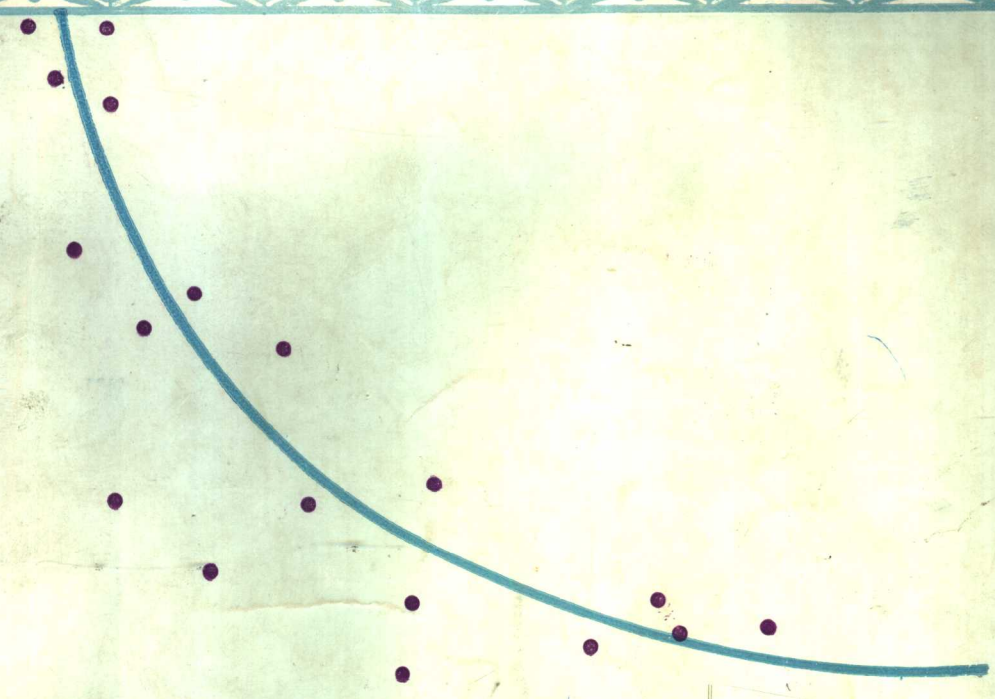
苑锡光 等编著

以电子计算机为工具的—

医用统计分析



四川科学技术出版社

以电子计算机为工具的

医用统计分析

苑锡光 等编著

四川科学技术出版社

一九八五年

内 容 简 介

本书通过各种实际例子介绍医学实践中一些常用的、基本的统计分析方法。共十五章。前六章是概率及统计学中一些必要的基础知识。第七至九章系统地介绍了对数据进行初步分析的方法。后面六章逐个地介绍多种统计方法的原理及其应用。对所需计算在列出公式后，即给出利用计算机统计程序所得计算结果，略去繁冗的计算过程，以突出各种方法的应用原理。

本书可供医学科学研究人员、高等医学院校师生及有关医学、生物科学人员参考。

序

本书是为医学科学工作者及有关生物科学工作者编写的一本关于统计分析方法的参考书，主要介绍那些宜于在电子计算机上实现的统计方法。往年，有不少需要大量、繁琐计算的统计方法，虽然他们在处理、分析资料时能提供较多的信息，但由于工具所限难以应用，使得这些统计方法较少被人了解。现在，电子计算机的普及使用为上述这些统计方法的实现提供了极为有利的条件。这种情况促使我们编写了这本书，希望能为医学科学（及有关生物科学）工作者掌握这些统计方法的原理和应用提供一些帮助。

本书的内容主要是医学实践中常用的一些基本的统计方法。另外还从医学统计工作实践中提出的一些问题出发，编入了一些内容，如列联表中的“残差分析”，方差分析中的“双因素不等例”方差分析等问题。再者，为了能正确地理解和使用这些统计方法，并为了解更多的统计方法创造条件，本书还介绍了概率论及数理统计中的一些必要的基础知识，并在附录中简单、通俗地介绍了所需数学基础。

在内容的编排上，为了有系统地掌握统计分析方法，我们采用了以概率统计基本概念为主干的编排方法（见第七至九章），看来这样做对系统地掌握有关统计方法可能是有益的。

在叙述的方式上，我们采用了直观的叙述方法，从实例出发介绍概念、原理和方法，尽量少用或不用数学论证和推导。在统计方法的叙述中尽量不涉及那些繁琐、费时的计算细节，而把重点放在方法的应用方面。着重阐明：它解决什么问题及其应用原理和使用条件，以利于读后能掌握使用统计分析这个工具。

如果阅读本书的同时，能在电子计算机上做些使用统计程序的练习，无疑是有益的。不过，即使手头没有这种条件，对阅读本书并无妨碍。总之，这是一本尽量采用直观易懂的方法介绍统计分析基本原理及如何使用的书籍。

