

PROBABILITY AND STATISTICS

概率论与数理统计

讲义

第 1-3 章；第 4 章至条件矩

Raisetsu41

2026 Autumn

26 October 2026

目录

第 1 章 概率、计数与分布函数	1
第 2 章 变换、期望与矩	31
第 3 章 常见分布族	56
第 4 章 联合分布与条件矩	116
英文概念与原理	124

第1章 概率、计数与分布函数

Probability, Counting and Distribution Functions

依据: CB 2e Ch1; 本学期 Chapter 1 资料

1.1 从随机试验到概率模型

From Random Experiments to Probability Models

随机试验 (random experiment) 的可能结果是明确的, 但试验前无法确定哪一个结果会发生。概率模型的基本结构为

随机试验 $\longrightarrow$ 样本空间 $S \longrightarrow$ 事件 $A \longrightarrow$ 概率 $P(A)$.

1.2 样本空间与事件

Sample Spaces and Events

Definition

(CB 2e §1.1)

For a random experiment, the **sample space** S is the set of all possible basic outcomes.

A set is **countable** if its elements can be placed in one-to-one correspondence with a subset of the integers.

An **event** is any subset of the sample space S . If the realized outcome belongs to $A \subseteq S$, we say that event A occurs.

The set containing no outcomes is denoted by $\emptyset$ and represents an impossible event.

样本空间可以是有限集、可数无限集或不可数集。事件发生, 是指试验实现的基本结果属于对应子集。

集合包含与相等满足

$$\begin{aligned} A \subseteq B &\iff (x \in A \Rightarrow x \in B), \\ A = B &\iff A \subseteq B \text{ 且 } B \subseteq A. \end{aligned}$$

证明两个集合相等时，通常使用双向包含。

1.3 集合运算

Set Operations

1.3.1 基本运算

Basic Operations

$$\begin{aligned} A \cup B &= \{x : x \in A \text{ 或 } x \in B\}, \\ A \cap B &= \{x : x \in A \text{ 且 } x \in B\}, \\ A^c &= \{x \in S : x \notin A\}. \end{aligned}$$

并、交运算满足交换律、结合律和分配律。

1.3.2 De Morgan 律

De Morgan's Laws

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c, \\ (A \cap B)^c &= A^c \cup B^c. \end{aligned}$$

对可数个集合，

$$\begin{aligned} \left(\bigcup_i A_i \right)^c &= \bigcap_i A_i^c, \\ \left(\bigcap_i A_i \right)^c &= \bigcup_i A_i^c. \end{aligned}$$

其中， $\bigcup_{i=1}^{\infty} A_i$ 表示至少属于一个 A_i ； $\bigcap_{i=1}^{\infty} A_i$ 表示属于每一个 A_i 。

1.4 样本空间的分割

Partitions of the Sample Space

Definition

(CB 2e §1.1)

Events $A_1, \dots, A_n$ are **mutually exclusive (pairwise disjoint)** if

$$A_i \cap A_j = \emptyset \quad (i \neq j).$$

Events $A_1, \dots, A_n$ are **collectively exhaustive** if

$$\bigcup_{i=1}^n A_i = S.$$

A collection $A_1, A_2, \dots$ is a **partition** of S if the sets are pairwise disjoint and collectively exhaustive.

分割是全概率公式与 Bayes 公式的基础。

1.5 sigma-代数

Sigma-Algebras

Definition

(CB 2e §1.1)

A collection $\mathcal{B}$ of subsets of S is a **σ -algebra** if

1. $\emptyset \in \mathcal{B}$;
2. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$;
3. $A_1, A_2, \dots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

由 De Morgan 律, σ -algebra 也对可数交封闭。幂集 2^S 与平凡 σ -algebra $\{\emptyset, S\}$ 都满足定义。

1.6 概率模型与概率公理

Probability Models and Axioms

完整模型为：a complete probability model is $(S, \mathcal{B}, P)$ ，其中 S 是 sample space， $\mathcal{B}$ 是 events 的集合， P 为事件赋予概率。

Definition

(CB 2e §1.2)

A **probability function** $P : \mathcal{B} \rightarrow [0, 1]$ satisfies:

1. $P(A) \geq 0$ for every $A \in \mathcal{B}$;
2. $P(S) = 1$;
3. if $A_1, A_2, \dots$ are pairwise disjoint, then

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

因此，一个完整的概率模型写作 $(S, \mathcal{B}, P)$ ，且

$$P : \mathcal{B} \rightarrow [0, 1].$$

三条要求依次称为 nonnegativity、normalization 与 countable additivity。

重要：建模含义

概率公理只约束概率如何运算，并不自动确定某个随机试验应采用哪一种概率模型。具体概率仍来自建模假设或数据。

1.7 概率公理的推论

Consequences of the Probability Axioms

基本结论为

$$\begin{aligned} P(\emptyset) &= 0, \\ 0 &\leq P(A) \leq 1, \\ P(A^c) &= 1 - P(A), \\ A \subseteq B &\Rightarrow P(A) \leq P(B). \end{aligned}$$

两事件公式为

$$\begin{aligned} P(B \cap A^c) &= P(B) - P(A \cap B), \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

1.8 分割公式与概率不等式

Partition Formula and Probability Inequalities

若 $\{C_i\}$ 是 S 的一个分割, 则

$$P(A) = \sum_i P(A \cap C_i).$$

Boole inequality

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i).$$

Bonferroni lower bound

$$P\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

当 $n = 2$ 时,

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

1.9 可数样本空间上的概率构造

Constructing Probabilities on Countable Sample Spaces

设

$$S = \{s_1, s_2, \dots\}$$

为有限或可数样本空间。若 $p_i \geq 0$ 且 $\sum_i p_i = 1$, 则

$$P(A) = \sum_{i: s_i \in A} p_i$$

定义了一个概率函数，即为每个基本结果分配概率质量，再对事件中的结果求和。

1.10 组合计数

Combinatorial Analysis

1.10.1 计数基本原理

Basic Principle of Counting

- **basic principle of counting**: 连续 k 个步骤分别有 $n_1, \dots, n_k$ 种选择时，总方案数为 $n_1 \cdots n_k$;
- **factorial**: $n! = n(n - 1) \cdots 1$ ，并约定 $0! = 1$ 。

从 n 个对象中抽取 r 个对象：

抽样方式	有序	无序
不放回	$\frac{n!}{(n-r)!}$	$\binom{n}{r} = \frac{n!}{r!(n-r)!}$
放回	n^r	$\binom{n+r-1}{r}$

1.10.2 隔板法与多项式系数

Stars and Bars and Multinomial Coefficients

隔板法（stars and bars）把 r 个相同对象分配到 n 个类别中，对应排列 r 个星号与 $n - 1$ 个隔板。

若 k 个位置中有 m 类元素，各类分别重复 $k_1, \dots, k_m$ 次，则不同排列数为

$$\frac{k!}{k_1!k_2! \cdots k_m!}.$$

1.11 等可能模型

Equally Likely Models

若有限样本空间的 N 个基本结果确实等可能，则

$$P(A) = \frac{|A|}{|S|}.$$

警告：使用条件

只有基本结果真正等可能时，才能使用“有利结果数除以总结果数”。无序结果不会因为表达方便而自动等可能。

1.12 条件概率

Conditional Probability

Definition

(CB 2e §1.3)

For events A and B with $P(B) > 0$, the **conditional probability** of A given B is

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

获知 B 已发生后，有效样本空间被限制到 B ； $P(A | B)$ 衡量 B 中同时属于 A 的概率质量占比。

1.12.1 乘法公式

Multiplication Rule

由定义直接得到

$$P(A \cap B) = P(A | B)P(B) = P(B | A)P(A).$$

对 $A_1, \dots, A_n$ 反复使用乘法公式，可得 **chain rule**：

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1) \prod_{i=2}^n P\left(A_i \left| \bigcap_{j=1}^{i-1} A_j\right.\right),$$

前提是式中的条件概率均有定义。

1.12.2 例 1: 连续抽到四张 A

Example 1: Drawing Four Aces in Succession

一副 52 张的标准扑克牌不放回抽取，前四张全是 A 的概率为

$$\begin{aligned} P(\text{first four cards are aces}) &= \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \cdot \frac{1}{49} \\ &= \frac{1}{\binom{52}{4}}. \end{aligned}$$

每次抽牌都会改变下一次的条件样本空间，因此各因子的分子、分母依次减少。

1.12.3 例 2: 不放回抽球

Example 2: Sampling Balls without Replacement

盒中有 r 个红球和 b 个蓝球。连续不放回抽取两个球，令

$$A = \{\text{first ball is red}\}, \quad B = \{\text{second ball is blue}\}.$$

则

$$P(A) = \frac{r}{r+b}, \quad P(B | A) = \frac{b}{r+b-1},$$

所以

$$P(A \cap B) = \frac{r}{r+b} \cdot \frac{b}{r+b-1}.$$

1.13 全概率公式

Law of Total Probability

若 $A_1, A_2, \dots$ 构成样本空间 S 的一个 partition，即两两互斥且并集为 S ，则对任意事件 B ，

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B | A_i)P(A_i).$$

它把事件 B 按照所有可能来源 A_i 拆开，再将各路径的概率相加。

Theorem

(CB 2e §1.3)

If $A_1, A_2, \dots$ form a partition of S , then for any event B ,

$$P(B) = \sum_i P(B | A_i)P(A_i).$$

1.13.1 例 3: 三个组别

Example 3: Three Groups

若 $P(B_i) = 1/3$, 且 $P(A | B_i) = i/6$, $i = 1, 2, 3$, 则

$$\begin{aligned} P(A) &= \sum_{i=1}^3 P(A | B_i)P(B_i) \\ &= \frac{1}{3} \left(\frac{1}{6} + \frac{2}{6} + \frac{3}{6} \right) \\ &= \frac{1}{3}. \end{aligned}$$

1.14 Bayes 公式

Bayes' Rule

若 $A_1, A_2, \dots$ 构成 partition 且 $P(B) > 0$, 则

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_j P(B | A_j)P(A_j)}.$$

Theorem

(CB 2e §1.3)

If $A_1, A_2, \dots$ form a partition and $P(B) > 0$, then

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_j P(B | A_j)P(A_j)}.$$

Bayes' rule reverses the direction of conditioning.

Interpretation

Bayes' rule 把条件方向从 $P(B | A_i)$ 反转为 $P(A_i | B)$:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

分母是由全概率公式得到的 $P(B)$ ，负责把各个未标准化权重归一化。

1.14.1 例 4：汽车保险风险分类

Example 4: Automobile Insurance Risk Classification

客户被分为高风险 H 、中风险 M 和低风险 L :

$$P(H) = 0.25, \quad P(M) = 0.25, \quad P(L) = 0.50,$$

一年内至少发生一次事故的条件概率为

$$P(A | H) = 0.25, \quad P(A | M) = 0.16, \quad P(A | L) = 0.10.$$

已知某客户发生事故，则其属于高风险客户的概率为

$$\begin{aligned} P(H | A) &= \frac{P(A | H)P(H)}{\sum_{T \in \{H, M, L\}} P(A | T)P(T)} \\ &= \frac{0.25^2}{0.25^2 + 0.16 \cdot 0.25 + 0.10 \cdot 0.50} \\ &\approx 0.410. \end{aligned}$$

观察到事故后，高风险类型的概率从 prior 0.25 更新到 posterior 0.410。

1.14.2 例 5: 色盲问题

Example 5: Color Blindness

设总体中男性人数与女性人数之比为 $r : 1$, 则

$$P(M) = \frac{r}{1+r}, \quad P(W) = \frac{1}{1+r}.$$

若色盲事件为 D , 且

$$P(D | M) = p, \quad P(D | W) = p^2,$$

则随机选出的色盲者为男性的概率为

$$\begin{aligned} P(M | D) &= \frac{p \frac{r}{1+r}}{p \frac{r}{1+r} + p^2 \frac{1}{1+r}} \\ &= \frac{r}{r+p}. \end{aligned}$$

1.15 三门问题

Monty Hall Problem

奖品等概率地藏在帘子 A, B, C 后。先选择 A , 主持人知道奖品位置, 并在另外两扇帘子中打开一扇空帘; 假设他打开了 B 。记

$$B^* = \{\text{host opens curtain } B \text{ and it is empty}\}.$$

主持人的行为规则意味着

$$P(B^* | A) = \frac{1}{2}, \quad P(B^* | B) = 0, \quad P(B^* | C) = 1.$$

由 Bayes' rule,

$$\begin{aligned} P(A | B^*) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} \\ &= \frac{1}{3}, \end{aligned}$$

$$P(C | B^*) = \frac{2}{3}.$$

Conclusion

换到 C 的获奖概率为 $2/3$, 高于坚持选择 A 的 $1/3$ 。这个结论依赖主持人“必定打开未选择的空帘, 并在有两扇空帘可选时等概率选择”的行为规则。

1.16 独立性

Independence

Definition

(CB 2e §1.3)

Events A and B are **independent** if

$$P(A \cap B) = P(A)P(B).$$

若 $P(B) > 0$, 则等价地有

$$P(A | B) = P(A),$$

即得知 B 发生不会改变 A 的概率。

1.16.1 独立与互斥

Independence and Mutual Exclusivity

二者含义不同:

- **Disjointness**: $A \cap B = \emptyset$, 两个事件不能同时发生;
- **Independence**: 一个事件是否发生不提供关于另一个事件的信息。

若 A, B 互斥且 $P(A) > 0, P(B) > 0$, 则

$$P(A \cap B) = 0 < P(A)P(B),$$

所以它们不可能独立。只有当至少一个事件概率为 0 时, 互斥与独立才可能同时成立。

1.16.2 例 6: 四次掷骰至少出现一次 6

Example 6: At Least One Six in Four Die Rolls

独立掷一枚公平骰子四次。使用 complement rule,

$$\begin{aligned} P(\text{at least one six}) &= 1 - P(\text{no six}) \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= \frac{671}{1296} \approx 0.518. \end{aligned}$$

更一般地, 若 $A_1, \dots, A_n$ 相互独立, 则其补事件也相互独立, 从而

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= 1 - P\left(\bigcap_{i=1}^n A_i^c\right) \\ &= 1 - \prod_{i=1}^n (1 - P(A_i)). \end{aligned}$$

若 $P(A_i) = p$, 则

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1 - p)^n.$$

1.17 独立性的进一步讨论

Further Discussion of Independence

1.17.1 独立性与补事件

Independence and Complements

Theorem

(CB 2e §1.3)

If A and B are independent, then

$$(A, B^c), \quad (A^c, B), \quad (A^c, B^c)$$

are also independent pairs.

例如，利用 $A = (A \cap B) \cup (A \cap B^c)$ 这一互斥分解，

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \\ &= P(A)[1 - P(B)] \\ &= P(A)P(B^c). \end{aligned}$$

其余两种情形可用相同方法证明。因此，独立事件的任意补集组合仍保持独立。

1.17.2 两两独立与相互独立

Pairwise and Mutual Independence

当事件数量不少于三个时，**pairwise independence** 不足以推出 **mutual independence**。

三个事件两两独立只要求

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(A \cap C) &= P(A)P(C), \\ P(B \cap C) &= P(B)P(C), \end{aligned}$$

但不保证

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Definition

(CB 2e §1.3)

$A_1, \dots, A_n$ are **mutually independent** if for every subcollection $A_{i_1}, \dots, A_{i_k}$,

$$P\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}).$$

Mutual independence 要求所有规模至少为 2 的子事件组都满足乘法关系；只检查每一对事件并不充分。

1.17.3 例 1: 两两独立但不相互独立

Example 1: Pairwise but Not Mutual Independence

设

$$S = \{a, b, c, d\},$$

四个结果等可能，并定义

$$A_1 = \{a, b\}, \quad A_2 = \{b, c\}, \quad A_3 = \{a, c\}.$$

每个事件的概率均为 $1/2$ ，任意两个事件的交集均只含一个结果，因此

$$\begin{aligned} P(A_i \cap A_j) &= \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} \\ &= P(A_i)P(A_j), \quad i \neq j. \end{aligned}$$

所以三个事件两两独立。然而

$$A_1 \cap A_2 \cap A_3 = \emptyset,$$

从而

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= 0 \\ &\neq \frac{1}{8} = P(A_1)P(A_2)P(A_3). \end{aligned}$$

故它们并不相互独立。

1.17.4 练习核对

Checking the In-Class Exercise

1. 盒中有 4 个红球、6 个蓝球，不放回抽取两个球：

$$P(\text{first red, second blue}) = \frac{4}{10} \cdot \frac{6}{9} = \frac{4}{15}.$$

2. 保险分类例中，

$$P(A) = 0.25(0.25) + 0.16(0.25) + 0.10(0.50) = 0.1525,$$

$$\begin{aligned} P(H \mid A) &= \frac{0.25(0.25)}{0.1525} \\ &= \frac{25}{61} \approx 0.410. \end{aligned}$$

3. 上述 A_1, A_2, A_3 的每个两两交集概率为 $1/4$ ，但三者交集概率为 0 ，因此两两独立而不相互独立。

1.18 随机变量

Random Variables

Definition

(CB 2e §1.4)

A **random variable** X is a function from the sample space S into the real numbers:

$$X : S \longrightarrow \mathbb{R}.$$

X 的所有可能取值构成其值域，通常记为 $\mathcal{X}$ 。随机变量并不是“随机变化的数”，而是先把每个原始结果 $s \in S$ 映射为数值 $X(s)$ ；随机性来自试验结果 s 的不确定性。

$$s \in S \quad \xrightarrow{X} \quad X(s) \in \mathbb{R}.$$

1.18.1 随机变量示例

Examples of Random Variables

Random experiment	Random variable
掷两枚骰子	X = 两个点数之和
抛硬币 25 次	X = 正面出现次数
测量作物反应	X = 单位面积产量
抽样调查 50 位选民	X = 支持某提案的人数

随机变量可以把复杂的原始结果压缩为当前分析所关心的 numerical summary。

1.18.2 随机变量诱导的概率

Probabilities Induced by a Random Variable

给定 $X : S \rightarrow \mathcal{X}$, 样本空间 S 上的概率会诱导出 X 取值上的概率。对 $x \in \mathcal{X}$,

$$P(X = x) = P(\{s \in S : X(s) = x\}).$$

更一般地, 对 $A \subseteq \mathcal{X}$,

$$P(X \in A) = P(\{s \in S : X(s) \in A\}).$$

右侧集合是 A 在映射 X 下的原像。因此, 对随机变量事件求概率, 本质上仍然是在样本空间 S 上求概率。

Notation

- 大写字母 X, Y, Z 表示随机变量;
- 小写字母 x, y, z 表示随机变量的可能取值或实现值。

1.18.3 例 2: 三次抛硬币

Example 2: Three Coin Tosses

公平硬币独立抛三次, 令 X 为正面次数。样本空间为

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

X 的概率分布为

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

因此

$$P(0 \leq X \leq 1) = P(X = 0) + P(X = 1) = \frac{4}{8},$$

而

$$P(0 \leq X < 1) = P(X = 0) = \frac{1}{8}.$$

对离散型随机变量，把 $\leq$ 改成 $<$ 可能改变事件本身，不能忽略端点。

1.19 分布函数

Distribution Functions

1.19.1 累积分布函数

Cumulative Distribution Function

Definition

(CB 2e §1.4)

The **cumulative distribution function** of X is

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

CDF 在所有实数 x 上都有定义，而不只是在 X 的值域 $\mathcal{X}$ 上定义。若 $a < b$ ，则

$$P(a < X \leq b) = F_X(b) - F_X(a).$$

1.19.2 例 3：三次抛硬币的 CDF

Example 3: CDF for Three Coin Tosses

对正面次数 X ，

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{8}, & 0 \leq x < 1, \\ \frac{3}{8}, & 1 \leq x < 2, \\ \frac{7}{8}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

这是一个阶梯函数。每个取值点 x_0 处的跳跃大小等于该点的概率质量：

$$F_X(x_0) - F_X(x_0^-) = P(X = x_0),$$

其中

$$F_X(x_0^-) = \lim_{x \uparrow x_0} F_X(x).$$

CDF 在跳跃点取跳跃后的函数值，因此是 **right-continuous**（右连续）的。

1.19.3 例 4：命中点到圆心的距离

Example 4: Distance from the Hit Point to the Center

射击命中半径为 R 的圆形靶，且命中点在靶面上按面积均匀分布。令 X 为命中点到圆心的距离。

当 $0 \leq x \leq R$ 时，事件 $\{X \leq x\}$ 对应半径为 x 的内圆，因此

$$\begin{aligned} F_X(x) &= \frac{\pi x^2}{\pi R^2} \\ &= \frac{x^2}{R^2}. \end{aligned}$$

完整的 CDF 为

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{R^2}, & 0 \leq x \leq R, \\ 1, & x > R. \end{cases}$$

Key point

“在圆面上按面积均匀” 不等于 “距离 X 在 $[0, R]$ 上均匀”。半径越大的环带面积越大，因此 $F_X(x)$ 是二次函数而不是 x/R 。

1.20 累积分布函数

Cumulative Distribution Function

1.20.1 CDF 的刻画

Characterization of a CDF

Theorem

(CB 2e §1.5)

Characterization of a cdf. A function F is a cdf if and only if:

1. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$;
2. F is nondecreasing;
3. F is right-continuous.

