生物科学
生物技术
系　列

Biostatistics

普通高等教育“十二五”规划教材

精品课程教材

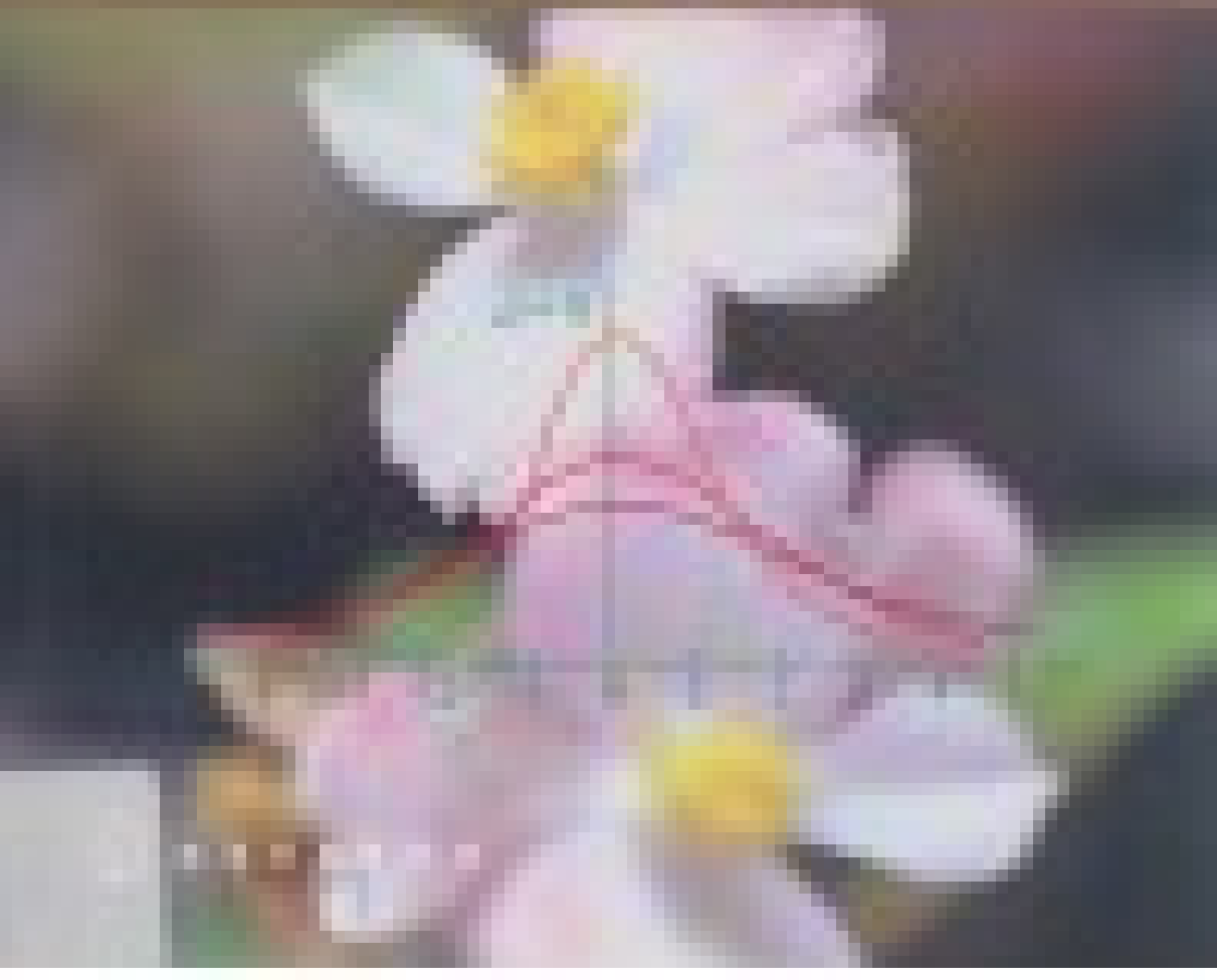
生物统计学

叶子弘　陈春　主编

化学工业出版社

Biostatistics

生物统计学



普通高等教育“十二五”规划教材　精品课程教材

生物统计学

叶子弘　陈　春　主　编
郑荣泉　张文英　副主编

化学工业出版社
·北京·

图书在版编目（CIP）数据

生物统计学/叶子弘，陈春主编．—北京：化学工业出版社，2011.12
普通高等教育“十二五”规划教材．精品课程教材
ISBN 978-7-122-12784-6

Ⅰ．生…　Ⅱ．①叶…②陈…　Ⅲ．生物统计-高等学校-教材　Ⅳ．Q-332

中国版本图书馆 CIP 数据核字（2011）第 231368 号

责任编辑：刘　畅　赵玉清　　文字编辑：李　瑾
责任校对：陶燕华　　装帧设计：尹琳琳

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）
印　　刷：北京永鑫印刷有限责任公司
装　　订：三河市万龙印装有限公司
787mm×1092mm　1/16　印张 17½　字数 456 千字　　2012 年 2 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：35.00 元

本书编写人员

主　　编： 叶子弘　陈　春

副 主 编： 郑荣泉　张文英

编写人员：（按姓氏汉语拼音排序）

陈　春（中国计量学院）

崔海峰（中国计量学院）

王　江（台州学院）

王　玉（浙江中医药大学）

王忠华（浙江万里学院）

叶子弘（中国计量学院）

张文英（长江大学）

郑荣泉（浙江师范大学）

前　言

生物统计学是一门探讨如何从不完整的信息中获取科学可靠的结论从而进一步进行生物学实验研究的设计、取样、分析、资料整理与推论的科学，不仅在传统生物学、医学、药学和农学中被广泛应用，在现代分子生物学、各种组学研究中也是不可缺少的工具。无论是试验设计、数据描述，还是试验结果的科学分析和推断，都需要以生物统计学的相关理论和方法为依据。因此，生物统计学是生命领域不同专业学生都应该掌握的一门重要工具课，是许多高等院校生物学、医学、药学和农学等专业的必修课程。生物统计学的理论性和实践性较强，且涉及的内容、公式和抽象概念较多，需要一定的数学基础和较强的逻辑推理能力，相对其他专业课程，具有一定难度。