

7

Continuous Distributions

连续分布

分布相当于理想化假设



我们仅仅是，川流不息河水里的，一个个涡漩。肉体灰飞烟灭，潮流浩浩荡荡。

We are but whirlpools in a river of ever-flowing water. We are not the stuff that abides, but patterns that perpetuate themselves.

—— 诺伯特·维纳 (Norbert Wiener) | 美国数学家 | 1894 ~ 1964



- ▶ `numpy.random.laplace()` 拉普拉斯分布随机数发生器
- ▶ `numpy.random.uniform()` 均匀分布随机数发生器
- ▶ `scipy.stats.beta()` Beta 分布
- ▶ `scipy.stats.beta.pdf()` Beta 分布概率密度函数
- ▶ `scipy.stats.chi2()` 卡方分布函数
- ▶ `scipy.stats.dirichlet()` Dirichlet 分布
- ▶ `scipy.stats.dirichlet.pdf()` Dirichlet 分布概率密度函数
- ▶ `scipy.stats.expon()` 指数分布函数
- ▶ `scipy.stats.laplace()` 拉普拉斯分布函数
- ▶ `scipy.stats.logistic()` 逻辑分布函数
- ▶ `scipy.stats.lognorm()` 对数正态分布函数
- ▶ `scipy.stats.norm()` 正态分布函数
- ▶ `scipy.stats.t()` 学生 t-分布函数
- ▶ `seaborn.histplot()` 绘制频率/概率直方图

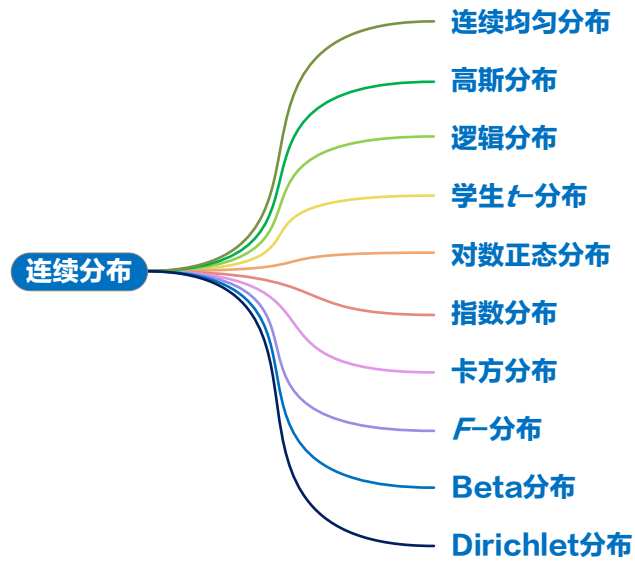
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



7.1 连续均匀分布：离散均匀分布的连续版

概率密度函数

如图 1 所示，连续随机变量 X 在区间 $[a, b]$ 内取得任意一个实数的概率密度函数满足：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (1)$$

则称 X 区间 $[a, b]$ 上服从**连续均匀分布** (continuous uniform distribution)。这个连续分布常记做 $\text{Uniform}(a, b)$ 或 $U(a, b)$ ，比如 $[0, 1]$ 区间上的均匀分布可以记做 $\text{Uniform}(0, 1)$ 或 $U(0, 1)$ 。

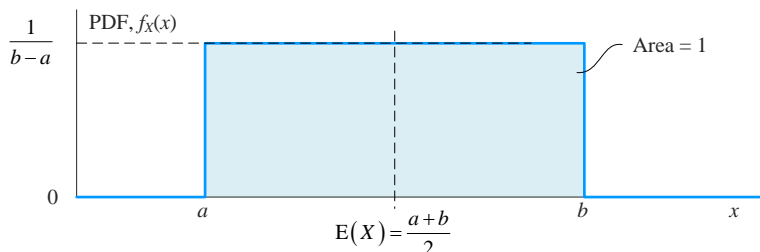


图 1. 随机变量 X 在 $[a, b]$ 上为均匀分布

期望、方差

服从 (1) 连续均匀分布 X 的期望和方差分别为：

$$E(X) = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12} \quad (2)$$

随机数

利用随机数发生器，我们可以获得满足连续均匀分布的随机数。图 2 (a) 所示为满足连续均匀分布随机数的直方图。

图 2 (b) 所示为随机数的**经验累积分布函数** (Empirical Cumulative Distribution Function, ECDF)。不难看出 ECDF 的取值范围为 $[0, 1]$ 。经验分布函数是在所有 n 个样本点上都跳跃 $1/n$ 的阶跃函数。对于某个特定样本，它的 ECDF 为样本中小于或等于该值的样本所占的比例。



我们在本书第 9 章还会提到经验累积分布函数 ECDF。

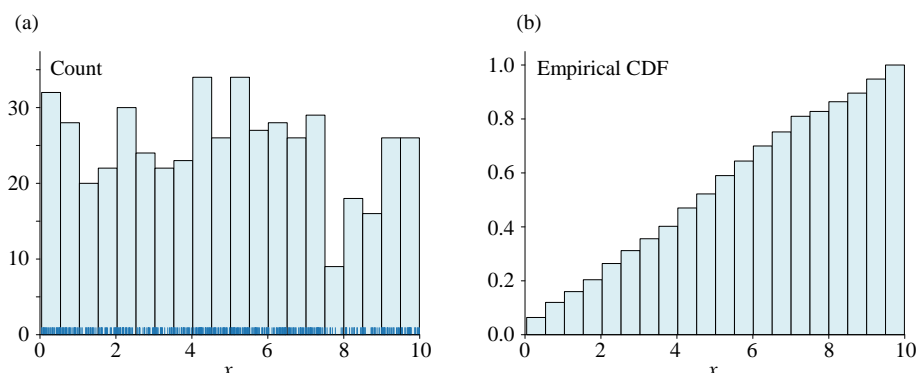


图 2. 满足连续均匀分布的随机数直方图、ECDF



Bk5_Ch07_01.py 代码绘制图 2。

7.2 高斯分布：最重要的概率分布，没有之一

高斯分布 (Gaussian distribution)，也叫**正态分布** (normal distribution)，仿佛是整个纷繁复杂宇宙表象下的终极秩序。实际上，高斯分布是由德国数学家和天文学家**亚伯拉罕·棣莫弗** (Abraham de Moivre) 于 1733 年首先提出。

➔ 高斯分布非常重要，“鸢尾花书”中回归分析、主成分分析、高斯朴素贝叶斯、高斯过程、高斯混合模型等等内容都和高斯分布有着密切的联系。本书第 9 ~ 13 章将从不同角度探讨高斯分布。

一元高斯分布

一元高斯分布 (univariate normal distribution) 的概率密度函数为：

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (3)$$

其中， μ 为均值/期望值， σ 为标准差。满足 (3) 的高斯分布常记做 $N(\mu, \sigma^2)$ 。

也就是说，连续随机变量 X 服从 $N(\mu, \sigma^2)$ ，即 $X \sim N(\mu, \sigma^2)$ ，则 X 的期望和方差为：

$$E(X) = \mu, \quad \text{var}(X) = \sigma^2 \quad (4)$$

图 3 所示为三个不同一元高斯分布 PDF、CDF 图像。可以发现，一元高斯分布 PDF 关于 $x = \mu$ 对称，当 x 远离 μ ，概率密度函数高度迅速下降。