一个函数能够成为某个随机变量的累积分布函数，当且仅当它满足左右端极限、单调不减和右连续三个条件。这三个条件不仅是必要条件，也是充分条件。

- 两端极限说明总概率从 0 累积到 1；
- 单调不减来自事件包含关系：

$$x_1 < x_2 \implies \{X \leq x_1\} \subseteq \{X \leq x_2\};$$

- 右连续表示从右侧逼近 x_0 时，CDF 回到 $F_X(x_0)$ 。

1.20.2 右连续性

Right-Continuity

当 $x \downarrow x_0$ 时，事件 $\{X \leq x\}$ 单调递减到 $\{X \leq x_0\}$ 。由概率测度对递减事件列的连续性，

$$\lim_{x \downarrow x_0} F_X(x) = P(X \leq x_0) = F_X(x_0).$$

因此 CDF 必须右连续。它不一定左连续；左侧的跳跃恰好记录点概率。

1.20.3 跳跃概率

Jump Probabilities

定义左极限

$$F_X(x_0^-) = \lim_{x \uparrow x_0} F_X(x).$$

则

$$P(X = x_0) = F_X(x_0) - F_X(x_0^-).$$

若 CDF 在 x_0 连续, 则该点没有概率质量:

$$F_X(x_0) = F_X(x_0^-) \implies P(X = x_0) = 0.$$

1.20.4 区间概率

Interval Probabilities

对任意随机变量,

$$P(a < X \leq b) = F_X(b) - F_X(a),$$

$$P(a \leq X \leq b) = F_X(b) - F_X(a^-).$$

其他端点形式也可由点概率推出:

$$P(a < X < b) = F_X(b^-) - F_X(a),$$

$$P(a \leq X < b) = F_X(b^-) - F_X(a^-).$$

当 CDF 在 a, b 都连续时, 四种开闭区间的概率相同; 对离散型随机变量则必须保留端点差异。

1.21 累积分布函数示例

Examples of CDFs

1.21.1 几何分布的 CDF

CDF of the Geometric Distribution

令 X 表示独立 Bernoulli 试验中第一次成功所需的试验次数，单次成功概率为 p 。则

$$P(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, \dots$$

对正整数 x ,

$$\begin{aligned} F_X(x) &= \sum_{i=1}^x (1 - p)^{i-1}p \\ &= 1 - (1 - p)^x. \end{aligned}$$

对任意实数 x ，完整形式可写为

$$F_X(x) = \begin{cases} 0, & x < 1, \\ 1 - (1 - p)^{\lfloor x \rfloor}, & x \geq 1. \end{cases}$$

CDF 只在正整数处跳跃，相邻整数之间保持不变。

1.21.2 Logistic 分布函数

Logistic CDF

考虑

$$F(x) = \frac{1}{1 + e^{-x}}.$$

其两端极限与导数为

$$\begin{aligned}\lim_{x \rightarrow -\infty} F(x) &= 0, \\ \lim_{x \rightarrow \infty} F(x) &= 1, \\ F'(x) &= \frac{e^{-x}}{(1 + e^{-x})^2} > 0.\end{aligned}$$

因此它连续、严格递增，并满足 CDF 的全部条件。

1.22 同分布随机变量

Identically Distributed Random Variables

Definition

(CB 2e §1.5)

Random variables X and Y are **identically distributed** if

$$P(X \in A) = P(Y \in A)$$

for every admissible set A .

若 X 与 Y 落入任意可测集合的概率都相同，则称二者同分布。它们可以定义在不同样本空间上，也可以是同一样本空间上的不同函数。

Theorem

(CB 2e §1.5)

X and Y are identically distributed if and only if

$$F_X(x) = F_Y(x)$$

for every real x .

一维随机变量的 CDF 唯一决定其概率分布。因此比较两个随机变量是否同分布，只需比较它们在所有实数处的 CDF。

1.22.1 例 1: 三次抛硬币中的正面数与反面数

Example 1: Numbers of Heads and Tails in Three Tosses

三次公平抛硬币，令

X = number of heads,

Y = number of tails.

X 与 Y 都服从参数为 $(3, 1/2)$ 的二项分布，所以同分布。但对每个具体试验结果，

$$X + Y = 3,$$

二者通常不相等，而且存在完全的函数关系。

重要: Identically distributed

同分布不表示 X 与 Y 是样本空间上的同一个函数，也不表示它们相互独立。它只说明两者的边缘概率规律相同。

1.23 概率质量函数

Probability Mass Function

1.23.1 定义

Definition

Definition

(CB 2e §1.6)

For a discrete random variable X , the **probability mass function** is

$$f_X(x) = P(X = x).$$

对离散型随机变量，PMF 直接把每个可能取值映射到该点的概率质量。它满足

$$\begin{aligned} f_X(x) &\geq 0, \\ \sum_x f_X(x) &= 1. \end{aligned}$$

随机变量的 support 为

$$\text{supp}(X) = \{x : P(X = x) > 0\}.$$

PMF 与 CDF 的关系是

$$\begin{aligned} F_X(x) &= \sum_{y \leq x} f_X(y), \\ f_X(x) &= F_X(x) - F_X(x^-). \end{aligned}$$

因此 PMF 的函数值正是 CDF 在相应取值点的跳跃高度。

1.23.2 几何分布的概率质量函数

Geometric PMF

对第一次成功等待时间,

$$f_X(x) = \begin{cases} (1-p)^{x-1}p, & x = 1, 2, \dots, \\ 0, & \text{otherwise.} \end{cases}$$

若 $1 \leq a \leq b$ 且 a, b 为整数, 则

$$\begin{aligned} P(a \leq X \leq b) &= \sum_{k=a}^b (1-p)^{k-1}p \\ &= (1-p)^{a-1} - (1-p)^b. \end{aligned}$$

1.24 概率密度函数

Probability Density Function

1.24.1 定义

Definition

Definition

(CB 2e §1.6)

A function f_X is a **probability density function** for a continuous random variable X when

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

for every x .

对连续型随机变量, PDF 不是点概率, 而是单位长度附近的概率密度; 从 $-\infty$ 积分到 x 才得到累计概率。

若 F_X 可微, 则

$$f_X(x) = F'_X(x).$$

反过来, 由密度积分可得区间概率:

$$\begin{aligned} P(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x) dx. \end{aligned}$$

1.24.2 点概率

Point Probabilities

对连续型随机变量,

$$P(X = x) = 0$$

对每个 x 都成立。因此

$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b).$$

重要: Density is not probability

$f_X(x)$ 不是 $P(X = x)$ 。密度值可以大于 1；只要 f_X 非负且曲线下的总面积为 1，它就是合法密度。概率必须通过对集合积分得到。

1.24.3 区间上的均匀密度

Uniform Density on an Interval

若

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{otherwise,} \end{cases}$$

则对 $a \leq c < d \leq b$,

$$\begin{aligned} P(c < X < d) &= \int_c^d \frac{1}{b-a} dx \\ &= \frac{d-c}{b-a}. \end{aligned}$$

密度为常数，因此概率与区间长度成正比。

1.25 归一化

Normalization

Theorem

(CB 2e §1.6)

A function f is a valid pmf or pdf if and only if:

1. $f(x) \geq 0$ for all x ;
2. the total mass equals one.

概率质量或概率密度必须处处非负，而且在整个取值空间上的总质量必须恰好等于 1。

分别写成

$$\sum_x f(x) = 1 \quad \text{for a pmf,}$$

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad \text{for a pdf.}$$

1.25.1 归一化原则

Normalization Principle

若 $h(x) \geq 0$ 且

$$0 < K = \int_{-\infty}^{\infty} h(x) dx < \infty,$$

则

$$f(x) = \frac{h(x)}{K}$$

是合法 PDF。 h 决定分布的形状，归一化常数 K 把曲线下的总面积调整为 1。

对离散情形同理：若 $h(x) \geq 0$ 且 $0 < \sum_x h(x) < \infty$ ，则用总和归一化即可得到 PMF。

1.25.2 例 2：求归一化常数

Example 2: Finding a Normalizing Constant

设

$$f(x) = \begin{cases} cx^2, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

由总面积为 1,

$$\begin{aligned} 1 &= \int_0^1 cx^2 dx \\ &= \frac{c}{3}, \end{aligned}$$

故 $c = 3$ 。于是

$$\begin{aligned} P\left(X > \frac{1}{2}\right) &= \int_{1/2}^1 3x^2 dx \\ &= 1 - \frac{1}{8} \\ &= \frac{7}{8}. \end{aligned}$$

1.25.3 例 3: Logistic 密度

Example 3: Logistic Density

由

$$F_X(x) = \frac{1}{1 + e^{-x}},$$

求导得到

$$f_X(x) = F'_X(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

区间概率既可由 CDF 作差，也可对 PDF 积分：

$$\begin{aligned} P(a < X < b) &= \int_a^b f_X(x) dx \\ &= F_X(b) - F_X(a). \end{aligned}$$

当密度表示存在时，CDF 与 PDF 携带同一分布的全部信息；二者只是累积表示与局部密度表示的区别。

1.26 离散与连续随机变量

Discrete and Continuous Random Variables

项目	Discrete random variable	Continuous random variable
核心函数	PMF $f_X(x) = P(X = x)$	PDF $f_X(x)$
点概率	$P(X = x) = f_X(x)$, 可以为正	$P(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} f_X(t)$	$F_X(x) = \int_{-\infty}^x f_X(t) dt$
区间概率	对区间内 PMF 求和	对区间上 PDF 积分
CDF 形状	通常为带跳跃的阶梯函数	若存在 PDF, 则 CDF 连续且绝对连续
与 CDF 的联系	PMF 是 CDF 的跳跃量	PDF 在可微点是 CDF 的导数

统一地说, CDF 对所有随机变量都存在; PMF 与 PDF 是适用于特定分布类型的更便捷表示。

第2章 变换、期望与矩

Transformations, Expectations and Moments

依据: CB 2e Ch2; 本学期 Chapter 2 资料

2.1 随机变量函数的分布

Distributions of Functions of a Random Variable

设随机变量 X 的分布已知, 并定义

$$Y = g(X).$$

目标是由 X 的分布求出 Y 的分布。首先要确定 g 在 X 的 support 上是否 one-to-one, 再选择离散求和、CDF method 或 change of variables。

2.1.1 离散变换

Discrete Transformations

Theorem

(CB 2e §2.1)

If X is discrete and $Y = g(X)$, then

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

同一个 y 可能对应多个 x 。因此, $Y = y$ 的概率等于所有满足 $g(x) = y$ 的原像点概率之和。

Example

(CB 2e §2.1)

设 X 在集合 $\{-2, -1, 0, 1, 2\}$ 上均匀分布, 令 $Y = X^2$ 。则

$$\begin{aligned}P(Y = 0) &= P(X = 0) = \frac{1}{5}, \\P(Y = 1) &= P(X = -1) + P(X = 1) = \frac{2}{5}, \\P(Y = 4) &= P(X = -2) + P(X = 2) = \frac{2}{5}.\end{aligned}$$

$g(x) = x^2$ 不是 one-to-one, 所以除 0 外, 每个 Y 的可能取值都要合并两个原像点的概率。

2.1.2 分布函数法

CDF Method

Definition

(CB 2e §2.1)

For $Y = g(X)$, begin with

$$F_Y(y) = P(g(X) \leq y).$$

Rewrite the event in terms of X , then differentiate when a density exists.

CDF method 直接处理事件 $\{g(X) \leq y\}$, 不要求 g 单调。它特别适用于非单调变换、最大值或最小值, 以及变换后 support 随 y 改变的情形。

求解时需要先把不等式改写为关于 X 的事件:

$$\begin{aligned}F_Y(y) &= P(Y \leq y) \\&= P(g(X) \leq y).\end{aligned}$$

若最后得到的 F_Y 可微, 则

$$f_Y(y) = F'_Y(y).$$

2.1.3 一一变量变换

One-to-One Change of Variables

Theorem

(CB 2e §2.1)

Suppose $Y = g(X)$ and g is strictly monotone and differentiable. If $x = g^{-1}(y)$, then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

for y in the transformed support.

绝对导数

$$\left| \frac{d}{dy} g^{-1}(y) \right|$$

是 one-dimensional Jacobian。它修正了变量变换对区间长度的伸缩，使变换前后的概率保持一致。

Example

(CB 2e §2.1)

设 $U \sim \text{Unif}(0, 1)$ ，并令

$$Y = -\log U.$$

由 $u = e^{-y}$ 可知 Y 的 support 为 $y > 0$ ，且

$$\left| \frac{du}{dy} \right| = e^{-y}.$$

因此

$$\begin{aligned} f_Y(y) &= f_U(e^{-y}) \left| \frac{du}{dy} \right| \\ &= e^{-y}, \quad y > 0. \end{aligned}$$

故

$$Y \sim \text{Exp}(1).$$

也可用 CDF method 验证：对 $y > 0$,

$$\begin{aligned} F_Y(y) &= P(-\log U \leq y) \\ &= P(U \geq e^{-y}) \\ &= 1 - e^{-y}. \end{aligned}$$

2.1.4 非一一变换

Non-One-to-One Transformations

Theorem

(CB 2e §2.1)

If g has inverse branches $x_j(y)$, then

$$f_Y(y) = \sum_j f_X(x_j(y)) \left| \frac{dx_j}{dy} \right|.$$

当一个 y 对应多个原像点时，每个 inverse branch 都贡献一项，最后把这些密度相加。

Example

(CB 2e §2.1)

令 $Y = X^2$ 。对 $y > 0$ ，两个 inverse branches 为

$$x_1(y) = \sqrt{y}, \quad x_2(y) = -\sqrt{y}.$$

相应的 Jacobian 绝对值均为 $1/(2\sqrt{y})$ ，所以

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}, \quad y > 0.$$

2.1.5 正态分布与卡方分布

Normal and Chi-Square Distributions

Theorem

(CB 2e §2.1)

If $Z \sim N(0, 1)$, then

$$Z^2 \sim \chi_1^2.$$

令 $Y = Z^2$ 。标准正态密度关于原点对称，因此对 $y > 0$,

$$\begin{aligned} f_Y(y) &= \frac{\phi(\sqrt{y}) + \phi(-\sqrt{y})}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y/2}. \end{aligned}$$

这正是自由度为 1 的 chi-square density。

进行变量变换时，应先写出 X 的 support，再求 $Y = g(X)$ 的 support，并列出全部 inverse branches。所得 PMF 或 PDF 还应满足非负性和总质量为 1。

2.2 期望

Expected Values

2.2.1 无意识统计学家法则

Law of the Unconscious Statistician (LOTUS)

Theorem

(CB 2e §2.2)

For a function g for which the expectation exists,

$$E[g(X)] = \sum_x g(x)p_X(x)$$

for discrete X , and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) dx$$

for continuous X .

该结论称为 **Law of the Unconscious Statistician (LOTUS)**。计算 $E[g(X)]$ 时，可以直接对 X 的分布求和或积分，不必先推导 $Y = g(X)$ 的分布。

例如，若 X 连续且二阶矩存在，则

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

2.2.2 期望的线性性质

Linearity of Expectation

Theorem

(CB 2e §2.2)

Whenever the expectations exist,

$$E[aX + bY + c] = aE[X] + bE[Y] + c.$$

Independence is not required.

更一般地，

$$E \left[\sum_{i=1}^n a_i X_i \right] = \sum_{i=1}^n a_i E[X_i].$$

期望的线性性只要求相关期望存在，不要求随机变量之间独立。独立性会在乘积期望或方差加法中发挥作用，但不是线性性的前提。

2.2.3 指示变量

Indicator Variables

Definition

(CB 2e §2.2)

For an event A , let $\mathbf{1}_A$ equal 1 when A occurs and 0 otherwise. Then

$$E[\mathbf{1}_A] = P(A).$$

示性变量把事件是否发生转化为取值为 0 或 1 的随机变量。若

$$N = \sum_{i=1}^n \mathbf{1}_{\{X_i \in A\}}$$

表示落入集合 A 的观测个数，则

$$\begin{aligned} E[N] &= \sum_{i=1}^n E[\mathbf{1}_{\{X_i \in A\}}] \\ &= \sum_{i=1}^n P(X_i \in A). \end{aligned}$$

若 $X_1, \dots, X_n$ identically distributed, 且 $P(X_i \in A) = p$, 则

$$E[N] = np.$$

这里仍不需要独立性；只要每一项的边缘概率为 p 即可。

2.2.4 方差

Variance

Definition

(CB 2e §2.2)

The variance of X is

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2.$$

方差度量随机变量围绕均值的平方离散程度。由定义可得

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$

平移常数 b 不改变离散程度，而缩放 a 会使方差乘以 a^2 。

2.2.5 均值最小化平方距离

The Mean Minimizes Squared Distance

Theorem

(CB 2e §2.2)

If $E[X^2] < \infty$, then the constant a minimizing

$$E[(X - a)^2]$$

is $a = E[X]$.

Proof

令 $\mu = E[X]$, 则

$$\begin{aligned} E[(X - a)^2] &= E[(X - \mu + \mu - a)^2] \\ &= E[(X - \mu)^2] + 2(\mu - a)E[X - \mu] + (\mu - a)^2 \\ &= \text{Var}(X) + (\mu - a)^2. \end{aligned}$$

第一项与 a 无关，第二项在且仅在 $a = \mu$ 时为 0，所以均值是平方损失下的最优常数预测。

2.3 诱导分布与原像

Induced Distributions and Inverse Images

2.3.1 变换诱导的分布

Distribution Induced by a Transformation

Definition

(CB 2e §2.1)

设随机变量 X 的分布已知, $Y = g(X)$ 。对 Y 的值域中的任意集合 A ,

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ &= P(X \in g^{-1}(A)), \end{aligned}$$

其中

$$g^{-1}(A) = \{x : g(x) \in A\}$$

称为集合 A 在 g 下的 **inverse image**。

因此, 求 Y 的分布可以理解为:

关于 Y 的事件 $\rightarrow$ 关于 X 的事件 $\rightarrow$ 利用 X 的分布求概率.

