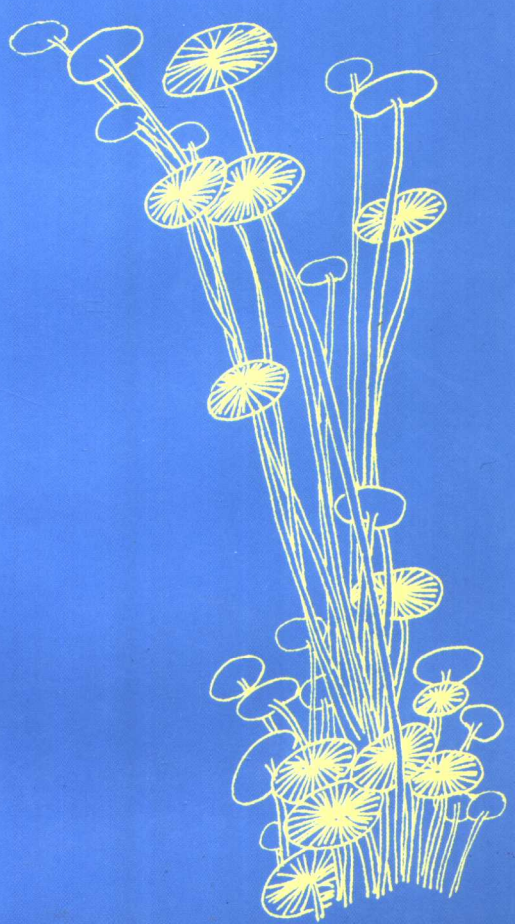
大连水产学院资助出版

生物统计

李盛德　查健禄　汪德洪　主编

SHENGWU TONGJI



大连海事大学出版社

大连水产学院资助出版

生 物 统 计

李盛德　查健禄　汪德洪　主编

大连海事大学出版社

图书在版编目(CIP)数据

生物统计 / 李盛德, 查健禄, 汪德洪主编 .—大连 : 大连海事大学出版社, 2002.11

ISBN 7-5632-1587-5

Ⅰ.生… Ⅱ.①李… ②查… ③汪… Ⅲ.生物统计 Ⅳ.Q-332

中国版本图书馆 CIP 数据核字(2002)第 071170 号

大连海事大学出版社出版

地址:大连市凌水桥 邮编:116026 电话:4728394 传真:4727996

http://www.dmupress.com E-mail:cbs@dmupress.com

大连海事大学印刷厂印装 大连海事大学出版社发行

幅面尺寸:185 mm × 260 mm 印张:15.25

字数:381 千字 印数:1 ~ 1 050 册

2002 年 11 月第 1 版 2002 年 11 月第 1 次印刷

责任编辑:陈 航 版式设计:陈 航

封面设计:王 艳 责任校对:李雪芳

定价:29.00 元

内容提要

本书内容包括在生物科学领域常用的数理统计基本概念、参数估计与假设检验、方差分析与试验设计、回归分析、协方差分析以及常用的多元统计分析方法。

本书着重基本概念的阐释及试验设计与统计分析方法的基本原理及其应用，并在当前生物类本科生可接受的数学程度内，力求做到论证严谨。

本书可作为有关生物类专业研究生教材或本科生的选修课教材，也可供有关生物科技工作者参考阅读。

前　言

本书的目的是作为有关生物类专业研究生的生物统计课程教材，本科生可选修其部分内容，也可作为生物科技工作者自学参考用书。

开设本门课程的基本目的是使学生在掌握常用试验设计及统计分析方法的同时，正确把握和深刻理解有关统计思想和概念，训练用正确的统计观点去观察和研究生物领域中有关问题的能力与习惯。

基于上述考虑，本书在编写时力求体现下述特点：

(1)对于基本概念，尽力阐释其统计意义与实际背景。对于统计分析方法，力求通过直观分析讲清其基本思路及实现途径，以理解方法的基本原理，并对有利于理解统计概念及实现方法的理论进行适当的证明，但不追求纯形式的理论推演。

(2)努力体现实用性。选材上包括数理统计、试验设计及多元分析在生物科技领域的常用方法，尽力阐述各种方法在应用中的一般性问题及其适用范围，以利依据实际问题的需要选用适当的方法，并分析、解释所得结论的统计意义与实际意义。

(3)在注重实用性的同时，注重思维的启迪。通过依实际问题的需要建立概念与方法，尽力阐述以数学思维方式观察、认识、描述、分析与解决问题的特征，并尽力从中提炼出具有一般意义的规律，以利学生理解、掌握本书未论述的统计分析方法。

