主编　张明芝　李红美　吕大兵

实用医学统计学与SAS应用

苏州大学出版社
Soochow University Press

实用医学统计学与SAS应用

张明芝　李红美　吕大兵　主编

苏州大学出版社

图书在版编目(CIP)数据

实用医学统计学与SAS应用 / 张明芝,李红美,吕大兵主编. —苏州: 苏州大学出版社, 2015.8
ISBN 978-7-5672-1442-2

Ⅰ.①实… Ⅱ.①张…②李…③吕… Ⅲ.①卫生统计-应用软件 Ⅳ.①R195.1-39

中国版本图书馆CIP数据核字(2015)第184763号

书　　名: 实用医学统计学与SAS应用

主　　编: 张明芝　李红美　吕大兵
责任编辑: 倪　青
装帧设计: 刘　俊

出版发行: 苏州大学出版社(Soochow University Press)
社　　址: 苏州市十梓街1号　**邮编**: 215006
印　　装: 常州市武进第三印刷有限公司
网　　址: www.sudapress.com
邮购热线: 0512-67480030
销售热线: 0512-65225020

开　　本: 889mm×1194mm　1/16　**印张**: 31.75　**字数**: 960千
版　　次: 2015年8月第1版
印　　次: 2015年8月第1次印刷
书　　号: ISBN 978-7-5672-1442-2
定　　价: 80.00元

前 言

医学统计学是医学类院校各层次、各专业的必修课程，编写一本适合医学生学习和科研需要的医学统计学教材是非常必要的。

《实用医学统计学与SAS应用》一书由苏州大学高水平新课程建设项目和专业学位硕士案例教材建设项目资助出版。“以促进教学改革、提高教学质量为目标，以增加新知识、新方法、新技术为重点，以医学和生命科学日常及科研工作实际需要为前提，以减少公式的数学证明、推导及减少手工计算为策略”是编写本教材的总的指导思想。

本书从医学科研和卫生工作的实际需要出发，详细介绍了各种统计学方法的应用范围、基本思想、公式计算、SAS程序的实现、结果解释等，着重介绍了解决问题的思路和方法，并注重教材的实用性。学生通过对本教材的学习，能够利用统计学知识和技能解决实际医学问题。

从内容上，本书分为两大部分：基础统计方法和高级统计方法。基础统计方法包括计量资料和计数资料的统计描述、统计图表、参数估计、t检验、方差分析、卡方检验、基于秩和检验的非参数统计方法、线性相关与回归，以及实验设计、调查设计、医学人口统计和寿命表，是面向本科生的；高级统计方法包括协方差分析、多重线性回归、Logistic回归分析、生存分析、聚类分析、判别分析、主成分分析和Meta分析，适合于研究生和本硕连读的医学生学习。在基础统计方法部分，每个章节还增加了“案例讨论”，针对各类统计方法容易被用错的情况提出问题，给出分析问题的思路和方法，以加深学生对知识的理解和应用。另外，为了帮助学生更好地掌握SAS软件包的应用，第二十四章介绍了SAS软件包的基础知识。

本教材具有以下五大特色：一是增加了部分常用的多元统计分析方法、高级统计分析方法及当前国际上卫生统计学的新方法；二是增加了国际上领先的最具权威性的SAS统计软件知识，针对书中的每道例题，均配有SAS程序及运行结果解释；三是减少了统计定理、公式的数学证明部分，重点讲解各统计方法的实际用处、对资料的要求及SAS运行结果的评价；四是本书例题较多彩用了编写人员近几年的科研数据；五是常用统计方法章节的标题以统计学方法及应用目的来命名，实用性强。

本教材适合于临床医学、法医学、放射医学、口腔医学、护理学、预防医学等专业的本科生和硕士研究生使用，同时可作为卫生与医学科研工作人员的卫生统计学参考书。

张明芝

2015年6月于苏州

目 录

第一章 绪 论

第一节 医学统计学的地位和作用

随着科学技术的不断进步，互联网和信息技术的高速发展，人类社会认识世界、改造世界的步伐明显加快。大数据（big data）时代的到来，使得数据证据（evidence）在科学研究中的重要性日益凸显。美国统计学家 William Edwards Deming 曾经说过："我们相信上帝，其他的请用数据说话（In God we trust, all others bring data）。"