2.3.2 原像与反函数

Inverse Image versus Inverse Function

inverse image 总是可以定义, 但 ordinary inverse function 只有在 g 为 one-to-one 时才存在。

例如, 若 $g(x) = x^2$, 则对 $y > 0$,

$$g^{-1}(\{y\}) = \{-\sqrt{y}, \sqrt{y}\}.$$

这里的 $g^{-1}(\{y\})$ 是一个包含两个点的集合, 而不是一个单值反函数。两个原像分支都会对 Y 在 y 附近的概率作出贡献。

警告: Common mistake / 常见错误

直接写 $x = g^{-1}(y)$, 却没有先检查 g 在 X 的 support 上是否一一对应, 会漏掉 inverse branches。例如 $Y = X^2$ 不能只保留 $x = \sqrt{y}$ 。

2.4 离散变换

Discrete Transformations

2.4.1 一般公式

General Formula

Theorem

(CB 2e §2.1)

若 X 为离散型随机变量, $Y = g(X)$, 则

$$p_Y(y) = P(Y = y) = \sum_{x:g(x)=y} p_X(x).$$

计算时可以先列出映射表, 再把落到同一 y 的概率相加。这里不需要 Jacobian, 因为离散分布比较的是点上的 probability mass, 而不是单位长度内的 density。

2.4.2 多对一变换示例

Many-to-One Example

Example

(CB 2e §2.1)

设

$$P(X = -1) = \frac{1}{4}, \quad P(X = 0) = \frac{1}{2}, \quad P(X = 1) = \frac{1}{4},$$

并令 $Y = X^2$ 。映射关系为:

x	$g(x) = x^2$	$P(X = x)$
-1	1	1/4
0	0	1/2
1	1	1/4

因此

$$P(Y = 0) = \frac{1}{2}, \quad P(Y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

求和检查给出 $P(Y = 0) + P(Y = 1) = 1$ 。

2.4.3 二项分布变换

Binomial Transformation

Example

(CB 2e §2.1)

设 $X \sim \text{Binomial}(n, p)$, 并定义

$$Y = n - X.$$

若 X 表示 n 次 Bernoulli 试验中的成功次数, 则 Y 表示失败次数。对 $y = 0, 1, \dots, n$, 方程 $n - x = y$ 给出 $x = n - y$, 所以

$$\begin{aligned} p_Y(y) &= p_X(n - y) \\ &= \binom{n}{n - y} p^{n - y} (1 - p)^y \\ &= \binom{n}{y} (1 - p)^y p^{n - y}. \end{aligned}$$

因此

$$Y \sim \text{Binomial}(n, 1 - p).$$

这个变换保留了 Binomial family, 只是把“成功”的定义互换, 使参数由 p 变为 $1 - p$ 。

2.5 分布函数法

CDF Method

2.5.1 一般步骤

General Procedure

Method

对任意变换 $Y = g(X)$, 先写

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y),$$

再把不等式 $g(X) \leq y$ 化成关于 X 的事件。若 F_Y 在相应区间可微, 则

$$f_Y(y) = F'_Y(y).$$

CDF method 尤其适合以下情况:

- g 不是 one-to-one;
- 同一个 y 有多个 inverse branches;
- 变换后的 support 变化较大;
- 与其直接求反函数, 不如先处理不等式更简单。

2.5.2 先确定支撑集

Support-First Principle

设 X 的 support 为

$$\mathcal{X} = \{x : f_X(x) > 0\},$$

则 $Y = g(X)$ 的 support 是像集

$$\mathcal{Y} = g(\mathcal{X}) = \{g(x) : x \in \mathcal{X}\}.$$

在求导或套用变换公式之前, 依次完成:

1. 写出 X 的 support $\mathcal{X}$;
2. 把 $\mathcal{X}$ 通过 g 映射为 $\mathcal{Y}$;
3. 检查 g 在 $\mathcal{X}$ 上是否单调;

4. 若不单调, 把 $\mathcal{X}$ 分割成若干单调区间;
5. 最终答案必须连同 $\mathcal{Y}$ 一起写出。

许多变量变换题的错误并非微积分错误, 而是 support error。

注: Boundary convention / 端点约定

对连续型随机变量, 给积分区间增减有限个端点不会改变概率。分布的支持集通常取闭集, 密度公式可以只在相应开区间内给出; 只要端点没有点质量, 端点处如何指定密度值就不影响分布。

2.5.3 平方变换

Square Transformation

设 $Y = X^2$ 。对 $y > 0$,

$$\begin{aligned} F_Y(y) &= P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

若 X 连续且 CDF 可微, 则

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}, \quad y > 0.$$

分子中两项分别来自 $x = \sqrt{y}$ 与 $x = -\sqrt{y}$ 两个 inverse branches。

2.5.4 均匀分布平方变换

Uniform-Square Example

Example

(CB 2e §2.1)

设

$$f_X(x) = \begin{cases} 1, & -\frac{1}{2} < x < \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

并令 $Y = X^2$ 。因为 $|X| < 1/2$, 所以 Y 的 support 为 $[0, 1/4]$ 。

当 $0 < y < 1/4$ 时,

$$\begin{aligned} F_Y(y) &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} 1 \, dx \\ &= 2\sqrt{y}. \end{aligned}$$

完整的 CDF 为

$$F_Y(y) = \begin{cases} 0, & y \leq 0, \\ 2\sqrt{y}, & 0 < y < \frac{1}{4}, \\ 1, & y \geq \frac{1}{4}. \end{cases}$$

因此

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{y}}, & 0 < y < \frac{1}{4}, \\ 0, & \text{otherwise.} \end{cases}$$

虽然 $f_Y(y)$ 在 $y \downarrow 0$ 时趋于无穷, 但它仍然是合法密度, 因为

$$\int_0^{1/4} \frac{1}{\sqrt{y}} \, dy = 2\sqrt{y} \Big|_0^{1/4} = 1.$$

密度值大于 1 或在某点附近无界并不违反概率公理; 真正需要检查的是密度非负且总积分为 1。

2.5.5 正弦平方变换

Sine-Square Example

Example

(CB 2e §2.1)

设

$$X \sim U(0, 2\pi), \quad Y = \sin^2 X.$$

首先由 $0 \leq \sin^2 X \leq 1$ 得到 Y 的 support 为 $[0, 1]$ 。对 $0 < y < 1$, 令

$$a = \arcsin(\sqrt{y}), \quad 0 < a < \frac{\pi}{2}.$$

事件 $\sin^2 X \leq y$ 等价于 $|\sin X| \leq \sqrt{y}$ 。在一个完整周期 $(0, 2\pi)$ 内，满足该条件的区间总长度为 $4a$ 。由于 X 在 $(0, 2\pi)$ 上均匀分布，

$$F_Y(y) = \frac{4a}{2\pi} = \frac{2}{\pi} \arcsin(\sqrt{y}), \quad 0 < y < 1.$$

求导得到

$$f_Y(y) = \frac{1}{\pi \sqrt{y(1-y)}}, \quad 0 < y < 1.$$

完整写法为

$$F_Y(y) = \begin{cases} 0, & y \leq 0, \\ \frac{2}{\pi} \arcsin(\sqrt{y}), & 0 < y < 1, \\ 1, & y \geq 1, \end{cases}$$

以及

$$f_Y(y) = \begin{cases} \frac{1}{\pi \sqrt{y(1-y)}}, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

该分布也称为 arcsine distribution，并且

$$Y \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right).$$

此例的关键不是直接解出某个单值反函数，而是先画出或想象 $\sin^2 x$ 在一个周期内的图像，找齐所有满足不等式的区间。

2.6 单调变换

Monotone Transformations

2.6.1 CDF 变换定理

CDF Theorem

Theorem

(CB 2e §2.1)

设 $Y = g(X)$, 且 g 在 X 的 support $\mathcal{X}$ 上单调。

若 g 严格递增, 则

$$F_Y(y) = F_X(g^{-1}(y)), \quad y \in \mathcal{Y}.$$

这是因为

$$g(X) \leq y \iff X \leq g^{-1}(y).$$

若 g 严格递减, 且 X 连续, 则

$$F_Y(y) = 1 - F_X(g^{-1}(y)), \quad y \in \mathcal{Y}.$$

此时不等号方向反转:

$$g(X) \leq y \iff X \geq g^{-1}(y).$$

注: Endpoint nuance / 端点细节

对一般随机变量, $P(X \geq a) = 1 - F_X(a^-)$ 。递减公式假设 X 连续, 因此 $F_X(a^-) = F_X(a)$ 。

2.6.2 密度变换定理

Density Transformation Theorem

Theorem

(CB 2e §2.1)

设 X 在 $\mathcal{X}$ 上具有连续密度 f_X , g 在 $\mathcal{X}$ 上单调, 且反函数 g^{-1} 在 $\mathcal{Y}$ 上可微, 则

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y}, \\ 0, & \text{otherwise.} \end{cases}$$

绝对值项是 one-dimensional Jacobian。它修正变换对横轴长度的拉伸或压缩, 并保证密度非负。

2.6.3 均匀分布与指数分布

Uniform-Exponential Relationship

Example

(CB 2e §2.1)

设 $X \sim U(0, 1)$, 并定义

$$Y = -\log X.$$

$g(x) = -\log x$ 在 $(0, 1)$ 上严格递减, 像集为 $(0, \infty)$, 反函数为

$$g^{-1}(y) = e^{-y}.$$

因此对 $y > 0$,

$$\begin{aligned} F_Y(y) &= P(-\log X \leq y) \\ &= P(X \geq e^{-y}) \\ &= 1 - e^{-y}, \end{aligned}$$

从而

$$f_Y(y) = e^{-y}, \quad y > 0.$$

所以 Y 服从均值为 1 的 exponential distribution。

2.6.4 Gamma 分布倒数变换

Inverted Gamma Transformation

Example

(CB 2e §2.1)

设 X 具有 shape-scale 参数化的 Gamma density:

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0,$$

并令

$$Y = \frac{1}{X}.$$

这里 $\mathcal{X} = (0, \infty)$, $\mathcal{Y} = (0, \infty)$, 且

$$g^{-1}(y) = \frac{1}{y}, \quad \left| \frac{d}{dy} g^{-1}(y) \right| = \frac{1}{y^2}.$$

代入单调密度变换公式:

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{1}{y}\right) \frac{1}{y^2} \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{-(\alpha+1)} \exp\left(-\frac{1}{\beta y}\right), \quad y > 0. \end{aligned}$$

该结果是 inverse-gamma distribution。最容易漏掉的部分是 Jacobian $1/y^2$ 。

2.7 分段单调变换

Piecewise Monotone Transformations

2.7.1 一般定理

General Theorem

Theorem

(CB 2e §2.1)

若 X 的 support (忽略 probability-zero set 后) 可分割为

$$A_1, \dots, A_k,$$

且 g 在每个 A_i 上单调并具有可微反函数 g_i^{-1} , 那么对 $y \in \mathcal{Y}$,

$$f_Y(y) = \sum_{i=1}^k f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|,$$

其中只对能够映到该 y 的 branches 求和。

它与离散变换公式具有相同结构:

在 y 处的总贡献 = $\sum$ 所有原像分支的贡献.

2.7.2 正态分布与卡方分布

Normal-Chi-Squared Relationship

Example

(CB 2e §2.1)

设

$$X \sim N(0, 1), \quad Y = X^2.$$

对 $y > 0$, 两个 inverse branches 为

$$x_1(y) = -\sqrt{y}, \quad x_2(y) = \sqrt{y},$$

并且

$$\left| \frac{d}{dy} x_1(y) \right| = \left| \frac{d}{dy} x_2(y) \right| = \frac{1}{2\sqrt{y}}.$$

记标准正态密度为

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

由 $\phi(-x) = \phi(x)$,

$$\begin{aligned} f_Y(y) &= \phi(-\sqrt{y}) \frac{1}{2\sqrt{y}} + \phi(\sqrt{y}) \frac{1}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi y}} e^{-y/2}, \quad y > 0. \end{aligned}$$

这正是自由度为 1 的 chi-squared density, 因此

$$X^2 \sim \chi_1^2.$$

2.8 概率积分变换

Probability Integral Transform

2.8.1 CDF 的广义逆

Generalized Inverse of a CDF

Definition

(CB 2e §2.1)

对单调不减的 CDF F , 定义 generalized inverse 为

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \quad 0 < u < 1.$$

即使 F 有平台区间或跳跃, 这一定义仍然是单值的。它既是 quantile 的严格定义, 也是 inverse-CDF simulation 的数学基础。

2.8.2 概率积分变换定理

Probability Integral Transform Theorem

Theorem

(CB 2e §2.1)

若 X 的 CDF F_X 连续, 则

$$U = F_X(X) \sim U(0, 1).$$

等价地, 对 $0 < u < 1$,

$$P(U \leq u) = u.$$

Proof idea

在 F_X 连续并采用 generalized inverse 的条件下,

$$\begin{aligned} P(F_X(X) \leq u) &= P(X \leq F_X^{-1}(u)) \\ &= F_X(F_X^{-1}(u)) \\ &= u. \end{aligned}$$

警告: Continuity matters / 连续性条件

若 X 为离散型随机变量, $F_X(X)$ 通常不是 $U(0, 1)$ 。例如 Bernoulli 随机变量的 $F_X(X)$ 只会取有限个值, 不可能具有连续均匀分布。

2.8.3 逆变换抽样

Inverse Transform Sampling

概率积分变换可以反向使用。若

$$U \sim U(0, 1),$$

并令

$$X = F^{-1}(U),$$

则 X 的 CDF 为 F 。直观上, U 随机选择一个 cumulative probability level, 而 F^{-1} 把这个概率水平映射回对应的 quantile。

算法可写为:

1. 生成 $U \sim U(0, 1)$;
2. 求解 $U = F_X(x)$;
3. 令 $X = F_X^{-1}(U)$ 。

2.8.4 生成指数随机变量

Generating an Exponential Random Variable

Example

(CB 2e §2.1)

设目标分布的 CDF 为

$$F(x) = 1 - e^{-x/\lambda}, \quad x \geq 0,$$

其中 $\lambda > 0$ 。令 $u = F(x)$, 则

$$u = 1 - e^{-x/\lambda}$$

给出

$$x = -\lambda \log(1 - u).$$

因此若 $U \sim U(0, 1)$, 则

$$X = -\lambda \log(1 - U)$$

服从均值为 λ 的 exponential distribution。由于 $1 - U \sim U(0, 1)$, 实践中也常写为

$$X = -\lambda \log U.$$

这解释了为什么很多随机数生成器只需要先产生 uniform random numbers, 就能进一步构造其他连续分布。

2.9 方法选择与错误检查

Method Selection and Error Checking

2.9.1 方法比较

Comparison of Methods

情形	推荐方法	核心操作
X 离散	PMF 求和	合并所有满足 $g(x) = y$ 的点 概率
g 单调且反函数简单	Density transformation	代入反函数并乘 Jacobian
g 非单调但可分段单调	Branch summation	对所有 inverse branches 的 密度贡献求和
support 或不等式较复杂	CDF method	先求 $P(g(X) \leq y)$ ，再求导
需要从目标分布抽样	Inverse CDF	令 $X = F^{-1}(U)$

2.9.2 变量变换检查清单

Transformation Checklist

Checklist

面对 $Y = g(X)$ 时，按以下顺序检查：

- 1. 写出 X 的 support $\mathcal{X}$;
- 2. 求像集 $\mathcal{Y} = g(\mathcal{X})$;
- 3. 判断 g 在 $\mathcal{X}$ 上是否 one-to-one;
- 4. 若不是，找全 inverse branches 或改用 CDF method;
- 5. 正确处理不等式方向与 Jacobian 的绝对值;
- 6. 把 PMF、CDF 或 PDF 连同 support 一起写出;
- 7. 检查 PMF 是否求和为 1，或 PDF 是否积分为 1。

2.9.3 高频错误

Frequent Errors

错误	后果	修正方式
未先写 $\mathcal{Y} = g(\mathcal{X})$	答案 support 错误	先映射端点、临界点与整个区间
非一一变换只取一个根	少算概率或密度	列出所有满足 $g(x) = y$ 的 roots
递减变换不反转不等号	CDF 写反	从事件 $g(X) \leq y$ 重新推导
漏掉 $ dg^{-1}/dy $	密度尺度错误	把反函数求导列为独立一步
未写 absolute value	递减变换得到负密度	Jacobian 必须取绝对值
只写区间内部公式	CDF 不完整	补上 support 外的 0 与 1
看到 density 大于 1 就判错	误解密度含义	检查积分，而非逐点高度

2.10 概念联系

Conceptual Connections

2.10.1 离散与连续情形的统一视角

A Unified View of Discrete and Continuous Cases

离散型与连续型变量变换的共同原则都是“收集所有原像贡献”：

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x)$$

与

$$f_Y(y) = \sum_i f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|$$

具有相同逻辑。连续情形额外出现 Jacobian，是因为 density 会随坐标尺度的改变而调整。

2.10.2 与前序内容的联系

Connection with Earlier Material

- CDF 提供最通用的事件概率表达；
- 随机变量函数的分布给出变量变换的基础框架；
- 本节将这一框架扩展到 multiple branches、support-first analysis 与 probability integral transform；
- inverse-CDF simulation 之后会直接用于统计模拟、Monte Carlo 方法和数值实验。

2.11 术语表

Terminology

English	中文	说明
transformation	变换	由 $Y = g(X)$ 定义新随机变量
induced distribution	诱导分布	X 的分布经 g 产生的 Y 的分布
inverse image	原像	所有被 g 映到目标集合的输入点
inverse function	反函数	仅在一一对应时存在的单值函数
support	支撑集	随机变量可能取值的范围
inverse branch	反函数分支	非单调函数在单调子区间上的反函数
Jacobian	雅可比因子	一维中为 $ dg^{-1}(y)/dy $
probability integral transform	概率积分变换	连续 X 满足 $F_X(X) \sim U(0, 1)$
generalized inverse	广义逆	$F^{-1}(u) = \inf\{x : F(x) \geq u\}$
inverse transform sampling	逆变换抽样	用 $F^{-1}(U)$ 生成目标分布样本

第3章 常见分布族

Common Families of Distributions

依据: CB 2e Ch3; 本学期 Chapter 3 资料

3.1 分布族与参数

Distribution Families and Parameters

3.1.1 什么是分布族

What Is a Distribution Family?

Definition

(CB 2e §3.2)

A statistical model usually specifies a family

$$\{f(x | \theta) : \theta \in \Theta\}.$$

Different choices of θ generate different probability laws.

A family organizes distributions with a common mathematical form.

一组由参数 θ 索引的概率分布称为 distribution family:

$$\mathcal{F} = \{f(x | \theta) : \theta \in \Theta\}.$$

对离散分布, f 表示 PMF; Θ 是允许的参数集合。 X 是随机变量, θ 在当前概率模型中是固定的模型参数, 二者不能混淆。

例如 $X \sim \text{Bin}(n, p)$: 给定 n, p 后, X 的具体值仍随机; 改变 p , 便得到同一分布族中的另一个分布。

3.1.2 四步识别法

A Four-Step Identification Rule

1. **机制**：独立重复试验、有限总体无放回抽样，还是固定区间内事件计数？
2. **停止条件**：固定试验次数，还是达到某个成功次数才停止？
3. **计数对象**：成功个数、总试验次数，还是失败个数？
4. **支持集**：最小值、最大值是什么？零有没有可能？

参数写清之后，**support** 也必须随之写清。比如超几何分布中，即使样本量是 K ，成功个数也未必能取遍 $0, \dots, K$ 。

3.2 伯努利分布

Bernoulli Distribution

3.2.1 一次二结果试验

One Trial with Two Outcomes

Definition

(CB 2e §3.2)

A random variable X has a Bernoulli(p) distribution if

$$X = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p, \end{cases} \quad 0 \leq p \leq 1.$$

The value 1 is conventionally called a success and 0 a failure.

设一次试验有两种结果，把关注的结果记为成功。成功概率为 p ，成功时 $X = 1$ ，失败时 $X = 0$ ，记作 $X \sim \text{Bernoulli}(p)$ 。

“成功”只是数学标签，不代表结果好。例如研究灯泡损坏时，可以把“损坏”定义为成功。

PMF 为

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

对 $0 < p < 1$ 可统一写成

$$f(x) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}.$$

$p = 0$ 或 $p = 1$ 时用前面的两个概率直接定义退化分布，避免机械处理 0^0 。

3.2.2 均值、方差与矩母函数

Mean, Variance, and MGF

Mean

X 只取 0 与 1，逐项相加：

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = p.$$

含 0 的项可以直接删去，这是指示变量型分布的共同特点。

Second Moment and Variance

指示变量满足 $X^2 = X$ ，因为 $0^2 = 0$ 、 $1^2 = 1$ 。于是

$$E(X^2) = E(X) = p,$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p).$$

用阶乘矩核对： $X(X-1) \equiv 0$ ，故 $E[X(X-1)] = 0$ ，于是

$$E(X^2) = E[X(X-1)] + E(X) = p,$$

与上面一致。

MGF

仍按定义求和，注意 $e^{t \cdot 0} = 1$ ：

$$M_X(t) = (1-p)e^0 + pe^t = 1-p+pe^t, \quad t \in \mathbb{R}.$$

只有两项，所以 M_X 对一切实数 t 都有限，这是六类分布中条件最好的一个。

Checking Moments with the MGF

$$M'_X(t) = pe^t, \quad M''_X(t) = pe^t,$$

因此 $M'_X(0) = p$ 、 $M''_X(0) = p$ ，从而

$$\text{Var}(X) = M''_X(0) - [M'_X(0)]^2 = p - p^2 = p(1 - p).$$

两条路线给出同一结果；做题时可用一种方法验证另一种。

Intuition

$p = 1/2$ 时不确定性最大，方差 $p(1 - p)$ 达到最大值 $1/4$ ； $p \rightarrow 0$ 或 $p \rightarrow 1$ 时结果越来越确定，方差趋于零。方差关于 p 与 $1 - p$ 对称，说明把“成功”与“失败”互换不改变不确定性。

3.3 二项分布

Binomial Distribution

3.3.1 固定次数下数成功个数

Fixed Trials and a Success Count

Definition

(CB 2e §3.2)

Let $X_1, \dots, X_n$ be independent Bernoulli(p) variables and define

$$Y = X_1 + \dots + X_n.$$

Then Y counts the number of successes in n independent trials.

For exactly y successes:

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0, 1, \dots, n.$$

We write

$$Y \sim \text{Bin}(n, p).$$

A fixed number of independent Bernoulli trials with common success probability is the defining binomial mechanism.

