

供研究生和七年制各专业用



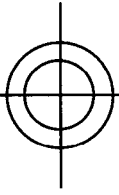
新世纪全国高等中医药院校创新教材

XIN SHI JI QUAN GUO GAO DENG ZHONG YI YAO YUAN XIAO
CHUANG XIN JIAO CAI

中医药统计学 与软件应用

主编 刘明芝 周仁郁

中国中医药出版社



新世纪全国高等中医药院校创新教材

中医药统计学与软件应用

(供研究生和七年制各专业用)

主 编 刘明芝 (湖南中医药大学)
周仁郁 (成都中医药大学)

中国中医药出版社
· 北 京 ·

图书在版编目(CIP)数据

中医药统计学与软件应用/刘明芝,周仁郁主编.
—北京:中国中医药出版社,2006.8
新世纪全国高等中医药院校创新教材
ISBN 7-80231-088-1

I. 中… II. ①刘…②…周 III. 中国医药学-医学统计-应用软件-中医学院-教材 IV. R2-32

中国版本图书馆 CIP 数据核字(2006)第 099807 号

中国中医药出版社出版
北京市朝阳区北三环东路 28 号易亨大厦 16 层
邮政编码:100013
传真:64405750
北京市卫顺印刷厂印刷
各地新华书店经销

*

开本 850×1168 1/16 印张 28.75 字数 676 千字
2006 年 8 月第 1 版 2006 年 8 月第 1 次印刷
书号 ISBN 7-80231-088-1 册数 5000

*

定价:35.00 元
网址 www.cptcm.com

如有质量问题请与本社出版部调换
版权专有 侵权必究

社长热线 010 64405720
读者服务部电话:010 64065415 010 84042153
书店网址:csln.net/qksd/

新世纪全国高等中医药院校创新教材

《中医药统计学与软件应用》编委会

- | | | |
|------|--------------|------------------|
| 主 编 | 刘明芝(湖南中医药大学) | 周仁郁(成都中医药大学) |
| 副主编 | 董 平(辽宁中医药大学) | 谢海林(山西中医学院) |
| | 吴翠珍(山东中医药大学) | 史周华(山东中医药大学) |
| | 赵文峰(河南中医学院) | 范薪生(贵阳中医学院) |
| | 王世钦(甘肃中医学院) | 林求诚(福建中医学院) |
| | 何 雁(江西中医学院) | 李大明(江西中医学院) |
| | 雷 磊(湖南中医药大学) | 崔相学(成都中医药大学) |
| | 尤昭玲(湖南中医药大学) | 李秀昌(长春中医药大学) |
| | 王泓午(天津中医药大学) | 孙 铭(首都医科大学中医药学院) |
| 编 委 | 张雪飞(湖北中医学院) | 王金虹(山西中医学院) |
| | 李 新(辽宁中医药大学) | 关明云(辽宁中医药大学) |
| | 王淑媛(长春中医药大学) | 于鹤丹(黑龙江中医药大学) |
| | 赵 莹(上海中医药大学) | 曹 敏(贵阳中医学院) |
| | 曹治清(成都中医药大学) | 杨胤清(成都中医药大学) |
| | 何迎春(湖南中医药大学) | 蔡 晶(福建中医学院) |
| | 郑洁刚(湖南中医药大学) | 黄爱武(湖南中医药大学) |
| | 汪旭升(广西中医学院) | 樊新荣(陕西中医学院) |
| | 姜 明(新疆医科大学) | 唐启义(浙江大学) |
| 学术秘书 | 何迎春(湖南中医药大学) | |

前 言

中医药统计学(Statistical for Traditional Chinese Medicine)是将数理统计学的原理和方法,应用于生物医药科研,特别是中医药科研,搜集、整理、分析资料,推断不确定现象客观数量规律的一门应用学科。贺石林教授1989年主编出版《中医科研设计与统计方法》一书,奠定了中医药统计学这一学科的基础。刘明芝主编的《中医药统计学》于1999年问世,构建了中医药统计学这一新兴的应用学科,经国家中医药管理局认可,学科名称“中医药统计学”的“国标_行标”代码为36070。2004年周仁郁主编出版了新世纪全国高等中医药院校本科规划教材《中医药统计学》,发展了这一应用学科。本书由全国高等中医药教材建设研究会组织,为填补七年制和研究生统计学教材空缺而撰写。

