TURING 图灵数学·统计学丛书

CAMBRIDGE

理论与数值计算相结合

Core Statistics

统计学核心方法及其应用

[英] 西蒙·N.伍德 著
(Simon N. Wood)
石丽伟 译

涵盖理解和运用参数统计方法所需核心知识
为数据分析构建新方法

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵数学·统计学丛书

Core Statistics

统计学核心方法及其应用

[英] 西蒙·N.伍德 著
(Simon N. Wood)
石丽伟 译

人民邮电出版社
北京

图书在版编目(CIP)数据

统计学核心方法及其应用/(英)西蒙·N. 伍德(Simon N. Wood)著；石丽伟译. —北京：人民邮电出版社, 2018.12

(图灵数学·统计学丛书)

ISBN 978-7-115-49746-8

Ⅰ. ①统… Ⅱ. ①西… ②石… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字（2018）第 238572 号

内 容 提 要

本书主要介绍了统计模型及统计推断中的问题，并引入极大似然法和贝叶斯方法来解答这些问题；概述 R 语言；简括极大似然估计的大样本理论，然后讨论应用该理论所涉及的数值方法；讲述贝叶斯计算所需的数值方法——马尔可夫链蒙特卡罗方法；介绍线性模型的理论及其应用.

本书适合具有数理知识基础、想要了解统计学核心方法和应用的读者阅读.

◆ 著　　[英] 西蒙·N. 伍德
◆ 译　　石丽伟
责任编辑　张海艳
责任印制　周昇亮
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编 100164　　电子邮件　315@ptpress.com.cn
网址 http://www.ptpress.com.cn
三河市祥达印刷包装有限公司印刷
◆ 开本：700×1000　1/16
印张：13.5
字数：273 千字　　2018 年 12 月第 1 版
印数：1–3 000 册　　2018 年 12 月河北第 1 次印刷
著作权合同登记号　图字：01-2017-0673 号

定价：69.00 元

读者服务热线：(010)51095186 转 600　　印装质量热线：(010) 81055316

反盗版热线：(010)81055315

广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩上
Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

前　言

本书主要面向具有数理知识的读者，他们学习过统计学和概率论的入门课程，并且想简要了解统计学中的核心方法及其应用. 这些方法并非完全建立在标准模型的基础上. 第 1 章简要介绍了本书所需的基本概率理论. 第 2 章讨论了统计模型及统计推断中的问题，并引入极大似然法和贝叶斯方法来解答这些问题. 第 3 章概述了 R 语言. 第 4 章简单概括了极大似然估计的大样本理论，第 5 章讨论了应用该理论所涉及的数值方法. 第 6 章涵盖了贝叶斯计算所需的数值方法，主要是马尔可夫链蒙特卡罗方法. 第 7 章简单介绍了线性模型的理论及其应用. 附录内容包括常用分布、矩阵计算和随机数生成等有用的信息. 本书既非百科全书，也非操作手册，书后的参考文献并不宽泛，其目的是为读者进一步学习提供最为有用的资源. 本书的目标是简要涵盖理解和运用参数统计方法所需的核心知识，为数据分析构建新的方法. 现代统计学是介于计算和理论之间的一门学科，本书体现了这一点.

感谢 Nicole Augustin、Finn Lindgren、剑桥大学出版社的编辑、巴斯大学“应用统计推断”课程的同学以及统计学博士培训学会“统计计算”课程的同学提出的宝贵意见和建议.

目　录

第 1 章　随机变量

1.1　随机变量概述

统计学的本质是从具有不可预测性的数据中提取信息，**随机变量**则是为这种可变性建立模型的数学工具. 在每一次观测中，随机变量随机取不同的值. 我们无法提前预测随机变量的精确取值，但是可以对可能的取值做出概率性的刻画. 也就是说，我们可以描述随机变量的取值的**分布**. 本章简要回顾应用随机变量时所涉及的专业知识，以及一些常用的结果. 详细论述见参考文献 [8]、[19].

