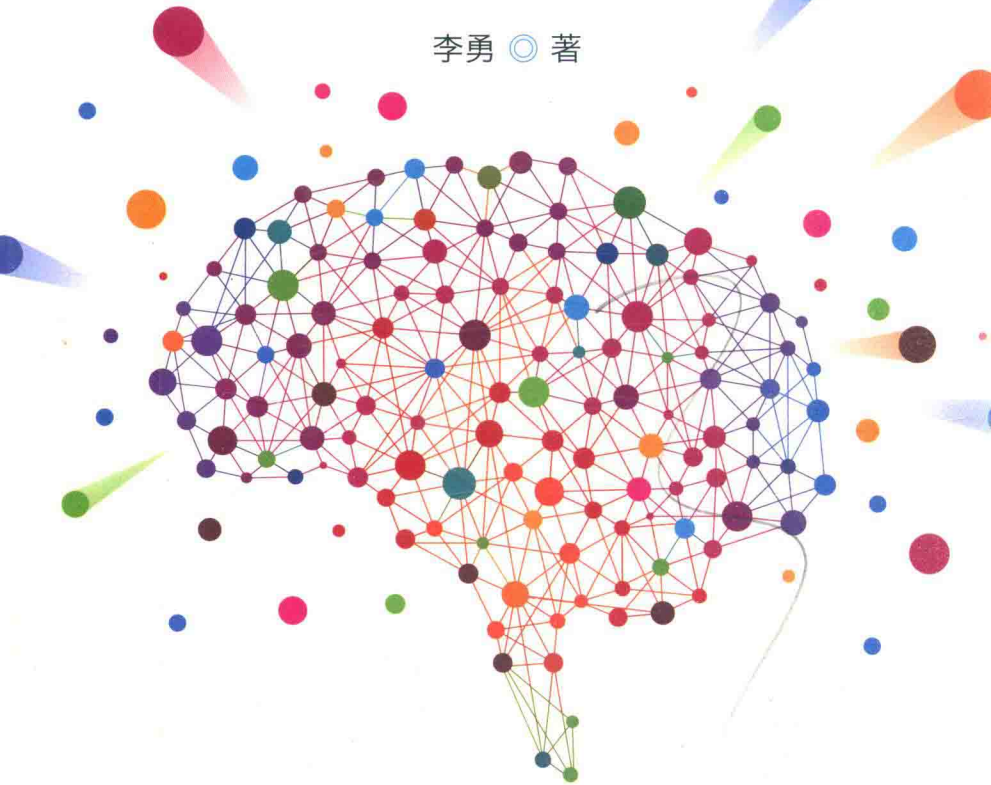
大数据时代的统计思想

李勇 ◎ 著

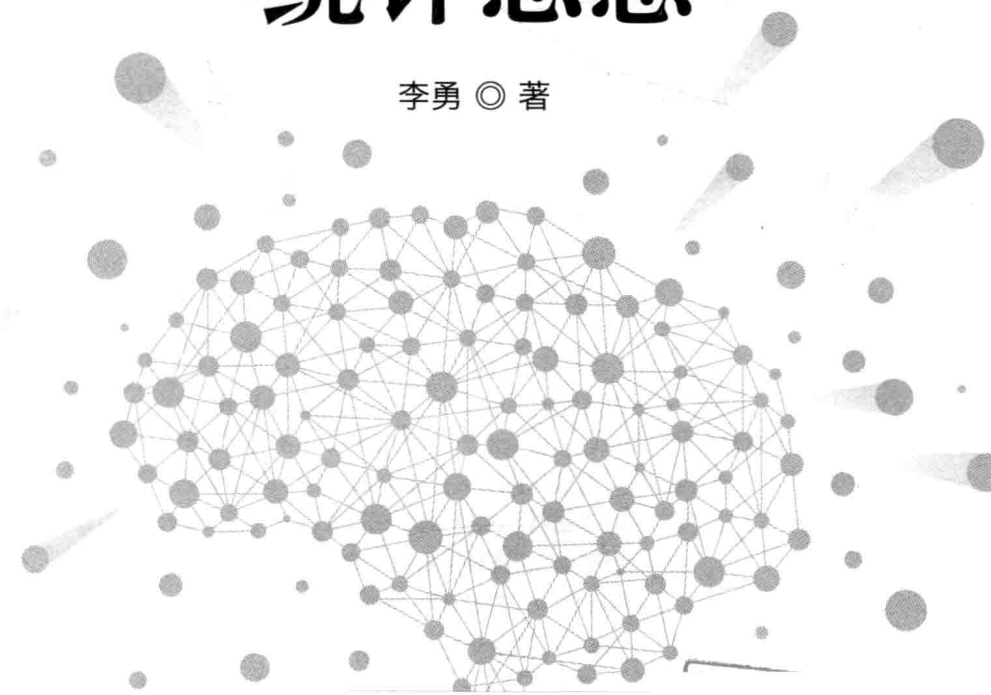


中国财经出版传媒集团
经济科学出版社
Economic Science Press

重庆工商大学
重庆市高校网络舆情与思想动态研究咨政中心 资助

大数据时代的统计思想

李勇 ◎ 著



中国财经出版传媒集团
经济科学出版社
Economic Science Press

图书在版编目（CIP）数据

大数据时代的统计思想/李勇著. —北京：经济科学出版社，2017.7

ISBN 978-7-5141-8182-1

Ⅰ.①大… Ⅱ.①李… Ⅲ.①统计学 Ⅳ.①C8

中国版本图书馆 CIP 数据核字（2017）第 158003 号

责任编辑：刘怡斐
责任校对：隗立娜
版式设计：齐 杰
责任印制：邱 天

大数据时代的统计思想

李 勇 著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

编辑部电话：010-88191348 发行部电话：010-88191522

网址：www.esp.com.cn

电子邮箱：esp@esp.com.cn

天猫网店：经济科学出版社旗舰店

网址：http://jjkxcbs.tmall.com

北京财经印刷厂印刷

三河市华玉装订厂装订

880×1230 32 开 7.5 印张 300000 字

2017 年 7 月第 1 版 2017 年 7 月第 1 次印刷

ISBN 978-7-5141-8182-1 定价：30.00 元

（图书出现印装问题，本社负责调换。电话：010-88191510）

（版权所有 侵权必究 举报电话：010-88191586

电子邮箱：dbts@esp.com.cn）