华西医科大学祝绍琪教授仔细地阅读了本书手稿，几乎逐字逐句地，乃至对本书的编写指导思想上，都认真、细致地给我们提出了具体意见。重庆师范学院罗哲明教授对本书的系统性、科学性等问题，也提了建设性的意见。对两位教授的有益帮助，我们衷心地感谢。另外第三军医大学科研处对本书的出版提供了很大的支持。数学教研室的梁正东、蔡昌启、贺玉梅等同志也为之付出了大量劳动。在此我们一并深致谢意。

本书原是为我校各科研究生开设数理统计课程所编讲义，经整理和充实写成。其中第十三及十五章，还有附录一及二系我室张英弟讲师编写。由于我们水平有限，如有错误或缺点敬请赐教，不胜感激。

范锡光 于第三军医大学数学教研室

一九八五年六月

目　录

引言……（1）

第一章　随机变量与概率……（4）

第一节　随机变量与事件……（4）

第二节　概率与频率……（6）

第三节　古典概型……（8）

第四节　条件概率……（9）

第五节　事件的独立性……（10）

第六节　n重贝努里试验……（12）

第二章　随机变量的分布与数字特征……（14）

第一节　离散随机变量的分布……（14）

第二节　二项分布与普阿松分布……（16）

第三节　连续随机变量的分布……（20）

第四节　正态分布与指数分布……（24）

第五节　随机变量的数字特征……（26）

第六节　大数定律与中心极限定理……（31）

第三章　随机向量……（33）

第一节　离散随机向量及联合概率分布……（33）

第二节　连续随机向量及联合密度函数……（36）

第三节　边际分布……（38）

第四节　条件分布……（41）

第五节　随机变量的独立性……（42）

第六节　随机向量的数字特征……（44）

第四章　样本及抽样分布……（48）

第一节　样本……（48）

第二节　样本直方图……（49）

第三节　统计量……（50）

第四节　抽样分布……（51）

第五章　参数估计……（56）

第一节　总体数字特征的估计……（56）

第二节　最大似然法与分布参数估计……（58）

第三节　区间估计……（61）

第四节　总体率的估计……（64）

第五节　普阿松分布参数的估计……（66）

第六节　正常值范围的估计……（67）

第六章　假设检验……（69）

第一节　参数的假设检验……（69）
第二节　总体率的假设检验……（74）
第三节　总体分布的拟合检验……（76）
第四节　似然比检验法……（78）
第七章　连续变量数据初步分析（一）
——单个变量……（80）
第一节　总体数字特征的估计……（80）
第二节　总体分布的正态性检验……（82）
第三节　总体均数及方差的检验及置信区间……（86）
第八章　连续变量数据初步分析（二）
——多总体同变量……（89）
第一节　两总体同变量数据初步分析……（89）
第二节　单因素方差分析……（94）
第九章　连续变量数据初步分析（三）
——同总体多变量……(101)
第一节　总体协方差及相关系数的估计……(101)
第二节　相关系数的显著性检验……(103)
第十章　二维分类数据分析——二维列联表分析……(105)
第一节　分类变量……(105)
第二节　二维列联表……(105)
第三节　分类变量的独立性检验……(108)
第四节　残差分析……(110)
第十一章　线性回归分析……(112)
第一节　线性回归方程……(113)
第二节　线性回归模型及有关参数估计……(115)
第三节　简单线性回归分析……(117)
第四节　多元线性回归分析……(127)
第五节　逐步回归分析……(133)
第十二章　方差分析……(136)
第一节　单因素方差分析的模型……(136)
第二节　双因素方差分析（一）
——固定型交叉分组……(142)
第三节　双因素方差分析（二）
——混合型交叉分组……(149)
第四节　双因素方差分析（三）
——混合型嵌套分组……(153)
第五节　多因素方差分析……(156)
第六节　双因素方差分析（四）
——不等例固定型交叉分组……(161)

第十三章　判别分析……(167)
第一节　线性判别函数……(167)
第二节　逐步判别法……(172)
第三节　贝叶斯公式及其在判别上的应用……(175)
第十四章　Logistic 分布及其应用……(178)
第一节　Logistic分布……(178)
第二节　S型曲线拟合……(179)
第三节　Logistic判别函数……(180)
第十五章　随机点过程分析初步……(182)
第一节　随机点过程……(182)
第二节　间隔密度函数的估计……(183)
第三节　序列相关系数及其估计……(185)
第四节　自相关函数及其估计……(187)
附录一　集合及其运算……(190)
附录二　排列与组合……(193)
附录三　矩阵及其运算……(196)
附录四　ATP及SDH 酶活性数据……(216)
附录五　正常成年人肺功能数据（五项指标）……(218)
附录六　五项生化指标数据……(220)
附录七　正常成年人肺功能数据（六项指标）……(223)
习题一至四……(226)
主要参考资料……(229)
主要名词中英文对照索引……(230)