1.2　累积分布函数

随机变量（r.v.）X 的累积分布函数（c.d.f.）是满足下式的函数 $F(x)$：

$$F(x)=\Pr(X\leqslant x).$$

即，$F(x)$ 给出了 X 的取值小于或等于 x 的概率. 显然，$F(-\infty)=0$，$F(\infty)=1$，并且 $F(x)$ 是单调函数. 该定义的一个有用的结论是，如果 F 是连续函数，那么 $F(X)$ 在 $[0,1]$ 上呈均匀分布：它取 0 和 1 之间任意值的概率是相等的. 这是因为

$$\Pr(X\leqslant x)=\Pr\{F(X)\leqslant F(x)\}=F(x)\Rightarrow\Pr\{F(X)\leqslant u\}=u.$$

（如果 F 是连续函数），那么后者是 $[0,1]$ 上的均匀随机变量的累积分布函数.

定义累积分布函数的反函数为 $F^{-}(u)=\min(x|F(x)\geqslant u)$. 当 F 为连续函数时，F^{-} 正是 F 在一般意义下的反函数. F^{-} 通常叫作 X 的**分位函数**. 如果 U 在 $[0,1]$ 上呈均匀分布，那么 $F^{-}(U)$ 的分布就是 X 的累积分布函数 F. 对于可计算的 F^{-}，在给定均匀随机偏差的产生方式的前提下，上述定义给出了任意分布下的随机变量的生成方法.

令 p 为 0 和 1 之间的一个数. X 的 p **分位数**是一个数值，X 小于或等于该值的概率是 p，即 $F^{-}(p)$. 分位数有广泛的应用，其中一个应用是验证 $x_1,x_2,\cdots,x_n$ 是否是累积分布函数为 F 的随机变量的观测值. 将 x_i 按顺序排列，把它们作为“观测分位数”. 这些点和理论上的分位点 $F^{-}\{(i-0.5)/n\}(i=1,\cdots,n)$ 共同绘制的图叫作**分位数–分位数图**（QQ 图）. 如果观测值来自于累积分布函数为 F 的分布，那么得到的 QQ 图应该接近直线.

1.3 概率函数与概率密度函数

在很多统计学方法中，描述随机变量取某个特定值的概率的函数比累积分布函数更有用. 为了探讨这类函数，首先需要区分取离散值（例如非负整数）的随机变量和取值为实数轴上的区间的随机变量.

对于离散型随机变量 X，**概率函数**（又叫**概率质量函数**）$f(x)$ 是满足下式的函数：

$$f(x) = \Pr(X = x).$$

显然，$0 \leqslant f(x) \leqslant 1$，并且因为 X 的取值一定存在，所以对 x 的所有可能取值（记为 x_i）求和可得 $\sum_i f(x_i) = 1$.

对于连续型随机变量 X，因为它所有可能的取值有无限个，所以取任意特定值的概率一般是 0，因此，概率函数对连续型随机变量不适用. 取而代之的是**概率密度函数**$f(x)$，它给出了 X 在 x 附近的单位区间内取值的概率，即 $\Pr(x - \Delta/2 < X < x + \Delta/2) \simeq f(x)\,\Delta$. 更加正式的定义是，对任意常数 $a \leqslant b$，

$$\Pr(a \leqslant X \leqslant b) = \int_a^b f(x)\mathrm{d}x.$$

显然，$f(x)$ 必须满足 $f(x) \geqslant 0$ 且 $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$. 注意，$\int_{-\infty}^{b} f(x)\mathrm{d}x = F(b)$，因此如果 F' 存在，那么 $F'(x) = f(x)$. 附录 A 给出了一些常用的标准分布的概率函数或概率密度函数.

除特别注明外，后续几节主要考虑连续型随机变量，用适当的求和代替积分，可以得到等价的对离散型随机变量适用的结果. 为了简洁起见，约定当自变量不同时，概率密度函数不同（例如，$f(y)$ 和 $f(x)$ 表示不同的概率密度函数）.