此外，随着生物技术的不断进步，统计分析数据量大增，仅依靠计算器进行数据分析和计算往往难度很大，且非常耗时。因此，本书根据生物统计学发展和教学应用特点，加强案例分析，提供SAS、DPS等实用统计分析软件的简要介绍并用例题加以说明，以便读者能够结合案例了解和掌握各种常用统计方法，并能够使用现代统计工具帮助完成相关分析。

本书由编写组集体撰写完成，具体章节分工为：第一章，叶子弘；第二章，王玉；第三、第四章，王江；第五、第十章，郑荣泉；第六章，王忠华；第七章，崔海峰、张文英；第八章和附录，陈春；第九章，张文英。全书由叶子弘和陈春统稿，叶子弘定稿。在本书的出版过程中，得到了化学工业出版社的大力支持，特此感谢。

由于作者水平的限制，书中疏漏和错误之处在所难免，敬请同行专家和读者批评指正，不胜感激。

叶子弘

2011年10月于杭州

目　录

第一章　导　　论

生命科学是一门实验科学。随着生物学的不断发展，对生物体的研究和观察已不再局限于定性的描述，而是需要从大量调查和测定数据中，应用统计学方法，分析和解释其数量上的变化，以正确制订试验计划，科学地对试验结果进行分析，从而做出符合科学实际的推断。而且，近年来分子生物学技术、测序水平等不断提高，为生物学研究带来了海量的数据资料，要对这些数据资料进行整理、分析，并得出科学合理的结论，均离不开统计工具的支持。国内外许多调查研究表明，科技期刊论著中统计学误用率相当高，工欲善其事，必先利其器。因此，有必要对统计知识和相关工具进行培训，使生命科学领域的教学研究人员或从事生物学相关工作的人员具有统计意识，能够进行合理的实验设计，选用适宜的统计方法，得出正确的统计推断。

一、生物统计学概论

1. 生物统计学的概念

统计学（statistics）：是一门数据（data）分析的科学，是研究数据的取样、收集、组织、总结、分析和表达的科学方法。统计的本业是消化数据，并产生有营养的结果——信息。统计学需要运用到大量的数学知识，数学为统计理论和统计方法的发展提供基础。但是，不能将统计学简单等同于数学。数学研究的是抽象的数量规律，统计学则是研究具体的、实际现象的数量规律；数学研究的是没有量纲或单位的抽象的数，统计学研究的是有具体实物或计量单位的数据；统计学与数学研究中所使用的逻辑方法不同，数学研究使用的主要是演绎法，统计学则是演绎法与归纳法相结合，占主导地位的是归纳。根据研究领域和研究对象，统计学又分为数理统计、经济统计、生物统计、医学统计、卫生统计等。