本书作为当代中医药统计学教学内容和方法的知识载体,作为中医药院校统计学教学的重要工具,在坚持三基(基础理论、基本知识、基本技能)和五性(思想性、科学性、先进性、启发性、适用性)的原则下,具有以下特色与创新:

(1)面向医药统计学教学和科研所需,结合中医药实际,内容新颖翔实。

本书将统计学理论和方法密切联系中医药科研实际,不但准确、系统阐述医药统计学的基本概念、基本原理和思维逻辑,而且注意吸取同类教材优点,反映医药统计学的新进展,引入新观点、新方法,内容覆盖医学、药学统计的主要方法,包括研究设计(含实验设计、调查设计,介绍了序贯设计、均匀设计、混料设计等方法)、基本统计方法(含 $P > \alpha$ 时假阴性率 β 的估算,曲线拟合,概率单位回归法计算半数致死量 LD_{50} ,联合用药效应判断,Ridit分析,Friedman M 检验,Cochran's Q 检验,游程检验,圆分布资料的参数估计、假设检验和相关性分析,等效性和非劣效性检验等)、多元分析(含Hotelling T^2 检验,多元方差分析,重复测量资料方差分析,二分类和多分类logistic回归,条件logistic回归,对数线性模型,Cox比例风险模型,典型相关分析)、诊断试验的ROC分析、新药临床试验、生存分析及流行病学方法、信度和效度分析、Meta分析,针对中医药研究常需处理模糊性概念、分析模糊性不确定性现象的特点,介绍了模糊数学综合评判的应用。阐明各种统计方法的用途、适用条件和注意事项,对容易混淆的概念和方法采用图示和综合归纳表进行引导和对比。

(2) 将统计学理论和方法与多种软件操作优化衔接,兼具教材和工具书功能。本书在介绍国际通用 SPSS14.0、SAS9.0 统计软件的同时,介绍适用于中医药专业的优秀统计软件 DPS7.55,编排方法主要贯穿 SPSS14.0,辅以 DPS7.55 解难,第二十七章专章介绍 SAS9.0 统计软件的应用。

统计软件的应用侧重数据文件的格式,操作思路和结果的解释,并在操作中说明了 SPSS13.0、14.0 与 SPSS11.5、12.0 版本的不同,为使用 SPSS13.0 以下版本的读者提供了方便;在第二十四章 Meta 分析中介绍了 RevMan 软件作 Meta 分析输出的森林图、漏斗图,以帮助读者扩大应用面。授课时,教师主要讲授中医药统计学的理论课,统计软件实验只要稍作引导,学生很容易掌握,可以自行完成。

(3) 注重统计分析方法的选择,第二十六章从不同方位对统计分析方法的选择进行了概括,纵横比较,广开思路。

(4) 适应面广,本书内容兼顾普及与提高,具有良好的适应性,可作为高等中医药院校各专业本科生、研究生统计课程教材,亦可作为培训和继续教育教材。不同层次的教学对象,可按学时选讲,利于继续学习和提高。

本书用 n 、 \bar{x} 、 s 等统计量表达计算公式,便于不方便使用统计软件时,采用具有统计功能的计算器进行计算。

全书文字规范流畅,论例配合科学,重点难点、突出,可作为生物医药科研工作者自学用书和工具书使用。

为方便教学,我们将书中例题的 SPSS 数据文件和 DPS 数据块放到网上,读者可输入网址 <http://www.chinadps.net/STCM.rar> 下载。如有问题,可发信给 cat0247@163.com 或 lmz829@126.com。

本书编写得到了全国高等中医药教材建设研究会和各高等中医药院校的大力支持,在此谨致衷心感谢。

本书吸收了众多专家学者的宝贵经验,凝聚了高等中医药院校统计教师的智慧,为方便教学和应用,以尽量少的篇幅容纳尽量多的实用知识,使读者增长知识和才干,作者呕心沥血,每一章节的内容及编排都饱含编著者的心血和辛劳。限于我们的学识和能力,虽然我们力图提高质量,书中难免有错漏和不妥之处,希望同行专家和广大读者提出宝贵意见,以便修订提高。