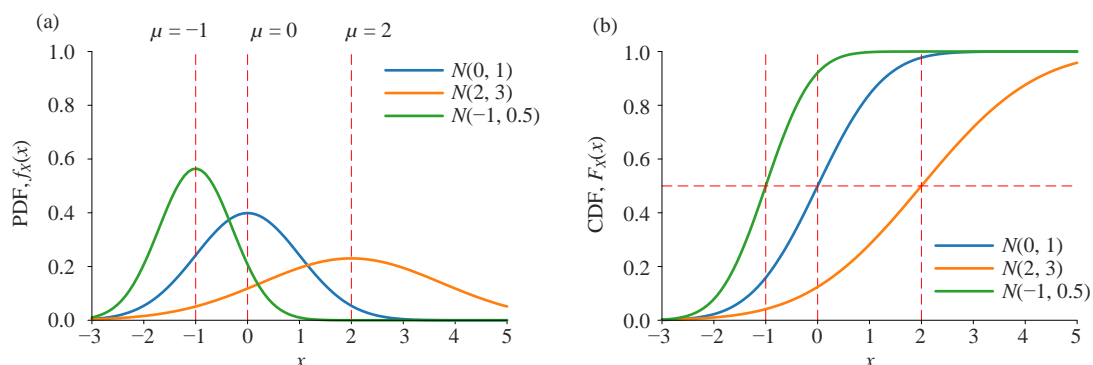


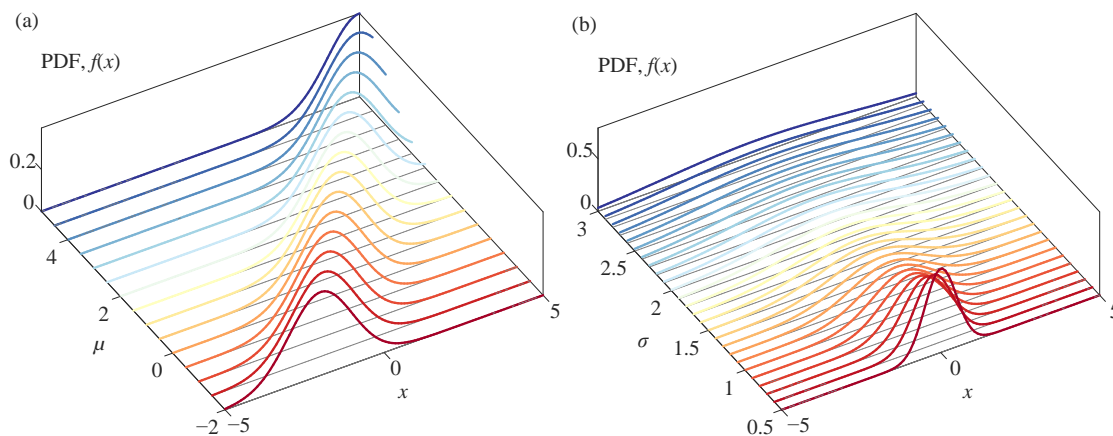
图 3. 三个正态分布 PDF 和 CDF



Bk5_Ch07_02.py 代码绘制图 3。

形状

μ 和 σ 两个参数确定了一元高斯分布 PDF 的位置和形状。如图 4 所示, μ 决定概率密度曲线 $p(x)$ 的位置, σ 影响曲线的胖瘦。特别地, 当 $\mu = 0$, 且 $\sigma = 1$ 时, 得到的高斯分布为**标准正态分布** (standard normal distribution)。

图 4. 均值 μ 和标准差 σ 分别对一元正态分布曲线形状影响

二元高斯分布

二元高斯分布 (bivariate Gaussian distribution), 也叫二元正态分布, 它的概率密度函数解析式如下:

$$f_{x_1, x_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}} \times \exp \left(-\frac{1}{2} \left[\frac{1}{(1-\rho_{1,2}^2)} \left(\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho_{1,2} \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right) \right] \right) \quad (5)$$

本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com

其中， μ_1 和 μ_2 分别为 X_1 和 X_2 的期望值， σ_1 和 σ_2 为 X_1 和 X_2 的标准差， $\rho_{1,2}$ 为两者的线性相关系数。

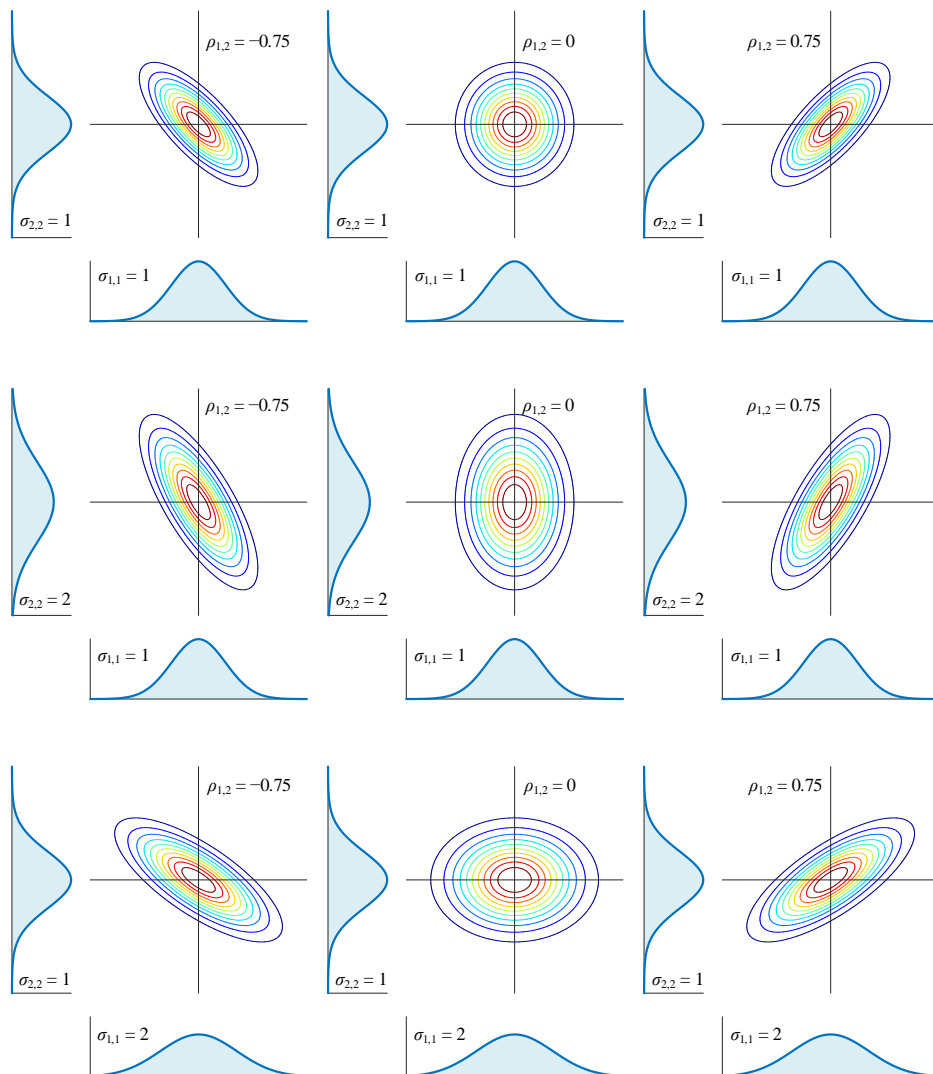
▲ 注意，上式中 $\rho_{1,2}$ 取值范围为 $(-1, 1)$ 。

➡ 相信大家已经在上式中看到椭圆！这是本书后续重要的线索之一。此外，我们在《数学要素》第 9 章专门介绍过这种椭圆形式。

连续随机变量 (X_1, X_2) 服从上述二元正态分布，记做：

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \underbrace{\begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}}_{\Sigma} \right) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (6)$$

图 5 所示为方差和相关性系数取不同值时，二元正态分布概率密度函数椭圆等高线以及边缘分布形状。注意，图中 $\sigma_{1,1}$ 和 $\sigma_{2,2}$ 代表方差，即标准差的平方。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 5. 方差和相关性系数取不同值时，二元正态分布概率密度函数椭圆等高线形态



本书第 10 章将专门以椭圆为视角讲解二元正态分布。

多元高斯分布

《矩阵力量》第 20 章用如下公式介绍过**多元高斯分布** (multivariate Gaussian distribution)，请大家据此回忆多元高斯分布 PDF 每个不同成分的含义：

$$\begin{aligned}
 d &= \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} && \text{Mahal distance} \\
 \|z\| &&& \text{z-score} \\
 z &= \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) && \text{Translate} \rightarrow \text{rotate} \rightarrow \text{scale} \\
 \left[\mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) \right]^T \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) &&& \text{Eigen decomposition} \\
 (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &&& \text{Ellipse/ellipsoid} \\
 f_x(\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \\
 &\downarrow \text{Distance} \rightarrow \text{similarity} && \downarrow \text{Normalization} \\
 &&& \downarrow \text{Multivariable calculus} \\
 &&& \downarrow \text{Scaling Eigenvalues}
 \end{aligned} \tag{7}$$



本书第 11 章深入讲解多元高斯分布。

拉普拉斯分布：

本节最后简要介绍**拉普拉斯分布** (Laplace distribution)。拉普拉斯分布的概率密度函数为：

$$f_x(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \tag{8}$$

形式上，拉普拉斯分布和高斯分布很类似，只不过拉普拉斯分布的 PDF 图像在对称轴处存在尖点。很容易发现，参数 μ 决定概率密度分布位置。如图 6 所示，参数 b 决定分布形状。

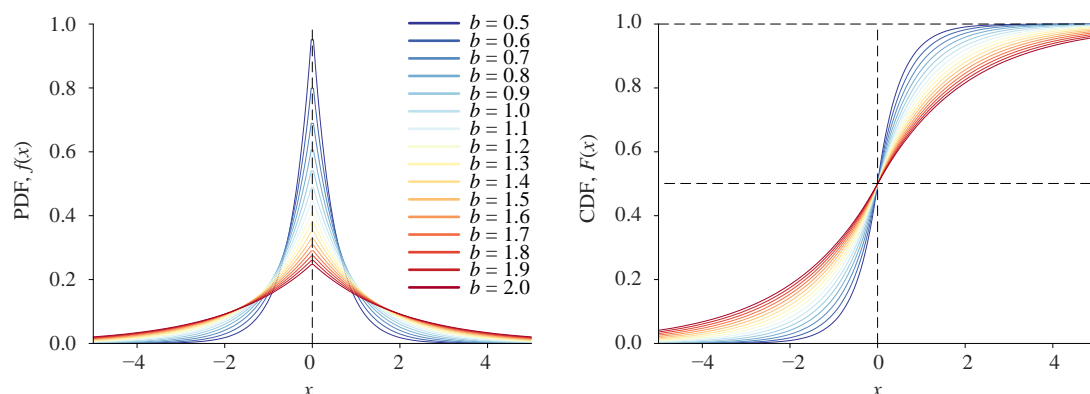


图 6. 拉普拉斯分布的 PDF 和 CDF

如果连续随机变量 X 满足 (8) 拉普拉斯分布, X 期望和方差为:

$$E(X) = \mu, \quad \text{var}(X) = 2b^2 \quad (9)$$

两个常用的拉普拉斯分布函数为 `scipy.stats.laplace()` 和 `numpy.random.laplace()`。



《数学要素》第 12 章分别讲解过高斯函数和拉普拉斯函数, 建议大家回顾。

7.3 逻辑分布: 类似高斯分布

一元逻辑分布 (univariate logistic distribution) 的 PDF 为:

$$f_X(x) = \frac{\exp\left(\frac{-(x-\mu)}{s}\right)}{s \left(1 + \exp\left(\frac{-(x-\mu)}{s}\right)\right)^2} \quad (10)$$