生物统计学（biostatistics）是数理统计在生物学研究中的应用，是用数理统计的原理和方法来解释和分析生物界各种现象和试验调查资料的科学。用统计学方法研究生命的学科，研究生物群体个体间的变异性和对生物性状观察过程中的误差进行研究。生物统计学不仅提供如何正确地设计科学试验和收集数据的方法，而且也提供如何正确地整理、分析数据，得出客观、科学的结论的方法。其主要目的是培养学生具有对试验资料进行统计分析处理的能力。生物统计学有助于探索生命科学内在的数量规律性。目前，生物统计学已经在持续发展与环境保护、资源保护与利用、生态学研究、分子生物学研究、高科技农业研究、生物制药技术、流行病规律研究与探索、数量遗传学研究、生物信息学研究等生命科学的分支领域有了广泛的应用。生物统计学作为统计学的一个分支，自身拥有一整套成熟的理论和应用体系，并在快速发展之中。

2. 生物统计学的主要内容

试验设计和统计分析是生物统计学的主要内容。

试验设计就是设计试验的过程，使得收集的数据适合于用统计方法分析，得出有效的和客观的结论。在研究工作进行之前，根据研究项目的需要，应用数理原理，作出周密安排，力求用较少的人力、物力和时间，最大限度地获得丰富而可靠的资料，通过分析得出正确的结论，明确回答研究项目所提出的问题。因此，任一试验问题都存在试验的设计和数据的统计分析，

二者是紧密相连的，因为统计分析方法依赖于试验所用的设计。在工农业生产和科学研究中，经常需要做试验，以求达到预期的目的。例如在工农业生产中希望通过试验达到高质、优产、低消耗，特别是新产品试验，未知的东西很多，要通过试验来摸索工艺条件或配方。科学合理的试验设计可以避免系统误差，控制、降低试验误差，无偏估计处理效应，从而对样本所在总体做出可靠、正确的推断。

重复、随机化和局部控制是试验设计的三个基本原则。重复是指基本试验的重复进行，通过重复使得试验误差可估，增加重复的次数可提高检测处理间差异的能力。随机化是指抽样或配置处理时必须使总体中任何一个个体都有同等的机会被抽取进入样本以及样本中任何一个个体都有同等的机会被分配到任何一个试验单元中。随机化保证了试验误差估计的有效性，减小主观判断对处理配置的影响。局部控制是用来提高试验精确度的一种方法。一个区组是一组同质的试验单元。

生物统计的主要作用是：①提供整理和描述数据资料的科学方法，确定某些性状和特征的数量特征。合理地进行调查或试验设计，科学地整理、分析所收集到的资料是生物统计的根本任务。②判断试验结果的可靠性，分析现象间的关系。例如检测了不同年龄居民的人体脂肪含量，通过相关分析，可以判断年龄与脂肪含量之间的关系，通过统计检验判断这种分析结果的可信度。③提供样本推断总体的方法。例如，想了解当代大学生的身高状况。由于大学生人数很多，不可能穷尽。因此，只能通过抽样进行分析，分析样本的身高状况来推断总体，可推断当代大学生的平均身高，男大学生与女大学生的平均身高以及二者之间是否有显著差异等等。④提供试验设计的一些重要原则。例如，要分析不同种植密度对水稻产量的影响。在进行试验设计时，不仅要考虑种植密度水平、产量测量方法和评价指标的确定，还要考虑水稻品种、种植地块条件、栽培措施等可能造成试验误差的来源，分析误差特性，适宜地采用重复、随机化和设置区组等措施来减小误差的影响。

3. 统计学发展概况

人类开始统计实践的时间虽然很早，但是统计学成为一门系统的学科，却是近代的事情，距今只有 300 余年的短暂历史。统计学的发展大致可划分为古典记录统计学、近代描述统计学和现代推断统计学三个阶段。