李勇，重庆工商大学重庆市高校网络舆情与思想动态研究咨政中心研究员，重庆工商大学统计系主任，重庆市高校网络舆情数据挖掘与应用统计研究所所长和重庆允升科技大数据研究中心主任。兼任全国工业统计教学研究会常务理事，企业经济统计学会常务理事兼秘书长；中国商业统计学会常务理事，数据科学与商业智能学会常务理事兼秘书长；中国统计教育学会理事，中国现场统计研究会资源与环境统计分会理事等。

主持（研）国家社（自）科基金面上项目6项，省部级课题10多项。公开发表学术论文60余篇，其中SCI/EI/ISTP/CSSCI/CSCD等收录40余篇。在科学出版社、经济科学出版社等出版著作8部。

此为试读，需要完整PDF请访问：www.ertongbook.com

序

大数据时代统计学理论体系的新特征

2017 年元旦刚过，阿尔法狗（AlphaGo）的变身马斯特（Master）一登场，连战 60 余场，将围棋高手们无一例外都“挑于马下”，使得人工智能和大数据再次火爆全球。而以数据为研究对象的统计学科，在当今人人都在谈论的大数据和机器智能的时代，将扮演什么样的“角色”？有人认为，统计学迎来了又一个“春天”；也有人认为，数据的出现促使统计学的产生，大数据的到来却是统计学的消亡。统计学科将何去何从？纵观统计学几百年的发展历程，尤其是近百年来的风云变幻，在大数据时代的今天，统计学科已经呈现出了什么样的新特征，其未来发展趋势如何？分析经典统计学体系的建立，剖析经典统计学体系的局限，透析经典统计学体系的突破，凸显大数据时代统计学的特征，预示统计学未来发展的趋势。

一、参数统计学体系的建立与完善

在 20 世纪，统计学一方面经 F. 高尔顿（F. Galton）、K. 皮尔逊（K. Pearson）、R. A. 费希尔（R. A. Fisher）、J. 奈曼（J. Neyman）和 E. 皮尔逊（E. Pearson）等对实际问题的研究，构建了统计学方法的基本框架；另一方面，在 K. R. 波莱尔（K. R. Borel）、H. L. 勒贝格（H. L. Lebesgue）、A. N. 柯尔莫哥洛夫（A. N. Kolmogorov）和

