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

单变量推断统计分册

颜 虹/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

单变量推断统计分册

颜 虹 / 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．单变量推断统计分册/颜虹主编．
—北京：中国统计出版社，2011.12
ISBN 978－7－5037－6469－1

Ⅰ.①中… Ⅱ.①颜… Ⅲ.①医学统计－中国－百科全书②医学统计－统计推断 Ⅳ.①R195.1－61

中国版本图书馆 CIP 数据核字(2012)第 000700 号

单变量推断统计分册

作　　者/颜　虹
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/787×1092mm　1/16
字　　数/480 千字
印　　张/21
版　　别/2012 年 3 月第 1 版
版　　次/2012 年 3 月第 1 次印刷
书　　号/ISBN 978－7－5037－6469－1/R.12
定　　价/53.00 元

前　言

统计推断是根据样本信息来推断总体特征的过程。本分册所介绍的单变量统计推断是通过单个变量的信息认识总体特征的统计方法，主要包括单变量的参数估计和假设检验。单变量统计推断既是统计推断的基本方法，也是多变量统计推断的基础，应用时常与单变量统计描述相结合，是数据分析不可或缺的内容。

本分册以条目形式介绍了单变量推断的常用方法，全书共计101个条目，每一条目尽可能地包含单变量推断的统计思想和原理、应用条件、实例分析等三部分内容。其中计量资料的假设检验以参数检验方法为主，如检验和方差分析，也包含了以方差分析为主的不同实验设计分析方法，同时还涉及到一些多变量分析的方法，如析因设计、交叉设计和正交设计等。圆形分布资料的分析方法以单样本资料区间估计和两样本资料假设检验为主，多为非参数统计方法。此外还涉及统计推断基本概念和统计量的介绍以及变量变换和剂量反应等方法的应用。另外，本分册注重介绍数据分布特征在统计推断中的应用，以区别于《描述性统计分册》；介绍方法的同时又结合实例分析，使本册实用性更强，更具参考价值。

本分册在编写过程中，力求简单、实用、易懂，语言文字简练，公式避繁就简，所用实例皆来源于医学科研和日常医疗卫生工作，易于理解和掌握，有较强的实用性。书中每一条目既与其他条目相互关联呼应，又可独立成节，易于检索。因此，该书可作为医学工作者进行数据分析的必备参考书。

本分册成稿之际，谨对各位编者和为本书付出心血的人士表示衷心感谢，同时感谢本书中所引用的有关专著和教材的作者。限于水平有限，加之时间仓促，书中难免存在不足之处，恳请同行和广大读者不吝赐教。

颜　虹

2011年10月

《中华医学统计百科全书》
专家指导委员会

《单变量推断统计分册》
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、必备案头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

目 录

参数估计

医学研究中常采用抽样研究的方法，即从某总体中随机抽取一个样本来进行研究，并根据样本提供的信息推断总体的特征。抽样研究中，总体参数(population parameter)是未知的，因此只能通过样本数据计算样本统计量，再用样本统计量来估计总体参数。例如用样本均数 $\bar{x}$ 估计总体均数 μ、样本方差 S^2 估计总体方差 σ^2 等。参数估计就是指用样本指标值(样本统计量)推断总体指标值(总体参数)。参数估计有点(值)估计(point estimation)和区间估计(interval estimation)两种方法。点估计就是用相应样本统计量直接作为其总体参数的估计值。区间估计是按预先给定的概率$(1-\alpha)$所确定的很可能包含未知总体参数的一个范围。

(曾令霞　潘媞华)

参数的点估计

设 θ 为总体参数，从该总体中随机抽取含量为 n 的样本 $x_1, x_2, x_3, \cdots, x_n$，得样本统计量 $\hat{\theta}$，直接用 $\hat{\theta}$ 来估计总体参数 θ，这就是总体参数的点估计。例如直接用随机样本的样本均数 $\bar{x}$ 作为总体均数 μ 的点估计值。例如，某市 2008 年所有 7 岁正常男童的身高值是一个总体，但总体的参数 μ——平均身高值未知。为此，随机抽取该市 10 名 7 岁正常男童，计算其平均身高 $\bar{x}=121.45$(cm)，标准差 $S=5.75$(cm)，这两个均为样本统计量。若用样本均数 $\bar{x}$ 作为总体均数 μ 的一个估计，用样本的标准差 S 作为总体标准差 σ 的一个估计，即认为该市所有 7 岁正常男童的平均身高为 121.45(cm)，标准差为 5.75(cm)。这就是点估计。

总体参数 μ 是未知的，但它是固定的值，并不是随机变量；而样本统计量是随机的，不同的样本所得结果是不相同的。如果另有一个研究者作同样的研究，随机抽取该市另

外10名7岁正常男童，计算其平均身高为 $\bar{x}=119.78$(cm)，并以此作为总体身高的点估计，也是可以的。点估计方法简单，但未考虑抽样误差，无法评价其估计错误的概率。

（曾令霞　潘媞华）

抽样误差

从同一总体中随机抽取样本量相同的若干样本来进行研究，由于总体中各观察单位间存在变异，样本只包含了总体中的部分观察单位，且每次抽样时抽到的观察单位不尽相同，因此，每次计算的样本统计量与总体参数间一般会存在差异；而且，来自同一总体的若干样本统计量(如多次抽样的均数)间，一般也会存在差异；这种由个体变异和抽样引起的差异，称抽样误差(sampling error)。因此，只要有个体变异，抽样就必将产生抽样误差，即抽样误差是不可避免的，但只要我们遵循随机化原则进行抽样，则抽样误差的分布有一定的规律，抽样误差的大小是可以估计的。