(1) 古典记录统计学　古典记录统计学的形成时间大致在 17 世纪中叶至 19 世纪中叶。由于天文学研究和政治科学的需要，初步建立了统计研究的方法和规则，并将概率论引进统计学，逐渐成为一项较成熟的方法。最初卓有成效地把古典概率论引进统计学的是法国天文学家、数学家、统计学家拉普拉斯。因此，后来比利时统计学家、数学家和天文学家凯特勒指出，统计学应从拉普拉斯开始。这一阶段还发展了大数定律，进行了大样本推断的尝试，建立了最小二乘法和高斯分布。这一阶段的主要代表性人物如下。

① 拉普拉斯 (P. S. Laplace)，1749—1827 年，法国天文学家、数学家、统计学家。他最早系统地把概率论方法运用到统计学研究中，建立了严密的概率数学理论，并应用到人口统计、天文学等方面的研究上，推广了概率论在统计中的应用。此外，他还明确了统计学的大数法则，进行了大样本推断的尝试。尽管其方法和结果还相当粗糙，但在统计发展史上，他利用样本来推断总体的思想方法，为后人开创了一条抽样调查的新路子。

② 高斯 (Gauss)，1777—1855 年，德国数学家。正态分布理论最早由法国数学家德莫佛 (De Moiver) 于 1733 年发现，后来高斯在进行天文观察和研究土地测量误差理论时又一次独立地发现了正态分布 (又称常态分布) 的理论方程，提出了“误差分布曲线”，后人为了纪念他，将正态分布也称为 Gauss 分布。1798 年，高斯完成最小二乘法的整个思考结构，正式发表于 1809 年。

（2）近代描述统计学　近代描述统计学形成期大致在19世纪中叶至20世纪上半叶。由于这种“描述”特色由一批原是研究生物进化的学者们提炼而成，因此历史上称他们为生物统计学派。“回归”的概念、卡方检验、回归与相关等均是在这一阶段提出和发现的。这一阶段的代表性人物如下。

① 高尔顿（F. Galton），1822—1911年，英国生物学家、统计学家。1882年Galton开设“人体测量实验室”，测量9337人的资料，探索能把大量数据加以描述与比较的方法和途径，引入了中位数、百分位数、四分位数、四分位差以及分布、相关、回归等重要的统计学概念与方法。1889年，高尔顿把总体的定量测定法引入遗传研究中，发表第一篇生物统计论文《自然界的遗传》。1901年创办了“Biometricka”（生物统计学报）杂志，首次明确“Biometry”（生物统计）一词。所以后人推崇高尔顿为生物统计学的创始人。

② 皮尔逊（Karl Pearson），1857—1936年，英国数学家、哲学家、统计学家。皮尔逊将生物统计学上升到通用方法论的高度。他探索了变异数据的处理，首创了频数分布表与频数分布图，提出了分布曲线。1900年，皮尔逊独立地又重新发现了χ^2分布，并提出了有名的“卡方检验法”。后经R·费歇尔补充，成为了小样本推断统计的早期方法之一。皮尔逊进一步发展了回归与相关的概念，得出至今仍被广泛使用的线性相关计算公式、回归方程式以及回归系数的计算公式。此外，在1897～1905年，皮尔逊还提出复相关、总相关、相关比等概念，不仅发展了高尔顿的相关理论，还为之建立了数学基础。

（3）现代推断统计学　现代推断统计学形成时间大致是20世纪初叶至20世纪中叶。人类历史进入20世纪后，无论社会领域还是自然领域都向统计学提出更多的要求。各种事物与现象之间繁杂的数量关系以及一系列未知的数量变化，单靠记录或描述的统计方法已难以奏效。因此，相继产生“推断”的方法来掌握事物总体的真正联系以及预测未来的发展。从描述统计学到推断统计学，是在农业田间试验领域中完成的。因此，历史上称之为农业试验学派。在这一阶段，发展了t检验与小样本思想、抽样分布、方差分析、试验设计等思想和方法，完善了SPSS、SAS等统计软件。这一阶段的主要代表性人物如下。