原句用 Y 表示成功计数；以下统一用 X 表示同一个计数变量，用 I_i 表示第 i 次试验的成功指示变量。

进行 n 次相互独立、成功概率均为 p 的 Bernoulli 试验，令 X 为成功次数：

$$X = \sum_{i=1}^n I_i, \quad I_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p), \quad X \sim \text{Bin}(n, p).$$

n 是固定非负整数， $0 \leq p \leq 1$ 。通常讨论 $n \geq 1$ ， $n = 0$ 时 $X = 0$ 。

三个关键条件是：固定次数、相互独立、成功概率相同。仅仅“两种结果”并不足以判断为二项分布。

3.3.2 推导概率质量函数

Deriving the PMF

任意一个含 x 次成功的特定排列，其概率为 $p^x(1-p)^{n-x}$ 。从 n 个位置选出成功位置，有 $\binom{n}{x}$ 种：

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

端点 $p = 0, 1$ 按退化分布解释。归一化来自二项式定理：

$$\sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n = 1.$$

3.3.3 均值的两种推导

Mean: Two Derivations

Route 1: Indicator Decomposition

把成功个数写成 n 个指示变量之和：

$$X = \sum_{i=1}^n I_i, \quad I_i = \mathbf{1}\{\text{第 } i \text{ 次成功}\}.$$

期望的线性性不要求独立性，因此可以逐项相加：

$$E(X) = \sum_{i=1}^n E(I_i) = \sum_{i=1}^n p = np.$$

Route 2: Direct Summation and a Combinatorial Identity

设 $q = 1 - p$ ，按定义展开：

$$E(X) = \sum_{x=0}^n x \binom{n}{x} p^x q^{n-x}.$$

$x = 0$ 的项为零。对 $x \geq 1$ 使用恒等式

$$x \binom{n}{x} = n \binom{n-1}{x-1},$$

它的含义是“先指定一个成功位置，再从剩余 $n - 1$ 个位置中选出其余成功位置”。代入并令 $y = x - 1$ ：

$$\begin{aligned} E(X) &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x q^{n-x} \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y q^{(n-1)-y} \\ &= np (p + q)^{n-1} = np. \end{aligned}$$

最后一步把求和识别为 $\text{Bin}(n-1, p)$ 的概率总和，等于 1。这条路线表明：把 x 吸收进组合数后， n 次试验的均值可化为“一个成功位置”的贡献。

3.3.4 方差

Variance

Route 1: Indicators and Covariance

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(I_i) + 2 \sum_{i < j} \text{Cov}(I_i, I_j).$$

相互独立使全部协方差为零，而 $\text{Var}(I_i) = p(1-p)$ ，故

$$\text{Var}(X) = np(1-p).$$

两个前提缺一不可：方差相加要求协方差为零，这里由独立性保证。若试验之间存在相关，上式不再成立。

Route 2: Factorial Moments

先算被组合数吸收的二阶乘积：

$$x(x-1) \binom{n}{x} = n(n-1) \binom{n-2}{x-2},$$

它来自“先指定两个成功位置”。于是

$$\begin{aligned} E[X(X-1)] &= \sum_{x=2}^n n(n-1) \binom{n-2}{x-2} p^x q^{n-x} \\ &= n(n-1)p^2 \sum_{y=0}^{n-2} \binom{n-2}{y} p^y q^{n-2-y} \\ &= n(n-1)p^2. \end{aligned}$$

再由 $E(X^2) = E[X(X-1)] + E(X)$ ：

$$\text{Var}(X) = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

两条路线给出同一结果。

3.3.5 矩母函数与可加性

MGF and the Sum Principle

由 Bernoulli 的 MGF 与独立和的乘积法则,

$$M_X(t) = \prod_{i=1}^n (1 - p + pe^t) = (1 - p + pe^t)^n, \quad t \in \mathbb{R}.$$

这个式子立刻给出一个原理: 若 $X \sim \text{Bin}(n, p)$ 与 $Y \sim \text{Bin}(m, p)$ 相互独立, 则

$$M_{X+Y}(t) = (1 - p + pe^t)^{n+m},$$

由 MGF 唯一性得 $X + Y \sim \text{Bin}(n + m, p)$ 。“成功次数相加仍是成功次数”, 这就是二项分布可加性的来源。

Checking Moments with the MGF

记 $u = 1 - p + pe^t$, 则 $u' = pe^t$:

$$M'(t) = npe^t u^{n-1},$$

$$M''(t) = npe^t u^{n-1} + n(n-1)p^2 e^{2t} u^{n-2}.$$

在 $t = 0$ 处 $u = 1$, $e^t = 1$, 于是

$$M'(0) = np, \quad M''(0) = np + n(n-1)p^2,$$

$$\text{Var}(X) = np + n(n-1)p^2 - n^2 p^2 = np(1-p).$$

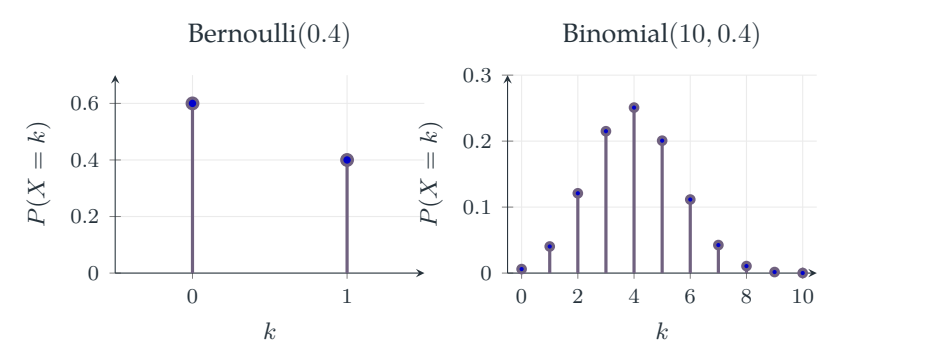
Example

(CB 2e §3.2)

10 次独立试验, 每次成功概率 0.2:

$$P(X = 3) = \binom{10}{3} (0.2)^3 (0.8)^7, \quad E(X) = 2, \quad \text{Var}(X) = 1.6.$$

均值等于 2 不代表每次实验都恰好观察到 2 次成功。



固定试验次数增加后，概率质量分散到更多整数位置；峰值落在均值 $np = 4$ 附近。

3.4 超几何分布

Hypergeometric Distribution

3.4.1 无放回抽样

Sampling without Replacement

Definition

(CB 2e §3.2)

Suppose a finite population has N objects, of which M are labeled “successes.” A sample of size K is drawn without replacement, and X is the number of successes sampled.

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}},$$

for feasible integer values of x .

总体共有 N 个对象，其中 M 个属于成功类。等可能地无放回抽取 K 个，令 X 为样本中的成功个数。

参数条件为 $N \geq 1$ 、 $0 \leq M \leq N$ 、 $0 \leq K \leq N$ ，均为整数。support 是

$$\max\{0, K - (N - M)\} \leq x \leq \min\{K, M\}.$$

上界不能超过样本量或总体成功数；若失败对象总数不足以填满样本，下界就不再是零。

3.4.2 用组合计数推导

Counting Samples

所有 K 元样本共有 $\binom{N}{K}$ 个。若恰有 x 个成功，应从 M 个成功中选 x 个、从 $N - M$ 个失败中选 $K - x$ 个：

$$P(X = x) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}.$$

support 外概率为零。没有额外乘排列数，因为分子分母都以无序样本为计数单位。分子求和等于分母，是 Vandermonde 恒等式的计数解释。

3.4.3 均值：对称性与指示变量

Mean: Symmetry and Indicators

令 $p = M/N$ ，把第 i 个抽样位置是否抽到成功类对象记为 I_i ，则 $X = \sum_{i=1}^K I_i$ ，但 $I_1, \dots, I_K$ 并不独立。

Marginal Probability

每个位置的边际成功概率都等于 p 。这一点的依据是对称性：在等可能抽样下，总体中每个对象落在第 i 个位置的机会都是 $1/N$ ，其中恰有 M 个成功类对象，因此

$$P(I_i = 1) = \frac{M}{N} = p, \quad i = 1, \dots, K.$$

由期望线性性（仍不要求独立性），

$$E(X) = \sum_{i=1}^K E(I_i) = Kp = \frac{KM}{N}.$$

均值与有放回抽样的二项均值相同：不放回只改变波动，不改变平均成功个数。

3.4.4 方差与有限总体修正

Variance and the Finite-Population Correction

Covariance Route

由于不独立，方差必须保留协方差项：

$$\text{Var}(X) = \sum_{i=1}^K \text{Var}(I_i) + 2 \sum_{i < j} \text{Cov}(I_i, I_j).$$

对两个不同位置，连续抽到成功类对象的概率为

$$E(I_i I_j) = P(I_i = 1, I_j = 1) = \frac{M}{N} \cdot \frac{M-1}{N-1} = \frac{M(M-1)}{N(N-1)},$$

第二个因子体现“已经抽走一个成功对象”。于是

$$\text{Cov}(I_i, I_j) = \frac{M(M-1)}{N(N-1)} - \frac{M^2}{N^2} = -\frac{p(1-p)}{N-1}.$$

协方差为负：一次抽到成功类会略微减少后续抽到成功类的机会。共有 $\binom{K}{2}$ 对位置，而 $2\binom{K}{2} = K(K-1)$ ，所以

$$\begin{aligned} \text{Var}(X) &= Kp(1-p) - K(K-1)\frac{p(1-p)}{N-1} \\ &= Kp(1-p) \left(1 - \frac{K-1}{N-1}\right) \\ &= \boxed{Kp(1-p) \frac{N-K}{N-1}}. \end{aligned}$$

Factorial-Moment Check

与二项分布对应，先算

$$E[X(X-1)] = \frac{K(K-1)M(M-1)}{N(N-1)},$$

它来自“任取两个位置都抽到成功对象”的计数。代入

$$\text{Var}(X) = E[X(X-1)] + E(X) - [E(X)]^2$$

并整理，得到同一结果。

$\frac{N-K}{N-1}$ 是 finite-population correction, 满足 $0 \leq \frac{N-K}{N-1} \leq 1$: 不放回抽样消除了部分独立波动, 故方差不大于相同 K, p 的二项方差。 $K = N$ 时样本即总体, 成功数恒为 M , 修正因子为零、方差为零; $N = 1$ 时含 $N - 1$ 的公式失效, 但分布本身退化, 方差仍为零。

3.4.5 生成函数与二项极限

Generating Functions and the Binomial Limit

超几何分布没有简洁的常用 MGF 闭式, 但可写成有限和

$$M_X(t) = \sum_x e^{tx} \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, \quad t \in \mathbb{R},$$

它对每个实数 t 都有限; “没有简洁闭式” 并不等于 “MGF 不存在”。

更实用的工具是阶乘矩母函数。使用恒等式 $\binom{x}{k} \binom{M}{x-k} = \binom{M}{k} \binom{M-k}{x-k}$ 与 Vandermonde 恒等式可得

$$E\binom{X}{k} = \frac{\binom{K}{k} \binom{M}{k}}{\binom{N}{k}}, \quad k = 0, 1, \dots, K.$$

也就是说 $E[(1+t)^X] = \sum_{k=0}^K E\binom{X}{k} t^k$ 的每个系数都容易计算: $k = 1$ 给出均值 Kp , $k = 2$ 给出 $E[X(X-1)]$, 与前面结果一致。

Relation to the Binomial

若 $N \rightarrow \infty$ 、 $M/N \rightarrow p$ 而 K 固定, 则

$$\frac{M}{N} \cdot \frac{M-1}{N-1} \rightarrow p^2, \quad \frac{N-K}{N-1} \rightarrow 1,$$

即位置之间的协方差趋于零、修正因子趋于 1, 超几何分布收敛到 $\text{Bin}(K, p)$ 。直观上, 总体很大时 “不放回” 几乎不改变总体比例, 抽样近似独立。反之, K/N 不可忽略时 (例如 $K/N = 40\%$) 就不能用二项近似。

3.4.6 例题与比较

Example and Comparison

Example

(CB 2e §3.2)

25 个对象中有 6 个成功类，抽 10 个，样本一个成功都没有的概率为

$$P(X = 0) = \frac{\binom{6}{0} \binom{19}{10}}{\binom{25}{10}} = \frac{13}{460} \approx 0.028261.$$

均值为 $10(6/25) = 2.4$ 。不能精确写成 $\text{Bin}(10, 6/25)$ ，因为抽样没有放回。

当 K/N 很小，总体比例几乎不随抽样改变，二项近似才可能合理。这里 $10/25 = 40\%$ ，不能忽略无放回影响。

3.5 泊松分布

Poisson Distribution

3.5.1 固定区间内数事件

Counting Events in an Interval

Definition

(CB 2e §3.2)

For $\lambda > 0$,

$$X \sim \text{Poisson}(\lambda)$$

if

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

其中 $\lambda > 0$ 是该区间的平均事件次数。 $\lambda = 0$ 时可单独定义 $X = 0$ 的退化分布。

归一化来自指数级数：

$$\sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

Poisson 经常用来描述某段时间、面积或长度中的事件数。在齐次 Poisson 过程假设下，若每单位时间平均发生 ν 次，长度为 h 的区间内计数参数为 $\lambda = \nu h$ 。

仅知道“平均每分钟多少次”，不足以保证真实计数一定服从 Poisson；模型还依赖发生机制，例如不相交区间增量独立、恒定发生率以及小区间多次发生概率可忽略等。

3.5.2 均值、方差与矩母函数

Mean, Variance, and MGF

Mean via the Ratio of Adjacent Probabilities

$$\frac{p_x}{p_{x-1}} = \frac{e^{-\lambda} \lambda^x / x!}{e^{-\lambda} \lambda^{x-1} / (x-1)!} = \frac{\lambda}{x},$$

即 $x p_x = \lambda p_{x-1}$ ($x \geq 1$)。把定义式中的 x 换成这个关系：

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x p_x = \lambda \sum_{x=1}^{\infty} p_{x-1} \\ &= \lambda \sum_{y=0}^{\infty} p_y = \lambda. \end{aligned}$$

全程只用到“概率总和为 1”，不需要处理阶乘。

Variance via Factorial Moments

对 $x(x-1)p_x$ 用同一递推关系两次：

$$x(x-1)p_x = \lambda^2 p_{x-2},$$

于是

$$E[X(X-1)] = \sum_{x=2}^{\infty} x(x-1)p_x = \lambda^2 \sum_{y=0}^{\infty} p_y = \lambda^2.$$

由 $E(X^2) = E[X(X-1)] + E(X)$ 得 $E(X^2) = \lambda^2 + \lambda$ ，因此

$$\text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

均值等于方差是 Poisson 分布最显著的特征，也是判断数据是否适合该模型的第一条线索。

MGF

$$\begin{aligned} M_X(t) &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} \exp\{\lambda e^t\} = \exp\{\lambda(e^t - 1)\}, \end{aligned}$$

对一切实数 t 有限。求导得

$$M'_X(t) = \lambda e^t M_X(t),$$

$$M''_X(t) = \lambda e^t M_X(t) + \lambda^2 e^{2t} M_X(t).$$

在 $t = 0$ 处 $M_X(0) = 1$ ，故 $M'_X(0) = \lambda$ 、 $M''_X(0) = \lambda + \lambda^2$ ，方差为 λ 。

Additivity

若 $X \sim \text{Poisson}(\lambda)$ 与 $Y \sim \text{Poisson}(\mu)$ 相互独立，则

$$M_{X+Y}(t) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)},$$

故 $X + Y \sim \text{Poisson}(\lambda + \mu)$ ：不相交区间的事件数相加仍是事件数，这正是 Poisson 分布适合描述计数的原因。

实际数据中若均值与方差差异很大，说明 Poisson 模型可能不合适；但二者相近也不足以证明模型正确。

3.5.3 时间换算与尾概率

Rescaling Time and Computing Tails

Example

(CB 2e §3.2)

设平均每 3 分钟接到 5 个电话。1 分钟内的参数为

$$\lambda = \frac{5}{3}.$$

没有电话的概率：

$$P(X = 0) = e^{-5/3} \approx 0.188876.$$

至少 2 个电话：

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-5/3} \left(1 + \frac{5}{3} \right) \approx 0.496332. \end{aligned}$$

不要把“至少 2”写成 $1 - P(X = 0)$ ，那是“至少 1”。

3.5.4 概率递推公式

A Recursion for Probabilities

设 $p_x = P(X = x)$ ，则

$$p_0 = e^{-\lambda}, \quad p_x = \frac{\lambda}{x} p_{x-1}, \quad x \geq 1.$$

来源是相邻两项 PMF 的比值。算多个相邻概率时，从 p_0 递推比重复计算阶乘更方便。

3.5.5 二项分布的泊松极限

The Binomial Limit

若 $X_n \sim \text{Bin}(n, p_n)$ ，且

$$n \rightarrow \infty, \quad p_n \rightarrow 0, \quad np_n \rightarrow \lambda,$$

则对每个固定非负整数 x ，

$$P(X_n = x) \rightarrow e^{-\lambda} \frac{\lambda^x}{x!}.$$

为看清原理，取 $p_n = \lambda/n$ ：

$$\begin{aligned} P(X_n = x) &= \frac{n(n-1) \cdots (n-x+1) \lambda^x}{n^x x!} \\ &\quad \times \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\ &\longrightarrow 1 \cdot \frac{\lambda^x}{x!} \cdot e^{-\lambda} \cdot 1. \end{aligned}$$

核心是“很多机会，每次概率很小，但平均总次数保持有限”，而不只是 n 大。

Example

(CB 2e §3.2)

5 页材料约有 1500 个单词，每个单词独立出现排字错误的概率为 $1/500$ 。错误数的精确模型为

$$X \sim \text{Bin}(1500, 1/500),$$

可用 Poisson(3) 近似。至多 2 个排字错误：

$$P(X \leq 2) \approx e^{-3} \left(1 + 3 + \frac{3^2}{2}\right) \approx 0.423190.$$

二项分布直接计算约为 0.422966。两者接近，但近似值不等于精确值。

3.5.6 泊松过程的原理

The Poisson Process Principle

把事件在时间上的发生记为计数过程 $\{N(t)\}_{t \geq 0}$ 。称它是发生率为 $\nu > 0$ 的齐次 Poisson 过程，如果满足：

1. $N(0) = 0$;
2. 增量独立：不相交时间区间上的计数相互独立；
3. 增量平稳：长度为 h 的区间内计数分布只依赖 h ；
4. 小区间至多发生一次： $P(N(h) = 1) = \nu h + o(h)$, $P(N(h) \geq 2) = o(h)$ 。

由此可推出两条常用结论。

Conclusion 1

$N(t) \sim \text{Poisson}(\nu t)$ 。把 $[0, t]$ 等分为 n 段，每段长度 t/n 。当 n 很大时，每段发生次数的期望为 $\nu t/n$ ，一段内发生两次以上的概率可忽略，各段又相互独立，因此 $N(t)$ 近似服从

$$\text{Bin}\left(n, \frac{\nu t}{n}\right).$$

令 $n \rightarrow \infty$ ，由 6.5 节的泊松极限定理得到 $N(t) \sim \text{Poisson}(\nu t)$ 。这也说明 6.5 节极限定理的用途：把“ n 次独立小概率试验”翻译成“连续时间上的稀有事件计数”。

Conclusion 2

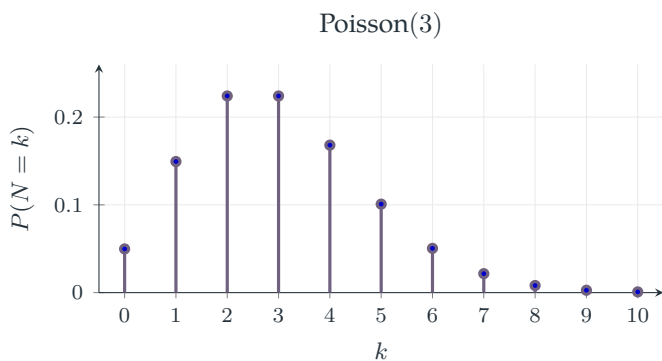
相邻事件的时间间隔服从指数分布。设 T 为第一次事件发生的时刻。事件在 t 之前尚未发生，等价于 $[0, t]$ 内计数为零：

$$P(T > t) = P(N(t) = 0) = e^{-\nu t},$$

所以 T 的分布函数为 $1 - e^{-\nu t}$ ，密度为 $\nu e^{-\nu t}$ 。指数分布与离散的几何分布承担同一角色——都是“等待第一次成功”，因此都具有无记忆性：

$$P(T > s + t \mid T > s) = e^{-\nu t} = P(T > t).$$

这解释了本章反复出现的结构：几何分布是离散时间上的等待时间，Poisson 分布是固定时间内的计数，二者通过指数分布相互联系。



参数 $\lambda = 3$ 同时是均值和方差；支持集从 0 起，没有固定的上界。

3.6 几何分布

Geometric Distribution

3.6.1 第一次成功所需总次数

Trials until the First Success

Definition

(CB 2e §3.2)

A discrete random variable X follows a geometric distribution with parameter p , $0 < p < 1$, if

$$P(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, \dots$$

Here X represents the number of independent Bernoulli trials required to obtain the first success.