H. 克拉美（H. Cramer）等的探索中，建立了基于现代数学框架下的统计学理论体系，将统计学的理论和方法构筑于概率测度的基础上。

（一）参数统计理论体系的建立

1. 参数统计体系的概率论基础

20 世纪 30 年代，A. N. 柯尔莫哥洛夫建立了基于测度的概率论公理化体系，将统计学视为归纳推理的一个数学模型。在引入概率空间（Ω，F，P）和事件 $A \in F$ 概念下，确定了概率论的基本问题是研究分布函数及其性质。将统计学的基本问题设定为：在统计结构（Ω，F，P）中，利用来自于概率分布族 P 中某一未知分布的样本数据 x_1，x_2，…，x_n 去推断这一分布？若概率分布族 P 仅依赖于某一参数（向量）θ，即 $P = \{P_\theta: \theta \in \Theta\}$，$\Theta$ 为参数空间；则将统计问题转化为利用样本推断参数 $\theta_0 \in \Theta$ 的参数统计推断问题。R. A. 费希尔在论文《理论统计的数学基础》（*On the mathematical foundations of theoretical statistics*，1921）中明确地阐述了统计学家的“三观”：基于数据建模、基于模型估计、基于估计推断，对应统计学的基本内容是：统计模型、模型估计、假设检验。费希尔统计学的主要目标在于：根据已知的样本信息，在给定的简单模型族中，去估计产生随机信息的模型。简单模型族一般指维数较低，仅与有限个参数有关的参数化模型。以极大似然法为参数估计的基本方法，构建了经典参数统计体系。1920 ~ 1960 是其发展的“黄金时期”，史称“Fisher 时代”。

2. 参数统计体系的推断理论基础

参数统计体系是基于模型的建模方法，其理论基础扎根于三个方面。首先，借助于分析学的逼近理论。维尔斯特拉斯（Weierstrass）证明了在有限区间上的任一连续函数都存在任意精度的多项式逼近。即定义在 $[a, b]$ 上的连续函数 $f(x)$，对 $\forall \varepsilon > 0$，存在多项式 $P(x)$，使得不等式 $|f(x) - P(x)| < \varepsilon$ 对所有 $x \in [a, b]$ 一致地成立。M. R. 费歇特（M. R. Frchet）证明了存在一列多项式几乎处

处收敛于定义在［a，b］上的几乎处处有限的可测函数$f(x)$。这是参数线性统计模型理论的基石。其次，立足于中心极限定理。最早由 A. 棣莫弗（A. De Moivre）和 P. S. 拉普拉斯（P. S. Laplace）提出，其成立的充要条件是林德伯格（Lindeberg）条件：分布函数为$F_k(x)$的独立随机变量序列$\xi_k(k=1, 2, \cdots)$，$E\xi_k=a_k$，$D\xi_k=\sigma_k^2(>0)$，记$B_n^2=\sum_{k=1}^{n}\sigma_k^2$，$\eta_n=\dfrac{\sum_{k=1}^{n}(\xi_k-a_k)}{B_n}$，则$\eta_n\xrightarrow{D}N(0, 1)$且$\max\limits_{1\leqslant k\leqslant n}\dfrac{\sigma_k^2}{B_n^2}\to 0$的充要条件是$\lim\limits_{n\to\infty}\dfrac{1}{B_n^2}\sum\limits_{k=1}^{n}\int_{|x-a_k|>\varepsilon B_n}(x-a_k)^2\mathrm{d}F_k(x)=0$，$\forall\varepsilon>0$。该定理指出了随机变量之和可以通过正态分布逼近，为参数统计推断的大样本性质奠定了理论基础。最后，依托于大数定律。最早的伯努利（Bernoulli）大数定律，表达了频率趋于概率，即：$\lim\limits_{n\to\infty}P\left(\left|\dfrac{\eta_n}{n}-p\right|\geqslant\varepsilon\right)=0$。在更一般的条件下，具有$\lim\limits_{n\to\infty}P\left(\left|\dfrac{1}{n}\sum\limits_{k=1}^{n}(\xi_k-E\xi_k)\right|\geqslant\varepsilon\right)=0$和强大数定律$\lim\limits_{n\to\infty}\dfrac{1}{n}\sum\limits_{k=1}^{n}(\xi_k-E\xi_k)=0(a.s.)$的成立。这些为参数统计体系的推断奠定了理论基础。