刘明芝 周仁郁

2006年8月

目 录

第一章 绪论	(1)
第一节 中医药统计学的意义和内容	(1)
第二节 统计学的几个概念	(1)
第三节 医药统计中的变量及资料类型	(3)
第四节 统计工作的基本步骤	(5)
第五节 关于统计软件的说明	(6)
第二章 实验设计	(12)
第一节 实验研究的三要素	(12)
第二节 误差知识	(15)
第三节 实验设计的基本原则	(17)
第四节 常用实验设计方法	(19)
第三章 定量资料的统计描述、参考值范围	(29)
第一节 频数分布表和频数分布图	(29)
第二节 描述集中趋势的指标	(31)
第三节 描述离散趋势的指标	(33)
第四节 正态分布	(35)
第五节 变量变换	(39)
第六节 容许区间与参考值范围	(39)
第七节 离群值的取舍	(42)
第四章 总体均数的估计、假设检验	(44)
第一节 总体均数的估计	(44)
第二节 假设检验	(48)
第三节 正态性检验、方差齐性检验	(51)
第四节 t 检验	(55)
第五节 假设检验应注意的问题	(64)
第六节 $P > \alpha$ 时检验效能 $1 - \beta$ 的估算	(65)
第七节 置信区间与假设检验的关系	(66)
第五章 方差分析	(68)
第一节 方差分析的基本思想	(68)
第二节 完全随机设计资料的方差分析、多重比较	(69)

第三节	配伍组设计资料的方差分析、多重比较	(72)
第四节	拉丁方设计资料的方差分析	(76)
第五节	析因设计资料的方差分析	(77)
第六节	交叉设计资料的方差分析	(78)
第七节	组内分组资料的方差分析	(79)
第八节	定量反应资料样本含量的估计	(80)
第六章	双变量相关与回归	(85)
第一节	直线相关	(85)
第二节	等级相关(秩相关)	(88)
第三节	直线回归	(90)
第四节	直线回归与相关的区别和联系	(95)
第五节	回归分析的应用	(96)
第六节	曲线回归	(97)
第七节	概率单位回归法计算半数致死量 LD_{50}	(99)
第八节	联合用药效应判断	(101)
第九节	协方差分析	(102)
第七章	两分类资料的统计描述与推断	(106)
第一节	常用相对数、动态数列	(106)
第二节	概率的运算法则	(109)
第三节	二项分布和 Poisson 分布	(112)
第四节	总体率的置信区间	(115)
第五节	样本率与总体率的比较	(116)
第六节	两样本比较的 z 检验	(117)
第七节	标准化法	(119)
第八章	$R \times C$ 表资料的分析	(121)
第一节	四格表的 Pearson χ^2 检验	(121)
第二节	四格表确切概率法	(125)
第三节	$R \times C$ 表资料的统计分析	(126)
第四节	McNemar 检验和 Kappa 检验	(130)
第五节	多样本率的两两比较	(133)
第六节	同一样本中两个或多个构成比的比较	(134)
第七节	计数资料比较时样本含量估计	(135)
第九章	非参数检验、Ridit 分析	(139)
第一节	非参数检验的意义	(139)
第二节	Wilcoxon 符号秩和检验(配对及单样本比较用)	(139)