① 戈塞特（W. S. Gosset），1876—1937年，英国统计学家。戈塞特长期从事实验和数据分析工作中，发现并提出了t分布，在1908年以“Student”笔名发表此项结果，故后人又称它为“学生氏分布”。后来，戈塞特又连续发表了“相关系数的概率误差”（1909年）、“非随机抽样的样本平均数分布”（1909年）、“从无限总体随机抽样平均数的概率估算表（1917年），等等”，这些论文的完成，为“小样本理论”奠定了基础。由于戈塞特开创的理论使统计学开始由大样本向小样本、由描述向推断发展，因此，有人把戈塞特推崇为推断统计学（尤其是小样本理论研究）的先驱者。

② 费歇尔（R. A. Fisher），1890—1962年，英国统计学家。费歇尔非常强调统计学是一门通用方法论，提出了假设无限总体的统计思想。他于1915年对相关系数的一般公式作了论证，1918年提出了方差和方差分析的概念并在1925年对方差和协方差分析进行了完整的叙述，1924年对t分布、χ^2分布和Z分布加以综合研究，1926年提出了试验设计的基本思想和方法，1938年与耶特斯合编了《F分布显著性水平表》。费歇尔在统计发展史上的地位是显赫的。这位多产作家的研究成果特别适用于农业与生物学领域，但他的影响已经渗透到一切应用统计学的领域，由此所提炼出来的推断统计学已越来越被广大领域所接受。

二、常用统计学术语

1. 总体（population）和样本（sample）

任何统计研究都必须首先确定观察单位，亦称个体或试验单元（experimental unit）。试验

单元是统计研究中最基本的单位，是试验处理实施的对象或观察对象，可以是一个人、一个家庭、一个地区、一个样品、一个采样点等。

总体（population）是根据研究目的确定的具有相同性质的试验单元的全体，或者说，是具有相同性质的所有试验单元某种观察值（变量值）的集合，包含所研究的全部个体（元素）。例如欲研究浙江省2011年在校大学生的血红蛋白含量，那么，试验单元是每一个浙江省2011年的在校大学生，分析的变量是血红蛋白含量，变量值（观察值）是血红蛋白含量测定值，则浙江省2011年在校大学生血红蛋白含量值构成一个总体。它的同质基础是同地区、同年份、同为在校大学生。

总体又分为有限总体和无限总体。有限总体是指在某特定的时间与空间范围内，同质研究对象的所有试验单元的某变量值的个数为有限个。如研究化工厂的废液排放对废液流经的5个湖泊的水质的影响，分析某大学某专业某班同学的身高。无限总体是抽象的，无时间和空间的限制，试验单元数是无限的，如研究碘盐对缺碘性甲状腺病的防治效果，该总体的同质基础是缺碘性甲状腺病患者，均用碘盐防治；该总体应包括已使用和设想使用碘盐防治的所有缺碘性甲状腺病患者的防治效果，没有时间和空间范围的限制，因而观察单位数无限，该总体为无限总体。

在实际工作中，所要研究的总体无论是有限的还是无限的，通常都是采用抽样研究。样本（sample）是按照随机化原则，从总体中抽取的有代表性的部分试验单元的变量值的集合，是实际研究中被分析的个体集合。如从上例的有限总体（浙江省2011年在校大学生）中，随机抽取200名在校大学生测定其血红蛋白含量，他们的血红蛋白含量值即为样本。从总体中抽取样本的过程为抽样，抽取的样本数量即为样本容量（sample size）。样本容量的大小将影响样本的分析误差。通常样本单位数大于30的样本可称为大样本，小于30的样本则称为小样本。在实际应用中，应该根据调查的目的认真考虑样本容量的大小。

2. 变量（variable）和常数（constant）

变量是指相同性质的事物间表现出差异性或差异特征的数据，常用小写的英文字母 x、y、z 等表示。变量值（value of variable）是指变量的观察结果。如研究某品种水稻的株高，株高是变量，而株高的测量结果即变量值。根据变量性质是否为连续分为连续变量和离散变量，如表示株高的变量为连续型变量，而表示性别的变量即为离散型变量。常数是指代表事物特征和性质的数值，通常由变量计算而来。