3. 参数统计体系的基础模型

参数统计推断的核心在于预先假定统计模型。为此，统计学家们提出了很多统计模型，常见基本模型族有二项分布族、Poisson 分布族、正态分布族、Gamma 分布族、Beta 分布族和 t 分布族等；更大的分布族如：指数型分布族、群族等。分布的随机变量经过变换作用形成群族，线性模型是其中最重要的一类基本群族，指模型参数与观测值呈线性关系的参数统计模型。最基础的首推 Gauss - Markov 模型$y=X\beta+\varepsilon$，$E(\varepsilon)=0$，$Cov(\varepsilon, \varepsilon)=\sigma^2 I$，若将误差协方差阵拓展到一般的$\sigma^2\sum$，就是一般线性模型。进一步将线性模型误差协方差阵中除一个标量外完全已知这一苛求条件放宽，使其

可以灵活选择其协方差阵的结构，就构建了线性混合模型 $y = X\beta + Zu + \varepsilon$，$E(u) = 0$，$Cov(u, u) = D$，$E(\varepsilon) = 0$，$Cov(\varepsilon, \varepsilon) = R$，其中 β 为固定效应，u 为随机效应。为探索运动的规律建立了线性时间序列等模型。这些基本模型本质上都是与正态分布族关联。

（二）参数统计学体系的完善

1. 参数线性统计模型的延拓

世界并不都是呈现正态性的，为了满足现实世界的需求，统计模型势必进一步拓展。一个重要修正是尼尔德和韦德伯恩（Nelder & Wedderburn，1972）将正态假定扩展到指数型分布族，提出了广义线性模型。其模型结构包括三部分：（1）随机成分：观测值 y_i 是相互独立的随机变量，服从指数型分布族，其方差可随均值变化。（2）系统成分：线性预测值与线性回归模型一样为 $\eta_i = x_i^T\beta$。（3）连接函数：是一个严格单调可导的函数，将因变量拟合值变换后等于线性预测值 $g(\mu_i) = x_i^T\beta$，其中 $\mu_i = E(y_i)$。若再将系统成分和连接函数拓展到包含随机效应，就构建广义线性混合模型 $g(\mu) = \eta = X\beta + Zu$，其中 β，u 分别为未知的固定效应和随机效应。进一步完善了参数线性模型体系。

2. 非线性统计模型的研究

线性的世界难以引发质变，现实世界本质是非线性的，统计模型的非线性化发展成为必然。首先是可线性化的非线性模型。譬如：Gompertz 曲线模型 $Y = ke^{-be^{-aX}}$，通过变换，将非线性模型转化为线性模型：$\ln\left[\ln\left(\frac{k}{Y}\right)\right] = \ln b - aX \rightarrow Y^* = b^* - aX$。

还有指数曲线模型、对数曲线模型、双曲线函数模型、多项式曲线模型等。其次是本质非线性模型的研究。如非线性回归模型的一般形式为 $Y = f(X, \theta) + e$，其中期望函数 f 是未知参数 θ 的非线性函数。期望函数可以不是未知参数和控制变量的显函数，甚至可能是由没有解析解的非线性微分方程或积分方程所确定。其估计方

法可采用非线性最小二乘 Gauss - Newton 迭代算法、Newton - Raphson 迭代算法、数值导数方法或无导数最优化程序计算最小二乘估计等。再如非线性时间序列模型的建立，汤家豪（H. Tong，1977）引入的门限自回归模型 $X_t = \sum_{i=1}^{k} \{b_{i0} + b_{i1}X_{t-1} + \cdots + b_{ip_i}X_{t-p_i} + \sigma_i\varepsilon_t\} I(X_{t-d} \in A_i)$，其中，$A_i$ 是（$-\infty$，∞）的一个分割。E. 恩格尔（E. Engle，1982）引进的自回归条件异方差模型 $X_t = \sigma_t\varepsilon_t$，$\sigma_t^2 = c_0 + b_1X_{t-1}^2 + \cdots + b_pX_{t-p}^2$，以及拓展的自回归条件异方差模型 $X_t = \sigma_t\varepsilon_t$，$\sigma_t^2 = c_0 + \sum_{i=1}^{p} b_iX_{t-i}^2 + \sum_{j=1}^{q} a_j\sigma_{t-j}^2$ 等。这些模型的引入，突破了统计模型的线性王国，进一步完善了参数统计学体系。