由于现已开设应用数学软件课程，本书未涉及常用统计软件的使用及作为上机实习资料的专业性较强的应用实例。书中仅附可经计算器完成的简单问题，作为掌握基本方法的练习。

本书是将 1995 年编写的生物数学讲义中的一部分，结合教学情况，并参阅众多相关的最新著作进行修订与充实后写成的。

本书是在大连水产学院有关部门的帮助、支持和资助下才得以完成并出版的，杨园、张立峰、万莹同志在组稿工作中付出很多劳动，在此一并致谢。

限于作者水平，不妥与错误之处在所难免，恳请读者批评指正。

作　者

2002 年 3 月

目 录

绪　论

实践是认识的来源，但认识并不是实践的直接产物。因此，常通过观察与试验来探索客观规律。首先应通过观察或试验收集必要的数据，这些数据常受到随机性因素的影响。下一步就是对收集到的数据进行整理、分析，以透过随机干扰对所研究问题中的规律做出某种形式的结论。在这个过程中存在许多数学问题，解决这些问题的理论与方法，就构成了数理统计的内容，故一般地可以说：

数理统计学是研究怎样用有效的方法去收集和使用带有随机性影响数据的一个数学分支。这里说明了数理统计的研究对象、研究内容与研究特点。

(1)数据必须带有随机性的影响，才能成为数理统计的研究对象。

这里所说的随机性的来源有二，一是问题中所涉及的对象为数很大，不可能对其全面研究，只能用“一定的方式”取其一部分进行观察。例如养殖扇贝 6 个月后，欲知其增重状况，我们不能将养殖扇贝全部取出计量，只能取其中一部分如 10 笼，由这 10 笼的计量结果，去估计所有扇贝的增重状况，在这里，随机性的影响就在于被取出的 10 笼是偶然的。

另一个来源是试验的随机误差，这是指那些在试验过程中未加控制、无法控制，甚至还不了解的因素对试验结果的影响。例如：某药物对真鲷育苗有作用，现欲通过试验来观察这种作用的程度，并选出适当的剂量供今后使用。而真鲷人工育苗状况除与该药物剂量有关外，还受到种鲷状况、水温、pH 值、操作水平等其他因素的影响，若在试验时对此未加或无法控制，势必对试验结果产生随机性影响。若从试验结果来看，使用剂量 t_1 较 t_2 好，那么这个结果在试验数据上的优势究竟是本质的，还是仅为随机误差的偶然性表现？这就需用数理统计的方法去分析。

如果由观察或试验所得的数据根本不存在随机性影响，例如在前述问题中，若将所有养殖扇贝均取出观察其增重状况，或将与真鲷人工育苗有关的所有因素均控制得如此严格(且以后推广也须如此)，以致使真鲷的育苗状况完全取决于该药物剂量(但这又是不可能的)，那就无须数理统计的分析方法了。

总之，所收集的数据是否有随机影响，是区别数理统计方法和其他数据处理方法的根本点。而且这里所说的带有随机性影响的数据是扬弃它们的实际意义经数学抽象的结果。

(2)“用有效的方式收集数据”是指能建立一个数学上可以处理并尽可能简便的模型来描述所得数据，且数据中包含尽可能多的与研究问题有关的信息。

例如为研究养殖虾的生长状况，从中取若干尾测其体长，在抽取时若刻意选取较大的，那所得数据就没有代表性，更谈不上有效了。若用一种纯随机方法抽取，则测得的体长大小分布状况反映了所有养殖虾体长的概率分布，从而可由此概率模型来描述所得数据。

又如在通过试验观察一些因素对某指标的影响时，处理与试验单元之间应如何搭配？当条件不允许做全面试验时，应如何选取部分试验以使收集到的试验数据更有代表性，且可建立简便又便于分析的模型？

这都是用有效方式收集数据所要研究的内容。这构成了数理统计的两个分支：抽样理论

与试验设计。

(3)“有效地使用数据”是指使用有效的方式去集中和提取所得数据的有关信息,以对所研究问题做出尽可能精确和可靠的推断。这里,所以只能做到“尽可能”而非绝对精确和可靠的推断,是因为数据受到随机性因素的影响,这种影响只能通过统计方法去估计或缩小其干扰作用,不能完全消除。而我们所做的推断又是对所研究问题的一个回答,并不仅限于所得数据的范围之内。

由前述分析已知,作为数理统计研究对象的带有随机性影响的数据已从其实际意义中超脱出来。因此,对它们进行有效收集与使用的方法具有广泛的应用性。例如一组试验数据只要其所受的随机性影响符合某个数学模型(如服从正态分布),就可用相应的统计分析方法分析,而不管这些数据的实际意义如何。但在将统计方法用于实际问题时,又必须对所论问题的专门知识有一定了解。这不仅有助于选用适当的统计方法,而且有助于对分析随机性数据所得的结论进行恰当解释。