生物医学有别于其他学科的一个显著特点是：研究对象具有生命现象，在不同类别、种系、个体之间的生命特征存在着千差万别的变异。从纷繁复杂、杂乱无章的生命现象或信息中发现和探索其特有的规律，是从事临床医学、预防医学、基础医学及其他生物医学学科工作者的使命和任务，即通过不断提高医学科研水平和医疗服务质量，降低疾病负担和死亡，保障人民的生命和健康。除了掌握专业知识外，统计学这门工具学科将有助于我们解决相关实践工作中所遇到的问题：如何做一个好的科研设计？如何记录或描述人类疾病的分布特征？如何研究影响疾病发生、发展的相关因素和机制？如何发现和验证新的临床治疗药物或治疗技术的疗效和副作用？如何科学地向大众呈现和传播研究成果？

在人类医学发展史上，运用统计学知识解决医学领域问题的事例数不胜数。其中最著名的案例就是"反应停（thalidomide）与短肢畸形（complete phocomelia）的关系研究"。

1960 年前后，在西欧诸国家，特别是德国与英国等国，新生儿患短肢畸形明显增加，其临床特点是四肢多处缺损，故称为"短肢畸形"或"海豹肢畸形"，还发生无耳、无眼、缺肾、心脏畸形等。此事当时就引起了医学界的关注。

一些学者根据西欧国家的医院或门诊记录对该病做了一些描述性研究。有资料表明，1959 年以前很少有关于该病的记录，但从 1959 年开始，有较多的记录，1960 年明显增加。德国、英国、美国、加拿大、日本、比利时等国家均有发生。

在短时间内许多国家出生畸形数量异常增加，这是一个不正常的现象，这也意味着人类的生殖细胞或胚胎发育正受到外界环境中某种致畸物的威胁。1961 年 12 月，澳大利亚新南威尔士一名医生 W. G. McBride 在著名的医学期刊 *Lancet* 的读者来信中提出，孕妇服用反应停可能与短肢畸形有关。在当时，反应停被认为是安全的止吐剂，可以防止妊娠呕吐，曾广泛应用。随后，Lenz W 等通过收集不同国家反应停销售量与短肢畸形的资料做相关分析后发现，反应停销售量与短肢畸形有统计学意义的关联（$r_s = 0.881$，$P = 0.004$，表 1-1）。

反应停的销售量与短肢畸形在时间分布上也有密切联系。图 1-1 显示：反应停从 1959 年开始在市场上销售，1960 年销售量迅速上升。1960 年底至 1961 年初，该病病例亦迅速增多。令人惊奇的是，这两条曲线正好相隔大约 3 个季度，这在医学上也可以得到解释，即这些病例的母亲在怀孕 1 个月左右正好有较强烈的妊娠反应，因而服用此药，而从服药到分娩正好相差约 9 个月。

Weicker H 等在 1962 年报告了关于反应停暴露与短肢畸形风险关系的病例对照研究结果：畸形儿的母亲有服用反应停史的为 68.0%，而对照组只有 2.2%，两者之间的关联强度比值比（odds ratio，OR）达到了 93.5，有高度统计学意义（$\chi^2 = 69.40$，$P < 0.0001$，见表 1-2）。

研究人员通过大量的数据收集和统计分析证实，反应停是短肢畸形的罪魁祸首。这件关于药物导致出生畸形重大公共卫生事件的发生，促使不少国家建立了先天性畸形监测系统，强化了临床新药在进入

市场之前的准入制度，即需要经过临床试验证明其安全性和有效性。这充分说明了医学统计学对医学科研发展的作用是巨大的。

目前，医学统计学已经成为临床医学、预防医学和基础医学等专业学生的必修课，医学研究者也愈来愈重视统计学方法的应用，统计思维和方法学已经渗透到医学研究和卫生研究决策之中。