引　言

研究任何科学问题都要涉及一些对象，对它们进行观察与分析。统计学中把所涉及对象的全体叫做总体，把每个对象叫做个体。

〔例1—1〕 要研究癌症患者的ATP酶活性，则全体癌症患者就是总体，每个癌症患者即一个体。

〔例1—2〕 调查健康成年人的肺活量及其与身高、体重的关系时，健康成年人即总体，每个健康成年人为一个体。

〔例1—3〕 调查、比较某些地区居民血型的分布情况时，某一地区全体居民可做为研究的总体，每个居民则为一个体。

〔例1—4〕 考察某单位流行性感冒的发病情况，则该单位全体人员即为总体。

〔例1—5〕 研究五种蛋白质对大白鼠的营养价值时，可以把分别食用各种蛋白质的大白鼠各视为一个总体，即有五个总体。

统计学是科学方法的一个分支，粗略地说，它是研究总体的性质的，下面对此予以必要的说明。

(一)

首先，我们关心的是总体本身的性质而不是个体的；而且这里的“性质”只是那些数量方面的性质，或说是数值性质，即那些可用数字表示的性质。如例1—1中研究的可以是（全体）癌症患者ATP酶活性的大小、范围或者其活性大于70者在全体患者中所占的比例，等等；但是既不研究任一特殊患者的ATP酶活性为什么偏高或偏低，也不去探索ATP酶其它方面的特性，如生物的或化学的性质等等。再如例1—3中我们集中注意的不是那个居民血型是O型还是A型的，而是关心各种血型的人数及其在全体居民中所占的比例，等等。

为研究这些数值性质，要对个体进行观测——直接进行观察、测量或经过试验进行观测。观测的结果得到一些数，我们把这些数叫做观测值。如从癌症患者观测到ATP酶活性有62、88、71、103、82……等等一些值。又如从健康成人测得他们的肺活量为3.55，4.426，4.227……升等等。上面这些数值序列都可看做是变量所取的值。所以，总体的数值性质表现为变量，或者说他们可用变量来表示。如ATP酶活性，肺活量、身高与体重，还有例1—5中大白鼠体重的增加量等等都是表示所关心的总体性质的变量。这些变量在不同的个体上可以取不同的值。

例1—3的问题中我们关心的是各个地区各种血型的人数及其所占比例，这两者对不同地区会有不同的值，它们就是表示我们研究的性质的变量。例1—4也属这种情况。

由此可见，研究总体的性质也就是研究表示这些性质的变量以及有关变量间的关系。而统计学处理的就是对这些变量（或即总体性质）进行观测或计数得到的数据。因为各观测值是与各个体一一对应的，所以，有时也把全体观测值乃至所研究的变量称为总体。

（二）

统计学研究的总体有一重要的特征，就是这些总体，从而组成它的所有个体，都时时经受着周围环境中起各种不同作用的、大量因素的影响。例如，医学研究中涉及的对象往往是人，健康的或患病者，实验动物以及可能是它们的部分组织等等。这些对象都时时经受着自然界与社会生活环境中大量因素的影响。而这些因素大多是我们无法控制的，甚至是我们研究的对象所固有的，比如生物界的个体差异就是个突出的例子。因此这些总体的性质就表现出很大的不确定性，或随机性。比如在测量健康成年人肺活量时，即使年龄、身高、体重、性别乃至职业都相同的人，他们的肺活量也有很大差别。又如同一药物治疗同一种疾病的患者，即使患者各方面条件都很相似，疗效也会有很大差异。总体性质的不确定性表现在即使条件完全相同，对其各个个体的某一变量的观测或实验结果仍会有明显差异，或者说仍是不确定的。

各种科学研究的对象显然并不都有很大的不确定性。比如研究某种金属的电阻，可以将温度与外界压强（影响电阻的两个因素）控制为定值，测定该金属不同长度与横截面积的电阻值，从而可找出电阻与长度及横截面积间的确定的关系式。这说明，在相同的温度、压强、长度与横截面积条件下，该金属的电阻有确定的值。而统计学研究的却是缺乏这种确定性的变量。

不过我们知道，如我们只观察很少数病人，便不能正确断定某药物对某种病有多大疗效；但当观察的病人足够多时，就可能得出该药物在某种程度上有疗效的看法。又如一次生育，生男生女难以肯定，但人口调查统计表明，整个社会人口中男与女人数的比总是很接近1:1的。再如测量一个正常人与一个癌症患者的ATP酶活性，可能发现正常人的ATP酶活性高于癌症患者；而如再测另外两个人，可能结果刚好相反，但大量观测结果表明绝大多数癌症患者的ATP酶活性是高于正常人的。实验科学中大量结果表明，单独个体观测结果的不确定性而累积结果却有某种确定性的情况是常常出现的。

