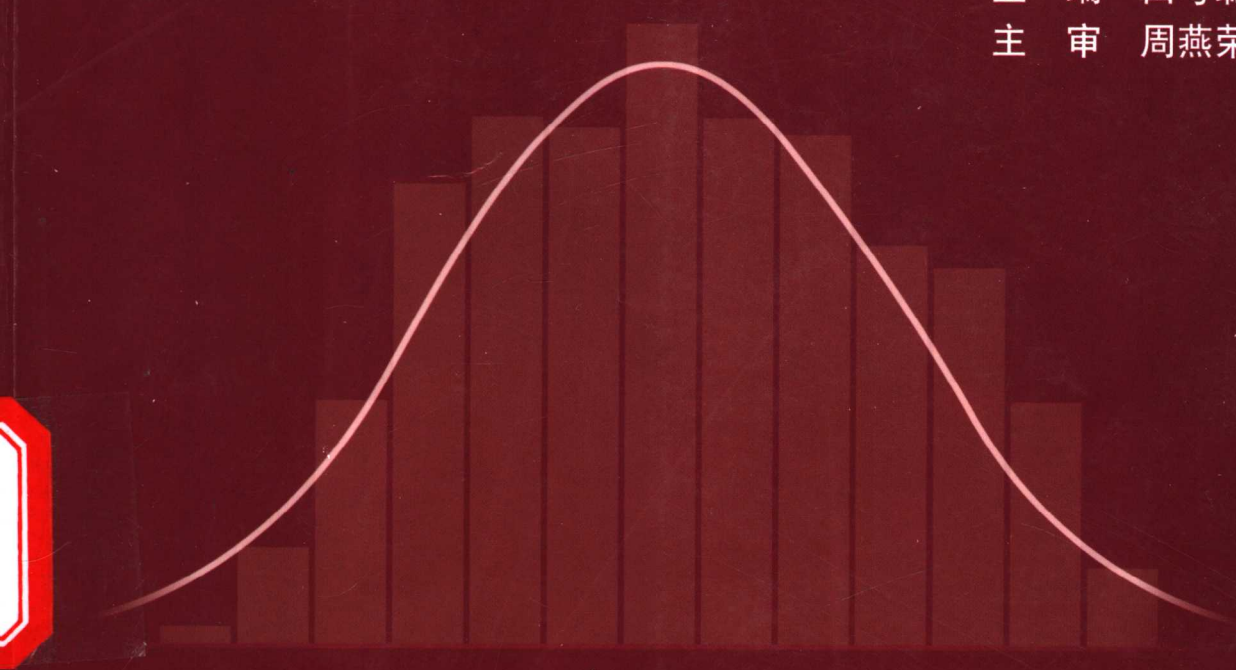
中国医学统计百科全书

描述性统计分册

总主编　徐天和
主　编　田考聪
主　审　周燕荣



人民卫生出版社

中国医学统计百科全书

描述性统计分册

总主编　徐天和
主　编　田考聪
主　审　周燕荣

人民卫生出版社

图书在版编目(CIP)数据

描述性统计分册/田考聪主编.—北京:
人民卫生出版社,2004.4
(中国医学统计百科全书)
ISBN 7-117-05955-9

Ⅰ.描… Ⅱ.田… Ⅲ.医学统计-中国-百科全书 Ⅳ.R311-61

中国版本图书馆 CIP 数据核字(2004)第 005830 号

中国医学统计百科全书

描述性统计分册

总 主 编:徐天和
主　　编:田考聪
出版发行:人民卫生出版社(中继线 67616688)
地　　址:(100078)北京市丰台区方庄芳群园 3 区 3 号楼
网　　址:http://www.pmph.com
E-mail:pmph@pmph.com
印　　刷:北京铭成印刷有限公司
经　　销:新华书店
开　　本:787×1092　1/16　　**印张**:11.5
字　　数:282 千字
版　　次:2004 年 5 月第 1 版　2004 年 5 月第 1 版第 1 次印刷
标准书号:ISBN 7-117-05955-9/R·5956
定　　价:24.00 元

