概率论与数理统计

授课教师：唐宏岩

前言

本讲义基于清华大学数学系唐宏岩老师于 2023-2024 学年秋季学期开设的《概率论与数理统计》课程，用于辅助同学们课后复习，助教尽量做到每周课后两天内更新。

由于时间与能力所限，本讲义可能不会出现大段的文字论述（但会包含重要的定义、定理与公式等）。但是，对许多基本概念的深入理解是非常有必要的，同学们可以在浏览时检查自己是否能够回忆起课上的内容，对掌握不够扎实的地方，鼓励大家查阅参考书或在课程群提问以解决问题。

由于此为教学团队第一年尝试整理讲义，诸如格式编排、内容完整性方面可能存在许多不足，欢迎大家联系我提出宝贵的意见与建议。

曹子尧

2023 年 9 月

目录

第一部分

初等概率论

第一章　事件的概率

1.1　概率的发展史

赌博中的 de Méré's Problem: 连续掷一个均匀六面骰 4 次，获得至少一次 “6” 的概率为 $1-(\frac{5}{6})^4 \approx 0.5177$；而连续掷两个均匀六面骰 24 次，获得至少一次 “对 6” 的概率为 $1-(35/36)^{24} \approx 0.4914$。

Pascal 和 Fermat 的通信中使用初等数学的方法，首创了概率论相当多的数学理论，虽然当时没有总结成通用的定理。

Laplace 创立了采用分析方法的分析概率论。

Kolmogorov 利用测度论方法发展了现代概率理论。

1.2　随机试验与事件

定义 1.1. 概率论中的随机试验指的是符合下面两个特点的试验：

1. 不能预先确知结果
2. 可以预测所有可能的结果

定义 1.2. 样本空间是指一个试验的所有可能结果的集合，常用 Ω 表示。

定义 1.3. 事件是样本空间的一个良定义的子集。

一次随机试验中，一个事件可能发生或不发生。

下面是一些常见的事件：

1. 全事件 Ω（必然事件）
2. 空事件 $\emptyset$（不可能事件）
3. 基本事件 $\{a\}$，其中 $a \in \Omega$，即仅包含单一试验结果的事件

1.3 事件的运算

由于事件是集合，因此事件之间可以进行集合之间的运算，如：

1. 余 $A^c = \Omega \setminus A$
2. 和 $A + B = A \cup B = (A^c \cap B^c)^c$
3. 差 $A - B = A \setminus B$
4. 积 $AB = A \cap B = (A^c \cup B^c)^c$

集合的 De Morgan's laws 也适用于事件：$(\bigcup_n A_n)^c = \bigcap_n A_n^c$。

事件的运算像集合的运算一样，可以用 Venn 图来表示。

1.4 概率的几种解释

对于概率这一数学概念，人们形成了几种从不同角度出发的解释：

1. 古典解释：基于等可能性的解释
2. 频率解释：基于大量重复试验的解释（频率学派采用的解释）
3. 主观解释：概率是一种对确信程度的度量（Bayes 学派采用的解释）

1.5 概率的公理化定义

我们用 2^Ω 表示 Ω 的幂集，即 Ω 的所有子集组成的集合。

定义 1.4. 事件集类 $\mathscr{F} \subset 2^\Omega$ 必须满足所谓 σ-代数的性质：

1. $\Omega \in \mathscr{F}$
2. $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$（对补运算的封闭性）
3. $A_i \in \mathscr{F}, \forall i \in \mathbb{N}^* \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathscr{F}$（对可列并的封闭性）

例 1.1. $\Omega = \{a, b, c, d\}$，以下是一些合法的事件集类：

1. $\mathscr{F}_1 = 2^\Omega$
2. $\mathscr{F}_2 = \{\Omega, \emptyset\}$
3. $\mathscr{F}_3 = \{\Omega, \emptyset, \{a, b\}, \{c, d\}\}$

定义 1.5. (Kolmogorov) 概率函数 $P : \mathscr{F} \to \mathbb{R}$ 是满足以下三条公理的映射：

1. $P(A) \geq 0, \forall A \in \mathscr{F}$
2. $P(\Omega) = 1$
3. $A_i \in \mathscr{F}, \forall i \in \mathbb{N}^*, A_i A_j = \emptyset, \forall i \neq j \Rightarrow P(\sum_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$（加法公理/可列可加性）

我们称 $(\Omega, \mathscr{F}, P)$ 是一个概率空间。

命题 1.1. 关于概率空间，有如下性质：

1. $P(A) \leq 1,\ \forall A \in \mathscr{F}$
2. $P(\emptyset) = 0$
3. $P(A) + P(A^c) = 1$
4. $A_i \in \mathscr{F}, \forall i \in \{1, 2, \cdots, n\},\ A_i A_j = \emptyset, \forall i \neq j \Rightarrow P(\sum_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$（有限可加性）
5. $A \subset B \Rightarrow P(A) \leq P(B)$（我们称事件 A 蕴含事件 B）
6. $P(A_1 + \cdots + A_n) = \sum_{i=1}^n P(A_i) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1}A_{i_2}\cdots A_{i_r}) + \cdots + (-1)^{n+1} P(A_1 \cdots A_n)$（容斥公式）

   特别地，$P(A + B) = P(A) + P(B) - P(AB)$。

例 1.2.（配对问题）

有 n 个人，每人有一顶帽子。现将所有帽子放到一起，再随机分配给每人一顶，考虑无人拿到自己的帽子的概率。

为此，设事件 A_i 为 “第 i 个人拿到自己的帽子”，则 $P(A_i) = 1/n$。

利用容斥公式，至少一人拿到自己帽子的概率为 $P(A_1 + \cdots + A_n) = \sum_{i=1}^n P(A_i) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1}A_{i_2}\cdots A_{i_r}) + \cdots + (-1)^{n+1} P(A_1 \cdots A_n)$,

其中 $\sum_{i_1<i_2<\cdots<i_r} P(A_{i_1}A_{i_2}\cdots A_{i_r}) = \frac{(n-r)!}{n!}\binom{n}{r} = \frac{1}{r!}$，即 $P(A_1 + \cdots + A_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{r+1}\frac{1}{r!} + \cdots + (-1)^{n+1}\frac{1}{n!}$。

所求概率 $P_n = 1 - P(A_1 + \cdots + A_n) = 1 - (1 - \frac{1}{2!} + \cdots + (-1)^{n+1}\frac{1}{n!}) \to e^{-1} (n \to \infty)$。

思考：恰有 k 个人拿到自己的帽子的概率?

1.6 条件概率

定义 1.6. 若 $P(B) > 0$，定义条件概率 $P(A|B) = \frac{P(AB)}{P(B)}$。

通常，我们计算条件概率的方法有两种：

1. 在缩小（受限）的样本空间（要求事件 B 发生）上，考虑事件 A 发生的概率
2. 根据定义计算

一种常用的形式是 $P(AB) = P(A|B)P(B) = P(B|A)P(A)$，这可以视作是求解两个事件的积的概率的方法（乘法法则）。

例 1.3. 掷一个均匀六面骰，$\Omega = \{1, 2, 3, 4, 5, 6\}, A = \{2, 3, 4, 5\}, B = \{1, 3, 5\}$，则 $P(A) = 4/6, P(B) = 3/6, P(AB) = 2/6, P(A|B) = \frac{P(AB)}{P(B)} = 2/3$。

例 1.4. 袋子中有 8 个红球和 4 个白球，无放回地取出两个球，利用组合数可知，两个都是红球的概率为 $\frac{\binom{8}{2}}{\binom{12}{2}}$。
用条件概率可以简化计算：$P(R_1R_2) = P(R_1)P(R_2|R_1) = \frac{8}{12} \times \frac{7}{11}$。

更一般地，我们有 $P(A_1A_2\cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$，常用于序贯发生的一系列事件的积的概率求解。

例 1.5. 回忆上一节的“配对问题”。我们有 $P(A_{i_1}A_{i_2}\cdots A_{i_r}) = P(A_{i_1})P(A_{i_2}|A_{i_1})\cdots P(A_{i_r}|A_{i_1}\cdots A_{i_{r-1}}) = \frac{1}{n} \times \frac{1}{n-1} \times \cdots \times \frac{1}{n-(r-1)} = \frac{(n-r)!}{n!}$。

命题 1.2. 对于给定的事件 B，$P(\cdot|B) : \mathscr{F} \to \mathbb{R}$ 是概率函数，即 $(\Omega, \mathscr{F}, P(\cdot|B))$ 仍是概率空间。

对于上述命题的证明，只需验证 $P(\cdot|B)$ 满足概率的三条公理即可。

这提示我们，条件概率也是一种概率，如果我们将 $P(A)$ 称为观察到事件 B 之前 A 的“先验概率”，则 $P(A|B)$ 就是相应的 “后验概率”。

一个常见的迷思是：观测到事件 A 已经发生后，是否可以说事件 A 发生的概率 $P(A) = 1$? 学过条件概率之后，我们知道答案是否定的，实际上是后验概率 $P(A|A) = 1$。

1.7 事件的独立性

定义 1.7. 若 $P(AB) = P(A)P(B)$，则称事件 A, B 相互独立。

如果 $P(B) > 0$，我们注意到 A, B 独立等价于 $P(A|B) = P(A)$。

命题 1.3. 若 A, B 独立，则 A^c, B 独立。

定义 1.8. 若 $P(ABC) = P(A)P(B)P(C)$，且 A, B, C 两两独立，则称事件 A, B, C 独立。

注意，仅有 A, B, C 两两独立，不能推出三者独立。

定义 1.9. 若对于事件列 $\{A_i\}_{i=1}^{\infty}$，任意取有限个事件 $A_{i_1}, A_{i_2}, \cdots, A_{i_r}$，都有 $P(A_{i_1}A_{i_2}\cdots A_{i_r}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_r})$，则称 $\{A_i\}_{i=1}^{\infty}$ 相互独立。

例 1.6. 每周开奖的彩票，各次中奖率均为 10^{-5} 且独立，问连续十年（520 周）不中奖的概率? 令事件 A_i 为第 i 周不中奖，则 $P(A_i) = 1 - 10^{-5}$，故 $P(A_1\cdots A_{520}) = (1-10^{-5})^{520} \approx 0.9948$。

定义 1.10. 若事件 A, B, E 满足 $P(AB|E) = P(A|E)P(B|E)$，则我们称 A, B 关于 E 条件独立。

注意，条件独立性和独立性之间没有蕴含关系。

1.8 Bayes 公式

定理 1.1. （全概率公式）
设 $\{B_i\}$ 是 Ω 的一个分割，即

1. $\sum_i B_i = \Omega$
2. $B_iB_j = \emptyset, \forall i \neq j$
3. $P(B_i) > 0, \forall i$

则 $P(A) = P(\sum_i (AB_i)) = \sum_i P(AB_i) = \sum_i P(A|B_i)P(B_i)$。
注：$\{B_i\}$ 可以是有限集合，或可数无穷集合。

例 1.7. 对于调查问卷中的敏感问题（如“你是否有过某病史”），被调查者可能会有所顾虑而做出虚假的回答。为保护被调查者的隐私，同时取得其信任，考虑引入一个“保护性问题”，即不具有敏感性的问题（如“你是否会游泳”），并让被调查者以抛硬币的方式，随机抽取一个问题回答。这样，抽到敏感问题的、确有过该病史的被调查者在回答“是”时也无须有病史暴露之虞。
设人群中，敏感问题答案为“是”的比例为 p（未知），保护性问题答案为“是”的比例为 q（假设已知），则若收集到 n 个被调查者的结果，其中 k 个为“是”，我们便有 $\frac{1}{2}p + \frac{1}{2}q \approx \frac{k}{n}$，可以据此得到 p 的估计。

定理 1.2. （Bayes 公式 / Bayes 准则）
设 $\{B_i\}$ 是 Ω 的一个分割，则 $P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$。

例 1.8. （假阳性悖论）
对于一种流行病，A 表示一个人检查呈阳性，B 表示此人确实患病。设 $P(B) = 10^{-4}, P(A|B) = 0.99, P(A|B^c) = 10^{-3}$，则一个检查呈阳性的人真的患病的概率仅为 $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)} \approx 9\%$。
如果再次检测仍呈阳性，且两次检测效率不变，结果彼此独立，则此人真的患病的概率为 $P(B|A_1A_2) = \frac{P(A_1A_2|B)P(B)}{P(A_1A_2|B)P(B)+P(A_1A_2|B^c)P(B^c)} = \frac{P(A_1|B)P(A_2|B)P(B)}{P(A_1|B)P(A_2|B)P(B)+P(A_1|B^c)P(A_2|B^c)P(B^c)} \approx 99\%$。

第二章　随机变量

2.1　一维随机变量

定义 2.1. 随机变量是样本空间上的实值函数。

注意，上述定义是不严格的。

更严谨的定义：若对于可测空间 $(\Omega, \mathscr{F})$ 和函数 $X : \Omega \to \mathbb{R}$，有 $\forall x \in \mathbb{R}, \{\omega | X(\omega) \leq x\} \in \mathscr{F}$，则称 X 是 $(\Omega, \mathscr{F})$ 上的随机变量。其中“可测空间”是指 $\mathscr{F}$ 是样本空间 Ω 上的 σ-代数。此处不要求“概率空间”，即随机变量的定义并不依赖概率测度 P 的存在。

例 2.1. 下表展示了两个随机变量。其中“像集”即 $\{X(\omega) | \omega \in \Omega\}$。

试验	样本空间 Ω	随机变量 X	像集
随机调查 50 人对某议题支持与否	$\Omega_1 = \{0, 1\}^{50}$	$X_1 =$ “1” 的个数	$\{0,1,\cdots,50\}$
随机抽取一名北京成年市民	$\Omega_2 =$ 所有北京成年市民之集	$X_2 =$ 其年收入	$\mathbb{R}$

注意，我们经常用“$X_1 = 20$”、“$X_2 > 100000$”等简化的记号来表示事件。例如，前者实际上指的是 $\{\omega \in \Omega_1 | X_1(\omega) = 20\}$。

诸如此类的试验结果集合需是事件，这体现出前述的随机变量严谨定义的意义。事实上，如果满足该严谨定义，则对于任意可测集 $I \subset \mathbb{R}$，都有 $\{\omega \in \Omega | X(\omega) \in I\} \in \mathscr{F}$。

随机变量是试验结果的数值摘要，起到一种概括的作用。随机变量的“随机”要素来自于样本点 $\omega \in \Omega$ 的随机选择。在实际应用中，随机变量常常比样本空间具有更直观的意义。

随机变量可以分为：

1. 离散型：至多可数多个取值
2. 连续型：区间型取值（非严格定义）
3. 其他

“其他”中的一个非常特殊的子类是所谓的混合型随机变量。

定义 2.2. 对于随机变量 X 和 $\mathbb{R}$ 的可测子集 I（例如 $I=(a,b]$），令 $X^{-1}(I)=\{\omega\in\Omega|X(\omega)\in I\}\subset\Omega$ 为 I 的原像集，我们定义记号 $P(X\in I)$ 表示“X 的取值在 I 中的概率”，其值为 $P(X^{-1}(I))$。

例如，$P(a<X\leq b)=P(\{\omega|X(\omega)\in(a,b]\})$。

定义 2.3. $F_X(x)=P(X\leq x),\forall x\in\mathbb{R}$ 称为随机变量 X 的累积分布函数（Cumulative Distribution Function, CDF）。下标 X 在无歧义时可省略。

我们有 $P(a<X\leq b)=F(b)-F(a)$。

例 2.2. 令 X 表示掷两个均匀六面骰所得的点数和，则 X 的分布表（详见 2.2 节）为

X	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

相应的 CDF 见图 2.1。

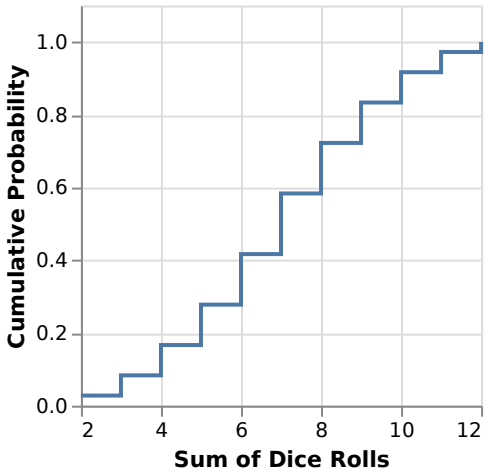

图 2.1: X 的 CDF 图象

注：由于软件限制，各个阶跃点的绘制方式不太规范，实际上从其左侧逼近应该为一个空圈，例如 $F(3)=3/36$ 而不是 $1/36$。另外，$\forall x<2,F(x)=0;\forall x\geq 12,F(x)=1$。

命题 2.1. CDF 的性质：

1. F 单调递增（未必严格单调递增）
2. $\lim_{x\to+\infty}F(x)=1,\lim_{x\to-\infty}F(x)=0$
3. F 右连续

可以证明，上述三条性质是任意函数 $F:\mathbb{R}\to\mathbb{R}$ 成为 CDF 的充要条件。

思考：如果我们将 CDF 的定义改为 $P(X<x)$，上述性质会如何变化？

命题 2.2. 若 X, Y 为随机变量，则 $aX + bY, XY, X/Y$ (需$Y \neq 0$) 都是随机变量。一般地，若 g 为可测函数，则 $g(X, Y)$ 是随机变量。

定义 2.4. 设 X_1, X_2 的 CDF 分别为 F_1, F_2，我们称 X_1 与 X_2 同分布，若 $\forall x \in \mathbb{R}, F_1(x) = F_2(x)$。

命题 2.3. 随机变量 X_1 与 X_2 同分布的一个充要条件是 $\forall$ 可测集$I \subset \mathbb{R}, P(X_1 \in I) = P(X_2 \in I)$。

注意，同分布不等价于“同变量”，即两个同分布的变量的取值不一定恒等。

例 2.3. 掷一次硬币，X 表示正面向上次数，Y 表示反面向上次数，显然 X 与 Y 同分布，但取值不等。

2.2 离散随机变量

定义 2.5. 离散随机变量 X 的*概率质量函数*（Probability Mass Function, PMF）f 是指该随机变量取各个可能值的概率，即 $f(x) = P(X = x), \forall x \in \mathbb{R}$。可以用*分布表*的形式展示各个可能取值与概率的对应关系。

命题 2.4. 如果离散随机变量 X 的所有可能取值为 $\{x_i\}$，则 X 的 PMF 具有如下性质：

1. $f(x_i) = p_i \geq 0, \forall i$
2. $\sum_i p_i = 1$
3. $F(x) = \sum_{x_i \leq x} f(x_i)$

定义 2.6. 离散随机变量 X 的*期望*定义为 $\mathrm{E}(X) = \sum_i x_i p_i$。
我们称 X 的期望*存在*，当且仅当 $\sum_i |x_i| p_i < +\infty$。
当期望存在时，其*方差*定义为 $\mathrm{Var}(X) = \sum_i (x_i - \mathrm{E}(X))^2 p_i = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$。
当方差有限时，称其算术平方根为 X 的*标准差*。

注意，通常我们所说的一个随机变量的*均值*指的就是期望。

*标准化*指的是对 X 作线性变换 $\frac{X-\mu}{\sigma}$，其中 μ 和 σ 分别为 X 的期望和标准差，得到均值为 0，标准差为 1 的随机变量。

对于可测函数 g，$g(X)$ 也是随机变量，其期望 $\mathrm{E}(g(X)) = \sum_i g(x_i) p_i$。

期望反映了随机变量的集中趋势，而方差反映了其分散程度。

2.3 常见离散分布

定义 2.7. 称一个随机变量 X 服从 *Bernoulli* 分布，若 $\exists p \in (0,1)$，X 的取值集合为 $\{0,1\}$，且 $P(X=1)=p, P(X=0)=1-p$。记作 $X \sim B(p)$。

$B(p)$ 中的 p 称为该 Bernoulli 分布的*参数*。后续介绍的其他分布同理。

我们常将两种取值分别称为“成功”和“失败”。

计算可得，若 $X \sim B(p)$，则 $\mathrm{E}(X)=p, \mathrm{Var}(X)=p(1-p)$。

定义 2.8. 称一个随机变量 X 服从**二项分布**，若 $\exists N \in \mathbb{N}^*$，$p \in (0,1)$，$X$ 的取值集合为 $\{0,1,\cdots,N\}$，且 $P(X=k)=\binom{N}{k}p^k(1-p)^{N-k}(k \in \{0,1,\cdots,N\})$。记作 $X \sim B(N,p)$。

我们常将 k 理解为“N 次独立 Bernoulli 试验中的成功次数”。

计算可得，若 $X \sim B(N,p)$，则 $\mathrm{E}(X)=Np, \mathrm{Var}(X)=Np(1-p)$。

定义 2.9. 称一个随机变量 X 服从 *Poisson* 分布，若 $\exists \lambda > 0$，X 的取值集合为 $\mathbb{N}$，且 $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}(k \in \mathbb{N})$。记作 $X \sim P(\lambda)$。

计算可得，若 $X \sim P(\lambda)$，则 $\mathrm{E}(X)=\lambda, \mathrm{Var}(X)=\lambda$。

对 Poisson 分布的一种常见理解是“一段时间内某个小概率事件发生的次数”所服从的分布。例如，观察时间 $(0,1]$ 内某路口的交通事故数 X，将 $(0,1]$ 区间等分成 n 个小区间，即 $l_i=(\frac{i-1}{n},\frac{i}{n}](i=1,2,\cdots,n)$。考虑到 n 很大时，每个区间的长度很小，我们作如下假设：

1. 每段区间内，至多发生一次事故
2. l_i 上发生一次事故的概率与区间长度（$1/n$）成正比，为 $p=\lambda/n$
3. 各区间内是否发生事故彼此独立

则 $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \to \frac{\lambda^k e^{-\lambda}}{k!}(n \to +\infty)$，即 $X \sim P(\lambda)$。

例 2.4. 设某医院平均每天出生婴儿数为 λ，则接下来 t 天内出生婴儿数服从参数为 $t\lambda$ 的 Poisson 分布。

对于一般的二项分布 $X \sim B(N,p)$，若 p 很小，N 很大，而 $\lambda=Np$ 不太大，则近似有 $X \sim P(\lambda)$，且近似误差不超过 $\min\{p, Np^2\}$。

进一步，若 N 次 Bernoulli 试验并非严格独立，但满足弱相依条件，则 Poisson 分布仍为一种较好的近似。

例 2.5.（配对问题）

A_i 表示第 i 个人拿到自己的帽子，则 $P(A_i)=1/n, P(A_i|A_j)=\frac{1}{n-1}(j \neq i)$，当 n 很大时，$1/n$

和 $\frac{1}{n-1}$ 很接近，可以认为满足弱相依条件。

记 X 为拿到自己帽子的人数，则 X 近似服从参数为 $\lambda = np = n \cdot \frac{1}{n} = 1$ 的 Poisson 分布，即 $P(X = k) \approx \frac{e^{-1}}{k!}$。

我们用常规做法检查这种近似是否合理。首先考虑指定的某 k 人，记事件 E 表示这 k 人拿到自己的帽子，事件 F 表示其余 $(n-k)$ 人未拿到自己的帽子，则 $P(EF) = P(E)P(F|E) = \frac{(n-k)!}{n!} \cdot P_{n-k}$，其中 P_{n-k} 为 $(n-k)$ 人随机拿帽子时无人拿对的概率。那么我们有 $P(X = k) = \binom{n}{k} P(EF) = \frac{1}{k!} P_{n-k} \to \frac{e^{-1}}{k!} (n \to +\infty)$。这说明前述的近似是较好的。

2.4 连续随机变量

定义 2.10. 对随机变量 X，若存在 $f : \mathbb{R} \to [0, +\infty)$，使得 $\forall$ 可测集 $I \subset \mathbb{R}$，都有 $P(X \in I) = \int_I f(x)\mathrm{d}x$，则称 X 为 *连续型随机变量*，f 称为其*概率密度函数* (Probability Density Function, PDF)。

命题 2.5. 连续随机变量 X 的 PDF 具有如下性质：

1. $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x \equiv 1$
2. $P(a < X \leq b) = \int_a^b f(x)\mathrm{d}x = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$
3. $P(X = a) \equiv 0, \forall a \in \mathbb{R}$
4. 若 f 在 x_0 处连续，则 $P(x_0 - \delta < X < x_0 + \delta) = \int_{x_0-\delta}^{x_0+\delta} f(t)\mathrm{d}t \approx f(x_0) \cdot 2\delta$
5. $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ 连续，且若 f 在 x 处连续，有 $F'(x) = f(x)$
6. PDF 若存在，则不唯一（可以修改其在任意零测集上的值，得到不同的 PDF）

定义 2.11. 连续随机变量 X 的*期望*定义为 $\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x$。

我们称 X 的期望存在，当且仅当 $\int_{-\infty}^{+\infty} |x| f(x)\mathrm{d}x < +\infty$。

当期望存在时，其*方差*定义为 $\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathrm{E}(x))^2 f(x)\mathrm{d}x = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$。

当方差有限时，称其算术平方根为 X 的*标准差*。

对于可测函数 g，$g(X)$ 也是随机变量，其期望 $\mathrm{E}(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x$。

2.5 常见连续分布

定义 2.12. 称一个连续型随机变量 X *服从均匀分布*，若其 PDF 为 $f(x) = \frac{1}{b-a} (x \in (a, b))$，$f$ 在其余各处取 0。记作 $X \sim U(a, b)$。

我们常将 $X \sim U(0,1)$ 称为随机数。

计算可得，若 $X \sim U(a,b)$，则 $\mathrm{E}(X) = \frac{a+b}{2}, \mathrm{Var}(X) = \frac{(b-a)^2}{12}$。

定义 2.13. 称一个连续型随机变量 X 服从正态分布，若其 PDF 为 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (\sigma > 0)$。记作 $X \sim N(\mu, \sigma^2)$。

计算可得，若 $X \sim N(\mu, \sigma^2)$，则 $\mathrm{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$。

著名的“经验法则”见图 2.2。

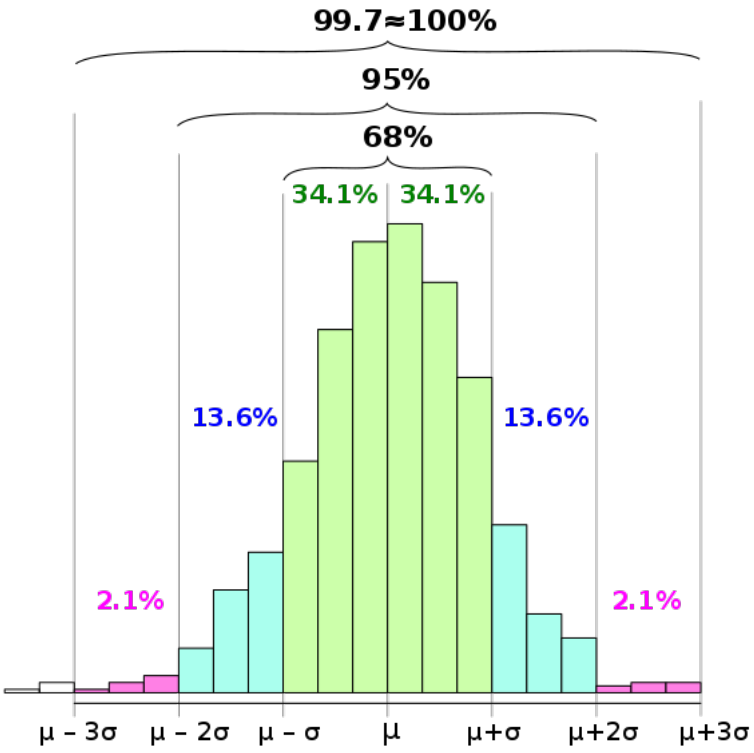

图 2.2: 经验法则

$X \sim N(\mu, \sigma^2)$ 的充要条件是 $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$。我们将 $N(0,1)$ 称为标准正态分布。

定义 2.14. 称一个连续型随机变量 X 服从指数分布，若其 PDF 为 $f(x) = \lambda e^{-\lambda x} (\lambda > 0, x > 0)$，$f$ 在其余各处取 0。记作 $X \sim Exp(\lambda)$。

指数分布常用于刻画等待时间、寿命等。

计算可得，若 $X \sim Exp(\lambda)$，则 $\mathrm{E}(X) = 1/\lambda, \mathrm{Var}(X) = 1/\lambda^2$。

指数分布有另一种符号约定，以 $\beta = 1/\lambda$ 为参数，一些数学软件可能采用此种约定。

指数分布的 CDF 为 $F(x) = 1 - e^{-\lambda x} (x > 0)$，所谓的“尾概率”为 $P(X > x) = 1 - F(x) = e^{-\lambda x} (x > 0)$。