其中, μ 为位置参数, s 为形状参数。

相比 PDF, 逻辑函数的 CDF 更常用:

$$F_X(x) = \frac{1}{1 + \exp\left(\frac{-(x-\mu)}{s}\right)} \quad (11)$$

图 7 所示为逻辑函数的 PDF 和 CDF 曲线随 b 变化。

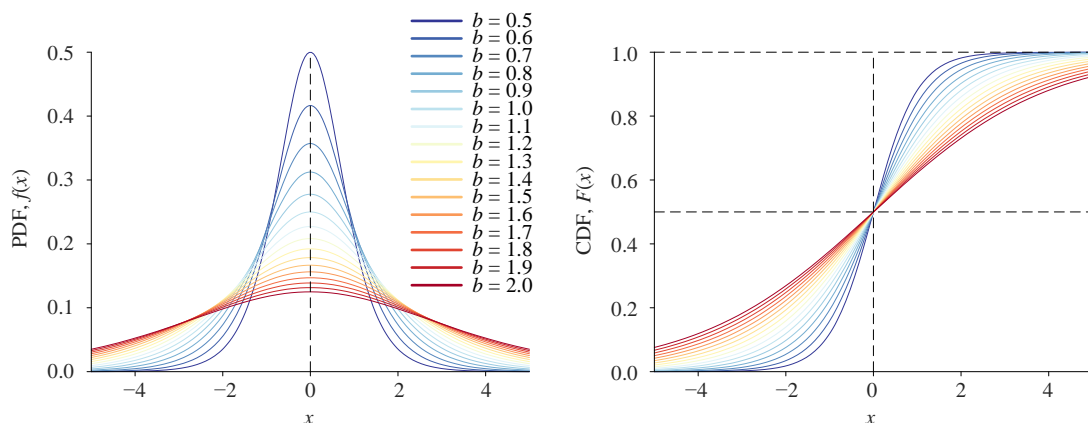


图 7. 逻辑分布 PDF 和 CDF

逻辑分布 vs 高斯分布

大家肯定已经发现，逻辑分布和高斯分布 PDF、CDF 长得很相似。为了比较逻辑函数和高斯函数，我们用标准正态分布 $N(0, 1)$ 的 PDF 和 CDF 图像，而逻辑分布的位置参数 $\mu = 0$ 。特别选取参数 s 使得逻辑分布 PDF 和标准正态分布 PDF 在 $x = 0$ 处高度一致。

如图 8 所示，相比标准正态分布，逻辑分布 PDF 中心部位“稍瘦”，而**厚尾** (fat tail)。厚尾，也叫肥尾，指的是和正态分布相比，尾部分布较厚的分布。下一节介绍的学生 t -分布就是典型的厚尾分布。

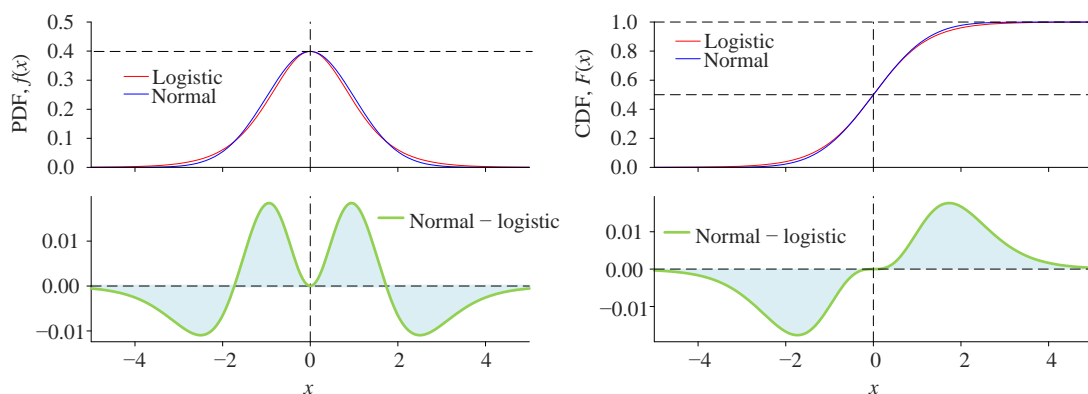


图 8. 比较逻辑函数和高斯函数



Bk5_Ch07_03.py 代码绘制图 7。

7.4 学生 t -分布：厚尾分布

学生 t -分布 (Student's t -distribution) 也称**学生分布**，或 t 分布，是由**戈赛特** (William Sealy Gosset) 于 1908 年提出的，Student 一词源自于他发表论文时用的化名。

学生 t -分布是常用的一类厚尾分布。学生 t -分布多应用于根据小样本数据来估计呈正态分布且方差未知的总体的均值，本书第 17 章将简要介绍相关内容。

一元学生 t -分布的 PDF 为：

$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad (12)$$

其中， ν 为**自由度** (number of degrees of freedom 或 df)， $\nu = n - 1$ ， n 为样本数； Γ 是 **Gamma 函数** (Gamma function)。

Gamma 函数

Gamma 函数是从阶乘的概念推广而来的，它将阶乘的概念推广到了实数和复数的范围。

ν 为正整数时，Gamma 方程类似于阶乘表达式，正整数 ν 的 Gamma 函数表达式为：

$$\Gamma(\nu) = (\nu - 1)! \quad (13)$$

ν 取特殊分数，比如 $1/2$ 和 $3/2$ 时， ν 的 Gamma 函数的值：

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi} \\ \Gamma\left(\frac{3}{2}\right) &= \frac{1}{2}\sqrt{\pi} \end{aligned} \quad (14)$$

图 9 所示为 Gamma 函数图像，其中红色 \times 是取正整数时 Gamma 函数的取值。

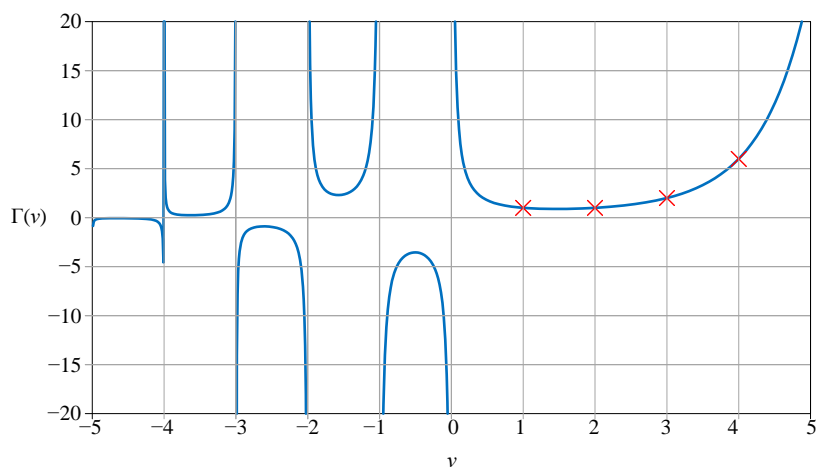


图 9. Gamma 函数图像

一般情况，当 ν 为偶数时，(12) 中系数部分为：

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2} \quad (15)$$

当 ν 为奇数时：

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3} \quad (16)$$

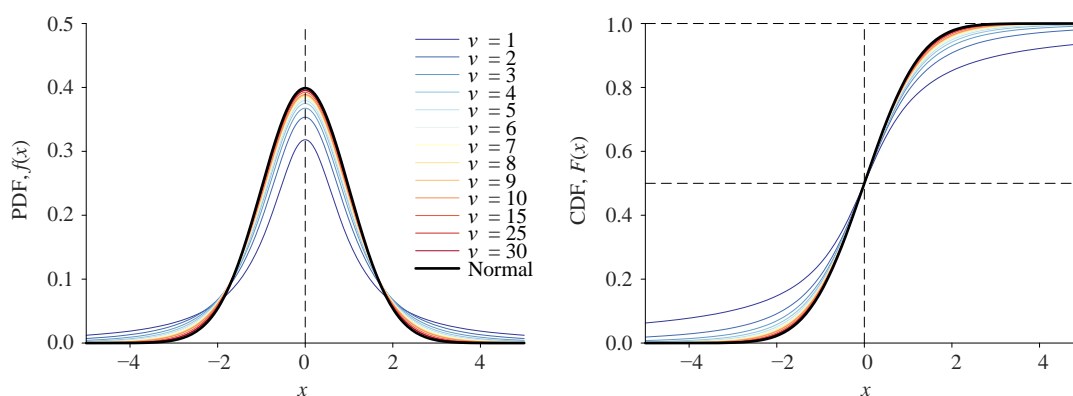
Gamma 函数存在如下递推关系：

$$\Gamma(\nu+1) = \Gamma(\nu) \cdot \nu \quad (17)$$

上式和 ν 取值无关。Gamma 函数在概率分布中具有重要的作用，尤其是在 Gamma 分布、卡方分布、 t 分布、Beta 分布、Dirichlet 分布等定义和性质中都涉及到 Gamma 函数。

