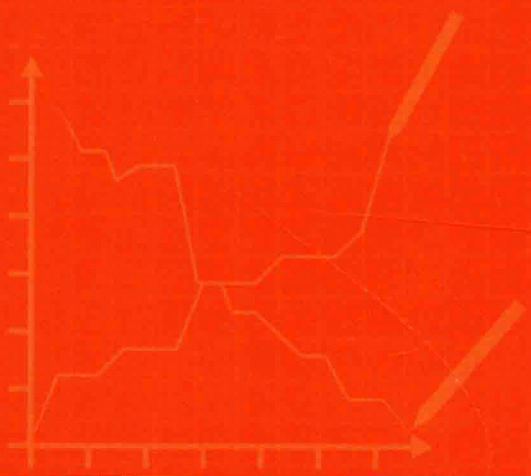
普通高等教育“十一五”国家级规划教材

Medical Statistics

医学统计学

仇丽霞 主编



第3版

中国协和医科大学出版社

医学统计学

（第3版）

主　审　刘桂芬

主　编　仇丽霞

副主编　刘玉秀　罗天娥

编者（按姓氏笔画排）：

尹　平　华中科技大学
王立芹　河北医科大学
仇丽霞　山西医科大学
艾自胜　同济大学
刘　艳　哈尔滨医科大学
刘玉秀　东部战区总医院
刘桂芬　山西医科大学
师先锋　山西医科大学汾阳学院
李济宾　中山大学肿瘤防治中心
李新华　贵州医科大学
张　燕　重庆医科大学
张业武　中国CDC信息中心
张丽荣　贵州医科大学
张岩波　山西医科大学
张彦琦　陆军军医大学
张晋昕　中山大学
陈长生　空军军医大学
陈平雁　南方医科大学
易　东　陆军军医大学
周　琴　复旦大学
罗天娥　山西医科大学
赵晋芳　山西医科大学
徐　涛　北京协和医学院
鄢艳晖　广东药科大学
郭　静　中国人民大学
黄高明　广西医科大学
萨　建　山西医科大学
曹文君　长治医学院
曹红艳　山西医科大学
阎玉霞　南方医科大学

秘书：萨　建　山西医科大学

中国协和医科大学出版社

图书在版编目（CIP）数据

医学统计学 / 仇丽霞主编. —3 版. —北京：中国协和医科大学出版社，2018.7
ISBN 978-7-5679-1038-6

Ⅰ. ①医… Ⅱ. ①仇… Ⅲ. ①医学统计 Ⅳ. ①R195.1

中国版本图书馆 CIP 数据核字（2018）第 057531 号

医学统计学（第 3 版）

主　　编：仇丽霞
责任编辑：吴桂梅

出版发行：**中国协和医科大学出版社**
（北京东单三条九号　邮编 100730　电话 65260431）
网　　址：www.pumcp.com
经　　销：新华书店总店北京发行所
印　　刷：北京新华印刷有限公司

开　　本：787×1092　1/16 开
印　　张：39.75
字　　数：700 千字
版　　次：2018 年 7 月第 3 版
印　　次：2018 年 7 月第 1 次印刷
定　　价：58.00 元

ISBN 978-7-5679-1038-6

再版前言

在信息急剧增长的背景下，如何利用统计学方法和计算机技术对健康医疗大数据进行合理的利用与挖掘，揭示海量医学知识中蕴藏的规律，如何制订科学研究计划、准确获取数据、精准地提炼信息并做好专业知识的客观解释，就成为本教材编写的宗旨与中心任务。

本教材根据数据分析、处理、识别和预测对医学生的基本要求，在《医学统计学》第2版国家级“十一五”规划教材内容优化组合基础上，再次组织21所高校及研究单位编写人员，结合医学改革与实践需求，进行了内容的更新、补充与修订。紧扣三基训练，重视学科知识结构，结合软件应用，强化统计分析方法的选择，特别注重结果的专业解释；深入浅出，融会贯通，是医学科研实践工作中非常实用的一本参考书。

全书共25章，主要分为四个部分。第一部分（第1~9章和第25章）主要介绍医学统计的基本理论与方法；第二部分（第10~14章和第23章）充实、完善增写了计数数据和重复测量数据的模型分析；第三部分（第15~18章）进一步强化了大数据背景下的研究设计理念，增加了量表设计等内容；第四部分（第19~22章及第24章）深入讨论了健康信息与医疗服务数据分析与应用。