总之，统计学研究的总体的性质虽有不确定性，但也有它确定性的一面；而正是这确定性的东西能对那不确定的一面给出某种确定的描绘。

（三）

大量的观测（或试验）的结果能帮助我们找出总体性质的有确定性的或叫有规律性的东西。是否可以对全体个体都一一地观测到呢？

当总体是有限的，即个体总数一定时，在一些简单情况下对全部个体进行观测是可能实现的。如例1—4，某单位人员数是有限的，对其全体人员一个不漏地进行诊断，这是可能的。而如例1—3，就可能遇到困难。因为不时有出生，不时也有死亡。既使指定某年、月、日，居民人数可认为是一确定的数，如果人口数量太大，也难以及时完成普查血型的任务。而对例1—1，如果指的是全人类中的癌症患者，而且有确定的日期（比如说今天），那么其个体的总数可以说是有一定值的，但显然是无法完成这种“普查”任务的。

以上这三种情况，不管个体总数有多大，相应的总体还是存在的，或是可设想其存在的。但如在例1—1中将未来的癌症患者也包括在内，则不只是个体总数无法确定，而且还有部分总体尚不存在！像这种总体叫做假设的总体。例1—5中，我们可能得到这样结论：

“甲种蛋白质的营养价值最高”，这个结论是对所有大白鼠而言的，包括那些并未食用这种蛋白质甚至尚未出生的在内！这个总体也是属假设的一类。显然对假设的总体要想对个体逐个观测一遍是完全不可能的。事实上，像例1—4那样的问题，我们研究的目的，往往是想从这一单位的调查结果推断更多单位的情况，所以研究的总体一般地并不是限于这一个单位的！总之，实践中只有借助于“大量”的观测来研究我们的问题才是现实可行的。但是这个量再大，也还是总体的一部分，如何由这一部分去合理地估计或推断总体的性质呢？这正是统计学的主要内容。

本书介绍医学研究中常用的几种基本的统计分析方法(第七至十五章)。为了了解这些方法的原理，在第一至三章中介绍了概率论中的部分基础知识，在第四至六章中介绍了数理统计的一些初步知识（这两部分并非系统的概率论和统计学，对此请参阅有关专著）。如已熟悉概率和统计的基本概念，可直接阅读第七至十五章，必要时查阅第一至六章有关内容。

第一章　随机变量与概率

第一节　随机变量与事件

总体的性质有不确定性，表示总体性质的变量也就有不确定性。如引言中所述，在测量健康成年人的肺活量时，即使年龄、身高、体重、性别与职业都相同的人，肺活量也会有很大差别。在测量之前肯定不了任一个体的肺活量值。又如，即使同父母所生子女，已知父母血型情况下，未测定之前也肯定不了任一子女的血型。所以这种变量的不确定性表现在：即使在相同条件下对总体中各个个体进行观测时，在观测之前不能肯定变量将取那一个值。我们把这种变量叫随机变量。

一、随机变量所取的值与样本空间

随机变量所取的值是对个体性质进行观测的结果。这种结果可分为两类，一类是测量或计数的结果，是可以进行某些运算的数；这时随机变量取的值就是各种数，小数、整数等。

另一类是把个体按某属性进行分类的观测结果，例如查明某人的血型为A型，又如诊断某人患有（或未患）某种疾病，等等。这时我们仍把观测结果叫做随机变量。这当然是最一般意义下的“变量”。并把结果适当地用一些数或符号去表示。如测定血型的结果可用数字1、2、3及4或用符号x_1、x_2、x_3及x_4分别表示血型为O、A、B及AB型。再如可用0及1分别表示某人未患及患有某种疾病，等等。这样做会为我们在研究形式上带来方便；只要注意到这样用的数字有时只有符号意义，是不能施以运算的。研究某一具体问题时，对个体的观测结果叫做样本点，全体样本点的集合叫做所研究的问题的样本空间。从随机变量的角度来看，观测结果是随机变量所取的值，所以样本点又叫样本值，而样本空间也就是：某一随机变量所有可能取的值的集合。以上对样本空间的两种定义实际上完全相同，为了阐述的方便将在不同的场合使用它们。

如用ξ表示引言中例1—2的变量——健康成年人的肺活量，显然ξ是一随机变量。则ξ可能取的值为大于零的实数，可以估计ξ有个最大限度，比如说15升。于是ξ取值的范围就在区间（0，15）内，这个区间内有无穷多个点，每个点与其他各点都有不同的值。它的样本空间S即为$S=\{x \mid 0<x<15\}$，其中x表示ξ的样本值或样本点。不过对这种情况我们一般不必写出其最大值，只把样本空间定为正的实数轴，即$S=\{x \mid 0<x<\infty\}$即可，这可更具有一般性。

又如例1—3的观测结果有四种血型，故此问题的样本空间即O、A、B及AB血型组成的集合。也可用η表示这个问题中的随机变量（血型），并用x_1、x_2、x_3及x_4分别表示O、A、B及AB血型，则η的样本空间即为$S=\{x_1, x_2, x_3, x_4\}$。而例1—4的样本空间即$\{0,1\}$，它只有两个点。

