概率论与数理统计

课堂笔记

作者：Ni, Lyu

组织：数据科学与工程学院, DaSE

时间：Spring, 2023

版本：1.0

ElegantLaTeX Program

All models are wrong, but some are useful.——G E P Box

目录

第 1 讲 随机事件、概率的定义、计算方法与性质

阅读章节

- ❏ Intro to Prob 1.1 / 1.2 / 1.6
- ❏ Prob&Stat 1.1 / 1.2 / 1.3

问题 1.1 假定有一个封闭盒，盒中共有 10 个球，我们从中取一个球，球是什么颜色的?

情况一： 如果盒子里有 5 个白球，5 个红球；

情况二： 如果盒子里有 10 个白球。

我们归纳一下上述两种情况的差异，如表1.3。

表 1.1: 两种情况比较

	结果	现象	特点
情况一	$\{W\}$或$\{R\}$	随机现象	(1) 结果不止一个；(2) 事先无法确定。
情况二	$\{W\}$	确定性现象	结果只有一个

注

- 记 $W=$ “抽中的球为白球”，$R=$ “抽中的球为红球”。
- W, R 为基本结果或样本点。

概率论研究的是随机现象。接下来，我们定义一些基本概念。

1.1 基本概念

定义 1.1 (样本空间)

随机现象的一切可能基本结果组成的集合。♣

注

- 样本空间是一个集合；
- 样本空间可分为离散样本空间和连续样本空间；
  - 离散样本空间：样本点的个数为有限或可列个。
  - 连续样本空间：样本点的个数为不可列或无限个。

例题 1.1

- 定点投篮
  - 规则一：10 个球中，投中的球数.
    $\Omega=\{0,1,\cdots,10\}$；
  - 规则二：投中 5 个球所花费的时间.
    $\Omega=\{t: t\geq 0\}$；

- 规则三：投中 5 个球，投球的总次数.
  $\Omega=\{k:k=5,\cdots\}$;
- 投硬币：$\Omega=\{正,反\}$;
- 手机的一天使用时间：$\Omega=\{t:0\leq t\leq 24\}$。

问题 1.2 以上定义的样本空间中，哪些是离散样本空间？哪些是连续样本空间？

定义 1.2 (随机事件)
随机现象的某些样本点组成的集合。 ♣

例题 1.2 考虑投掷一颗六面骰子，样本空间为 $\Omega=\{1,2,3,4,5,6\}$。若我们感兴趣的是“骰子出现奇数点”，那么随机事件为 $A=\{1,3,5\}$。

注

- 通常用大写英文字母来表示随机变量；
- 基本事件：Ω 中单个元素组成的子集；
- Ω 必然事件：Ω 的最大子集；
- ϕ 不可能事件：Ω 的最小子集（即空集 ϕ）。

1.2 事件间的关系

表 1.2: 事件间的关系

事件间的关系	集合间的关系	Venn 图
事件 A 发生必然导致事件 B 发生	$A\subset B$	
事件 A 与事件 B 相等	$A=B\Longleftrightarrow A\subset B$ 且 $A\supset B$	
事件 A 与事件 B 不可能同时发生	A 与 B 互不相容且 $AB=\phi$	

1.3 事件间的运算

表 1.3: 事件间的运算

事件间的运算	集合间的运算	Venn 图
事件 A 与 B 中至少有一个发生	并：$A \cup B$	
事件 A 与 B 同时发生	交：$A \cap B = AB$	
事件 A 发生而 B 不发生	差：$A - B$	
事件 A 不发生	对立事件：$\overline{A} = \Omega - A$	

1.4 事件的运算性质

1. 交换律

$$A \cup B = B \cup A \qquad AB = BA$$

2. 结合律

$$(A \cup B) \cup C = A \cup (B \cup C) \qquad (AB)\,C = A\,(BC)$$

3. 分配律

$$(A \cup B) \cap C = AC \cup BC \qquad (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

4. 对偶律

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \qquad \overline{A \cap B} = \overline{A} \cup \overline{B}$$

注

1.
$$\overline{\bigcup_{i=1}^{n} A_i} = \bigcap_{i=1}^{n} \overline{A_i} \qquad \overline{\bigcap_{i=1}^{n} A_i} = \bigcup_{i=1}^{n} \overline{A_i}$$

2.
$$\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i} \qquad \overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i}$$

1.5 概率的定义

定义 1.3 (事件域)

设 Ω 为一样本空间。$\mathcal{F}$ 为 Ω 的某些子集所组成的集合。如果 $\mathcal{F}$ 满足以下条件

1. $\Omega \in \mathcal{F}$;
2. 若 $A \in \mathcal{F}$，则 $\overline{A} \in \mathcal{F}$;
3. 若 $A_n \in \mathcal{F}, n = 1, 2, \ldots$，则可列并 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$。

那么，称 $\mathcal{F}$ 为样本空间 Ω 的一个事件域，又称为 σ 域或 σ 代数. ♣

定义 1.4 (概率)

设 Ω 的一个样本空间，且 $\mathcal{F}$ 为 Ω 的一个事件域。

如果对任一事件 $A \in \mathcal{F}$，定义在 $\mathcal{F}$ 上的一个实值函数 $P(A)$ 满足以下条件

1. 非负性公理：若 $A \in \chi$，则 $P(A) \geq 0$;
2. 正则性公理：$P(\Omega) = 1$;
3. 可列可加性公理：若 $A_1, A_2, \ldots, A_n, \ldots$ 互不相容，则
$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

则称 $P(A)$ 为事件 A 的概率，称三元素 $\{\Omega, \mathcal{F}, P\}$ 为概率空间。♣

注

- 概率的公理化定义是由俄国数学家柯尔莫格罗夫提出的;
- 概率的公理化定义没有告诉我们如何确定概率。

1.6 概率的计算

在概率的公理化体系建立之前，概率可以用其他的定义方法：频率定义、古典定义、几何定义和主观定义。

1. 确定概率的频率方法
   - 【基本思想】
     - 与事件 A 有关的随机现象可大量重复进行;
     - 在 n 次重复试验中，记 $n(A)$ 为事件 A 出现的次数（也称频数），称频率
       $$f_n(A) = \frac{n(A)}{n}$$
       为事件 A 的概率。
     - 人们长期实践表明：随着试验次数 n 增加，频率 $f_n(A)$ 会稳定在某一常数 α 附近，称这个常数为频率的稳定值。**这个稳定值就是我们所定义的概率**。
   - 【优点】

    - 可供想象的具体值（简单）;
    - 在试验重复次数 n 较大时，频率是概率的一个近似值。
- 【缺点】
    - 在现实世界里，人们无法把一个试验无限次地重复下去。

问题 1.3 为什么我们的键盘是按这样排列的?

2. 确定概率的古典方法：在经验事实的基础上，对事件 A 的可能性进行逻辑分析后得出该事件的概率。
- 【基本思想】
    - 所设计的随机现象只有有限个样本点，设为 n 个。
    - 每个样本点发生的可能性相等（称等可能性）。
    - 若事件 A 含有 k 个样本点，则是假 A 的概率为
$$P(A) = \frac{k}{n}.$$

例题 1.3 投掷两枚硬币，如何求出现一正一反的概率?

解: 样本空间为
$$\Omega = \{(\text{正}, \text{正}),(\text{正}, \text{反}),(\text{反}, \text{正}),(\text{反}, \text{反})\}.$$
于是，样本空间 Ω 共有 4 个样本点。事件“出现一正一反”含有的样本点的个数为 2。因此，所求的概率为 $1/2$。

- 古典方法是概率轮发展初期确定概率的常用方法，所得的概率又称为古典概率。
- 在古典概率的计算中，经常使用排列组合工具。

注

排列：从 n 个不同的元素中任取 $r(r \leq n)$ 个元素拍成一列（考虑元素先后出现次序）。这种排列数为
$$P_n^r = n \times (n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$
特别地，$r = n$ 时，称 $P_n^n = n!$ 为全排列。

重复排列：从 n 个不同元素中每次去处一个，放回后在取下一个，如此连续取 r 次所得的排列。这种重复排列数共有 n^r 个。注意：这里允许 $r > n$。

组合：从 n 个不同元素中任取 $r(r \leq n)$ 个元素并成一组（不考虑元素间的先后次

序)。这种组合数为

$$C_n^r = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

特别地，(1) $0! = 1$；(2) $\binom{n}{0} = 1$ ；(3) $\binom{n}{r} = \binom{n}{n-r}$。

重复组合 ：从 n 个不同元素中每次取出一个，然后放回后再取下一个。如此连续取 r 次所得的组合。这种重复组合数为 $\binom{n+r+1}{r}$

问题 1.4 倘若总共有 3 个数，抽 2 次，求重复组合数为多少?

例题 1.4(不放回抽样) 一批产品共有 9 件，其中 3 件是不合格品，6 件事合格品。从中不放回地随机抽出 4 件。试求事件 $A_m=$ “取出 4 件产品中有 m 个不合格品” 的概率，$m=0,1,2,3$。

解:

$$\begin{aligned}
P(A_0) &= \frac{C_6^4}{C_9^4} = \frac{15}{126} = \frac{5}{42} \\
P(A_1) &= \frac{C_6^3 \cdot C_3^1}{C_9^4} = \frac{60}{126} = \frac{20}{42} \\
P(A_2) &= \frac{C_6^2 \cdot C_3^2}{C_9^4} = \frac{45}{126} = \frac{15}{42} \\
P(A_3) &= \frac{C_6^1 \cdot C_3^3}{C_9^4} = \frac{6}{126} = \frac{2}{42}
\end{aligned}$$

注

- 这个问题中 m 可以看作一个**随机变量**;
- m 取 $0,1,2,3$ 四种情况必有之一;
- m 的概率分布为

m	0	1	2	3
$P(A_m)$	$\frac{5}{42}$	$\frac{20}{42}$	$\frac{15}{42}$	$\frac{2}{42}$

例题 1.5(放回抽样) 从中有放回地随机抽出 4 件。试求事件 $B_m=$ “取出 4 件产品中有 m 个不合格品” 的概率。

解:

$$\begin{aligned}
P(B_0) &= \left(1-\frac{3}{9}\right)^4 = \left(\frac{2}{3}\right)^4 = \frac{16}{81} \\
P(B_1) &= \frac{4\times 3^1\times 6^3}{9^4} = 4\times\frac{3}{9}\times\left(\frac{6}{9}\right)^3 = \frac{32}{81} \\
P(B_2) &= \frac{C_4^2\times 3^2\times 6^2}{9^4} = 6\times\left(\frac{3}{9}\right)^1\times\left(\frac{6}{9}\right)^2 = \frac{24}{81} \\
P(B_3) &= \frac{C_4^3\times 3^3\times 6^1}{9^4} = 4\times\left(\frac{3}{9}\right)^3\times\left(\frac{6}{9}\right)^1 = \frac{8}{81} \\
P(B_4) &= \left(\frac{3}{9}\right)^4 = \frac{1}{81}
\end{aligned}$$

例题 1.6(盒子模型) 设有 n 个球，每个球都等可能地被放到 N 个不同盒子中的任一个。每个盒子所放球数不限。试求：

- 指定的 $n(n\le N)$ 个盒子中有一个球的概率 p_1；
- 恰好有 $n(n\le N)$ 个盒子各有一个球的概率 p_2；

解:

- $p_1=\frac{n!}{N^n}$；
- $p_2=\frac{C_N^n n!}{N^n}=\frac{P_N^n}{N^n}=\frac{N!}{N^n(N-n)!}$.

例题 1.7(生日问题) n 个人的生日全不相同的概率 p_n 是多少?

解:

3. 确定概率的几何方法
   - 【基本思想】
     - 如果一个随机现象的样本空间 Ω 充满某个区域，其度量（长度、面积或提及）大小用 S_Ω 表示；
     - 任意一点落在度量相同的子区域内（可能位置不同）是等可能的；
     - 若事件 A 为 Ω 中某个子区域，且其度量大小可用 S_A 表示，则事件 A 的概率为
     $$P(A)=\frac{S_A}{S_\Omega}.$$
     这个概率称为几何概率。

例题 1.8(会面问题) 甲乙两人约定下午 6 点到 7 点之间在某处会面，并约定先到者应等候另一个人 20 分钟，过时即可离去。求两人能会面的概率。

解: 设 x, y 分别表示甲、乙到达约会地点的事件。事件 A = "两人能会面" ⇔ $|x-y| \leq 20$. 可以计算

$$\begin{aligned} S_\Omega &= 60^2 = 3600 \\ S_A &= 3600 - 2 \times \frac{1}{2} \times 40^2 = 2000 \\ P(A) &= \frac{S_A}{S_\Omega} = \frac{2000}{3600} = \frac{5}{9}. \end{aligned}$$

例题 1.9(蒲丰投针) 平面上画有间隔为 $d(d>0)$ 的等距平行线，向平面任意投掷一枚长为 l 的针（l<d)，求针与任一平行线相交的概率。

解: 设 x 为针的中点与最近一条平行线的距离，ϕ 为针与此直线间的夹角。于是，x 和 ϕ 满足

$$0 \leq x \leq \frac{d}{2}, \quad 0 \leq \phi \leq \pi.$$

所以，

$$S_\Omega = \frac{\pi d}{2}.$$

而针与平行线相交，⇔ $x \leq \frac{l}{2}\sin\phi$。所以，

$$S_A = \int_0^\pi \frac{l}{2}\sin\phi \mathrm{d}\phi = \frac{l}{2}\left(-\cos\phi|_0^\pi\right) = \frac{l}{2} \times 2 = l.$$

因此，所求的概率为

$$P(A) = \frac{S_A}{S_\Omega} = \frac{l}{\frac{\pi d}{2}} = \frac{2l}{\pi d}.$$

注

- 当 l 与 d 已知时，$P(A)$ 可计算；
- 如果已知 $P(A)$，那么可以近似 π。(蒲丰投针实验是近似 π 值的有名实验)
- 只要设计一个随机试验，使得一个事件的概率与某个未知数有关，然后重复试验，以频率估计概率，即可求出未知数的近似值。
- 这个方法称为**随机模拟法**。

1.7 概率的性质

1. 不可能事件的概率为零，$P(\emptyset) = 0$；
2. 【有限可加性】若有限个事件 $A_1, A_2, \ldots, A_n$ 互不相容，则有
$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$
3. 【对立事件的概率】对任一事件 A，有 $P(\overline{A}) = 1 - P(A)$。
4. 【概率的单调性】

(a). 引理：若 $A \supset B$，则 $P(A-B)=P(A)-P(B)$；

(b). 推论：若 $A \supset B$，则 $P(A) \geq P(B)$；

(c). 性质：$\forall A, B$，有 $P(A-B)=P(A)-P(AB)$；

5. 【概率的加法公式】对任意两个事件 A, B，有

$$P(A \cup B)=P(A)+P(B)-P(AB)$$

(a). 【半可加性】对任意两个事件 A, B，有

$$P(A \cup B) \leq P(A)+P(B).$$

(b). 推论：对任意 n 个事件 $A_1, A_2, \ldots, A_n$，有

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i).$$

6. 【概率的连续性】

定义 1.5 (极限事件)

(a). 对 $\mathcal{F}$ 中任一单调不减的事件序列 $F_1 \subset F_2 \subset \cdots \subset F_n \subset \cdots$，称可列并且 $\bigcup_{n=1}^{\infty} F_n$ 为 $\{F_n\}$ 的极限事件记为

$$\lim_{n\to\infty} F_n=\bigcup_{n=1}^{\infty} F_n.$$

(b). 对 $\mathcal{F}$ 中任一单调不增的事件序列 $E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$，称可列并且 $\bigcap_{n=1}^{\infty} E_n$ 为 $\{E_n\}$ 的极限事件记为

$$\lim_{n\to\infty} E_n=\bigcap_{n=1}^{\infty} E_n.$$

♣

定义 1.6 (上、下连续)

对 $\mathcal{F}$ 上的一个概率 P，

(1) 若其对 $\mathcal{F}$ 中任一单调不减的事件序列 $\{F_n\}$ 均成立，

$$\lim_{n\to\infty} P(F_n)=P(\lim_{n\to\infty} F_n)$$

则称概率 P 是下连续的。

(2) 若其对 $\mathcal{F}$ 中任一单调不增的事件序列 $\{E_n\}$ 均成立，

$$\lim_{n\to\infty} P(E_n)=P(\lim_{n\to\infty} E_n)$$

则称概率 P 是上连续的。

♣

性质 连续性

若 P 为事件域 $\mathcal{F}$ 上的概率，则 P 既是下连续的又是上连续的。

证明 先证 P 的下连续性.

设 $\{F_n\}$ 是 $\mathcal{F}$ 中一个单调不减的事件序列，即

$$\lim_{n\to\infty} F_n = \bigcup_{n=1}^{\infty} F_n$$

若定义 $F_0=\phi$，则

$$\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty}(F_i - F_{i-1})$$

由于 $F_{i-1}\subset F_i$ ，显然 $(F_i - F_{i-1})$ 两两不相容，再由可列可加性得

$$\begin{aligned}
P(\bigcup_{i=1}^{\infty} F_i) &= P(\bigcup_{i=1}^{\infty}(F_i - F_{i-1})) \\
&= \sum_{i=1}^{\infty} P(F_i - F_{i-1}) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i - F_{i-1}) \\
&= \lim_{n\to\infty} P(\bigcup_{i=1}^{n}(F_i - F_{i-1})) \\
&= \lim_{n\to\infty} P(F_n)
\end{aligned} \tag{1.1}$$

所以，概率 P 的下连续性由此得证。

再证概率 P 的上连续性。

设 $\{E_n\}$ 是单调不增的事件序列，则 $\{\overline{E_n}\}$ 为单调不减的事件序列，由概率的下连续性可得

$$\begin{aligned}
1-\lim_{n\to\infty} P(E_n) &= \lim_{n\to\infty}(1-P(E_n)) \\
&= \lim_{n\to\infty} P(\overline{E_n}) \\
&= P(\bigcup_{n=1}^{\infty}\overline{E_n}) \\
&= P(\overline{\bigcap_{n=1}^{\infty} E_n}) \\
&= 1-P(\bigcap_{n=1}^{\infty} E_n)
\end{aligned} \tag{1.2}$$

因此，

$$\lim_{n\to\infty} P(E_n) = P(\bigcap_{n=1}^{\infty} E_n) = P(\lim_{n\to\infty} E_n)$$

这就证明了概率 P 的上连续性。

注 概率的公理化定义中，可以将可列可加性换成有限可加性和下连续性.

第 2 讲 条件概率的定义、三大公式及独立性

阅读章节

- ❑ Intro to Prob　1.3 1.4 1.5
- ❑ Prob&Stat　1.4 1.5

在疫情期间，当我核酸检测出现“阳性”之后，为什么我还需要二次检测?

问题 2.1 当我发现“阳性”之后，我真正感染新冠的概率是多少?

注

- 随机事件 A = “我感染了新冠”;
- 随机事件 B = “我的核酸是阳性”。
- “我感染了新冠的概率”与“核酸阳性的**条件下**，我感染新冠的概率”有什么差异?

2.1 条件概率的定义

定义 2.1 (条件概率)

设 A 与 B 是样本空间 Ω 中的两个事件，若 $P(B)>0$，则称

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

为“在事件 B 条件下，事件 A 的条件概率”，简称条件概率，记作 $P(A|B)$。 ♣

问题 2.2

1. $P(A|B)$ 与 $P(A)$ 不同。那么有何不同?
2. 条件概率是概率吗? 为什么?