教材（第3版）能再次出版，要衷心感谢广大读者多年来的厚爱与好评，衷心感谢为教材整体规划付出心血的刘桂芬教授，为教材编写履行传帮带的一、二版教材全体编委，特别感谢2版教材吕桦、凌莉、于浩、田考聪、韩少梅、金水高、余红梅、肖琳、吴艳乔、宿庄编委的传承，感谢为教材编审默默奉献的专家教授，为教材出版、联络、编辑和修改的老师与出版社同仁，为教材复核、校对的博士研究生饶华祥、硕士研究生潘金花、李文瀚、郭强、于智凯、任浩、张壮。值此出版之际，谨致以最真挚的谢意！

鉴于我们编写能力有限，教材中难免存有纰漏之处，诚望同仁与广大读者批评斧正。

仇丽霞

2018年1月于山西医科大学

目 录

第一章 绪 论

重点掌握：

1. 医学统计工作的基本步骤。
2. 常见的医学统计资料的类型。
3. 医学统计学常用的几个基本概念：总体与样本、随机误差、概率与频率。

医学统计学（medical statistics）是描述、归纳、探索医学数据分布特征和解释数据规律的一门学科，是科研工作者运用概率论与数理统计原理，进行数据的获取、存储及管理和分析，评价人类健康水平，探索疾病发生与发展规律，进行预测评价的方法，是循证实践中数据挖掘不可或缺且起关键作用的一种技术手段。统计学已不仅仅是对数据进行观察、测量、记录和归纳，更重要的是利用统计方法对研究事物做出科学合理的决策与推断，以帮助我们更好地认识和掌握群体及个体健康变化的规律。

第一节 医学统计工作的作用及基本步骤

众所周知，事物的量化有助于人们提高对事物认识的准确性和深度。科学研究是通过对数据的量化分析，探索未知世界的一种认识活动，与统计有着不解之缘。医学研究的主要对象是人，人是世界上最复杂的生命体，不但具有生物性，还具有社会性；不但有生理活动，还有心理活动，个体变异错综复杂。医学统计学不仅是人类健康与疾病数据分析必备的科学方法，而且其作用横跨健康与疾病研究的各个学科，纵贯生命研究的全过程。其工作步骤大体分为：

一、医学研究设计

研究设计（design）是医学统计工作过程的一个重要内容，它是医学科研工作的第一步，是对医学科学研究过程、内容及具体实施方法和步骤的总设想或安排。设计就是针对具体的研究目的或问题，确定研究对象和观察单位（个体），明确分析指标。根据是否施加干预因素，如何获取数据，怎样进行资料的清理和分析，如何控制误差，预期分析结果有哪些等提出详尽的实施方案和技术路线，做好周密的考虑和安排。其内容分为：①专业设计：它主要反映研究者对专业知识掌握的能力和程度，与科研课题或项目的深度及水平有关；②统计设计：反映研究者对统计知识、技术正确应用的程度和科学研究的能力，主要与科研工作的质量有关。怎样才能以较少的人力、物力和财力获取准确、可靠的科学结论，搞好研究设计是最关键的一个内容，也是探索新知识、验证新理论、推广新方法，甚至利用大数据做出

新决策等必不可少的手段。

二、搜集资料

搜集资料（collection of data）是统计工作的基础，它直接影响科研工作的质量。其任务是研究者按研究设计要求，获取准确、可靠、有价值的数据，并做好数据质量管理与评估。若所获取的数据不准确、不完善，其结论有可能是悖论，失去研究的意义与价值。数据的准确性是指观察、测量、记录、储存、转移或计算的数据，均无虚假差错之处，尽可能做到界限明确、真实、可靠，不造成混淆。数据的完整性是指用来研究分析的项目没有遗漏、重复和缺失。数据搜集的及时性是指资料在一定条件下，按规定的时间完成采集、储存和管理等。