数理统计在应用中的作用在于通过事物的外在数量上的表现,透过其中的随机性干扰,去探索、揭示事物的潜在规律性。但对事物为什么存在这样或那样的规律性的确认与解释,数理统计无能为力,只能依靠所论问题的专业知识。但这并不降低它的意义,由于事物的本质规律性往往隐藏很深,不易为人们觉察,而其外在数量上的表现则易于引起人们的注意,因此,在人们对事物的内在机理认识尚不充分,并且一直在探索其规律性的过程中,数理统计常能起到引导人们由事物外在的数量规律性去探究其内在规律性的先导作用。

在有了一定的理论用于生产实践时,为了探究对一种产品的某质量指标有影响的因素有哪些,哪些是主要的,影响有多大,何种因素状态水平是该产品的最优先生产条件,都要进行试验,就是把有关因素固定在若干水平上做试验,去观察感兴趣的指标值。所得试验结果必然受到大量随机因素的影响,只有运用统计分析方法才能回答前述问题。

在生物科学的有关领域中,由于生命现象常以大量重复的形式出现,又受到多种外界环境和内在因素的随机干扰,因此,不仅各种统计分析方法必然成为生物科学领域的研究工作和生产实践中的常规手段,而且一些统计分析方法即源于生物科学领域的实际问题。例如遗传学中的 Mendel 定律就是根据观察资料提出的定律。一些水产养殖物也要通过试验来确定其最适宜的生长条件。而分析试验数据的一种极重要的方法——方差分析法,就是 R. A. Fisher 等在 1923 年 ~ 1926 年期间,由田间试验开始发展起来的。因此,在近年发展起来的生物数学中,最早的一个分支就是生物统计,而生物统计实质上就是数理统计的分析方法在生物领域中的应用。

第1章　基本概念

1.1　总体与样本

如果要了解依某新技术养殖虾苗6个月后虾的生长状况，我们不能将所有虾取出进行观察，只能依一定方式从中取出部分进行观察，并依观察结果，推断由此技术养殖的所有虾的生长状况。若又知由原技术养殖的所有虾的生长状况，即可推断出两种养殖技术的差异。

在数理统计中，我们将研究的问题所涉及对象的全体构成的集合，称为总体。总体中的每个成员称为个体。从总体中抽取一些个体的行为，称为抽样。抽得的每个个体称为样品，抽得的个体的集合称为样本。样本中所含个体的个数称为样本容量。容量为 n 的样本常简称 n 样本。

不难知道这些概念在上例的具体意义。

(1) 我们研究总体时，并不是研究构成该总体的所有个体本身，而是研究依研究目的确定的某些特征或指标。例如虾的生长状况，我们关心的是虾的体长 X，X 的取值的全体，就构成了我们研究的总体。显然，虾的生长状况的优劣，并不在于区间$(0, L)$本身，而应由 X 在其取值范围内的概率分布确定，因此，X 是一个随机变量，总体的本质是该随机变量 X 的概率分布。因此，我们总是将总体与随机变量 X 及其概率分布等同起来，常称为总体 X，或依其分布称为正态总体等。

可见，随机变量 X 的取值范围无关紧要，虾的体长 X 的取值范围我们可认为是$(0, +\infty)$甚至$(-\infty, +\infty)$。

我们根据研究目的，来研究总体 Y 或总体(X, Y)等，依据随机变量的维数分别称为一维总体，二维总体 ……p 维总体。

(2) 总体的大小依据我们的研究目的确定。在上例中，如果我们只是研究一个养虾场，就其各种条件，该新技术在本场的养虾状况，那么总体就是这个养虾场运用该技术养殖的所有虾；如果我们要研究的是该技术的推广，那总体就是今后推广范围内的所有养殖虾。

我们又依据总体所含个体的多少，将总体分为有限总体与无限总体，离散型总体与连续型总体。

(3) 样本的两重性。在总体 X 中抽取一个个体为样本时，由于抽样不能预言该样本的取值，所以样本是一个随机变量。在抽样后，得到该随机变量的一次实现，即随机变量的一个观测值，这称为样本的两重性。在总体 X 中抽取 n 样本 $X_1, X_2, \cdots, X_n$，抽样后，得到 $X_1, X_2, \cdots, X_n$ 的实现 $x_1, x_2, \cdots, x_n$。