独立重复 Bernoulli 试验，成功概率固定为 $0 < p < 1$ 。令 X 为第一次成功发生时的总试验次数，记作 $X \sim \text{Geom}(p)$ 。

原因是：前 $x - 1$ 次全部失败，第 x 次成功。此处 X 最小为 1，不是 0。

令 $q = 1 - p$ ，则

$$P(X > n) = q^n, \quad n = 0, 1, \dots,$$

因为前 n 次全部失败。对整数 $n \geq 1$,

$$P(X \leq n) = 1 - q^n.$$

若写一般实数 x 的 CDF, $x < 1$ 时为零, $x \geq 1$ 时为 $1 - q^{\lfloor x \rfloor}$ 。离散 CDF 是阶梯函数。

3.6.2 矩母函数、均值与方差

MGF, Mean, and Variance

记 $q = 1 - p$ 。

Normalization

PMF 求和为 1，由等比级数保证：

$$\sum_{x=1}^{\infty} q^{x-1} p = p \cdot \frac{1}{1-q} = 1.$$

Mean: Differentiating the Geometric Series

先记住基本级数 ($|q| < 1$):

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q}, \quad \sum_{x=1}^{\infty} x q^{x-1} = \frac{d}{dq} \frac{1}{1-q} = \frac{1}{(1-q)^2}.$$

于是

$$E(X) = \sum_{x=1}^{\infty} x q^{x-1} p = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}.$$

技巧在于把 x 看作用 q 求导产生的因子，因而不必逐项化简 xq^{x-1} 。这与 Poisson 分布用相邻概率比值消去 x 是同一思想。

Variance via Factorial Moments

对同一级数求二阶导：

$$\sum_{x=1}^{\infty} x(x-1) q^{x-2} = \frac{2}{(1-q)^3},$$

两边乘 pq ：

$$E[X(X-1)] = \sum_{x=1}^{\infty} x(x-1) q^{x-1} p = \frac{2q}{p^2}.$$

由 $E(X^2) = E[X(X-1)] + E(X)$,

$$E(X^2) = \frac{2q}{p^2} + \frac{1}{p},$$

$$\text{Var}(X) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q + p - 1}{p^2} = \frac{q}{p^2},$$

最后一步用 $p + q = 1$ 。

MGF and a Derivative Check

仍由等比级数,

$$M_X(t) = \sum_{x=1}^{\infty} e^{tx} q^{x-1} p = \frac{pe^t}{1 - qe^t}, \quad qe^t < 1.$$

收敛条件必须保留, 即 $t < -\log q$; $p = 1$ ($q = 0$) 时 $X \equiv 1$, MGF 为 e^t 且对一切实数有限, 属于退化情形。对 M_X 求导:

$$M'_X(t) = \frac{pe^t}{(1 - qe^t)^2}, \quad M''_X(t) = \frac{pe^t(1 + qe^t)}{(1 - qe^t)^3}.$$

在 $t = 0$ 处 $1 - q = p$, 所以

$$M'_X(0) = \frac{1}{p}, \quad M''_X(0) = \frac{1 + q}{p^2} = \frac{2 - p}{p^2},$$

$$\text{Var}(X) = \frac{2 - p}{p^2} - \frac{1}{p^2} = \frac{1 - p}{p^2},$$

与阶乘矩路线结果一致。

3.6.3 无记忆性

Memorylessness

Theorem

(CB 2e §3.2)

对非负整数 n, m ,

$$\boxed{P(X > n + m \mid X > n) = P(X > m)}.$$

Proof

$$P(X > n + m \mid X > n) = \frac{q^{n+m}}{q^n} = q^m = P(X > m).$$

等价地，当 $s \geq t$ 是非负整数时，

$$P(X > s \mid X > t) = P(X > s - t).$$

已经等了 n 次且仍未成功，未来等待时间的分布和刚开始时一样。关键是每次成功概率没有随历史改变，且试验独立，不是“已经失败很多次，下次必然更容易成功”。

The Converse Also Holds

设取值于 $\{1, 2, \dots\}$ 的随机变量满足无记忆性，记 $g(n) = P(X > n)$ ，则条件化为

$$g(n + m) = g(n)g(m).$$

取 $m = 1$ 并反复使用，得 $g(n) = [g(1)]^n = q^n$ ，其中 $q = P(X > 1)$ ；于是

$$P(X = n) = g(n - 1) - g(n) = q^{n-1}(1 - q),$$

这正是几何分布。所以几何分布是 $\{1, 2, \dots\}$ 上唯一的无记忆分布，对应连续时间上唯一的无记忆分布——指数分布。

Example

(CB 2e §3.2)

设灯泡每天独立地以 $p = 0.001$ 的概率损坏，把“损坏”定义为成功。灯泡使用超过 30 天的概率为

$$P(X > 30) = (0.999)^{30} \approx 0.970431.$$

已使用 20 天且未坏，再用超过 30 天的条件概率也为 0.999^{30} 。这是恒定日损坏概率模型的结果；真实灯泡若老化改变故障概率，就不一定无记忆。

3.6.4 改为数第一次成功前的失败数

Counting Failures Instead

若令 $Y = X - 1$, 则

$$P(Y = y) = q^y p, \quad y = 0, 1, \dots,$$

$$E(Y) = \frac{q}{p}, \quad \text{Var}(Y) = \frac{q}{p^2}, \quad M_Y(t) = \frac{p}{1 - qe^t}.$$

这也是一些教材和软件的 **geometric** 参数化。因为平移不改变方差，两种定义的均值相差 1，方差相同。

对该定义，适合写无记忆性为

$$P(Y \geq n + m \mid Y \geq n) = P(Y \geq m).$$

不要把总次数定义中的严格大于号机械照搬到失败数定义，否则容易差 1。

3.7 负二项分布

Negative Binomial Distribution

3.7.1 第 r 次成功所需总次数

Trials until the r -th Success

Definition

(CB 2e §3.2)

In independent Bernoulli(p) trials, let X be the trial on which the r th success occurs.

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots$$

This is one common negative-binomial parameterization.

设独立试验成功概率为 $0 < p < 1$ ，目标成功次数 r 是正整数。令 X 为取得第 r 次成功所需的总试验次数。

若 $X = x$:

- 最后一次必须成功;
- 前 $x - 1$ 次恰有 $r - 1$ 次成功;
- 共有 $x - r$ 次失败。

前 $x - 1$ 次试验的成功位置有 $\binom{x-1}{r-1}$ 种, 每一种完整排列的概率为 $p^r(1-p)^{x-r}$; 相乘就得到上述 PMF。

组合数不是 $\binom{x}{r}$, 因为最后一次成功的位置已被固定。

3.7.2 等待时间之和

A Sum of Waiting Times

从第 $j - 1$ 次成功到第 j 次成功之间的等待时间 (按总试验次数计) 记为 G_j 。由独立重复试验假设, $G_1, \dots, G_r$ 相互独立且都服从 $\text{Geom}(p)$, 而总试验次数正是这些等待时间之和:

$$X = G_1 + \dots + G_r.$$

Mean and Variance

由期望的线性性与方差的可加性 (后者要求相互独立):

$$E(X) = \sum_{j=1}^r E(G_j) = \frac{r}{p},$$

$$\text{Var}(X) = \sum_{j=1}^r \text{Var}(G_j) = \frac{r(1-p)}{p^2}.$$

MGF

由独立和的乘积法则与 7.2 节的结果,

$$M_X(t) = \prod_{j=1}^r \frac{pe^t}{1 - qe^t} = \left(\frac{pe^t}{1 - qe^t} \right)^r, \quad t < -\log q,$$

其中 $q = 1 - p$ 。 $r = 1$ 时回到几何分布; $p = 1$ 时 $X \equiv r$, 属退化情形。

MGF Verified Directly from the PMF

令 $x = r + y$ ($y \geq 0$):

$$\begin{aligned} M_X(t) &= \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r q^{x-r} \\ &= (pe^t)^r \sum_{y=0}^{\infty} \binom{r+y-1}{y} (qe^t)^y \\ &= \frac{(pe^t)^r}{(1 - qe^t)^r}, \end{aligned}$$

最后一步用了负二项级数

$$\sum_{y=0}^{\infty} \binom{r+y-1}{y} z^y = \frac{1}{(1-z)^r}, \quad |z| < 1,$$

它同时也是 PMF 归一化 (在 $t = 0$ 处取 $z = q$) 的依据。两条路线给出同一 MGF, 说明“等待时间之和”的解释与 PMF 完全一致。

3.7.3 第 r 次成功前的失败次数

Failures before the r -th Success

Definition

(CB 2e §3.2)

$$Y = X - r,$$

the number of failures before the r th success. Then

$$P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y, \quad y = 0, 1, 2, \dots$$

令 $Y = X - r$, 把总试验次数中的 r 次成功扣掉, 剩下的就是失败次数。此时 support 为 $\{0, 1, 2, \dots\}$ 。组合数也可写成 $\binom{y+r-1}{r-1}$, 因为 $\binom{y+r-1}{y} = \binom{y+r-1}{r-1}$ 。

对应矩与 MGF:

$$E(Y) = \frac{r(1-p)}{p}, \quad \text{Var}(Y) = \frac{r(1-p)}{p^2},$$

$$M_Y(t) = \left(\frac{p}{1 - (1-p)e^t} \right)^r.$$

Example

(CB 2e §3.2)

成功概率 $p = 0.2$ ，等到第 3 次成功。总次数均值为 $3/0.2 = 15$ ，之前失败数均值为 $15 - 3 = 12$ ，两者方差都为 $3(0.8)/(0.2)^2 = 60$ 。

总次数恰好为 5 的概率：

$$P(X = 5) = \binom{4}{2} (0.2)^3 (0.8)^2 = 0.03072.$$

3.7.4 为什么一般不具有无记忆性

Why It Is Usually Not Memoryless

当 $r > 1$ ，条件“尚未达到 r 次成功”并没有说明已经取得多少次成功，剩余所需成功数会受到过去信息影响。

例如 $r = 2, p = 1/2$ ， X 是第二次成功的总等待次数：

$$P(X > 2) = \frac{3}{4}, \quad P(X > 4) = \frac{5}{16}.$$

因此

$$P(X > 4 | X > 2) = \frac{5/16}{3/4} = \frac{5}{12} \neq \frac{3}{4} = P(X > 2).$$

几何分布无记忆，不能因此推广到所有负二项分布。

3.7.5 过度离散与泊松极限

Overdispersion and the Poisson Limit

Poisson-Gamma Mixture

若事件的发生率本身是随机的，计数就会比 Poisson 更分散。设

$$Y | \Lambda \sim \text{Poisson}(\Lambda),$$

其中 Λ 服从形状为 r 、比率为 β 的伽马分布。把 Poisson 的 PMF 与伽马密度相乘后对 λ 积分，被积函数的核恰好仍是伽马密度，因而可以直接积出：

$$P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y, \quad p = \frac{\beta}{1+\beta},$$

这正是失败数口径的负二项分布；伽马的均值 $r/\beta = r(1-p)/p$ 与 Y 的均值一致。

Overdispersion

混合之后

$$E(Y) = \frac{r(1-p)}{p}, \quad \text{Var}(Y) = \frac{r(1-p)}{p^2} = \frac{E(Y)}{p} > E(Y),$$

因为 $0 < p < 1$ 。Poisson 分布要求均值等于方差，而负二项分布允许方差大于均值；实际计数数据中若样本方差远大于样本均值，负二项分布往往更合适。

The Poisson Limit

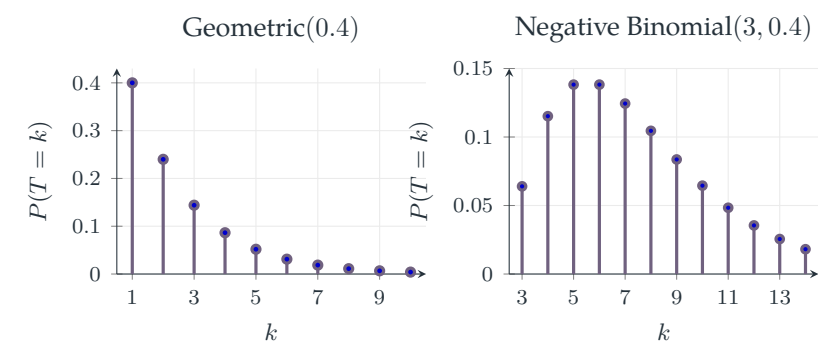
固定 $r(1-p) \rightarrow \lambda$ ，令 $r \rightarrow \infty$ 、 $p \rightarrow 1$ 。由失败数的 MGF

$$M_Y(t) = \left(\frac{p}{1 - qe^t} \right)^r, \quad q = 1 - p,$$

取对数并用 $\log(1-x) \approx -x$ ：

$$r \log \frac{p}{1 - qe^t} = r \log(1 - q) - r \log(1 - qe^t) \rightarrow \lambda(e^t - 1),$$

故 $M_Y(t) \rightarrow \exp\{\lambda(e^t - 1)\}$ ，即负二项分布收敛到 Poisson(λ)。这说明 Poisson 是“发生率几乎不随机”的极限：机会极多、单次概率极小、平均次数稳定时，混合效应消失，回到 Poisson。



图中 T 都计入最后一次成功；第 3 次成功不可能早于第 3 次试验。

3.8 从求和到积分

From Sums to Integrals

3.8.1 密度函数与概率

Density and Probability

Definition

(CB 2e §3.3)

若存在非负可积函数 f ，使随机变量 X 满足

$$P(X \in B) = \int_B f(x) \, dx$$

对一切可测集 B 成立，则称 X 为连续型随机变量， f 为其概率密度函数。
与离散分布最大的区别在于：离散分布的 PMF 直接给出点概率，而连续分布的密度**不是**点概率，只有积分才有概率含义。也就是说，

$f(x)$ 是概率密度，不是概率.

由此推出两条常用事实。

Zero Point Probability

取 $B = \{x_0\}$, 则 B 的长度为零, 故

$$P(X = x_0) = \int_{x_0}^{x_0} f(x) dx = 0.$$

所以连续型随机变量取到任一指定值的概率都是零。这解释了为什么问“ X 恰好等于 3”没有意义, 而问“ X 落在 3 附近”才有意义。

Endpoints Do Not Matter

因为单点概率为零,

$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b).$$

做题时不必纠结开区间与闭区间。

3.8.2 密度与分布函数

Density and CDF

Definition

(CB 2e §3.3)

连续型随机变量的分布函数为

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

于是密度与分布函数互为“积分与求导”的关系:

- 由密度得分布函数: $F(x) = \int_{-\infty}^x f(t) dt$;
- 由分布函数得密度: 在 F 可导的点处 $f(x) = F'(x)$ 。

这条互推关系是解决连续型问题的基本工具: 求“ X 落在某区间”的概率, 既可以积密度 (得 $F(b) - F(a)$), 也可以直接读分布函数的值。

Remark

判断一个函数是否为密度，只需检查两条：非负、积分为 1。若函数含待定常数，就先用归一化条件定出常数——这是最常考的一类小题。

3.8.3 用积分求矩

Moments as Integrals

对任意函数 g ,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx,$$

这与离散情形的 $\sum_x g(x)f(x)$ 完全平行，即连续版的无意识统计学家法则。特别地，

$$E(X) = \int xf(x) dx, \quad \text{Var}(X) = E(X^2) - [E(X)]^2,$$

$$M_X(t) = E(e^{tX}) = \int e^{tx}f(x) dx.$$

需要注意：期望存在的条件是

$$\int_{-\infty}^{\infty} |x|f(x) dx < \infty.$$

若该积分发散，即使把积分理解为对称截断的极限能得到 0，也不能说 $E(X) = 0$ ，而是 $E(X)$ 不存在。柯西分布正是这种情形（见第 6 节）。这是连续型与离散型都会遇到、但更容易被忽略的细节。

3.8.4 两个基础积分

Two Fundamental Integrals

后面每一类连续分布都归结为下面两个积分，先把它们推清楚。

The Gaussian Integral

记

$$I = \int_{-\infty}^{\infty} e^{-x^2/2} dx.$$

这个积分的原函数不能用初等函数表示，但可以通过“先平方、再用极坐标”算出来。把 I^2 写成二重积分，并让两个变量各自独立：

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy.$$

作极坐标代换 $x = r \cos \theta$ 、 $y = r \sin \theta$ ，则

$$x^2 + y^2 = r^2, \quad dx dy = r dr d\theta, \quad r \in [0, \infty), \theta \in [0, 2\pi).$$

于是

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta \\ &= 2\pi \left[-e^{-r^2/2} \right]_0^{\infty} = 2\pi. \end{aligned}$$

因此

$$I = \int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}.$$

由同一个代换还可得到一般形式：对 $a > 0$ ，

$$\int_{-\infty}^{\infty} e^{-ax^2/2} dx = \sqrt{\frac{2\pi}{a}},$$

只需令 $x = u/\sqrt{a}$ 。后面推导正态分布、卡方分布时都要用它。

The Gamma Function

定义

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \quad \alpha > 0.$$

用分部积分可得递推公式。取 $u = x^{\alpha-1}$ 、 $dv = e^{-x}dx$ ，则 $du = (\alpha-1)x^{\alpha-2}dx$ 、 $v = -e^{-x}$ ，

$$\begin{aligned}\Gamma(\alpha) &= [-x^{\alpha-1}e^{-x}]_0^\infty + (\alpha-1) \int_0^\infty x^{\alpha-2}e^{-x}dx \\ &= (\alpha-1)\Gamma(\alpha-1),\end{aligned}$$

边界项为零： $x \rightarrow \infty$ 时指数衰减快于幂增长； $x \rightarrow 0$ 时要求 $\alpha > 1$ 。等价地写成

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha).$$

再结合 $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$ ，得到整数情形

$$\Gamma(n) = (n-1)!.$$

即伽马函数是阶乘的连续推广，这正是它在连续型分布中反复出现的原因。

3.8.5 连续型求矩的四种技术

Four Techniques for Continuous Moments

技术	适用情形	关键动作
换元化为标准积分	密度含幂函数与指数	用 $\int_0^\infty x^{a-1}e^{-x}dx = \Gamma(a)$
配方与高斯积分	正态型被积函数	把指数配方，再用 $\int e^{-u^2/2}du = \sqrt{2\pi}$
对称性	密度关于某点对称	奇函数部分积分为零，直接得均值
MGF 求导	已有闭式 MGF	求 $M'_X(0)$ 与 $M''_X(0)$

这四种技术对应离散情形中的组合恒等式、指示变量分解、阶乘矩与递推公式。做题时先看密度的形状，再选技术：出现 $x^{\alpha-1}e^{-x/\beta}$ 就用伽马函数；出现 $e^{-(x-\mu)^2/(2\sigma^2)}$ 就配方；密度关于某点对称就先判均值。

3.9 均匀分布

Uniform Distribution

3.9.1 密度与分布函数

Density and CDF

Definition

(CB 2e §3.3)

若 X 的密度为

$$f(x) = \frac{1}{b-a}, \quad a < x < b,$$

则称 X 服从区间 (a, b) 上的均匀分布，记作 $X \sim \text{Uniform}(a, b)$ 。

均匀分布的含义是“落在任意等长子区间的概率相同”，因为概率只与区间长度成正比：

$$P(c < X < d) = \frac{d-c}{b-a}, \quad a \leq c < d \leq b.$$

分布函数按区间分三段：

$$F(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & x \geq b. \end{cases}$$

在 a, b 两点 F 不可导，但这不影响密度：改变有限个点的函数值不改变积分，因此密度的定义允许这种“个别点例外”。

3.9.2 用对称性求均值

Mean by Symmetry

密度在 (a, b) 上为常数，故分布的“重心”就是区间中点，于是

$$E(X) = \frac{a+b}{2}.$$

也可以直接积分核对：

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{a+b}{2}.$$

对称性的用法值得记住：若密度关于 $x = c$ 对称且期望存在，则 $E(X) = c$ 。理由是 把 $x - c$ 代入期望后被积函数为奇函数。这里 $\int |x|f(x)dx < \infty$ ，所以对称性可以直接使用（与柯西分布的区别见第 6 节）。

3.9.3 方差

Variance

先算二阶矩：

$$E(X^2) = \frac{1}{b-a} \int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3},$$

第二步用了立方差公式 $b^3 - a^3 = (b-a)(a^2 + ab + b^2)$ 。于是

$$\begin{aligned} \text{Var}(X) &= \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4(a^2 + ab + b^2) - 3(a^2 + 2ab + b^2)}{12} \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

方差只依赖区间长度，与位置无关：把区间整体平移，分散程度不变。

3.9.4 矩母函数与线性变换

MGF and Linear Transform

$$M_X(t) = \frac{1}{b-a} \int_a^b e^{tx} dx = \frac{e^{tb} - e^{ta}}{t(b-a)}, \quad t \neq 0.$$

在 $t = 0$ 处表达式为 $0/0$ ，但由

$$e^{tb} - e^{ta} = t(b-a) + \frac{t^2(b^2 - a^2)}{2} + \dots$$

两边除以 $t(b-a)$ 并令 $t \rightarrow 0$ ，得 $M_X(0) = 1$ ，与 $M_X(0) = 1$ 的一般性质一致。顺便还能读出 $M'_X(0) = (a+b)/2$ ，与前面求得的均值相符。

Linear Transforms Stay Uniform

设 $U \sim \text{Uniform}(0, 1)$, $a < b$, 令

$$X = a + (b - a)U.$$

用分布函数法求 X 的密度: 对 $a < x < b$,

$$F_X(x) = P(a + (b - a)U \leq x) = P\left(U \leq \frac{x - a}{b - a}\right) = \frac{x - a}{b - a},$$

求导得

$$f_X(x) = \frac{1}{b - a}, \quad a < x < b,$$

即 $X \sim \text{Uniform}(a, b)$ 。这说明均匀分布是位置与尺度族的成员: 位置参数为 a 、尺度参数为 $b - a$ (见第 7 节)。

Example

设 $X \sim \text{Uniform}(0, 1)$, 则 $E(X) = 1/2$ 、 $\text{Var}(X) = 1/12$ 。若 $Y = 3 + 4X$, 则 $Y \sim \text{Uniform}(3, 7)$, 从而

$$E(Y) = \frac{3 + 7}{2} = 5, \quad \text{Var}(Y) = \frac{(7 - 3)^2}{12} = \frac{4}{3}.$$