总 主 编 徐天和
主　　编 田考聪
副 主 编 易　东　王洁贞　潘晓平
主　　审 周燕荣
编　　委 （以姓氏笔画为序）
王　玖　王昌玲　王洁贞　王润华　尹全焕
冯丽云　田考聪　李会庆　李炽民　刘隆健
张菊英　易　东　周燕荣　施学忠　唐　军
徐天和　潘晓平
秘 书 长 祁爱琴
学术秘书 王　玖　石德文

序

随着医学科学研究的不断深入发展，对科研设计与科研数据的统计处理方法提出了越来越高的要求，现代医学统计科学的新理论、新方法也在不断涌现，但迄今为止国内尚无一部大型医学统计专著。中国医院统计杂志社发起并组织国内百余名从事医学统计研究与教学的专家、学者编撰的这部《中国医学统计百科全书》填补了这一空白。

该书的各位编写者都是工作在医学统计研究与教学第一线的专家、学者，其中归国留学人员占60%，他们为国内带回了最新研究成果。我衷心祝贺本书的出版，并为我国医学统计学界人才辈出、长江后浪推前浪的喜人形势而感欣慰。我相信该书的出版发行一定会大大推动我国医学科研事业的发展。

2003年10月

编者说明

统计描述，作为医学统计学的重要组成部分之一，是进行统计推断的基础。它是在收集、整理数据的基础上，通过相应的统计量以及统计图和统计表来描述资料某些特征的统计方法。一般说来，进行统计描述应遵循这样的原则：根据分析目的和资料类型选择恰当的统计量以及统计图和统计表来描述资料的统计特征。

统计描述涉及较多的概率论与数理统计学的基本概念。如果对这些基本概念缺乏清楚的认识，在面对实际问题时就难以对需要处理的资料有一个全面的把握，也就无从选择正确的统计量来描述资料的统计特征，在进行统计分析时，对如何选择统计方法就会感到无所适从。对于广大实际工作者而言，没有必要要求他们全面系统地掌握概率论与数理统计学知识。为了解决这个矛盾，本书旨在以最简明的文字，深入浅出、准确地阐明统计描述所涉及的数学概念，并尽可能地选用了恰当的实例，以使读者能在较短的时间内对自己所要解决的问题有一个清楚的认识，从而达到正确进行统计描述、选择正确的统计分析方法的目的。

随着医学研究在广度和深度上的迅猛发展，对医学统计学提出的要求也越来越高，就统计描述而言，所涉及的内容也越来越丰富。作为一本工具书，在这个分卷里，我们尽量地收集了描述性统计的内容，同时还对近年来在医学领域中使用较多的一些描述性内容作了较为详细的介绍。全书以条目格式编写，主要介绍基本概念、重要理论、科学依据、计算方法、推断结论，并辅之以必要的实例，使读者便于检索，易于理解和掌握，具有很强的实用性。

由于涉及面广、工作量大、经验不足，加之水平有限，本书中的缺点错误在所难免，敬请读者批评指正。

田考聪

2003年6月

序　言

世界各国尤其是发达国家都非常重视百科全书的编纂工作。它是衡量一个国家科学技术发展水平乃至综合国力的标志之一。随着时代的前进和科技的进步，特别是党的十一届三中全会以来，我国的社会主义现代化建设和文化科学事业蓬勃发展。统计科学和医学统计工作的发展进入了一个新阶段。统计方法作为一种获取信息和科学研究的工具与作出决策的依据，其重要性正被越来越多的人们所认识。为总结我国改革开放以来统计科学和统计工作的成果经验，吸收与传播现代医学统计科学的新理论、新方法与新成果，迫切需要编撰一部具有中国特色的医学统计百科全书，以填补国内这一空白。为此，中国医院统计杂志社发起并任总主编单位，组织全国部分高等医学院校及有关医疗卫生机构的百余名统计学教授、专家，经过6年的辛勤劳动，发挥集体智慧，共同编撰了《中国医学统计百科全书》。