3. 参数模型稳健性的提出

参数统计学体系的核心是基于模型驱动，即分析的前提是对数据的模型事先作出假定。若实际模型与假定的参数模型存在偏离时，基于假定模型推断的可信度如何？换句话说，基于理想化模型的最优化推断，随着参数的微小波动是否具有连续性。固定的单一模型假定，会产生极大的估计错误风险，自然的一个解决方案是假定数据来自于一个模型族，而非单一模型。为此，E. P. 博克斯（E. P. Box，1953）和 P. 休伯特（P. Huber，1954）提出了参数统计方法的稳健性。譬如 M 估计，是由极大似然估计引申出的稳健统计估计：指样本 X_1，X_2，…，X_n 对参数 θ 的估计 $\hat{\theta}$，满足 $\sum_{i=1}^{n} \frac{\partial\rho(X_i, \hat{\theta})}{\partial\hat{\theta}} = 0$ 或 $\sum_{i=1}^{n}\rho(X_i, \hat{\theta}) = \min_{\theta}\sum_{i=1}^{n}\rho(X_i, \theta)$，$(\rho(x, \theta) \geqslant 0)$。稳健性的研究，更进一步完善了参数统计学的推断理论体系。

二、参数统计学体系存在的尴尬和修正

参数统计模型一方面从线性延拓到广义线性，再扩大到非线性模型；另一方面，从单一模型假设到模型族的拓展，进行了稳健性

的探索，形成了较为完善的经典参数统计学体系。但面对现实复杂、高维数据的建模和分析，仍然存在不少的困境。

（一）参数统计学体系存在的尴尬

1. 维数灾难问题的存在

当将参数统计的推断方法推广到多变量的高维空间时，发现在低维空间表现优良的方法，在高维空间中往往失效，存在 R. 贝尔曼（R. Bellman，1961）的维数灾难问题。譬如：p 维单位超立方体上均匀分布数据的子立方体邻域，覆盖数据的部分 r 所需要的子立方体边长为 $e_p(r)=r^{\frac{1}{p}}$，在高维时不再具有局部性；高维空间中稀疏选择的另一个问题是所有样本点更靠近样本空间的边界，使得靠近训练样本边沿的预测更加困难；在高维空间中所有可用的训练样本稀疏地散布在 p 维输入空间，若 1 维变量样本量为 N，达到同等稠密样本所需的样本量呈指数倍增长为 N^p。维数灾难问题的存在，说明经典参数统计学体系存在一定的缺陷。

2. 不适定问题的发现

尽管有很多问题的形式化理论解存在，却不能利用有限的计算或信息等现实资源获得，这就是不适定问题。J. S. 阿达玛（J. S. Hadamard，1902）早就指出若算子方程 $Af(t)=F(x)$ 的解不满足存在性、唯一性和稳定性中的任一性质，则解该算子方程属于不适定问题。当时认为不适定问题仅存在于理论中，但随着计算机的发展，人们发现有一大类实际问题也属于不适定问题。尤其是经典参数统计体系的核心问题：利用样本数据 X_1，X_2，…，X_n 构造经验分布函数 $F_n(x)$ 逼近未知分布函数 $F(x)$，则通过求解积分方程 $\int_{-\infty}^{x} p(t)\mathrm{d}t=F(x)$ 获得密度函数 $p(x)$ 的密度估计问题属于不适定问题。这再一次揭示了经典参数统计学体系存在的严重缺陷。

3. 算法复杂度的提出

将统计学构筑于柯尔莫哥洛夫形式化公理体系上，尽管奠定了

此为试读，需要完整PDF请访问：www.ertongbook.com

统计学的数学严密化，却回避了随机性的本质问题。直到所罗曼诺夫（Solomonoff，1960）探索归纳推理的本质和 A. N. 柯尔莫哥洛夫（1965）与查德（Chaitin，1966）研究随机性的本质时，终于才提出了算法复杂度来构造随机性模型。算法复杂度是度量实现这个算法的最小程序长度的指标，利用演算论可在不计有界项的差别的意义下定义的。一个长度为 l 的二进制数据串是随机的，若不存在任何复杂度远小于 l 的算法能够产生这个数据串。若对数据串的描述不能被计算机压缩，则这个数据串具有一个随机序列的所有性质；这就说明，若能够在很大程度上压缩对一个给定数据串的描述，则所使用的算法就描述了数据的内在性质。而算法复杂度已经成为学习机器归纳推理的主要工具。由此可见，经典参数统计学体系存在诸多的尴尬，要实现统计学作为探索从数据所反映的规律的一门科学，还需进一步的修正完善。