定理 2.1 (乘法公式)

若 $P(B)>0$，则

$$P(AB)=P(B)\cdot P(A|B)$$

证明 乘法公式可由条件概率的定义直接推出。

推论 2.1

若 $P(A_1A_2\cdots A_{n-1})>0$，则

$$P(A_1A_2\cdots A_n)=P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

证明 令 $B_i=A_1A_2\cdots A_i, i=1,2,\cdots,n$。

一方面，我们想要证明所定义的 $P(B_i)$ 均大于零。因为 $P(B_{n-1})=P(A_1A_2\cdots A_{n-1})>0$ 而且

$$\begin{aligned} P(B_i) &= P(A_1A_2\cdots A_i) \\ &\geq P(A_1A_2\cdots A_i\cap A_{i+1}\cdot A_{n-1}) \\ &= P(B_{n-1})>0, i=1,2,\cdots,n-1. \end{aligned}$$

另一方面，我们证明 $P(A_1A_2\cdots A_n)=P(A_1A_2\cdots A_{n-1})\cdot P(A_n|A_1A_2\cdots A_{n-1})$. 因为 $P(B_{n-1})=P(A_1A_2\cdots A_{n-1})>0$，所以

$$\begin{aligned} P(B_n) &= P((A_1A_2\cdots A_{n-1})\cap A_n) \\ &= P(B_{n-1}\cap A_n) \\ &= P(B_{n-1})\cdot P(A_n|B_{n-1}) \\ &= P(B_{n-1})\cdot P(A_n|A_1A_2\cdots A_{n-1}). \end{aligned}$$

类似地，我们有 $P(B_i)=P(B_{i-1})P(A_i|B_{i-1}), i=2,3,\cdots,n-2.$ 由此得证。

例题 2.1(罐子模型) 设罐子里有 b 个黑球，r 个红球。每次随机取出一个球，取出后将原球放回，同时放入 c 个同色球和 d 个异色球。若连续从罐子里取出三个球，求三个球中有两个红球、一个黑球的概率。

解 记 B_i 为"第 i 次取出的是黑球"，R_j 为"第 j 次取出的是红球"。由乘法公式，可知所求的概率为

$$\begin{aligned} P(B_1R_2R_3) &= P(B_1)P(R_2|B_1)P(R_3|B_1R_2)=\frac{b}{b+r}\cdot\frac{r+d}{b+r+c+d}\cdot\frac{r+d+c}{b+r+2c+2d}, \\ P(R_1B_2R_3) &= P(R_1)P(B_2|R_1)P(R_3|R_1B_2)=\frac{r}{b+r}\cdot\frac{b+d}{b+r+c+d}\cdot\frac{r+d+c}{b+r+2c+2d}, \\ P(R_1R_2B_3) &= P(R_1)P(R_2|R_1)P(B_3|R_1R_2)=\frac{r}{b+r}\cdot\frac{r+c}{b+r+c+d}\cdot\frac{b+2d}{b+r+2c+2d}. \end{aligned}$$

罐子模型的四种变形

1. **不放回抽样**：$c=-1$ 且 $d=0$.
$$P(B_1R_2R_3)=P(R_1B_2R_3)=P(R_1R_2B_3)=\frac{br(r-1)}{(b+r)(b+r-1)(b+r-2)}$$
前次抽样结果会影响后次抽样结果；抽取的黑、红球个数确定，概率不依赖其抽球次序。
2. **放回抽样**：$c=0$ 且 $d=0$.
$$P(B_1R_2R_3)=P(R_1B_2R_3)=P(R_1R_2B_3)=\frac{br^2}{(b+r)^3}$$
前次抽样结果不会影响后次抽样结果；抽取的概率相等。
3. **传染病抽样**：$c>0$ 且 $d=0$.
$$P(B_1R_2R_3)=P(R_1B_2R_3)=P(R_1R_2B_3)=\frac{br(r+c)}{(b+r)(b+r+c)(b+r+2c)}$$
每次取出球后会增加下一次取到同色球的概率；每次发现一个传染病患者，以后都会增加再传染的概率
4. **安全抽样**：$c=0$ 且 $d>0$.
$$\begin{aligned}P(B_1R_2R_3)&=\frac{b}{b+r}\cdot\frac{r+d}{b+r+d}\cdot\frac{r+d}{b+r+2d}\\P(R_1B_2R_3)&=\frac{r}{b+r}\cdot\frac{b+d}{b+r+d}\cdot\frac{r+d}{b+r+2d}\\P(R_1R_2B_3)&=\frac{r}{b+r}\cdot\frac{r}{b+r+d}\cdot\frac{b+2d}{b+r+2d}\end{aligned}$$
每当事故发生，安全工作就抓紧，下次再发生事故的概率就会减少；反之，没有事故，发生事故的概率增大。

注

- 只要 $d=0$ 且抽取的黑球和红球个数确定，则所求的概率不依赖抽球的次序；
- 当 $d>0$ 时，则所求的概率与抽球的次序有关。

2.2 全概率公式

定义 2.2 (样本空间的分割)

对样本空间 Ω，如果有 n 个事件 $D_1,D_2,\cdots,D_n$ 满足：

1. 诸 D_i 互不相容：$\forall i\neq j, D_i\cap D_j=\phi$;
2. $\bigcup_{i=1}^n D_i=\Omega$;

则称 $\{D_1,D_2,\cdots,D_n\}$ 为样本空间 Ω 的一组分割。♣

例题 2.2 用 x 表示电视机的彩色浓度。则

- 样本空间 $\Omega=(-\infty,+\infty)$;
- 令 $D_1=\{|x-m|\leq a\}$ 代表一等品；
- $D_2=\{a<|x-m|\leq 2a\}$ 代表二等品；

- $D_3=\{2a<|x-m|\le 3a\}$ 代表三等品；
- $D_4=\{|x-m|>3a\}$ 代表不合格品。
- $\mathcal{D}=\{D_1,D_2,D_3,D_4\}$ 是一个分割。

定理 2.2 (全概率公式)

设 $B_1,B_2,\cdots,B_n$ 为样本空间 Ω 的一个分割，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{i=1}^n B_i=\Omega$。如果 $P(B_i)>0$，则对任一事件 A，有

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$$

♡

证明 因为

$$A=A\Omega=A\left(\bigcup_{i=1}^{n}B_i\right)=\bigcup_{i=1}^{n}(AB_i)$$

因为 $B_1,B_2,\cdots,B_n$ 互不相容，因此，$AB_1,AB_2,\cdots,AB_n$ 互不相容.
由可加性知，

$$P(A)=P\left(\bigcup_{i=1}^{n}(AB_i)\right)=\sum_{i=1}^{n}P\left(AB_i\right)=\sum_{i=1}^{n}P(B_i)P(A|B_i).$$

于是，定理得证。

注

1. 最简单的情况：如果 $0<P(B)<1$，那么 $P(A)=P(B)P(A|B)+P(\overline{B})P(A|\overline{B})$。
2. 在定理中，“$B_1,B_2,\cdots,B_n$ 为样本空间 Ω 的一个分割”可改为“$B_1,B_2,\cdots,B_n$ 互不相容，且 $A\subset\bigcup_{i=1}^n B_i$”，全概率公式仍成立。
3. 若 $B_1,B_2,\cdots$ 可列个事件互不相容，且 $A\subset\bigcup_{i=1}^\infty B_i$，则全概率公式仍成立。

例题 2.3 摸彩模型 设在 n 张彩票中有一张可中大奖，求第二个人摸出中奖彩票的概率是多少?

解 设 $A_i=$“第 i 个人摸到中奖彩票”，$i=1,2,\cdots,n$。第一个人模出中奖彩票的概率为

$$P(A_1)=\frac{1}{n},$$

而其未摸出中奖彩票的概率为

$$P(\overline{A_1})=\frac{n-1}{n}.$$

在第一个人是否模出中奖彩票的条件下，第二个人能否摸出中奖彩票 A_2 的条件概率是不同的，即

- 如果第一个人中奖，第二个人一定不会中奖，即
$$P(A_2|A_1)=0.$$
- 如果第一个人未中奖，第二个人中奖的机会更大，即
$$P(A_2|\overline{A_1})=\frac{1}{n-1}$$
；

由全概率公式可知

$$P(A_2)=P(A_1)P(A_2|A_1)+P(\overline{A_1})P(A_2|\overline{A_1})=\frac{1}{n}\cdot 0+\frac{n-1}{n}\cdot\frac{1}{n-1}=\frac{1}{n}$$

。

注 类似地，可以推导出

$$P(A_i)=\frac{1}{n}, i=1,2,\cdots,n.$$

这表明，摸到中奖彩票的机会与先后次序无关。

2.3 贝叶斯公式

定理 2.3 (贝叶斯公式)

设 $B_1,B_2,\cdots,B_n$ 为样本空间 Ω 的一个分割，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{i=1}^n B_i=\Omega$. 如果 $P(A)>0, P(B_i)>0, i=1,2,\cdots,n$，则

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}, i=1,2,\cdots,n$$

♡

证明 由条件概率的定义可知

$$P(B_i|A)=\frac{P(AB_i)}{P(A)}=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}.$$

于是定理得证。

例题 2.4 某地区居民的肝癌发病率为 0.0004，现用甲胎蛋白法进行普查。医学研究表明，化验结果是可能存在错误的。已知患有肝癌的人其化验结果 99% 呈阳性，而没患肝癌的人其化验结果 99.9% 呈阴性。现某人的检查结果呈阳性，问他真的患肝癌的概率是多少?

解 记 B 为事件“被检查者患有肝癌”，A 为事件“检查结果呈阳性”。根据题目可知，

$$\begin{aligned} P(B) &= 0.0004, \quad P(\overline{B})=1-0.0004=0.9996 \\ P(A|B) &= 0.99, \quad P(A|\overline{B})=1-0.999=0.001 \end{aligned}$$

我们关心的问题是求 $P(B|A)$。由贝叶斯公式得

$$\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(B)P(A|B)+P(\overline{B})P(A|\overline{B}} \\ &= \frac{0.0004\times 0.99}{0.0004\times 0.99+0.9996\times 0.001}=0.2837. \end{aligned}$$

这表明了，在检查结果呈阳性的人中，真患肝癌的人约为 28.37%。

注 分析原因：肝癌的发病率很低。在 10000 个人中约有 4 人患肝癌，而约有 9996 人不患肝癌。对 10000 个人用甲胎蛋白法进行检查，按其错检的概率可知，99996 个不患肝癌的人约有 $9996\times 0.001=9.996$ 个呈阳性。另外 4 个真患肝癌者的检查报告中约为 $4\times 0.99=3.96$ 个呈阳性。从 13.956 个呈阳性者中，真患肝癌的 3.96 人约占 28.37%。

例题 2.5(例题 1.4 续) 对首次检查呈阳性的人群再进行复查。此时 $P(B)=0.2837$，再次使用

贝叶斯公式可得

$$\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(B)P(A|B)+P(\overline{B})P(A|\overline{B}} \\ &= \frac{0.2847\times 0.99}{0.2847\times 0.99+0.7153\times 0.001} = 0.9975. \end{aligned}$$

接下来，我们用数学方式来讲一个寓言故事“狼来了”。

例题 2.6(狼来了) 设事件 A 为“小孩说谎”，事件 B 为“小孩可信”。

不妨设村民过去对这个小孩的印象为

$$P(B)=0.8 \quad P(\overline{B})=0.2.$$

第一次说谎后，村民对他的信任程度有所改变。我们假设可信的孩子说谎的概率为 $P(A|B)=0.1$，不可信的孩子说谎的概率为 $P(A|\overline{B})=0.4$.

利用贝叶斯公式，计算小孩第一次说谎后，其可信的概率为

$$P(B|A)=\frac{P(B)P(A|B)}{P(B)P(A|B)+P(\overline{B})P(A|\overline{B})}=\frac{0.8\times 0.1}{0.8\times 0.1+0.2\times 0.5}=0.444.$$

在一次说谎后，村民对小孩的信任程度由 0.8 变为 0.444. 即

$$P(B)=0.444 \quad P(\overline{B})=0.556.$$

第二次说谎后，村民对他的信任程度为

$$P(B|A)=\frac{0.444\times 0.1}{0.444\times 0.1+0.556\times 0.5}=0.138.$$

2.4 独立性

例题 2.7 回顾：罐子模型中有回放机制 VS 无回放机制

假设罐子里有 r 个红球，有 b 个黑球。令 R_1 为事件“第一个人抽到红球”，R_2 为事件“第二个人抽到红球”。

在**有回放**机制下，“第二个人是否能够抽到红球”不受到“第一个人是否抽到红球”结果的影响。于是，

$$\begin{aligned} P(R_1) &= \frac{r}{r+b} \\ P(R_2) &= \frac{r}{r+b} \\ P(R_1R_2) &= \frac{r^2}{(r+b)^2} = \frac{r}{r+b}\cdot\frac{r}{r+b} = P(R_1)\cdot P(R_2). \end{aligned}$$

在**无回放**机制下，“第二个人是否能够抽到红球”会受到“第一个人是否抽到红球”结果的影响。

$$
\begin{aligned}
P(R_1) &= \frac{r}{r+b} \\
P(R_2) &= P(R_1)P(R_2|R_1) + P(\overline{R_1})P(R_2|\overline{R_1}) \\
&= \frac{r}{r+b}\cdot\frac{r-1}{r+b-1} + \frac{b}{r+b}\cdot\frac{r}{r+b-1} = \frac{r}{r+b} \\
P(R_1R_2) &= \frac{C_r^2}{C_{r+b}^2} = \frac{r!/(2!(r-2)!)}{(r+b)!/(2!(r+b-2)!)} = \frac{r(r-1)}{(r+b)(r+b-1)} \neq P(R_1)\cdot P(R_2).
\end{aligned}
$$

比较上述两个例子，我们可以定义“独立性”的概念。**独立性**指的是一个事件的发生（与否）不影响另一个事件的发生（与否）。

定义 2.3 (独立性)

如果

$$P(AB) = P(A)P(B)$$

成立，则称事件 A 与 B 相互独立，简称 A 与 B 独立。否则，称 A 与 B 不独立或相依。♣

性质 若事件 A 与 B 独立，则

1. A 与 $\overline{B}$ 独立；
2. $\overline{A}$ 与 B 独立；
3. $\overline{A}$ 与 $\overline{B}$ 独立；

证明 这里我们仅证明第一个性质，另外两条性质供学生课后自行证明。由于概率的性质可知，

$$P(A\overline{B}) = P(A) - P(AB) = P(A) - P(A)P(B) = P(A)(1-P(B)) = P(A)P(\overline{B}).$$

于是，性质得证。

定义 2.4 (两两独立性 vs 相互独立)

设有三个事件 A, B, C。如果

$$
\begin{cases}
P(AB) = P(A)P(B) \\
P(AC) = P(A)P(C) \\
P(BC) = P(B)P(C)
\end{cases}
$$

则称 A, B, C 两两独立。若还有

$$P(ABC) = P(A)P(B)P(C),$$

则称 A, B, C 相互独立。 ♣

注

- A, B, C 两两独立 $\not\Rightarrow A, B, C$ 相互独立.
  例题 2.8 反例 设又一个均匀的正四面体，其第一面染成红色，第二面染成白色，第三面染成黑色，而第四面染成有红、白、黑三种颜色。现在以 A, B, C 分别记为投一次四面体出现红、白、黑颜色的事件，则由于四面体种有两面染有红色，因此，$P(A) = P(B) = P(C) = 1/2$。另外，容易算出
  $$P(AB) = P(BC) = P(AC) = 1/4.$$
  所以，A, B, C 两两独立。但是，
  $$P(ABC) = 1/4 \neq 1/8 = P(A)P(B)P(C).$$
  因而，A, B, C 不相互独立。
- A, B, C 相互独立 $\not\Rightarrow A, B, C$ 两两独立.
  例题 2.9 反例 请同学课后自行构建一个反例。

例题 2.10 设 A, B, C 三个事件相互独立，那么事件 $A \cup B$ 与 C 相互独立。

证明 因为 $P(A \cup B)C = AC \cup BC$. 所以，

$$\begin{aligned}
P((A \cup B)C) &= P(AC \cup BC) \\
&= P(AC) + P(BC) - P(AC \cap BC) \\
&= P(A)P(C) + P(B)P(C) - P(A)P(B)P(C) \\
&= (P(A) + P(B) - P(A)P(B)) \cdot P(C) \\
&= P(A \cup B)P(C).
\end{aligned}$$

因此，$A \cup B$ 与 C 相互独立。

定义 2.5 (n 个事件的独立性)

设有 n 个事件 $A_1, A_2, \cdots, A_n$，对任意的 $1 \leq i < j < k < \cdots \leq n$，如果

$$\begin{cases} P(A_iA_j) = P(A_i)P(A_j) \\ P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k) \\ \vdots \\ P(A_1A_2\cdots A_n) = \prod_{i=1}^{n} P(A_i) \end{cases}$$

则称此 n 个事件 $A_1, A_2, \cdots, A_n$ 相互独立。♣

定义 2.6 (条件独立性)

设有三个事件 A, B, C，且 $P(C) > 0$。如果

$$P(AB|C) = P(A|C)P(B|C)$$

则称在给定事件 C,A 与 B 是条件独立的。♣

注

- 若 $P(B \cap C) > 0$, 给定 C，A 与 B 是条件独立的等价于 $P(A|B \cap C) = P(A|C)$.
- 独立性无法推导出条件独立性；反之亦然。

例题 2.11 考虑独立地投掷两枚公平的硬币，即所有结果都是等可能的。令

$$\begin{aligned} H_1 &= \{\text{第一枚硬币正面}\} \\ H_2 &= \{\text{第二枚硬币正面}\} \\ D &= \{\text{两枚硬币的结果不一致}\} \end{aligned}$$

于是，H_1 和 H_2 是独立的。但是，

$$P(H_1|D) = \frac{1}{2}, \quad P(H_2|D) = \frac{1}{2}, \quad P(H_1 \cap H_2|D) = 0.$$

所以，$P(H_1 \cap H_2|D) \neq P(H_1|D)P(H_2|D)$。这意味着 H_1 和 H_2 不是条件独立的。

例题 2.12 有两枚硬币，一红一蓝。我们从中等概率地随机选择一枚，并独立地进行两次投掷。假定这两枚硬币是有偏的。蓝色的硬币正面朝上的概率为 0.99，而红色的硬币正面朝上的概率为 0.01.

令 B 为事件“蓝色的硬币被选中”，H_i 为事件“第 i 次投掷的结果为正面朝上”。在选定硬币后，H_1 和 H_2 是独立的。于是，

$$P(H_1 \cap H_2|B) = P(H_1|B)P(H_2|B) = 0.99^2.$$

另一方面，H_1 和 H_2 不是独立的。这是因为，如果第一次投掷的结果为正面朝上，者导致我们会猜测，选中的是蓝色硬币，这样第二次投掷正面朝上的概率就更大。从数学公式中，

我们可以计算

$$\begin{aligned} P(H_1) &= P(H_2) = P(B)P(H_1|B) + P(\overline{B})P(H_1|\overline{B}) = 0.5 \cdot 0.99 + 0.5 \cdot 0.01 = 0.5. \\ P(H_1 \cap H_2) &= P(B)P(H_1 \cap H_2|B) + P(\overline{B})P(H_1 \cap H_2|\overline{B}) = 0.5 \cdot 0.99^2 + 0.5 \cdot 0.01^2 = 0.4901. \end{aligned}$$

因此，$P(H_1 \cap H_2) \neq P(H_1)P(H_2)$，即 H_1 和 H_2 不独立。

第 3 讲　一元随机变量的定义、累积分布函数，概率分布列及常见离散型随机变量

阅读章节

- ❑ Intro to Prob　2.1 2.2
- ❑ Prob&Stat　2.1 2.4

3.1　一元随机变量的定义

问题 3.1 什么是随机变量?