由有限个或可数个值组成的样本空间叫离散的，相应的随机变量叫离散随机变量。如例1—3中的η及例1—4中的随机变量都是离散随机变量。如样本空间为整个实数轴或其一个

区间则此样本空间叫连续的，相应的随机变量叫连续随机变量。如例 1 — 2 中的ξ即连续随机变量。

二、事　件

随机事件是样本空间的子集，简称事件。如例 1 — 2 中，$E=\{x \mid 3.5\leqslant x<4\}$ 即一事件，它表示随机变量ξ的值（健康成年人肺活量）落在区间〔3.5, 4) 内，这区间是样本空间$(0,\infty)$的一个子集。事件E也可记做$(3.5\leqslant\xi<4)$。

又如例 1 — 3 中，样本空间 $\{x_1, x_2, x_3, x_4\}$ 的任一子集都是一事件。如 $E_1=\{x_1\}$ 及 $E_6=\{x_2, x_3\}$，等等都是事件。E_1 表示随机变量η（血型）取值为 x_1（O型）的事件，E_1 也可记为$(\eta=x_1)$。E_6表示η取值为x_2或x_3（即A或B型）的事件，也可记为$(\eta=x_2, x_3)$。对离散样本空间来说，每个样本点都是一个事件。

在一次观测中，如随机变量的观测值为某一事件的样本点，我们就说这个事件“出现”或“发生”了，否则就说这个事件没有发生。如测得一健康成年人肺活量为3.75升，则上述事件E发生了。如测得某居民的血型为A型，则上述事件E_6发生了，而事件E_1却未发生。

在每次观测中必定出现的事件叫必然事件，如例 1 — 3 血型测定中，事件 $S=\{x_1, x_2, x_3, x_4\}$ 就是必然事件。在各次观测中必不出现的事件叫不可能事件，如例 1 — 2 测定肺活量时，“健康成年人肺活量 = 100升”就是不可能事件。一般地我们规定样本空间 S 及空集ϕ都是事件，于是S就是必然事件而ϕ则为不可能事件。

三、事件间的关系

对样本空间S中的事件E、F、……等等，我们定义一些关系及事件。

如E中每个样本点都包含在F中，即$E\subset F$或$F\supset E$，则称事件 F 包含事件 E，这时事件E发生必导致事件F发生，但F发生不一定导致E的发生。如前面例 1 — 3 中用E表示事件“血型为A型或B型”，即$E=\{x_2, x_3\}$，用F表示事件$\{x_2, x_3, x_4\}$，则$E\subset F$，但如有一人血型为AB型，则F发生，但E未发生。

如有$E\subset F$，且$F\subset E$也成立，则称 E 等于F，即$E=F$。相等的事件必同时发生。

对于事件 E，由样本空间中所有不包含在 E中的样本点组成之事件叫做 E 的对立事件或逆事件，记为$\overline{E}$。如上例中$\overline{E}$即$\{x_1, x_4\}$。显然E也是$\overline{E}$的对立事件，所以它们是互逆的。

用EF 表示所有同时属于E及F 的样本点的集合。事件 EF 叫做F 与 F 的交（或积），也记做$E\cap F$，它表示E与F同时发生的事件。

用 $E+F$ 表示所有 E与 F 的样本点组成的集合（如有共同的样本点只算一个），叫做 E 与 F 的并（或和），也记做$E\cup F$，它表示E或F至少有一个发生的事件。如前面例 1 — 3 中，用E表示事件$\{x_2, x_3\}$，F表示$\{x_3, x_4\}$，则$EF=\{x_3\}$，$E+F=\{x_2, x_3, x_4\}$。

如果 E 与 F 的交是空集，即 $EF=\phi$，则称 E与 F 互不相容，这表示它们不可能同时发生。对立与不相容不同。对立事件必不相容，不相容者不一定是对立的。注意到$E+\overline{E}=S$，所以一次测观的结果，对立事件中必有一个发生（且仅有一个），但不相容者可能同时都不发生。如例 1 — 3 中 $E=\{x_2, x_3\}$，则$\overline{E}=\{x_1, x_4\}$，而$G=\{x_1\}$就与 E 是互不相容的。每次观测结果非x_2或x_3即x_1或x_4，故E与$\overline{E}$必有一个（且仅有一个）发生，但E与G可能都不发生。

用$E-F$表示包含在E中而不在F中的样本点的集合，这个集合叫做E与F的差，它表示E发生而F不发生的事件。例如$S-E=\overline{E}$。

第二节 概率与频率

那一事件在某次观测中会出现是不能确定的，但各种事件出现的可能性却是可以确定的，而且往往是有大小之分的。比如在例1—3所述血型普查工作中，从一般工作经验可以知道，事件（O型）出现的可能性比事件$E_4=\{x_4\}$（AB型）出现的可能性要大。又如例1—2所说检查成年人肺活量时，出现小于3.5升(事件A)的可能性比出现在3.5～4.5升之间（事件B）的可能性要小些，等等。

