计百科全书》适应了我国医学科学研究发展对统计分析方法的需要，本书的出版，势必会大大促进我国现代医学的发展。本书既是我国医学统计工作者、医疗卫生统计信息工作者、高等医学院校师生以及广大医务工作者必备的大型医学统计参考工具书，也适合于医学各不同层次和不同专业的读者阅读。我相信本书的出版，不仅对于促进我国医学统计发展，促进我国与国际生物医学统计间的交流，繁荣社会主义先进文化具有重要意义，而且该书也必定会成为广大医学科学研究工作者的良师益友，故欣然为之作序。

编者的话

近年来，医学统计科学发展迅速，如遗传统计、多水平模型、结构方程模型、健康量表等新的统计理论与方法不断涌现，并被应用到医学科研实践中。这些新的统计理论与方法在医学科学研究中的不断拓展应用，要求广大的医学科技工作者在工作中必须学习和掌握这些新知识。所以，怎样使这些新的统计理论与方法易于被广大的医学科技工作者接受和使用，以提高医疗卫生工作质量，成为统计学专家的首要解决的任务。为此，组织编纂一部适合于广大医学科技工作者学习和使用的工具书，成为当前形势之必需。《中华医学统计百科全书》(下文简称“全书”)正是基于这样的背景而孕育产生的。

编纂“全书”的想法一经提出，就得到了国内高等医学院校和科研院所的统计学专家们的赞同。专家们云集一堂，进行商讨，达成共识——要集全国高等医学院校和科研院所的统计学专家之力，编纂出一部内容全面、概念精确、表述完整、接近世界医学统计学先进水平、编辑形式简洁的大型医学统计学工具书。2008年，“全书”开始酝酿筹备，几经讨论，搭成框架条目，确定编写格式，并开始全面着手编写，终于于2011年初编纂出初稿。值得欣喜的是，在中国统计出版社的大力支持下，“全书”项目先后成功申报了国家出版基金(项目编号2011C_2—003)和全国统计科学研究(计划)课题(立项编号2011LY080)，皆荣获批准。有了国家出版基金和全国统计科学研究(计划)课题的支持，“全书”的编纂工作如虎添翼，更上台阶。

通过国内外数十所大学、医学院校与医学科研院所近百位统计学专家教授的共同努力，“全书”终于能够付梓成册，得以与广大读者见面，编者倍感欣慰。“全书”既全面介绍了医学统计学的基本理论、基本知识与方法，又介绍了大量新的统计理论与方法，对生物医学统计的传统方法及最新进展进行了全面梳理，同时还改变了目前医学统计著述中普遍存在的统计方法或指标之间缺乏相互联系，过于分散与单一的现象。这就形成了“全书”的特点：全面、系统、实用、前沿。

“全书”共8个分册：描述性统计分册、单变量推断统计分册、多元统计分册、非参数统计分册、管理与健康统计分册、医学研究与临床统计设计分册、健康测量分册、遗传统计分册，均由著名高校医学统计学教授担纲主编，同

时聘请国内外知名医学统计教授担任顾问。可谓举全国名校之力，集百家精英之长。在编写过程中，专家们严谨认真，精益求精，在注重科学性、知识性、先进性、可读性的前提下，紧紧把握医学科学研究与医疗卫生工作的特殊性和复杂性，精心研究论证各种统计理论与方法在医学领域的适用性与应用条件。为了便于读者学习和理解应用，书中不仅有理论分析，还提供了实例运用，并把计算机软件程序应用于其中，对统计方法或体系的科学性与可行性进行检验，使统计理论与医学实际得到紧密结合。在每一分册的内容安排上，遵循从简单到复杂、从一般到先进、从传统到前沿的原则，使各分册在内容上既相互衔接补充，融为一体，又能各自独立成册。为方便读者查阅，书中各条目层次分明，结构严谨，醒目易读，是广大医学科学工作者学习和使用、案头必备的大型医学统计工具用书。

“全书”在编写过程中，引用了相关专著及教材的部分资料，在此对引用资料的原作者表示衷心感谢！引用资料中多数已在书中注出，也有部分没有一一注出，对于没有注出的部分，在此敬请原作者给予谅解！中国统计出版社教材编辑部和滨州医学院的领导及同仁们为“全书”的编辑和出版付出了大量心血，在此致以诚挚感谢！