健康与疾病及相关数据的来源主要有：

1. 常规保存的工作记录　如人口健康信息、出勤记录、职工流动、工伤、出生、死亡登记等，职业病报表、恶性肿瘤报告卡、气象数据、环境 PM 2.5 等监测数据等，它是群体健康状况研究的一个重要来源。但由于没有严密的设计，有时会给分析带来诸多不便；量大、来源广泛、结构复杂、漏报、重复和缺失是最常见的问题，除资料搜集过程中应加强专业性较高的技术督查外，尚应注重标准化数据的循序积累。

2. 住院和家庭医疗活动适时积累的健康管理大数据　如呈爆炸性增长的、源自不同地区个体诊所、社区、医院、保健、康复等医疗机构的预约挂号信息、电子病历、医药、X 线、磁共振成像及 CT 等影像记录、日常收费财务运营，生物医药、基因测序、安全监控和不良反应监测等资料，可根据研究目的与结构化、非结构化和无结构化数据类型，进行专业大数据的整合、分析和利用。

3. 卫生健康服务信息系统　主要指由卫生健康机构根据国家有关部门规定统一管理，或监测哨点逐级上报的内容。如法定传染病报告系统、食品安全风险监测系统、食品安全国家标准跟踪评价及意见反馈平台、地方病、寄生虫病及职业病防控、精神卫生疾病管理、计划免疫、慢性病综合防控、死因和出生缺陷监测，各种健康生命体征监测、社区居民健康档案和医院工作报表等。它可为了解居民健康状况、拟订卫生健康服务规划、合理配置医疗卫生资源等提供科学依据。

4. 专项调查与实验研究资料　专项调查或实验研究一般指为解决某个（些）问题或验证某个（些）假说等所进行的专门研究。如全国 7~10 岁儿童龋患率现场调查、某地 2 型糖尿病调查、全国膳食营养水平调查、某地中学生视力状况研究、联合用药对难治性类风湿关节炎疗效的多中心临床研究等。

5. 公共共享资料　指为研究工作需要取自公开发布的报告、专业参考文献、基因数据库、商业数据库、人口、公安、保险等与人类健康相关的数据资料等。

三、清理资料

清理资料（cleaning data）是按设计要求将一些分散的、表现个体特征的原始数据系统化、条理化，以便更好地揭示所研究事物的内在规律。它往往结合网络平台规划、数据采集

策略、存储数据督查、研究项目要求、分析模型方法和具体研究过程而进行。应考虑：

1. 资料核查 除获取数据时调查员自查、调查员间互查和专业人员核查外，清理资料尚应对储存、抽取、转移和加载的全部裸数据进行逻辑检查和数字核准等。包括对原始调查项目的审核，缺失数据的检查，误填、漏填项目的核准、修正，数据类型以及编码等问题的考虑，它是保证和提高数据分析质量的前提。

2. 数据类型与特点 从观察或实验研究获取的任何结果，都必须结合专业知识转变为数据后才能进行统计分析。无论是字符型、还是数值型变量，均可用二维结构矩阵来表述。一般情况下，行表示观察单位，列表示分析指标或变量。Excel、SAS、SPSS、R 等软件均可以此形式作为数据分析的基本格式。其中观察单位（observation unit）亦称个体（subject），随研究目的而异，可以是一个人、一个组群、一份样品、一个采样点、一毫升水或一个病室等。对观察单位的一次测量，可获得一个记录，也可对同一个体进行多次观测，获取几个测量记录，但每个观察单位所得数据，一般放在同一行上。主要研究指标可以是研究项目、混杂因素，也可以是研究对象的基本特征；可以是分析变量，也可以是分组变量或协变量等。它们可由测量结果直接录入，也可以从卫生健康大数据转移生成新的数据库。

3. 数据编码 数据汇总时，应由专业人员根据专业要求进行编码。编码技术包括：①设计编码：如调查问卷设计中数值型变量值的位数、取值范围控制，不详数据的编码，定性数据的数量化，连续变量的离散化等；②过录与检查编码：无论是数据采集、储存、转移还是分析，专业技术人员均应在搜集到数据后进行项目内容、编码和数据核准检查，并保存到确定的位置上选择两名以上责任心强的数据管理人员，进行双份一致性检查，对发现的错误进行纠正并记录。

4. 设计分组 一般应据研究目的、数据结构、指标类型、观察单位数目、专业与学科习惯用法及其指标间的关联性等来考虑分组。大体分为质量分组和数量分组。质量分组即按研究事物的类别属性特征进行归组，如性别、疗法。数量分组即按研究指标观测值数量大小来归组，详见第 2 章第 1 节。