《中国医学统计百科全书》是一部大型医学统计参考工具书，主要读者对象是全国医学统计工作者和医疗卫生单位统计信息工作者以及高等医学院校的师生，预计在工作中需要查阅这部百科全书的读者将远远超过这一范围。全书包括描述统计、推断统计、非参数统计、多元统计、统计设计和健康与管理统计等内容，选材着重其在医学上的应用。由于近年来统计理论与方法发展十分迅速，高效、实用的新方法不断出现，像多变量分析方法和非参数统计分析等较高层次的统计方法，在应用上日益普遍，本书以较大的篇幅重点进行了阐述。以便既体现统计理论与方法的完整性，又反映学科的先进水平。全书用条目形式撰写，一个条目介绍一种统计方法，包括问题的背景、方法的理论和直观依据、实施步骤和实例等。由于书的性质，不可能对方法的数学理论根据作严格的推导，但也避免把本书写成一种“菜谱”式的东西。因为，为了用好一种统计方法，对其背景和依据

有些了解，是很重要的。我觉得这是本书写作的一个重要特点，这样使本书能比较全面而确切地介绍了医学统计科学的重要内容与最新研究成果，又不失其以应用为主的特性。另外，本书的编撰方法科学性强，层次分明，结构严谨，既突破了传统的辞典式编撰方法，又汲取了辞典的某些特点。本书强调实用性，深入浅出，具有知识性、可读性、可查性和适用性，它适合于医学各专业、不同层次和不同专业需要的读者阅读。作为医学界、医学统计理论界和医疗卫生统计信息部门的一部大型专业工具书，我相信它会成为这方面专业人士书架上常备之书，对推动我国医学统计科学和统计工作的发展作出积极的贡献，故乐为之序。

陈峰

2003年7月

目　录

医学统计学 …… (1)
概率 …… (3)
随机变量及其分布 …… (6)
总体与样本 …… (8)
参数与统计量 …… (9)
误差 …… (11)
统计工作步骤 …… (13)
资料类型 …… (16)
数值资料与频数表 …… (17)
分类资料与分类表 …… (19)
数据的预处理 …… (20)
未检出值的估计 …… (21)
异常数据的发现与处理 …… (23)
有效数字与数字舍入规则 …… (26)
相对数 …… (27)
率 …… (28)
比 …… (29)
标准化率 …… (31)
动态数列 …… (34)
危险度 …… (36)
相对危险度 …… (37)
归因危险度和归因危险度百分比 …… (39)
人群归因危险度和人群归因危险度百分比 …… (40)
标准死亡比 …… (41)
比值比 …… (43)
诊断性试验的评价 …… (45)
平均数 …… (50)
算数平均数 …… (51)
几何均数 …… (53)
众数 …… (55)
调和平均数 …… (56)
中位数和百分位数 …… (58)

变异指标 …… (61)
二项分布 …… (68)
超几何分布 …… (72)
Poisson 分布 …… (75)
负二项分布 …… (78)
均匀分布 …… (81)
正态分布 …… (84)
χ^2 分布 …… (87)
t 分布 …… (90)
F 分布 …… (92)
Γ 分布 …… (95)
指数分布 …… (99)
威布尔分布 …… (103)
圆形分布 …… (108)
统计表与统计图 …… (113)
参考值范围 …… (121)
正态分布法 …… (123)
百分位数法 …… (125)
容许区间法 …… (126)
混杂样本剖析法 …… (128)
最可能数法 …… (131)
多元分析法 …… (134)
附录一 统计用表 …… (136)
附表 1 二项分布的概率值 …… (136)
附表 2 标准正态分布密度函数曲线下的面积 …… (141)
附表 3 χ^2 界值表 …… (142)
附表 4 t 界值表 …… (145)
附表 5-1 F 界值表(方差分析用,$P=0.05$) …… (147)
附表 5-2 F 界值表(方差分析用,$P=0.01$) …… (151)
附表 6 F 界值表(方差齐性检验用) …… (155)
附表 7 $\beta=1$ 时 Γ 分布界值表 …… (157)
附表 8 圆形分布的 r 界值表 …… (162)
附表 9 K 值表 …… (163)
附表 10 Bessel 函数 …… (164)
附表 11 圆形正态分布的分布函数表(平均角 $\theta=180°$) …… (165)
附表 12 Poisson 分布中实际数与预期数之比的界值表 …… (167)
附表 13 Poisson 分布 μ 的可信区间 …… (168)
附表 14 正态分布容许区间的系数 K …… (168)
附录二 英汉医学统计学词汇 …… (169)
附录三 汉英医学统计学词汇 …… (171)