概率是对事件出现可能性大小的一种度量，事件A出现的概率记为$P(A)$。概率是与频率紧密联系着的。在对某一随机变量的观测过程中，某事件出现的次数m与观测（或试验）总次数n的比，叫该事件出现的频率f，即

$$\text{某事件出现的频率 } f=\frac{\text{该事件出现的次数}}{\text{观测总次数}}=\frac{m}{n}$$

显然，频率最大不会超过1，最小不会小于零，即$0\leqslant f\leqslant 1$。我们看一个例。投掷一枚硬币，看它落下后出现正面还是反面。投掷结果只有两个（样本点）即**正**与**反**。假设这硬币是理想的均匀对称的，于是**正**与**反**两个事件出现的可能性是相同的。在任一次投掷中，出现那个事件是不能事先确定的，比如说第一次事件**正**出现了，第二次就不一定是事件**反**出现。不过有人做过试验，据报道投掷到1500次时，事件**正**出现的频率为0.4927；投掷到2800次时得事件**正**出现的频率为0.5018；还有人投掷到24000次，得频率0.5005。这些试验结果表明，事件**正**出现的频率，随着试验次数的增加，它在0.5的附近摆动，并且摆动的幅度越来越小。可见这频率的稳定值是客观存在的，它就是事件**正**出现可能性的一种度量。

一般地，当试验（或观测）次数很大时，如事件A出现的频率稳定地在某一数值p附近摆动，而且随着试验次数的增加，摆动的幅度越来越小，则称数值p为事件A出现的概率，记为$P(A)=p$。数值p就是事件A出现可能性大小的一种度量。这是概率的统计定义。如上述投掷硬币例中即有$P(\text{正})=0.5$。

从随机变量来说，任一事件的概率也就是随机变量取某些值的概率。如在本节开始一段所说的例1—2中两种可能性的比较即可记为

$$P(\xi<3.5)<P(3.5<\xi<4.5);$$

而例1—3中两种血型出现可能性大小亦可记为$P(\eta=x_1)>P(\eta=x_4)$，等等。

随机变量的最重要的特征就是，对它样本空间内任一事件都对应着一个确定的概率。随机变量有随机性，对任一个体它将取那一个值，事先不能确定；但是它取那一个值或在那一区间内取值的概率却都是确定的。所以说，概率为随机变量的不确定性给出了一种定量的描述。

上述概率定义，也提供了近似地计算概率的一种方法。如前面例1—3中的问题（用η表示血型），假设调查了某地区$n=500,000$个居民的血型，结果如表1—1所列。如果调查之前充分地考虑了调查对象的代表性，则根据观测次数足够大，实践中可以把所得频率做为概率来使用，即可以认为

表 1 — 1　　某地居民血型分布

血　　型 η	x_1(O型)	x_2(A型)	x_3(B型)	x_4(AB型)
人　　数 m	152,450	137,550	161,650	48,350
频　率 $f=\frac{m}{n}$	0.3049	0.2751	0.3233	0.0967

$$P(\eta=x_1)=0.3049,\quad P(\eta=x_2)=0.2751$$
$$P(\eta=x_3)=0.3233,\quad P(\eta=x_4)=0.0967$$

显然，这种计算概率的方法是很不经济的，而且往往是不可能的。统计学的任务之一就是通过适量的观测（或试验）对所关心的某些概率给出合理的估计及推断。另外，在某些假设或已知理论的条件下，还可推导出一些事件的概率值。为了区别，我们将把由观测或试验得到的概率称为经验概率。

下面介绍概率的一些性质。由频率取值的范围可知，在给定的样本空间S中：

（1）对任何事件A，必有$0\leqslant P(A)\leqslant 1$；

（2）对必然事件S及不可能事件ϕ，分别有$P(S)=1$及$P(\phi)=0$。

（3）概率具有可加性。即对r个两两互不相容的事件A_1、A_2、$\cdots A_r$，它们的和事件的概率等于各个事件概率的和，写成式子就是

$$P(A_1\cup A_2\cup\cdots\cup A_r)=P(A_1)+P(A_2)+\cdots+P(A_r)。$$

例如，在$r=2$的情况下，假设进行了n次观测，其中A_1发生了m_1次，A_2发生了m_2次。由于A_1与A_2互不相容，可知$A_1\cup A_2$发生了m_1+m_2次，于是事件$(A_1\cup A_2)$发生的频率为

$$f=\frac{m_1+m_2}{n}=\frac{m_1}{n}+\frac{m_2}{n}$$

即A_1与A_2各自发生的频率的和。所以应有

$$P(A_1\cup A_2)=P(A_1)+P(A_2)$$

同理对任意有限个两两互不相容的事件来说可加性也成立。另外，由此可推知事件A的逆事件$\bar{A}$的概率为

$$P(\bar{A})=1-P(A)$$

这是因为A与$\bar{A}$不相容，而且$A\cup\bar{A}=S$，故知$P(A)+P(\bar{A})=P(A\cup\bar{A})=P(S)=1$。