自由度

图 10 所示为 ν 从 1 变化到 30 时，学生 t -分布 PDF 和 CDF 图像。图 10 中黑色的曲线对应正态分布。当自由度 ν 不断提高时，厚尾现象逐渐消失，学生 t -分布逐渐接近标准正态分布（黑色）。很明显，学生 t -分布的偏度为 0。

图 10. 学生 t -分布 PDF 和 CDF 随自由度变化

Bk5_Ch07_04.py 代码绘制图 10。

多元学生 t -分布

类似 (7) 给出的多元高斯分布，多元学生 t -分布的概率密度函数为：

$$f_{\mathbf{z}}(\mathbf{x}) = \frac{\Gamma[(\nu + D)/2]}{\Gamma(\nu/2)\nu^{D/2}\pi^{D/2}|\Sigma_t|^{1/2}} \left[1 + \frac{1}{\nu} \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma_t^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{\text{Ellipse}} \right]^{-(\nu+D)/2} \quad (18)$$

其中， ν 为自由度， D 为维数。相信大家在上式中也看到了椭圆。

上式中 Σ_t 和多元高斯分布的协方差矩阵关系为：

$$\Sigma_t = \frac{\nu}{\nu - 2} \Sigma \quad (19)$$

7.5 对数正态分布：源自正态分布

定义

如果随机变量 X 的对数 $\ln X$ 服从正态分布，则 X 服从**对数正态分布** (logarithmic normal distribution)。

对于 $x > 0$ ，对数正态分布的 PDF 为：

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (20)$$

其中， μ 是 X 对数的平均值， σ 是 X 对数的标准差。

如果 X 满足 (20) 的对数正态分布，则 X 期望和方差为：

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \text{var}(X) = [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2) \quad (21)$$

图像

图 11 给出对数正态分布的图像。对数正态分布的最大特点是右偏，即正偏。对于右偏的对数正态分布，其平均值大于其众数。



大家将会在《数据有道》一册看到对数正态分布的应用。

▲ 再次强调，对数正态分布的随机变量取值只能为正值。

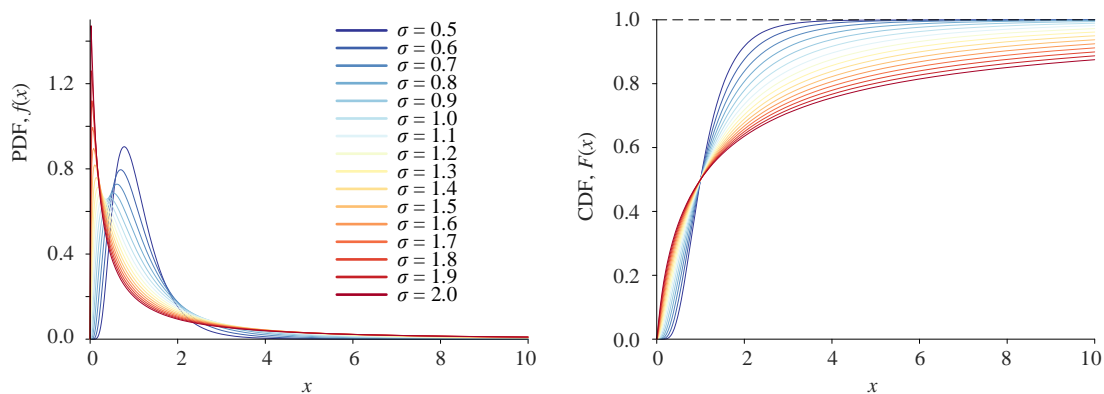


图 11. 对数正态分布的 PDF 和 CDF

图 12 对比正态分布和对数正态分布。

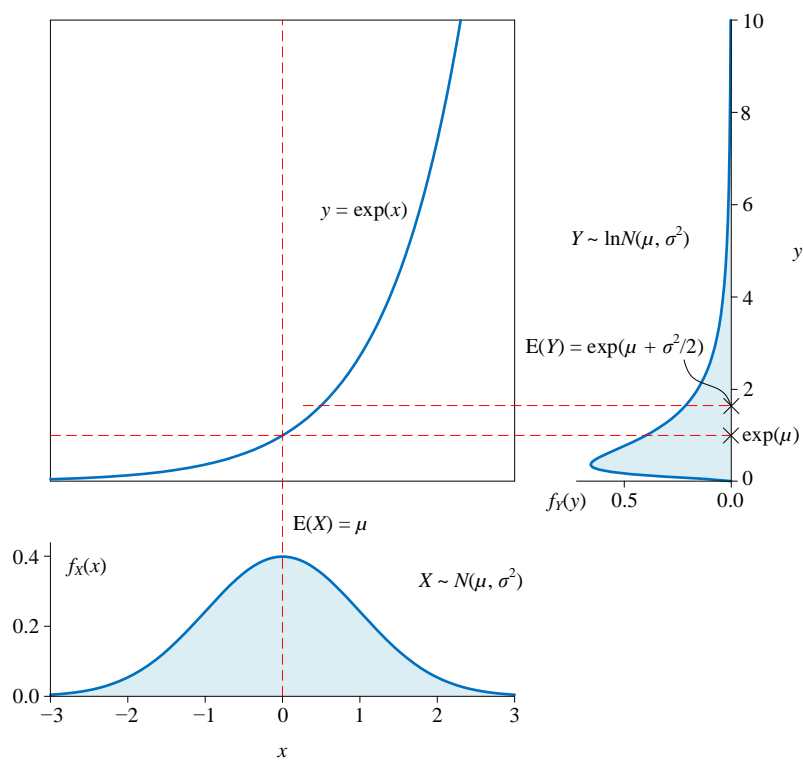


图 12. 比较正态分布和对数正态分布



Bk5_Ch07_05.py 代码绘制图 11。Bk5_Ch07_06.py 代码绘制图 12

7.6 指数分布：泊松分布的连续随机变量版

定义

指数分布 (exponential distribution) 和本章第 5 章介绍的泊松分布息息相关。

与泊松分布相比，指数分布重要特点是随机变量连续。而泊松分布是针对随机事件发生次数定义的，发生次数是离散的。

指数分布的概率密度函数为：

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (22)$$

指数分布的期望和方差分别为：

$$E(X) = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2} \quad (23)$$

图像

图 13 所示为 λ 取不同值时，指数分布 PDF 和 CDF 图像。

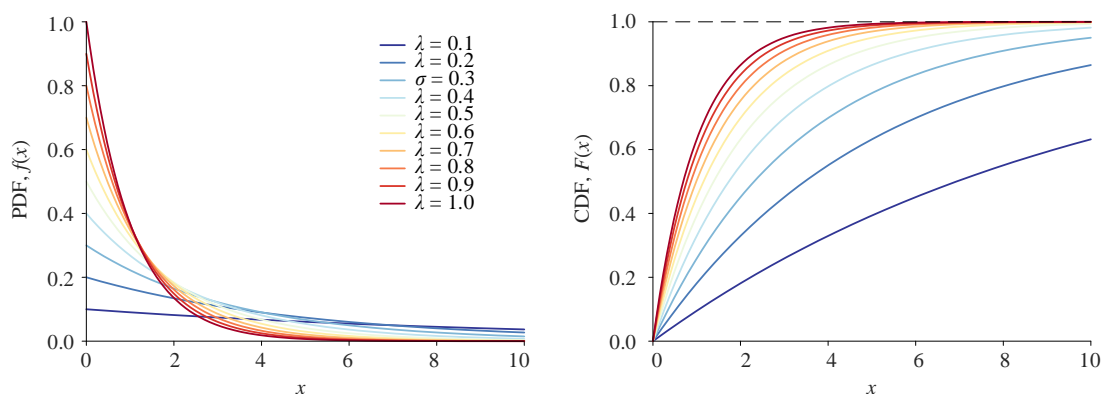


图 13. λ 取不同值时，指数分布 PDF 和 CDF 图像



Bk5_Ch07_07.py 代码绘制图 13。

7.7 卡方分布：若干 IID 标准正态分布平方和

定义

卡方分布 (chi-square distribution 或 χ^2 -distribution) 先是德国统计学家**赫尔默特** (Friedrich Robert Helmert) 在 1875 年提出。