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,妨碍了人们对事物本质的认识。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、CPD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。可见一项研究课题,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。通常是针对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推断总体。因此，统计学的主体是统计推断。它是根据研究目的和资料性质，利用样本统计量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 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):其直观意义是描述随机事件发生的可能性大小的度量。下面,我们将给出用于建立概率理论体系的严密数学定义。设(Ω,F)是可测空间,对每一集$A\in F$,定义实值集合函数P,它满足如下三个条件:① 对每一$A\in F$有$0\leqslant P(A)\leqslant 1$(非负性);② 对必然事件$P(\Omega)=1$(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。

例1　在生男生女问题中,其样本空间为$\Omega=\{\omega_1,\omega_2\}$,$\omega_1=$(男),$\omega_2=$(女),基本事件为$A_1=\{\omega_1\}$,$A_2=\{\omega_2\}$。已知$F=\{\Phi,A_1,A_2,\Omega\}$,这里$\Phi$表示"既不生男,也不生女"这一事件,$\Omega$表示"生男或生女"这一事件。对$F$中的集合,对应一实数$P$,定义如下:$P(\Phi)=0$;$P(A_1)=P(A_2)=\frac{1}{2}$;$P(\Omega)=1$。显然,$P$满足上述定义中的①、②、③,因此,$P$是$(\Omega,F)$上的概率。

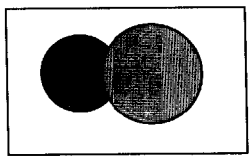
作为上述关于概率定义的特殊情况,一般常用的有古典概率、几何概率。为便于理解,这里给出古典概率的定义:设有一随机试验,其所有可能出现的结果有n个(n为有限数),这n个基本事件是两两互不相容的,且发生的可能性均相等,而事件A恰包含其中的f个结果,则事件A(简称事件A,下同)发生的概率为

$$P(A)=f/n \tag{1}$$

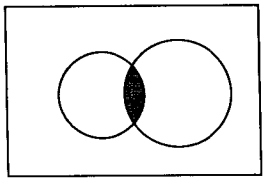
其统计意义是:设在相同条件下,独立重复进行n次试验,事件A出现f次,则称f/n为事件A出现的频率。当n逐渐增大时,频率f/n始终在某一常数P的左右作微小摆动,就称P为事件A的概率,记作$P(A)=P$。在许多实际问题中,当概率不易求得时,只要n充分大,可以将频率作为概率的估计值。

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,通常将$P\leqslant 0.05$或$P\leqslant 0.01$称为小概率事件,表示某事件发生的可能性很小。

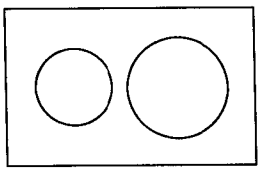
事件之间的相互关系及其示意图如下:



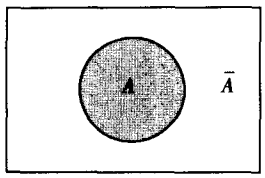
$A+B=${事件A和B至少有一个发生}



$AB=${事件A和B同时发生}



A和B互不相容



$\bar{A}$ = {事件 A 的对立事件}

概率运算法则

·加法法则：

$$P(A+B)=P(A)+P(B)-P(AB) \tag{2}$$

如 A 和 B 互不相容

$$P(A+B)=P(A)+P(B) \tag{3}$$

这一法则可以推广到有限个互不相容的事件。

$$P(A)+P(\bar{A})=1 \tag{4}$$

·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 $P(B|A)$。

对于任意两事件 A 和 B 同时发生的概率为：

$$P(AB)=P(B)P(A|B) \tag{5}$$

$$P(AB)=P(A)P(B|A) \tag{6}$$

如 A（或 B）发生与否并不影响 B（或 A）的概率，就说 A 和 B 相互独立，则事件 AB 的概率等于 A 的概率与 B 的概率之积，即

$$P(AB)=P(A)P(B) \tag{7}$$

这一法则可推广到有限个相互独立的事件

例 2　对某地区 2000 名 60 岁以上老年人的疾病调查中发现：710 人患有高血压，315 人患有糖尿病，150 人同时患有高血压及糖尿病。

A = {患有高血压病}

B = {患有糖尿病}

$A+B$ = {患有高血压或糖尿病}

AB = {既患有高血压又患有糖尿病}

$\bar{A}$ = {未患有高血压病}

由古典概率的定义，当调查数 n 充分大时，可以将频率作为概率

$P(A)=710/2000=0.355$；

$P(B)=315/2000=0.1575$；

$P(AB)=150/2000=0.075$；

$P(A+B)=P(A)+P(B)-P(AB)=0.4375$；

$P(\bar{A})=1-P(A)=0.645$。

例 3　据估计成年人口中约 15% 有高血压，又有成年人口中约 75% 不觉得血压高。同时估计人口中约 6% 患有高血压而不觉有病。如果一成年人认为自己没有高血压，此人实际有病的概率是多少。

A = {不觉有病}，B = {患有此病}

则 $P(A)=0.75$，$P(B)=0.15$，

$P(AB)=0.06$

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

$=0.06/0.75=0.08$

全概率公式：

$$A_i \cap A_j=\Phi_{(i\neq j)},\ \bigcup_{i=1}^{n}A_i=\Omega,\ P(A_i)>0$$

$$P(B)=\sum P(B|A_i)P(A_i) \tag{8}$$

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

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n}P(A_i)P(B|A_i)} \tag{9}$$

例 4　在对某单位生产的注射器进行检验中发现：在 1000 只中有 20 只是不合格产品，现不放回地抽取 2 只注射器，问第 2 次抽取到的是不合格产品的概率。

A = {第一次抽到的是不合格品}

B = {第二次抽到的是不合格品}

$A_1=A$，$A_2=\bar{A}$

$P(A_1)=20/1000$

$P(A_2)=980/1000$

$P(B|A_1)=19/999$

$P(B|A_2)=20/999$

$P(B)=20/1000\times19/999+980/1000\times20/999$

$=0.02$

例5　某地区根据多年的某种病史分析，将病例分为甲状腺功能低下(H_1)、正常(H_2)与亢进(H_3)三类，并得：$P(H_1)=0.15$，$P(H_2)=0.65$，$P(H_3)=0.20$。若仅按“食欲亢进”指标(A)而言，根据积累的资料得：$P(A|H_1)=0.08$，$P(A|H_2)=0.10$，$P(A|H_3)=0.61$。今有某病员新进食量大增，试分别求其甲状腺功能低下，正常与亢进的概率。

$$P(H_1)P(A|H_1)=0.15\times0.08=0.0120$$

$$P(H_2)P(A|H_2)=0.65\times0.10=0.0650$$

$$P(H_3)P(A|H_3)=0.20\times0.61=0.1120$$

$$\sum P(H_i)P(A|H_i)=0.1990$$

由 Bayes 公式

$$P(H_1|A)=0.0120/0.1990=0.0603$$

$$P(H_2|A)=0.0650/0.1990=0.3266$$

$$P(H_3|A)=0.1220/0.1990=0.6131$$

故仅按“食欲亢进”一项指标而言，该病员患甲状腺功能亢进的可能性最大，大约2倍于正常，10倍于低下。

（易　东　尹全焕）

随机变量及其分布

随机变量(random variable)的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。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。但在许多实际问题中,有时很难精确地求出其分布函数,或只须知道一、二个描述分布的特征参数,常用的有数学期望、方差、矩等等。