3. 参数（parameter）和统计量（statistic）

为了描述总体和样本的数量特征，需要计算出几个特征数。描述总体特征的数值叫参数（parameter）。总体参数一般未知，通过样本进行估计。常用希腊字母表示参数，例如表示总体平均数的 μ，表示总体方差的 σ^2。从样本中计算所得的特征数叫统计量（statistic）。常用拉丁字母表示统计量，例如用 $\bar{x}$ 表示样本平均数，用 S 表示样本标准差。如果 $X_1, X_2, \Lambda, X_i, \Lambda, X_n$ 是总体的样本，统计量是样本的已知函数 $g(X_1, X_2, \Lambda, X_i, \Lambda, X_n)$，它不包含总体分布的任何未知参数。

4. 准确性（accuracy）和精确性（precision）

准确性是指在调查或试验中某一试验指标或性状的观测值与其真值接近的程度，也称为准确度。设某一试验指标或性状的真值为 μ，观测值为 x，若 x 与 μ 相差的绝对值 $|x-\mu|$ 小，则观测值 x 的准确性高；反之则低。精确性指调查或试验中同一试验指标或性状的重复观测值彼此接近的程度，也称为精确度，描述多次测定值的变异程度。若观测值彼此接近，即任意二个观测值 x_i、x_j 相差的绝对值小，则观测值精确性高；反之则低。由于真值常常不知道，所以准确性不易度量，但利用统计方法可度量精确性。

5. 无偏性（unbiasedness）和有效性（efficiency）

评价参数估计量优劣时，通常用无偏性和有效性来衡量。如果估计量 $\hat{\theta}$ 是参数 θ 的点估计，并有 $E(\hat{\theta})=\theta$，则 $\hat{\theta}$ 是 θ 的无偏估计（unbiased estimation）。对于某个参数，可能存在若干个无偏估计量。这些无偏估计量并不都是等效的。如果参数的两个无偏估计量和的方差分别为 $\sigma^2(\hat{\theta}_1)$ 和 $\sigma^2(\hat{\theta}_2)$，并且 $\sigma^2(\hat{\theta}_1)<\sigma^2(\hat{\theta}_2)$，那么无偏估计量 $\hat{\theta}_1$ 比无偏估计量 $\hat{\theta}_2$ 更有效。

6. 误差（error）和错误（mistake）

在科学试验中，试验指标除受试验因素影响外，还受到许多其他非试验因素的干扰，从而产生误差。误差是指试验中不可控制因素引起的观测值偏离真值的差异。误差分为随机误差（random error）与系统误差（systematic error）。随机误差是由许多无法控制的内在和外在的偶然因素所造成，虽然在试验过程中尽可能地控制一致但难以做到绝对一致，如试验对象的初始条件、管理措施等。随机误差带有偶然性质，只能通过控制试验条件尽可能地减小，但无法完全消除。随机误差影响试验的精确性。系统误差是由于试验条件未获得良好控制，使得试验结果出现一致性地变化趋势。如，由于称量设备未做有效校准，使得称量结果一致性地偏低或偏高。系统误差影响试验的准确性，可通过良好的试验设计进行控制。错误是指试验过程中人为作用引起的差错，是完全应该和可以避免的。

三、概率

在自然界、生产实践和科学试验中，有些事件是可预言其结果的，即在保持条件不变的情况下，重复进行试验，其结果总是确定的，必然发生（或必然不发生）。而有些事件是事前不可预言其结果的，即在保持条件不变的情况下，重复进行试验，其结果未必相同，这类事件称为必然事件（certain event），用 Ω 表示。例如，掷一枚质地均匀对称的硬币，其结果可能是出现正面，也可能出现反面，这类事件称为随机事件（random event），通常用 A、B、C 等来表示。随机事件发生的可能性大小即概率（probability）。事件 A 的概率记为 P（A）。

在相同条件下进行 n 次重复试验，如果随机事件 A 发生的次数为 m，那么 m/n 称为随机事件 A 的频率（frequency）；当试验重复数 n 逐渐增大时，随机事件 A 的频率越来越稳定地接近某一数值 p，那么就把 p 称为随机事件 A 的概率。

根据概率的定义，概率有如下基本性质：

① 对于任何事件 A，有 $0\leqslant P(A)\leqslant 1$；

② 必然事件的概率为 1，即 $P(\Omega)=1$；

③ 不可能事件的概率为 0，即 $P(\phi)=0$。

若随机事件的概率很小，例如小于 0.05、0.01、0.001，称之为小概率事件。小概率事件虽然不是不可能事件，但在一次试验中出现的可能性很小，不出现的可能性很大，以至于实际上可以看成是不可能发生的。在统计学上，把小概率事件在一次试验中看成是实际不可能发生的事件称为小概率事件实际不可能性原理，亦称为小概率原理。小概率事件实际不可能性原理是统计学上进行假设检验（显著性检验）的基本依据。