例如，例 1 — 3 中设用E_1、E_2、E_3及E_4分别表示$\eta=x_1$、x_2、x_3及x_4等事件，并设$A=E_1\cup E_3$，$B=E_1\cup E_2$，则由表 1 — 1 可知

$$P(A)=P(E_1)+P(E_3)=0.6282,$$
$$P(\bar{A})=1-0.6282=0.3718,$$
$$P(\bar{B})=P(E_3)+P(E_4)=0.42,\ \text{等等。}$$

再者由$\sum_{i=1}^{4}P(E_i)=P(S)=1$可知对样本点为有限的样本空间而言，所有单个样本点事件的和的概率必为 1 。

第三节 古 典 概 型

如果样本空间S只有有限个样本点，而且每个样本点出现的可能性相同，我们讨论其中事件的概率问题，情况就比较简单。这种等可能概型叫做古典概型。设$S=\{x_1, x_2, \cdots, x_n\}$，因每个样本点都是一个事件，而且它们出现的可能性相同，所以有

$$1=P(S)=P(x_1)+P(x_2)+\cdots+P(x_n)=nP(x_i)$$

$$P(x_i)=\frac{1}{n}; \qquad i=1, 2, \cdots, n$$

如果事件A由m个样本点组成，则由概率的可加性知

$$P(A)=\frac{m}{n}=\frac{A\text{ 中包含的样本点数}}{\text{样本点总数}}$$

这就是古典概型中事件A的概率的计算公式。

〔例 1 — 6〕. 一袋子里装有 8 个球，5 个白的，3 个黑的。这些球的大小、重量乃至表面的光洁度都相同，用手去摸时无法分辨是那一个。于是经适当混合后，摸出任一个球的可能性都相同（以下再谈到摸球问题时都作同样假设，不再重复），所以摸到任一球的概率都是1/8。试求摸到一个黑球的概率。

我们把 8 个球编上号，白球编为①至⑤号，黑球编为⑥至⑧号。于是样本空间由①至⑧等八个样本点组成。记摸到黑球为事件A，则$A=\{⑥, ⑦, ⑧\}$。故$P(A)=\frac{3}{8}$。

〔例 1 — 7〕 一个骰子各面分别涂有 1、2、3、4、5 及 6 个点。假设它是密度均匀的正六面体，投掷它时各面出现（朝上）的可能性相等（以下说到骰子时都作同样假设不再重复）。显然这样本空间由六个样本点组成，易见

$$P\text{（出现 3 点）}=\frac{1}{6}, \quad P\text{（出现 3 或 6 点）}=\frac{1}{3}$$

$$P\text{（出现任何一个点）}=1, \quad P\text{（同时出现 3 点与 6 点）}=0。$$

〔例 1 — 6（续）〕 如果一次摸两个球出来，摸到一对黑球的概率是多少？这个问题与前面例 1 — 6 中提出的主要不同点是：这里我们观测的对象是两个球为一组的。这样的组共有$C_8^2=28$种可能，它们是

①②	①③	①④	①⑤	①⑥	①⑦	①⑧
②③	②④	②⑤	②⑥	②⑦	②⑧	③④
③⑤	③⑥	③⑦	③⑧	④⑤	④⑥	④⑦
④⑧	⑤⑥	⑤⑦	⑤⑧	⑥⑦	⑥⑧	⑦⑧

注意，组合是没有顺序性的，①②与②①是同一个组。所以样本空间由上列 28 个样本点组成，其中各样本点出现的可能性相同。两个都是黑球的组（样本点），即两个编号都是 6、7 或 8 的那些组，共有 3 个。用组合数计算即$C_3^2=3$。所以

$$P\text{（出现一对黑球）}=\frac{C_3^2}{C_8^2}=\frac{3}{28}$$

上面两种问题（摸球及掷骰子）因其简单易懂，常被用来说明一些概率中的基本概念。

第四节 条 件 概 率

上节中列举了有关概率的一些简单计算，这些计算都是在一定条件下进行的，条件不同时概率值就可能有变化。比如例 1 — 6 （续）中，如分两次摸这些球，第一次摸出一球，不再放回。第一次摸得白球的概率为5/8，在第一次摸得白球条件下（此时袋中还有 4 个白球和三个黑球），第二次再摸得白球的概率就变为4/7了。同一个“摸得白球”的概率，有了不同的值，这是因为在第二种情况下多了一个条件：先已摸出了一个白球。这种在已知一事件A发生的条件下，另一事件B发生的概率叫条件概率，记为$P(B|A)$。