用变换规则核对: $E(3 + 4X) = 3 + 4 \cdot \frac{1}{2} = 5$, $\text{Var}(3 + 4X) = 16 \cdot \frac{1}{12} = \frac{4}{3}$, 两者一致。

3.10 伽马分布族

The Gamma Family

3.10.1 定义与归一化常数

Definition and Normalizing Constant

Definition

(CB 2e §3.3)

对参数 $\alpha > 0$ 、 $\beta > 0$ ，若密度为

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0,$$

则称 X 服从伽马分布，记作 $X \sim \text{Gamma}(\alpha, \beta)$ 。其中 α 是形状参数， β 是尺度参数。

Where the Normalizing Constant Comes From

分母不是凭空规定的，而是由伽马积分算出来的。作代换 $x = \beta u$ ($dx = \beta du$):

$$\begin{aligned} \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx &= \int_0^\infty (\beta u)^{\alpha-1} e^{-u} \beta du \\ &= \beta^\alpha \int_0^\infty u^{\alpha-1} e^{-u} du = \Gamma(\alpha) \beta^\alpha. \end{aligned}$$

两边同除 $\Gamma(\alpha)\beta^\alpha$ 即得密度积分为 1。这也说明 $\Gamma(\alpha)\beta^\alpha$ 正是归一化常数。

3.10.2 均值

Mean

关键技巧是把 x 吸收进幂次： $x \cdot x^{\alpha-1} = x^\alpha$ ，于是积分重新变成伽马积分：

$$\begin{aligned} E(X) &= \int_0^\infty x \cdot \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta} dx \\ &= \frac{\Gamma(\alpha+1)\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} = \alpha\beta. \end{aligned}$$

倒数第二步把积分识别为参数 $\alpha+1$ 的伽马积分，所以结果是 $\Gamma(\alpha+1)\beta^{\alpha+1}$ ；最后一步用 $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ 。

3.10.3 方差

Variance

同样把 x^2 吸收进幂次：

$$E(X^2) = \frac{\Gamma(\alpha+2)\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} = \alpha(\alpha+1)\beta^2,$$

其中 $\Gamma(\alpha+2) = (\alpha+1)\Gamma(\alpha+1) = \alpha(\alpha+1)\Gamma(\alpha)$ 。于是

$$\text{Var}(X) = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

两个参数的分工由此明确： α 决定形状（偏斜程度）， β 决定尺度；均值与方差都随 β 线性或平方增长。

3.10.4 矩母函数

MGF

把 e^{tx} 与 $e^{-x/\beta}$ 合并：

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} \exp\left\{-x\left(\frac{1}{\beta} - t\right)\right\} dx.$$

当 $t < 1/\beta$ 时括号内为正，令

$$\frac{1}{\beta'} = \frac{1}{\beta} - t \implies \beta' = \frac{\beta}{1 - \beta t} > 0,$$

则积分仍是伽马积分，值为 $\Gamma(\alpha)(\beta')^\alpha$ 。因此

$$M_X(t) = \frac{\Gamma(\alpha)(\beta')^\alpha}{\Gamma(\alpha)\beta^\alpha} = \left(\frac{1}{1 - \beta t}\right)^\alpha, \quad t < \frac{1}{\beta}.$$

收敛条件必须写出：当 $t \geq 1/\beta$ 时指数不再衰减，MGF 不存在（取值为 $+\infty$ ）。

Checking Moments with the MGF

记 $u = 1 - \beta t$ ，则 $M = u^{-\alpha}$ 、 $du/dt = -\beta$ ，

$$M'_X(t) = \alpha\beta u^{-\alpha-1}, \quad M''_X(t) = \alpha(\alpha+1)\beta^2 u^{-\alpha-2}.$$

在 $t = 0$ 处 $u = 1$ ，所以

$$M'_X(0) = \alpha\beta, \quad M''_X(0) = \alpha(\alpha+1)\beta^2,$$

$$\text{Var}(X) = \alpha(\alpha + 1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2,$$

与直接积分的结果一致。

3.10.5 指数分布

Exponential Distribution

取 $\alpha = 1$, 由 $\Gamma(1) = 1$ 得

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0, \quad X \sim \text{Exponential}(\beta).$$

它正是 $\text{Gamma}(1, \beta)$, 因此可直接读出

$$E(X) = \beta, \quad \text{Var}(X) = \beta^2, \quad M_X(t) = \frac{1}{1 - \beta t}, \quad t < \frac{1}{\beta}.$$

Direct Integration Check

用分部积分算 $E(X)$:

$$E(X) = \int_0^\infty \frac{x}{\beta} e^{-x/\beta} dx = \left[-x e^{-x/\beta} \right]_0^\infty + \int_0^\infty e^{-x/\beta} dx = \beta,$$

边界项为零。同理

$$E(X^2) = \int_0^\infty \frac{x^2}{\beta} e^{-x/\beta} dx = 2\beta^2,$$

故 $\text{Var}(X) = 2\beta^2 - \beta^2 = \beta^2$, 与伽马分布的公式一致。

CDF and Survival Function

直接积分:

$$F(x) = 1 - e^{-x/\beta}, \quad P(X > x) = e^{-x/\beta}, \quad x > 0.$$

Memorylessness

对任意 $s, t > 0$ ，按条件概率定义并用生存函数：

$$\begin{aligned} P(X > s + t \mid X > s) &= \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{e^{-(s+t)/\beta}}{e^{-s/\beta}} = e^{-t/\beta} = P(X > t). \end{aligned}$$

Remark

无记忆性的现实含义是：若元件寿命服从指数分布，则“已经用了 s 小时没坏”这一信息不改变剩余寿命的分布。这与 2026-10-05 笔记中的几何分布完全对应——几何分布是离散时间上的无记忆等待时间，指数分布是连续时间上的对应物。反过来也成立：指数分布是 $[0, \infty)$ 上唯一具有无记忆性的连续分布。

3.10.6 卡方分布

Chi-Squared Distribution

取 $\alpha = \nu/2$ 、 $\beta = 2$ ：

$$f(x) = \frac{1}{\Gamma(\nu/2)2^{\nu/2}} x^{\nu/2-1} e^{-x/2}, \quad x > 0,$$

记作 $X \sim \chi^2(\nu)$ ， ν 称为自由度。由伽马分布的公式立得

$$E(X) = \nu, \quad \text{Var}(X) = 2\nu, \quad M_X(t) = (1 - 2t)^{-\nu/2}, \quad t < \frac{1}{2}.$$

Why the Name Chi-Squared

设 $Z \sim N(0, 1)$ ，先算 Z^2 的 MGF。把标准正态密度代入并合并指数：

$$\begin{aligned} M_{Z^2}(t) &= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2(1-2t)/2} dz. \end{aligned}$$

当 $t < 1/2$ 时 $1 - 2t > 0$ ，用 1.4 节的一般高斯积分（取 $a = 1 - 2t$ ）：

$$M_{Z^2}(t) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{2\pi}{1-2t}} = (1-2t)^{-1/2}, \quad t < \frac{1}{2}.$$

这正是 $\text{Gamma}(1/2, 2) = \chi^2(1)$ 的 MGF。也就是说，一个标准正态变量的平方服从自由度为 1 的卡方分布。

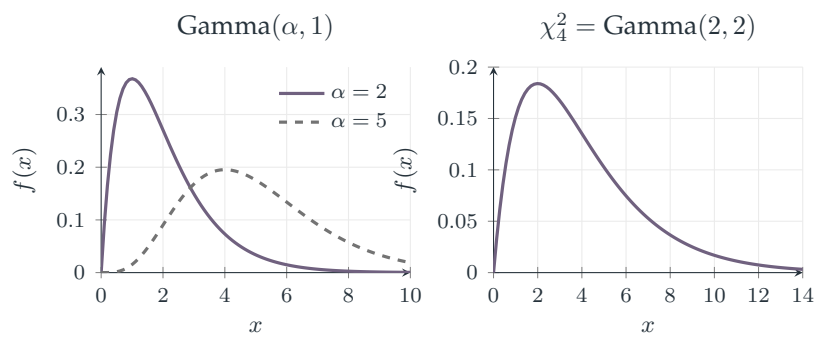
再设 $Z_1, \dots, Z_\nu$ 相互独立且都服从 $N(0, 1)$ 。由独立和的乘积法则，

$$M_{\sum_i Z_i^2}(t) = \prod_{i=1}^\nu (1 - 2t)^{-1/2} = (1 - 2t)^{-\nu/2},$$

由 MGF 的唯一性得

$$\sum_{i=1}^\nu Z_i^2 \sim \chi^2(\nu).$$

这就是卡方分布名称的由来：它是 ν 个独立标准正态平方之和。卡方分布是后面区间估计与假设检验的核心分布，而这段 MGF 计算说明它在正态模型下是自然出现的，不是外加的定义。



相同尺度下，形状参数越大，峰值越向右； χ^2_4 是形状为 2、尺度为 2 的 Gamma 分布。

3.11 正态分布

Normal Distribution

3.11.1 定义

Definition

Definition

(CB 2e §3.3)

若密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R},$$

则称 X 服从参数为 μ 与 σ^2 的正态分布, 记作 $X \sim N(\mu, \sigma^2)$ 。

由 1.4 节的高斯积分 (取 $a = 1/\sigma^2$) 可以验证它确实积分为 1:

$$\int_{-\infty}^{\infty} f(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2\pi\sigma^2} = 1.$$

当 $\mu = 0$ 、 $\sigma = 1$ 时称为标准正态分布 $N(0, 1)$, 其密度记作 φ 、分布函数记作 Φ :

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \int_{-\infty}^z \varphi(u) du.$$

3.11.2 标准化

Standardization

Definition

(CB 2e §3.3)

设 $X \sim N(\mu, \sigma^2)$, 令

$$Z = \frac{X - \mu}{\sigma},$$

称 Z 为 X 的标准化变量。

Why $Z \sim N(0, 1)$

用分布函数法。对任意 z ,

$$F_Z(z) = P\left(\frac{X - \mu}{\sigma} \leq z\right) = P(X \leq \mu + \sigma z) = F_X(\mu + \sigma z),$$

其中 F_X 是 X 的分布函数。两边对 z 求导 (复合函数法则, 内层导数为 σ):

$$f_Z(z) = \sigma f_X(\mu + \sigma z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

正是标准正态密度。因此

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

这一步把任意正态的计算化为查一张标准正态表，是正态分布最重要的操作。

Symmetry

标准正态密度是偶函数，因此对任意 z ,

$$\Phi(-z) = 1 - \Phi(z).$$

Proof

作代换 $u = -t$:

$$\Phi(-z) = \int_{-\infty}^{-z} \varphi(t) dt = \int_z^{\infty} \varphi(u) du = 1 - \int_{-\infty}^z \varphi(u) du = 1 - \Phi(z).$$

于是只需记住 $z > 0$ 时的 $\Phi(z)$ ，负值用上式换算。这也是查表时常用的技巧。

3.11.3 均值与方差

Mean and Variance

Mean via Symmetry

先确认期望存在: $\int |x|f(x)dx < \infty$ ，因为指数衰减快于幂增长。然后作代换 $u = x - \mu$:

$$\begin{aligned} E(X) - \mu &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu) e^{-(x-\mu)^2/(2\sigma^2)} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} u e^{-u^2/(2\sigma^2)} du = 0, \end{aligned}$$

最后一步中被积函数是奇函数、积分区间对称。因此

$$E(X) = \mu.$$

Remark

“期望存在”这一句不能省略。柯西分布的被积函数同样关于原点对称，但 $\int |x|f(x)dx$ 发散，因此不能由对称性断言其均值为零——它根本没有均值。

Variance via Integration by Parts

作代换 $u = (x - \mu)/\sigma$:

$$E[(X - \mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-u^2/2} du.$$

计算该积分时取 $dv = u e^{-u^2/2} du$ 、 $v = -e^{-u^2/2}$ ，分部积分得

$$\begin{aligned} \int_{-\infty}^{\infty} u^2 e^{-u^2/2} du &= \left[-u e^{-u^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-u^2/2} du \\ &= 0 + \sqrt{2\pi} = \sqrt{2\pi}, \end{aligned}$$

边界项为零。因此

$$\text{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = \sigma^2.$$

两个参数的含义由此明确： μ 是位置参数（决定中心）， σ 是尺度参数（决定分散程度）。

3.11.4 矩母函数

MGF

由定义并配方。关键是把指数部分写成完全平方：

$$tx - \frac{(x - \mu)^2}{2\sigma^2} = \mu t + \frac{\sigma^2 t^2}{2} - \frac{(x - \mu - \sigma^2 t)^2}{2\sigma^2}.$$

把它代回积分，并把与 x 无关的因子提出来：

$$M_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{\mu t + \sigma^2 t^2/2} \int_{-\infty}^{\infty} e^{-(x - \mu - \sigma^2 t)^2/(2\sigma^2)} dx.$$

括号内的积分正是正态密度（均值 $\mu + \sigma^2 t$ 、方差 σ^2 ）在全直线上的积分，等于 $\sqrt{2\pi}\sigma$ 。于是

$$M_X(t) = \exp \left\{ \mu t + \frac{\sigma^2 t^2}{2} \right\}, \quad t \in \mathbb{R}.$$

它对一切实数 t 有限, 这是连续型分布中条件最好的情形之一。

Checking Moments with the MGF

求导得

$$M'_X(t) = (\mu + \sigma^2 t) M_X(t),$$

$$M''_X(t) = [\sigma^2 + (\mu + \sigma^2 t)^2] M_X(t).$$

在 $t = 0$ 处 $M_X(0) = 1$, 故

$$M'_X(0) = \mu, \quad M''_X(0) = \sigma^2 + \mu^2, \quad \text{Var}(X) = \sigma^2.$$

3.11.5 线性组合的封闭性

Linear Combinations

Theorem

(CB 2e §3.3)

若 $X_1, \dots, X_n$ 相互独立, $X_i \sim N(\mu_i, \sigma_i^2)$, $a_1, \dots, a_n$ 为常数, 则

$$\sum_{i=1}^n a_i X_i \sim N \left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2 \right).$$

Proof

由 MGF 的乘积法则 (独立性给出 $E[\prod e^{ta_i X_i}] = \prod E[e^{ta_i X_i}]$),

$$\begin{aligned} M_{\sum a_i X_i}(t) &= \prod_{i=1}^n M_{X_i}(a_i t) \\ &= \prod_{i=1}^n \exp \left\{ a_i \mu_i t + \frac{a_i^2 \sigma_i^2 t^2}{2} \right\} \\ &= \exp \left\{ t \sum a_i \mu_i + \frac{t^2}{2} \sum a_i^2 \sigma_i^2 \right\}, \end{aligned}$$

这正是 $N(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2)$ 的 MGF，由唯一性即得结论。

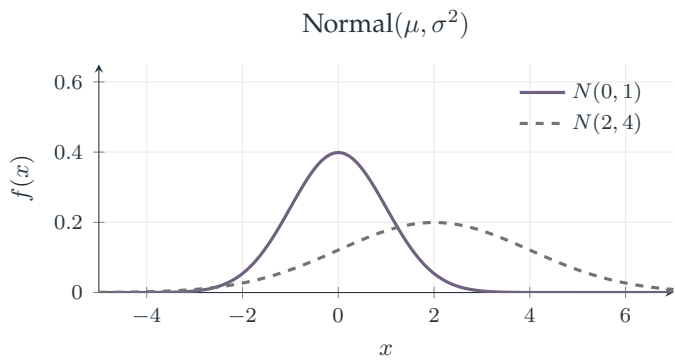
Corollary

- 独立正态变量之和仍为正态；
- 若 $X_1, \dots, X_n$ 独立同分布于 $N(\mu, \sigma^2)$ ，取 $a_i = 1/n$ 得

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

样本均值的方差随 n 减小，这是后面统计推断大量使用正态分布的根本原因；

- 正态性在“加常数、乘常数”下保持，这就是正态分布属于位置与尺度族的含义。



μ 移动中心； σ 变大时曲线更宽、更低，面积仍为 1。

3.12 贝塔分布

Beta Distribution

3.12.1 定义

Definition

Definition

(CB 2e §3.3)

对 $\alpha > 0, \beta > 0$, 若密度为

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1,$$

则称 $X \sim \text{Beta}(\alpha, \beta)$, 其中

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

称为贝塔函数。因为取值被限制在 $(0, 1)$ 内, 贝塔分布常用来描述比例、比率或“某个概率本身”的不确定性, 例如产品合格率、某地区患病率。

Remark

贝塔分布与二项分布关系密切: 若成功概率本身随机且服从贝塔分布, 则观测到的成功次数服从贝塔—二项分布。这也是它在贝叶斯统计中常作为“概率的先验分布”的原因。

3.12.2 与伽马函数的关系

Relation to the Gamma Function

Theorem

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

Proof

从右边出发, 把两个伽马积分写成二重积分:

$$\Gamma(\alpha)\Gamma(\beta) = \int_0^\infty \int_0^\infty u^{\alpha-1} v^{\beta-1} e^{-(u+v)} du dv.$$

作代换

$$u = st, \quad v = s(1-t), \quad s > 0, 0 < t < 1.$$

其雅可比行列式为

$$\left| \frac{\partial(u, v)}{\partial(s, t)} \right| = \begin{vmatrix} t & s \\ 1-t & -s \end{vmatrix} = -st - s(1-t) = -s,$$

取绝对值得 $du dv = s ds dt$ 。同时

$$u + v = s, \quad u^{\alpha-1} v^{\beta-1} = s^{\alpha+\beta-2} t^{\alpha-1} (1-t)^{\beta-1}.$$

于是二重积分化为两个一元积分的乘积：

$$\begin{aligned} \Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty s^{\alpha+\beta-1} e^{-s} ds \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt \\ &= \Gamma(\alpha + \beta) B(\alpha, \beta), \end{aligned}$$

两边除以 $\Gamma(\alpha + \beta)$ 即得结论。

有了这个关系，贝塔分布的矩就完全可以用伽马函数的递推公式计算，不必每次做积分。

3.12.3 均值与方差

Mean and Variance

Mean

$$\begin{aligned} E(X) &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^\alpha (1-x)^{\beta-1} dx = \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} \\ &= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}, \end{aligned}$$

最后一步用 $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ 与 $\Gamma(\alpha+\beta+1) = (\alpha+\beta)\Gamma(\alpha+\beta)$ 。

Second Moment and Variance

同理

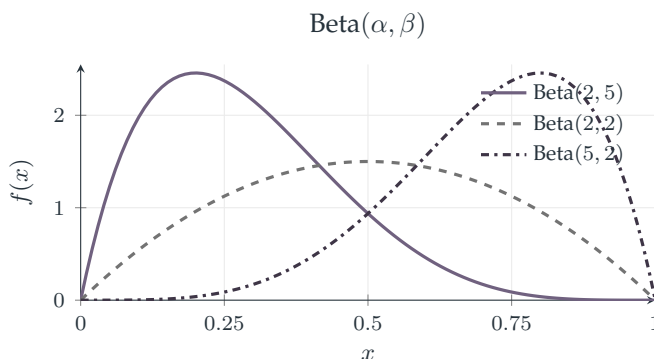
$$E(X^2) = \frac{B(\alpha+2, \beta)}{B(\alpha, \beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)},$$

于是

$$\begin{aligned}\text{Var}(X) &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\ &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.\end{aligned}$$

Remark

- Beta(1, 1) 的密度为常数 1，即 Uniform(0, 1)；
- 均值 $\frac{\alpha}{\alpha+\beta}$ 只依赖两参数之比；固定比值而同时增大 α, β ，方差变小，密度向均值集中；
- 贝塔分布没有常用的简洁 MGF 闭式，求矩一般走上面的积分路线。



交换 α 与 β 会使曲线关于 $x = 1/2$ 镜像；两参数相等时分布对称。

3.13 柯西分布与重尾

Cauchy Distribution and Heavy Tails

3.13.1 密度与分布函数

Density and CDF

Definition

(CB 2e §3.3)

密度为

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}$$

的分布称为柯西分布，记作 $\text{Cauchy}(0, 1)$ 。

它的分布函数可以直接积出来：

$$F(x) = \int_{-\infty}^x \frac{dt}{\pi(1+t^2)} = \frac{1}{2} + \frac{1}{\pi} \arctan x,$$

因为 $\arctan$ 的导数为 $1/(1+x^2)$ ，且由对称性 $F(0) = 1/2$ 。

3.13.2 期望不存在

No Expectation

Theorem

柯西分布的一阶矩不存在，即

$$E|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \infty.$$

Proof

由密度为偶函数，

$$E|X| = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx = \frac{1}{\pi} [\ln(1+x^2)]_0^{\infty} = \infty.$$

由于 $\int |x|f(x)dx$ 发散，期望 $E(X)$ 没有定义。

Remark

初学者常犯的错误是看到密度关于 0 对称就直接写 $E(X) = 0$ 。这里必须区分两个不同的量：

- 对称截断积分（主值）：

$$\lim_{R \rightarrow \infty} \int_{-R}^R \frac{x}{\pi(1+x^2)} dx = 0,$$

因为被积函数是奇函数，正负部分逐项抵消；

- 期望：定义为 $\int xf(x)dx$ ，只有在 $\int |x|f(x)dx < \infty$ 时才存在。

柯西分布的 $\int |x|f(x)dx$ 发散，所以它**没有期望**，而不是“期望等于零”。同理，它的方差与更高阶矩也都不存在。

柯西分布因此是三条重要结论的反例：

1. 对称不等于期望为零，还要求期望存在；
2. 大数定律与中心极限定理都需要矩条件，柯西分布是它们失效的经典例子；
3. 对柯西分布用样本均值估计中心是无效的：样本均值与单个观测同分布，不会随样本量增大而集中。