定义 3.1

定义在样本空间 Ω 上的实值函数 $X = X(\omega), \omega \in \Omega$，称 X 为随机变量。♣

注 通常用大写字母 X, Y, Z 表示随机变量，其取值用小写字母 x, y, z 等表示。

例题 3.1 考虑抛一枚硬币三次的结果。记硬币正面朝上为 “1”，而硬币反面朝上为 “0”。我们考虑样本空间及硬币正面朝上的次数。

表 3.1: 抛三枚硬币的结果

抛三次硬币的结果	正面朝上的次数 X
$\omega_1 =$ （反，反，反）	0
$\omega_2 =$ （正，反，反）	1
$\omega_3 =$ （反，正，反）	1
$\omega_4 =$ （反，反，正）	1
$\omega_5 =$ （正，正，反）	2
$\omega_6 =$ （正，反，正）	2
$\omega_7 =$ （反，正，正）	2
$\omega_8 =$ （正，正，正）	3

这里，一枚硬币抛三次，正面朝上的次数就是所定义的**随机变量**。

注

1. 不同的样本点对应不同的实数；
2. 多个样本点对应同一个实数；
3. 样本点可以用数值表示，也可以不用数值表示；但随机变量一定是数值型。

问题 3.2 随机变量与变量有什么区别?

3.2 累积分布函数

定义 3.2

设 X 为一个随机变量。对任意实数 x，称

$$F(x) = P(X \leq x)$$

为随机变量 X 的累积分布函数，简称分布函数。也称 X 服从 $F(x)$，记 $X \sim F(x)$。♣

注 任一随机变量 X 都有一个分布函数。

问题 3.3 为什么专门需要计算 $P(X \leq x)$?

性质 任一分布函数 $F(x)$ 都具有如下三条基本性质：

1. 单调性：$F(x)$ 是定义在整个实数轴 $(-\infty, \infty)$ 上的单调非减函数，即对任意 $x_1 < x_2, F(x_1) \leq F(x_2)$。
2. 有界性：对任意的 x，有 $0 \leq F(x) \leq 1$，且
$$\begin{cases} F(-\infty) = \lim\limits_{x\to-\infty} F(x) = 0 \\ F(\infty) = \lim\limits_{x\to\infty} F(x) = 1 \end{cases}$$
3. 右连续性：$F(x)$ 是 x 的右连续函数，即对任意的 x_0，有
$$\lim_{x\to x_0+0} F(x_0) = 0, \quad F(x_0 + 0 =)F(x_0).$$

证明

1. 设事件 $A=\{\omega: X(\omega) \leqslant x_1\} \quad B=\{\omega: X(\omega) \leqslant x_2\}$。若 $x_1<x_2$，那么 $A \subset B$。根据概率的单调性，
$$P\left(x \leqslant x_1\right)=P(A) \leqslant P(B)=P\left(X \leqslant x_2\right).$$
即 $F(x_1) \leq F(x_2)$。
2. 因为 $F(x)=P(X \leq x)$ 表示事件 $X \leq x$ 的概率。所以，$0 \leq F(x) \leq 1$。根据 $F(x)$ 的单调性可知，$F(\lfloor x\rfloor) \leqslant F(x) \leqslant F(\lceil x\rceil)$。对于整数 m 和 n，
$$\lim_{x \rightarrow-\infty} F(x)=\lim_{m \rightarrow-\infty} F(m) \quad \lim_{x \rightarrow \infty} F(x)=\lim_{x \rightarrow \infty} F(n)$$
存在。由于概率的可列可加性，
$$\begin{aligned}
1 &= P(-\infty<X<+\infty) \\
&= P\left(\bigcup_{i=-\infty}^{+\infty}\{i-1<X \leqslant i\}\right) \\
&= \sum_{i=-\infty}^{+\infty} P(i-1<X \leqslant i) \\
&= \lim_{n \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{i=m}^{n} P(i-1<X \leqslant i)=\lim_{n \rightarrow \infty} \lim_{m \rightarrow-\infty} P(m<X \leqslant n) \\
&= \lim_{n \rightarrow \infty} F(n)-\lim_{m \rightarrow-\infty} F(m) \Rightarrow 1 \geqslant \lim_{n \rightarrow \infty} F(n)=1+\lim_{m \rightarrow \infty} F(m) \geqslant 1
\end{aligned}$$
因此，$\lim\limits_{n \rightarrow \infty} F(x)=1=\lim\limits_{x \rightarrow \infty} F(x) \quad ; \lim\limits_{m \rightarrow-\infty} F(m)=0=\lim\limits_{x \rightarrow \infty} F(x)$。
3. 因为 $F(x)$ 是有界、单调函数，所以其任一点 x_0 的右极限 $F(x_0+0)$ 存在。考虑数列 $x_1>x_2>\ldots>x_n>\ldots>x_0$，当 $x_n \Rightarrow x_0 (n \Rightarrow \infty)$ 时，
$$\begin{aligned}
F\left(x_1\right)-F\left(x_0\right)=P\left(x_0<x \leqslant x_1\right)=P\left(\bigcup_{i=1}^{\infty}\left\{x_{i+1}<x \leqslant x_i\right\}\right) \\
=\sum_{i=1}^{\infty} P\left(x_{i+1}<x \leqslant x_i\right)=\sum_{i=1}^{\infty} F\left(x_i\right)-F\left(x_{i+1}\right) \\
=\lim_{n \rightarrow \infty} \sum_{i=1}^{n}\left[F\left(x_i\right)-F\left(x_{i+1}\right)\right]=F\left(x_1\right)-\lim_{n \rightarrow \infty} F\left(x_n\right)
\end{aligned}$$
因此，$F\left(x_0\right)=\lim_{n \rightarrow \infty} F\left(x_n\right)=F\left(x_0+0\right)$。

注

1. 这三条基本性质也是判断某个函数是否为某个随机变量的分布函数的充要条件。
2. 利用分布函数，我们可以容易地计算概率
$$\begin{aligned}
P(a<X \leq b) &= F(b)-F(a) \\
P(X=a) &= F(a)-F(a-0) \\
P(X>b) &= 1-F(b)
\end{aligned}$$
3. 特别地，$F(x)$ 在 ab 处连续时，有 $F(a-0)=F(a) \qquad F(b-0)=F(b)$。

例题 3.2 对于反正切函数

$$F(x)=\frac{1}{\pi}\left(\arctan x+\frac{\pi}{2}\right),-\infty<x<\infty.$$

我们可以发现

1. $F(x)$ 是连续且严格单调增函数;
2. $F(\infty)=1$ 且 $F(-\infty)=0$;
3. $F(x)$ 是某个随机变量的分布函数，该分布为柯西分布。
4. 设 X 为一个服从柯西分布的随机变量，则

$$\begin{aligned}P(-1\le X\le 1) &= F(1)-F(-1)\\ &= \frac{1}{\pi}(\arctan(1)-\arctan(-1))\\ &= \frac{1}{2}.\end{aligned}$$

3.3 概率分布列

定义 3.3

设 X 是一个离散随机变量，如果 X 的所有可能取值 $x_1,x_2,..,x_n,...$，则称 X 取 x_i 的概率

$$p_i=p(x_i)=P(X=x_i)\quad i=1,2,...,n,...$$

为 X 的概率分布列，或称概率质量函数（p.m.f）。♣

概率分布列的通常写法为

X	x_1	x_2	$\cdots$	x_n	$\cdots$
P	p_1	p_2	$\cdots$	p_n	$\cdots$

例题 3.3 掷两颗骰子，其样本空间含有 36 个样本点 $\Omega=\{(x,y):x,y=1,2,\dots,6\}$。在 Ω 上我们定义 3 个随机变量 X,Y,Z。

- X 为两颗骰子的点数之和，则其分布列为

- Y 为 6 点的骰子个数，则其分布列为

- Z 为最大点数，则其分布列为

性质 任一概率分布列 $\{p_i\}$ 都具有如下两条基本性质：

1. 非负性：$p(x_i) \geq 0, i = 1, 2, \cdots$；
2. 正则性：$\sum_{i=1}^{\infty} p(x_i) = 1$。

注 上述两个性质也是判断某个函数是否为随机变量的概率分布列的充要条件。

问题 3.4 分布函数与分布列有什么关系？

3.4 常见的离散随机变量

3.4.1 单点分布

定义 3.4

常量 c 可作为仅取一个值的随机变量 X，即

$$P(X = c) = 1.$$

这个分布称为单点分布或退化分布。

3.4.2 二项分布

在介绍二项分布之前，这里先介绍一个重要的概念——伯努利试验。

定义 3.5

1. 设有两个试验 E_1 和 E_2。假如试验 E_1 的任一结果（事件）与试验 E_2 的任一结果（事件）都是相互独立的事件，则称这两个试验相互独立的。
2. 如果 E_1 的任一结果，E_2 的任一结果，$\cdots$，E_n 的任一结果都是相互独立的事件，则称试验 $E_1, E_2, \cdots, E_n$ 相互独立。

3. 如果这 n 个独立试验还是相同的，则称为 n 重独立重复试验。
4. 如果在 n 重独立重复试验中，每次试验的可能结果有两个（A 和 $\overline{A}$），则称这种试验为 n 重伯努利试验。

♣

定义 3.6

假定伯努利试验中成功（事件 A）概率为 p。记 X 为 n 重伯努利试验中成功的次数，其分布列为

$$P(X=k)=C_n^k p^k(1-p)^{(n-k)}, k=0,1,2,\cdots,n.$$

称这个分布为二项分布。记 $X\sim b(n,p)$。

♣

注

1. 二项分布的分布列中每一项 $C_n^k p^k(1-p)^{(n-k)}$ 恰好是 n 次二项式 $(p+(1-p))^n$ 的展开式中 $k+1$ 项。这正是其名称的由来。

3.4.3 二点分布

二点分布是一种特殊的二项分布，即 $n=1$。

定义 3.7

假定伯努利试验中成功（事件 A）概率为 p。记 X 为一次伯努利试验中成功的次数，其分布列为

$$P(X=k)=p^k(1-p)^{(1-k)}=\begin{cases} p, k=1 \\ 1-p, k=0. \end{cases}, k=0,1.$$

称这个分布为二点分布（或伯努利分布）。记 $X\sim b(1,p)$。

♣

注 二项分布与二点分布之间的关系：服从二项分布的随机变量可以分解为 n 个独立同为二点分布的随机变量之和，即设 $X\sim b(n,p)$ 且 $X_i \overset{\text{i.i.d}}{\sim} b(1,p), i=1,2,\cdots,n$，有 $X=\sum_{i=1}^n X_i$.

3.4.4 几何分布

定义 3.8

假定伯努利试验中成功（事件 A）概率为 p。记 X 为伯努利试验首次成功的次数，其分布列为

$$P(X=k)=(1-p)^{(k-1)}p, k=1,2,\cdots.$$

称这个分布为几何分布。记 $X\sim Ge(p)$。

♣

性质 （无记忆性）设 $X \sim Ge(p)$，则对任意正整数 m 和 n，有

$$P(X > m+n|X > m) = P(X > n).$$

证明 因为 X 的概率分布列为 $P(X=k) = (1-p)^{k-1}p, k = 1, 2, \cdots,$，所以，

$$P(X > n) = \sum_{k=n+1}^{+\infty} (1-p)^{k-1}p = p \cdot \frac{(1-p)^n}{p} = (1-p)^n.$$

因此，对于任意正整数 m 和 n，条件概率为

$$P(X > m+n|X > m) = \frac{P(X > m+n, X > m)}{P(X > m)} = \frac{P(X > m+n)}{P(X > m)} = \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n$$

3.4.5 负二项分布

定义 3.9

假定伯努利试验中成功（事件 A）概率为 p。记 X 为伯努利试验第 r 次成功的次数，其分布列为

$$P(X=k) = C_{k-1}^{r-1}(1-p)^{(k-r)}p^r, k = r, r+1, r+2, \cdots.$$

称这个分布为负二项分布。记 $X \sim Nb(r,p)$。♣

注 负二项分布与几何分布的关系：负二项分布的随机变量可以分解为 r 个独立同分布的几何分布随机变量之和，即如果 $X \sim Nb(r,p)$，$X_i \overset{\text{i.i.d}}{\sim} Ge(p)$，那么，$X = \sum_{i=1}^{r} X_i$.

3.4.6 泊松分布

定义 3.10

假定一个离散随机变量 X，其分布列为

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}, k = 0, 1, 2, \cdots$$

其中，参数 $\lambda > 0$。称随机变量 X 的概率分布为泊松分布，记 $X \sim P(\lambda)$.♣

注 泊松分布是有法国数学家 Siméon-Denis Poisson 教授提出的。

问题 3.5 函数

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}, k = 0, 1, 2, \cdots$$

是一个随机变量的分布列吗?

证明 要论证一个函数是一个随机变量的分布列，就是要证明该函数满足非负性和正则性。

- 要证明非负性，即 $P(X=k) \geq 0$。一个显然的结果 $P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda} > 0, k = 0, 1, 2, \cdots$。这是因为参数 $\lambda > 0$。

- 要证明正则性，即 $\sum_{k=0}^{\infty} P(X=k)=1$。于是，
$$1=\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!}=e^{-\lambda}e^{\lambda}.$$
也就是说，要证明
$$\sum_{k=0}^{+\infty}\frac{\lambda^k}{k!}=e^{\lambda}.$$
以下给出具体的证明：令 $f(x)=e^{\lambda x}$ 可知
$$\begin{aligned} f'(x) &= \lambda e^{\lambda x}, \\ f''(x) &= \lambda^2 e^{\lambda x}, \\ &\vdots \\ f^{(k)}(x) &= \lambda^k e^{\lambda x}. \end{aligned}$$
利用在 $x=x_0$ 泰勒展开可知
$$f(x)=f(x_0)+\sum_{k=1}^{\infty}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k.$$
所以，取 $x=1$ 且 $x_0=0$ 时，有
$$e^{\lambda}=f(1)=f(0)+\sum_{k=1}^{\infty}\frac{f^{(k)}(0)}{k!}=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}.$$

注

- 泊松分布是一种常见的离散分布。
- 常与单位时间上的计算过程相关。

问题 3.6 泊松分布与二项分布有什么关系？

例题 3.4 学校（东门）单位时间内的车流量，也就是考虑单位时间内通过 k 辆车的概率。
我们将单位时间（以一小时为例）切分为 n 个小时段，n 非常大。
接下来提两个理想化假设：

- $\frac{1}{n}$ 非常小，在一个小时段内仅可通过一辆车；
- 在各个小时段内是否通过车是互不影响的。假设通过一辆车的概率为 p_n。在 n 个时段中，有 k 个时段中通过了一辆车，而余下的 $n-k$ 个时段中并没有通过。

定理 3.1

在 n 重伯努利试验中，记事件 A 在一次试验中发生的概率为 p_n（与试验次数 n 有关），如果当 $n\to\infty$ 时，有 $np_n\to\lambda$，则

$$\lim_{n\to\infty} C_n^k p_n^k(1-p_n)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda}.$$

证明 令 $\lambda_n=np_n$，即 $p_n=\frac{\lambda_n}{n}$，可得

$$\begin{aligned}C_n^k p_n^k(1-p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!}(\frac{\lambda_n}{n})^k(1-\frac{\lambda_n}{n})^{n-k}\\ &= \frac{\lambda_n^k}{k!}(1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{k-1}{n})(1-\frac{\lambda_n}{n})^{n-k}\end{aligned}$$

对固定的 k 有

$$\lim_{n\to\infty}\lambda_n=\lambda,$$

于是有

$$\lim_{n\to\infty}(1-\frac{\lambda_n}{n})^{n-k}=e^{-\lambda}$$

$$\lim_{n\to\infty}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})=1$$

从而

$$\lim_{n\to\infty} C_n^k p_n^k(1-p_n)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda},$$

对任意 k 均成立。

第 4 讲 一元随机变量的概率密度函数及常见连续型随机变量

阅读章节

- ❑ Intro to Prob 3.1 3.2 3.3
- ❑ Prob&Stat 2.1 2.5

4.1 概率密度函数

相较于离散型随机变量，本课程所定义的连续型随机变量通常是狭义的。如果一个随机变量的取值能够充满一个区间 (a, b)，这里 $a < b$ 是两个实数，也允许无穷值。

例题 4.1 学生到达教室的时间？

定义 4.1

设随机变量 X 的分布函数为 $F(x)$, 如果存在实数轴上的一个非负可积函数 $p(x)$，使得对任意实数 x 有

$$F(x) = \int_{-\infty}^{x} p(t)dt,$$

那么称 $p(x)$ 为 X 的概率密度函数（probability density function, p.d.f.)。称随机变量 X 为连续型随机变量，其分布函数 $F(x)$ 是连续分布函数。♣

性质 连续型随机变量 X 的概率密度函数 $p(x)$，其具有

1. 非负性 $p(x) \geqslant 0$;
2. 正则性 $\int_{-\infty}^{\infty} p(x)dx$。

注 上述两条也是判别某个函数是否为密度函数的充要条件。

例题 4.2 向区间 $(0, a)$ 上任意投点，用 X 表示其坐标。设这个点落在 $(0, a)$ 中任一小区间的概率与这个小区间的长度成正比，而与小区间位置无关。求 X 的分布函数和密度函数。

解 记 X 的分布函数为 $F(x)$, 则

- 当 $x < 0$ 时，因为 $\{X \leq x\}$ 是不可能事件, 所以 $F(x) = P(X \leq x) = 0$;

- 当 $x \geq a$ 时，因为 $\{X \leq x\}$ 是必然事件, 所以 $F(x) = P(X \leq x) = 1$;
- 当 $0 \geq x < a$ 时，$F(x) = P(X \leq x) = P(0 \leq X \leq x) = kx$，其中 k 为比例系数。因为 $F(a) = ka$，所以 $k = \frac{1}{a}$。于是，分布函数为

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

令 X 的密度函数 $p(x)$,

- 当 $x < 0$ 或 $x > a$ 时，$p(x) = F'(x) = 0$;
- 当 $x > a$ 时，$p(x) = F'(x) = 0$,
- 当 $0 < x < a$,

$$p(x) = F'(x) = \frac{1}{a}.$$

而在 $X = 0$ 和 $x = a$ 处，$p(x)$ 可取任意值，一般就近取值为宜，这不会影响概率的计算。于是，X 的概率密度是

$$p(x) = \begin{cases} \dfrac{1}{a}, & 0 < x < a \\ 0, & \text{其他} \end{cases}$$

注 这个分布就是区间 $(0, a)$ 上的均匀分布，记 $U(0, a)$。

例题 4.3 验证

$$f(x)=\begin{cases}\dfrac{1}{2\sqrt{x}}, & 0<x<1\\ 0, & \text{其他}\end{cases}$$

是否为一个随机变量 X 的概率密度函数?