若 n 个相互独立的随机变量 Z_1, Z_2, \dots, Z_k 均服从标准正态分布，即：

$$Z_i \sim N(0,1), \quad \forall i=1,\dots,k \quad (24)$$

这 n 个随机变量的平方和构成一个新的随机变量 X ， X 服从自由度为 k 的卡方分布：

$$X = \sum_{i=1}^k Z_i^2 \sim \chi_k^2 \quad (25)$$

其中， k 称为自由度。自由度为 k 的卡方分布一般标记为 χ_k^2 。

如果随机变量 X 满足 (25) 的卡方分布， X 的期望值和方差为：

$$E(X) = k, \quad \text{var}(X) = 2k \quad (26)$$

图像

如图 14 所示，卡方分布的值均为正值，且呈现右偏态，随着自由度 n 的增大，卡方分布趋近于正态分布。当自由度大于 30 时，已经非常类似于正态分布。

不知道大家看到 (25)，是否想到马氏距离的平方？



我们将在本书第 23 章讲解马氏距离时用到卡方分布。

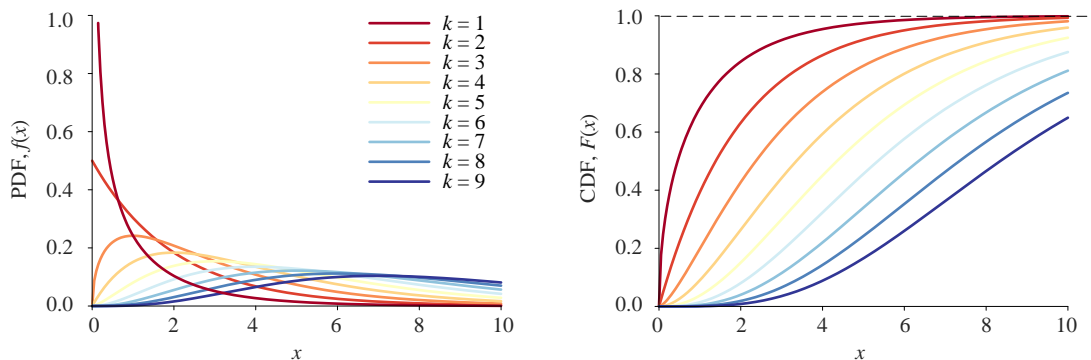


图 14. 卡方分布 PDF 和 CDF



Bk5_Ch07_08.py 代码绘制图 14。

7.8 F -分布：和两个服从卡方分布的独立随机变量有关

定义

F -分布是两个服从卡方分布的独立随机变量各除以其自由度后的比值的抽样分布。

如果随机变量 X 满足参数为 d_1 和 d_2 的 F -分布，记做 $X \sim F(d_1, d_2)$ 。随机变量 X 为：

$$X = \frac{S_1/d_1}{S_2/d_2} \quad (27)$$

其中，随机变量 S_1 和 S_2 分别服从自由度为 d_1 、 d_2 的卡方分布。

如果 $X \sim F(d_1, d_2)$ ， X 的 PDF 为：

$$f_X(x; d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{(d_1+d_2)}{2}} \quad (28)$$

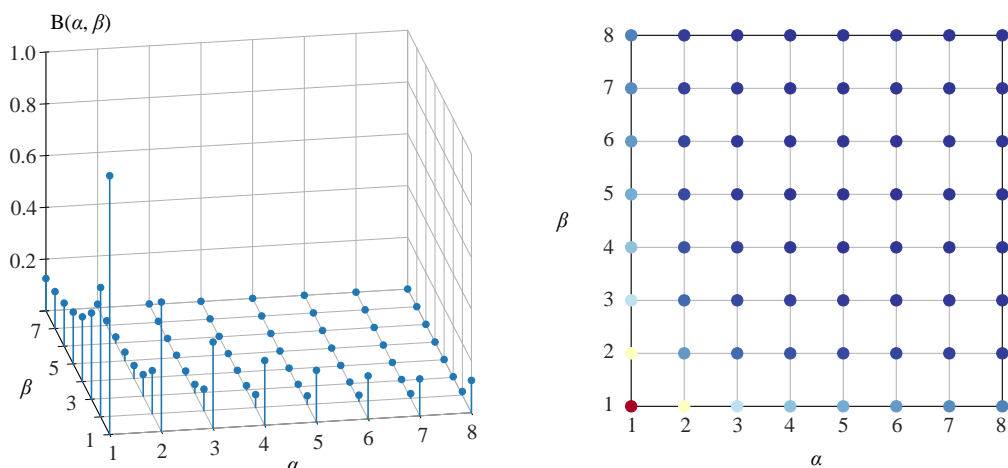
其中， $B()$ 叫做 Beta 函数。 $B(\alpha, \beta)$ 函数和 Gamma 函数的关系为：

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (29)$$

请大家特别注意上式中的积分式，我们将在本书第 21 章讲解贝叶斯推断时用到这个积分式。

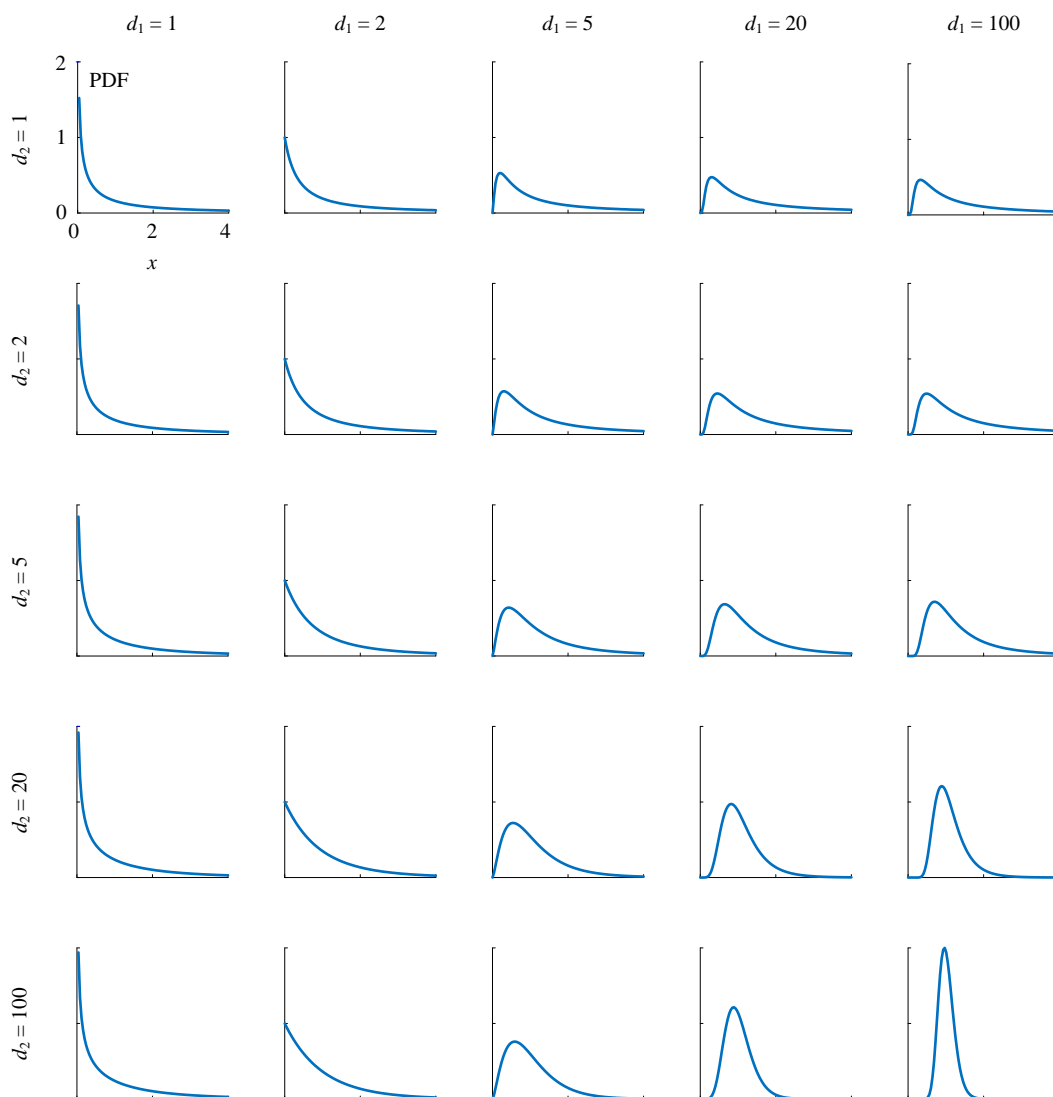
图像

图 15 所示为 $B(\alpha, \beta)$ 函数取值随 α 和 β 变化的火柴梗图、三维散点图。下一节的 Beta 分布中也会用到 $B(\alpha, \beta)$ 函数。

图 15. $B(\alpha, \beta)$ 函数取值火柴梗图、三维散点图

如图 16 所示, F -分布是一种非对称分布, 且 d_1 、 d_2 的位置不可随意互换。

在“鸢尾花书”中, F -分布将用在《数据有道》中的**方差分析** (analysis of variance, ANOVA) 和线性回归显著性检验。

图 16. F 分布 PDF 形状随 d_1 和 d_2 变化

Bk5_Ch07_09.py 代码绘制图 16。

7.9 Beta 分布：概率的概率

贝叶斯推断 (Bayesian inference) 是数据科学和机器学习重要的数学工具，而 Beta 分布在贝叶斯推断中扮演重要角色。

定义

Beta 分布定义在 $(0, 1)$ 或 $[0, 1]$ 区间的连续概率分布，它有两个参数 α 、 β 。Beta(α, β) 分布的概率密度函数为：

$$f_x(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (30)$$

其中， $x^{\alpha-1}(1-x)^{\beta-1}$ 决定 PDF 曲线的形状。

大家可能已经注意到，这个 PDF 概率密度曲线有两个区间，原因是 α 、 β 当取不同值时 x 的取值范围不同。举个例子，当 α 、 β 均为 0.1 时，Beta 分布的定义域为 $(0, 1)$ 。

相信大家已经在上述解析式中看到了 $B(\alpha, \beta)$ 函数。利用 $B(\alpha, \beta)$ ，(30) 可以写成：

$$f_x(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (31)$$

而 $B(\alpha, \beta)$ 是让 $x^{\alpha-1}(1-x)^{\beta-1}$ 成为概率密度函数的归一化因子。白话说， $B(\alpha, \beta)$ 让 PDF 曲线和横轴围成的图形面积为 1。