3.13.3 线性变换下的封闭性

Closure under Linear Transformations

设 $X \sim \text{Cauchy}(0, 1)$, $\mu \in \mathbb{R}$ 、 $\sigma > 0$, 令 $Y = \mu + \sigma X$ 。用分布函数法：

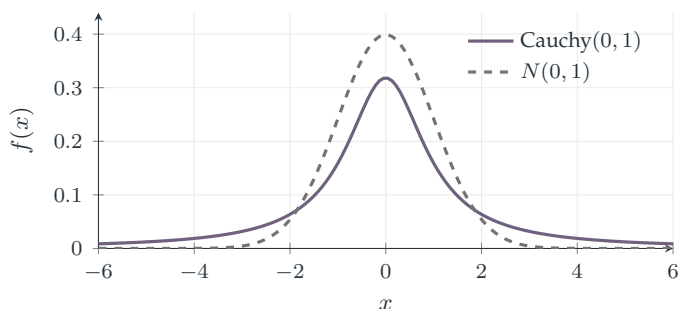
$$F_Y(y) = P\left(X \leq \frac{y - \mu}{\sigma}\right) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{y - \mu}{\sigma},$$

求导得

$$f_Y(y) = \frac{\sigma}{\pi [\sigma^2 + (y - \mu)^2]},$$

这仍是柯西型密度。可见柯西分布构成一个位置与尺度族，而且这个性质与矩是否存在无关——推导只用了分布函数，没有用到期望。

中心相同、尾部不同



柯西密度在中心较低、两侧下降较慢；图形对称仍不足以保证均值存在。

3.14 位置与尺度族

Location-Scale Families

3.14.1 定义与密度公式

Definition and Density Formula

Definition

(CB 2e §3.3)

设 f 是某个密度的标准形式 (通常中心为 0、尺度为 1), Z 的密度为 f 。若对 $\mu \in \mathbb{R}$ 、 $\sigma > 0$, 随机变量 $X = \mu + \sigma Z$ 的密度可写成

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right),$$

则称这族分布构成位置与尺度族: μ 是位置参数, σ 是尺度参数。

Deriving the Formula

设 Z 的分布函数为 F , 则

$$F_X(x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = F\left(\frac{x - \mu}{\sigma}\right).$$

两边对 x 求导, 内层导数为 $1/\sigma$:

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right).$$

Remark

这条公式是把“标准化”反过来用的结果: 标准化把 X 变成 $(X - \mu)/\sigma$, 这里把标准成员变成一般成员, 两者互为逆运算。因此只要记住标准形式的密度, 就能写出整族的密度; 公式中的因子 $1/\sigma$ 来自换元时 dx 与 dz 的比例关系。

3.14.2 矩的变换规则

Moment Rules

对 $X = \mu + \sigma Z$, 由期望的线性性与方差的性质,

$$E(X) = \mu + \sigma E(Z), \quad \text{Var}(X) = \sigma^2 \text{Var}(Z).$$

若 Z 的均值为 0、方差为 1（标准成员），则 $E(X) = \mu$ 、 $\text{Var}(X) = \sigma^2$ 。这就是“位置与尺度族”这个名称的统计含义。

3.14.3 族内成员

Members of the Family

分布族	标准成员	位置参数	尺度参数
正态	$N(0, 1)$	μ	σ
均匀	$\text{Uniform}(0, 1)$	a	$b - a$
柯西	$\text{Cauchy}(0, 1)$	位置	尺度
伽马（固定形状 α ）	—	—	β

Remark

伽马分布的 α 是形状参数而不是位置参数，所以它只在“固定 α ”的意义下属于尺度族；贝塔分布的取值区间固定在 $(0, 1)$ ，也不构成位置与尺度族。判断一个族是否属于位置与尺度族，看的是“平移与伸缩能否把标准成员变成族内任意成员”。

3.15 由变量变换得到的分布

Distributions from Transformations

3.15.1 韦布尔分布

Weibull Distribution

指数分布描述恒定失效率下的等待时间。对指数变量作幂变换，可以使失效率随已经等待的时间变化。令 $E \sim \text{Exp}(1)$ ，取

$$T = \eta E^{1/\gamma}, \quad \gamma > 0, \quad \eta > 0.$$

由于变换严格递增，当 $t \geq 0$ 时，

$$\begin{aligned} S_T(t) &= \mathbb{P}(T > t) = \mathbb{P}(E > (t/\eta)^\gamma) = e^{-(t/\eta)^\gamma}, \\ F_T(t) &= 1 - e^{-(t/\eta)^\gamma}. \end{aligned}$$

$t < 0$ 时 $F_T(t) = 0$ 。对 CDF 求导，得到

$$T \sim \text{Weibull}(\gamma, \eta), \quad f_T(t) = \frac{\gamma}{\eta} \left(\frac{t}{\eta} \right)^{\gamma-1} e^{-(t/\eta)^\gamma}, \quad t > 0.$$

支持集外密度为零。 γ 是形状参数 (shape parameter), η 是尺度参数 (scale parameter), 与 T 具有相同单位。若课件从 $X \sim \text{Exp}(\beta)$ 出发定义 $Y = X^{1/\gamma}$, 则 $X = \beta E$, 从而 $Y = \beta^{1/\gamma} E^{1/\gamma}$; 对应的 Weibull 尺度是 $\eta = \beta^{1/\gamma}$, 不能直接写成 β 。

矩由指数变量的积分得到

对实数 $k > -\gamma$,

$$\begin{aligned} \mathbb{E}T^k &= \eta^k \mathbb{E}E^{k/\gamma} = \eta^k \int_0^\infty u^{k/\gamma} e^{-u} du \\ &= \eta^k \Gamma\left(1 + \frac{k}{\gamma}\right). \end{aligned}$$

条件 $k > -\gamma$ 保证零点附近的积分有限; 正整数阶矩均存在。因此

$$\mathbb{E}T = \eta \Gamma\left(1 + \frac{1}{\gamma}\right), \quad \text{Var}(T) = \eta^2 \left\{ \Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right) \right\}.$$

失效率 (hazard function) 表示已经存活至 t 的条件下, 每单位时间即时失效的速率, 定义为 $h(t) = f_T(t)/S_T(t)$ 。代入得到

$$h(t) = \frac{\gamma}{\eta} (t/\eta)^{\gamma-1}.$$

$\gamma = 1$ 时为尺度 η 的指数分布, 失效率恒定; $\gamma > 1$ 时失效率递增, $0 < \gamma < 1$ 时递减。这里的失效率是条件速率, 不能直接当作一个区间上的概率。

所有正整数阶矩存在, 并不保证 MGF 在零点附近存在。观察尾部积分中的指数 $st - (t/\eta)^\gamma$ 可知: $\gamma > 1$ 时 MGF 对所有实数 s 有限; $\gamma = 1$ 时要求 $s < 1/\eta$; $0 < \gamma < 1$ 时任何 $s > 0$ 都发散, 因而没有包含零点开邻域的 MGF。

3.15.2 对数正态分布

Lognormal Distribution

令 $Z \sim N(\mu, \sigma^2)$, 其中 $\sigma > 0$, 并令 $X = e^Z$ 。则 $X > 0$, 而 $\log X = Z$ 服从正态分布。对 $x > 0$,

$$F_X(x) = \mathbb{P}(e^Z \leq x) = \mathbb{P}(Z \leq \log x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right).$$

$x \leq 0$ 时 CDF 为零。链式法则给出密度

$$X \sim \text{Lognormal}(\mu, \sigma^2), \quad f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \quad x > 0.$$

其中 $1/x$ 来自 $d(\log x)/dx$; 它是变量替换的导数因子。参数 μ 和 σ^2 分别描述 $\log X$ 的均值和方差, 并非 X 本身的均值和方差。

借助正态变量的 MGF 求矩

对任意实数 k , 由 $X^k = e^{kZ}$ 得

$$\mathbb{E}X^k = M_Z(k) = \exp\left(k\mu + \frac{k^2\sigma^2}{2}\right).$$

因此 $\mathbb{E}X = e^{\mu+\sigma^2/2}$ 、 $\mathbb{E}X^2 = e^{2\mu+2\sigma^2}$, 相减得到

$$\text{Var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

这里调用的是正态变量 Z 的 MGF, 不是对数正态变量 X 的 MGF。事实上, 任意 $s > 0$ 时

$$\mathbb{E}e^{sX} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[se^z - \frac{(z - \mu)^2}{2\sigma^2}\right] dz = \infty.$$

因为 se^z 在 $z \rightarrow \infty$ 时超过任意二次多项式, 右尾积分发散。由此可见, 全部整数阶矩存在仍不保证 MGF 在零点附近有限。

3.15.3 拉普拉斯分布

Laplace Distribution

拉普拉斯分布也称双指数分布 (double exponential distribution)。取位置参数 $\mu \in \mathbb{R}$ 、尺度参数 $b > 0$, 定义

$$X \sim \text{Laplace}(\mu, b), \quad f_X(x) = \frac{1}{2b} e^{-|x-\mu|/b}, \quad x \in \mathbb{R}.$$

密度以 μ 为中心对称，两侧分别按指数速度衰减；两侧积分各为 $1/2$ ，所以总积分为 1 。CDF 按中心两侧积分得到

$$F_X(x) = \begin{cases} \frac{1}{2}e^{(x-\mu)/b}, & x \leq \mu, \\ 1 - \frac{1}{2}e^{-(x-\mu)/b}, & x > \mu. \end{cases}$$

绝对值变换必须汇总两条原像

令 $Y = |X - \mu|$ 。每个 $y > 0$ 对应 $x = \mu + y$ 与 $x = \mu - y$ ，两条逆变换的导数绝对值均为 1 ，故

$$\begin{aligned} f_Y(y) &= f_X(\mu + y) + f_X(\mu - y) \\ &= \frac{1}{2b}e^{-y/b} + \frac{1}{2b}e^{-y/b} = \frac{1}{b}e^{-y/b}, \quad y > 0. \end{aligned}$$

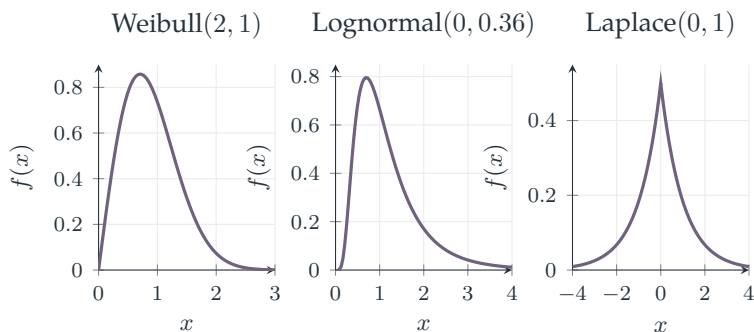
所以 $Y \sim \text{Exp}(b)$ ；只取正半轴的一条原像会漏掉一半概率。由对称性且绝对一阶矩有限， $\mathbb{E}X = \mu$ ；由 $Y^2 = (X - \mu)^2$ ，

$$\text{Var}(X) = \mathbb{E}(X - \mu)^2 = \mathbb{E}Y^2 = 2b^2.$$

若将 $W = X - \mu$ 的 MGF 积分在零点处分为两段，则在 $|s| < 1/b$ 时

$$\begin{aligned} M_W(s) &= \frac{1}{2b} \int_{-\infty}^0 e^{(s+1/b)w} dw + \frac{1}{2b} \int_0^{\infty} e^{(s-1/b)w} dw \\ &= \frac{1}{2(1+bs)} + \frac{1}{2(1-bs)} = \frac{1}{1-b^2s^2}. \end{aligned}$$

故 $M_X(s) = e^{\mu s}/(1 - b^2s^2)$ ，存在区间为 $|s| < 1/b$ 。左右两段同时收敛才构成有效的 MGF。



三种变换给出不同的形态：单峰且右偏、长右尾，以及以中心为尖点的双侧指数。

3.16 分布族之间的构造关系

Connections between Distribution Families

从计数到等待：先固定什么，再决定随机变量

n 个独立 $\text{Bernoulli}(p)$ 指示变量相加得到 $\text{Binomial}(n, p)$ ；当 $n \rightarrow \infty$ 、 $p_n = \lambda/n$ ，固定 k 的概率趋于 $e^{-\lambda} \lambda^k / k!$ ，得到 Poisson 极限。这里固定的是平均计数 $np_n = \lambda$ ，不是让 p 固定不变而单独增大 n 。

若把时间划分为宽度 Δ 的小格，每格成功概率取 $p_\Delta = \rho\Delta$ ，令 T_Δ 为首次成功所在的格数，则实际等待时间 $W_\Delta = \Delta T_\Delta$ 满足

$$\mathbb{P}(W_\Delta > t) = (1 - \rho\Delta)^{\lfloor t/\Delta \rfloor} \rightarrow e^{-\rho t}, \quad \Delta \rightarrow 0.$$

因此缩放后的几何等待趋于尺度 $1/\rho$ 的指数等待。参数和时间必须同时缩放；离散的试验次数不能直接当作连续时间。

在速率为 ρ 的同一 Poisson 过程中，令 $N(t)$ 为截至 t 的计数， T_m 为第 m 次事件到达时间， m 为正整数。则

$$\{T_m \leq t\} = \{N(t) \geq m\}, \quad \mathbb{P}(T_m \leq t) = 1 - e^{-\rho t} \sum_{j=0}^{m-1} \frac{(\rho t)^j}{j!}.$$

对 t 求导后，中间项相消，留下

$$f_{T_m}(t) = \frac{\rho^m t^{m-1} e^{-\rho t}}{(m-1)!}, \quad t > 0.$$

这正是 $\text{Gamma}(m, 1/\rho)$ 的密度。固定时间数到达次数用 Poisson；固定目标次数问到达时间用 Gamma；两种提问可以借事件等价关系转换。

从平方、比例和指数化建立连续分布联系

独立标准正态的平方和为 $\chi_\nu^2 = \text{Gamma}(\nu/2, 2)$ 。若 $G_1 \sim \text{Gamma}(\alpha, \beta)$ 、 $G_2 \sim \text{Gamma}(\delta, \beta)$ 相互独立且具有相同尺度，则 $G_1/(G_1 + G_2) \sim \text{Beta}(\alpha, \delta)$ ；相同尺度条件不能遗漏。正态变量指数化得到 Lognormal，指数变量作正幂变换得到 Weibull，Laplace 关于中心取绝对值得到 Exponential。每个结论都指定了原随机变量、参数条件及变换，不能只记分布名称之间的箭头。

3.17 例题

Worked Examples

3.17.1 例：由归一化条件定常数并求矩

设

$$f(x) = \begin{cases} c(1 - x^2), & |x| < 1, \\ 0, & |x| \geq 1. \end{cases}$$

(1) 求 c 。由归一化条件

$$1 = \int_{-1}^1 c(1 - x^2)dx = c \left[x - \frac{x^3}{3} \right]_{-1}^1 = c \cdot \frac{4}{3},$$

故 $c = \frac{3}{4}$ 。

(2) 求 $E(X)$ 。密度关于 0 对称，且在 $[-1, 1]$ 外为零，故 $\int |x|f(x)dx < \infty$ ，所以

$$E(X) = 0.$$

(3) 求 $\text{Var}(X)$ 。

$$\begin{aligned} E(X^2) &= \frac{3}{4} \int_{-1}^1 x^2(1-x^2)dx = \frac{3}{4} \left[\frac{x^3}{3} - \frac{x^5}{5} \right]_{-1}^1 \\ &= \frac{3}{4} \cdot 2 \left(\frac{1}{3} - \frac{1}{5} \right) = \frac{3}{4} \cdot \frac{4}{15} = \frac{1}{5}. \end{aligned}$$

所以 $\text{Var}(X) = 1/5$ 。

3.17.2 例：指数分布的寿命问题

设某设备寿命 $X \sim \text{Exponential}(\beta = 500)$ (单位：小时)。

(1) 平均寿命与标准差。

$$E(X) = 500, \quad \text{Var}(X) = 500^2, \quad \sigma_X = 500.$$

(2) 寿命超过 800 小时的概率。

$$P(X > 800) = e^{-800/500} = e^{-1.6} \approx 0.2019.$$

(3) 已知已用 300 小时仍正常，再用超过 800 小时的概率。由无记忆性，

$$P(X > 1100 \mid X > 300) = P(X > 800) \approx 0.2019,$$

不必重新计算生存函数。这也说明：如果实际数据表现出“越旧越容易坏”的趋势，指数分布就不再合适，需要换成失效率递增的模型。

3.17.3 例：正态分布的概率计算

设 $X \sim N(100, 15^2)$ ，例如智商分数的常见设定。

(1) $P(85 < X < 130)$ 。标准化后

$$\begin{aligned} P(85 < X < 130) &= \Phi\left(\frac{130-100}{15}\right) - \Phi\left(\frac{85-100}{15}\right) \\ &= \Phi(2) - \Phi(-1) = \Phi(2) + \Phi(1) - 1 \\ &\approx 0.9772 + 0.8413 - 1 = 0.8185. \end{aligned}$$

(2) $P(X > 120)$ 。

$$P(X > 120) = 1 - \Phi\left(\frac{120 - 100}{15}\right) = 1 - \Phi(1.333) \approx 0.0912.$$

(3) 为什么 $P(X = 120) = 0$ 。连续型分布的单点概率为零，所以求概率必须针对区间。这一条与离散分布形成对照：离散型的 $P(X = 120)$ 可能为正。

3.17.4 例：伽马分布的概率

设 $X \sim \text{Gamma}(2, 3)$ ，由 $\Gamma(2) = 1$ 得

$$f(x) = \frac{x e^{-x/3}}{9}, \quad x > 0.$$

求 $P(X > 6)$ 。先用分部积分算对应积分（取 $u = x$ 、 $dv = e^{-x/3}dx$ 、 $v = -3e^{-x/3}$ ）：

$$\begin{aligned} \int_6^\infty x e^{-x/3} dx &= \left[-3x e^{-x/3} \right]_6^\infty + 3 \int_6^\infty e^{-x/3} dx \\ &= 18e^{-2} + 9e^{-2} = 27e^{-2}. \end{aligned}$$

因此

$$P(X > 6) = \frac{27e^{-2}}{9} = 3e^{-2} \approx 0.4060.$$

均值 $E(X) = \alpha\beta = 6$ ，但 $P(X > 6) \approx 0.406 < 0.5$ ：伽马分布是右偏的，均值被右侧长尾拉大，所以超过均值的概率小于一半。这与正态分布的对称情形不同。

3.17.5 例：贝塔分布建模比例

设某生产线的合格率 $\theta \sim \text{Beta}(2, 3)$ 。

(1) 均值与方差。由 5.3 节的公式，

$$E(\theta) = \frac{2}{2+3} = 0.4,$$

$$\text{Var}(\theta) = \frac{2 \cdot 3}{(2+3)^2(2+3+1)} = \frac{6}{25 \cdot 6} = 0.04.$$

(2) 合格率超过 0.5 的概率。先算归一化常数：

$$B(2, 3) = \frac{\Gamma(2)\Gamma(3)}{\Gamma(5)} = \frac{1 \cdot 2}{24} = \frac{1}{12},$$

故 $f(\theta) = 12\theta(1 - \theta)^2$ 。作代换 $u = 1 - \theta$ ：

$$\begin{aligned} P(\theta > 0.5) &= 12 \int_{0.5}^1 \theta(1 - \theta)^2 d\theta \\ &= 12 \int_0^{0.5} (1 - u)u^2 du = 12 \left[\frac{u^3}{3} - \frac{u^4}{4} \right]_0^{0.5} \\ &= 12 \left(\frac{1}{24} - \frac{1}{64} \right) = \frac{5}{16} = 0.3125. \end{aligned}$$

均值是 0.4，但合格率超过 0.5 的概率只有 0.3125，两者并不矛盾：分布右偏，长尾在右侧把均值拉高，而中位数偏左。

3.17.6 例：柯西分布与样本均值

设 $X_1, X_2, \dots$ 独立同分布于 $\text{Cauchy}(0, 1)$ 。反复使用 6.3 节的线性变换封闭性可知，样本均值

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

仍服从 $\text{Cauchy}(0, 1)$ 。因此无论 n 多大， $\bar{X}_n$ 都不会向 0 集中，用它估计分布中心是无效的。这提醒我们：平均值只有在矩条件成立时才是可靠的中心估计。

第4章 联合分布与条件矩

Joint Distributions and Conditional Moments

依据: CB 2e Ch4; 本学期 Chapter 4 资料 pp.1–31

4.1 随机向量与联合分布函数

Random Vectors and Joint CDFs

4.1.1 随机向量

Random Vectors

Definition

(PS-SL04P01 p.4)