（二）参数统计学体系的修正

1. 非参数统计模型的发展

参数统计学体系其实质是在一个由有限个参数决定的范围较窄的密度集合中估计密度函数，其推理模式以极大似然方法为主。一个自然的修正方案是进一步扩大可选密度函数集，将参数拓展到非参数可选密度函数集，建立非参数统计学体系。F. 罗森布拉特（F. Rosenblatt，1956）率先改进直方图法提出了 Rosenblatt 估计 $f_R(x)=\frac{1}{nh}\#\{i:\ X_i\in I_x,\ 1\leqslant i\leqslant n\}$，$I_x=\left[x-\frac{h}{2},\ x+\frac{h}{2}\right]$，帕尔逊（Parzen，1962）进一步提出 Parzen 核估计 $f_R(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h_n}\right)$，创立了估计密度函数的核估计方法，为非参数统计学奠定了基础。随后更多密度估计方法（最近邻密度估计、小波密度估计、经验似然估计、正交级数密度估计和条件密度局部线性估计等）的引入，大大拓展了密度函数的估计方法，形成了从传统的非参数检验发展到非

参数密度估计、条件密度估计、非参数回归模型、密度比模型、非参数时间序列模型等现代非参数统计模型方法体系。

2. 半参数统计模型的建立

非参数模型对总体分布限制较少，具有较大的灵活性和稳健性，但若是用于解决高维问题时，仍将面临“维数灾难”问题，为此提出了半参数统计模型。譬如恩格尔（Engle，1986）等提出的部分线性模型 $Y_i=\beta_0^T X_i+g(U_i)+\varepsilon_i$，$i=1$，2，…，$n$，由线性部分 $\beta_0^T X_i$ 和非线性部分 $g(U_i)$ 这两部分组成。又如单指标模型 $Y_i=g(\beta_0^T X_i)+\varepsilon_i$，$i=1$，2，…，$n$ 通过 p 维协变量 X 的线性组合将所有协变量投影到一维线性空间上拟合一个一元函数，一定程度上避免了多元非参数回归中出现的“维数灾难”问题。半参数统计模型体系已经建立了以部分线性模型、单指标模型、变系数模型、可加模型等为基本模型以及拓展模型：部分线性单指标模型、部分线性变系数模型、部分线性可加模型、单指标变系数模型、部分变系数单指标模型等。

3. 基于数据建模的统计体系完善

基于模型驱动统计体系的局限性，J. 图基（J. Tukey）提出探索性数据分析，强调从数据角度出发进行统计分析的观点，突破以模型为基础的统计分析观点，开创了基于数据建模的统计思想。随着非参数统计模型体系和半参数统计模型体系的发展，大大拓展了参数统计模型体系，逐渐完成了统计特殊推理模式向统计通用推理模式的转化，构建了基于数据建模的经典统计学体系。

三、经典统计学体系的突破和发展

（一）经典统计学体系的突破

现代统计学经历了参数——非参数统计学体系的建立和发展，形成了较为完善的基于数据建模的经典统计学体系；但面对海量数

据和高维复杂数据，以及从小样本数据中学习等难题，还是相形见绌。随着计算机和大数据时代的到来，统计问题的规模和复杂性剧增，再次引发了统计科学的革命。

1. 通用统计推理体系的建立

由费希尔开创的经典统计学体系发展较快，但主要是关于参数的特殊统计推理模式，至于一般的通用统计推理研究相对滞后。其实，格利文科（Glivenko）和康特里（Cantelli，1933）早就证明了，当样本量 $n\to\infty$ 时，经验分布函数 $F_n(x)$ 依概率收敛于真实分布 $F(x)$。即：$\sup_x |F_n(x)-F(x)| \xrightarrow[n\to\infty]{P} 0$。柯尔莫哥洛夫发现了收敛性的渐近精确收敛的指数速率：$\lim_{n\to\infty} P\{\sqrt{n}\sup_x |F_n(x)-F(x)| < \varepsilon\} = 1-2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2\varepsilon^2k^2}$，和 $P\left\{\lim_{m\to\infty}\sup_{n>m}\sup_x \sqrt{\frac{2n}{\ln\ln n}}\,|F_n(x)-F(x)| = 1\right\} = 1$。由此可以得出分布函数的估计界：

$$F_n(x)-\sqrt{-\frac{\ln\eta}{2n}} \leqslant F(x) \leqslant F_n(x)+\sqrt{-\frac{\ln\eta}{2n}}$$

$$F_n(x)-\sqrt{\frac{\ln\ln n}{2n}} \leqslant F(x) \leqslant F_n(x)+\sqrt{\frac{\ln\ln n}{2n}}$$