表 1-1　不同国家反应停销售量与短肢畸形的关系

国家	反应停销售量(kg)	短肢畸形例数
奥地利	207	8
比利时	258	26
英国	5769	349
荷兰	140	25
挪威	60	11
葡萄牙	37	2
瑞士	113	6
西德	30099	5000

$r_s = 0.881$，$P = 0.004$

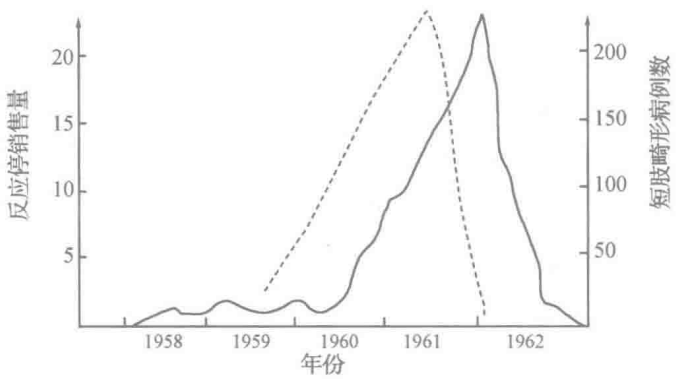

图 1-1　西德反应停销售量(虚线)与短肢畸形病例数(实线)的时间分布曲线

表 1-2　反应停与短肢畸形关系的病例对照研究（Weicker H，1962）

服用反应停史	畸形儿母亲(%)	对照(%)	合计
有	34(68.0)	2(2.2)	36
无	16(32.0)	88(97.8)	104
合计	50	90	140

$\chi^2 = 69.40$，$P < 0.0001$

第二节　医学统计学的定义与内容

一、统计学的来源及概念

“统计学”一词来源于古希腊的城邦统计(state statistics)。《流行病学词典》(John M. Last，*A Dictionary of Epidemiology*)中关于统计学的定义是“The science and art of dealing with variation in data through collection，classification，and analysis in such a way as to obtain reliable results”，即统计学是一门处

理数据中变异性的科学与艺术，内容包括收集、整理、分析、解释和表达数据，以获得可靠的结果。

二、医学统计学的概念

随着人类社会的不断发展，统计学作为一个工具学科已经逐步在各个学科中得到应用，如农业、工业和医学等。医学统计学（medical statistics）是一门运用统计学尤其是数理统计学的原理和方法，研究医学科研及卫生工作中有关数据的收集、整理、分析的学科。国内有“医学统计学”、“卫生统计学（health statistics）”等不同名称。医学统计学与卫生统计学的原理和方法完全相同。医学统计学的应用侧重于临床医学、基础医学、口腔医学、中医学等学科的非公共卫生方面的研究。卫生统计学的应用更侧重于医学与卫生学等公共卫生研究领域，如居民健康状况及卫生服务领域；国际上更普遍使用“生物统计学（biostatistics）”这一名称。生物统计学的应用涵盖整个生物学范围，对统计方法的重视和数理研究也是生物统计学区别于医学统计学与卫生统计学的另一显著特点。

三、统计学发展简史

统计学是建立在科学理论和试验研究基础上的一门学科。它的发展过程大致分成三个阶段：17 世纪到 18 世纪中叶的古典统计学、18 世纪中叶到 19 世纪末叶的近代统计学以及 19 世纪末叶至今的现代统计学。

（一）古典统计学

古典统计学的代表人物是德国的 Conring（1606—1681）。从 1660 年 11 月 20 日起，他在希尔姆斯特大学以“国势学”为题，讲授欧洲各国领土、人口、财政等信息，通过比较的方法来研究国家的盛衰变化。他是国势学派的创始人。1749 年，他在《近代欧洲各国国势学纲要》一书中首次提出“统计学”（statistic）一词。随后，政治算术学派的创始人 John Graunt（1620—1674），通过收集死亡方面的数据，编制了世界上第一张寿命表。古典统计学的另一个重要分支——概率论学派的真正发展是从 17 世纪开始的。其基础理论的奠定者为法国数学家 Blaise Pascal（1623—1662）和 Pierre de Fermat（1606—1665），他们经常就赌博中的得分问题进行交流，从而引进类似概率论的基本理论——数学期望的计算。