由于编者水平有限，书中难免会存在错误和不足之处，恳请广大读者提出宝贵意见。

最后，感谢您学习和使用“全书”，希望它能使您开卷有益。

总主编　徐天和

前　言

《描述性统计分册》作为《中华医学统计百科全书》诸分册之一，主要介绍了有关描述性统计的基本知识与基本理论。描述性统计是统计学的重要组成部分，是统计数据处理的基础。它主要研究在收集、整理数据的基础上，如何选择和利用图形、表格及相应的统计指标对数据的特征进行统计描述。

随着医学研究在广度和深度上的迅猛发展，广大读者对医学统计学提出的要求也越来越高，单就统计描述而言，所涉及的内容也越来越丰富和广泛，读者们也亟需了解和掌握大量的统计描述方法以应用于医学科研工作。在《描述性统计分册》的编纂中，我们以此为指导，尽量全面地收集编写了描述性统计的内容，以完整反映描述统计的全貌。因此，该分册既包括常用的描述性统计指标、常见的概率分布、统计资料的类型、数据的预处理等内容，又包括描述性统计所涉及到的概率论与数理统计学的有关基本概念和基本知识，同时在书中还介绍了进行统计描述时如何根据分析目的和数据资料的类型来选择恰当的统计量、统计图和统计表等，以满足读者的多方面需要。

《描述性统计分册》与《中华医学统计百科全书》其他分册一样，采用条目形式编写，内容安排上循序渐进，前后呼应，结构合理，层次清楚，便于读者查阅和阅读。整个分册以介绍基本概念、基本理论、计算方法为主，同时还辅以必要的应用实例，使统计理论与实践案例得以结合，既便于检索，又易于理解和掌握，因此具有更强的实用性。

鉴于水平所限，时间又较为仓促，尽管全体编者严肃认真，尽心竭力，书中之缺点错误仍在所难免，敬请广大读者给予批评指正。

田考聪

2011 年 10 月

目 录

医学统计学

医学统计学(medical statistics)是运用统计学原理与方法研究医学现象数字资料的搜集、整理、分析与推断的一门学科。统计学以数量说明事物的本质和发展规律,是认识社会现象与自然现象的重要工具,是一门应用性很强的学科。统计研究的特点是在质与量的辩证统一中研究现象和过程的数量表现,并以数量反映质的特征。其目的在于取得真实有效的科学结论,并通过搜集、归纳、分析和解释大量数据来完成这一使命。由于事物的数量表现既受本质规律的制约,又受许多偶然因素的影响,往往这些偶然因素(不确定性)掩盖了必然性,妨碍了对事物本质的认识。在医学现象中,人体、生物体以及与人体有关的各种社会、自然现象更是千差万别,具有广泛的变异性,因而有必要运用统计方法这一工具透过偶然现象来探测其规律性。因此,有学者认为统计学是处理资料中变异性的一门科学。

一些杰出的统计学家从19世纪20年代开始创立了概率论、数理统计基础,包括参数估计、假设检验、相关与回归分析、抽样理论等;近代的非参数方法、多元分析、数学模型等大大丰富了医学统计学的研究方法,使得医学统计学作为一门新兴的应用学科而建立起来。特别是计算机技术的高速发展,为医学研究在空间(因素或变量空间)广度上(横向发展)和时间深度上(纵向发展)提供了有力的工具,使复杂的运算得以实现,多因素分析得以开展,能方便地进行大量的信息储存与检索、模拟抽样等。近几年来,不少多元分析的计算程序相继问世,并形成软件包,更是加快了分析的速度,拓宽了应用范围。我国统计学家创立的秩和比法、交叉积差法等方法也丰富了统计方法的内容。此外,模糊数学、灰色理论及运筹学等又为定量研究提供了思路。