例 2.6. 设某医院平均每天出生婴儿数为 λ，现在观察到一名婴儿出生，则接下来 t 天内有婴儿出生的概率为 $P(X \le t)$，其中 X 表示到下一个婴儿出生所需等待的时间。
记 $N(t)$ 为 t 天内出生婴儿数,我们已经知道 $N(t) \sim P(t\lambda)$,则 $P(X > t) = P(N(t) = 0) = e^{-\lambda t}$，故 $P(X \le t) = 1 - e^{-\lambda t}$。我们发现 X 服从参数为 λ 的指数分布。

我们从另一个角度理解指数分布。

首先引入*失效率*或*危险率*的概念。设 X 为连续型随机变量（表示某种零件的寿命），其 CDF 为 $F(x)$，且 $F(0) = 0$。考虑条件概率 $P(x < X < x + \mathrm{d}x | X > x) = \frac{P(x<X<x+\mathrm{d}x)}{P(X>x)} = \frac{F(x+\mathrm{d}x)-F(x)}{1-F(x)} \approx \frac{F'(x)}{1-F(x)}\mathrm{d}x$，即“年龄”为 x 的零件不能继续工作的条件概率密度为 $\frac{F'(x)}{1-F(x)}$，我们称其为瞬时失效率 $\lambda(x)$，则 $F(x) = 1 - e^{-\int_0^x \lambda(t)\mathrm{d}t}$。

在“无老化”假设下，即 $\lambda(t) \equiv \lambda$ 不随时间变化，则 $F(x) = 1 - e^{-\lambda t}(x > 0)$，$X$ 服从指数分布。

指数分布有所谓“无记忆性”: $P(X > t+s | X > s) = \frac{P(X>t+s)}{P(X>s)} = e^{-\lambda t} = P(X > t)(t, s > 0)$。

“无老化”假设并不总是成立。为此，我们可以进行一定程度的改进，例如令 $\lambda(x) = \alpha\frac{x^{\alpha-1}}{\beta^\alpha}(x > 0, \alpha, \beta > 0$ 为常数)，则 $F(x) = 1 - e^{-(\frac{x}{\beta})^\alpha}(x > 0)$，称之为 *Weibull* 分布。当 $\alpha = 1$ 时，Weibull 分布退化为参数为 $1/\beta$ 的指数分布。

总览至此我们介绍过的各个分布的参数，可以将其大致分为以下几类:

1. 位置参数：决定了分布平移到的位置，通常在 PMF/PDF 中体现为 $f(x) = g(x - \cdot)$ 的形式，如正态分布的参数 μ
2. 尺度参数：决定了分布伸缩的程度，通常在 PMF/PDF 中体现为 $f(x) = g(\frac{x}{\cdot})$ 的形式，如正态分布的参数 σ、Weibull 分布的参数 β
3. 形状参数：决定了分布的形状，如 Weibull 分布的参数 α

2.6 随机变量的函数

对于随机变量 X 和可测函数 g，$Y = g(X)$ 也是随机变量。特别地，若 X 为离散型随机变量，则 Y 也离散。但若 X 为连续型随机变量，Y 未必连续。

例 2.7. $X \sim Exp(\lambda)$，$Y = \begin{cases} 0, & X \le t_0, \\ 1, & X > t_0, \end{cases}$ 其中 $t_0 > 0$ 为常数，则 $Y \sim B(e^{-\lambda t_0})$。

例 2.8. 设 X 为连续型随机变量，PDF 为 $f(x)$，考虑 $Y = X^2$。
从 CDF 入手，$\forall y > 0, P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f(x)\mathrm{d}x$，我们有 Y 的 PDF 为 $l(y) = \frac{\mathrm{d}}{\mathrm{d}y}P(Y \le y) = \frac{1}{2\sqrt{y}}(f(\sqrt{y}) + f(-\sqrt{y}))(y > 0)$。
特别地，若 $X \sim N(0, 1)$，称 Y 服从自由度为 1 的 χ^2-分布，读作“卡方分布”。

若 $Y = g(X)$ 为随机变量，我们可以计算 Y 的分布如下：

- $P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y))$
- $P(Y \leq y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y]))$

第三章　联合分布

3.1　随机向量

定义 3.1. 称 $(X_1, X_2, \cdots, X_n) : \Omega \to \mathbb{R}^n$ 为（n 维）随机向量，若 $\{X_i\}_{i=1}^n$ 均为随机变量。

定义 3.2. n 维随机向量的（联合）（累积）分布函数（CDF）定义为 $F(x_1, \cdots, x_n) = P(X_1 \leq x_1, \cdots, X_n \leq x_n), \forall (x_1, \cdots, x_n) \in \mathbb{R}^n$。

对于 $n = 2$（二元分布）的情形，我们常用 (X, Y) 来表示随机向量，对应的 CDF 为 $F(x, y)$。

3.2　离散分布

定义 3.3. 称 n 维随机向量 $(X_1, \cdots, X_n)$ 是离散的，当且仅当 $\{X_i\}_{i=1}^n$ 均为离散随机变量。离散随机向量 $(X_1, \cdots, X_n)$ 的（联合）概率质量函数（PMF）定义为 $f(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n), \forall (x_1, \cdots, x_n) \in \mathbb{R}^n$。

命题 3.1. 离散随机向量 $(X_1, \cdots, X_n)$ 的 PMF 具有如下性质：

1. $f(x_1, \cdots, x_n) \geq 0, \forall (x_1, \cdots, x_n) \in \mathbb{R}^n$
2. $\sum_{x_i \in \{X_i(\omega) | \omega \in \Omega\}, \forall i \in \{1, \cdots, n\}} f(x_1, \cdots, x_n) \equiv 1$

注意第 2 条性质中求和的项数为至多可数，原因是有限个至多可数集的笛卡尔积仍是至多可数集。

例 3.1. 设 $\{B_i\}_{i=1}^n$ 为 Ω 的一个分割（分割的定义见 1.8 节），$P(\omega \in B_i) = p_i \geq 0, \forall i \in \{1, \cdots, n\}$，$\sum_{i=1}^n p_i = 1$。

进行 N 次独立试验，设 $\forall i \in \{1, \cdots, n\}$，有 X_i 个试验结果落在 B_i 中，则若 $k_1 + \cdots + k_n = N$，其中 k_i 均为非负整数，我们有 $P(X_1 = k_1, \cdots, X_n = k_n) = \binom{N}{k_1, \cdots, k_n} p_1^{k_1} \cdots p_n^{k_n}$。其中 $\binom{N}{k_1, \cdots, k_n} = \frac{N!}{k_1! \cdots k_n!}$ 为多项式 $(a_1 + \cdots + a_n)^N$ 中 $a_1^{k_1} \cdots a_n^{k_n}$ 项的系数。

我们称 $(X_1, \cdots, X_n)$ 服从多项分布。

3.3 连续分布

定义 3.4. 对 n 维随机向量 $(X_1,\cdots,X_n)$，若存在 $f:\mathbb{R}^n\to[0,+\infty)$，使得 $\forall$ 可测集 $Q\subset\mathbb{R}^n$，都有 $P((X_1,\cdots,X_n)\in Q)=\int_Q f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n$，则称 $(X_1,\cdots,X_n)$ 为连续型随机向量，f 称为其（联合）概率密度函数（PDF）。

命题 3.2. 连续随机向量 $(X_1,\cdots,X_n)$ 的 PDF 具有如下性质：

1. $\int_{\mathbb{R}^n} f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n\equiv 1$
2. 以 $n=2$ 为例，$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(t,s)\mathrm{d}s\mathrm{d}t$, $f(a,b)=\frac{\partial^2 F}{\partial x\partial y}(a,b)$, a.e.

其中 a.e. 表示“几乎处处”。

例 3.2. 矩形域上的均匀分布的 PDF：$f(x,y)=\begin{cases}\dfrac{1}{(b-a)(d-c)}, & (x,y)\in(a,b)\times(c,d),\\ 0, & \text{其他.}\end{cases}$

例 3.3. 二元正态分布 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ 的 PDF：
$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_1}{\sigma_1})^2+(\frac{y-\mu_2}{\sigma_2})^2-2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2})},\forall(x,y)\in\mathbb{R}^2,\sigma_1,\sigma_2>0,|\rho|<1$。
令 $\boldsymbol{x}=\begin{bmatrix}\frac{x-\mu_1}{\sigma_1}\\ \frac{y-\mu_2}{\sigma_2}\end{bmatrix}$, $W=\frac{1}{1-\rho^2}\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix}$, $W=A^{\mathrm{T}}A$ 为正定矩阵 W 的 Cholesky 分解，则
$-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_1}{\sigma_1})^2+(\frac{y-\mu_2}{\sigma_2})^2-2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2})=-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}W\boldsymbol{x}=-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x}=-\frac{1}{2}(A\boldsymbol{x})^{\mathrm{T}}(A\boldsymbol{x})$。
上述 Cholesky 分解的结果为 $A=\frac{1}{\sqrt{1-\rho^2}}\begin{bmatrix}1 & -\rho\\ 0 & \pm\sqrt{1-\rho^2}\end{bmatrix}$ 或 $A=\frac{1}{\sqrt{1-\rho^2}}\begin{bmatrix}-1 & \rho\\ 0 & \pm\sqrt{1-\rho^2}\end{bmatrix}$。

3.4 边际分布

对 n 维随机向量 $(X_1,\cdots,X_n)$，称 $F_i(x)=P(X_i\le x)=P(X_i\le x,-\infty<X_j<+\infty,\forall j\neq i)$ 为 X_i 的边际分布。

例如，若 $n=2$，随机向量 (X,Y) 有 CDF $F(x,y)$，则 X 的边际分布为 $F_X(x)=P(X\le x)=P(X\le x,Y\in\mathbb{R})=\lim\limits_{y\to+\infty}P(X\le x,-\infty<Y\le y)=\lim\limits_{y\to+\infty}F(x,y)$。

若 $n=3$，随机向量 (X,Y,Z) 有 CDF $F(x,y,z)$，则 X 的边际分布为 $F_X(x)=\lim\limits_{y,z\to+\infty}F(x,y,z)$，$(X,Y)$ 的边际分布为 $F_{X,Y}(x,y)=P(X\le x,Y\le y)=P(X\le x,Y\le y,-\infty<Z<+\infty)=\lim\limits_{z\to+\infty}F(x,y,z)$。

例 3.4. 设二维随机向量 (X,Y) 的 CDF 为 $F(x,y)$，则 $\forall a,b\in\mathbb{R},P(X>a,Y>b)=1-F_X(a)-F_Y(b)+F(a,b)$。

对于离散型随机向量, 以 $n=2$ 为例, 定义边际 *PMF* 为 $P(X=x)=\sum\limits_{y}P(X=x,Y=y)$。

对于连续型随机向量, 以 $n=2$ 为例, 设联合 PDF 为 $f(x,y)$, 则 $F_X(x)=P(X\le x,Y\in\mathbb{R})=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}f(t,s)\mathrm{d}s\mathrm{d}t$, 则 X 的边际 *PDF* 为 $f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y$。

例 3.5. 随机向量 (X,Y) 服从二元正态分布 $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$, 则 $f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$, 即 $X\sim N(\mu_1,\sigma_1^2)$。同理 $Y\sim N(\mu_2,\sigma_2^2)$。

3.5 条件分布

以 $n=2$ 为例说明条件分布的概念, 考虑随机向量 (X,Y)。

对于离散型随机向量, 设联合 PMF 为 $P(X=a_i,Y=b_j)=p_{ij}\ge0,\sum_{i,j}p_{ij}\equiv1$, 则在 $Y=b_j$ 条件下的 X 的条件 *PMF* 为 $P(X=a_i|Y=b_j)=\frac{P(X=a_i,Y=b_j)}{P(Y=b_j)}=\frac{p_{ij}}{\sum_k p_{kj}}$。条件 PMF 满足 $\sum_i P(X=a_i|Y=b_j)\equiv1,\forall j$。

对于连续型随机向量, 设联合 PDF 为 $f(x,y)$, 首先考虑条件概率 $P(X\le x|y\le Y\le y+\mathrm{d}y)=\frac{P(X\le x,y\le Y\le y+\mathrm{d}y)}{P(y\le Y\le y+\mathrm{d}y)}=\frac{\int_{-\infty}^{x}\int_{y}^{y+\mathrm{d}y}f(t,s)\mathrm{d}s\mathrm{d}t}{\int_{y}^{y+\mathrm{d}y}f_Y(s)\mathrm{d}s}$, 对 x 求导得 X 在 $y\le Y\le y+\mathrm{d}y$ 条件下的条件 PDF 为 $\frac{\int_{y}^{y+\mathrm{d}y}f(x,s)\mathrm{d}s}{\int_{y}^{y+\mathrm{d}y}f_Y(s)\mathrm{d}s}\to\frac{f(x,y)}{f_Y(y)}(\mathrm{d}y\to0)$。

定义 3.5. 对于连续型随机向量 (X,Y), 设联合 PDF 为 $f(x,y)$, 若 $f_Y(y)>0$, 则称 X 在 $Y=y$ 条件下的条件 *PDF* 为 $f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$。

可以验证 $f_{X|Y}(x|y)$ 满足 PDF 的各性质。

相应的条件 *CDF* 为 $F_{X|Y}(a|y)=P(X\le a|Y=y)=\int_{-\infty}^{a}f_{X|Y}(x|y)\mathrm{d}x$。

我们熟知的各个定理均有适用于连续型随机向量的版本:

1. $f(x,y)=f_{X|Y}(x|y)f_Y(y)=f_{Y|X}(y|x)f_X(x)$（乘法法则）
2. $f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y=\int_{-\infty}^{+\infty}f_{X|Y}(x|y)f_Y(y)\mathrm{d}y$（全概率公式）
3. $f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}=\frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty}f_{X|Y}(x|y)f_Y(y)\mathrm{d}y}$（Bayes 公式）

例 3.6. 随机向量 (X,Y) 服从二元正态分布 $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$, 则 $f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}=\frac{1}{\sqrt{2\pi}\sigma_2}\frac{1}{\sqrt{1-\rho^2}}e^{-\frac{(y-(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)))^2}{2(1-\rho^2)\sigma_2^2}}$, 即 $Y|X=x\sim N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),(1-\rho^2)\sigma_2^2)$。

3.6 独立性

定义 3.6. 设二维随机向量 (X,Y) 的 CDF 为 $F(x,y)$, 若 $F(x,y)=F_X(x)F_Y(y),\forall x,y\in\mathbb{R}$, 则称 X,Y 相互独立。

可以证明，对于二维离散型（或连续型）随机向量 (X,Y)，X,Y 相互独立的充要条件是 $f(x,y)=f_X(x)f_Y(y),\forall x,y\in\mathbb{R}$，其中 $f(x,y)$ 为联合 PMF（或 PDF）。

定义 3.7. 设 n 维随机向量 $(X_1,\cdots,X_n)$ 的 CDF 为 $F(x_1,\cdots,x_n)$，若 $F(x_1,\cdots,x_n)=F_1(x_1)\cdots F_n(x_n),\forall x_1,\cdots,x_n\in\mathbb{R}$，则称 $X_1,\cdots,X_n$ 相互独立。

可以证明，对于 n 维离散型（或连续型）随机向量 $(X_1,\cdots,X_n)$，$X_1,\cdots,X_n$ 相互独立的充要条件是 $f(x_1,\cdots,x_n)=f_1(x_1)\cdots f_n(x_n),\forall x_1,\cdots,x_n\in\mathbb{R}$，其中 $f(x_1,\cdots,x_n)$ 为联合 PMF（或 PDF）。

定理 3.1.

1. 若 $X_1,\cdots,X_n$ 相互独立，则 $\forall m\in\{1,\cdots,n-1\}$，可测函数 g_1,g_2，有 $Y_1=g_1(X_1,\cdots,X_m)$ 与 $Y_2=g_2(X_{m+1},\cdots,X_n)$ 相互独立。
2. 若 n 维连续型随机向量 $(X_1,\cdots,X_n)$ 的联合 PDF $f(x_1,\cdots,x_n)=g_1(x_1)\cdots g_n(x_n),\forall x_1,\cdots,x_n\in\mathbb{R},g_i:\mathbb{R}\to[0,+\infty),\forall i\in\{1,\cdots,n\}$，则 $X_1,\cdots,X_n$ 相互独立且 $\forall i\in\{1,\cdots,n\}$，$X_i$ 的边际 PDF f_i 与 g_i 相差常数因子。

例 3.7. 设 (X,Y) 服从如图 3.1 的三角形域上的均匀分布，即 $f(x,y)=\begin{cases}c, & (x,y)\in\text{三角形域},\\ 0, & \text{其他}\end{cases}$，则 X,Y 不独立。

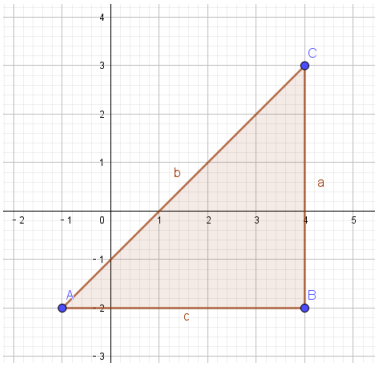

图 3.1: 三角形域上的均匀分布