如果 α 、 β 都是大于 1 的正整数， $B(\alpha, \beta)$ 可以展开写成：

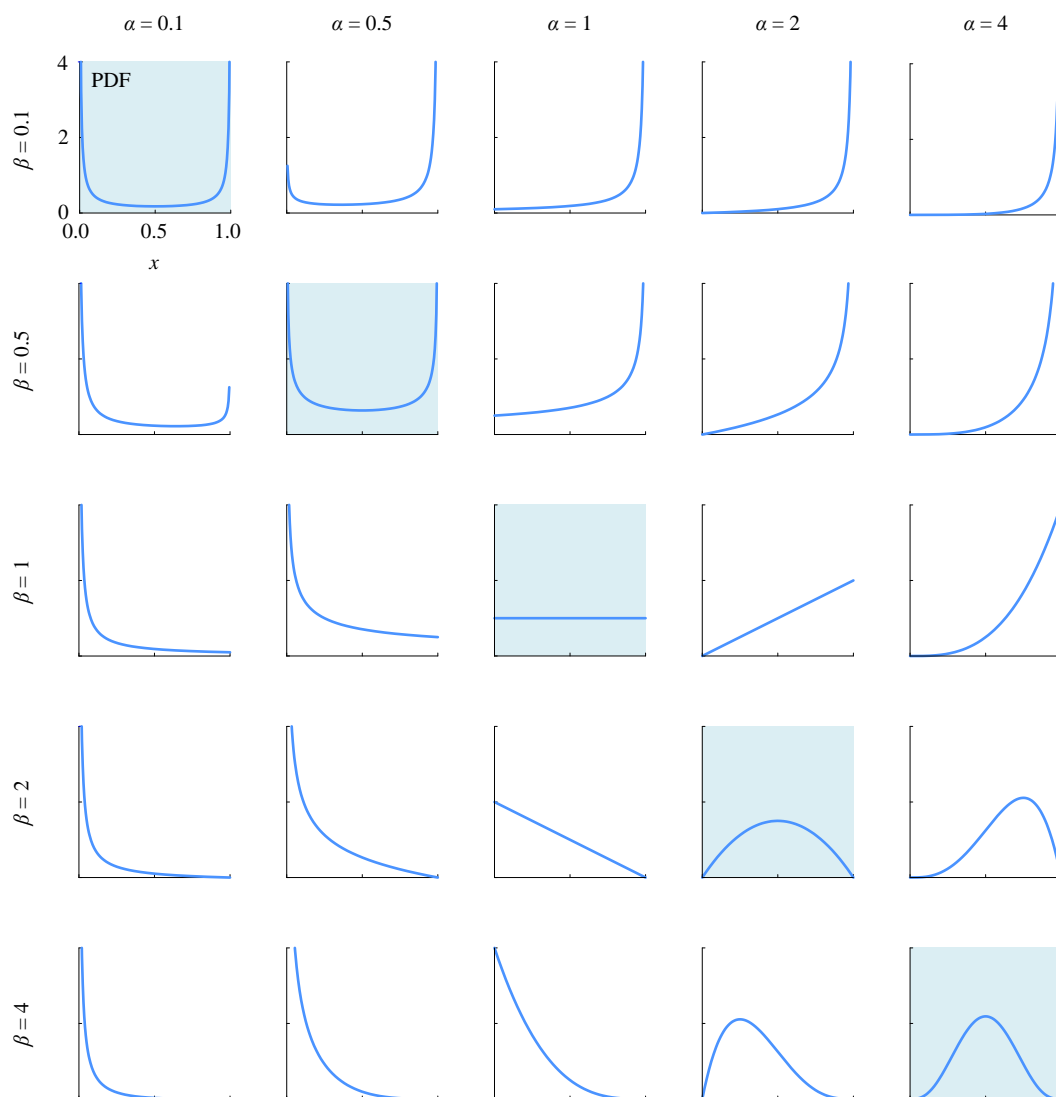
$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \quad (32)$$

图像

图 17 所示为参数 α 、 β 取不同值时 Beta 分布 PDF 图像。

容易发现 Beta(α, β) 分布实际上代表了一系列分布。举个例子，连续均匀分布 $U(0, 1)$ 便是 Beta(1, 1)。

请大家特别注意图 17 对角线上的图像，即 $\alpha = \beta$ ，这些 PDF 图像对称，对应的分布相当于 Beta(α, α)。本书第 21 章将用到 Beta(α, α) 这个分布。

图 17. 参数 α 、 β 取不同值时 $\text{Beta}(\alpha, \beta)$ 分布 PDF 图像

Bk5_Ch07_10.py 绘制图 17。代码还绘制 $\text{Beta}(\alpha, \beta)$ 分布的 CDF 图像。



在 Bk5_Ch07_10.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 $\text{Beta}(\alpha, \beta)$ 两个参数值，观察 PDF 曲线变化。请大家参考 Streamlit_Bk5_Ch07_10.py。此外，请大家选取前文某个概率分布，做一个类似的 App。

众数 vs 期望

如果 X 服从 $\text{Beta}(\alpha, \beta)$ 分布， X 的期望为：

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad (33)$$

我们常常用到的是 $\text{Beta}(\alpha, \beta)$ 分布的众数：

$$\frac{\alpha - 1}{\alpha + \beta - 2}, \quad \alpha, \beta > 1 \quad (34)$$

众数是概率密度函数曲线最大值所在位置。这一点在本书后文的贝叶斯推断格外重要，请大家注意。

推导期望

推导 $\text{Beta}(\alpha, \beta)$ 的期望其实很容易，我们甚至不需要积分。

连续随机变量 X 的期望为：

$$E(X) = \int_x x \cdot f_X(x) dx \quad (35)$$

将 $\text{Beta}(\alpha, \beta)$ 的概率密度函数代入上式，得到：

$$\begin{aligned} E(X) &= \int_x x \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_x x^{\alpha} (1-x)^{\beta-1} dx}_{\text{Beta}(\alpha+1, \beta)} \end{aligned} \quad (36)$$

容易看出来，上式中积分部分可以整理成为 $\text{Beta}(\alpha + 1, \beta)$ 分布的 PDF 解析式。缺的就是归一化系数。

补充这个归一化系数，上式可以写成：

$$\begin{aligned} E(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha + \beta + 1)} \underbrace{\int_x \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + 1)\Gamma(\beta)} x^{\alpha} (1-x)^{\beta-1} dx}_{=1} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha + \beta + 1)} \end{aligned} \quad (37)$$

根据 Gamma 函数的递推关系 $\Gamma(\nu + 1) = \Gamma(\nu) \cdot \nu$ ，上式进一步整理为：

$$\begin{aligned} E(X) &= \frac{\cancel{\Gamma(\alpha + \beta)}}{\cancel{\Gamma(\alpha)} \cancel{\Gamma(\beta)}} \frac{\cancel{\Gamma(\alpha)} \cdot \alpha \cdot \cancel{\Gamma(\beta)}}{\cancel{\Gamma(\alpha + \beta)} \cdot (\alpha + \beta)} \\ &= \frac{\alpha}{\alpha + \beta} \end{aligned} \quad (38)$$

方差、标准差

Beta(α, β) 的方差为：

$$\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (39)$$

Beta(α, β) 的标准差为方差的平方根：

$$\text{std}(X) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad (40)$$

为了方便和下文的 Dirichlet 分布对照，令

$$\alpha_0 = \alpha + \beta \quad (41)$$

Beta(α, β) 的可以进一步写成：

$$\begin{aligned} \text{var}(X) &= \frac{\alpha(\alpha_0 - \alpha)}{\alpha_0^2(\alpha_0 + 1)} \\ &= \frac{\frac{\alpha}{\alpha_0} \left(1 - \frac{\alpha}{\alpha_0}\right)}{\alpha_0 + 1} \end{aligned} \quad (42)$$

7.10 Dirichlet 分布：多元 Beta 分布

Dirichlet 分布也叫狄利克雷分布，它本质上是**多元 Beta 分布** (multivariate Beta distribution)。Dirichlet 分布常作为贝叶斯统计的先验概率。

Dirichlet 分布概率密度函数为：

$$f_{X_1, \dots, X_K}(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i - 1}, \quad \sum_{i=1}^K x_i = 1 \quad (43)$$

注意， x_i ($i = 1, 2, \dots, K$) 的取值范围为 $[0, 1]$ ，而且它们的和为 1。这个分布常记做 $\text{Dir}(\alpha)$ 或 $\text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_K)$ 。本书后文在贝叶斯推断中，会用 θ 代替 x 。

K 元 $B()$ 函数的定义为：

$$B(\alpha_1, \dots, \alpha_K) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \quad (44)$$

举个例子

当 $K = 3$ 时, x_1 、 x_2 、 x_3 满足:

$$x_1 + x_2 + x_3 = 1 \quad (45)$$

并且, x_1 、 x_2 、 x_3 都在区间 $[0, 1]$ 内。显然, x_1 、 x_2 、 x_3 在一个平面上。

白话说, $x_1 + x_2 + x_3 = 1$ 好比三维空间撑起的一张“画布”, 概率密度等高线则必须画在这张画布上。

本节下文将采用五种可视化方案展示 Dirichlet 分布概率密度函数。如图 18 所示, 这五种可视化方案主要分成两大类。由于 (45) 等式关系, 给定 x_1 、 x_2 , 则 x_3 确定。因此, 我们可以用图 18 (a) 的 x_1x_2 平面展示 Dirichlet 分布 PDF 图像。