医学统计学的形成与发展,与自然科学、社会科学有着密切的联系,如数学、物理学、生物学、医学、系统科学、环境科学、社会学、心理学和计算机科学等。同时,医学统计学的发展又成为促进其他学科发展的有力工具。例如,统计推断的思维逻辑与合理的统计设计、统计分析方法的引入,使流行病学中的描述流行病学、实验流行病学、理论流行病学以及临床流行病学(DME)等有了丰富的内涵和方法学基础。

医学统计学的基本内容包括实验设计和数据处理两大部分,主要有以下几方面:

1 统计研究设计

医学研究在作调查设计与实验设计时,除了从专业上考虑外,还必须根据统计学要求进行周密设计,以保证实验结果的准确性、可靠性、严密性和可重复性。一个好的设计可以用较少的人力、物力和时间取得丰富而可靠的资料。例如一项研究课题,当其研究

目标确立之后，统计设计就是从研究的部署、实施，直到实验结果的解释，进行系统安排，这是实现研究目标的重要前提和保障。主要的设计方案有：配对设计、完全随机设计、随机区组设计、交叉设计、析因设计、拉丁方设计、正交设计、序贯设计、均匀设计、系统分组设计等。此外从专业角度出发，分为临床试验、现场试验、动物实验设计等。

2 统计描述

统计描述是数据处理的必经之途。利用统计指标及统计表、图描述资料的某些特征，为进一步作统计推断奠定基础。通常是根据科研设计获得的数据，按照明确的统计工作步骤，进行数据预处理。按分类变量、数值变量分别计算有关的样本统计量。如均数、标准差、比、率、危险度等指标来描述资料的某些特征。

3 单变量统计推断

统计分析的目的是由样本对总体的性质、特征和规律进行估计和推断。因此，统计学的主体是统计推断，包括总体参数估计和假设检验两大部分。它是根据研究目的和资料性质，利用样本统计量及其分布规律对总体特征或性质进行估计或推断的统计方法。常用的单变量统计推断方法有总体参数的点估计、区间估计、t 检验、F 检验、χ^2 检验、U 检验、非参数检验等。

4 多变量统计分析

由于医学现象的发生、发展和变化是多种因素在一定条件下相互影响、相互制约而产生的综合效应。为了充分利用医学资料众多因素的综合信息，分析健康状况及疾病的发生、变化、转归、预后等内在联系的客观规律，作出科学的符合实际的结论，需要运用涉及多个变量的统计分析方法——多元统计分析。其主要内容包括：多元线性回归、逐步回归、判别分析、聚类分析、主成分分析、因子分析、典型相关分析、Cox 回归、logistic 回归等。

5 预测与综合评价

在疾病防治过程中，经常需要进行多种检测结果的综合评定、选择治疗方案、进行效果预测预报等。医学统计学提供了必要的方法与手段，主要包括：时间序列模型、线性回归预测、灰色预测、先验信息条件下的统计决策、序贯决策、后验信息的统计决策、统计质量管理、综合指数评价、灰色系统法及 Meta 分析等。

21 世纪是高度发达的信息时代，医学科学的发展对统计方法会有更高的要求。在病因学探讨、临床效果及方法学评价、健康与疾病的预测等方面都需要医学统计学的介入。通过信息库与应用软件，提供可靠的基础数据，利用统计方法对信息进行加工、提炼，排除和减弱偶然因素的干扰，显示和突出事物的本质，为推断与决策提供可靠的统计信息。总之，在新的世纪里，医学统计学的应用领域将更加广阔，与医学实际的联系将更为紧密。

（田考聪　周燕荣）

概　率

概率(probability):是概率论最基本的概念之一,直观意义是描述随机事件发生的可能性大小的度量。

概率的严密数学定义:设(Ω,F)是可测空间,对每一集$A\in F$,定义实值集合函数P,它满足如下3个条件:① 对每一$A\in F$有$0\leqslant P(A)\leqslant 1$(非负性);② 对必然事件$\Omega$,$P(\Omega)=1$(规范性);③ 对任意$A_i\in F$,$A_i\cap A_j=\phi$,$i\neq j$恒有$P(\bigcup_{i=1}^{\infty}A_i)=\sum_{i=1}^{\infty}P(A_i)$(可列可加性),则称实值集函数$P$为$(\Omega,F)$上的概率,$P(A)$就称为事件$A$的概率。