第三节	Wilcoxon 检验和 Mann-Whitney U 检验(两组比较用)	(141)
第四节	Kruskal-Wallis H 检验(多组比较用)	(144)
第五节	Friedman M 检验、Cochran's Q 检验(配伍资料比较用)	(146)
第六节	Ridit 分析	(149)
第七节	游程检验	(154)
第十章	圆分布资料的分析	(156)
第一节	圆分布资料及其特点	(156)
第二节	圆分布资料的参数估计和假设检验	(157)
第三节	圆分布资料的相关性分析	(166)
第十一章	诊断试验的评价	(170)
第一节	诊断试验准确度的常用指标	(170)
第二节	诊断试验的 ROC 分析	(173)
第十二章	序贯设计及其试验分析	(178)
第一节	序贯试验的特点、分类和适用范围	(178)
第二节	经典序贯试验	(180)
第三节	成组序贯试验	(188)
第十三章	新药临床试验	(193)
第一节	临床试验设计	(193)
第二节	新药临床试验	(194)
第三节	等效性和非劣效性检验	(197)
第四节	新药临床研究的病例数	(200)
第十四章	调查设计与调查质量评价	(203)
第一节	调查设计	(203)
第二节	调查质量评价	(209)
第十五章	横断面研究、病例对照研究、队列研究	(216)
第一节	横断面研究	(216)
第二节	病例对照研究	(217)
第三节	队列研究	(223)
第十六章	Hotelling T^2 检验、多元方差分析、重复测量资料方差分析	(227)
第一节	多元计量资料的常用统计量	(227)
第二节	Hotelling T^2 检验和多元方差分析	(228)
第三节	重复测量资料的方差分析	(233)
第十七章	多重线性回归及多元相关分析	(238)
第一节	多重线性回归	(238)
第二节	多元相关分析	(244)
第十八章	logistic 回归与对数线性模型	(247)
第一节	两分类 logistic 回归	(247)

第二节	多分类 logistic 回归	(257)
第三节	条件 logistic 回归	(262)
第四节	对数线性模型	(263)
第十九章	聚类分析和判别分析	(268)
第一节	聚类分析	(268)
第二节	判别分析	(273)
第二十章	主成分分析与因子分析	(280)
第一节	主成分分析	(280)
第二节	因子分析	(284)
第二十一章	生存分析	(288)
第一节	生存分析的基本概念和设计要求	(288)
第二节	生存分析的常用单因素非参数方法	(289)
第三节	Cox 回归	(296)
第二十二章	正交设计、均匀设计、混料设计及其试验分析	(301)
第一节	正交设计及试验结果分析	(301)
第二节	均匀设计及其试验结果分析	(312)
第三节	混料设计及其试验结果分析	(315)
第二十三章	典型相关分析	(322)
第二十四章	Meta 分析	(330)
第二十五章	模糊综合评判	(341)
第一节	综合评判的正问题	(341)
第二节	综合评判的逆问题	(350)
第二十六章	统计方法的选择、统计表和统计图	(354)
第一节	统计分析方法的选择	(354)
第二节	统计设计方法的选择	(359)
第三节	统计表和统计图	(360)
第二十七章	SAS 统计软件的应用	(364)
第一节	实验设计	(364)
第二节	计量资料分析	(369)
第三节	分类资料分析	(376)
第四节	秩和检验、Ridit 分析	(378)
第五节	相关与回归	(380)
第六节	多元分析	(387)
第七节	生存分析	(390)
第八节	正交、均匀、混料设计试验结果分析	(391)
第二十八章	DPS 软件的基础应用	(396)

第一节 计量资料的分析	(396)
第二节 资料的分析	(401)
第三节 秩和检验、Ridit 分析	(403)
第四节 线性相关与回归、协方差分析	(406)
练习题及参考答案	(410)
附 统计用表	(430)

第一章

绪 论

第一节 中医药统计学的意义和内容

中医药统计学(Statistics for Traditional Chinese Medicine)是将数理统计学的原理和方法应用于生物医药特别是中医药科研,搜集、整理和分析资料,推断不确定现象客观数量规律的一门应用学科。

中医药统计学的主要内容包括统计学基本理论和统计分析方法,统计分析方法包括统计设计、统计描述和统计推断、因素间的关系、生存分析、多元分析等。中医药统计学与用统计学原理和方法研究生物医学现象的医学统计学(Medical Statistics)内容兼容,更体现中医药特色。

中医药统计学给医药研究和探索提供了数量方法,可以帮助医药工作者周密地进行统计学设计,以正确的方式收集可靠的数据,去粗取精,描述资料的统计特征,去伪存真,推断事物内在的联系和规律,对分析的结果合理解释,作出科学的专业结论;在医药研究过程中,保证分组具有可比性,样本含量合适,研究结果适用于观察样本以外的同类对象;阅读中医药文献时,能分析中医药文献的科学价值,正确理解文献中统计处理的意义,推断文献的科学性和可信性;阅读国内外医药学术书刊时,能判断别人研究结果可信程度和局限性,吸收和借鉴其经验。