现通过一个均数的抽样实验的例子，具体说明抽样误差的规律及其度量方法。假设将100名正常非孕、未哺乳妇女血红蛋白含量(g/L)为实验总体，其 $\mu=129.97$g/L、$\sigma=11.43$g/L。见表1。

表1　实验总体（$\mu=129.97$g/L、$\sigma=11.43$g/L）

编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)
1	131	13	138	25	133	37	127	49	138
2	127	14	144	26	138	38	120	50	131
3	121	15	129	27	121	39	152	51	116
4	149	16	135	28	151	40	125	52	121
5	126	17	114	29	138	41	130	53	145
6	118	18	125	30	119	42	126	54	113
7	115	19	112	31	119	43	122	55	118
8	114	20	127	32	111	44	129	56	134
9	133	21	139	33	112	45	138	57	133
10	156	22	136	34	143	46	149	58	138
11	130	23	143	35	148	47	116	59	133
12	130	24	146	36	113	48	112	60	123

续表

编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)	编号	血红蛋白含量(g/L)
61	113	69	140	77	138	85	154	93	112
62	136	70	115	78	142	86	123	94	139
63	122	71	125	79	128	87	136	95	134
64	120	72	126	80	124	88	136	96	129
65	132	73	131	81	130	89	129	97	114
66	123	74	142	82	144	90	133	98	114
67	138	75	112	83	116	91	134	99	143
68	142	76	143	84	134	92	130	100	148

采用完全随机抽样的方法，随机抽取 $n=10$ 的样本。

例

编号	34	55	63	64	67
血红蛋白含量(g/L)	143	118	122	120	138
编号	73	77	88	89	96
血红蛋白含量(g/L)	131	138	136	129	129

求得 $\bar{x}=130.40$ g/L，$S=8.45$ g/L。

第二个样本：

编号	3	8	34	41	54
血红蛋白含量(g/L)	121	114	143	130	113
编号	56	72	89	92	94
血红蛋白含量(g/L)	134	126	129	130	139

求得 $\bar{x}=127.90$ g/L，$S=9.80$ g/L。

如此反复抽取 100 个 $n=10$ 样本，其均数如表 2。

表 2　100 个样本均数

130.4	124.2	123.6	128.5	128.5	129.6	135.0	129.0	127.5	137.0
127.9	135.3	135.0	132.6	131.5	124.9	129.1	125.4	132.0	129.9
130.3	130.3	130.5	129.6	129.5	121.5	125.3	133.3	129.3	129.9
132.0	130.5	127.3	123.9	127.1	128.0	125.7	135.6	127.1	124.1
132.1	130.3	125.6	128.1	128.1	131.5	130.8	127.8	134.8	131.3
137.3	126.1	131.8	130.6	133.8	127.4	132.3	125.7	129.1	125.9
129.7	132.1	127.8	126.6	127.6	128.6	125.1	131.3	132.5	128.7
128.6	127.0	131.6	132.2	129.1	129.6	129.8	132.3	129.1	124.2
128.4	141.5	126.4	125.1	128.2	132.7	129.3	133.3	127.9	137.7
130.0	136.0	133.8	132.4	129.3	131.9	128.3	133.9	128.1	128.9

采用距法进行正态性检验，$D=0.064$，$P=0.2$，说明样本均数呈正态分布。即，从正态总体中随机抽样含量相同的样本，样本均数的分布呈正态分布。当总体不是正态分布，只要样本含量足够大，样本均数的分布亦呈正态分布，因此我们可用样本均数的离散程度的指标，样本均数的标准差 $\sigma_{\bar{x}}$（又称标准误）来衡量抽样误差的大小。

$$\sigma_{\bar{x}}=\sqrt{\frac{\sum(\bar{x}-\mu)^2}{k}}$$

其中 k 为抽样次数，本例 $k=100$，$\sigma_{\bar{x}}=3.462$ g/L，表示从该实验总体中随机抽取 $n=10$ 的样本，其抽样误差为 3.462 g/L。

（曾令霞　潘媞华）

标准误

通常，将样本统计量的标准差称为标准误（standard error，SE）。样本均数的标准差也称均数的标准误（standard error of mean，SEM），它反映样本均数间的离散程度，也反映样本均数与相应总体均数间的差异，因而说明了均数抽样误差的大小。进行抽样研究时，通常我们只有一个样本，理论上均数的标准误为：

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{1}$$

前述抽样实验中，$\sigma=11.43$g/L，$n=10$，按公式（1）得 $\sigma_{\bar{x}}=3.61$g/L。实际工作中，由于总体标准差 σ 通常未知，用样本标准差 S 来估计。因此均数标准误的估计值为：

$$S_{\bar{x}}=\frac{S}{\sqrt{n}} \tag{2}$$

如抽样实验的第一个样本，$\bar{x}=130.40$g/L，$S=8.45$g/L，代入公式（2），得 $S_{\bar{x}}=\frac{8.45}{\sqrt{10}}=2.67$。

标准误可反映抽样误差的大小，可以用它来求总体均数的可信区间及均数间比较的假设检验。

（曾令霞　潘媞华）