四、概率分布

如果表示试验结果的变量 x，只可能取有限个或至多可列个值，且以各种确定的概率取这些不同的值，则称 x 为离散型随机变量（discrete random variable），如出生小孩的性别，豌豆的花色。如果表示试验结果的变量 x，其可能取值为某范围内的任何数值，且 x 在其取值范围内的任一区间中取值时，其概率是确定的，则称 x 为连续型随机变量（continuous random variable），如身高、产量。事件的概率表示了一次试验某一个结果发生的可能性大小。若要全面了解试验，则必须知道试验的全部可能结果及各种可能结果发生的概率，即必须知道随机试验的概率分布（probability distribution）。根据随机变量性质的不同，分为离散型概率分布和

连续型概率分布。

1. 离散型概率分布

设 x 是一个离散型随机变量，它的所有可能取值为 $x_i(i=1,2,\cdots)$。若 x 取这些值的概率为：

$$P(x=x_i)=p(i=1,2,\cdots) \tag{1-1}$$

则称式（1-1）为离散型随机变量 x 的概率分布（或概率函数），简称分布。

下表为概率分布的表格形式，为离散型随机变量的概率分布列：

x	x_1	x_2	$\cdots$	x_n	$\cdots$
$P(x=x_i)$	p_1	p_2	$\cdots$	p_n	$\cdots$

由概率定义知，p_i 满足如下性质：

① $p_i \geqslant 0$ $(i=1, 2, \cdots)$；

② $\sum_i p_i = 1$。

当给定了 x_i 及 p_i $(i=1, 2, \cdots)$，我们就能很好地描述随机变量 x 了，因为我们已经知道了它所取的值，以及它取这些值的概率。

几种常见的离散型随机变量的概率分布。

（1）两点分布或（0-1）分布　若随机变量 x 的概率分布为：

$$P(x=1)=p,\ P(x=0)=1-p=q,\ \text{其中 } 0<p<1$$

则称随机变量 x 服从两点分布［或（0-1）分布］，记为：

$$x\sim (0,\ 1)$$

【例 1-1】 掷一枚硬币的试验，观察其正反面的结果，令

$$x=\begin{cases}1, & \text{结果为正面}\\ 0, & \text{结果为反面}\end{cases}$$

则有 $P(x=1)=1/2$，$P(x=0)=1/2$，故随机变量 x 服从两点分布。

（2）二项分布　若随机变量 x 的分布为：

$$P(x=k)=C_n^k p^k(1-p)^{n-k},\ (k=0,1,2,\cdots,n),\ \text{其中 } 0<p<1$$

则称随机变量 x 服从以 n，p 为参数的二项分布，记为：

$$x\sim B(n,\ p)$$

显然，概率 $C_n^k p^k(1-p)^{n-k}$ 就是 n 重伯努利试验中事件 A 发生 k 次的概率，且两点分布就是二项分布在 $n=1$ 时的特殊情况。

【例 1-2】 设一批产品共 1000 个，其中有 10 个次品，采用有放回抽样方式随机抽取 100 个样品，求样品中次品数 x 的概率分布。

解：从产品中任取一件为次品的概率 $p=0.01$，采用有放回抽样方式，每一次是否取到次品是相互独立的，因此样品中次品数 x 的可能取值为：0，1，2，…，100 且

$$P(x=k)=C_{100}^k 0.01^k(1-0.01)^{100-k},\ k=0,\ 1,\ 2,\ \cdots,\ 100$$

（3）泊松分布（Poisson 分布）　若随机变量 x 的概率分布为：

$$P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda},\ (k=0,1,2,\cdots,),\ \text{其中 } \lambda>0 \text{ 为常数}$$

则称 x 服从参数为 λ 的泊松分布，记为：

$$x\sim P(\lambda)$$

在客观世界中，服从泊松分布的随机变量是常见的，如一页书中印刷错误出现的个数；一块试验地中发生病虫害的头数；公共汽车站到来的乘客数等都服从或近似服从泊松分布。泊松