(4) 简单随机样本。在总体 X 中抽样，目的是通过样本 $X_1, X_2, \cdots, X_n$ 来研究总体，自然要求样本应很好地反映总体信息，为此对抽样应有一定的要求。最常见的是简单随机抽样，它要求抽取的样本满足如下的要求：

1) 要有代表性，即要求每一个体都有同等的机会被抽入样本，这便意味着每一样品 X_i 与总体 X 有相同的分布。

2) 要有独立性,即每次抽取的结果不受其他各次抽取的影响,也不影响其他各次的抽取,这便意味着 $X_1, X_2, \cdots, X_n$ 相互独立。

由简单随机抽样获得的样本,叫做简单随机样本。今后,只讨论简单随机样本,故也简称为样本。这时,样本 $X_1, X_2, \cdots, X_n$ 是相互独立的具有同一分布的随机变量,简称为独立同分布样本。

由此可知,若总体 X 的分布函数是 $F(x)$,则其独立同分布样本 $X_1, X_2, \cdots, X_n$ 的联合分布函数为 $F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$。

若总体 X 有分布密度 $f(x)$,则 $X_1, X_2, \cdots, X_n$ 的联合分布密度为:

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

1.2 总体分布

由 1.1 知,总体 X 的本质特性由 X 的概率分布刻画,下面介绍在生物领域中,最常见的总体分布。

1.2.1 两点分布

在许多问题中我们常只关心总体中的个体是否具有某一性状 A,例如一批鱼苗放养一定时期后,其中鱼苗是否成活或是否患病。这时,常可定义随机变量

$$\chi = \begin{cases} 1 & \text{有性状 A} \\ 0 & \text{无性状 A} \end{cases}$$

总体 χ 的概率分布 $f(x) = P(\chi = x) = p^x q^{1-x}, x = 0,1$,其中 $p + q = 1, 0 \leqslant p \leqslant 1$,叫做两点分布或 0-1 总体。也常记为 $\chi \sim \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$ 或 $\chi \sim B(1,p)$。

易知 $E\chi = p, D\chi = pq$。

例如在鱼病试验中,在两水箱分别放置 40 尾鱼,其中均有 20 尾病鱼。在其中一个水箱施用某种药物治疗,一定时间后,该水箱仅剩 5 尾病鱼;另一水箱中仍有 20 尾病鱼。可定义随机变量 $\chi = \begin{cases} 1 & \text{未愈} \\ 0 & \text{病愈} \end{cases}$,则用某药物治疗的总体 $\chi \sim \begin{pmatrix} 1 & 0 \\ \frac{1}{8} & \frac{7}{8} \end{pmatrix}$,未用某药物治疗的总体 $\chi \sim \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$。

可见,虽两总体均是 0-1 总体,但概率分布中的参数 p 不同,这反映了两总体间的差异。

1.2.2 二项分布

在许多问题中,常关心由 0-1 总体 $\chi \sim \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$,得到独立同分布的 n 样本 $\chi_1, \chi_2, \cdots, \chi_n$,

其中性状 A 出现的次数 $\chi = \sum_{i=1}^{n} \chi_i$，这是由 0-1 总体派生出的一个总体，常称为二项总体 χ，记为 $\chi \sim B(n,p)$。

显然二项总体 χ 的概率分布 $f(x) = P(\chi = x) = C_n^x p^x q^{n-x}$，$x = 0,1,2,\cdots,n$，$0 \leqslant p \leqslant 1$，$p + q = 1$。

易知 $E\chi = np$，$D\chi = npq$。

例 1.1 已知豌豆的颜色受一对等位基因 y(黄)，g(绿) 的控制，且 y 显于 g。现将两纯合亲本 yy，gg 杂交的 F_1 代自交，得到 556 粒 F_2 代的豌豆，求其中黄色豌豆数的概率分布。

解 定义 $\chi = \begin{cases} 1 & F_2 \text{ 代子体呈黄色} \\ 0 & F_2 \text{ 代子体呈绿色} \end{cases}$，则由 Mendel 遗传定律知 $\chi \sim \begin{pmatrix} 1 & 0 \\ 0.75 & 0.25 \end{pmatrix}$。所以，556 粒 F_2 代豌豆中的黄色豌豆数 $\chi \sim B(556,0.75)$，即

$$P(\chi = x) = C_{556}^x \times 0.75^x \times 0.25^{556-x} \qquad x = 0,1,2,\cdots,556$$

$$E\chi = 556 \times 0.75 = 417, \quad D\chi = 556 \times 0.75 \times 0.25 = 104$$