（二）近代统计学

在近代统计学时期涌现出一大批杰出的统计学家。国势学派后期代表人物，德国统计学家 August Ludwig Von Schlozer（1735—1809）在 1804 年出版的《统计学原理》一书中提出建议：以确切的数字代替笼统的叙述；August Niemann（1761—1832）于 1807 年出版了《统计学与国势学纲要》，他认为统计学是将收集到的事实加以整理和记述的学问。

近代统计学——政治算术学派在人口统计、保险统计、医学统计、卫生统计及农业统计等方面涌现出许多有代表性的学者。其中，法国著名医师 Pierre-Charles-Alexandre Louis（1787—1872）倡导用统计方法研究卫生或医疗问题，他被尊称为“医学统计之父”。他认为，随着观察例数的增多，疗效逐步显现，即用数据表达临床疗效，提出医学上的“数量法”。Willam Farr（1807—1883）在英国首创了人口和死亡的常规资料收集方法，并提出了一系列公共卫生领域的重要概念，如标准化率、人年、剂量反应关系等，促进了统计学在公共卫生领域的应用。John Snow（1813—1858）从 1848 年起对伦敦地区发生的霍乱流行进行了深入研究，编制了统计地图，提出了霍乱暴发是通过水传播的论断，并成功地控制了霍乱的流行，这是统计学在公共卫生与疾病控制领域应用的典范。

在此阶段，统计学理论创新获得较大发展。例如，法国数学家 Legendre 提出最小二乘法；德国数学家 Gauss 根据天文观察和土地测量的经验发现，观察值 x 与真正值 μ 的误差变异服从正态分布（高斯分布），他用极大似然法推出了测量误差的概率分布公式。法国数学家 Laplace 最早系统地把数学分析方法运用到概率论研究中，将微积分应用于概率分布理论和分析方法，建立了严密的概率数学理论；他发明了大数法则，并进行了大样本推断的尝试。

（三）现代统计学

19 世纪末叶至今，英、法、美等国在统计学研究方面成果丰硕，名家辈出。作为生物统计学派的奠基人，英国人类学家 Galton 除了将正态分布理论应用到人类遗传和进化问题外，还提出了相关、回归的概念；Karl Pearson（1857—1936）是公认的现代统计学之父，他对统计学的学术和人才培养贡献巨大。他初步建立了系统的数据分析统计学方法，如提出描述生物变异程度的指标——标准差、卡方检验方法、拟合优度检验，发展了相关和回归指标计算方法。他一直热衷于统计学的教育和统计人才的培养工作。他于 1894 年在剑桥大学开设了统计理论课，1901 年创建了世界上最权威的生物统计学杂志 *Biometrika*，1911 年创建了世界上第一个应用统计系。W. S. Gosset（1876—1937）曾经是爱尔兰都柏林一家啤酒厂的酿酒师，1906—1907 年，他有机会去 Pearson 那里学习统计学。1908 年，他以 Student 作为笔名发表了著名的小样本抽样误差理论分布的文章，开创了小样本抽样统计推断的新纪元。R. A. Fisher 被认为是 20 世纪贡献最大的统计学家。他的工作量多、质高、面广。他建立了方差分析的理论和方法，提出了实验设计控制误差的方法，创立了检验理论和估计理论等统计理论体系（假设检验），使得推断统计学成为数理统计的主流。美国统计学家 J. Neyman（1894—1981）的最大贡献在于使统计学建立在严格的数学基础上，如提出“区间估计”概念，在数学上完备了“假设检验”和“区间估计”的理论体系。

随着计算机技术的发展，统计学理论和算法日趋复杂，更新更加迅速。生存分析方法、多元回归、广义线性模型、EM 算法（expectation maximization algorithm）和贝叶斯理论等已经广泛用于医学科研实践中。