5. 预分析 根据不同的应用软件，建立数据库时应注意的问题略有不同，但都可进行变量分布、数据特征描述、探索性研究等预分析，以利于更好地揭示研究事物内部的共性和对比组之间的差异性，也是分层分组分析、选择统计方法与指标、客观解释研究结果的重要线索。

四、分析资料

分析资料（analysis of data）亦称统计分析，包括统计描述、统计推断与统计预测。统计描述（statistical description）指按研究设计要求，从多维、多角度、按特定模式构建关系型数据库，实现数据可视化；计算反映事物特征的指标，并用适当的统计图表来概括，以阐明事物现象的水平及其内在联系。统计推断（statistical inference）指根据抽样原理，在概括随机样本特征及问题条件和模型假定的基础上，根据假设检验和估计概率大小对所研究总体或总体间的特征做出推断和解释，其理论和方法构成了数理统计学的主要内容。统计预测（statistical prediction）指在大数据整合基础上，运用统计方法探索研究事物变量之间的本质

联系，从而由某事物现象的过去和现在，对未来的发展变化趋势和方向做出判断。

统计工作的四步是紧密联系、环环相扣、不可分割的整体，任一阶段有缺陷都将造成一定的损失，甚至导致科学研究工作失败。

第二节 医学统计学中常用的几个基本概念

一、同质与变异

同质性（homogeneity）指研究事物现象存在的共性，它是统计研究的基础，是资料清理和分析的前提。任何源于事实的数据，皆应以组内尽可能相同或相近，对比组间具有均衡可比性为前提。

尽管在同质总体中，不同个体某指标的观测值间经常存在不确定性。这种同质群体中自然状态下个体值间千差万别、参差不齐的现象称为变异（variation）。变异是客观存在的，是绝对的，而同质是相对的。统计学就是处理不确定性（变异）的科学与艺术，在此基础上描述同一总体的同质性，揭示不同总体的异质性。

二、医学统计资料的类型

欲了解某年某市10岁健康男童身高水平，凡在该市居住两年以上，年龄满10周岁，排除了患有影响身高疾患的男童，均可作为本次研究的调查对象。每个男童就是一个观察单位。观察单位的研究特征称作变量（variable），变量的观察结果称为变量值或观测值。如本次研究中，身高是研究特征，对每个男童身高测量的结果称为身高变量值，简称身高值或变量值。对变量取值的过程就是测量，而取值所需的标准称为测量尺度，它是获取正确、稳定、一致测量结果的条件。

（一）常用的测量尺度

1. 名义尺度（nominal scale） 指变量的结果是按某事物属性分类来进行测量的，如性别变量：男、女；血型变量：A型、B型、O型、AB型，所用符号与属性一一对应，同一符号内各变量值的本质相同。

2. 顺序尺度（ordinal scale） 指变量值不但可以分类，而且各类之间有某种特征程度上的不同，可用数学上大于或小于来表达它们之间的关系，如疗效评价中的无效、好转、显效、痊愈；工作面污染程度的轻度、中度和重度等。显然，评价尺度可改变，但它们的顺序或等级不变。有时用序数或秩次1，2，…，R来表示，相邻序次间级差相等，但实际资料相邻类别间的数量级差却不一定相同，且难以精确量化。

3. 区间尺度（interval scale） 指用数量大小来度量某种特征，其变量值 X 可以是实数轴上的一个连续区间，任意两个取值之间可有无穷多个数值，表现为连续型随机变量，如身高（cm）、体重（kg）、血压（kPa）、呼吸次数（次/分）等。变量值 X 也可以是整数范围内的随机变量，如育龄妇女生育子女数、患龋齿数等。

4. 比数尺度（ratio scale） 指以比值、比率等来度量某种特征，如中性粒细胞占白细胞

数总数的百分比、体质指数、某指标治疗后占治疗前百分比等。

（二）变量分类

根据变量值是定量或定性的特征，将变量分为定量变量、分类变量。

1. 定量变量（quantitative variable） 观察单位的变量值是定量的，表现为数值的大小，一般有度量衡单位，可按区间和比数尺度测得。分为离散型（discrete）和连续型（continuous）两种：离散型定量变量是指测量值只取整数，如育龄妇女生育的子女数、患龋齿数等；若测量值是区间内任意值，则称为连续型定量变量，如身高、体重等。由一组同质的定量变量值所组成的资料也称为定量资料（quantitative data）。