G－C－K 定理的概率测度形式可表示为

$$\sup_{P\in P_0} P\{\sup_{A_x\in F} |P(A_x)-\nu_n(A_x)| > \varepsilon\} < 2\exp\{-2\varepsilon^2 n\}$$

其中，$\nu_n(A_x)$ 为事件 $A_x=(-\infty, x)$ 的频率。这一理论在当时仅作为统计理论的一个内在技术，人们并没有意识到它暗示了比参数统计更一般的统计推理新思想，直到 20 世纪 60 年代末，模式识别问题的经验风险最小化原则（ERM）的建立，拓展到实函数的结构风险最小化原则（SRM）的构建，从而才为通用统计推理的基本体系奠定了基础。

2. 统计学习理论基础的奠定

在通用统计推理原则基础上，为了建立学习理论的统计学基

础，将大数定律随着样本数据 $n\to\infty$，均值序列 $\frac{1}{n}\sum_{k=1}^{n}\xi_k\to E\xi_k$ 的核心思想拓广到函数空间中，建立了关于均值一致收敛于数学期望充要条件的函数空间中的大数定律：对于函数集 $Q(z,\ \alpha)$，$\alpha\in\Lambda$ 中所有函数和集合 P 中所有分布函数 $F=F(z)$，一致双边收敛 $\sup_{\alpha\in\Lambda}\left|\int Q(z,\ \alpha)\mathrm{d}F(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\ \alpha)\right|\xrightarrow[a.\ s.]{n\to\infty}0$ 成立和一致单边收敛 $\sup_{\alpha\in\Lambda}\left(\int Q(z,\ \alpha)\mathrm{d}F(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\ \alpha)\right)\xrightarrow[n\to\infty]{a.\ s.}0$ 成立的充要条件分别是在 $Q(z,\ \alpha)$，$\alpha\in\Lambda$ 中的 ε 熵和在 $Q^*(z,\ \overline{\alpha})$，$\overline{\alpha}\in\Lambda$ 中的 ε 熵满足

$$\frac{H^{\Lambda}(\varepsilon;\ n)}{n}\xrightarrow{n\to\infty}0,\quad\forall\varepsilon>0$$

其中，ε 熵 $H^{\Lambda}(\varepsilon;\ n)=E\ln N^{\Lambda}(\varepsilon;\ z_1,\ \cdots,\ z_n)$；且 $Q^*(z,\ \overline{\alpha})=\begin{cases}0 & z<\overline{\alpha}\\ 1 & z\geqslant\overline{\alpha}\end{cases}$，满足 $Q(z,\ \alpha)\geqslant Q^*(z,\ \overline{\alpha})$，$\int[Q(z,\ \alpha)-Q^*(z,\ \overline{\alpha})]\mathrm{d}F(z)<\varepsilon$。提出函数集容量的一些新概念：函数集的 VC 熵、生长函数和 VC 维等，描述了函数集的多样性，建立了学习机器收敛速率 VC 界的系列理论，得到学习方法的推广性取决于容量而不是维数，有效克服了维数灾难问题，奠定了统计学习的理论基础。

（二）基于算法建模统计体系的形成

随着计算机功能的大大提高，突破了统计学家的“三观”，提出以预测而不是模型为统计学的新目标，在计算机等领域开发了许多基于算法的计算方法，譬如：神经网络、支持向量机、决策树、随机森林、boosting 和概率图模型等算法模型，能更好地处理复杂数据。