统计学在我国的发展也经历了一个曲折的过程。中国最早的统计活动出现在原始社会末期，结绳刻契为我国统计的萌芽。随着社会经济的发展和社会的变革，春秋战国时期统计思想十分活跃，并产生了最初的统计分析。《尚书·禹贡》中的九州表是国家学术界公认的最早的统计史料。19 世纪末，统计学在我国作为一门学科开始被接受。1903 年，我国最早的统计学教材正式出版；20 世纪 40 年代，中国数理统计学理论研究也逐渐取得突破。其中著名的统计学家有许宝騄教授（1910—1970），他在 Neyman-Pearson 理论、参数估计理论等方面取得了卓越成就。著名的遗传学家、生物统计学家李景均教授（1912—2013）1948 年出版了《群体遗传学》，在学术界影响较大。20 世纪中叶，李光荫教授、许世谨教授、薛仲三教授和郭祖超教授是我国卫生统计学的奠基人。改革开放后，我国统计学迎来了前所未有的机遇，在国际统计学的舞台上涌现出一批著名的中国学者。我国生物统计学理论研究和发展仍然任重道远。

第三节　基本概念

一、观察单位

观察单位（observation unit）又称为研究个体或研究对象，是科学研究工作中最基本的单位。数据的获得来自于每个观察单位或研究对象。在进行科研设计和数据收集时，要根据研究目的确定研究对象或最基本的观察单位。研究目的不同，其观察单位或数据收集的载体相异，它可以是一个人、一只动物、一个家庭、一个社区或一个国家，也可以是一种器官、一种组织、一个细胞等。例如，研究儿童青少年学生的近视率时，其观察单位是每个学生；研究家庭的年收入时，其观察单位就是每个家庭；研究乳腺癌肿瘤组织、癌旁组织及正常组织 P53 基因的突变率时，研究单位则是每个研究对象的某个组织。

二、同质和变异

通俗地讲，同质（homogeneity）是指根据研究目的所确定的性质相同的人或事物。物以类聚，人以群分。而从科研设计的角度来说，同质是指规定研究对象在某些性质上相同或者指对研究指标有影响的主

要因素相同。例如,在研究2015年某地7岁男童身高这项调查中,我们规定了相同年龄、性别、地区和调查年份,这些因素也是影响儿童身高的主要因素。在科学研究中,要求研究对象同质也是为了最大限度地控制混杂因子的影响,减少研究偏倚,使研究结果更接近于真实情况。

变异(variability)是指即使规定了同质的对象,其测量值或观察结果也不尽相同。在前述的例子中,尽管我们规定了年龄、性别、地区和调查年份,但由于每个儿童的遗传背景、营养、生长环境等因素不尽相同,因此他们的身高参差不齐。这种生物学现象在自然界中普遍存在。古人云:"一母生九子,九子各不同。"探索变异数据或现象背后的原因不仅是统计学这门学科的主要任务,更是医学工作者的使命。遗传、环境、社会心理因素都在错综复杂地影响着每个个体的健康和生命质量。如:每个不同的个体感染结核杆菌后所导致的结局各不相同;同一病理类型的肿瘤,相同的病期,相同的年龄、性别,用同一种抗肿瘤疗法,其生存时间却不完全相同,其背后的原因是什么?这些都值得我们去探索和追根求源。目前,肿瘤个体化治疗已经取得了很大的进展。例如,乳腺癌、肺癌靶向治疗药物的使用,大大提高了具有基因治疗靶点的肿瘤患者的生存率。

从哲学角度讲,同质指的是矛盾的普遍性,而变异则是指矛盾的特殊性。世界上没有两片完全相同的叶子。医学科学研究者既要从纷繁复杂的变异现象中总结普遍性的规律,也要充分考虑每个研究个体与众不同的特性,辩证和客观地理解同质和变异。