我国《药品注册管理办法》规定,新药临床试验必须自始至终有统计学人员参与;生物医药实验室研究、临床研究、流行病学探索和医药公共事业管理都要寻求统计学家的帮助;许多医药杂志都邀请统计学家审稿等,都说明统计学思维和方法已经渗透到医药研究和管理决策之中。中医药统计学是现代医药科研定量分析的重要工具,是认识医药数量规律的必备知识。

第二节 统计学的几个概念

1. 总体与样本 总体是根据研究目的确定的性质相同的研究对象的全体(population)。总体的基本单位(case)是个体(individual)。根据不同的研究目的,个体可以是一个人、一个家庭、一个地区、一只大白鼠、一个检测样品、一个采样点等。没有个体,总体就不存在;没有总体,也就无法确定个体。总体又有有限总体和无限总体,有明确范围(如空间、时间)限制,个体数有限的总体称有限总体。不易划分确切范围,个体数无限的总体称为无限总体。

同一总体的个体之间存在差异,称为个体差异(individual variation)或变异(variation)。变异普遍存在,例如,同为正常发育的10岁男孩,各人身高有不同;又如,病情相同的患者服用

相同的药物,其疗效不尽相同。可见变异具有不确定性。总体所要求的同质是指同一总体中的个体大同小异。

总体的性质或特征由其各个个体而定,要研究总体的性质或特征,须对它的个体进行观测,由于有个体差异,最好是对每个个体都加以观察。但是,这样做要花费很大的人力和财力,而且有时是不允许的,例如带有破坏性的药品检验不可能逐一进行。统计学的方法是通过样本来推断总体。

样本(sample)是从总体中随机抽取的部分个体。样本中所包含的个体数目即样本例数,称为样本含量(sample size)或样本大小、样本量。从总体中抽取样本的过程称为抽样(sampling),抽样的目的是用样本推断总体,所以样本要有足够的样本含量及可靠性、代表性。样本含量应在统计设计时进行估计。样本的可靠性是指样本中每一个观察单位是否确属于既定的同质总体。同质应考虑时间、空间、非实验因素的干扰,诊断标准等条件尽量相同。样本的代表性是指样本是否能充分反映总体的真实情况。为得到有代表性的样本,要求抽样和分配必须遵循随机化(randomization)的原则,保证总体中的每个个体都有同等的机会被抽到和分配,不受研究者的主观意愿或客观无意识的偏性影响。

2. 参数与统计量、抽样误差 刻画总体特征的指标称为总体参数(parameter),参数一般用小写的希腊字母表示,如用 μ 表示总体均数, σ 表示总体标准差, ρ 表示总体相关系数等。总体参数是常数。

在抽到样本含量为 n 的一个样本以后,样本中 n 个个体的观察值是 n 个具体的数值。通过对样本数据进行统计分析所产生的样本统计指标称为统计量(statistical variable)。统计量是不包含任何未知参数的样本观测值的函数,一般用英文字母表示。样本统计量不仅可用来估计总体参数,而且是统计推断的基础。

由于总体中的个体存在差异,在抽样过程中,即使从同一总体中随机抽取含量相等的若干样本,由不同样本计算的同一样本统计量(如样本均数)会有所不同,这些样本统计量间的差异同时反映了样本统计量与相应总体参数的差异,由随机抽样造成的样本统计量与相应总体参数的差异称为抽样误差(sampling error)。例如,从某一总体中,随机抽取样本含量为 n 的样本,算得样本平均数 \bar{x} ,用来估计总体平均数 μ ,但 μ 与 \bar{x} 有差异,这种差异是由于抽样误差造成的。因为个体差异客观存在,所以抽样误差是不可避免的。

3. 概率与频率 对某项变异进行观察或测量得到的指标称为变量(variable)。变量的观察结果称为变量值(value of variable)或观察值(observed value)。由观察值构成资料(data)。统计学中研究的变量是随机变量(random variable),具有如下特性:①在一次观察下,取值具有不确定性。②在大量重复试验中,变量的各种可能取值发生的可能性具有一定的统计规律。