1)离散型分布:当ξ的一切可能取值为$x_1, x_2, \cdots, x_n \cdots$,则称$\xi$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$\xi$的分布列,亦称为$\xi$的概率函数。对离散随机变量,如下列出更为直观:

ξ	x_1	x_2	$\cdots$	x_n	$\cdots$
$P(\xi=x_n)$	p_1	p_2	$\cdots$	p_n	$\cdots$

$p_n \geqslant 0(n=1,2,\cdots), \sum_{i=1}^{\infty} p_n=1$

例1　设小白鼠接受一定剂量的某种毒物处理后,有80%死亡,即每只小白鼠的死亡概率为0.8,生存概率为0.2。若每组各用甲、乙、丙三只小白鼠做实验,则生存数ξ的概率分布为:

生存数ξ	0	1	2	3
$P(\xi=x_n)$	0.512	0.384	0.096	0.008

由分布函数定义,离散型分布函数为:

$$F(x)=\sum_{\xi=0}^{x} P(\xi)$$

于是有:

ξ	0	1	2	3
$F(x)$	0.512	0.896	0.992	1

2)连续型分布:若存在非负函数$f(x)$,$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x<\infty$,使随机变量$\xi$取值于任一区间$(a,b)$的概率为

$$P\{a<\xi<b\}=\int_a^b f(x)\mathrm{d}x$$

则称ξ为连续型随机变量。$f(x)$称为ξ的分布概率函数,$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t$为分布函数。即连续型随机变量的取值充满某一空间,且满足一定的分布规律。这里

$$f(x)\geqslant 0, \int_{-\infty}^{+\infty} f(t)\mathrm{d}t=1$$

例如,某市成年男性的血红蛋白量就是一连续型随机变量,其概率密度函数为

$$f(x)=\frac{1}{1.22\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-14.18)^2}{2(1.22)^2}}$$

分布函数为

$$F(X)=\int_{-\infty}^{X}\frac{1}{1.22\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-14.18)^2}{2(1.22)^2}}\mathrm{d}t$$

此例中的血红蛋白量服从的是一种正态分布,其中$\overline{X}=14.18$是平均数,$s=1.22$为标准差。常见的离散型分布有二项分布、Poisson分布、超几何分布等等;连续型分布有正态分布、指数分布等等。

对随机变量的描述也可采用特征参数,而常用的有数学期望、方差。

数学期望:设有一个含量为n的样本,其观察值分别为$x_1, x_2, \cdots, x_k$,对应的频数

为 $f_1, f_2, \cdots, f_k$，则均数为：

$$\overline{X} = \frac{\sum f_i x_i}{\sum f_i} = \sum (x_i)\frac{f_i}{n}, \sum f_i = n$$

$$i = 1, 2, \cdots, k$$

当 n 充分大时，频率 f_i/n 接近于概率。受这个加权均数的启发，我们定义离散型随机变量 X 的数学期望 $E(X)$ 为

$$E(X) = \sum XP(X)$$

式中 $\sum P(X) = 1$，并要求 $E(X)$ 为一确定的数值。由此可见，数学期望就是总体均数，常简记 μ。

类似地，对具有密度函数 $f(x)$ 的随机变量，有

$$\mu = E(X) = \int_{-\infty}^{+\infty} Xf(X)\mathrm{d}X$$

这里同样要求 $E(X)$ 为一确定的数值。

方差：数学期望刻画了随机变量的集中位置，为了刻画它的变异程度，下面定义随机变量的方差。

对于离散型随机变量 X 方差为：

$$V(X) = E[X - E(X)]^2 = \sum (X - \mu)^2 P(X)$$

对于连续型随机变量 X 方差为：

$$V(X) = \int_{-\infty}^{+\infty} (X - \mu)^2 f(X)\mathrm{d}X$$

（易　东　尹全焕）