解 【思路】验证 $f(x)$ 是否满足非负性和正则性。具体证明过程由同学们课后证明。

注 比较密度函数与分布列的异同:

1. 已知概率分布列或概率密度函数，可以求出概率分布函数或概率值;
2. 离散随机变量的分布函数 $F(x)$ 总是右连续的阶梯函数；而连续随机变量的分布函数 $F(x)$ 一定是整个数轴上的连续函数，其增量为
$$F(x+\Delta x)-F(x)=\int_x^{x+\Delta x}p(t)dt\to 0(\Delta x\to 0).$$
3. 离散型随机变量 X 在其可能取值的点 $x1,x2,...,xn,...,$ 上的概率不为 0，而连续随机变量 X 在 $(-\infty,+\infty)$ 上任一点 a 的概率恒为 0，即 $P(x=a)=\int_a^a p(x)dx=0$。这可以作为以下观点的一个例子：不可能事件的概率为 0，但概率为 0 的事件不一定是不可能事件。类似地，必然事件的概率为 1，但概率为 1 的事件不一定是必然事件。
4. 连续型随机变量 $P(a\le X\le b)=P(a<X\le b)=P(a\le X<b)=P(a<X<b)$，但离散型随机变量并不满足这个性质，要“点点计较”。
5. 由于在若干点上改变概率密度函数 $p(x)$ 的值并不影响其积分值，从而不影响分布函数 $F(x)$ 的值。这意味着一个连续分布的密度函数不唯一。

4.2 常见的连续随机变量

4.2.1 均匀分布

定义 4.2

设定义在区间 (a,b) 的一个随机变量 X，其中 $a<b$ 为两个未知参数。X 的概率密度函数为

$$p(x)=\frac{1}{b-a},a<x<b.$$

称 X 的分布为均匀分布。记 $X\sim U(a,b)$。♣

4.2.2 正态分布

定义 4.3

设定义在区间 $(-\infty,\infty)$ 的一个随机变量 X。X 的概率密度函数为

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

称 X 的分布为正态分布。记 $X\sim N(\mu,\sigma^2)$，其中参数 $\mu\in R,\sigma^2>0$。♣

注

1. 正态分布是最早由法国数学家棣莫弗（Abraham de Moivre）在近似二项分布时得到的，后由德国数学家高斯（Carolus Fridericus Gauss）在测量误差时导出。因高斯的工作对后世的贡献巨大，所以，正态分布又称高斯分布。
2. 概率密度函数 $p(x)$ 是一条钟型曲线，特点为：中间高，两边低，左右对称。
3. 正态分布的两个参数 μ 和 σ^2 是决定密度函数位置和形状，称 μ 为位置参数，σ^2 是尺度参数。

这里很自然我们构建一个正态分布类，即

$$\mathcal{P}=\{N(\mu,\sigma^2):\mu\in R,\sigma^2>0\}.$$

其中有个极为特殊的正态分布——标准正态分布，即 $\mu=0,\sigma^2=1$。这里我们具体讲解一下这个特殊的正态分布。**注**

1. 标准正态分布的密度函数为

$$\phi(z)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}z^2\right\};$$

2. 标准正态分布的分布函数为

$$\Phi(z)=\int_{-\infty}^{z}\phi(x)\mathrm{d}x;$$

3. 标准正态分布的概率计算常用公式：
   (a). $\Phi(-z)=P(Z\le -z)=P(Z\ge z)=1-\Phi(z)$

(b). $P(Z > z) = 1 - \Phi(z)$
(c). $P(a < Z < b) = \Phi(b) - \Phi(a)$
(d). $P(|Z| < c) = 2\Phi(c) - 1(c \geq 0)$

定理 4.1

若随机变量 $X \sim N(\mu, \sigma^2)$，则 $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$.

证明 记 X 和 Z 的分布函数分别为 $F_X(x)$ 和 $F_Z(z)$，密度函数分别为 $p_X(x)$ 和 $p_Z(z)$. 则由分布函数的定义可知

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P(\frac{X-\mu}{\sigma} \leq z) \\ &= P(X \leq \mu + \sigma\mu) \\ &= F_X(\mu + \sigma\mu) \end{aligned} \tag{4.1}$$

由于正态分布函数是严格单调递增且处处可导。因此

$$\begin{aligned} p_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) \\ &= \frac{\mathrm{d}}{\mathrm{d}z} F_X(\mu + \sigma z) \\ &= p_X(\mu + \sigma z) \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mu+\sigma z-\mu)^2} \cdot \sigma \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \end{aligned} \tag{4.2}$$

由此可得

$$U = \frac{x-\mu}{\sigma} \sim N(0,1)$$

注 3σ 原则：

1. $P(\mu - \sigma < X < \mu + \sigma) = 2\Phi(1) - 1 \approx 0.6826$；
2. $P(\mu - 2\sigma < X < \mu + 2\sigma) = 2\Phi(2) - 1 \approx 0.9545$；
3. $P(\mu - 3\sigma < X < \mu + 3\sigma) = 2\Phi(3) - 1 \approx 0.9973$。

4.2.3 指数分布

定义 4.4

设一随机变量 X，其密度函数为

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

则称 X 的分布为指数分布，记 $X \sim e^{\lambda}$，其中参数 $\lambda > 0$。

根据随机变量的密度函数，可以计算其分布函数为

$$F_X(x) = \begin{cases} \int_0^x p(t)dt = \int_0^x \lambda e^{-\lambda t}dt = e^{-\lambda t}|_0^x = 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

类似于泊松分布，指数分布也具有无记忆性。

性质 （无记忆性）

如果随机变量 $X \sim e^{\lambda}$，则对任意 $s > 0, t > 0$ 有

$$P(X > t + s | X > s) = P(X > t)$$

证明 因为 $X \sim e^{\lambda}$，所以 $P(X > s) = e^{-\lambda s}, s > 0$。又因为 $\{X > s + t\} \subset \{X > s\}$，于是，条件概率

$$P(X > s + t | X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

泊松分布与指数分布有非常紧密的关系，我们利用以下一个例子来说明。

例题 4.4 如果某设备在长为 t 的时间 $[0, t]$ 内发生故障的次数 $N(t)$（与时间长度 t 有关）服从参数为 λt 的泊松分布，则相继两次故障之间的时间间隔 T 服从参数为 λ 的指数分布。

解 设 $N(t) \sim P(\lambda t)$，即

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}, k = 0, 1, \cdots$$

注意到两次故障之间的时间间隔 T 是非负随机变量且事件 $\{T > t\}$ 说明此设备在 $[0, t]$ 内没有发生故障。即 $\{T > t\} = \{N(t) = 0\}$. 由此可得，

- 当 $t < 0$ 时，有
$$F_T(t) = P(T \leq t) = 0$$
；
- 当 $t \geq 0$ 时，有
$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - P(N(t) = 0) = 1 - e^{-\lambda t}$$
。

因此，$T \sim \text{Exp}\{\lambda\}$，即相继两次故障之间的时间间隔 T 服从参数为 λ 的指数分布。

4.2.4 伽马分布

定义 4.5

称

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

为伽马函数。

根据伽马函数的定义，可以证明伽马函数的一些常用性质。

性质

1. $\Gamma(1)=1,\quad \Gamma(\frac{1}{2})=\sqrt{\pi}$;
2. $\Gamma(\alpha+1)=\alpha\Gamma(\alpha),\quad$ 特别地，$\Gamma(n+1)=n\Gamma(n)=n!$。

证明

1. 令 $x=u^2$，于是，

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^{+\infty}x^{-\frac{1}{2}}e^{-x}dx\\ &= \int_0^{+\infty}u^{-1}e^{-u^2}d(u^2)\\ &= 2\int_0^{+\infty}e^{-u^2}du\\ &= \int_{-\infty}^{+\infty}e^{-u^2}du\\ &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}}}e^{-\frac{u^2}{2(\frac{1}{2})}}du\cdot\sqrt{\pi}\\ &= \sqrt{\pi}.\end{aligned}$$

2. 这里只需要证明 $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$，即

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^{+\infty}x^{\alpha}e^{-x}dx\\ &= \int_0^{+\infty}(-x^{\alpha})\,d\left(e^{-x}\right)\\ &= -x^{\alpha}e^{-x}|_0^{+\infty}+\int_0^{+\infty}e^{-x}d\left(x^{\alpha}\right)\\ &= \int_0^{+\infty}\alpha x^{\alpha-1}e^{-x}dx\\ &= \alpha\Gamma(\alpha).\end{aligned}$$

基于伽马函数，我们来定义伽马分布。

定义 4.6

假设 X 为一随机变量，其密度函数为

$$p(x)=\begin{cases}\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & ,x\geq 0\\ 0 & ,x<0\end{cases}$$

则称其分布为伽马分布，记作 $X\sim Ga(\alpha,\lambda)$，其中 $\alpha>0$ 为形状参数，$\lambda>0$ 为尺度参数。♣

注 当 $\alpha=1$ 时，$Ga(1,\lambda)=e^{\lambda}$。

以下例子讲解了泊松分布与伽马分布之间的关系，和之前讲解过的关于泊松分布与指数分布之间的关系的证明过程类似，供学生课后自学。

例题 4.5 若在 $(0,t)$ 内发生冲击的次数 $N(t)$ 服从参数为 λt 的泊松分布，试证明第 n 次冲击来到的时间 S_n 服从伽马分布 $Ga(n,\lambda)$。

证明 因为事件"第 n 次冲击来到的时间 S_n 小于等于 t"等价于事件"$(0,t)$ 内发生冲击的次数 $N(t)$ 大于等于 n"，即

$$\{S_n\leq t\}=\{N(t)\geq n\}.$$

于是，S_n 的分布函数为

$$F(t)=P(S_n\leq t)=P(N(t)\geq n)=\sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t}.$$

令 $a_k=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$, $b_k=\frac{\lambda^k}{\Gamma(k)}\int_t^{+\infty}x^{k-1}e^{-\lambda x}dx$。由于

$$\begin{aligned}
b_k &= \frac{\lambda^k}{\Gamma(k)}\int_t^{+\infty}x^{k-1}e^{-\lambda x}dx\\
&= \frac{\lambda^k}{(k-1)!}\int_t^{+\infty}-\frac{1}{\lambda}x^{k-1}de^{-\lambda x}\\
&= \frac{\lambda^k}{(k-1)!}(-\frac{1}{\lambda}x^{k-1}e^{-\lambda x}|_t^{+\infty}+\frac{k-1}{\lambda}\int_t^{+\infty}x^{k-2}e^{-\lambda x}dx)\\
&= \frac{\lambda^k}{(k-1)!}(\frac{1}{\lambda}t^{k-1}e^{-\lambda t}+\frac{k-1}{\lambda}\int_t^{+\infty}x^{k-2}e^{-\lambda x}dx)\\
&= \frac{\lambda^{k-1}}{(k-1)!}t^{k-1}e^{-\lambda t}+\frac{\lambda^{k-1}}{(k-2)!}\int_t^{+\infty}x^{k-2}e^{-\lambda x}dx\\
&= a_{k-1}+b_{k-1}
\end{aligned}$$

可得 $a_{k-1}=b_k-b_{k-1}$，则

$$\begin{aligned}
\sum_{k=1}^{n-1}a_k &= \sum_{k=1}^{n-1}(b_{k+1}-b_k)\\
&= (b_2-b_1)+\cdots+(b_n-b_{n-1})=b_n-b_1\\
&= \frac{\lambda^n}{\Gamma(n)}\int_t^{+\infty}x^{n-1}e^{-\lambda x}dx-a_0.
\end{aligned}$$

且

$$\begin{aligned} b_1 &= \frac{\lambda}{\Gamma(1)}\int_t^{+\infty} e^{-\lambda x}dx \\ &= \lambda\int_t^{+\infty} e^{-\lambda x}dx \\ &= -e^{-\lambda x}|_t^{+\infty} = e^{-\lambda t} = a_0 \end{aligned}$$

由此，我们有

$$\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!}e^{-\lambda t} = \frac{\lambda^n}{\Gamma(n)}\int_t^{+\infty} x^{n-1}e^{-\lambda x}dx$$

因此，

$$\begin{aligned} F(t) &= \sum_{k=n}^{+\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t} \\ &= 1-\frac{\lambda^n}{\Gamma(n)}\int_t^{+\infty} x^{n-1}e^{-\lambda x}dx \\ &= 1-\int_t^{+\infty}\frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}dx \\ &= \int_0^t\frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}dx \end{aligned}$$

所以，$S_n \sim Ga(n,\lambda)$.

4.2.5 贝塔分布

定义 4.7

称

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx, a>0, b>0$$

为贝塔函数。

根据贝塔函数的定义，可以证明贝塔函数的一些常用性质。

性质

1.
2. $B(a,b) = B(b,a)$;
3. $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$。

证明 由伽马函数的定义可知

$$\Gamma(a)\Gamma(b) = \int_0^{+\infty}\int_0^{+\infty} x^{a-1}y^{b-1}e^{-(x+y)}dxdy.$$

作变量变换

$$\begin{cases} x = uv \\ y = u(1-v) \end{cases} \Longrightarrow \begin{cases} u = x + y \\ v = \frac{u}{x} \end{cases} \Longrightarrow J = \begin{vmatrix} v & u \\ (1-v) & -u \end{vmatrix} = -u$$

故

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^1 \int_0^{+\infty} (uv)^{a-1}(u(1-v))^{b-1}e^{-u}ududv \\ &= \int_0^{+\infty} u^{a+b-1}e^{-u}du \cdot \int_0^1 v^{a-1}(1-v)^{b-1}dv \\ &= \Gamma(a+b)B(a,b). \end{aligned}$$

基于贝塔函数，我们来定义贝塔分布。

定义 4.8

假设一随机变量 X，其密度函数为

$$p(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & ,0 < x < 1 \\ 0 & ,other \end{cases}$$

则称 X 的分布为贝塔分布，记 $X \sim Be(a,b)$，其中 $a > 0, b > 0$ 均为形状参数。♣

注 特别地，当 $a = 1, b = 1$ 时，$Be(1,1) = U(0,1)$。

第 5 讲 期望、方差及其他特征数

阅读章节

- ❏ Intro to Prob 2.4
- ❏ Prob&Stat 2.2 2.3 2.7

5.1 数学期望

问题 5.1 分赌本问题 甲、乙两人赌技不相上下。两人进行赌博，各出赌注 50 法郎，每局中无平局。他们约定：谁先赢三局，则赢得全部赌本 100 法郎。当甲赢了二局，乙赢了一局，因故要中止赌博。现问这 100 法郎如何分才算公平?

解

1. 甲得 $\frac{2}{3}$, 乙得 $\frac{1}{3}$。这个方案看似是合理的。但是，如果甲赢了一局，乙赢了零局，那么这种分法的结果是：甲得 100 法郎, 乙得 0 法郎, 这样是否合理？
2. 设想再赌下去，则甲最终所得 X 为一个随机变量，其可能取值为 0 或 100，再赌两局必可结束。接下来两局的结果为 $\{$甲甲, 甲乙, 乙甲, 乙乙$\}$。这四种情况中有三种情况：甲赢了 100 法郎。只有一种情况：乙赢下 100 法郎，因为两人赌技不相上下，所以，在两局后甲的收益是一个随机变量，其分布列为

X	100	0
P	$\frac{3}{4}$	$\frac{1}{4}$

于是，甲的“期望”收益应为

$$100 \times \frac{3}{4} + 0 \times \frac{1}{4} = 75(\text{法郎})$$

相应地，乙的“期望”收益应为 25（法郎）。

定义 5.1 (数学期望)

1. 假设 X 为一个离散随机变量，其分布列记为 $p_i = P(X = x_i) \quad i = 1.2, \cdots, n, \cdots$。如果 $\sum_{i=1}^{\infty} |x_i| p_i < \infty$，则称
$$E(x) = \sum_{i=1}^{\infty} x_i p(x_i)$$
为随机变量 X 的（数学）期望或均值。
但若 $\sum_{i=1}^{\infty} |x_i| p_i$ 不收敛，则称 X 的期望不存在
2. 假设 X 为一个连续随机变量，其密度函数 $p(x)$。如果 $\int_{-\infty}^{+\infty} |x| p(x) dx < \infty$，则称
$$E(x) = \int_{-\infty}^{+\infty} x p(x) dx$$
为随机变量 X 的（数学）期望或均值。

但若 $\int_{-\infty}^{+\infty}|x|p(x)$ 不收敛，则称 X 的数学期望不存在。♣

例题 5.1 如果 X 的分布列为

X	-1	0	1
P	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

那么 $E(X)=0$。而 X^2 的分布列为

X	0	1
P	$\frac{1}{3}$	$\frac{2}{3}$

那么 $E(X^2)=\frac{2}{3}$。

定理 5.1

若随机变量 X 的分布用分布列 $p(x_i)$ 或密度函数 $p(x)$ 表示，则 X 的某一函数 $g(X)$ 的数学期望为

$$E(g(x))=\begin{cases}\sum_i g\left(x_i\right)p\left(x_i\right), & \text{在离散场合,}\\ \int_{-\infty}^{+\infty} g(x)p(x)dx, & \text{在连续场合.}\end{cases}$$

♡

性质

1. 若 c 是常数，则 $E(c)=c$；
2. 对任意常数 a，有 $E(aX)=aE(X)$；
3. 对任意的两个函数 $g_1(x)$ 和 $g_2(x)$，有
$$E(g_1(X)\pm g_2(X))=E(g_1(X))\pm E(g_2(X)).$$

5.2 随机变量的方差与标准差

例题 5.2 考虑 X 和 Y 两个随机变量。

$$P(X=k)=\begin{cases}1/3, & k=-1;\\ 1/3, & k=0;\\ 1/3, & k=1;\end{cases}\qquad P(Y=k)=\begin{cases}1/3, & k=-10;\\ 1/3, & k=0;\\ 1/3, & k=10;\end{cases}$$

我们可以计算 $E(X)=E(Y)=0$，但是这两个随机变量是不同的：Y 的取值波动比 X 取值的波动大。

问题 5.2 我们如何刻画随机变量取值的波动大小呢?

定义 5.2 (方差与标准差)

若随机变量 X^2 的数学期望 $E(X^2)$ 存在，则称偏差平方

$$E(X-E(X))^2$$

为随机变量 X 的方差，记为

$$\begin{aligned} Var(X) &= E(X-E(X))^2 \\ &= \begin{cases} \sum\limits_i \left(x_i - E(x)\right)^2 p\left(x_i\right), & \text{在离散场合} \\ \int_{-\infty}^{+\infty} (x-E(X))^2 p(x)dx, & \text{在连续场合} \end{cases} \end{aligned}$$

称方差的平方根 $\sqrt{\mathrm{Var}(X)}$ 为随机变量 X 的标准差，记为 $\sigma(x)$ 或 σ_x。♣

性质

1. $Var(X)=E(X^2)-(E(X))^2$;
2. 常数的方差为零，即若 c 为常数，则 $Var(c)=0$;
3. 若 a,b 为常数，则 $Var(aX+b)=a^2Var(X)$.