概率(probability)和频率(frequency)都是度量某一事件发生的可能性大小的数值。

若事件 A 在 n 次独立重复试验中发生了 m 次,则称 m 为频数(counts),称比值 m/n 为事件 A 在 n 次试验中出现的频率或相对频数(relative frequency)。医药工作中通常所说的有效率、治愈率等都是频率。显然, $0 \leq m/n \leq 1$ 。频率是统计量,反映的是样本特征,刻画样本中事件出现的可能性大小,样本不同,其值也不同,具有偶然性。

随着试验次数增多,频率呈现出稳定在一个确定常数的特征。频率的稳定性说明随机事

件出现的可能性是事物本身固有的一种客观属性。

概率常用 P 表示。概率的统计定义:频率所稳定的常数称为概率。概率刻画的是总体中随机事件出现的可能性大小,一个随机变量 A 发生的概率是一个参数,具有必然性,是客观存在的。当试验次数 n 够大时,可用频率作为概率的近似值。

概率 P 值介于 0 与 1 之间。必然事件的概率为 1,不可能事件的概率为 0。概率越接近 1,表明其发生的可能性越大,概率越接近 0,则其发生的可能性越小。

4. 概率分布与抽样分布 任何一个随机变量各种可能取值发生的概率之和恒为 1。一个随机变量各种可能取值发生的概率所呈现的统计规律,称为该随机变量的概率分布(probability distribution),简称分布(distribution)。概率分布由随机变量各可能的取值与其对应的概率共同构造,完整地刻画随机变量的统计规律性。

总体中有很多个体,究竟哪些个体在抽样中被抽到,要依机会而定,因此,样本是随机变量。因统计量是依赖于样本的,故统计量也是随机变量,所以就有一定的概率分布,统计量的概率分布又称为抽样分布。常用的抽样分布有样本均数的分布、 χ^2 分布、 t 分布、 F 分布。在统计推断中,抽样分布充分发挥着主要作用。统计学研究的基本方法是通过样本构造统计量,再通过抽样分布的规律,对样本来自的总体进行分析和推断。

5. 统计描述与统计推断 统计描述(statistical description)和统计推断(statistical inference)是统计分析(statistical analysis)的主要内容。

统计描述是对原始资料的一种概括,即把分散而不好理解的原始数据,通过统计指标、统计图、统计表等方法来描述资料的特征(分布的形状、集中趋势、离散趋势)。

统计推断包括参数估计(parameter estimation)与假设检验(hypothesis testing),是在考察和表明抽样误差的前提下,推断总体特征,为研究者下专业结论时提供科学的依据。

统计描述和统计推断两者常需结合使用。

第三节 医药统计中的变量及资料类型

一、医药统计中的变量类型

变量的分类方法很多,按取值的结果可分为离散(discrete)型和连续(continuous)型两类。

离散型变量取值为有限个或可列无穷多个。例如采用某新疗法对 10 名病人进行试治,其治愈、有效、无效的人数为可列整数,可能取值为 0, 1, ..., 10 中的一个数,具有不连续性,为离散型变量。

连续型变量的可取值为某一区间的任何数值。例如,从同一批号规格为每袋重 12g 的玉叶解毒冲剂中任抽一袋称重,其重量可能是 12.1g、11.9g、...; 7 岁男孩的身高可能是 120.0cm,也可能是 119.5cm...等,这类变量的可取值充满某一个区间,具有连续性。

原始资料中的变量,按在统计分析中所起的作用可分为标识变量(identifiers variable)和分析变量(construe variable)。标识变量如“编号”、“住院号”,主要用于数据管理,包括数据的核对与增删等,是研究记录中的重要内容,但不是随机变量。分析变量又分为自变量(independ-

ent variable) 和应变量(response variable), 是数据分析的主要内容。应变量也称因变量, 即反应变量, 是研究的观测指标, 其取值受自变量的影响而具有随机性。自变量是影响应变量的变量, 常为试验因素(即处理因素、干预因素)、分组变量(between groups variables)等。与应变量有关的影响因素称为协变量(covariate), 属于自变量。

变量按观察指标的性质分类如下:

- 定量变量, 亦称计量变量、数值变量或尺度变量(用等距尺度或比例尺度度量的变量)
- 分类变量 {
 - 无序分类变量, 亦称名义变量(用名义尺度度量的变量)
 - 有序分类变量, 亦称等级变量、顺序度量(用顺序尺度度量的变量)