〔例 1 — 6 （续）〕 分两次摸球，每次摸出一球，不再放回。记第一次摸得白球的事件为A，第二次摸得黑球的事件为B。（注意这里有次序问题！）试分别计算事件$C=A\cap B$的概率，事件A的概率，事件B的概率及在事件A发生的条件下事件B发生的概率$P(B|A)$。

先考虑样本空间。由题意，观测是有先后顺序的，两个球为一个对象。所有可能摸到的结果就有（前后二数分别表示第一、二次摸到的球的编号）

(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)
(2,1)	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)
(3,1)	(3,2)	(3,4)	(3,5)	(3,6)	(3,7)	(3,8)
(4,1)	(4,2)	(4,3)	(4,5)	(4,6)	(4,7)	(4,8)
(5,1)	(5,2)	(5,3)	(5,4)	(5,6)	(5,7)	(5,8)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,7)	(6,8)
(7,1)	(7,2)	(7,3)	(7,4)	(7,5)	(7,6)	(7,8)
(8,1)	(8,2)	(8,3)	(8,4)	(8,5)	(8,6)	(8,7)

即共 $8\times 7=56$个样本点，各样本点出现的可能性是相等的。

符合事件$C=AB$的样本点是第一个数为 1 至 5 且第二个数为 6 至 8 的那些，共有15个。故$P(C)=P(AB)=\frac{15}{56}$。

再考查$P(A)$。符合事件A的样本点是第一个数为 1 至 5 而不论第二个数是几的那些，共有35个，故$P(A)=\frac{35}{56}=\frac{5}{8}$。

同样地，符合事件B的样本点是第二个数是 6，7 或 8 而不论第一个数是几的那些，共有21个，故$P(B)=\frac{21}{56}=\frac{3}{8}$。

事件$(B|A)$与以上三者不同，它要求在事件A发生的条件下来考虑问题。也就是限制在前五行中去考虑，故此时样本空间可视为只包含前五行共35个样本点。其中符合事件$(B|A)$的样本点，即第二个数为 6、7 或 8 的共有15个。故$P(B|A)=\frac{15}{35}=\frac{3}{7}$。

当然，这一概率能直接计算。如第一次摸得白球，袋中就剩下 4 个白球与 3 个黑球，所以$P(B|A)=\frac{3}{7}$。但前面的分析可以帮助我们找出这几个概率之间的关系：

$$P(B|A)=\frac{15}{35}=\frac{15/56}{35/56}=\frac{P(AB)}{P(A)}$$

这个等式虽是从特例推出，但是有普遍意义的，所以我们

定义 设A和B是样本空间S中两个事件，且$P(A)>0$，则在事件A已发生条件下事件B发生的条件概率定义为

$$P(B|A)=\frac{P(AB)}{P(A)} \qquad (1-1)$$

上式的另一形式：

$$P(AB)=P(A)P(B|A), \qquad (1-2)$$

叫做概率的乘法定理。而且如$P(B)>0$，还可定义$P(A|B)$，这时有

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

第五节 事件的独立性

上节中我们通过实例看到事件A发生条件下，事件B发生的概率$P(B|A)$，与没有这个条件时事件B发生的概率$P(B)$不相等了。这说明一事件发生与否，对另一事件发生的概率可能有影响。在上节例1—6（续）中采用了无放回摸球，如果采用有放回摸球，情况又如何呢？

〔例1—8〕 在例1—6（续）中采用有放回摸球，求：

（1）在第一次摸得白球（事件A）条件下，第二次摸出黑球（事件$B|A$）的条件概率；

（2）第二次摸出黑球（事件B）的概率。

这里与例1—6（续）中不同之处为“有放回摸球”，即第一次摸得几号球，第二次仍有可能摸到它。所以就多了8个样本点，即（1，1）、（2，2）、…、（8，8）。于是有

（1）$P(A)=\frac{40}{64}=\frac{5}{8}$，$P(AB)=\frac{15}{64}$，$P(B|A)=\frac{15/64}{5/8}=\frac{3}{8}$

（2）$P(B)=\frac{24}{64}=\frac{3}{8}$

注意这里的$P(B|A)=P(B)$，即事件A是否发生，对事件B发生的概率没有影响。显然，这是因为我们采用了“有放回地摸球”，第二次摸球时袋中两种球各占的比例与第一次者相同，当然第一次的事件对第二次者就没有影响了。在这种情况下我们称这两个事件是统计独立的，简称独立的。即我们

定义 在样本空间S中的两个事件A与B，如满足下列任一关系，则称它们是相互独立的，或简称独立的。

（1）$P(A|B)=P(A)$

（2）$P(B|A)=P(B)$

（3）$P(AB)=P(A)P(B)$ （1—3）

此三式既可用于检验两事件是否相互独立，也可用于概率计算。特别第三式，它表示如二事件相互独立，则它们同时发生的概率等于各自概率的乘积。