证明

定理 5.2 (Chebyshev 不等式)

设随机变量 X 的数学期望和方差都存在，则对任意常数 $\varepsilon>0$, 有

$$P(|x-Ex|\geqslant\varepsilon)\leq\frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

♡

注 事件 $|x-Ex|\geqslant\varepsilon$ 称为大偏差，其概率 $P(|x-Ex|\geqslant\varepsilon)$ 称为大偏差发生概率。

证明 (仅考虑连续情况) 其密度函数 $p(x)$, 记 $E(x)=\mu$,

$$\begin{aligned} P(|X-\mu|\geqslant\varepsilon) &= \int_{\{x:|x-\mu|\geqslant\varepsilon\}} p(x)dx \\ &\leqslant \int_{\{x:|x-\mu|\geqslant\varepsilon\}} \frac{(x-\mu)^2}{\varepsilon^2}p(x)dx \\ &\leqslant \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(x-\mu)^2p(x)dx \\ &= \frac{\mathrm{Var}(X)}{\varepsilon^2}. \end{aligned}$$

这里介绍三个常见随机变量期望和方差的计算方法。

例题 5.3 如果 $X \sim b(n,p)$，那么 $E(X) = np$ 且 $Var(X) = np(1-p)$。

解 X 的分布列为

$$P(X=k) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, 2, \cdots, n.$$

于是，X 的期望为

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} kP(X=k) \\
&= \sum_{k=0}^{n} k \cdot C_n^k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&\overset{k'=k-1}{=} np \cdot \sum_{k'=0}^{n-1} \frac{(n-1)!}{(k')!(n-1-k')!} p^{k'} (1-p)^{n-1-k'} \\
&= np,
\end{aligned}$$

因为 X 的方差 $Var(X) = E(X^2) - (E(X))^2$，所以，需要计算 X^2 的期望。而 X^2 的期望为

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k^2 P(X=k) \\
&= \sum_{k=0}^{n} k^2 \cdot C_n^k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k^2 \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \cdot \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \cdot \sum_{k=1}^{n} k \cdot \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&\overset{k'=k-1}{=} np \cdot \sum_{k'=0}^{n-1} (k'+1) \frac{(n-1)!}{(k')!(n-1-k')!} p^{k'} (1-p)^{n-1-k'} \\
&= np \cdot ((n-1)p + 1) \\
&= n(n-1)p^2 + np,
\end{aligned}$$

于是，

$$Var(X) = E(X^2) - (E(X))^2 = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p).$$

例题 5.4 如果 $X \sim N(\mu, \sigma^2)$，那么 $E(X) = \mu$ 且 $Var(X) = \sigma^2$。

解 X 的密度函数为

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

于是，X 的期望为

$$\begin{aligned}
E(X) &= \int_0^{\infty} xp(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x \\
&\overset{z=x-\mu}{=} \int_{-\infty}^{\infty} (z+\mu)\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z)^2\right\}\mathrm{d}z \\
&= \int_{-\infty}^{\infty} z\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z)^2\right\}\mathrm{d}z + \mu\int_{-\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z)^2\right\}\mathrm{d}z \\
&= \mu.
\end{aligned}$$

而 X 的方差为

$$\begin{aligned}
Var(X) &= E(X-E(X))^2 \\
&= \int_{-\infty}^{\infty} (x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x \\
&\overset{z=x-\mu}{=} \int_{-\infty}^{\infty} z^2\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}z^2\right\}\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} -\sigma^2 z\mathrm{d}\exp\left\{-\frac{1}{2\sigma^2}z^2\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\cdot\left(-\sigma^2 z\exp\left\{-\frac{1}{2\sigma^2}z^2\right\}\Big|_{-infty}^{\infty} + \int_{-\infty}^{\infty}\sigma^2\exp\left\{-\frac{1}{2\sigma^2}z^2\right\}\mathrm{d}z\right) \\
&= \int_{-\infty}^{\infty} \sigma^2\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}z^2\right\}\mathrm{d}z \\
&= \sigma^2.
\end{aligned}$$

例题 5.5 如果 $X \sim Ga(\alpha, \lambda)$，那么 $E(X) = \frac{\alpha}{\lambda}$ 且 $Var(X) = \frac{\alpha}{\lambda^2}$。

解 X 的密度函数为

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp\{-\lambda x\}, x > 0.$$

这里我们先考虑 X^k 的期望

$$\begin{aligned}
E(X^k) &= \int_0^{\infty} x^k p(x)\mathrm{d}x \\
&= \int_0^{\infty} x^k \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp\{-\lambda x\}\mathrm{d}x \\
&= \int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+k-1}\exp\{-\lambda x\}\mathrm{d}x \\
&= \int_0^{\infty} \frac{(\alpha+k-1)\times\cdots\times(\alpha)}{\lambda^k}\cdot\frac{\lambda^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1}\exp\{-\lambda x\}\mathrm{d}x \\
&= \frac{(\alpha+k-1)\times\cdots\times(\alpha)}{\lambda^k}.
\end{aligned}$$

于是，X 的期望为

$$E(X)=\frac{\alpha}{\lambda}$$

. 而 X 的方差为

$$Var(X)=E(X^2)-(EX)^2=\frac{(\alpha+1)\alpha}{\lambda^2}-\frac{\alpha^2}{\lambda^2}=\frac{\alpha}{\lambda^2}.$$

注 在计算期望时，一个最常用的方法是“合成概率函数”。

5.3 k 阶矩

定义 5.3

设 X 为随机变量，k 为正整数。如果以下的数学期望都存在，则

1. 称
$$\mu_k=E(X^k)$$
为 X 的 k 阶原点矩。
2. 称
$$\nu_k=E((X-E(X))^k)$$
为 X 的 k 阶中心矩。

♣

注

1. μ_1 是数学期望；
2. ν_2 是方差；
3. k 阶矩存在时，$k-1$ 阶矩也存在，从而低于 k 的各阶矩都存在。
4. 中心矩和原点矩的关系

$$\nu_k=E(X-E(X))^k=E(X-\mu_1)^k=\sum_{i=0}^{k}C_k^i\mu_i(-\mu_1)^{k-i}.$$

以下介绍三个无量纲的指标。

5.4 变异系数

定义 5.4

设随机变量 X 的二阶矩存在，则称比值

$$C_v(X) = \frac{\sqrt{Var(X)}}{E(X)} = \frac{\sigma(X)}{E(X)}.$$

为 X 的变异系数。

♣

注 变异系数是一个无量纲的量，从而消除量纲对波动的影响。

5.5 偏度系数

定义 5.5

设随机变量 X 的前三阶矩阵存在，则比值

$$\beta_s = \frac{\nu_3}{\nu_2^{3/2}} = \frac{E(X - E(X))^3}{(\mathrm{Var}(X))^{3/2}}$$

称为 X 或分布的偏度系数，简称偏度。

♣

注

1. 偏度 β_s 是描述分布偏离对称程度的一个特征数；
2. $\beta_s < 0$ 时，称左偏，又称负偏；
3. $\beta_s > 0$ 时，称右偏，又称正偏；
4. 对于连续型随机变量，当密度函数 $p(x)$ 关于其数学期望对称，则其三阶中心矩 ν_3 必为 0，从而 $\beta_s = 0$。

- 当密度 $p(x)$ 关于数学期望堆成时,$V_3 = 0 \Rightarrow \beta_S = 0$.
- $\beta_S < 0$ 时，称左偏
- $\beta_S > 0$ 时，称右偏

5.6 峰度系数

定义 5.6

设随机变量 X 的前四阶存在，则

$$\beta_k = \frac{\nu_4}{\nu_2^2} - 3 = \frac{E(X - E(X))^4}{(\operatorname{Var}(X))^2} - 3$$

称为 X 或分布的峰度系数，简称峰度。♣

注

1. 峰度是描述分布是否存在“尖峰后尾”现象的一个特征数；
2. 正态分布 $N(\mu, \sigma^2)$ 中 $\nu_2 = \sigma^2$ 且 $\nu_4 = 3\sigma^4$（请同学们课后证明）。于是，$\beta_k = 3\sigma^4/\sigma^4 - 3 = 0$。这表明了峰度系数 β_k 是相对于正态分布而言的超出量。
3. $\beta_k > 0$ 表示“厚尾”分布；
4. $\beta_k < 0$ 表示“薄尾”分布；
5. 偏度和峰度都是描述分布形状的特征数。

5.7 分位数

定义 5.7

设连续随机变量 X 的分布函数为 $F(x)$，密度函数为 $p(x)$。对任意 $p \in (0,1)$,，称满足条件

$$F\left(x_p\right) = \int_{-\infty}^{x_p} p(x) dx = p$$

的 x_p 为此分布的 p 分位数，又称下侧 p 分位数。
称满足条件

$$1 - F\left(x_p'\right) = \int_{x_p'}^{+\infty} p(x) dx = p$$

的 x_p' 为此分布的上侧 p 分位数。♣

5.8 中位数

定义 5.8

设连续随机变量 X 的分布函数为 $F(x)$，密度函数为 $p(x)$，称 $p = 0.5$ 时的 p 分位数 $x_{0.5}$ 为此分布的中位数，即

$$F\left(x_{0.5}\right)=\int_{-\infty}^{x_{0.5}} p(x) d x=0.5$$

♣

例题 5.6 若 X 服从参数为 λ 的指数分布，求其中位数。

解 X 的分布函数为

$$F(x)=1-\exp\{-\lambda x\}.$$

设 $x_{0.5}$ 为所求的中位数，即

$$F(x_{0.5})=1-\exp\{-\lambda x_{0.5}\}=0.5.$$

则

$$x_{0.5}=\ln(2)/\lambda.$$

第 6 讲 一元随机变量函数的分布

阅读章节

❑ Intro to Prob 2.3 ❑ Prob & Stat 2.6

问题 6.1 本节要解决的问题是：如果 X 是一个已知分布的随机变量，且我们知道某种函数（或者某种映射关系）的具体形式 $g:\Re\to\Re$，所以，只要 $g(x)$ 不是一个常值函数，$g(X)$ 通常仍是一个随机变量。因此，我们想要知道 $Y=g(X)$ 的分布是怎样的？能否求出其分布的形式？

6.1 离散随机变量函数的分布

已知 X 为一个离散型随机变量，其分布列为 欲求 $Y=g(X)$ 的分布？

X	x_1	x_2	$\cdots$	x_n	$\cdots$
P	$p(x_1)$	$p(x_2)$	$\cdots$	$p(x_n)$	$\cdots$

1. $Y=g(X)$ 也是一个离散随机变量；
2. Y 的分布列为

Y	$g(x_1)$	$g(x_2)$	$\cdots$	$g(x_n)$	$\cdots$
P	$p(x_1)$	$p(x_2)$	$\cdots$	$p(x_n)$	$\cdots$

3. 如果 $g(x_1),g(x_2),\cdots,g(x_n),\cdots$ 中有某些值相等时，则把那些相等的值分别合并，并把对应的概率相加即可。

例题 6.1 倘若 X 的分布列为

X	-2	-1	0	1	2
P	0.2	0.1	0.1	0.3	0.3

于是，$Y=X^2+X$ 的分布列为

Y	0	2	6
P	0.2	0.5	0.3

笔记

6.2 连续随机变量函数的分布

6.2.1 $g(X)$ 是一个离散型随机变量

倘若 $Y = g(X)$ 是一个离散型随机变量，我们只需要将 Y 的所有可能取值一一列出，在球 Y 的取各个可能值的概率。

例题 6.2 若 X 服从正态分布 $N(\mu, \sigma^2)$，则

$$Y = \begin{cases} 0 & , X < \mu \\ 1 & , X \geq \mu \end{cases}$$

那么，Y 服从二点分布 $b(1, 0.5)$。

6.2.2 $g(\cdot)$ 是严格单调函数

当 $g(x)$ 是一个关于 x 的严格单调函数，我们有以下定理。

定理 6.1

设 X 是连续随机变量，其密度函数为 $p_X(x)$。$Y = g(X)$ 是另一个连续随机变量。若 $y = g(x)$ 严格单调，其反函数 $h(y)$ 有连续导函数，则 $Y = g(X)$ 的密度函数为

$$p_Y(y) = \begin{cases} p_X\left[h(y)\right]\left|h'(y)\right|, a < y < b \\ 0, \text{其他} \end{cases}$$

其中，$a = \min\left\{g(-\infty), g(+\infty)\right\}, b = \max\left\{g(-\infty), g(+\infty)\right\}$

证明 不妨设 $g(x)$ 是一个严格单调增函数，这时它的反函数 $h(y)$ 也是严格单调增函数，且 $h'(y) > 0$。

首先，考虑 Y 的取值范围。记 $a = \min\{g(-\infty), g(\infty)\}$，$b = \max\{g(-\infty), g(\infty)\}$。于是，$Y = g(X)$ 仅在区间 (a, b) 取值。其次，考虑 Y 的分布函数和密度函数。

- 若 $y \leq a$，那么 $P(Y \leq y) = 0$，则 $p_Y(y) = 0$；
- 若 $y \geq b$，那么 $P(Y \leq y) = 1$，则 $p_Y(y) = 0$；
- 若 $a < y < b$，那么 Y 的分布函数为

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(h(g(X)) \leq h(y)) \\ &= P(X \leq h(y)) \\ &= \int_a^{h(y)} p_X(x) \mathrm{d}x \end{aligned}$$

Y 的密度函数为

$$\begin{aligned} p_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}F(y) = \frac{\mathrm{d}}{\mathrm{d}y}P(Y \le y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{h(y)} p_X(x)\mathrm{d}x \\ &= p_X(h(y)) \cdot h'(y). \end{aligned}$$

类似地，因为 $g(x)$ 是严格单调减函数，其反函数 $h(y)$ 也是严格单调减函数，所以 $h'(y) < 0$。因此，结论中 $h'(y)$ 需要绝对值符号。

笔记

例题 6.3 设随机变量 X 服从正态分布 $N(\mu, \sigma^2)$，则当 $a \neq 0$ 时，有 $Y = aX+b \sim N(a\mu+b, a^2\sigma^2)$.

解 以下我们从 $a > 0$ 和 $a < 0$ 两个方面来证明。

1. 若 $a > 0$，$y = g(x) = ax + b$ 是严格增函数。仍在 $(-\infty, +\infty)$ 上取值，其反函数 $x = h(y) = \frac{y-b}{a}$。由上述定理可知，

$$\begin{aligned} p_Y(y) &= p_X(h(y))\,|h'(y)| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(\frac{y-b}{a} - \mu)^2\right\} \cdot \frac{1}{a} \\ &= \frac{1}{\sqrt{2\pi(a\sigma)^2}}\exp\left\{-\frac{1}{2(a\sigma)^2}(y - a\mu - b)^2\right\}. \end{aligned}$$

   因此，$Y \sim N((a\mu + b, a^2\sigma^2)$。

2. 若 $a < 0$，证明结果类似，学生可以在课后进行补充。

笔记

例题 6.4 设随机变量 $X \sim N(\mu,\sigma^2)$，则 $Y=e^X$ 的概率密度函数为

$$p_Y(y)=\begin{cases}\frac{1}{\sqrt{2\pi}y\sigma}e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}} & , \quad y>0\\ 0 & , \quad y\leqslant 0\end{cases}$$

解 因为 $y=g(x)=e^x$ 是严格单调递增函数，它仅在 $(0,+\infty)$ 上取值，其反函数 $x=h(y)=lny$，而 $h'(y)=\frac{1}{y}$，根据定理可知，

$$\begin{aligned}p_Y(y) &= p_X(h(y))\,|h'(y)|\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(lny-\mu)^2\right\}\cdot\frac{1}{y}\\ &= \frac{1}{\sqrt{2\pi}y\sigma}\exp\left\{-\frac{1}{2\sigma^2}(lny-\mu)^2\right\}\end{aligned}$$

注 这个分布称为对数正态分布 $\omega(\mu,\sigma^2)$。

例题 6.5 设 $X\sim Ga(\alpha,r)$，则当 $k>0$ 时，有 $Y=kX\sim Ga(\alpha,\frac{r}{k})$。

解 因为 $k>0$，所以 $y=kx$ 是严格增函数，它仍在 $(0,+\infty)$ 上取值，其反函数 $x=\frac{y}{k}$。

1. 当 $y<0$ 时，$p_Y(y)=0$;
2. 当 $y\geq 0$ 时，我们根据上述定理有

$$\begin{aligned}p_Y(y) &= p_X\left(\frac{y}{k}\right)\cdot\frac{1}{k}\\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{y}{k}\right)^{\alpha-1}e^{-\lambda\frac{y}{k}}\cdot\frac{1}{k}\\ &= \frac{\left(\frac{\lambda}{k}\right)^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\frac{\lambda}{k}y}\end{aligned}$$

   即 $Y\sim Ga(\alpha,\frac{\lambda}{k})$。

推论 6.1

若随机变量 X 的分布函数 $F_X(x)$ 为严格单调增的连续函数，其反函数 $F_X^{-1}(y)$ 存在，则 $Y=F_X(X)$ 服从 $(0,1)$ 上的均匀分布 $U(0,1)$。

证明 首先注意到 $Y=F_X(X)$ 是一个随机变量，于是，我们需要求其分布函数。由于根据分布函数的有界性，分布函数 $F_X(x)$ 仅在 $[0,1]$ 区间上取值，故当 $y<0$ 时，因为 $\{F_X(X)\leq y\}$ 是不可能事件，所以

$$F_Y(y)=P(Y\leqslant y)=P(F_X(X)\leqslant y)=0$$

当 $y\geq 1$ 时，因为 $\{F_X(X)\leq y\}$ 是必然事件，所以

$$F_Y(y)=P(Y\leqslant y)=P(F_X(X)\leqslant y)=1.$$

当 $0 \le y < 1$ 时，有

$$\begin{aligned} F_Y(y) &= P(Y \leqslant y) \\ &= P\left(F_X(X) \leqslant y\right) \\ &= P\left(F_X^{-1}\left(F_X(X)\right) \leqslant F_X^{-1}(y)\right) \\ &= P\left(X \leqslant F_X^{-1}(y)\right) \\ &= F_X\left(F_X^{-1}(y)\right) \\ &= y \end{aligned}$$

综上所述，$Y = F_X(X)$ 的分布函数为

$$F_Y(y) = \begin{cases} 0 & , y < 0 \\ y & , 0 \le y < 1 \\ 1 & , y \geqslant 1 \end{cases}$$

因此，$Y \sim U(0,1)$。

注

1. 任一个连续随机变量 X 都可通过其分布函数 $F(x)$ 与均匀分布随机变量 U 有关联。
2. $X \sim Exp(\lambda)$，其分布函数 $F(x) = 1 - e^{-\lambda x}$。于是，
$$U = 1 - e^{-\lambda x} \Rightarrow x = \frac{1}{\lambda} \ln \frac{1}{1-u}.$$
这表明了由均匀分布 $U(0,1)$ 的随机数 u_i 可得指数分布 $Exp(\lambda)$ 的随机数 $x_i = \frac{1}{\lambda} \cdot \ln \frac{1}{1-u_i} \quad i = 1,2,\cdots,n,\cdots$，这是 Monte Carlo 法的基础。

6.2.3 $g(\cdot)$ 是其他特殊形式

例题 6.6 $X \sim N(0,1)$，求 $Y = X^2$ 的密度函数。

解 先求 Y 的分布函数 $F_Y(y)$。由于 $Y = X^2 \ge 0$，故当 $y \le 0$ 时，有 $F_Y(y) = 0$，从而 $P_Y(y) = 0$.

当 $y > 0$ 时，有

$$F_Y(y) = P(Y \leqslant y) = P\left(X^2 \leqslant y\right) = P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

因此，Y 的分布函数为

$$F_Y(y) = \begin{cases} 2\Phi(\sqrt{y}) - 1 & , y > 0 \\ 0 & , y \le 0 \end{cases}$$

再用求导的方式求出 Y 的密度函数

$$p_Y(y) = \begin{cases} \varphi(\sqrt{y}) \cdot y^{-\frac{1}{2}} & , y > 0 \\ 0 & , y \leqslant 0 \end{cases} = \begin{cases} \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} & , y > 0 \\ 0 & , y \leqslant 0 \end{cases}$$

称 Y 服从自由度为 1 的卡方分布，记 $\chi^2(1)$。

注 可以发现，这个分布也是伽马分布，即 $Ga(\frac{n}{2}, \frac{1}{2}) = \chi^2(n)$。

第 7 讲 随机向量及常见的分布

阅读章节

- ❑ Intro to Prob 2.5 3.4
- ❑ Prob & Stat 3.1 3.2

例题 7.1

1. 每个儿童的身高 $X_1(\omega)$ 和体重 $X_2(\omega)$ 构成一个二维随机变量。
2. 每个家庭的衣食住行的花费占其家庭总收入的百分比，其中 $X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega)$ 分别表示衣食住行，则 (X_1, X_2, X_3, X_4) 就是一个四维随机变量。

定义 7.1

如果 $X_1(\omega), X_2(\omega), ..., X_n(\omega)$ 是定义在同一个样本空间 $\Omega = \{\omega\}$ 上的 n 个随机变量，则称

$$\boldsymbol{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))'$$

为 n 维 (或 n 元) 随机变量或随机向量。♣

注 所有定义的随机向量都是**列向量**。

问题 7.1 如何刻画一个随机向量?