1. 以神经网络为基础的算法模型

基于算法建模的第一步是生理学家 F. 罗森布拉特（1958）提

出的一种线性分类学习机器——感知机。由输入空间（特征空间）$X\subseteq R^n$ 到输出空间 $Y=\{+1, -1\}$ 的函数 $f(x)=\text{sign}(wx+b)$ 组成，在所有线性分类模型 $\{f|f(x)=wx+b\}$ 中，将问题转化为求解由误分类点定义的损失函数 $L(w, b)=-\sum_{x_i\in M} y_i(wx+b)$ 最小化问题，以求得分离超平面。感知机其实由两层神经元组成，神经元模型是神经网络最基本的成分，最简单的是 M-P 神经元模型。感知机只拥有一层功能神经元，学习能力非常有限，要解决非线性可分问题，需考虑使用多层功能神经元。迄今最成功的神经网络学习算法是误差逆传播（BP）算法，其实质是将最小均方算法推广到非线性可微神经元组成的多层前馈网络，其学习能力比单层感知机强多了。单纯增加模型的复杂度，使得其训练效率偏低，易于过拟合。好在随着云计算和大数据时代的到来，计算能力的大幅提高可缓解训练的低效性，训练数据的大幅增加可降低过拟合风险。因此，以深度学习为代表的深层复杂神经网络模型登上了时代的历史舞台。

2. 以支持向量机为核心的算法模型

感知机利用误分类最小策略求解分离超平面，只能解决训练数据集线性可分，且解不唯一。利用间隔最大化求最优解的支持向量机，具有唯一解，且可以解决线性和非线性数据集。支持向量机学习方法根据训练数据是否线性可分，其模型分为：线性可分支持向量机、线性支持向量机和非线性支持向量机。线性可分支持向量机是通过间隔最大化或求解相应的凸二次规划问题

$$\min_{w,b}\frac{1}{2}\|w\|^2,\ s.t.\ y_i(wx_i+b)-1\geqslant 0,\ i=1, 2, \cdots, N$$

得到分离超平面 $w^*x+b^*=0$ 和分类决策函数 $f(x)=sign(w^*x+b^*)$。

线性支持向量机引入松弛变量 $\xi_i\geqslant 0$，通过软间隔最大化或求解相应的凸二次规划问题 $\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i$，$s.t.\ y_i(wx_i+b)\geqslant 1+\xi_i$，$\xi_i\geqslant 0$，$i=1, 2, \cdots, N$，得到分离超平面 $w^*x+b^*=0$ 和

分类决策函数 $f(x) = \text{sign}(w^* x + b^*)$。

非线性支持向量机利用核函数技巧，核函数 $k(x, z)$ 指输入空间 X 与特征空间 H（Hilbert 空间），存在映射：$\phi(x)$：$X \to H$，满足内积 $k(x, z) = \phi(x)\phi(z)$，$\forall x, z \in X$。通过核函数与软间隔最大化或求解相应的凸二次规划问题

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{N} \alpha_i,$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0, \ 0 \leqslant \alpha_i \leqslant C, \ i = 1, 2, \cdots, N$$

学习得到分类决策函数 $f(x) = \text{sign}(\sum \alpha_i^* y_i^* k(x, x_i) + b^*)$。支持向量机的核技巧实质是将输入空间映射到特征空间，在特征空间中寻求分离超平面，而特征空间通常是高维，甚至无穷多维。以支持向量机为代表的基于算法建模的统计学习方法成为大数据分析的重要方向。

3. 概率图模型

理论上进行概率推理，只需要联合概率分布即可，但现实中，其计算复杂度是变量个数的指数倍增长。概率图模型是基于图的表示作为高维空间上紧凑编码复杂分布的基础，由于图形语言可以呈现我们需要编码表述的实际分布的结构，使得所需结构变量可以大大减少，以降低推理计算的复杂度，使得模型能够有效地构造和利用。概率图模型是支持构建智能系统三个关键功能（表示、推理和学习）的重要框架，其基本类型是基于有向图模型的 Bayes 网和基于无向图模型的 Markov 网。Bayes 网是有向无圈图，节点代表随机变量，节点间的边代表变量之间的直接依赖关系；每个节点附有一个概率分布，根节点 X 是边缘分布 $p(X)$，非根节点是条件概率分布 $p(X_i \mid \pi(X_i))$，Bayes 网的联合概率分布分解为 $p(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} p(X_i \mid \pi(X_i))$。Markov 网源自于统计物理学中的 Markov 随机场，边表示一种不受网络中任何其他变量影响的交互影响，将

此为试读，需要完整PDF请访问：www.ertongbook.com