随机向量 (random vector) 是从样本空间到 $\mathbb{R}^n$ 的映射 $\mathbf{X} = (X_1, \dots, X_n)^\top$, 其中每个坐标 X_j 都是随机变量。

例如, 同一名学生的数学成绩 X 与统计成绩 Y 构成 (X, Y) 。仅知道两门成绩各自的分布, 不能判断高数学分是否往往伴随高统计分; 还需两者的联合分布 (joint distribution)。

4.1.2 联合 CDF 与边缘分布

Joint CDF and Marginals

Definition

(PS-SL04P01 p.5)

二元随机向量 (X, Y) 的联合分布函数 (joint cumulative distribution function) 为

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \quad (x, y) \in \mathbb{R}^2.$$

两个坐标各自的分布称为边缘分布 (marginal distributions)。令另一个坐标的上界趋于无穷, 便可从联合 CDF 取出边缘 CDF:

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y), \quad F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y).$$

由联合分布可以恢复边缘分布；反过来不行。取 $U \sim U(0, 1)$, (U, U) 与 $(U, 1-U)$ 的两个边缘均为 $U(0, 1)$ ，但前一对完全同向变动，后一对之和恒为 1。(PS-SL04P01 p.16)

4.1.3 矩形区域的概率

Rectangle Probabilities

设 $a < b, c < d$ 。先画出矩形 $(a, b] \times (c, d]$ ，再对左侧与下侧作容斥：

$$P(a < X \leq b, c < Y \leq d) = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c).$$

式中 $F(b, d)$ 覆盖整个左下象限；减去 $X \leq a$ 和 $Y \leq c$ 的两块后，左下角被重复减去，须加回一次。此式也适用于有离散概率质量的变量；不需假定连续。(PS-SL04P01 p.6)

Example

$$P(1 < X \leq 2, -1 < Y \leq 3) = F(2, 3) - F(1, 3) - F(2, -1) + F(1, -1).$$

把四个坐标先写在矩形四角，可减少符号错位。

4.2 离散型联合分布

Discrete Joint Distributions

4.2.1 联合 PMF 与支持集

Joint PMF and Support

Definition

(PS-SL04P01 p.7)

对离散型 (X, Y) ，联合概率质量函数 (joint PMF) 为

$$\begin{aligned} p_{X,Y}(x, y) &= P(X = x, Y = y), \\ p_{X,Y}(x, y) &\geq 0, \\ \sum_x \sum_y p_{X,Y}(x, y) &= 1. \end{aligned}$$

支持集是所有 $p_{X,Y}(x,y) > 0$ 的坐标对。若 $A \subseteq \mathbb{R}^2$, 则 $P((X,Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x,y)$ 。计算前先列出满足事件的格子, 而不是把两个边缘概率相乘。

4.2.2 对另一坐标求和

Marginal PMFs

Theorem

(PS-SL04P01 p.8)

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

固定 $X = x$ 后, 不同 $Y = y$ 的事件互斥且并成 $\{X = x\}$, 所以沿另一坐标求和。表格中的“行和”“列和”就是两个边缘 PMF。

Example

下表每格是一个坐标对的概率; 最右栏、最下行是边缘概率。(PS-SL04P01 pp.9–11)

$X \backslash Y$	0	1	$p_X(x)$
0	0.20	0.30	0.50
1	0.10	0.40	0.50
$p_Y(y)$	0.30	0.70	1

$P(X + Y \geq 1) = 1 - P(X = 0, Y = 0) = 0.80$ 。 $E(XY) = 0.40$, 因为仅 $(1, 1)$ 对乘积有贡献; 而 $EX = 0.50$ 、 $EY = 0.70$, 故 $E(XY) \neq EX EY$ 。

4.2.3 联合函数的期望

Functions of a Random Vector

若 $E|g(X,Y)| < \infty$, 则

$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p_{X,Y}(x,y).$$

取 $g(x, y) = x, y, xy$, 依次得到 $EX, EY, E(XY)$ 。均值只用各自边缘即可求; 混合矩 $E(XY)$ 一般需要联合分布。期望的线性性 $E(aX + bY) = aEX + bEY$ 不要求独立, 乘积拆为 $E(XY) = EX EY$ 则不能随意使用。(PS-SL04P01 pp.10–11,19)

4.3 连续型联合分布

Continuous Joint Distributions

4.3.1 联合密度

Joint Density

Definition

(PS-SL04P01 p.12)

若存在非负函数 $f_{X,Y}$, 使 $\iint_{\mathbb{R}^2} f_{X,Y}(x, y) dx dy = 1$, 并且对可测区域 A 有

$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dx dy,$$

则 (X, Y) 有联合密度 (joint PDF)。特别地,

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) dv du.$$

在相应混合偏导存在之处, $f_{X,Y} = \partial^2 F_{X,Y} / (\partial x \partial y)$ 。单点 (x, y) 的概率为 0; 密度值不是单点概率。两个边缘各自有密度, 不保证联合密度存在: (U, U) 的全部概率都落在直线 $x = y$ 上, 不能用平面上的普通密度描述。

4.3.2 边缘密度与积分限

Marginal Densities and Integration Limits

Theorem

(PS-SL04P01 p.13)

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx.$$

先在 xy 平面标出联合支持集，再固定一个坐标，沿另一个坐标的可行区间积分。非矩形支持集的积分界会随固定坐标改变。

Example

取 $f_{X,Y}(x,y) = 2$ ，定义域 $0 < x < y < 1$ ，其余为 0。区域是单位正方形对角线 $y = x$ 上方的三角形，面积 $1/2$ ；常数密度 2 使总概率为 1。固定 x 时， y 从 x 到 1；固定 y 时， x 从 0 到 y 。因此

$$\begin{aligned} f_X(x) &= \int_x^1 2 dy = 2(1-x), \quad 0 < x < 1, \\ f_Y(y) &= \int_0^y 2 dx = 2y, \quad 0 < y < 1. \end{aligned}$$

联合密度虽在三角形内为常数，边缘密度却不是均匀分布。（PS-SL04P01 p.14）

4.3.3 区域概率与期望

Probabilities and Expectations over Regions

沿用三角形密度。事件 $X < 1/2$ 对应 $0 < x < 1/2$ 、 $x < y < 1$ ，故

$$P(X < 1/2) = \int_0^{1/2} \int_x^1 2 dy dx = \frac{3}{4}.$$

事件 $X + Y < 1$ 再给出上界 $y < 1 - x$ 。要求 $x < 1 - x$ ，即 $0 < x < 1/2$ ，于是

$$P(X + Y < 1) = \int_0^{1/2} \int_x^{1-x} 2 dy dx = \frac{1}{2}.$$

若 $E|g(X,Y)| < \infty$ ，连续型联合期望为

$$E[g(X,Y)] = \iint g(x,y) f_{X,Y}(x,y) dx dy.$$

例如

$$E(XY) = \int_0^1 \int_0^y 2xy dx dy = \int_0^1 y^3 dy = \frac{1}{4}.$$

三个积分的上、下限都来自同一支持集 $0 < x < y < 1$ ，但事件额外约束改变了可积区域。(PS-SL04P01 pp.15,17-18)

4.4 条件分布

Conditional Distributions

4.4.1 条件质量与条件密度

Conditional PMF and PDF

Definition

(PS-SL04P01 p.22)

若 (X, Y) 离散且 $p_Y(y) > 0$ ，则

$$p_{X|Y}(x | y) = P(X = x | Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

若 (X, Y) 有联合密度，且 $f_Y(y) > 0$ ，则在适用的 y 上定义

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

对固定的 y ，条件 PMF 沿 x 求和为 1，条件 PDF 沿 x 积分为 1。连续型事件 $\{Y = y\}$ 的概率通常是 0；条件密度不是把 $P(X = x, Y = y)$ 除以 $P(Y = y)$ ，而是上式中联合密度与边缘密度的比。

同一等式可倒过来用： $p_{X,Y} = p_{X|Y}p_Y$ ，或 $f_{X,Y} = f_{X|Y}f_Y$ 。若已知条件分布与一个边缘分布，便能重建联合分布。(PS-SL04P01 p.23)

4.4.2 离散条件分布例

A Discrete Conditioning Example

在上面的 2×2 表中， $P(Y = 1) = 0.70$ 。条件于 $Y = 1$ 后，仅保留 $Y = 1$ 一列并除以列和：

$$\begin{aligned} P(X = 0 | Y = 1) &= \frac{0.30}{0.70} = \frac{3}{7}, \\ P(X = 1 | Y = 1) &= \frac{0.40}{0.70} = \frac{4}{7}. \end{aligned}$$

$3/7 + 4/7 = 1$ 。若误用行和 $P(X = 0) = 0.50$ 作分母，所得不再是“已知 $Y = 1$ ”的条件分布。(PS-SL04P01 p.24)

4.4.3 连续条件分布例

A Continuous Conditioning Example

对三角形密度， $f_Y(y) = 2y$ ($0 < y < 1$)。固定 $Y = y$ 后， X 只能落在 $(0, y)$ ，故

$$f_{X|Y}(x | y) = \frac{2}{2y} = \frac{1}{y}, \quad 0 < x < y.$$

换言之， $X | (Y = y) \sim U(0, y)$ 。

检查 $\int_0^y (1/y) dx = 1$ 。条件支持集随 y 改变；若将积分上限错写为 1，密度便不能归一化。(PS-SL04P01 p.25)

4.5 条件矩

Conditional Moments

4.5.1 条件期望

Conditional Expectation

Definition

(PS-SL04P01 pp.26–28)

给定 $X = x$ ，对 Y 的条件分布求普通期望，得到 $m(x) = E[g(Y) | X = x]$ ：

$$m(x) = \begin{cases} \sum g(y)p_{Y|X}(y | x), & \text{离散型,} \\ \int_{-\infty}^y g(y)f_{Y|X}(y | x) dy, & \text{连续型.} \end{cases}$$

把随机的 X 代入函数 m ，得到随机变量 $E[g(Y) | X] = m(X)$ 。给定 $X = x$ 时，它是一个数；在尚未观察 X 前，它随 X 变化。取 $g(y) = y$ 得 $E(Y | X = x)$ 。条件期望对可积变量仍满足线性性。(PS-SL04P01 p.27)

Example

在离散表中, $P(X = 0) = P(X = 1) = 1/2$ 。因 $Y \in \{0, 1\}$,

$$E(Y | X = 0) = P(Y = 1 | X = 0) = \frac{3}{5},$$

$$E(Y | X = 1) = P(Y = 1 | X = 1) = \frac{4}{5}.$$

因此 $E(Y | X)$ 取 $3/5$ 或 $4/5$, 取值由 X 决定。按 X 的概率再平均, $E[E(Y | X)] = (1/2)(3/5) + (1/2)(4/5) = 0.70 = EY$ 。此处只是用数值检验; 一般的迭代期望定理将在后续内容系统证明。(PS-SL04P01 pp.29–30)

4.5.2 条件方差

Conditional Variance

Definition

(PS-SL04P01 p.31)

对固定且条件分布有定义的 x ,

$$\text{Var}(Y | X = x) = E(Y^2 | X = x) - [E(Y | X = x)]^2.$$

条件均值描述知道 X 后 Y 的中心; 条件方差描述此后仍存在的波动。二者均可随 x 变化。对三角形例子, $X | (Y = y) \sim U(0, y)$, 所以 $E(X | Y = y) = y/2$ 、 $\text{Var}(X | Y = y) = y^2/12$ 。

英文概念与原理

English Definitions and Principles

CHAPTER 1

Probability, Counting and Distribution Functions

1. Experiments, events and probability

A sample space contains the outcomes of an experiment. Events belong to a sigma-algebra, which contains the full space and is closed under complements and countable unions. In a finite space, all unions of its atoms form such a family.

A probability measure is nonnegative, assigns one to the full space and is countably additive on disjoint events.

Disjointness is a set relation; independence is a probability relation. Complementation, inclusion–exclusion and the union bound do not require independence. Counting favorable outcomes divided by all outcomes is valid only when the chosen elementary outcomes are equally likely.

2. Counting and sampling mechanisms

Permutations retain order and combinations ignore it. Repeated indistinguishable objects require division by their multiplicity factorials. Required consecutive strings can be treated as blocks, subject to overlap checks.

Unlabeled equal-size groups require division both for internal order and for exchange of entire groups; labeled groups do not require the latter factor.

Sampling without replacement changes subsequent success chances and leads to a hypergeometric count. Independent fixed-probability draws lead to a binomial count. Numerator and denominator in a counting probability must use consistent ordering conventions.

3. Conditioning, total probability and Bayes' theorem

Conditional probability restricts and renormalizes the sample space. A partition consists of disjoint events covering it. Total probability sums the contributions of all possible types; Bayes' theorem reverses a conditional direction through joint probabilities.

The prior is a type probability before observation, the likelihood is the observation probability under that type, and the posterior is the updated type probability.

In screening, prevalence is the prior, sensitivity is the true-positive rate, and specificity the true-negative rate. The positive predictive value is a posterior, not sensitivity. Conditional independence does not imply marginal independence when observations share an unknown type.

Pairwise independence also need not imply mutual independence.

4. Random variables, PMFs, PDFs and CDFs

A random variable maps outcomes to numbers. Its support specifies feasible values. A discrete PMF assigns nonnegative point masses summing to one. An absolutely continuous PDF is nonnegative and integrates to one; interval probabilities are integrals and point probabilities vanish. A density may exceed one.

The CDF is $P(X \leq x)$ for every distribution. It is nondecreasing, right-continuous, and has limits zero and one. Its jump is point mass, while the left limit excludes that point. Survival is $1 - F(x)$. Always validate a proposed distribution before using it; differentiating an arbitrary function does not make it a PDF.

CHAPTER 2

Transformations, Expectations and Moments

1. Transformations: support before algebra

For $Y = g(X)$, first find the image of the original support. The CDF method rewrites $g(X) \leq y$ as events in X and works even for nonmonotone maps. Inverse-density formulas must sum every feasible branch and include the absolute inverse derivative. The feasible branch count can change across intervals.

The Jacobian compensates for stretching of the coordinate axis. An affine transformation has density $f_X((y - b)/a)/|a|$. Check the new support, nonnegativity, normalization and CDF continuity. Density values at finitely many continuous endpoints do not affect probabilities.

2. Quantiles and the probability integral transform

A quantile inverts cumulative probability, using the generalized inverse if necessary. A continuous median solves $F(m) = 1/2$ and need not equal the mean. For continuous strictly increasing F , the probability integral transform makes $F(X)$ uniform.

Conversely, applying the generalized inverse to a uniform variable generates the target CDF. Discrete $F(X)$ need not be uniform, so conditions matter.

3. Expectation, moments and variance

Expectation is a probability-weighted average. LOTUS computes $E[g(X)]$ directly using the original PMF or PDF, without deriving the transformed law. Raw moments are EX^k ; central moments center at the mean. Variance is $EX^2 - (EX)^2$ and standard deviation is its square root.

Expectation is linear without independence. Variance scales by the square of a multiplier and is invariant to shifts. A finite expectation requires absolute integrability; divergent positive and negative parts cannot cancel. Symmetry alone does not guarantee a mean.

Mixture raw moments are weighted component moments, but mixture variance also reflects differences between component means.

4. MGFs: existence, differentiation and identification

The MGF is $E(e^{tX})$. When finite on a neighborhood of zero, derivatives yield raw moments and the MGF uniquely identifies the distribution. Its convergence domain is part of the answer. Affine changes transform the argument and add an exponential factor; independent sums multiply MGFs.

Bounded support gives finiteness for all real arguments. Exponential and Laplace tails restrict the domain. A normal MGF exists everywhere, while a Cauchy MGF does not exist near zero. A lognormal has all power moments but no finite MGF for positive arguments. Recognizing Gamma or Gaussian integrals often simplifies derivations.

CHAPTER 3

Common Families of Distributions

1. Discrete families from experiment mechanisms

A Bernoulli indicator records one success. A binomial variable counts successes in a fixed number of independent constant-probability trials. Geometric and negative-binomial variables stop at success one or success r .

State whether the count includes all trials or only failures; the final success is fixed and only earlier positions are arranged.

Hypergeometric counts sample without replacement, with support limited by available successes and failures. Draws are not independent. Poisson counts have nonnegative integer support and intensity λ ; their event assumptions must be checked rather than inferred from a large trial number alone.

2. Discrete moments and MGFs

The table uses geometric and negative-binomial failure counts. Their MGFs require $t < -\ln(1 - p)$; shifting to total trials adds e^t or e^{rt} . Bernoulli, binomial and Poisson MGFs exist for all real t , as does any finite-support MGF.

Negative dependence from sampling without replacement gives the hypergeometric finite-population variance correction $(N - n)/(N - 1)$ for $N > 1$.

Poisson thinning conditions on the original random count, applies a binomial retention model, and sums over the count to obtain $\text{Poisson}(\lambda p)$. Zero truncation

instead conditions on a positive value and renormalizes by $1 - e^{-\lambda}$; it changes both moments and the distribution family.

3. Continuous families: intervals, waiting and normality

A continuous uniform law has constant density on an interval. An exponential law with scale β has rate $1/\beta$, mean β and variance β^2 . Its survival function gives memorylessness; a nontrivial mixture of unequal rates loses that property.

The Gamma density defines shape a and scale β , with mean $a\beta$ and variance $a\beta^2$. Rate notation replaces scale by its reciprocal in both the exponent and the normalizing constant. Shape one is exponential, and integer shape corresponds to accumulated exponential waits.

A Gamma waiting-time tail can be expressed as a Poisson count probability.

In $N(\mu, \sigma^2)$, the second parameter is variance and standardization divides by the standard deviation. Standard-normal CDF differences give interval probabilities. The sum of ν independent squared standard normals is chi-square with ν degrees of freedom, equivalently $\text{Gamma}(\nu/2, 2)$.

4. Beta, transformed families and heavy tails

Beta distributions model values in $(0, 1)$ using two shape parameters and a Beta-function normalizer; $\text{Beta}(1, 1)$ is uniform. A Weibull law arises by taking a positive power of an exponential variable, and its usual scale is the corresponding power of the exponential scale.

A lognormal is the exponential of a normal variable. Its parameters describe the logarithm, so its mean is not the exponential of the log mean. Laplace distributions have a location and scale with variance twice scale squared.

A standard Cauchy has such heavy tails that its mean does not exist; a symmetric principal value is not a valid expectation.

5. Continuous MGFs and their domains

Completing the square in the normal MGF integral shifts the normal mean to $\mu + \sigma^2 t$ and leaves the factor $e^{\mu t + \sigma^2 t^2/2}$. A Gamma integral instead changes its exponential decay coefficient to $1/\beta - t$.

The table lists both formulas and convergence domains. The uniform formula has a removable singularity at zero. Beta MGFs exist everywhere because the

support is bounded, even without a simple elementary formula. Cauchy and lognormal laws do not have the required finite MGF neighborhood.

Equal means and variances alone do not identify a distribution; MGF uniqueness requires agreement on a neighborhood of zero.

CHAPTER 4

Joint Distributions and Conditional Moments

1. Joint distributions and marginalization

A random vector has a joint law describing simultaneous values, not merely separate marginals. A joint CDF is defined for every law. Its rectangle increments give rectangle probabilities and must be nonnegative.

An ordinary joint PDF exists only for laws absolutely continuous with respect to area; continuous marginals alone do not guarantee it.

Marginalization sums or integrates out the other coordinate. Integration bounds come from the joint support. In the triangle $0 < x < y < 1$, fixing x leaves y between x and one, whereas fixing y leaves x between zero and y .

2. Independence and the support

Independence requires factorization for every pair of measurable events, equivalently for all joint CDF values, all joint masses, or almost everywhere for densities. A single counterexample disproves independence, but one successful point cannot prove it. The support must also be consistent with factorization.

Identical marginals do not determine a joint law. The pairs (U, U) and $(U, 1 - U)$ share uniform marginals but place probability on different lines. Joint expectations sum or integrate the function against the joint law. Independence implies factorization of product moments when defined, but the converse is false.

3. Conditional distributions as normalized slices

A conditional PMF divides joint mass by the conditioning marginal mass. A conditional PDF uses a density ratio where the marginal density is positive; it does not divide by a zero-probability point event. Holding one coordinate fixed gives a slice whose feasible support must be normalized.

Multiplying a marginal by a conditional reconstructs the joint density. One can then integrate for the other marginal and reverse the conditioning. Independence is an additional property under which conditioning leaves the other distribution unchanged. Conditional densities must integrate to one over their slice.

4. Conditional means and conditional variances

The conditional mean given $X = x$ is an ordinary expectation under a conditional law and is a function $m(x)$. Replacing x by X makes $E(Y | X)$ a random variable.

Conditional variance likewise depends on the conditioning value and equals the conditional second moment minus the squared conditional mean.

Recognized conditional families allow their usual moment formulas. The selected exercises check by explicit sums or integrals that averaging the conditional means recovers the marginal mean. Core coverage ends with the joint and conditional distributions and moments in the available notes.

Later inference topics are excluded; the older minimum question is only a CDF-and-independence bridge.