例 1.2 n 个生物个体在体积 V 的空间分布。

设生物个体不群居，且每个个体以相同的概率出现在空间的任一体积相同的部分，求在体积为 D 的样方中出现的生物个体数的概率分布。

解 由题知，任一个体出现在体积为 D 的样方中的概率 $p = \frac{D}{V}$。若定义

$$\chi = \begin{cases} 1 & \text{子体在样方中} \\ 0 & \text{子体不在样方中} \end{cases}$$

则 $\chi \sim \begin{pmatrix} 1 & 0 \\ \frac{D}{V} & 1 - \frac{D}{V} \end{pmatrix}$，从而，$n$ 个个体在此样方中的个数 $\chi \sim B\left(N, \frac{D}{V}\right)$，$E\chi = N\frac{D}{V} = \frac{N}{V}D$，$D\chi = \frac{ND}{V}\left(1 - \frac{D}{V}\right)$。

例如，用显微镜检查某溶液中的细菌数，在 118 个格子中共有 352 个细菌，则 1 个格子内的细菌数 $\chi \sim B\left(352, \frac{1}{118}\right)$，而在 118 个格子中，格子内有 x 个细菌的理论数应为

$$118P(\chi = x) = 118C_{352}^x \left(\frac{1}{118}\right)^x \left(\frac{117}{118}\right)^{352-x}$$

1.2.3 泊松(Poisson)分布

若总体 X 的概率分布 $f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$，$x = 0,1,2,\cdots,n$，则称 X 为泊松总体。记为 $X \sim \pi(\lambda)$。

易知 $EX = \lambda$，$DX = \lambda$。

由概率知识可知，当二项总体 $X \sim B(n,p)$ 的 n 很大，p 很小时，

$$C_n^x p^x (1-p)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0,1,2,\cdots,n$$

其中 $\lambda = np$。这个关系可记为 $X \sim \pi(np)$。

这表明:泊松分布是描述小概率事件 A 在大量重复独立试验中,出现次数 X 的概率分布。

在例 1.2 中,$n = 352, p = \frac{1}{118}, \lambda = \frac{352}{118} = 2.983$,可认为在一个格子中的细菌数 $X \sim \pi(2.983)$。由此计算 $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ 要较由二项分布计算简单许多,特别地,$P(X = x+1) = P(X = x)\frac{\lambda}{x+1}$,在依次计算 $P(X = x), x = 0,1,2,\cdots,n$ 时,更为简便。

例 1.3 已知经辐射处理,种子的突变率 $p = 0.005$,现观察 100 粒种子,有 2 粒种子发生突变的概率是多少,又欲观察到 1 粒种子发生突变的概率为 90%,至少应观察多少粒种子。

解 由 $p = 0.005$ 很小,$n = 100$ 很大,$np = 0.5$,故在 100 粒种子中的突变数

$$X \sim \pi(0.5)$$

$$P(X = 2) = \frac{e^{-0.5}0.5^2}{2!} = 0.075\ 8$$

由 $P(X \geqslant 1) = 1 - P(X = 0) = 1 - e^{-np} = 0.9$,得

$$n = \frac{\ln(1-0.9)}{-p} = \frac{\ln 10}{0.005} = 460.517 \approx 461$$

由上述可知 0-1 总体、二项总体、泊松总体三者之间有着密切的联系。其中二项总体、泊松总体都是由 0-1 总体派生出来的。二项总体 $B(n,p)$、泊松总体 $\pi(\lambda)$,其中的参数 $n, p, \lambda(\lambda = np)$,都与它们对应的 0-1 总体有关。因此,在讨论二项总体 $B(n,p)$ 与泊松总体 $\pi(\lambda)$ 时,必须认清它们对应的 0-1 总体。

1.2.4 正态分布

若总体 X 的概率密度 $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < \mu < +\infty, \sigma > 0, -\infty < x < +\infty$,则称 X 服从正态分布,记为 $X \sim N(\mu,\sigma^2)$,或称 X 为正态总体 $N(\mu,\sigma^2)$。其分布函数

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

易知 $EX = \mu, DX = \sigma^2$。

经分析后可知,正态总体的概率密度 $f(x)$ 具有下述特征:

(1) 当 $x = \mu$ 时,$f(x)$ 的值最大,$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$。

(2) $f(x)$ 随 $\left|\frac{x-\mu}{\sigma}\right|$ 单调递减。

(3) 曲线 $f(x)$ 关于直线 $x = \mu$ 对称,且以 x 轴为渐近线。

(4) 曲线 $f(x)$ 在 $x = \mu \pm \sigma$ 处,各有一拐点。

$f(x)$ 的图形见图 1.1。

当 $\mu = 0, \sigma^2 = 1$ 时,对应的正态分布 $N(0,1)$ 叫做标准正态分布,记为 $U \sim N(0,1)$。

概率密度 $\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u}{2}}$。

分布函数 $\phi(u) = P(U \leqslant u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{x^2}{2}}\mathrm{d}x$。