1.4 随机向量

从单次观测中很难得到有用的信息. 有效的统计分析需要多重观测和同时处理多元随机变量的能力. 因此，我们需要概率密度函数的多元形式. 二维的情形能够充分阐释所需的概念，因此考虑随机变量 X 和 Y.

设 Ω 是 $x-y$ 平面上的任意区域，X 和 Y 的**联合概率密度函数**$f(x,y)$ 是满足下式的函数：

$$\Pr\{(X,Y) \in \Omega\} = \iint_{\Omega} f(x,y)\mathrm{d}x\mathrm{d}y. \tag{1.1}$$

因此，$f(x,y)$ 在 x,y 的取值是 $x-y$ 平面上**单位面积**的概率. 设 ω 是包含点 x,y 的面积为 α 的小区域，那么 $\Pr\{(X,Y) \in \omega\} \simeq f_{xy}(x,y)\,\alpha$. 同单变量的概率密度函数一样，$f(x,y)$ 是非负的，并且在 $\mathbb{R}^2$ 上的积分值为 1.

例 图 1-1 给出了下式中的联合概率密度函数的图像.

$$f(x,y)=\begin{cases} x+3y^2/2, & 0<x<1, 0<y<1, \\ 0, & \text{其他}. \end{cases} \tag{1.2}$$

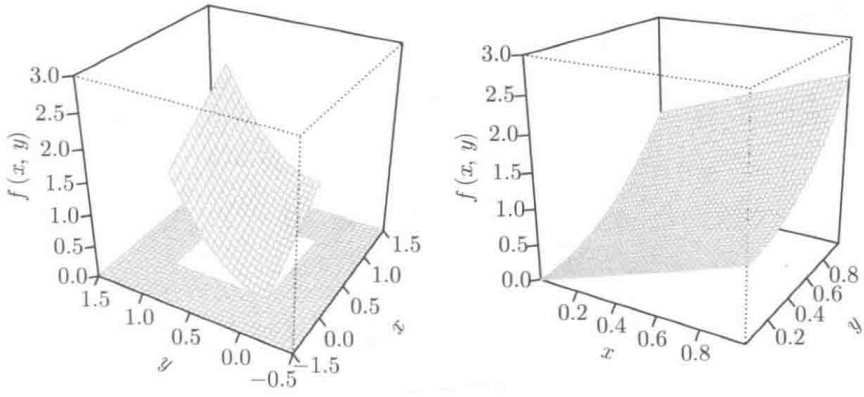

图 1-1 式 (1.2) 的联合概率密度函数. 左图: $[-0.5, 1.5] \times [-0.5, 1.5]$ 上的概率密度函数. 右图: 概率密度函数的非 0 部分

该概率密度函数下的两个概率值的估计如图 1-2 所示.

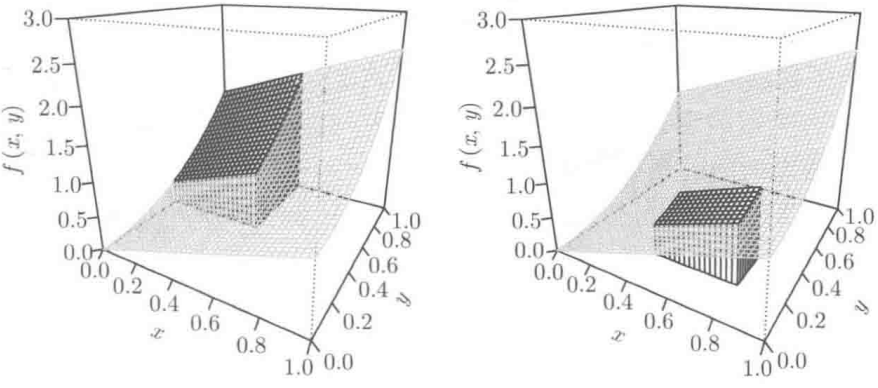

图 1-2 灰色部分表示用式 (1.2) 的联合概率密度函数估计概率值. 左图: 黑色部分的体积用于计算 $\Pr[X < 0.5, Y > 0.5]$. 右图: $\Pr[0.4 < X < 0.8, 0.2 < Y < 0.4]$