此外, 我们还可以用图 18 (b) 所示的可视化方案。这实际上是**重心坐标系** (barycentric coordinate system)。



“鸢尾花书”《可视之美》专门讲解过重心坐标系, 请大家参考。

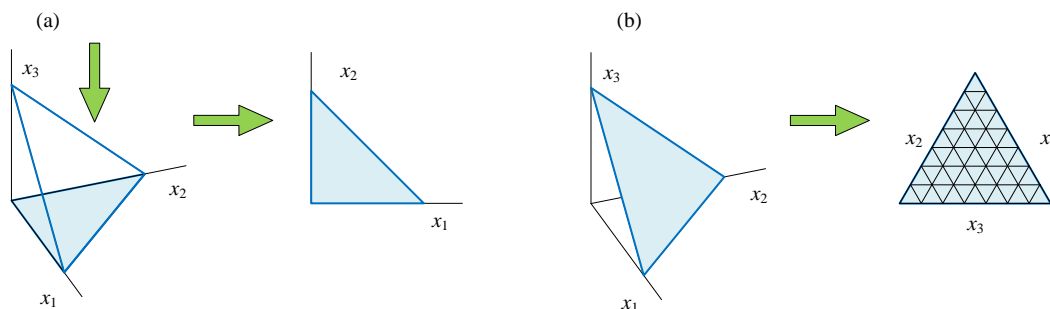


图 18. 可视化方案原理

Dirichlet 分布非常重要, 因此我们下文用图 19~图 23 五种可视化方案展示 Dirichlet 分布的分布特征。

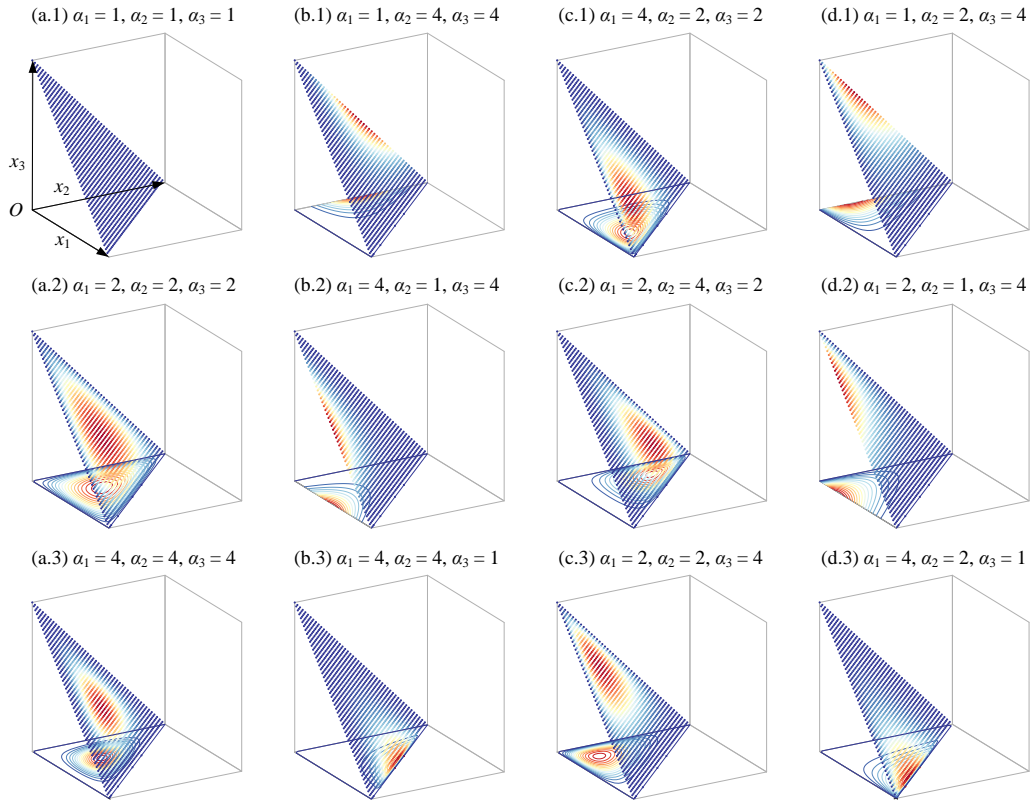
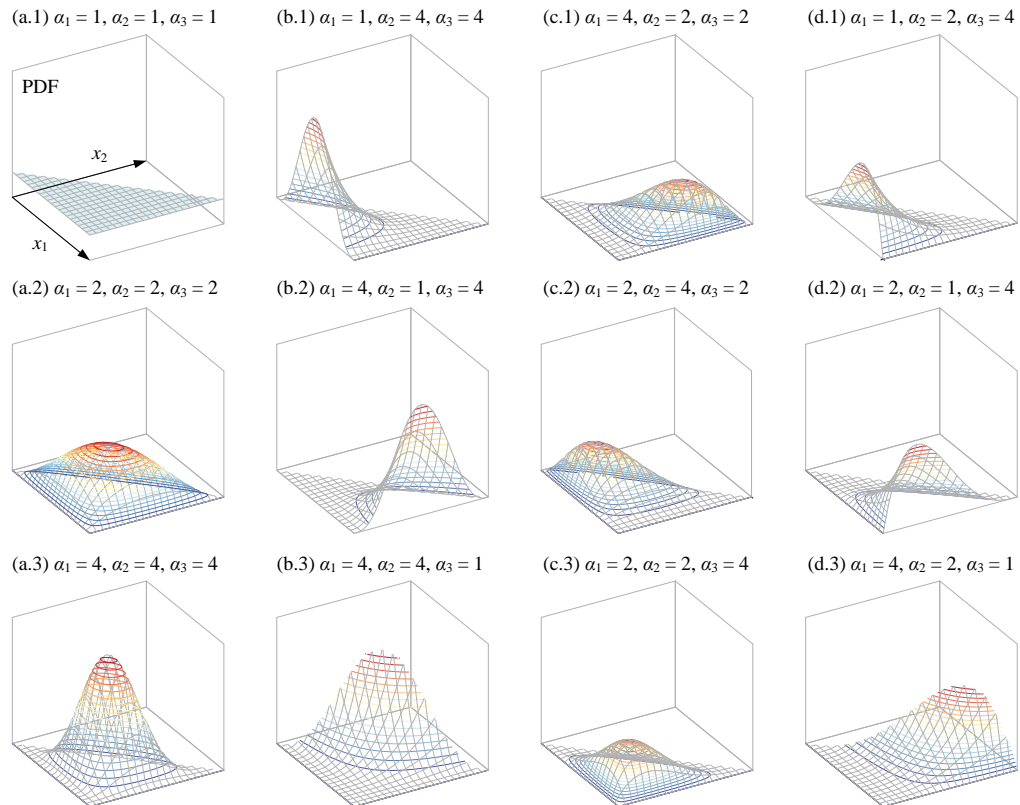