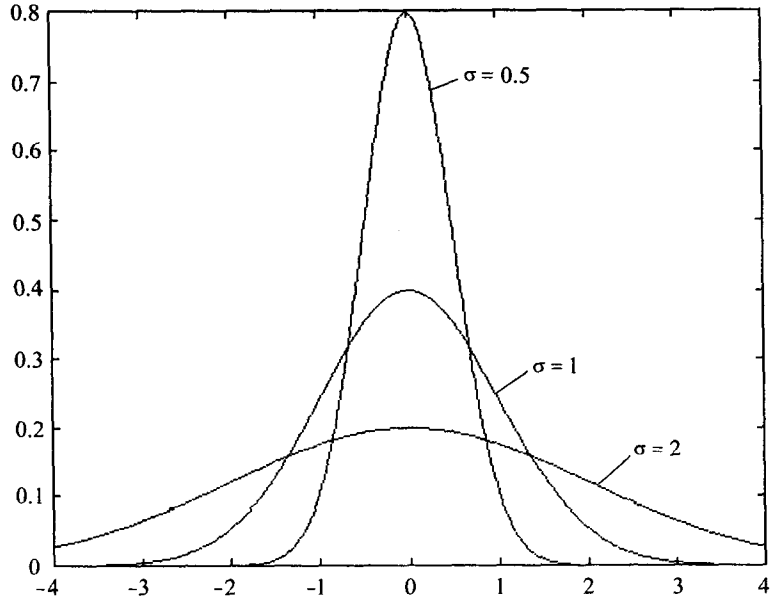


图 1.1 $\mu = 0;\sigma = 0.5,1,2$ 时的正态总体概率密度函数图形

正态分布函数图形见图 1.2。

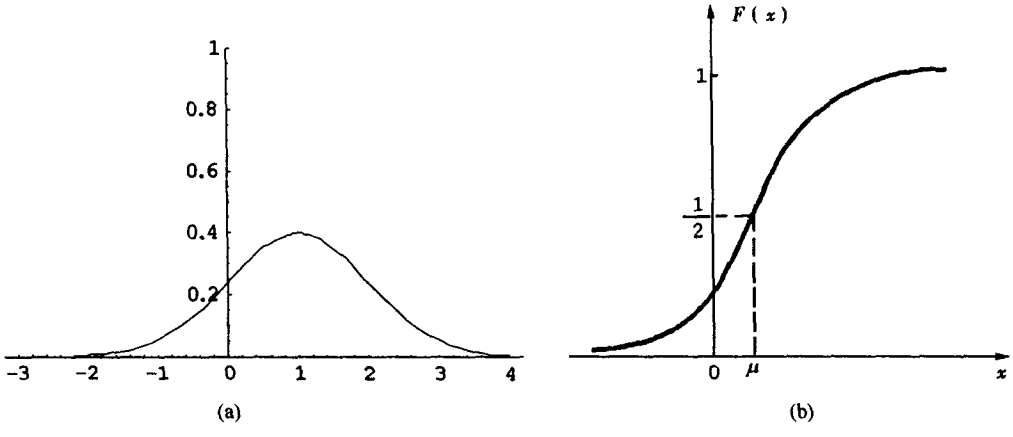


图 1.2 正态分布密度函数、分布函数图

显然,标准正态分布具有下述性质:

$$\varphi(u) = \varphi(-u)$$

$$\phi(-u) = 1 - \phi(u)$$

若 $X \sim N(\mu,\sigma^2)$,则 $U = \dfrac{x-\mu}{\sigma} \sim N(0,1)$,并称对 X 的这一变换为标准化变换。由此可见,标准正态分布函数 $\phi(u)$ 是正态分布计算的基础,且只要已知 $u > 0$ 时 $\phi(u)$ 的值即可,因此人们编制了 $u > 0$ 时的 $\phi(u)$ 函数值表(见附表 1)。例如:

$$\phi(1) = 0.841\,3, \quad \phi(1.64) = 0.949\,5$$

$$\phi(-0.83) = 1 - \phi(0.83) = 1 - 0.796\,7 = 0.203\,3$$

$$P(-0.83 < U < 1.64) = \phi(1.64) - \phi(-0.83) = 0.746\,2$$

反之,若给定概率 p,$\phi(u) = p$,亦可从表中查得相应的 u 值,例如 $\phi(u) = 0.99$,由附表1查得 $\phi(2.32) = 0.989\,8$,$\phi(2.33) = 0.990\,1$,用线性内插法可得 $\phi(2.327) = 0.99$,故 $u = 2.327$。