1.4.1 边缘分布

继续沿用 X 和 Y 的例子, 忽略其中一个变量, X 或 Y 的概率密度函数可以通过 $f(x,y)$ 来计算. 在给定 $-\infty < Y < \infty$ 的条件下, X 的概率密度就是 X 的**边缘概率密度函数**. 由概率密度函数的定义显然可以得到

$$f(x)=\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y,$$

$f(y)$ 的定义同理.

1.4.2 条件分布

假设已知 Y 取定值 y_0，那么关于 X 的分布，我们有什么结论？因为 X 和 Y 的联合概率密度函数是 $f(x,y)$，所以在给定 $Y=y_0$ 的条件下，我们预计 x 的密度与 $f(x,y_0)$ 成正比，即

$$f(x|Y=y_0)=kf(x,y_0),$$

其中 k 是常数. 如果 $f(x|y)$ 是一个概率密度函数，那么它一定能够取到积分值 1. 因此

$$k\int_{-\infty}^{\infty} f(x,y_0)\mathrm{d}x=1\Rightarrow kf\left(y_0\right)=1\Rightarrow k=\frac{1}{f\left(y_0\right)},$$

其中 $f\left(y_0\right)$ 表示 y 取 y_0 时的边缘密度. 因此我们有：

定义 如果 X 和 Y 的联合概率密度函数是 $f(x,y)$，那么在 $Y=y_0$ 的条件下，X 的**条件密度**是

$$f(x|Y=y_0)=\frac{f(x,y_0)}{f\left(y_0\right)},\tag{1.3}$$

假设 $f\left(y_0\right)>0$.

注意，当 Y 取定值 y_0 时，这是随机变量 X 的概率密度函数. 在意义明确的前提下，为了简洁起见, 可以用 $f(x|y_0)$ 代替 $f(x|Y=y_0)$. 显然，在给定 X 时，Y 的条件分布有类似的定义：$f\left(y\,|\,x_0\right)=f\left(x_0,y\right)/f\left(x_0\right)$. 联合概率密度函数和条件概率密度函数之间的关系如图 1-3 所示.

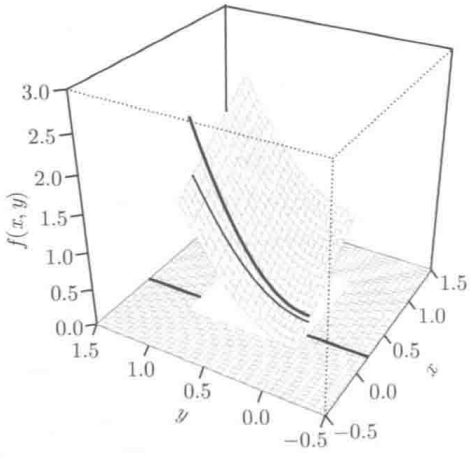

图 1-3 条件概率密度函数 $f\left(y|0.2\right)$. 灰色面表示联合概率密度函数 $f\left(x,y\right)$, 黑色细线表示 $f\left(0.2,y\right)$, 黑色粗线表示 $f\left(y|0.2\right)=f\left(0.2,y\right)/f_x\left(0.2\right)$

在统计学中，常常利用 $f(x,y)=f(x|y)f(y)$ 将联合概率密度替换为条件概率密度，但当维数超过 2 时，结论不能直接推广. 以下是 3 个较为常用的例子.

(1) $f(x,z|y)=f(x|z,y)f(z|y)$.

(2) $f(x,z,y)=f(x|z,y)f(z|y)f(y)$.

(3) $f(x,z,y)=f(x|z,y)f(z,y)$.