二、医药统计中的资料类型

对应于变量按观察指标性质的分类, 医药统计资料的类型见表 1-1。

表 1-1 医药统计资料的类型及其特点

资料类型	特点	变量类型	实例
定量资料 (数值资料)	每个个体都能观察到一个观察指标的数值, 有度量衡单位	连续型或离散型 数值变量	身高(cm), 体重(kg), 细胞数(个), 人口数(人)
分类资料 (包括无序分类和有序分类资料)	每个个体的观察结果是属性或分类, 不是数值, 清点各属性或分类中个体的数目, 才得出分类资料的数据	离散型	
无序分类资料 (计数资料)	每个个体观察结果的属性或分类间无大小顺序之分。二分类时为对立的两类属性, 多分类时为不相容的多类属性, 类与类之间界限清楚, 不会错判	离散型 无序分类变量	性别(男, 女)为二分类; 血型(A、B、O、AB)为多分类
有序分类资料 (等级资料)	多分类, 按程度、档次等级顺序分别计数, 各等级之间有程度的差别, 按大小顺序排列。每个观察单位无确切定量	离散型 有序分类变量	疗效(痊愈, 显效, 有效, 无效)

1. 定量资料 又称计量资料(measurement data)、数值资料(numerical data)或尺度资料(scale data), 是对观察对象测量指标的数值大小所得的资料, 观察指标是定量的, 表现为数值大小。

2. 分类资料 将研究因素按属性类别或档次分组, 清点各组个体的个数(频数), 将分类变量的分组标志及各组频数同时给出, 便得到了分类资料(categorical data)。分类资料有无序分类资料和有序分类资料两种:

(1) 无序分类资料(unordered categorical data): 又称计数资料(enumeration data)。无序分类资料的属性或类型互斥, 无次序之分。根据属性或类型分组的多少又可分为两种: ① 两分类资料: 它是将观察单位按两种对立的属性分类, 如按死亡和生存、治愈和未愈、有效和无效等两种分类计数。② 多分类资料: 它是将观察单位按多种互斥的属性分类, 清点各类观察单位的个数, 如血型按 A、B、AB、O 四种分类计数。

(2) 有序分类资料(ordered categorical data): 将观察单位按某种属性的不同程度、档次或

等级顺序分类或分组后,清点各顺序分组的观察单位数所得的资料为有序分类资料,亦称等级资料(ranked data)。例如,化验结果按-,+,++,+++分组计数等,因有档次或等级的差别,可以用 $X=0,1,2,3$ 的等级区别表示,等级之间的差别并非确切定量差别,可以是量的差别,也可以是质的差别,给人以“半定量”的概念,故又称半计量资料。

根据分析研究的需要,定量资料可以转化为分类资料,定量资料和有序分类资料可以互相转化,例如,白细胞计数属于定量资料;如果按白细胞总数正常与不正常分为两组,然后清点各组人数,便成为两分类资料;如果按白细胞总数过高、正常、低下分为三组,然后清点各组人数,就成为有序分类资料;将有序资料的各个类别分别赋予一个合适的数量,可转化为定量资料。须指出的是:有序分类资料转化为定量资料,存在如何变换才合理的问题,例如,不咳嗽,轻度咳嗽,中度咳嗽,重度咳嗽的4个等级,可以分别赋予0、1、2、3的得分,也可以分别赋予0、1、3、6的得分。研究者可依据自身的经验判断作出不同的变换,也可采用等级(秩)得分或 Ritdit 值(见第九章)的方法赋值。

医药统计中的资料除按研究因素的性质分为定量和分类资料之外,对呈周期性变化规律或表示在圆周上位置特点的资料,称为圆分布资料;随访慢性病例或保健对象远期效果的资料,常有未达到预定观察终点的不完全数据特点,称为随访资料;等等。

对任何一份资料进行统计分析之前,首先要分辨其资料类型和数据特点,因为不同类型的资料,统计分析方法不同。

第四节 统计工作的基本步骤

统计工作分为设计,搜集资料,整理资料,分析资料四个基本步骤:

1. 设计 设计(design)就是制定完成科研目标的技术路线和实施方案,对包括研究目的,观察指标,资料搜集、整理、分析全过程总的安排,是最关键的一步。科研设计包括专业设计与统计设计(statistical design),专业设计是针对专业问题所进行的研究设计,统计设计是针对统计数据收集、整理和分析所进行的设计。两种设计的对比见表1-2。

表 1-2 专业设计与统计设计对比

	专业设计	统计设计
依据	运用现代科学技术与医药专业知识进行设计	运用数理统计的原理和方法进行设计
目的	主要解决调查或实验观察结果的创新性、先进性和应用性	主要解决抽样误差的问题,保证样本的代表性和组间的可比性,实现调查或实验观察结果的可靠性、科学性、可重复性和经济性
意义	决定学术水平的高低及科研成果的大小	有效地控制误差和提高检验效率
内容	①选择研究题目;②明确观察指标及研究方法;③确定诊断标准及观察对象	①对照形式;②选择调查或实验的设计方案;③随机抽样和均衡分配;④样本含量及抽样误差大小的估计;⑤盲法

统计设计和专业设计相辅相成,缺一不可。专业是基础,专业设计方法因课题而异,是科研创造性与学术水平的决定因素。统计设计是保证专业设计布局合理性和科研结果可靠性的

关键。医药论文中的统计学处理应交代统计研究设计的名称和主要做法。

统计设计的内容和方法详见第二章、第十三章、第十四章、第十五章。

2. 搜集资料 搜集资料即取得准确可靠的原始资料,是统计分析的前提与基础,收集必须严格按科研设计方案,及时、准确、完整、清楚、统一收集,要注意防止误差及偏倚影响观察结果的精确性,确保统计资料的科学性和可靠性。例如随机抽样,要注意审查样本是否从同质总体中随机抽取,样本含量是否适当,资料是否完整,数据是否准确,是否是同时期的资料,诊断标准和疗效判断标准要有科学依据并要求统一,避免研究者的主观偏见。

3. 整理资料 整理资料包括:①对原始数据进行逻辑和计算审核。②设计分组,包括质量分组(如病人按性别、职业、病情轻重分组)和数量分组(如按年龄、体重、身高分组)两种。分组的原则是合理,具有同质性,先细后粗,先分后合。③拟订整理表。整理表是统计表的前身,是过渡性表格,作归纳汇总用。宜尽量把整理表设计成复合表及相关表,以免漏掉两指标间存在的相关信息。④编制统计表。

表 1-3 某病感染按性别、年龄整理表(问题表)

	性别		年龄(岁)				合计
	男	女	0~	5~	10~	15~	
调查人数							
感染人数							

表 1-4 某病感染按性别、年龄整理表(正确表)

年龄组	男		女	
	调查人数	感染人数	调查人数	感染人数
0~				
5~				
10~				
15~				
合计				

整理资料存在的问题主要表现在:①编制整理表时,未将有联系的项目安排适当,不便说明事物之间相互的规律性。如表 1-3,由于将性别与年龄这两个有联系的项目分割开来,计算不出不同性别、年龄的感染率,丧失了有价值的信息,为了克服上述缺点,应采用表 1-4 的整理表。②分组不合理;③有意将过高或过低数据筛选掉或无意舍弃掉,使原始资料所提供的信息变样;④统计表编制不规范(例见第二十六章表 26-6)。⑤计算错误。

4. 分析资料 分析资料是指进行统计描述和统计推断。常见的问题有:抽样研究未作统计检验,只凭表面数据大小即下结论;统计分析方法选择不当或统计方法交代不清;对有关联事物的分析提不出数据,找不出规律;将偶然因素当作因果分析;忽视对比条件,未作标准化处理;将构成比当作率分析;发病率与患病率相混淆,死亡率与病死率相混淆等。

统计工作的四个基本步骤是互相联系,不可分割的,任何步骤的缺陷都将影响统计工作的效果。

强调指出:统计工作必须持严肃认真、实事求是的科学态度,反对伪造和篡改统计数据!

第五节 关于统计软件的说明

一、关于统计软件的使用

统计软件能迅速完成统计的复杂计算。一般说来,使用统计软件,需要掌握“分析方法、数据格式、读取结论”三个方面的基本技能。只学会统计软件的操作方法,并不见得就会进行