笔记

7.1 随机向量的联合分布函数

定义 7.2

对任意 n 个实数 $x_1, x_2, \cdots, x_n$，n 个事件 $\{X_1 \leqslant x_1\}, \{X_2 \leqslant x_2\}, ..., \{X_n \leqslant x_n\}$ 同时发生的概率

$$\begin{aligned} F(x_1, x_2, \cdots, x_n) &= P(\{X_1 \leqslant x_1\} \cap \{X_2 \leqslant x_2\} \cap \cdots \cap \{X_n \leqslant x_n\}) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n) \end{aligned}$$

称为 n 维随机向量 $(x_1, x_2, \cdots, x_n)'$ 的联合分布函数。♣

注 对于二维随机向量 $(X,Y)'$，其联合分布函数为

$$F(x,y)=P(X\leq x,Y\leq y)$$

定理 7.1 (二维随机变量联合分布函数的性质)

任一二维联合分布函数 $F(x,y)$ 必具有如下性质:

1. 单调性：$F(x,y)$ 分别对 x 或 y 是单调非减的，即

   当 $x_1<x_2$ 时，有
   $$F\left(x_1,y\right)\leq F\left(x_2,y\right);$$
   当 $y_1<y_2$ 时，有
   $$F\left(x,y_1\right)\leq F\left(x,y_2\right).$$
2. 有界性: 对任意的 x 和 y，有 $0\leqslant F(x,y)\leqslant 1$ 且
   $$\begin{aligned}F(-\infty,y)&=\lim_{x\to-\infty}F(x,y)=0,\\F(x,-\infty)&=\lim_{y\to-\infty}F(x,y)=0,\\F(\infty,\infty)&=\lim_{x,y\to\infty}F(x,y)=1.\end{aligned}$$
3. 右连续性：对每个变量都是右连续的，即
   $$\begin{aligned}F(x+0,y)&=F(x,y)\\F(x,y+0)&=F(x,y)\end{aligned}$$
4. 非负性：对任意的 $a<b,c<d$ 有
   $$P(a<X\leq b,c<Y\leqslant d)=F(c,d)-F(a,d)-F(b,c)+F(a,c)\geqslant 0$$

♡

注

1. 具有以上四条性质的二元函数 $F(x,y)$ 一定是某个二维随机变量的分布函数。
2. 性质 (4) 是二维随机向量所特有的，也是合理的，但性质 (4) 不能由前三条性质推出，因此需要单独列出。

例题 7.2 设二元函数为

$$G(x,y)=\begin{cases}0 & x+y<0\\1 & x+y\geqslant 0\end{cases}$$

可以证明其满足前三个性质（这里留给学生课后完成），但该函数不满足性质 4。考虑在 $(-1,-1),(1,-1),(-1,1),(1,1)$ 所围成的正方形上的面积。

笔记

7.2 边际分布函数

定义 7.3 (边际分布函数)

如果二维随机变量 (X,Y) 的联合分布函数为 $F(x,y)$，那么，称 $F_X(x)=P(X\leq x)=P(X\leq x,Y<\infty)=\lim_{y\to\infty}F(x,y)$ 为 X 的边际分布。类似地，称

$$F_Y(y)=F(\infty,y)$$

为 Y 的边际分布。♣

例题 7.3 设二维随机变量 (X,Y) 的联合分布函数为

$$F(x,y)=\begin{cases}1-e^{-x}-e^{-y}+e^{-x-y-\lambda xy}, x>0,y>0.\\ 0,\text{其他}.\end{cases}$$

这个分布被称为二维指数分布，其中参数 $\lambda>0$。求 X 和 Y 的边际分布函数。

笔记

7.3 联合分布列与边际分布列

定义 7.4

如果二维随机变量 (X,Y) 只取有限个或可列个数对 (x_i,y_i) 则称 (X,Y) 为二维离散随机变量，称

$$p_{ij}=P\left(X=x_i,Y=y_j\right)\quad i,j=1.2,\cdots$$

为 (X,Y) 的联合分布列，也可用如下表格形式记联合分布列。♣

性质 联合分布列的基本性质：

1. 非负性：$p_{ij}\geq 0$;
2. 正则性：$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}p_{ij}=1$.

定义 7.5 (边际分布列)

如果二维随机便利 (X,Y) 的联合分布列 $\{P(X=x_i,Y=y_j)\}$ 中，称

$$P(X=x_i)=\sum_{j=1}^{\infty}P(X=x_i,Y=y_j),i=1,2,\cdots$$

为 X 的边际分布列。类似地，称

$$P(Y=y_j)=\sum_{i=1}^{\infty}P(X=x_i,Y=y_j),j=1,2,\cdots$$

为 Y 的边际分布列。♣

注 我们通常采用表格的形式来表示联合分布列和边际分布列，即

X	Y					
	y_1	y_2	$\cdots$	y_j	$\cdots$	
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	$p_{1\cdot}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	

1. 联合分布列：$p_{ij}=P(X=x_i,Y=y_j)$;
2. 边际分布列：$p_{i\cdot}=P(X=x_i)$，$p_{\cdot j}=P(Y=y_j)$。

7.3.1 联合密度函数

定义 7.6

如果存在二元非负函数 $p(x,y)$，使得二维随机变量 (X,Y) 的分布函数 $F(x,y)$ 可表示为

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}p(u,v)\mathrm{d}v\mathrm{d}u$$

则称 (X,Y) 为二维连续随机变量，称 $p(x,y)$ 为 (X,Y) 的联合密度函数。♣

注

1. 在 $F(X,Y)$ 偏导数存在的点上有
$$p(x,y)=\frac{\partial^2}{\partial x\partial y}F(x,y).$$
2. 给定联合密度函数 $p(x,y)$，若 G 为平面上的一个区域，则事件 $\{(X,Y)\in G\}$ 的概率可以表示为 G 上对 $p(x,y)$ 的二重积分
$$P((X,Y)\in G)=\iint_G p(x,y)\mathrm{d}y\mathrm{d}x.$$

性质 联合密度函数的基本性质：

1. 非负性：$p(x,y) \geq 0$；
2. 正则性：$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\mathrm{d}y\mathrm{d}x = 1$。

定义 7.7 (边际密度函数)

如果二维连续随机变量 (X,Y) 的联合密度函数为 $p(x,y)$，称

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

为 X 的边际密度函数。类似地，称

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

为 Y 的边际密度函数。

例题 7.4 设二维随机变量 (X,Y) 的联合密度函数为

$$p(x,y) = \begin{cases} 1, & 0 < x < 1, |y| < x \\ 0, & \text{其他}. \end{cases}$$

求：

1. 边际密度函数 $p_X(x)$ 和 $p_Y(y)$；
2. $P(X < 1/2)$ 和 $P(Y > 1/2)$。

笔记

7.4 随机变量的独立性

笔记

定义 7.8 (独立性)

设 n 维随机变量 $(X_1, X_2, \cdots, X_n)'$ 的联合分布函数 $F(x_1, x_2, \cdots, x_n)$，$F_i(x_i)$ 为 X_i 的边际分布函数。如果对任意 n 个实数 $x_1, x_2, \cdots, x_n$ 有

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_i(x_i)$$

则称 $X_1, X_2, \cdots, X_n$ 相互独立。

1. 在离散随机变量场合，如果对其任意 n 个取值 $x_1, x_2, \cdots, x_n$ 有

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

   则称 $X_1, X_2, \cdots, X_n$ 相互独立。

2. 在连续随机变量场合，如果对其任意 n 个实数 $x_1, x_2, \cdots, x_n$，有

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p_i(x_i)$$

   则称 $X_1, X_2, \cdots, X_n$ 相互独立。

♣

例题 7.5 若二维随机变量 (X, Y) 的联合密度函数为

$$p(x,y) = \begin{cases} 8xy, & 0 \le x \le y \le 1 \\ 0, & \text{其他} \end{cases}$$

问 X 和 Y 是否相互独立?

笔记

7.5 常见的多维随机变量的分布

7.5.1 多项分布

定义 7.9

进行 n 次独立重复实验，如果每次实验有 r 个互不相容的结果：$A_1, A_2, \cdots, A_r$ 之一发生，且每次是试验中 A_i 发生的概率为 $p_i = P(A_i), i = 1, 2, \cdots, r$，且 $p_1+p_2+\cdots+p_r = 1$。记 X_i 为 n 次独立重复试验中 A_i 出现的次数，$i = 1, 2, \cdots, r$，则 $(X_1, X_2, \cdots, X_r)$ 取值 $(x_1, x_2, \cdots, x_r)$ 的概率，即 A_1 出现 x_1 次，A_2 出现 x_2 次,$\cdots$,A_r 出现 x_r 的概率为

$$P\left(X_1 = x_1, X_2 = x_2, \cdots, X_r = x_r\right) = \frac{n!}{x_1!x_2!\cdots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}$$

其中 $n = n_1 + n_2 + \cdots + n_r$。称这个联合分布列为多项分布，又称为 r 项分布。记 $M(n, p_1, p_2, \cdots, p_r)$。♣

注

1. 典型例子：投掷 r 面骰子。
2. 当 $r = 2$ 时，即为二项分布。
3. r 项分布是 $r-1$ 维随机变量的分布。

接下来，我们用一个例子来讨论多项分布与二项分布之间的关系。

例题 7.6 考虑三项分布 $M(n, p_1, p_2, p_3)$ 实质上是一个二维随机变量 (X, Y) 的分布，其联合分布为

$$P(X = i, Y = j) = \frac{n!}{i!j!(n-i-j)!} p_1^i p_2^j \left(1 - p_1 - p_2\right)^{n-i-j}, \begin{cases} i, j = 0, 1, \cdots, n \\ i + j \leq n \end{cases}$$

于是，X 的边际分布为

$$\begin{aligned} P(X = i) &= \sum_{j=0}^{n-i} \frac{n!}{i!j!(n-i-j)!} p_1^i p_2^j \left(1 - p_1 - p_2\right)^{n-i-j} \\ &= \frac{n!}{i!(n-i)!} p_1^i \left(1 - p_1\right)^{n-i} \sum_{j=0}^{n-i} \frac{(n-i)!}{j!(n-i-j)!} \frac{p_2^j \left(1 - p_1 - p_2\right)^{n-i-j}}{\left(1 - p_1\right)^{n-i}} \\ &= \frac{n!}{i!(n-i)!} p_1^i \left(1 - p_1\right)^{n-i} \sum_{j=0}^{n-i} \frac{(n-i)!}{j!(n-i-j)!} \left(p_2^*\right)^j \left(1 - p_2^*\right)^{n-i-j} \\ &= \frac{n!}{i!(n-i)!} p_1^i \left(1 - p_1\right)^{n-i}, \end{aligned}$$

其中，$p_2^* = \frac{p_2}{1-p_1}$。因此，$X \sim b\left(n, p_1\right)$。

注

1. 三项分布的一维边际分布为二项分布；
2. 多项分布的一维边际分布为二项分布；

7.5.2 多维均匀分布

定义 7.10

设 D 为 R^n 中的一个有界区域，其度量为 S_D。如果多维随机变量 $(X_1, X_2, \cdots, X_n)'$ 的联合密度函数为

$$p(x_1, x_2, \cdots, x_n) = \begin{cases} \frac{1}{S_D}, & (x_1, x_2, \cdots, x_n) \in D \\ 0, & \text{其他} \end{cases}$$

则称 $X_1, X_2, \cdots, X_n$ 服从 D 上的多维均匀分布，记为 $X_1, X_2, \cdots, X_n \sim U(D)$。♣

注 二维均匀分布所描述的随机现象就是向平面区域 D 中随机投点。如果该点坐标 (X, Y) 落在 D 的子区域 G 中概率只与 G 的面积有关，而与 G 的位置无关，则

$$P((X,Y) \in G) = \iint_G p(x,y)dxdy = \iint_G \frac{1}{S_D}dxdy = \frac{S_G}{S_D}$$

7.5.3 多维正态分布

定义 7.11

若 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)'$ 为一个 n 维随机变量，其密度函数为

$$p(x_1, x_2, \cdots, x_n) = p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

称 $\boldsymbol{X}$ 满足 n 元正态分布，记 $X \sim N_n(\boldsymbol{\mu}, \Sigma)$。♣

注

1. 当 $n=1$ 时，$\boldsymbol{\mu} = \mu_1$，$\Sigma = \sigma_1^2$，一元正态分布的密度函数为

$$p(x) = (2\pi\sigma_1^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right\}$$

2. 当 $n=2$ 时，$\boldsymbol{\mu} = (\mu_1, \mu_2)'$ 且

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

所以，Σ 的行列式为

$$|\Sigma| = \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)^2 = (1-\rho^2)\sigma_1^2\sigma_2^2$$

而它的逆矩阵为

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{|\Sigma|}\Sigma^* \\ &= \frac{1}{|\Sigma|}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \end{aligned}$$

其中，Σ^* 是 Σ 的伴随矩阵。所以，二元正态分布的密度函数为

$$
\begin{aligned}
p(\boldsymbol{x}) &= p(x_1,x_2) = (2\pi)^{-\frac{2}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\
&= (2\pi)^{-1}\left((1-\rho^2)\sigma_1^2\sigma_2^2\right)^{-1/2} \\
&\quad \cdot\exp\left\{-\frac{1}{2}((x_1,x_2)'-(\mu_1,\mu_2)')'\frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2}\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}((x_1,x_2)'-(\mu_1,\mu_2)')\right\} \\
&= (2\pi)^{-1}\left((1-\rho^2)\sigma_1^2\sigma_2^2\right)^{-1/2} \\
&\quad \cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\cdot\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}
\end{aligned}
$$

例题 7.7 二维正态分布的边际分布为一元分布。

解 $(X_1,X_2)'$ 的联合密度函数为

$$
\begin{aligned}
p(x_1,x_2) &= (2\pi)^{-1}\left((1-\rho^2)\sigma_1^2\sigma_2^2\right)^{-1/2} \\
&\quad \cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\cdot\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}
\end{aligned}
$$

于是，X 的边际密度函数为

$$
\begin{aligned}
p_{X_1}(x_1) &= \int_{-\infty}^{+\infty} p(x_1,x_2)dx_2 \\
&= \int_{-\infty}^{+\infty}(2\pi)^{-1}\left((1-\rho^2)\sigma_1^2\sigma_2^2\right)^{-1/2} \\
&\quad \cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\cdot\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}\mathrm{d}x_2 \\
&= \int_{-\infty}^{+\infty}(2\pi)^{-1/2}\left((1-\rho^2)\sigma_2^2\right)^{-1/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\cdot\left(\frac{(x_2-\mu_2)}{\sigma_2}-\rho\frac{(x_1-\mu_1)}{\sigma_1}\right)^2\right\}\mathrm{d}x_2 \\
&\quad \cdot(2\pi\sigma_1^2)^{-1/2}\cdot\exp\left\{(1-\rho^2)\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right\} \\
&= (2\pi\sigma_1^2)^{-1/2}\cdot\exp\left\{(1-\rho^2)\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right\}
\end{aligned}
$$

其中，第三个等式成立的原因是

$$
\begin{aligned}
&\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\cdot\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_1-\mu_1)^2}{\sigma_1^2} \\
=\ &\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\cdot\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\rho^2\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\rho^2\frac{(x_1-\mu_1)^2}{\sigma_1^2} \\
=\ &\left(\frac{(x_2-\mu_2)}{\sigma_2}-\rho\frac{(x_1-\mu_1)}{\sigma_1}\right)^2+(1-\rho^2)\frac{(x_1-\mu_1)^2}{\sigma_1^2}.
\end{aligned}
$$

所以，X 的边际分布为 $N(\mu_1,\sigma_1^2)$。

第 8 讲 多维随机变量的特征数

阅读章节

- ❏ Intro to Prob 2.5 3.4 4.2
- ❏ Prob&Stat 3.4

8.1 多维随机变量函数的数学期望

笔记

定理 8.1

若二维随机变量 (X,Y) 的分布用联合分布列 $P(X=x_i,Y=y_j)$ 或用联合密度函数 $p(x,y)$ 表示，则 $Z=g(X,Y)$ 的数学期望为

$$E(Z)=\begin{cases}\sum_i\sum_j g\left(x_i,y_j\right)P\left(X=x_i,Y=y_j\right), & \text{在离散场合}\\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)P(x,y)\mathrm{d}x\mathrm{d}y & \text{在连续场合}\end{cases}$$

注

1. 当 $g(X,Y)=X$ 时，X 的数学期望为
$$\begin{aligned}E(X)&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xp(x,y)\mathrm{d}y\mathrm{d}x\\&=\int_{-\infty}^{+\infty}xp_X(x)\mathrm{d}x\end{aligned}$$
2. 当 $g(X,Y)=(X-EX)^2$ 时，X 的方差为
$$\begin{aligned}Var(X)&=E(X-EX)^2\\&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(X))^2p(x,y)\mathrm{d}y\mathrm{d}x\\&=\int_{-\infty}^{+\infty}(x-E(X))^2p_X(x)\mathrm{d}x\end{aligned}$$

3. 当 $g(X,Y)=Y$ 时，Y 的数学期望为

$$
\begin{aligned}
E(Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yp(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} yp_Y(y)\mathrm{d}y
\end{aligned}
$$

4. 当 $g(X,Y)=(Y-EY)^2$ 时，Y 的方差为

$$
\begin{aligned}
Var(Y) &= E(Y-EY)^2 \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y-E(Y))^2 p(x,y)\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} (y-E(Y))^2 p_Y(y)\mathrm{d}y
\end{aligned}
$$

例题 8.1 在长度为 a 的线段上任取两个点 X 与 Y，求此两点间的平均长度。

解 因为 X 与 Y 均服从 $(0,a)$ 上的均匀分布且 X 与 Y 相互独立，所以 (X,Y) 的联合密度函数为

$$
p(x,y)=\begin{cases}\dfrac{1}{a^2} & ,0<x,y<a \\ 0 & ,\text{其他}\end{cases}
$$

两点间的平均长度为

$$
\begin{aligned}
E(|X-Y|) &= \int_0^a\int_0^a |x-y|\cdot\frac{1}{a^2}\mathrm{d}x\mathrm{d}y \\
&= \frac{1}{a^2}\left(\int_0^a\int_0^x (x-y)\mathrm{d}y\mathrm{d}x+\int_0^a\int_x^a (y-x)\mathrm{d}y\mathrm{d}x\right) \\
&= \frac{1}{a^2}\left(\int_0^a \left(xy-\frac{1}{2}y^2\right)\Big|_0^x \mathrm{d}x+\int_0^a \left(\frac{1}{2}y^2-xy\right)\Big|_x^a \mathrm{d}x\right) \\
&= \frac{1}{a^2}\int_0^a \left(\frac{1}{2}x^2+\frac{1}{2}a^2-ax+\frac{1}{2}x^2\right)\mathrm{d}x \\
&= \frac{1}{a^2}\left(\frac{1}{2}a^2x-\frac{1}{2}ax^2+\frac{1}{3}x^3\Big|_0^a\right) \\
&= \frac{1}{a^2}\left(\frac{1}{2}a^3-\frac{1}{2}a^3+\frac{1}{3}a^3\right) \\
&= \frac{a}{3}.
\end{aligned}
$$

以下我们介绍一些期望和方差的性质。

性质

1. 期望与求和可交换：设 (X,Y) 是二维随机变量，则有

$$E(X+Y)=E(X)+E(Y).$$

注

$$E\left(\sum_{i=1}^n X_i\right)=\sum_{i=1}^n E(X_i)$$

2. 多个独立随机变量的期望与方差的简便计算公式。

(a). 若随机变量 X 和 Y 相互独立，则有

$$E(XY)=E(X)E(Y).$$

证明 这里仅证明 (X,Y) 为连续随机变量，离散随机变量的证明过程供学生课后自行完成。其密度函数为 $p(x,y)$。因为 X 和 Y 相互独立，所以 $p(x,y)=p_X(x)p_Y(y)$。令 $g(X,Y)=XY$，则有

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp_X(x)p_Y(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} xp_X(x)\mathrm{d}x\cdot\int_{-\infty}^{\infty} yp_Y(y)\mathrm{d}y \\ &= E(X)E(Y). \end{aligned}$$

注 若 $X_1,X_2,\cdots,X_n$ 相互独立，则有

$$E\left(\prod_{i=1}^{n}X_i\right)=\prod_{i=1}^{n}E(X_i).$$

(b). 若随机变量 X 与 Y 相互独立，则有

$$Var(X\pm Y)=Var(X)+Var(Y).$$

证明 根据定义，

$$\begin{aligned} Var(X\pm Y) &= E(X\pm Y-(E(X)\pm E(Y)))^2 \\ &= E((X-E(X))\pm(Y-E(Y)))^2 \\ &= E((X-E(X))^2+(Y-E(Y))^2\pm 2(X-E(X))(Y-E(Y))) \\ &= E(X-E(X))^2+E(Y-E(Y))^2\pm 2E((X-E(X))(Y-E(Y))) \\ &= E(X-E(X))^2+E(Y-E(Y))^2 \\ &= Var(X)+Var(Y), \end{aligned}$$

最后一个等式成立是因为 $E((X-E(X))(Y-E(Y)))=E(XY)-E(X)E(Y)=0$.