1.4.3 贝叶斯定理

从上一小节可知

$$f(x,y)=f(x|y)f(y)=f(y|x)f(x).$$

重组上式的后两项可以得到

$$f(x|y)=\frac{f(y|x)f(x)}{f(y)}.$$

这个重要的结论叫作**贝叶斯定理**，在该定理的基础上形成了一个完整的统计学模型体系，见第 2 章和第 6 章.

1.4.4 独立性和条件独立性

对于随机变量 X 和 Y，如果 $f(x|y)$ 的取值不依赖于 y 的取值，那么在统计意义上 x **独立于** y. 由此可以推导出

$$\begin{aligned}f(x)&=\int_{-\infty}^{\infty}f(x,y)\,\mathrm{d}y=\int_{-\infty}^{\infty}f(x\,|\,y)f(y)\,\mathrm{d}y\\&=f(x\,|\,y)\int_{-\infty}^{\infty}f(y)\,\mathrm{d}y=f(x\,|\,y),\end{aligned}$$

反过来可以得到 $f(x,y)=f(x\,|\,y)f(y)=f(x)f(y)$. 显然反之结论也成立，因为由 $f(x,y)=f(x)f(y)$ 可以得到 $f(x\,|\,y)=f(x,y)/f(y)=f(x)f(y)/f(y)=f(x)$. 一般来说：

当且仅当联合概率（密度）函数等于边缘概率（密度）函数的乘积，即 $f(x,y)=f(x)f(y)$ 时，随机变量 X 和 Y 相互独立.

在建模时假设随机向量的元素相互独立通常能够简化统计推断. 与之相比，假设元素**独立同分布**更加简便，但是适用性较差.

在很多实际应用中，建模时无法将一组观测值看作独立的，但是可以将其看作**条件独立**的. 大量的现代统计学研究致力于利用各种各样的条件独立性为非独立数据建立有用的模型，从而实现其计算的可行性.

考虑一个随机变量序列 $X_1,X_2,\cdots,X_n$，并且令 $\boldsymbol{X}_{-i}=(X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_n)^{\mathrm{T}}$. 条件独立性的一种简单形态是一阶马尔可夫性，

$$f(x_i|\boldsymbol{x}_{-i})=f(x_i|x_{i-1}).$$

也就是说，X_{i-1} 完全决定了 X_i 的分布，因此，**给定** X_{i-1}，X_i 是独立于序列的其余变量的. 由此可得

$$f(\boldsymbol{x}) = f(x_n|\boldsymbol{x}_{-n})\, f(\boldsymbol{x}_{-n}) = f(x_n|x_{n-1})\, f(\boldsymbol{x}_{-n})$$
$$= \cdots = \prod_{i=2}^{n} f(x_i|x_{i-1})\, f(x_1),$$

利用此公式常常可以极大地减少计算量.

1.5 均值和方差

尽管了解如何全面刻画随机变量的分布有其重要性，但是在很多情况下只需要了解其一阶或二阶性质就足够了. 概率密度函数为 $f(x)$ 的随机变量 X 的**均值**或**期望值**的定义是

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x.$$

因为上述积分对 x 的所有可能的取值按其发生的相对频率进行加权，所以我们可以将 $E(X)$ 理解为由 X 的观测值构成的一个无限序列的均值.

期望的定义对 X 的任意函数 g 都适用：

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x.$$

定义 $\mu = E(X)$，那么最为有用的一个函数是 $(X-\mu)^2$，这个函数用来计算 X 和它的均值之间的平方差，由此可以给出 X 的**方差**的定义：

$$\mathrm{var}(X) = E\left\{(X-\mu)^2\right\}.$$

X 的方差可以衡量 X 分布的分散程度. 虽然方差便于计算，但是由于它的单位是 X 的单位的平方，这使得它的解释性稍差. **标准差**是方差的平方根，因此标准差和 X 的数量级相同.