概率的古典定义:如果一个试验满足:①试验只有有限个基本结果;②试验的每个基本结果出现的可能性是一样的。这样的试验,称为古典试验。对于古典试验中的随机事件A,它的概率定义为:$P(A)=f/n$,n表示该试验中所有可能出现的基本结果的总数,f表示事件A包含的试验基本结果数。这样定义的概率称为古典概率。

概率的统计定义:设在相同条件下,独立重复进行n次试验,事件A出现f次,则称f/n为事件A出现的频率。当n逐渐增大时,频率f/n始终在某一常数P的左右作微小摆动,称P为事件A的概率,记作$P(A)=P$。在许多实际问题中,当概率不易求得时,只要n充分大,可以将频率作为概率的估计值。

在一定条件下,肯定发生的事件称为必然事件,肯定不发生的事件称为不可能事件。可能发生也可能不发生的事件称为随机事件或偶然事件。必然事件的概率等于1,不可能事件的概率等于0,随机事件的概率介于0和1之间。概率越接近1,表示事件发生的可能性越大;概率越接近于0,表示事件发生的可能性越小。统计上的许多结论都是带有概率意义的,如果事件A满足:$P(A)<0.05$或$P(A)<0.01$,则称事件A为小概率事件,表示事件A发生的可能性很小。

概率运算法则

加法法则:$P(A+B)=P(A)+P(B)-P(AB)$。如A和B互不相容,$P(A+B)=P(A)+P(B)$,这一法则可以推广到有限个互不相容的事件。

乘法法则:在事件A已发生的条件下,事件B发生的概率,就称为事件B的条件概率,记作$P(B\mid A)$。任意两事件A和B同时发生的概率为:

$$P(AB)=P(B)P(A\mid B)$$

$$P(AB)=P(A)P(B \mid A)$$

若 A(或 B)发生与否并不影响 B(或 A)是否发生,则称 A 和 B 相互独立。若 A 与 B 相互独立,则:

$$P(AB)=P(A)P(B)$$

这一法则可推广到有限个相互独立的事件。

全概率公式:若 $A_i \cap A_j=\phi\ (i\neq j)$,$\bigcup_{i=1}^{n} A_i=\Omega$,$P(A_i)>0$,则:

$$P(B)=\sum P(B \mid A_i)P(A_i)$$

Bayes 公式:若 $A_j \cap A_i=\phi\ (i\neq j)$,$\bigcup_{i=1}^{n} A_i=\Omega$,$P(A_i)>0$,则在事件 B 出现的条件下 $P(B)>0$,事件 A_i 出现的概率为:

$$P(A_i \mid B)=\frac{P(A_i)P(BA_i)}{\sum_{i=1}^{n} P(A_i)P(BA_i)}$$

(易 东)

随机变量

随机变量(random variable):直观意义是在随机试验中被测出的具有一定概率分布的量,它是随机事件的数量化。对随机变量进行描述的重要工具是随机变量的分布。下面给出其用于建立分布理论体系的严密数学定义:设(Ω,F,P)是一个概率空间,对于$\omega\in\Omega$,$\xi(\omega)$是一个实值的单值函数,若对于任一实数 x,$\{\omega:\xi(\omega)<x\}\in F$ 是一随机事件,则称 $\xi(\omega)$为随机变量,而 $F(x)=P\{\xi(\omega)<x\}$称为 $\xi(\omega)$的分布函数。由定义可看出随机变量 $\xi(\omega)$总是联系着一个概率空间,即服从一定的概率分布。随机变量按其取值形式的不同分为不同的类型,常见的有离散型和连续型两种,均可用分布函数表达随机变量的概率性质。

1 离散型分布

若 ξ 的一切可能取值为 $x_1,x_2,\cdots,x_n,\cdots$,则称 ξ 为离散型随机变量。

如令 $p_n=P(\xi=x_n)(n=1,2,\cdots)$,称 $p_1,p_2,\cdots,p_n,\cdots$为 ξ 的分布列,亦称为 ξ 的概率函数。对离散随机变量,如下列出更为直观: