

# 标题 title

作者 *author*

2023 年 8 月 27 日

# 前言

# 目录

前言	i
第一部分 科学的逻辑	1
第一章 合情推理	2
§1.1 回顾：命题逻辑的演绎推理	2
§1.2 合情推理的数学模型	4
1.2.1 似然，合情推理的原则	4
1.2.2 似然与概率	6
§1.3 合情推理的归纳强论证	8
1.3.1 先验与基率谬误	8
1.3.2 归纳强论证	9
1.3.3 有效论证和归纳强论证的比较	12
第二章 Markov 链与决策	15
§2.1 Markov 链	15
§2.2 Markov 奖励过程 (MRP)	19
§2.3 Markov 决策过程 (MDP)	22
§2.4 隐 Markov 模型 (HMM)	26
2.4.1 评估问题	27
2.4.2 解释问题	28
第二部分 信息与数据	30
第三章 信息论基础	31

§3.1 熵	31
3.1.1 概念的导出	31
3.1.2 概念与性质	34
3.1.3 熵与通信理论	39
§3.2 Kullback-Leibler 散度	42
3.2.1 定义	42
3.2.2 两个关于信息的不等式	44
3.2.3 在机器学习中的应用：语言生成模型	45
§3.3 附录：Shannon 定理的证明	46
§3.4 习题	47
§3.5 章末注记	49
<b>第四章 Johnson-Lindenstrauss 引理</b>	<b>51</b>
§4.1 机器学习中的数据	51
§4.2 矩法与集中不等式	52
§4.3 J-L 引理的陈述与证明	56
§4.4 J-L 引理的应用	60
§4.5 习题	61
§4.6 章末注记	61
<b>第五章 差分隐私</b>	<b>62</b>
§5.1 数据隐私问题	62
§5.2 差分隐私的定义与性质	64
§5.3 差分隐私的应用	68
5.3.1 随机反应算法	68
5.3.2 全局灵敏度与 Laplace 机制	69
5.3.3 DP 版本 Llyod 算法	71
§5.4 差分隐私与信息论	72
§5.5 习题	73
§5.6 章末注记	73
<b>第三部分 决策与优化</b>	<b>74</b>
<b>第六章 凸分析</b>	<b>75</b>

§6.1 决策与优化的基本原理 . . . . .	75
6.1.1 统计决策理论 . . . . .	75
6.1.2 优化问题 . . . . .	76
6.1.3 例子：网格搜索算法 . . . . .	79
§6.2 凸函数 . . . . .	81
§6.3 凸集 . . . . .	84
6.3.1 基本定义和性质 . . . . .	84
6.3.2 分离超平面定理 . . . . .	86
<b>第七章 对偶理论 . . . . .</b>	<b>88</b>
§7.1 条件极值与 Lagrange 乘子法 . . . . .	89
§7.2 Karush–Kuhn–Tucker 条件 . . . . .	92
§7.3 Lagrange 对偶 . . . . .	95
7.3.1 Lagrange 定理 . . . . .	95
7.3.2 弱对偶定理，强对偶定理 . . . . .	99
§7.4 应用：支持向量机 (SVM) . . . . .	103
<b>第八章 不动点理论 . . . . .</b>	<b>106</b>
§8.1 Banach 不动点定理 . . . . .	106
§8.2 Brouwer 不动点定理 . . . . .	109
§8.3 不动点的一般视角 . . . . .	112
<b>第四部分 逻辑与博弈 . . . . .</b>	<b>113</b>
<b>第九章 动态博弈 . . . . .</b>	<b>114</b>
§9.1 输赢博弈 . . . . .	114
§9.2 随机博弈 (Markov 博弈) . . . . .	119
<b>第十章 静态博弈 . . . . .</b>	<b>125</b>
§10.1 正则形式博弈 . . . . .	125
10.1.1 生成对抗网络 . . . . .	126
10.1.2 混合策略 . . . . .	128
§10.2 不完全信息博弈 (Bayes 博弈) . . . . .	129

<b>第五部分 认知逻辑</b>	<b>134</b>
<b>第十一章 模态逻辑基础</b>	<b>135</b>
§11.1 模态逻辑的起源 . . . . .	135
11.1.1 三段论 . . . . .	135
11.1.2 非经典逻辑 . . . . .	136
§11.2 模态语言 . . . . .	137
§11.3 Kripke 语义与框架语义 . . . . .	140
§11.4 模态可定义性 . . . . .	145
<b>第十二章 认知逻辑与共同知识</b>	<b>147</b>
§12.1 “泥泞的孩童”谜题 . . . . .	147
§12.2 认知逻辑的基本模型与性质 . . . . .	149
12.2.1 “泥泞的孩童”再回顾 . . . . .	153
12.2.2 Aumann 结构 . . . . .	154
§12.3 对不一致达成一致 . . . . .	155
§12.4 Rubinstein 电子邮件博弈 . . . . .	158
<b>附录 A 线性代数基础</b>	<b>162</b>
§A.1 线性空间 . . . . .	162
§A.2 线性映射 . . . . .	166
§A.3 矩阵 . . . . .	171
§A.4 双线性型与二次型 . . . . .	176
§A.5 带内积的线性空间 . . . . .	180
§A.6 行列式 . . . . .	185
§A.7 算子范数与谱理论 . . . . .	187
<b>附录 B 微分学基础</b>	<b>193</b>
§B.1 点集拓扑 . . . . .	193
B.1.1 度量空间, 范数 . . . . .	193
B.1.2 开集与闭集 . . . . .	196
B.1.3 紧集, 收敛性, 完备性 . . . . .	199
B.1.4 连续映射 . . . . .	202
B.1.5 与实数有关的性质 . . . . .	205
§B.2 一元函数的微分学 . . . . .	207

B.2.1	导数与微分的定义 . . . . .	207
B.2.2	微分学基本定理 . . . . .	210
§B.3	多元函数的微分学 . . . . .	212
B.3.1	微分、偏导数与导数的定义 . . . . .	213
B.3.2	微分学基本定理 . . . . .	219
B.3.3	隐函数定理 . . . . .	220
附录 C	概率论基础	224
§C.1	从朴素概率论到公理化概率论 . . . . .	224
C.1.1	Kolmogorov 概率论 . . . . .	224
C.1.2	条件概率, 独立性 . . . . .	227
§C.2	随机变量, 分布函数 . . . . .	231
§C.3	随机变量的数字特征, 期望 . . . . .	231
§C.4	多元正态分布 . . . . .	231

# 第一部分

## 科学的逻辑



## 第二部分

### 信息与数据

## 第三部分

### 决策与优化

## 第四部分

### 逻辑与博弈

## 第五部分

### 认知逻辑

## 附录 C 概率论基础

本附录主要介绍 Kolmogorov 概率论，讨论只局限在数学层面，不涉及概率论的哲学讨论。

### §C.1 从朴素概率论到公理化概率论

#### C.1.1 Kolmogorov 概率论

朴素的概率论通常讨论两种极端的情况，一个是可以用数数的方式来计算概率的情况，比如说掷骰子，另一个是用面积的方式来计算概率的情况，比如在随机选一个圆周上的点。这两个情况分别对应了古典概型和几何概型。

我们先给一些术语。考虑一个随机试验，它的所有可能结果组成的集合称为样本空间，记为  $\Omega$ 。样本空间的元素称为样本点，通常记为  $\omega$ 。样本空间的某些子集被称为事件。我们来看看这些概念在朴素的概率论中都具体是什么。

**例 C.1 (古典概型)** 考虑先后掷两个骰子的情况。样本空间为

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}.$$

样本点为  $(i, j)$ ，表示第一个骰子掷出  $i$  点，第二个骰子掷出  $j$  点。“第一个骰子掷出  $i$  点”这个事件可以表示为  $A_i = \{(i, j) : 1 \leq j \leq 6\}$ 。“第一个骰子掷出  $i$  点，第二个骰子掷出  $j$  点”这个事件可以表示为  $B_{ij} = \{(i, j)\}$ 。

**例 C.2 (几何概型)** 考虑随机选一个圆周上的点的情况。如果用弧度来表示圆周上的点，那么样本空间为

$$\Omega = [0, 2\pi).$$

样本点为  $\omega$ ，表示选出点的弧度。事件  $A = [0, \pi)$  表示选出了上半圆周，事件  $B = [0, \pi/2) \cup [\pi, 3\pi/2)$  表示选出了右上或左下的  $1/4$  圆周。

那么，如何定义概率呢？朴素地说，概率是某个事件出现的可能性占总可能的比例。

对于古典概型，我们简单认为每个样本点出现的概率都是相同的，也就是说，如果用  $p_\omega$  表示样本点  $\omega$  出现的概率，那么对任意  $\omega \in \Omega$ ，都有  $p_\omega = 1/|\Omega|$ 。于是，对于任意事件  $A$ ，它发生的概率为

$$\sum_{\omega \in A} p_\omega = \frac{|A|}{|\Omega|}.$$

例如在上面掷骰子的例子中， $p_\omega = 1/36$ ， $A$  发生的概率为  $1/6$ ， $B$  发生的概率为  $1/36$ 。

对于几何概型，不能再用古典概型的方式定义概率。一段长为  $2\pi$  的圆弧上，一个点的长度当然是 0，所以选到一个点的概率是 0。计算选到上半圆周的概率，就是把所有上半圆周上的点的概率加起来，任意多个 0 相加依然还是 0，所以这样的定义出来的概率永远是零，这样是不可行的。

朴素的直觉告诉我们，选到上半圆周的概率是  $1/2$ ，因为上半圆周刚好占了半个圆周。所以几何概型的概率定义利用了体积的概念。事件  $A$  的概率定义为

$$\frac{\text{事件 } A \text{ 对应的体积}}{\text{样本空间 } \Omega \text{ 对应的体积}}.$$

这里体积应该按照广义上来理解，一维集合的体积就是长度，二维集合的体积就是面积，三维集合的体积就是体积，以此类推。例如在上面圆周的例子中， $A$  对应的体积（长度）为  $\pi$ ， $\Omega$  对应的体积（长度）为  $2\pi$ ，所以  $A$  发生的概率为  $1/2$ 。同理， $B$  的概率也是  $1/2$ 。

几何概型的定义非常微妙，因为我们并不知道如何定义“体积”。我们来看一个有趣的例子。

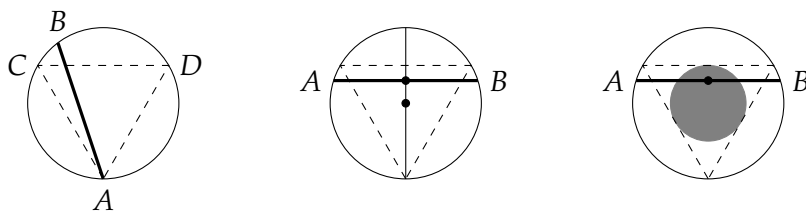
**例 C.3 (Bertrand 悖论)** 考虑一个圆，它的半径为 1。现在我们随机地在圆上取一个弦，那么这个弦的长度超过  $\sqrt{3}$ （即圆内接正三角形的边长）的概率是多少？我们给出三种答案。

**解答 1.** 不妨固定弦的其中一个点  $A$ ，那么另一个点  $B$  可以在圆上等可能取。以  $A$  为顶点作圆内接正三角形  $ACD$ ，弦的长度超过  $\sqrt{3}$  等价于  $B$  在弧  $CD$  上，所以概率为  $1/3$ 。

**解答 2.** 弦长只与它到圆心的距离有关系，与方向无关。弦长超过  $\sqrt{3}$  等价于它到圆心的距离小于  $1/2$ ，所以概率为  $1/2$ 。

**解答 3.** 弦被它的中点唯一确定，弦长大于  $\sqrt{3}$  等价于中点落在一个半径为  $1/2$  的同心小圆内，所以概率为同心小圆面积比上大圆面积，即  $(1/2)^2 = 1/4$ 。

三种解答的示意图见下（从左到右分别是解答 1 到 3）：



因此，我们需要一个更加严格的定义来描述概率。首先注意到，概率应该是一个函数，它的值域是  $[0, 1]$ 。那么，它的定义域应该是什么呢？我们已经看到，概率应该定义在事件上，而非样本点上。那么，概率可以定义在任意事件上吗？这个问题的答案非常微妙，我们不在这里讨论。这里只是指出，我们关心的并不总是任意事件，而是一类被  $\sigma$ -代数所刻画的事件。