2. 分类变量（categorical variable） 其变量值表现为不同的属性、特征或类别，分为无序分类和有序分类变量两种。若按名义尺度测量的属性特征归类，也称定性变量（qualitative variable）或无序分类变量。其中，若定性变量值表现为相互对立的两种类别，称为二项分类变量（binary variable），如性别；若定性变量值表现为互不相容的多个类别，称为多项分类变量，如血型。将按照事物属性特征分组，并清点各组观察单位数而得到的资料称为定性资料（qualitative data），如95例某患者中，50例男性、45例女性，其中A型18例、B型19例、O型35例、AB型23例的资料。若按顺序尺度测量的类别或程度归类，则称顺序变量（ordinal variable），或有序分类变量，也称等级变量，如尿红细胞检验结果为-、±、+、++、+++、++++。将以变量值的等级或程度分组，清点各组观察单位数而得到的资料，称为等级资料（ordinal data）或半定量资料，如360名职工眼底动脉硬化检查结果：正常326例、轻度硬化18例、中度硬化13例、重度硬化3例。半定量资料在类别或程度间既有分类的不同，也有量的差异。

实际问题分析中，研究目的不同，资料类型也可以转化。如某医师测得10名3岁儿童血红蛋白含量（g/L）结果如下：108、110、116、95、109、87、92、113、120、116，该资料为定量资料；若只考虑是否贫血，可按临床参考值整理为无贫血者≥110g/L 5例、贫血者<110g/L 5例，即转化为定性资料；欲分析贫血的严重程度，则整理为无贫血者≥110g/L 5例，轻度贫血者90~110g/L 4例，中度贫血者≤90g/L 1例，转化为等级资料。

三、总体与样本

总体（population）是根据研究目的所确定的同质的所有观察单位某种变量值的集合，如某地某年所有15岁健康男孩身高值。据总体中观察单位数（N）是否可知，分为有限总体和无限总体。有限总体指总体的观察单位数是有限的或可知的，如调查时点某地区的户籍人口数。无限总体是指总体中的观察单位数是无限的或不可知的，如空气中SO_2的浓度的观察单位。反映总体特征的指标称作参数（parameter）。医学研究中，划清总体的同质范围，确定研究对象是非常重要的。根据研究目的，从研究总体中随机抽取、对总体有代表性、反映总体特征的部分观察单位，称为样本（sample）。样本中的观察单位数称作样本例数（n）或样本量（sample size）。从研究总体中按一定的概率规则，抽取部分观察单位的研究方法，称作随机抽样。随机抽样（random sampling）不是随意选择（purpose selection），所谓随机是指研究总体的各观察单位按其在总体中的分布特征，被抽到样本中的机会均等且互不影

响。只有这样，才有可能保证抽到的样本有代表性，它是统计推断的基础，统计学中将反映样本特征的指标称作统计量（statistic），也称为参数估计值。

四、误差

任何研究总是期望对总体做出客观、可靠、真实的评价。但在实际工作中，调查结果可能会受到各种因素的干扰与影响而偏离真实情况，将实测值与真实值之差称为误差（error）。统计研究中，据其产生的原因分为随机误差和非随机误差两类。

1. 随机误差　包括抽样误差和随机测量误差等。①抽样误差指随机抽样研究中，由于抽样而引起的样本统计量与总体参数间的差异，其大小随样本不同而改变，它也是一个随机变量；②随机测量误差（random measurement error）指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。即使严密控制研究条件，但由于一些偶然因素或就目前医疗技术水平尚无法控制的因素，也可使实测值与真实值不同。实际工作中，测量误差也无法避免，但应控制在容许误差范围之内。