注

I. 若 $X_1,X_2,\cdots,X_n$ 相互独立，则有

$$Var\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}Var(X_i).$$

II. 若 $X_1,X_2,\cdots,X_n$ 是独立同分布的随机变量，且方差存在，$Var(X_1)=\sigma^2$，则其算术平均数的方差为

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}.$$

例题 8.2 设一袋中装有 m 个颜色各不相同的球，每次从中任取一个，有放回地摸取 n 次，以 X 表示在 n 次摸球中摸到球的不同颜色的数目，求 $E(X)$。

解 令 X_i 表示是否第 i 种颜色的球在 n 次摸球中至少被摸到一次，即

$$X_i = \begin{cases} 1, & \text{摸到过}, \\ 0, & \text{未摸到}, \end{cases} \quad i = 1, 2, \cdots, m.$$

于是，第 i 种颜色的球在 n 次摸球中未被摸到过的概率为

$$P(X_i = 0) = \left(1 - \frac{1}{m}\right)^n.$$

令 X 为 n 次摸球中不同颜色的数目，则 $X = \sum_{i=1}^m X_i$。于是，

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^m X_i\right) \\ &= \sum_{i=1}^m E(X_i) \\ &= \sum_{i=1}^m P(X_i = 1) \\ &= \sum_{i=1}^m \left(1 - \left(1 - \frac{1}{m}\right)^n\right) \\ &= m\left(1 - \left(1 - \frac{1}{m}\right)^n\right). \end{aligned}$$

笔记

8.2 协方差与相关系数

8.2.1 协方差

在考虑多个随机变量时，每一个随机变量的期望与方差是我们关注的特征数之外，两个随机变量之间的关系也是我们关心的一个特征数。以下我们介绍特征数——协方差。

定义 8.1 (协方差)

设 (X,Y) 是一个二维随机变量，若

$$E\left((X-E(X))(Y-E(Y))\right)$$

存在，则称其为 X 与 Y 的协方差或相关（中心）矩，并记为 $Cov(X,Y) = E\left((X-E(X))(Y-E(Y))\right)$ ♣

注

1. 特别地，$Cov(X,X) = Var(X)$；
2. 通过协方差可以判断两个随机变量之间的关系，即
   - $Cov(X,Y) > 0$ 时，称 X 与 Y 正相关
   - $Cov(X,Y) < 0$ 时，称 X 与 Y 负相关
   - $Cov(X,Y) = 0$ 时，称 X 与 Y 不相关（毫无关联/非线性关系）

以下介绍协方差的一些性质。

性质

1. $Cov(X,Y) = E(XY) - E(X)E(Y)$；
2. X 与 Y 独立 $\Rightarrow Cov(X,Y) = 0$，反之不然。注 令 $X \sim N(0,\sigma^2), Y = X^2$。我们可以计算
$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) = 0.$$
3. $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$。注 对任意 n 个随机变量 $X_1, X_2, \cdots, X_n$，有
$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} Cov\left(X_i, X_j\right).$$
4. 若 X 与 Y 不相关，则
$$E(XY) = E(X)E(Y), \quad 且 \quad Var(X \pm Y) = Var(X) + Var(Y).$$
5. 协方差的计算与次序无关，即 $Cov(X,Y) = Cov(Y,X)$。
6. 任意随机变量 X 与常数 a 的协方差为零，即 $Cov(X,a) = 0$。
7. 对任意常数 a,b，有 $Cov(aX,bY) = ab \cdot Cov(X,Y)$。
8. 对于任意三个随机变量 X,Y,Z，有
$$Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$$

8.2.2 相关系数

定义 8.2 (相关系数)
设 (X,Y) 是一个二维随机变量，且 $Var(X)=\sigma_X^2>0, Var(Y)=\sigma_Y^2>0$，则称

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

为 X 与 Y 的（线性）相关系数。♣

注 从另一个角度来看相关系数，令 $E(X)=\mu_X$，$E(Y)=\mu_Y$。将原始的随机变量进行标准化，即

$$X^*=\frac{X-\mu_X}{\sigma_X},\quad Y^*=\frac{Y-\mu_Y}{\sigma_Y}.$$

经过标准化后的两个随机变量的协方差为

$$\begin{aligned}Cov(X^*,Y^*) &= Cov\left(\frac{X-\mu_X}{\sigma_X},\frac{Y-\mu_Y}{\sigma_Y}\right)\\ &= \frac{1}{\sigma_X\sigma_Y}Cov(X,Y)\\ &= Corr(X,Y).\end{aligned}$$

例题 8.3 证明二维正态分布 $(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho)$ 的相关系数为 ρ。

解 首先，计算协方差

$$\begin{aligned}Cov(X,Y) &= E((X-E(X))(Y-E(Y)))\\ &= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho^2)}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)\\ &\quad\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_1\sigma_2}+\frac{(y-\mu_Y)^2}{\sigma_2^2}\right]\right\}\mathrm{d}x\mathrm{d}y\\ &= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho^2)}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)\cdot\\ &\quad\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}-\rho\frac{y-\mu_Y}{\sigma_Y}\right)^2+\left(\sqrt{1-\rho^2}\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}\mathrm{d}x\mathrm{d}y\end{aligned}$$

令

$$u=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{x-\mu_X}{\sigma_X}-\rho\frac{(y-\mu_Y)}{\sigma_Y}\right),\quad 且\quad v=\frac{y-\mu_Y}{\sigma_Y}$$

则

$$\begin{cases}x-\mu_X=\sigma_X\left(u\sqrt{1-\rho^2}+\rho v\right),\\ y-\mu_Y=\sigma_Y v,\end{cases}\quad 且\quad |J|=\begin{vmatrix}\sqrt{1-\rho^2}\sigma_X & \rho\sigma_Y\\ 0 & \sigma_Y\end{vmatrix}=\sqrt{1-\rho^2}\sigma_X\sigma_Y$$

$$\begin{aligned} Cov\left(X,Y\right) &= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2\left(1-\rho^2\right)}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(\sigma_X\left(u\sqrt{1-\rho^2}+\rho v\right)\right)\left(\sigma_Y v\right) \\ &\quad \cdot \exp\left\{-\frac{1}{2}\left[u^2+v^2\right]\right\}\cdot\sqrt{1-\rho^2}\cdot\sigma_X\sigma_Y \mathrm{d}u\mathrm{d}v \\ &= \frac{\sigma_X\sigma_Y}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(uv\sqrt{1-\rho^2}+\rho v^2\right)\exp\left\{-\frac{1}{2}\left(u^2+v^2\right)\right\}\mathrm{d}u\mathrm{d}v \end{aligned}$$

注意到

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} uv\exp\left\{-\frac{1}{2}\left(u^2+v^2\right)\right\}\mathrm{d}u\mathrm{d}v = \int_{-\infty}^{+\infty} u\exp\left\{-\frac{1}{2}u^2\right\}\mathrm{d}u\cdot\int_{-\infty}^{+\infty} v\exp\left\{-\frac{1}{2}v^2\right\}\mathrm{d}v = 0$$

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} v^2\exp\left\{-\frac{1}{2}\left(u^2+v^2\right)\right\}\mathrm{d}u\mathrm{d}v &= \int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2}\left(u^2\right)\right\}\mathrm{d}u\cdot\int_{-\infty}^{+\infty} v^2\exp\left\{-\frac{1}{2}v^2\right\}\mathrm{d}v \\ &= 2\pi\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}u^2\right\}\mathrm{d}u\int_{-\infty}^{+\infty} v^2\cdot\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}v^2\right\}\mathrm{d}v = 1\cdot E\left(v^2\right)\cdot 2\pi = 2\pi. \end{aligned}$$

则

$$Cov(X,Y) = \frac{\sigma_X\sigma_Y}{2\pi}\cdot\rho\cdot 2\pi = \rho\sigma_X\sigma_Y$$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = \rho.$$

定理 8.2 (Schwarz 不等式)

对任意二维随机变量 (X,Y)，若 X 与 Y 的方差都存在，且记 $\sigma_X^2 = Var(X), \sigma_Y^2 = Var(Y)$，则有

$$(Cov(X,Y))^2 \le \sigma_X^2\sigma_Y^2.$$

证明 不妨设 $\sigma_X^2 > 0$，设函数

$$\begin{aligned} g(t) &= E(t(X-E(X))+(Y-E(Y)))^2 \\ &= t^2\sigma_X^2 + 2tCov\left(X,Y\right)+\sigma_Y^2. \end{aligned}$$

因为 $g(t)$ 是一个非负随机变量的期望，所以，$g(t)$ 恒非负。而且，这个二次函数的开口向上，其只有一个或零个零根。所以，其判别式小于或等于零，即

$$(2Cov(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leqslant 0.$$

也就是说，

$$(Cov(X,Y))^2 \leqslant Var(X)Var(Y).$$

性质

1. $|Corr(X,Y)| \le 1$；
2. $Corr(X,Y) = \pm 1 \Leftrightarrow X$ 与 Y 之间几乎处处有线性关系，即存在 $a \neq 0$ 与 b，使得

   $$P(Y = aX+b) = 1$$

   其中，当 $Corr(X,Y) = 1$ 时，有 $a > 0$；当 $Corr(X,Y) = -1$ 时，有 $a < 0$。

   注

- $Corr(X,Y)=0$，则称 X 与 Y 不相关；
- $Corr(X,Y)=+1$，则称 X 与 Y 完全正相关；
- $Corr(X,Y)=-1$，则称 X 与 Y 完全负相关；
- $0<|Corr(X,Y)|<1$，则称 X 与 Y 有“一定程度”的线性关系。

3. 在二维正态分布 $N\left(\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho\right)$ 场合，不相关与独立是等价的。

8.3 期望向量与协方差矩阵

定义 8.3

记 n 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)'$，若其每个分量的数学期望都存在，则称

$$E(\boldsymbol{X})=\left(E\left(X_1\right),E\left(X_2\right),\cdots,E\left(X_n\right)\right)'$$

为 n 维随机向量的数学期望向量，而称

$$E(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))'=\begin{pmatrix} Var\left(X_1\right) & Cov\left(X_1,X_2\right) & \cdots & Cov\left(X_1,X_n\right) \\ Cov\left(X_2,X_1\right) & Var\left(X_2\right) & \cdots & Cov\left(X_2,X_n\right) \\ \vdots & \vdots & & \vdots \\ Cov\left(X_n,X_1\right) & Cov\left(X_n,X_2\right) & \cdots & Var\left(X_n\right) \end{pmatrix}$$

为该随机向量的方差—协方差矩阵，记为 $Cov(\vec{X})$

定理 8.3

n 维随机向量的协方差矩阵 $Cov(\boldsymbol{X})=\left\{Cov\left(X_i,X_j\right)\right\}_{n\times n}$ 是一个对称非负定矩阵。

第 9 讲 多维随机变量函数的分布

阅读章节

❑ Intro to Prob 4.1 4.3 4.5 ❑ Prob & Stat 3.3

问题 9.1 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)'$ 为 n 维随机变量，且已知其联合分布的信息。已知多元函数 $\boldsymbol{g}:R^n\to R^k$。则 $\boldsymbol{Y}=\boldsymbol{g}(X_1,X_2,\cdots,X_n)$ 也是一个 k 维随机变量。求 $\boldsymbol{Y}$ 的联合分布是什么?

这里我们将介绍 g 三种的常见函数形式。

9.1 可加性（卷积公式）

“加法”是一种最为常见的运算方法。对多个随机变量进行求和是实际场景中具有广泛应用。例如：我们考虑学生每个月在衣、食、行及其他的预算，其中每一项均可以看作一个随机变量，而这四个随机变量的和可以表示该生每个月的消费预算，具有实际意义。因此，如何求多个随机变量的和的分布这里将会给出一般的计算方法。首先，我们先看一个例子。

例题 9.1 有两个服从泊松分布且相互独立的随机变量，其和仍旧服从泊松分布。即设 $X\sim P(\lambda_1),Y\sim P(\lambda_2)$ 且 X 与 Y 独立，则 $X+Y\sim P(\lambda_1+\lambda_2)$。

证明 令这两个随机变量的和为 Z，即 $Z=X+Y$。可以注意到 Z 的取值范围为 $0,1,2,\cdots$，即所有非负整数。而事件 $Z=k$ 可看作一些互不相容事件的并，即，$\{Z=k\}=\cup_{i=0}^{k}\{X=i,Y=k-i\}$。于是，对于任意非负整数 k，有

$$\begin{aligned}
P(Z=k) &= P\left(\cup_{i=0}^{k}\{X=i,Y=k-i\}\right)\\
&= \sum_{i=0}^{k}P\left(X=i,Y=k-i\right)\\
&= \sum_{i=0}^{k}P(X=i)\cdot P(Y=k-i)\\
&= \sum_{i=0}^{k}\frac{\lambda_1^i}{i!}e^{-\lambda_1}\cdot\frac{\lambda_2^{k-i}}{(k-i)!}e^{-\lambda_2}\\
&= \frac{(\lambda_1+\lambda_2)^k}{k!}e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^i\cdot\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{k-i}\\
&= \frac{(\lambda_1+\lambda_2)^k}{k!}e^{-(\lambda_1+\lambda_2)}
\end{aligned}$$

这表明 $X+Y\sim P(\lambda_1+\lambda_2)$。

注

1. 上述这个性质也可以推广到有限个独立的泊松分布的和。

2. 上述例子表明：两个独立且服从同一类随机变量的和仍服从该分布。由此，我们需要用一个专有名词来表述分布所独有的一种性质。

定义 9.1 (可加性)

如果某个分布满足服从该分布的多个独立随机变量的和仍服从该类分布，那么称该分布具有可加性。

对于更为一般的求两个随机变量和的形式，我们这里提供以下定理来解决。

定理 9.1 (卷积公式)

设 X 与 Y 是两个相互独立的连续随机变量，其密度函数分别为 $p_X(x)$ 和 $p_Y(y)$，则其和 $Z = X + Y$ 的密度函数为

$$
\begin{aligned}
p_Z(z) &= \int_{-\infty}^{\infty} p_X(z-y)p_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\mathrm{d}x
\end{aligned}
$$

证明 考虑到 $Z = X + Y$ 的分布函数为

$$
\begin{aligned}
F_Z(z) &= P(Z \le z) = \iint_{x+y\le z} p_X(x)p_Y(y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-y} p_X(x)\mathrm{d}x\right)p_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} p_X(t-y)\mathrm{d}t\right)p_Y(y)\mathrm{d}y \\
&= \int_{-\infty}^{z}\left(\int_{-\infty}^{\infty} p_X(t-y)p_Y(y)\mathrm{d}y\right)\mathrm{d}t
\end{aligned}
$$

由此可得，Z 的密度函数为

$$
p_Z(z) = \int_{-\infty}^{\infty} p_X(z-y)p_Y(y)\mathrm{d}y.
$$

令 $z - y = x$，则可得

$$
p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\mathrm{d}x.
$$

注

1. **卷积**名称的由来：在泛函分析中，卷积指的是通过两个可积函数 f 和 g，利用积分运算来生成一个新的函数，即

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

2. 在上述定理中假定两个随机变量 X 和 Y 是相互独立的。但是，对于不独立的 X 和 Y，只需要把边际密度函数的乘积改为联合密度函数即可。
3. 上述定理给出的是连续场合的卷积公式，在离散场合同样适用。值得注意的是，连续场合下的密度函数应被替换为分布列，而求积分运算应被替换为求和。

根据卷积公式，我们应用于两个例子中。

例题 9.2 正态分布的可加性 设随机变量 $X \sim N\left(\mu_1, \sigma_1^2\right), Y \sim N\left(\mu_2, \sigma_2^2\right)$, 且 X 与 Y 独立。证明

$$Z = X + Y \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

证明 首先，$Z = X + Y$ 仍在 $(-\infty, +\infty)$ 上取值。其次，利用连续场合下的卷积公式，可得

$$p_Z(z) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2\sigma_1^2}\left(z - y - \mu_1\right)^2\right\} \exp\left\{-\frac{1}{2\sigma_2^2}\left(y - \mu_2\right)^2\right\} \mathrm{d}y.$$

注意到

$$\begin{aligned}
& \frac{1}{\sigma_1^2}\left(z - y - \mu_1\right)^2 + \frac{1}{\sigma_2^2}\left(y - \mu_2\right)^2 \\
= & \frac{1}{\sigma_1^2} y^2 - \frac{2}{\sigma_1^2}\left(z - \mu_1\right) y + \frac{1}{\sigma_1^2}\left(z - \mu_1\right)^2 + \frac{1}{\sigma_1^2} y^2 - \frac{2}{\sigma_2^2}\mu_2 z + \frac{1}{\sigma_2^2}\mu_2^2 \\
= & \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) y^2 - 2y\left(\frac{z - \mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right) + \frac{1}{\sigma_1^2}\left(z - \mu_1\right)^2 + \frac{1}{\sigma_2^2}\mu_2^2 \\
= & \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\left(y - \frac{\frac{z-\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)^2 - \frac{\left(\frac{z-\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)^2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} + \frac{1}{\sigma_2^2}\left(z - \mu_1^2\right) + \frac{1}{\sigma_2^2}\mu_2^2 \\
= & A\left(y - \frac{B}{A}\right)^2 - \frac{\frac{\sigma_2^2}{\sigma_1^2}}{\sigma_1^2 + \sigma_2^2}\left(z - \mu_1\right)^2 - \frac{\frac{\sigma_1^2}{\sigma_2^2}}{\sigma_1^2 + \sigma_2^2}\mu_2^2 - \frac{2}{\sigma_1^2 + \sigma_2^2} \cdot \left(z - \mu_1\right)\mu_2 + \frac{1}{\sigma_1^2}\left(z - \mu_1\right)^2 + \frac{1}{\sigma_2^2}\mu_2^2 \\
= & A\left(y - \frac{B}{A}\right)^2 + \frac{1}{\sigma_1^2 + \sigma_2^2}\left(z - \mu_1 - \mu_2\right)^2,
\end{aligned}$$

其中，

$$A = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \quad 且 \quad B = \frac{z - \mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}.$$

于是，我们有

$$\begin{aligned}
p_Z(z) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}\left(z - y - \mu_1\right)^2 - \frac{1}{\sigma_2^2}\left(y - \mu_2\right)^2\right)\right\} \mathrm{d}y \\
&= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} \exp\left\{-\frac{1}{2} \cdot \frac{\left(z - \mu_1 - \mu_2\right)^2}{\sigma_1^2 + \sigma_2^2}\right\} \cdot \int_{-\infty}^{+\infty} \exp\left\{-\frac{A}{2}\left(y - \frac{B}{A}\right)^2\right\} \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left\{-\frac{1}{2} \cdot \frac{\left(z - \mu_1 - \mu_2\right)^2}{\sigma_1^2 + \sigma_2^2}\right\}
\end{aligned}$$

其中 $\int_{-\infty}^{+\infty} \exp\left\{-\frac{A}{2}\left(y - \frac{B}{A}\right)^2\right\} \mathrm{d}y$ 看成一个正态随机变量的核，即 $N\left(\frac{B}{A}, \frac{1}{A}\right)$。

因此，Z 的密度函数为

$$P_Z(z) = \frac{1}{\sqrt{2\pi\left(\sigma_1^2 + \sigma_2^2\right)}} \exp\left\{-\frac{1}{2} \cdot \frac{\left(z - \mu_1 - \mu_2\right)^2}{\sigma_1^2 + \sigma_2^2}\right\}.$$

也就是说，$Z \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$。

注 这个结论也可推广至有限个独立的正态分布随机变量的线性组合。也就是说，若 $X_i \sim$

$N(\mu_i,\sigma_i^2), i=1,2,\cdots,n$，诸 X_i 之间相互独立且 $a_1,a_2,\cdots,a_n$ 为 n 个非零常数，则

$$\sum_{i=1}^{n} a_i X_i = a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim N(\mu_0,\sigma_0^2),$$

其中，$\mu_0=\sum_{i=1}^{n} a_i\mu_i$，$\sigma_0^2=\sum_{i=1}^{n} a_i^2\sigma_0^2$。

问题 9.2 如果 $(X_1,X_2)'$ 服从二元正态分布随机变量，且 $corr(X_1,X_2)=\rho\neq 0$，那么 $X-Y$ 的分布是什么?