三、总体和样本

总体(population)是指根据研究目的所确定的同质观察单位或其观察值的集合。总体可分成无限总体和有限总体两类。无限总体(indefinite population)是指不限制研究的时间和地点等因素,其所包含的观察单位是无限的,如研究成年人的血压。有限总体(definite population)是指限制了时间、地点等时空因素,研究对象的数目是有限、可数的。即使是有限总体,其所包含的观察单位也可能是巨大的,我们不大可能对总体中的每个观察单位一一进行研究,这样也不切实际。例如,研究某个兵工厂所生产弹药的质量如何时,我们往往采取抽样研究(sampling study),即从总体中随机抽取部分观察单位组成样本,由样本信息推断总体信息(图1-2)。从总体的全部观察单位中随机抽取的部分观察单位(某项特征的观测值)的集合叫作样本(sample)。样本中所含的观察单位数叫样本含量(sample size)(或样本量、样本大小),一般用 n 表示。抽样研究是医学统计学的核心思想。其关键步骤是所抽取的样本要具有代表性。一句西方谚语说得好:要想知道牛肉的味道是坚韧的,也不必吃掉一头牛(you do not have to eat the whole ox to know that the meat is tough!)。

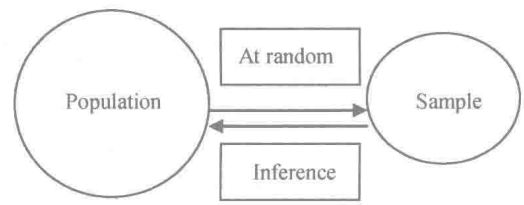

图1-2 抽样研究示意图

四、参数和统计量

参数(parameter)是相对于总体的特征值,又称为总体参数,是由总体中全部观察值计算出来的特征值,是固定的常数,但往往未知(如总体均数 μ、总体率 π、总体标准差 σ 等);而统计量(statistics)是相对于样本的特征值,又称为样本统计量,是由样本的所有观察单位计算出来的特征值,其值因每次抽样样本的不同而异,是可知的(如样本均数 $\bar{x}$、样本率 p 等)。

五、变量和资料类型

总体或样本所对应的观察单位的某项特征(或指标)称为变量(variable),对变量的观测值或观察值称为变量值,变量值构成资料(data)。资料可分为以下几种类型:

（一）数值变量（numerical variable）资料

数值变量资料又称定量资料或计量资料，用定量的方法对每个观察单位的某项定量指标测得对应的数据，一般有度量衡单位。如空腹血糖（mmol/L）、体重（kg）、血压（kPa）等资料。

（二）分类变量（categorical variable）资料

分类变量资料又称定性资料或计数资料，是定性特征（或指标）的变量值。定性指标的变量值为某种属性或某种类别。分类变量资料包含二分类变量资料与多分类变量资料。

1. 二分类变量资料

观察单位某项定性特征（或指标）分为两种不同性质或类别的资料，称为二分类变量资料。例如，性别按男女分类，某样本中各个观察单位性别的取值为男或女；肺癌病人治疗 2 年后的结局分为两类：生存和死亡，某样本各个肺癌病人治疗 2 年后结局的取值为生存或死亡；某种医学检验的结果分为阳性、阴性；等等。上述男和女的取值、生存和死亡的取值、阳性结果和阴性结果的取值均属于二分类变量资料。二分类变量资料不定义为等级资料。

2. 多分类变量资料

观察单位某项定性特征（或指标）的取值为多种不同性质或类别的资料，称为多分类变量资料。它又可分为多分类有序变量资料及多分类无序变量资料。

（1）多分类有序变量资料：指按观察单位某项特征（或指标）的不同程度（等级）分组的多分类变量资料，其分类往往有等级强弱关系，也称等级资料或半定量资料（semi-quantity data）。如某血清反应根据反应强度分为 -、±、+、++、+++、++++ 共六个等级，各等级的取值是有序的。一般来说，等级资料的等级划分界限不是十分清楚，人为因素较多。

（2）多分类无序变量资料：指观察单位某项特征（或指标）在各组间是无序的，不反映等级关系，其分类界限非常清楚。如 ABO 血型资料，分为 A、B、AB、O 四组，四组间无等级关系。