又如 $P(u > \mu) = 0.05$,由 $P(u > \mu) = 1 - P(u \leqslant \mu) = 1 - \phi(u) = 0.05$,得 $\phi(u) = 0.95$,查附表1可得 $u = 1.645$。此值常称为标准正态分布的(上)0.05分位点。

一般地,若总体 X 的分布函数 $F(x)$ 连续单调,对给定的 $\alpha \in (0,1)$ 满足 $P(X > x) = \alpha$ 的 x 的值叫做 X 的上 α 分位点,记为 x_α,即 $P(X > x_\alpha) = \alpha$。

当 $\alpha = 0.5$ 时,常称 $x_{0.5}$ 为 X 的中位点,常记为 MX。

当总体 X 的概率密度 $f(x)$ 为偶函数时,满足 $P(|X| > x) = \alpha$ 的值 x,叫做 X 的双侧 α 分位点,记为 $x_{\frac{\alpha}{2}}$,即 $P(|X| > x_{\frac{\alpha}{2}}) = \alpha$。

下列性质显然成立:

(1) $P(X < x_\alpha) = 1 - \alpha$, $P(|X| < x_{\frac{\alpha}{2}}) = 1 - \alpha$, $P(X < x_{1-\alpha}) = \alpha$。

(2) x_α, $x_{\frac{\alpha}{2}}$ 随 α 单调减少,且 $x_{\frac{\alpha}{2}} > x_\alpha$。

对标准正态总体的分位点通常用 z_α, $z_{\frac{\alpha}{2}}$ 或 u_α, $u_{\frac{\alpha}{2}}$ 表示。经查附表1可求得标准正态总体的常用的 α 分位点:

$$z_{0.05} = z_{\frac{0.1}{2}} = 1.645, \quad z_{0.005} = z_{\frac{0.01}{2}} = 2.576$$

$$z_{0.01} = 2.326, \quad z_{\frac{0.05}{2}} = 1.96$$

这里要注意 $z_{0.05}$ 与 $z_{\frac{0.1}{2}}$ 在意义上的差别。

当 $X \sim N(\mu,\sigma^2)$ 时,经标准化得 $U = \dfrac{X-\mu}{\sigma} \sim N(0,1)$,即可由附表1查得所需的值。例如

$$X \sim N(4,100^2)$$

$$P(-192 < X \leqslant 200) = P\left(\frac{-192-4}{100} < \frac{X-4}{100} \leqslant \frac{200-4}{100}\right) =$$

$$P\left(-1.96 < \frac{X-4}{100} \leqslant 1.96\right) = \phi(1.96) - \phi(-1.96) =$$

$$2\phi(1.96) - 1 = 2 \times 0.975 - 1 = 0.95$$

一般情况下,可求得

$$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$$

$$P(\mu - 2.58\sigma < X < \mu + 2.58\sigma) = 0.99$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997\,3$$

这表明对正态总体 $N(\mu,\sigma^2)$ 在抽样前不能预言 X 的取值,但可知 X 的取值 x 在区间 $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ 内的概率为95%。

由概率知识知,正态总体具有良好的性质,主要有

定理 1.1 若 $X_i \sim N(\mu_i,\sigma_i^2)$,且相互独立,$i = 1,2,\cdots,n$,则

$$X = \sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2\right) \qquad c_i \text{ 是常数}$$

推论 1.1 若 $X_1 \sim N(\mu_1,\sigma_1^2), X_2 \sim N(\mu_2,\sigma_2^2)$ 且相互独立，则

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

推论 1.2 若 $X_i \sim N(\mu,\sigma^2), i = 1,2,\cdots,n$，且相互独立，则 $\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$。

定理 1.2 （独立同分布的中心极限定理）

若 $X_i (i = 1,2,\cdots,n)$ 为独立同分布的随机变量，且 $EX_i = \mu, DX_i = \sigma^2 \neq 0$，则 $Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$ 的分布函数 $F_n(x)$ 对任意的 x，满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} < x\right) = \phi(x)$$

其中，$\phi(x)$ 是标准正态分布 $N(0,1)$ 的分布函数。这时，常称 $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$ 渐近服从标准正态分布。

上述定理表明，不仅若干个相互独立的正态随机变量的线性组合仍是正态随机变量，而且具有相同期望与方差的 n 个独立同分布的随机变量的和渐近服从正态分布。甚至这些随机变量有不同的期望与方差，这个结论仍在一定条件下成立。这就是说，无论各个独立的随机变量 $X_i, i = 1,2,\cdots,n$ 服从什么分布，在定理的条件下，它们的和 $\sum_{i=1}^{n} X_i$，当 n 很大时就近似服从正态分布，这是正态分布在理论与应用中都占有重要地位的基本原因。在生物科学领域的应用中正态分布尤为重要。