图 19. 用涂色三维散点可视化 Dirichlet 分布图像



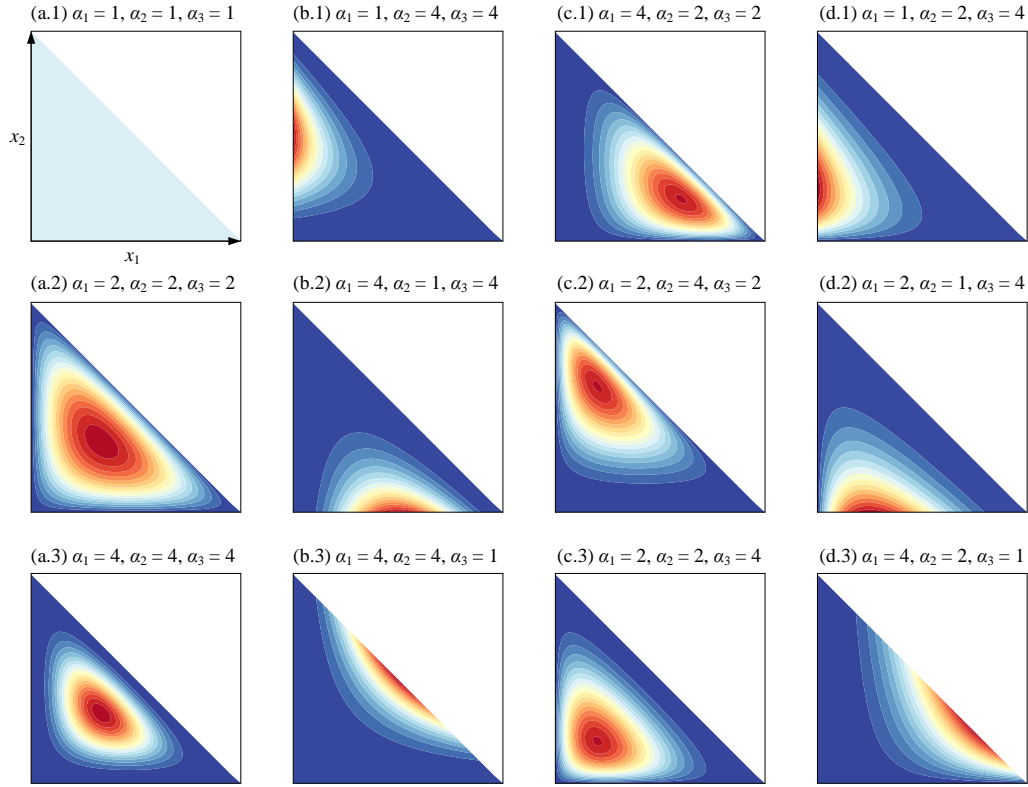
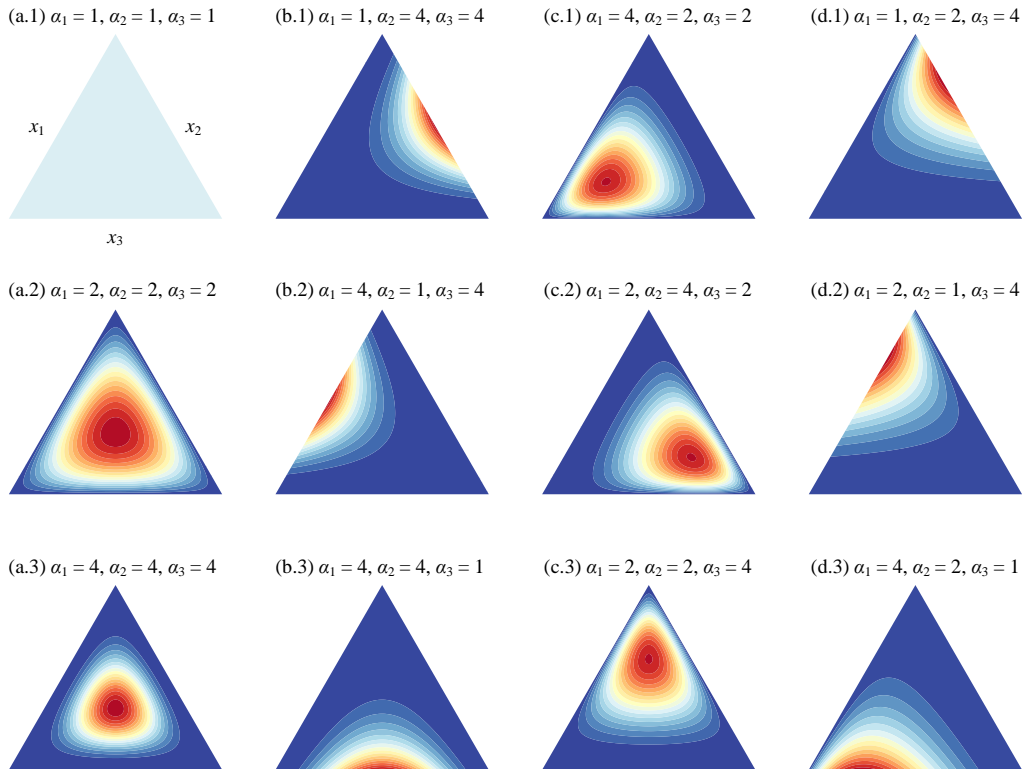
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 20. 基于 x_1x_2 平面的 Dirichlet 分布 PDF 三维等高线, z 轴为 PDF 取值图 21. x_1x_2 平面等高线中的 Dirichlet 分布 PDF 等高线

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 22. 重心坐标系中的 Dirichlet 分布 PDF 等高线

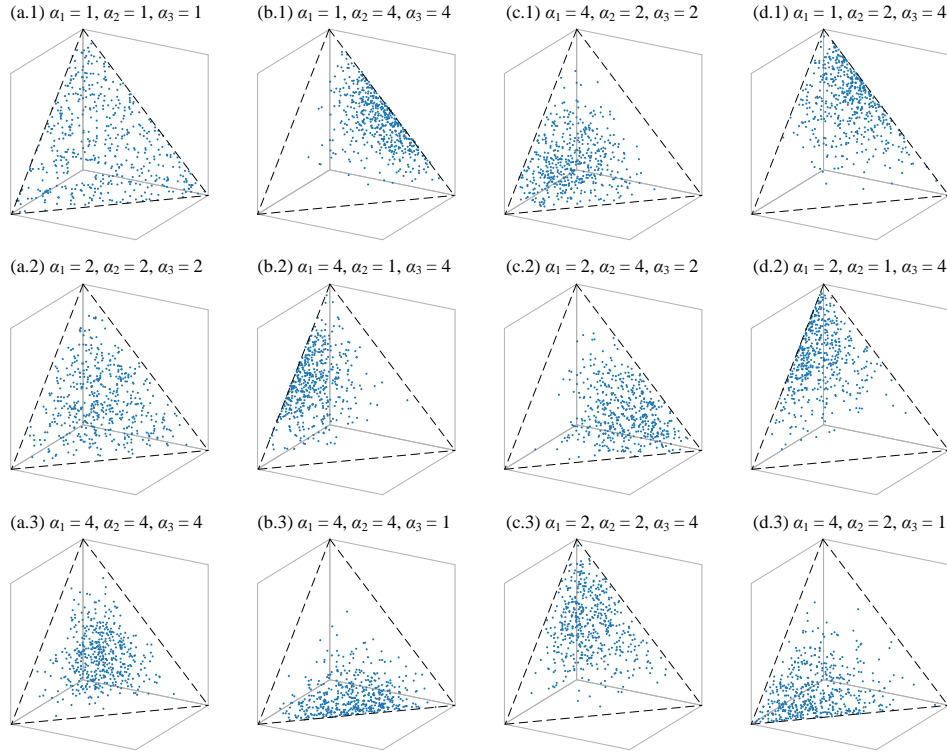


图 23. 满足 Dirichlet 分布的随机数

边缘分布

Dirichlet 分布的边缘分布服从 Beta 分布：

$$X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i) \quad (46)$$

其中：

$$\alpha_0 = \sum_{i=1}^K \alpha_i \quad (47)$$

以图 19 中 (d) 组为例，三个 Dirichlet 分布的边缘分布 PDF 如图 24 所示。

X_i 的期望为：

$$E(X_i) = \frac{\alpha_i}{\sum_{k=1}^K \alpha_k} = \frac{\alpha_i}{\alpha_0} \quad (48)$$

X_i 的众数为：

$$\frac{\alpha_i - 1}{\sum_{k=1}^K \alpha_k - K} = \frac{\alpha_i - 1}{\alpha_0 - K}, \quad \alpha_i > 1 \quad (49)$$

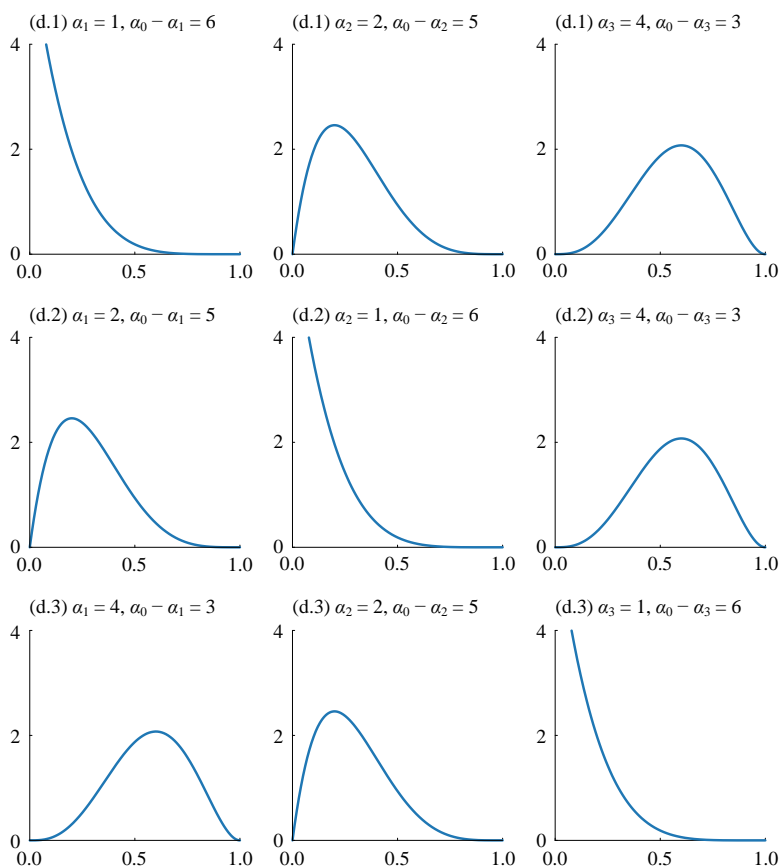


图 24. 三个 Dirichlet 分布的边缘分布



Bk5_Ch07_11.py 绘制图 19、图 20、图 21、图 23、图 24。



在 Bk5_Ch07_11.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 Dirichlet($\alpha_1, \alpha_2, \alpha_3$) 三个参数值，观察 PDF 曲面变化。请大家参考 Streamlit_Bk5_Ch07_11.py。



《统计至简》一册整体来看，高斯分布更为重要，但是它不是本章的重点。这一章最重要的分布有两个——Beta 分布、Dirichlet 分布。它俩分别对应本书第 5 章的二项分布、多项分布。这四个分布在本书后续贝叶斯推断中将扮演重要角色。