随机误差的出现是随机的，分布是有规律的，其值可正可负、可大可小，当观察次数足够大时，随机误差服从正态分布。统计推断就是依其分布规律由样本对总体做出推断的。

2. 非随机误差　指所得资料偏离研究的真实情况，致使推断、预测出现的偏差，包括系统误差和过失误差。系统误差（systematic error）或偏倚（bias）可产生于设计人员、调查者或调查对象；也可由于设计考虑不当，资料搜集不准，储存、转移、汇总、计算有误等造成。一般带有倾向性，如有恒向、恒量、周期性或有特定的变化规律。其产生原因复杂，贯穿于研究全过程并对研究结果有影响，又很难用统计方法评价其影响的大小，必须依靠科学的研究设计，正确的资料搜集，科学的数据管理，合理完善的分析计算，严谨的工作态度与作风，将其减小或控制在最小容许范围内。过失误差指由于工作人员责任心不强，检查核对制度不严，或故意修改等而造成的检查、记录、观察、录入数据错误等而产生。过失误差是错误，一般应杜绝。一旦发生应彻底纠正，否则有可能得出悖论。

五、频率与概率

频率（frequency）是指在相同条件下，进行有限 n 次重复试验，某随机事件 A 发生次数与 n 次试验的比值，其值介于 0～1 之间。如某地妇产医院 2015 年记录在册的出生人数 120 名，其中男婴 70 名，男婴占 58.33%，称作男婴的出生频率，频率随样本变化而改变。

概率（probability）是描述某随机事件 A 发生可能性大小的度量，常记作 P，可用小数或百分数表示。医学研究中，绝大多数属随机现象。如手术治疗甲状腺癌的疗效有治愈、好转、无效、死亡 4 种，但对于一个初诊的甲状腺癌患者而言，手术治疗后究竟会是哪种疗效结果，是一个不确定的随机事件，亦称偶然事件，简称事件。若将患者转归为“治愈”记作事件 A，则治愈的概率记为 P（A），简记为 P，这是一个很有意义的、研究者颇为关心的未知数值。当事件一定发生，则 $P=1$，称必然事件；当事件一定不发生，则 $P=0$，称不可能事件；当事件发生的可能性为 $0<P<1$ 时称随机事件，P 值越接近于 1，该事件发生的可能性就越大。统计研究中，不确定事件的结果均具有概率性，习惯上把 $P\leqslant 0.05$ 的随机事件

称作小概率事件，表明该随机事件发生的可能性很小，在一次试验或一次调查中，可认为该事件几乎不可能发生，而“小概率事件”是统计推断的重要依据。

虽然随机事件 A 在一次试验中可能出现，也可能不出现，但在多次重复试验中，它呈现出明显的规律性。假设在相同条件下，独立地进行 n 次重复试验，随 n 的不断增大，频率逐渐趋稳，可将稳定的频率作为事件 A 概率的估计值。但 n 较小时，频率波动较大，用之估计概率是不可靠的。

第三节 大数据背景下的医学统计学

随着互联网技术的飞速发展，健康与医疗实践步入大数据时代，医学科研面临新的机遇与挑战。医学大数据顾名思义就是在健康促进及医疗领域中，快速增长、数量极其庞大、获取、存储、管理和分析均大大超出了传统数据管理和分析软件能力范围的数据集合。它具有海量、多样、快速和易变、真实、复杂和价值的特征。大量性（volume）指数据总量能达10~99TB；多样性（variety）泛指结构化（structured）、半结构化（semi-structured）和非结构化（unstructured）等数据类型，其来源也具有多样化；伴随数据快速性（velocity）增长的特征，数据流还呈现不稳定的随日、季节或特定事件触发而周期性峰值波动的特征，也称之为易变性（variability）；真实性（veracity）亦称数据质量，这些数据都是日常工作的客观实际记录；复杂性（complexity）体现在数据的获取、管理、操作和分析中，如何从中提取、转换、加载、连接、揭示关联等，对数据管理和统计分析提出了新问题；价值（value）指合理运用大数据分析后揭示出来的信息与知识，它与前 5 个特性有本质的不同。若前 5 个特性被认为是数据工作者具体实践中面临的挑战，价值则是征服这些挑战后获得的回报。同时，也提示在大数据背景下，设置变量时要尽量保证测量值信息充分。