**定义 C.1 ( $\sigma$ -代数)** 设  $\Omega$  是一个集合， $\mathcal{F}$  是  $\Omega$  的子集的集合。如果  $\mathcal{F}$  满足

1.  $\Omega \in \mathcal{F}$ ;
2. 如果  $A \in \mathcal{F}$ ，则  $A$  的补集  $\Omega \setminus A \in \mathcal{F}$ ;
3. 如果  $A_1, A_2, \dots \in \mathcal{F}$ ，则  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ 。

则称  $\mathcal{F}$  是  $\Omega$  上的一个  $\sigma$ -代数。

在样本空间中，我们要求事件也形成一个  $\sigma$ -代数。这样的  $\sigma$ -代数称为事件域，记为  $\mathcal{F}$ ，关于这一定义的哲学讨论，可以见第 1 章。接下来，我们给出 Kolmogorov 概率论的公理化定义。

**定义 C.2 (概率空间，概率测度)** 设  $\Omega$  是一个集合， $\mathcal{F}$  是  $\Omega$  上的一个  $\sigma$ -代数。如果函数  $\Pr: \mathcal{F} \rightarrow [0, 1]$  满足

1. 正则性:  $\Pr(\Omega) = 1$ ;
2. 可列可加性: 如果  $A_1, A_2, \dots \in \mathcal{F}$  是两两不相交的事件，则

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i),$$

则称  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $\Pr$  称为概率测度或概率。

容易证明，概率有如下性质：

**命题 C.1** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间，则：

1.  $\Pr(\emptyset) = 0$ ;
2. 单调性：对任意的  $A, B \in \mathcal{F}$ ，如果  $A \subseteq B$ ，则  $\Pr(A) \leq \Pr(B)$ ;
3. 有限可加性：对两两不相交的  $A_1, A_2, \dots, A_n \in \mathcal{F}$ ，有

$$\Pr\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \Pr(A_i).$$

他们的证明都不困难，我们略去。

对于古典概型来说，我们容易写出它的概率空间。此时事件域恰好为所有  $\Omega$  的子集的集合，概率测度的定义也就是我们之前的定义。对于几何概型来说，这一定义需要克服很多技术层面的困难，我们不在这里讨论。

### C.1.2 条件概率，独立性

接下来，我们讨论条件概率与独立性。我们还是看先后掷两个骰子的例子。如果掷完第一个骰子，我们马上观察结果，然后再掷第二个骰子，问第一个骰子是  $i$ ，第二个是  $j$  的概率是多少？如果继续套用原来的概率空间，我们很快就会觉得不对劲。此时，第一个骰子完全没有随机性！所以朴素的直觉告诉我们，这里的概率应该有另一个依赖于第一次投骰子结果的定义，这样的概率就是条件概率。

我们直接给出一般情况下条件概率的定义。

**定义 C.3 (条件概率)** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $A, B \in \mathcal{F}$  是两个事件，且  $\Pr(A) > 0$ 。则称

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

是事件  $B$  在事件  $A$  发生的条件下发生的条件概率。

以上定义要求  $A$  发生概率为正，然而  $A$  是零概率的时候也是可能有条件概率的。例如，从  $[0, 1] \times [0, 1]$  中均匀地随机选一个点  $(X, Y)$ ，观察它的横坐标  $X$ ，不管什么样的  $x$ ， $X = x$  的概率都是 0。然而，从朴素的直觉来看，条件在  $X = x$  上， $Y > 1/2$  的概率不仅存在，而且应该是  $1/2$ 。在附录 C.2 中，我们会针对一类特殊的事件，给出此时条件概率的定义。

我们继续看投两个骰子的例子。假设事件  $A$  是“第一个骰子是  $i$ ”，事件  $B$  是“第二个骰子是  $j$ ”。我们可以计算出  $\Pr(B|A) = \Pr(B) = \frac{1}{6}$ 。如果单看数学计算，这是一个非常神奇的式子：条件在  $A$  上和不条件在  $A$  上概率是一样的！从朴素的直觉来说，这件事情却



并不神秘，因为第一个骰子的结果和第二个骰子的结果是不应该有关系的。我们把这种现象称为独立性。更一般地，对任意事件  $A, B$ ，如果  $\Pr(A) > 0$ ，那么

$$\Pr(B|A) = \Pr(B) \iff \frac{\Pr(A \cap B)}{\Pr(A)} = \Pr(B) \iff \Pr(A \cap B) = \Pr(A) \Pr(B).$$