(1) 许多量所受的随机影响，都是由大量彼此独立的随机影响叠加而成，且每个随机因素的影响均很小，故这种随机误差近似地服从正态分布。例如，鱼池中同一种鱼的体长、体重等都良好地服从正态分布。

(2) 一些非正态量，在一定条件下可逼近于正态量。例如 $X \sim B(n,p)$，由于 $X = \sum_{i=1}^{n} X_i$，$X_i \sim \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$ 且相互独立，$i = 1,2,\cdots,n$，因此，当 $n \to \infty$ 时，$\frac{X - np}{\sqrt{npq}}$ 渐近服从 $N(0,1)$。当 n 很大时，有 $\frac{X - np}{\sqrt{npq}} \dot{\sim} N(0,1)$。

(3) 正态随机变量的一些函数的概率分布，也在理论与应用中有重要的地位。例如在计算毒性试验的半致死剂量中用到的对数正态分布及数理统计中最常用的 χ^2 分布、t 分布、F 分布等。

例 1.4 设 $X \sim N(\mu,\sigma^2)$，求 $Y = e^X$ 函数的概率密度。

解 由 Y 的分布函数 $F_Y(y)$，当 $y \leqslant 0$ 时，$F_Y(y) = 0$；当 $y > 0$ 时，

$$F_Y(y) = P(Y \leqslant y) = P(e^X \leqslant y) = P(X \leqslant \ln y) = \int_{-\infty}^{\ln y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

于是，Y 的概率密度为

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}x}F_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma y}\mathrm{e}^{-\frac{(\ln y-\mu^2)}{2\sigma^2}} & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

称此分布为对数正态分布(见图 1.3)，记为 $Y \sim LN(\mu,\sigma^2)$。

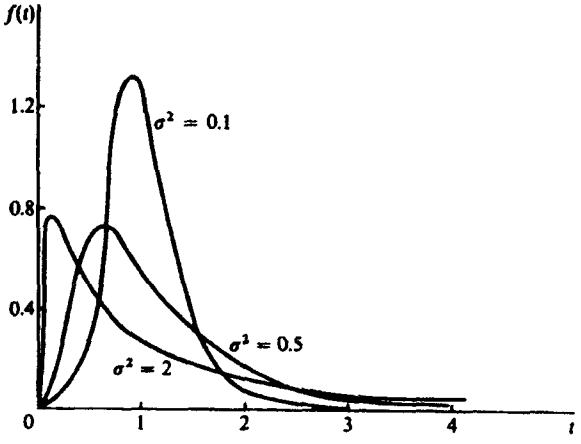


图 1.3　$\mu = 0$ 时的对数正态密度曲线

可以求得 $EY = \mathrm{e}^{\mu+\frac{\sigma^2}{2}}$，$DY = \mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2}-1)$，$MY = \mathrm{e}^{\mu}$。

类似可证：若 $\ln X \sim N(\mu,\sigma^2)$，则 $X \sim LN(\mu,\sigma^2)$。

1.2.5　由正态总体导出的重要分布

1.2.5.1　χ^2 分布

定义 1.1　若 $X_i \sim N(0,1)$ 且相互独立，$i = 1,2,\cdots,n$，则称 $\chi^2 = \sum_{i=1}^{n} X_i^2$ 为服从自由度 n 的 χ^2 分布，记为 $\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$。自由度(degree freedom) 常记为 df，是指其中独立正态变量的个数。

其概率密度 $f(x)$ 需用 Γ 函数表示，本书从略，图形见图 1.4。

对给定的 α 值，χ^2 分布的上 α 分位点 $\chi_\alpha^2(n)$ 可由自由度 n 及 α 值在附表 3 中查得，例如 $\chi_{0.1}^2(10) = 15.987$，$\chi_{0.9}^2(10) = 4.865$。

χ^2 分布具有下列性质：

(1) $E\chi^2 = n$，$D\chi^2 = 2n$。

(2) χ^2 分布的可加性。若 $\chi_1^2 \sim \chi^2(n_1)$，$\chi_2^2 \sim \chi^2(n_2)$ 且相互独立，则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$。可由 χ^2 分布的定义直观理解。

(3) 若 $X \sim \chi^2(n)$，则对任意的 x，有

$$\lim_{n\to\infty} P\left\{\frac{X-n}{\sqrt{2n}} \leqslant x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$