当今，互联网技术的飞速发展，为我们从不同角度、更细致、更全面地利用爆炸性增长的健康、医疗大数据提供了可能，同时也激发了医学科研工作者们进行真实世界研究（real world study，RWS）的欲望。数据的整合与分析告诉了我们什么？绝不仅仅局限于验证样本资料是否满足了所提出的某种假设，而是促使研究者利用大数据挖掘各种感兴趣的关联，并进而比较、分析、归纳，通过真实世界研究做出科学决策。现有的数据库软件和建立在假设基础上的统计方法，已远远不能满足目前健康与医疗大数据海量信息的采集、储存、管理与利用的要求。健康与医疗大数据对提高医疗质量、降低患者风险、节省医疗资源与成本、提高服务效率等发挥巨大的作用，同时对统计工具和算法等也提出了更高的要求。统计软件是管理数据、计算分析、模拟和实现统计过程的一类应用软件，是统计方法应用的重要载体，在健康与医疗数据分析中具有重要的地位，是统计学发展中不可割裂的一个重要组成部分。怎样才能从健康与医疗大数据中提取所需信息，帮助研究者合理运用统计分析方法与技术，并对研究结果给予恰如其分的评价和解释，进而做出科学决策，正是医学统计学所要解决的新问题。

大数据背景下统计分析大体归为：①可视化分析（analytic visualizations）：是数据分析最基本的要求，它可直观地展示数据的基本特征；②目标分析：根据样本统计量，考虑按设

定的风险对总体特征做出推断或者进行预测分析。其中，预测性分析（predictive analysis）是大数据分析方法中最有价值的一种，根据可视化和数据挖掘结果做出预测性判断，需要在深化数据内部算法、挖掘有价值信息的基础上，整合数据量较大且满足一定精度要求的相关历史资料，确定统计预测方法，建立预测模型，进行群体或个体短期、长期预测，并完成预测误差分析。

总之，医学统计学是医学研究领域非常活跃的一门学科。它以其独特的思维方式渗透到健康与医药研究及管理等各个领域，它既是一门专业基础理论课，也是健康与医疗大数据分析利用不可或缺的方法技能课。

针对大数据背景下统计分析应用中存在的问题，本教材拟回避繁杂的数学公式推导，侧重于加强统计学基本概念的理解，有机结合实际应用问题，以统计软件结果的解释作为学生重点掌握的内容。基础章节可结合课堂简单讲解公式与计算，更好地帮助研究者提升科研设计和数据有效利用及分析的能力。

作为新型医学人才，基于大数据背景，不仅需要掌握基本的医学统计学原理与方法，更应结合健康与医疗等专业知识，更好地利用有价值的数据信息，进而完善医学理论，做好医学研究设计，合理地选择统计方法，恰当地解释分析结果，不断修正我们对生物医学现象的认识。

本教材拟从群体与个体卫生健康大数据利用入手，强化健康与医疗评价及其风险影响研究设计（观察性研究、实验性研究、临床试验研究设计、量表研制与设计等，详见第15~18章）；系统介绍统计方法的基本原理及分析结果的正确表述等（第1~14章和第25章）；在普通高等教育“十一五”国家级规划教材《医学统计学》第2版基础上，进一步更新与完善了群体与个体健康评价常用指标及其模型分析（第20~23章），充实并修改了传染病监测数据分析方法（第24章），浅显地引入健康测量指标模型分析理念（第19章），为有效利用爆炸式增长的健康与医疗大数据信息提供了新策略。

小　结

1. 统计学是运用概率论和数理统计的基本原理和方法，研究数据的搜集、清理和分析的一门学科，它是关于研究设计和数据分析的学问，是通过对生物医药研究中有变异数据的获取、存储、管理、转移与分析，汲取健康与医疗信息，解释生物医学现象，并对其结果给予恰当评价的一门科学与艺术。

2. 医学统计工作的基本步骤　分为：①医学研究设计；②搜集资料；③清理资料；④统计分析。统计分析包括统计描述、统计推断和统计预测等内容。统计描述是按研究设计要求，构建关系数据库，实现数据可视化；计算并用统计图表概括反映数据特征。统计推断是根据随机抽样原理，在概括样本信息的基础上，根据其概率大小，对所研究总体的特征做出推断，它是数理统计学方法的主要内容；统计预测是基于大数据整合与分析，阐明变量间的本质联系及其发展变化趋势，做出健康与医学科学决策的一种方法。

3. 根据变量的观察结果不同，将变量分为定量变量和分类变量；将统计分析资料分为