最后一个式子并不要求  $\Pr(A) > 0$ ，因此我们用它作为独立性的定义，这样定义可以不依赖条件概率。

**定义 C.4 (独立性)** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $A, B \in \mathcal{F}$  是两个事件。如果  $\Pr(A \cap B) = \Pr(A) \Pr(B)$ ，则称事件  $A$  和  $B$  相互独立。

一般地，给定一个事件族  $\mathcal{A} \subseteq \mathcal{F}$ ，如果对任意的有限个不同的  $A_1, A_2, \dots, A_n \in \mathcal{A}$ ，都有

$$\Pr\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \Pr(A_i),$$

则称事件族  $\mathcal{A}$  中的事件是相互独立的。

我们在定义中还给出了多个事件相互独立的定义，这一定义是说不管挑出其中多少个事件，他们都应该满足交的概率等于概率的积。这并不等价于任意两个事件都相互独立，我们看下面的例子。

**例 C.4** 两个人进行石头剪刀布游戏，每个人独立等概率地出剪刀石头布。考虑下面三个事件： $A = \{\text{甲出了石头}\}$ ， $B = \{\text{乙出了剪刀}\}$ ， $C = \{\text{甲赢}\}$ 。

容易算出， $\Pr(A \cap B) = \Pr(A) \Pr(B) = 1/9$ ， $\Pr(A \cap C) = \Pr(A) \Pr(C) = 1/9$ ， $\Pr(B \cap C) = \Pr(B) \Pr(C) = 1/9$ ，所以  $A, B, C$  两两独立。但是  $A, B, C$  不是相互独立的： $\Pr(A \cap B \cap C) = 1/9 \neq 1/27 = \Pr(A) \Pr(B) \Pr(C)$ 。

这个例子说明，三个事件的独立性远比他们任意两个之间的独立性要复杂，三个事件可能放在一起才会出现不独立的情况。对于一般情况，这样的现象更加普遍，所以我们多个事件的独立性定义是要求任意有限个事件都独立，而不是任意两个事件都独立。

最后，我们给出条件概率的一些性质。

**命题 C.2** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间，那么

1. 对任意  $A \in \mathcal{F}$  满足  $\Pr(A) > 0$ ， $\Pr(\cdot|A)$  也是一个概率测度；
2.  $\Pr(|\Omega) = \Pr(\cdot)$ ，
3. 对任意  $A \in \mathcal{F}$  满足  $\Pr(A) > 0$ ， $\Pr(A|A) = 1$ 。

以上性质的证明都很简单，我们就不给出了。

**定理 C.1 (全概率公式)** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $A_1, A_2, \dots \in \mathcal{F}$  是一列两两不相交的事件，且  $\Pr(A_i) > 0$ ， $\bigcup_{i=1}^{\infty} A_i = B$ ，则对任意的  $C \in \mathcal{F}$ ，有

$$\Pr(C|B) = \sum_{i=1}^{\infty} \Pr(C|A_i) \Pr(A_i).$$

特别地，对于有限个  $A_i$ ，这一定理也成立。

**证明** 注意到

$$\Pr(C) = \Pr(C \cap B) = \Pr\left(C \cap \bigcup_{i=1}^{\infty} A_i\right) = \Pr\left(\bigcup_{i=1}^{\infty} (C \cap A_i)\right) = \sum_{i=1}^{\infty} \Pr(C \cap A_i).$$

最后一个等号是因为  $C \cap A_i$  两两不相交。另一方面，

$$\Pr(C \cap A_i) = \Pr(C|A_i) \Pr(A_i),$$

所以

$$\Pr(C) = \sum_{i=1}^{\infty} \Pr(C|A_i) \Pr(A_i).$$

对于有限个  $A_i$ ，只需要把无穷求和改成有限求和，利用有限可加性即可即可。  $\square$

全概率公式是一种分而治之的思想，它把一个复杂的事件分解成若干个简单的事件，然后再把简单的事件的概率加起来。我们来看一个例子。

**例 C.5** 从装有  $w$  个白球和  $b$  个黑球的盒子中随机地取出一个球，不放回，再取出一个球。问第二个球是白球的概率是多少？

设事件  $A$  是“第一个球是白球”，事件  $B$  是“第二个球是白球”。我们有

$$\begin{aligned} \Pr(B) &= \Pr(B|A) \Pr(A) + \Pr(B|\bar{A}) \Pr(\bar{A}) \\ &= \frac{w-1}{w+b-1} \cdot \frac{w}{w+b} + \frac{w}{w+b-1} \cdot \frac{b}{w+b} \\ &= \frac{w}{w+b}. \end{aligned}$$