线性变换的均值和方差

由期望的定义随即可以得到：如果 a 和 b 是有限的实常数，那么 $E(a+bX) = a + bE(X)$. $a+bX$ 的方差需要稍多一点的推导：

$$\mathrm{var}(a+bX) = E\left\{(a+bX-a-b\mu)^2\right\}$$
$$= E\left\{b^2(X-\mu)^2\right\} = b^2 E\left\{(X-\mu)^2\right\} = b^2 \mathrm{var}(X).$$

如果 X 和 Y 是随机变量，那么 $E(X+Y)=E(X)+E(Y)$. 为了得到此式，假设它们的联合概率密度是 $f(x,y)$，那么

$$\begin{aligned}E(X+Y)&=\int(x+y)f(x,y)\mathrm{d}x\mathrm{d}y\\&=\int xf(x,y)\mathrm{d}x\mathrm{d}y+\int yf(x,y)\mathrm{d}x\mathrm{d}y=E(X)+E(Y).\end{aligned}$$

这个结果对 X 和 Y 的分布没有做任何假设. 如果假设 X 和 Y 相互独立，那么由以下的推导可以得到 $E(XY)=E(X)E(Y)$：

$$\begin{aligned}E(XY)&=\int xyf(x,y)\mathrm{d}x\mathrm{d}y\\&=\int xf(x)yf(y)\mathrm{d}x\mathrm{d}y\quad\text{（根据独立性）}\\&=\int xf(x)\mathrm{d}x\int yf(y)\mathrm{d}y=E(X)E(Y).\end{aligned}$$

注意，只有当 X 和 Y 的联合分布为高斯分布时，上式反之才成立.

方差不像均值一样具有良好的可加性（除非 X 和 Y 相互独立），因此我们需要**协方差**的概念：

$$\mathrm{cov}(X,Y)=E\{(X-\mu_x)(Y-\mu_y)\}=E(XY)-E(X)E(Y),$$

其中，$\mu_x=E(X),\mu_y=E(Y)$. 显然 $\mathrm{var}(X)\equiv\mathrm{cov}(X,X)$，并且如果 X 和 Y 相互独立，那么 $\mathrm{cov}(X,Y)=0$（因为由独立性可得 $E(XY)=E(X)E(Y)$）.

现令 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 分别表示有相同行数并且元素为有限定值的一个矩阵和一个向量，并令 $\boldsymbol{X}$ 表示一个随机向量. 那么 $E(\boldsymbol{X})=\boldsymbol{\mu}_x=\{E(X_1),E(X_2),\cdots,E(X_n)\}^{\mathrm{T}}$，并且随即有 $E(\boldsymbol{AX}+\boldsymbol{b})=\boldsymbol{A}E(\boldsymbol{X})+\boldsymbol{b}$. 对 $\boldsymbol{X}$ 的二阶性质进行一个有效的总结需要综合考虑它的元素的方差和协方差. 这些可以写成（对称的）方差–协方差矩阵 $\boldsymbol{\Sigma}$，其中 $\Sigma_{ij}=\mathrm{cov}(X_i,X_j)$，这意味着

$$\boldsymbol{\Sigma}=E\left\{(\boldsymbol{X}-\boldsymbol{\mu}_x)(\boldsymbol{X}-\boldsymbol{\mu}_x)^{\mathrm{T}}\right\}. \tag{1.4}$$

一个非常有用的结果是

$$\boldsymbol{\Sigma}_{AX+b}=\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}, \tag{1.5}$$

这个结果可简单证明如下：

$$\begin{aligned}\boldsymbol{\Sigma}_{AX+b}&=E\left\{(\boldsymbol{AX}+\boldsymbol{b}-\boldsymbol{A\mu}_x-\boldsymbol{b})(\boldsymbol{AX}+\boldsymbol{b}-\boldsymbol{A\mu}_x-\boldsymbol{b})^{\mathrm{T}}\right\}\\&=E\left\{(\boldsymbol{AX}-\boldsymbol{A\mu}_x)(AX-\boldsymbol{A\mu}_x)^{\mathrm{T}}\right\}\end{aligned}$$