当然，资料类型的划分不是绝对的，它们之间是可以相互转化的。定量资料可以转化为定性资料或等级资料，定性资料也可以数量化转化为定量资料。例如，健康调查简表 SF-36 中把健康状况分为“非常好”、“较好”、“一般”、“差”、“非常差”五个等级，应划归为等级资料。但若将这五个等级数量化，分别将它们赋值为 5、4、3、2、1，就转化为定量资料了。不同的资料类型要用不同的统计方法进行处理。

六、频率与概率

一个试验（如某种治疗、医学检查、医学观察、医学调查、实验室实验等），有两种对立的结果：事件 A（试验结果可能出现的某种现象）发生或不发生。一次随机试验结果，事件 A 发生或不发生完全是偶然的；观察大量试验，事件 A 的发生呈现统计规律性的结果，此种试验称为随机试验，此种事件称为随机事件。

（一）频率（relative frequency）

将随机试验重复 n 次，n 次试验中随机事件 A 共发生 m 次，则 $\frac{m}{n}$ 表示随机事件 A 发生的频率。

（二）概率（probability）

随机事件发生的可能性大小称为概率，记作 P。

1. 概率的统计定义

当试验次数 n 趋向于无穷大时，频率 $\frac{m}{n}$ 的极限值为概率，即 $P=\lim\limits_{n\to\infty}\frac{m}{n}$。在实际中，当 n 很大（大样本）时，往往用频率近似估计概率。

2. 概率的古典定义

一个随机试验，有 n 种等可能的结果，其中有利于事件 A 发生的结果数为 m，则事件 A 发生的概率等

于$\frac{m}{n}$。例如，随机事件为抛掷一枚质量分布均匀的硬币，有两种等可能的结果：硬币落下时数字朝上或数字朝下。事件 A 定义为硬币落下时数字朝上，有利于数字朝上的等可能的结果数为 1，则抛掷一枚硬币出现数字朝上的概率为$\frac{m}{n}=\frac{1}{2}=0.5$。

由以上概率的两个定义可看出，$0\leqslant P\leqslant 1$。当 $P=0$ 时，称为不可能事件；当 $P=1$ 时，称为必然事件。当 P 很小时，称为小概率事件(rare event)。例如，$P\leqslant 0.05$ 或 $P\leqslant 0.01$ 的随机事件，试验 100 次，平均发生数不超过 5 次或 1 次，一般认为是几乎不会发生的事件。小概率推断原理：一般认为，小概率事件($P\leqslant 0.05$ 或 $P\leqslant 0.01$)在一次抽样中是不会发生的。

第四节 统计工作的基本步骤

统计工作可分为统计设计(design)、资料收集(collection of data)、资料整理(data sorting)和资料分析(analysis of data)共四个步骤。

一、统计设计

统计或医学科研工作是一项比较复杂的系统工程，在正式实施这项工作之前，要有详细的规划和设计。科学的设计是决定一项科研工作成败的关键。如果统计设计存在明显缺陷，即使统计方法再高深，所得的结论也是不可靠的。所以有的学者提出，学习医学统计学不仅仅是学会数据的处理与分析方法，同时应该掌握这门学科中所蕴含的许多科学的科研设计思想。

非常遗憾的是，即使在今天，仍有许多研究人员并不真正重视科研设计中所蕴含的统计设计元素，如随机化、对照、样本含量的大小等问题。往往都是等数据收集完成，遇到具体困难了，才去咨询统计专业人员。英国著名统计学家、现代统计学的奠基人之一 Fisher 曾指出，做完实验以后再去找统计学家无异于请他做尸体解剖，他能做的事情就是告诉你这项实验失败的原因是什么。因此，在研究之前一定要查阅大量文献，必要时可咨询统计学专家，做好周密的设计。