这里  $\bar{A}$  指的是  $A$  的补集，即“第一个球是黑球”。

**定理 C.2 (贝叶斯公式)** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $A, B \in \mathcal{F}$  且  $\Pr(A), \Pr(B) > 0$ ，则

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}.$$

这一公式的证明几乎是显然的，我们略去。

一个特别重要的推论被称为链式法则，它是 *Bayes* 网络的基础。

**推论 C.1 (链式法则)** 设  $(\Omega, \mathcal{F}, \Pr)$  是一个概率空间， $A_1, A_2, \dots, A_n \in \mathcal{F}$ ，且  $\Pr(A_1 \cap A_2 \cap \dots \cap A_n) > 0$ ，则

$$\begin{aligned} & \Pr(A_1 \cap A_2 \cap \dots \cap A_n) \\ &= \Pr(A_1) \Pr(A_2|A_1) \Pr(A_3|A_1 \cap A_2) \cdots \Pr(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned}$$

我们也看一个例子。

**例 C.6 (Pólya 的罐子)** 一个罐子装有  $w$  个白球和  $b$  个黑球，随机取出一个，观察它的颜色，放回，再放回相同颜色的  $c$  个球，再随机取一次，重复上述操作，如此反复  $n$  次，问每一次都取到白球的概率是多少？

设事件  $A_i$  是“第  $i$  次取出的球是白球”。我们有

$$\begin{aligned} \Pr(A_1) &= \frac{w}{w+b}, \\ \Pr(A_2|A_1) &= \frac{w+c}{w+b+c}, \\ \Pr(A_3|A_1 \cap A_2) &= \frac{w+2c}{w+b+2c}, \\ &\dots \\ \Pr(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}) &= \frac{w+nc}{w+b+nc}. \end{aligned}$$

所以

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \frac{w}{w+b} \cdot \frac{w+c}{w+b+c} \cdots \frac{w+nc}{w+b+nc}.$$

**注.** 在概率论中，我们经常要讨论事件的交，所以我们通常会把  $A \cap B$  简记为  $AB$ 。

另外，我们也经常要讨论一个关于  $\omega$  的陈述  $Q(\omega)$  定义的事件  $\{\omega \in \Omega : Q(\omega)\}$ ，在 Pólya 的罐子的例子中，事件  $A_1$  其实就是由陈述  $Q(\omega)$ ：“ $\omega$  中第一次取出的球是白球”定义的事件。在这种情况下，我们将这一事件简记为  $\{Q\}$ ，它的概率就是  $\Pr(\{Q\})$  或者简记为  $\Pr(Q)$ 。此时，事件交的概率也经常以逗号的形式写出，例如， $\Pr(A_1 A_2)$  我们会记为  $\Pr(\text{第一次取出的球是白球, 第二次取出的球是白球})$ 。这样的记号更直观，并且在随机变量部分会经常使用。

**§C.2 随机变量，分布函数**

**§C.3 随机变量的数字特征，期望**

**§C.4 多元正态分布**

## 参考文献

- [Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.
- [CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.
- [Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.
- [Inf] Information | Etymology, origin and meaning of information by etymonline. <https://www.etymonline.com/word/information>.
- [Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.
- [KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [LLG<sup>+</sup>19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.
- [McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.
- [RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

- [Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.
- [Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.
- [Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.
- [Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.
- [Uff22] Jos Uffink. Boltzmann’s Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.
- [李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

# 索引

$\sigma$ -代数, 226

Bayes 网络, 230

事件域, 226

全概率公式, 229

条件概率, 227

样本点, 224

样本空间, 224

概率, 226

概率测度, 226

概率空间, 226

独立性, 228

贝叶斯公式, 229

链式法则, 230