二、资料收集

卫生统计资料有三个来源：①统计报表，如死因报表、法定传染病报表、职业病报表、医院工作报表等，这些报表由管理部门统一设计、逐级上报。有些地区还建立了恶性肿瘤发病报告制度，这为肿瘤疾病的监测和预防积累了重要的基础资料。②经常性工作记录，如卫生监测记录、健康检查记录、门诊病历、住院病历等。③专项实验或专题调查。如果现有的资料不能解决实际的科学假设，就需要专门收集相关的人群或实验资料，如某个基因位点多态性与肺癌发病风险的关系研究。

目前，随着计算机储存技术、互联网技术的发展，近 30 年来，许多发达国家已经建立了每个公民关于出生、住院、门诊药物使用和疾病诊断相关的人群健康数据库，这为研究相关疾病和健康问题提供了重要的基础资料。

目前，关于大数据的挖掘和研究方兴未艾。生物医学大数据的特点是“4V”准则：Volume，数据体量巨大，为 TB 到 PB，包括变量大于样本量的高维数据；Variety，数据类型繁多，如图片、地理信息等；Value，价值密度低，商业价值高；Velocity，需要处理速度快。

三、资料整理

资料整理也称数据处理。医学数据的统计处理涉及医学专业知识、统计专业知识、处理数据的经验

和技巧，是一门高深的艺术。在资料整理过程中，原始数据的录入、数据的管理、统计软件的使用都是必须重视的关键环节。随着计算机软件（如 Epidata、Excel、Foxpro、Access 等）在数据处理中的广泛运用，数据管理过程变得更为便捷。现代的数据处理技术流程一般包括调查表的双遍编码、双人双遍录入、双遍核查、逻辑校对，其目的是保证将收集的资料比较准确地录入数据库中，这也是数据处理过程中的质量控制措施。

四、资料分析

资料收集完成后，最后的工作是资料分析。它包括两个方面的主要内容：统计描述和统计推断。统计描述是指采用统计图、统计表及相关的统计指标对数据的分布特征进行简单的描述；统计推断是指由样本信息推断总体信息的过程，包括参数估计和假设检验。

第五节　学习医学统计学应注意的问题

医学统计学是医学专业学生的专业基础课，学好该课程可为其他专业课程的学习打下必要的统计学基础，为毕业后从事卫生及相关领域的研究和实际工作提供分析问题、解决问题的方法。为此，学习本门课程时，应该注意以下几个问题：

（1）重点应放在卫生统计学基本概念和基本原理的理解和掌握。对于任何一门学科来说，其基本概念和基本原理都是整个学科体系的基石，统计学也不例外。只有深刻理解和掌握这些基本概念和基本原理，才能举一反三，运用这些原理和方法解决卫生实践中的实际问题。

（2）重点应放在基本统计方法的适用条件、用途及注意事项的理解和掌握上。对于一般的卫生工作者而言，并不需要深究统计公式的推导过程和死记硬背统计概念与公式，重点要放在一些基本统计方法的适用条件、用途及注意事项的理解和掌握上，即一些基本统计方法在资料具备什么条件下可用、用来解决什么问题、使用时应注意什么问题（包括统计分析结果如何正确解释与表达）等的理解和掌握。

（3）要能结合实际案例，充分运用数据库和统计软件解决数据录入、资料处理、结果解释等问题。

（4）要注意学习医学统计学不仅仅是学习如何进行数据统计的技术，更重要的是学会用统计思维方法进行医学科研设计，合理解释统计结果。

（5）在大数据来临的时代，医学统计学技术非常重要。但在收集、整理和处理科研数据时，一定要尊重原始数据，要培养学术诚信的优良品德，而不能为一己私利或论文的发表而篡改或伪造数据。

练 习 题

一、最佳选择题

1. 观察单位为研究中的（　　）。

A. 样本　　B. 总体　　C. 影响因素　　D. 个体

2. 统计学中所说的总体是指（　　）。

A. 任意想象的研究对象的全体　　B. 根据研究目的确定的研究对象的全体

C. 根据地区划分的研究对象的全体　　D. 根据时间划分的研究对象的全体

3. 统计学中的样本是指（　　）。