笔记 留给学生课后思考。

例题 9.3 设随机变量 $X\sim G_a(\alpha_1,\lambda), Y\sim G_a(\alpha_2,\lambda)$ 且 X 与 Y 独立。证明

$$X+Y\sim G_a(\alpha_1+\alpha_2,\lambda).$$

证明 首先，$Z=X+Y$ 仍在 $(0,+\infty)$ 上取值，所以当 $z\leqslant 0$ 时，$p_Z(z)=0$。而当 $z>0$ 时，

$$\begin{aligned}
p_z(z) &= \int_0^z \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)P(\alpha_2)}(z-y)^{\alpha_1-1}e^{-\lambda(\lambda-y)}y^{\alpha_2-1}e^{-\lambda y}\mathrm{d}y \\
&= \frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^z (z-y)^{\alpha_1-1}y^{\alpha_2-1}\mathrm{d}y \\
&= \frac{\lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^z \left(1-\frac{y}{z}\right)^{\alpha_1-1}\left(\frac{y}{z}\right)^{\alpha_2-1}\mathrm{d}y \\
&= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}z^{\alpha_1+\alpha_2-1}e^{-\lambda z}
\end{aligned}$$

因此，$Z\sim \mathrm{Ga}(\alpha_1+\alpha_2,\lambda)$。

注

1. 这个结论也可以推广至有限个独立的伽马分布随机变量的和。
2. 由于指数分布是一种特殊的伽马分布，即 $Exp(\lambda)=Ga(1,\lambda)$。那么 m 个独立同分布的指数分布随机变量之和为伽马分布。
3. 由于卡方分布是一种特殊的伽马分布，即 $\chi^2(n)=Ga\left(\frac{n}{2},\frac{1}{2}\right)$。那么 m 个独立的卡方分布随机变量之和仍为卡方分布。

结论 这里我们总结一下，具有可加性的常见分布。

1. 二项分布：$b(n,p)*b(m,p)=b(n+m,p)$.（留作学生课后自学内容）
2. 泊松分布：$P(\lambda_1)*P(\lambda_2)=P(\lambda_1+\lambda_2)$.
3. 正态分布：$N(\mu_1,\sigma_1^2)*N(\mu_2,\sigma_2^2)=N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.
4. 伽马分布：$Ga(\alpha_1,\lambda)*Ga(\alpha_2,\lambda)=Ga(\alpha_1+\alpha_2,\lambda)$.

9.2 极值分布

max 和 min 是两种常见的运算，其广泛的应用于风险管理问题中。比如：上海地区 2023 年最高气温达到 40 摄氏度的概率有多大？这里我们利用两个例题来阐述在不同的条件下如何计算极值的分布。

例题 9.4 设 $X_1, X_2, \cdots, X_n$ 是相互独立的 n 个随机变量，若 $Y = \max\{X_1, X_2, \cdots, X_n\}$，在以下情况下求 Y 的分布。

1. 若 $X_i \sim F_i(x), i = 1, 2, \cdots, n$，则 $Y = \max\{X_1, X_2, \cdots, X_n\}$ 的分布函数为
$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(\max\{X_1, X_2, \cdots, X_n\} \le y) \\ &= P(X_1 \le y, X_2 \le y, \cdots, X_n \le y) \\ &= P(X_1 \le y)P(X_2 \le y) \cdots P(X_n \le y) \\ &= \prod_{i=1}^{n} F_i(y). \end{aligned}$$
2. 若诸 X_i 同分布，即 $X \sim F(x), i = 1, 2, \cdots, n$，则 Y 的分布函数为
$$F_Y(y) = (F(y))^n .$$
3. 若诸 X_i 为连续随机变量，且诸 X_i 同分布，即 X_i 的密度函数为 $p(x), i = 1, 2, \cdots, n$，则 Y 的分布函数仍为
$$F_Y(y) = (F(y))^n .$$
而 Y 的密度函数为
$$\begin{aligned} p_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= n\,(F(y))^{n-1}\, p(y). \end{aligned}$$
4. 若 $X_i \sim Exp(\lambda), i = 1, 2, \cdots, n$，则 Y 的分布函数为
$$F_Y(y) = \begin{cases} (1 - e^{-\lambda y})^n, & y \ge 0 \\ 0, & y < 0 \end{cases}$$
而 Y 的密度函数为
$$p_Y(y) = \begin{cases} n(1 - e^{-\lambda y})^{n-1} \cdot \lambda e^{-\lambda y}, & y \ge 0 \\ 0, & y < 0 \end{cases}$$

例题 9.5 设 $X_1, X_2, \cdots, X_n$ 是相互独立的 n 个随机变量，若 $Z = \min\{X_1, X_2, \cdots, X_n\}$，在以下情况下求 Z 的分布。

1. 若 $X_i \sim F_i(x), i=1,2,\cdots,n$，则 $Z=\min\{X_1,X_2,\cdots,X_n\}$ 的分布函数为

$$\begin{aligned} F_Z(z) &= P(Z \le z) \\ &= P(\min\{X_1,X_2,\cdots,X_n\} \le z) \\ &= 1-P(\min\{X_1,X_2,\cdots,X_n\} > z) \\ &= 1-P(X_1>z, X_2>y,\cdots,X_n>z) \\ &= 1-P(X_1>z)P(X_2>z)\cdots P(X_n>z) \\ &= 1-\prod_{i=1}^{n}(1-F_i(z)). \end{aligned}$$

2. 若诸 X_i 同分布，即 $X \sim F(x), i=1,2,\cdots,n$，则 Z 的分布函数为

$$F_Z(z)=1-(1-F(z))^n.$$

3. 若诸 X_i 为连续随机变量，且诸 X_i 同分布，即 X_i 的密度函数为 $p(x), i=1,2,\cdots,n$，则 Z 的分布函数仍为

$$F_Z(z)=1-(1-F(z))^n.$$

而 Z 的密度函数为

$$\begin{aligned} p_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) \\ &= n\,(1-F(z))^{n-1}\,p(z). \end{aligned}$$

4. 若 $X_i \sim Exp(\lambda), i=1,2,\cdots,n$，则 Z 的分布函数为

$$F_Z(z)=\begin{cases} 1-e^{-n\lambda z}, & z \ge 0 \\ 0, & z<0 \end{cases}$$

而 Z 的密度函数为

$$p_Y(y)=\begin{cases} n\lambda e^{-n\lambda z}, & z \ge 0 \\ 0, & z<0 \end{cases}$$

9.3 变量变换法

9.3.1 二维情况

设二维随机变量 (X,Y) 的联合密度函数为 $p(x,y)$，如果函数

$$\begin{cases} u=g_1(x,y) \\ v=g_2(x,y) \end{cases}$$

有连续偏导数，且存在唯一的反函数

$$\begin{cases} x=x(u,v) \\ y=y(u,v) \end{cases}$$

其变换的雅克比行列式

$$J=\frac{\partial(x,y)}{\partial(u,v)}=\begin{vmatrix}\frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}\end{vmatrix}=\left(\frac{\partial(u,v)}{\partial(x,y)}\right)^{-1}=\begin{vmatrix}\frac{\partial u}{\partial x} & \frac{\partial u}{\partial y}\\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y}\end{vmatrix}^{-1}\neq 0$$

若

$$\begin{cases}U=g_1(X,Y)\\ V=g_2(X,Y)\end{cases}$$

则 $(U,V)'$ 的联合密度函数为

$$p(u,v)=p(x(u,v),y(u,v))\cdot|J|.$$

注 这个方法实际上就是二重积分的变量变换法。

例题 9.6 设随机变量 X 与 Y 独立同分布，都服从正态分布 $N(\mu,\sigma^2)$。记

$$\begin{cases}U=X+Y\\ V=X-Y\end{cases}$$

试求 $(U,V)'$ 的联合密度函数，且问 U 与 Y 是否独立?

解 因为

$$\begin{cases}u=x+y,\\ v=x-y\end{cases}$$

的反函数为

$$\begin{cases}x=\dfrac{u+v}{2},\\ y=\dfrac{u-v}{2},\end{cases}$$

则

$$J=\begin{vmatrix}\frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}\end{vmatrix}=\begin{vmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & -\frac{1}{2}\end{vmatrix}=-\frac{1}{2}.$$

所以，$(U,V)'$ 的联合密度函数为

$$\begin{aligned}p(u,v)&=p(x(u,v),y(u,v))|J|\\&=p_X\left(\frac{u+v}{2}\right)p_Y\left(\frac{u-v}{2}\right)\left|-\frac{1}{2}\right|\\&=\frac{1}{2\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{((u+v)/2-\mu)^2}{2}\right\}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{((u-v)/2-\mu)^2}{2\sigma^2}\right\}\\&=\frac{1}{4\pi\sigma^2}\exp\left\{-\frac{(u-2\mu)^2+v^2}{4\sigma^2}\right\}\\&=\frac{1}{\sqrt{2\pi(2\sigma^2)}}\exp\left\{-\frac{(u-2\mu)^2}{2\cdot(2\sigma^2)}\right\}\cdot\frac{1}{\sqrt{2\pi\cdot(2\sigma^2)}}\exp\left\{-\frac{v^2}{2\cdot(2\sigma^2)}\right\}.\end{aligned}$$

根据联合密度函数可知，$U\sim N(2\mu,2\sigma^2)$，$V\sim N(0,2\sigma^2)$。同时，由 $p(u,v)=p_U(u)p_V(v)$ 可知，U 与 V 相互独立。

注 作为变量变换法的一种变形，**增补变量法**也是常用的方法，为求出二维随机变量 (X,Y) 的

函数

$$U = g(X, Y)$$

的密度函数，需要增补一个新的随机变量

$$V = h(X, Y).$$

如何增补这个随机变量是该方法中的难点，通常令 $V = X$ 或 Y 可以解决大部分的问题。

其基本解法是

- 利用变量变换法求出 $(U, V)'$ 的联合密度函数 $p(u, v)$；
- 对 $p(u, v)$ 关于 v 积分，从而得到 U 的边际密度函数。

以下我们给出两个常用的公式，请同学们课后自行学习增补变量法后将证明过程不全。

例题 9.7 积的公式 设随机变量 X 与 Y 相互独立，其密度函数分别为 $p_X(x)$ 和 $p_Y(y)$。则 $U = XY$ 的密度函数为

$$p_U(u) = \int_{-\infty}^{\infty} p_X(u/v) p_Y(v) \frac{1}{|v|} \mathrm{d}v.$$

证明

例题 9.8 商的公式 设随机变量 X 与 Y 相互独立，其密度函数分别为 $p_X(x)$ 和 $p_Y(y)$。则 $U = X/Y$ 的密度函数为

$$p_U(u) = \int_{-\infty}^{\infty} p_X(uv) p_Y(v) |v| \mathrm{d}v.$$

证明

9.3.2 n 维情况（选修）

设 n 维随机变量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)'$ 的联合密度函数为 $p(x_1,x_2,\cdots,x_n)$。如果变换

$$\begin{cases} y_1=g_1\left(x_1,x_2,\cdots,x_n\right) \\ y_2=g_2\left(x_1,x_2,\cdots,x_n\right) \\ \vdots \\ y_n=g_n\left(x_1,x_2,\cdots,x_n\right) \end{cases}$$

有连续偏导数，且存在唯一的逆变换

$$\begin{cases} x_1=h_1\left(y_1,y_2,\cdots,y_n\right) \\ x_2=h_2\left(y_1,y_2,\cdots,y_n\right) \\ \vdots \\ x_n=h_n\left(y_1,y_2,\cdots,y_n\right) \end{cases}$$

其变换的雅克比行列式

$$J=\left|\frac{\partial\left(x_1,\cdots,x_n\right)}{\partial\left(y_1,\cdots,y_n\right)}\right|=\left|\left(\frac{\partial x_i}{\partial y_j}\right)\right|$$

则

$$\boldsymbol{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}=\begin{pmatrix} g_1(X_1,X_2,\cdots,X_n) \\ g_2(X_1,X_2,\cdots,X_n) \\ \vdots \\ g_n(X_1,X_2,\cdots,X_n) \end{pmatrix}$$

的联合密度函数为

$$p_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)=p_{\boldsymbol{X}}\left(g_1(\boldsymbol{y}),g_2(\boldsymbol{y}),\cdots,g_n(\boldsymbol{y})\right)\cdot|J|$$

例题 9.9 利用矩阵的技巧，我们重新来看一下例9.6。从矩阵的角度来看，

$$\begin{pmatrix} u \\ v \end{pmatrix}=\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

记

$$A=\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

可以计算其逆矩阵为

$$A^{-1}=\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$$

于是，雅可比行列式为

$$|J|=|A^{-1}|=-\frac{1}{2}.$$

因为

$$(X,Y)' \sim N_2\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}\right),$$

其联合密度函数为

$$\begin{aligned}
p(x,y) &= (2\pi)^{-2/2}|\sigma^2 I_2|^{-1/2}\exp\left\{-\frac{1}{2}\left(\begin{pmatrix}x\\ y\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)'(\sigma^2 I_2)^{-1}\left(\begin{pmatrix}x\\ y\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\right\}\\
&= (2\pi\sigma^2)^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\begin{pmatrix}x\\ y\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)'\left(\begin{pmatrix}x\\ y\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\right\}
\end{aligned}$$

所以，(U,V) 的联合密度函数为

$$\begin{aligned}
p(u,v) &= (2\pi\sigma^2)^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(A^{-1}\begin{pmatrix}u\\ v\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)'\left(A^{-1}\begin{pmatrix}u\\ v\end{pmatrix}-\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\right\}\cdot|J|\\
&= (2\pi\sigma^2)^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\begin{pmatrix}u\\ v\end{pmatrix}-A\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)'(A^{-1})'(A^{-1})\left(\begin{pmatrix}u\\ v\end{pmatrix}-A\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\right\}\cdot\frac{1}{2}\\
&= (2\pi\sigma^2)^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\begin{pmatrix}u\\ v\end{pmatrix}-A\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)'(AI_2A')^{-1}\left(\begin{pmatrix}u\\ v\end{pmatrix}-A\begin{pmatrix}\mu\\ \mu\end{pmatrix}\right)\right\}\cdot\frac{1}{2}\\
&= (4\pi\sigma^2)^{-1}\exp\left\{-\frac{1}{2}\left(\begin{pmatrix}u\\ v\end{pmatrix}-\begin{pmatrix}2\mu\\ 0\end{pmatrix}\right)'\begin{pmatrix}2\sigma^2 & 0\\ 0 & 2\sigma^2\end{pmatrix}^{-1}\left(\begin{pmatrix}u\\ v\end{pmatrix}-\begin{pmatrix}2\mu\\ 0\end{pmatrix}\right)\right\}
\end{aligned}$$

所以，

$$\begin{pmatrix}U\\ V\end{pmatrix} \sim N_2\left(\begin{pmatrix}2\mu\\ 0\end{pmatrix}, \begin{pmatrix}2\sigma^2 & 0\\ 0 & 2\sigma^2\end{pmatrix}\right)$$

结论 若 $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$，对于任意常数矩阵 $A_{m,n}$，有 $\boldsymbol{Y} = A\boldsymbol{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A')$.

第 10 讲 条件分布与条件期望

阅读章节

❑ Intro to Prob 3.5 3.6 ❑ Prob & Stat 3.5

例题 10.1 （Simpson 悖论）某一年北美某高校的商学院和法学院录取率如表10.1所示。从数据来看，在录取时该高校是否存在性别歧视？因女生的性别，而导致学生不予录取。

表 10.1: 某高校两个学院的总录取人数

	全部男生	全部女生
录取	209	143
未录取	95	110
录取率	67.43%	56.52%

商学院和法学院实际的录取人数如表10.2所示。无论在法学院还是在商学院，女生的录取率都高于男生的录取率。

表 10.2: 某高校两个学院各自的录取人数

	法学院		商学院	
	男生	女生	男生	女生
录取	8	51	201	92
未录取	45	101	50	9
录取率	15.09%	33.55%	80.08%	91.09%

问题 10.1 但是为什么从总体上来看，我们会认为女生的录取低？

笔记

10.1 离散场合下的条件分布

设二维离散随机变量 (X,Y) 的联合分布列为

$$p_{ij}=P\left(X=x_i,Y=y_j\right),i=1,2,\cdots;\quad j=1,2,\cdots.$$

定义 10.1

对一切使 $P\left(Y=y_j\right)=p_{\cdot j}=\sum_{i=1}^{\infty}p_{ij}>0$ 的 y_j，称

$$\begin{aligned}p_{i|j}&=P\left(X=x_i|Y=y_j\right)\\&=\frac{P\left(X=x_i,Y=y_j\right)}{P\left(Y=y_j\right)}\\&=\frac{p_{ij}}{p_{\cdot j}}\end{aligned}$$

为给定 $Y=y_j$ 条件下 X 的条件分布列，$i=1,2,\cdots$。♣

定义 10.2

给定 $Y=y_j$ 条件下 X 的条件分布函数为

$$F\left(x|y_j\right)=\sum_{x_i\le x}P\left(X=x_i|Y=y_j\right)=\sum_{x_i\le x}p_{i|j}$$

♣

例题 10.2 设随机变量 X 与 Y 独立，且 $x\sim P\left(\lambda_1\right),Y\sim P\left(\lambda_2\right)$。在已知 $X+Y=n$ 的条件下，求 X 的条件分布。

解 在例9.1中已经证明了 $X+Y\sim P(\lambda_1+\lambda_2)$。所以

$$\begin{aligned}P(X=k|X+Y=n)&=\frac{P(X=k,X+Y=n)}{P(X+Y=n)}\\&=\frac{P(X=k,Y=n-k)}{P(X+Y=n)}\\&=\frac{P(X=k)\cdot P(Y=n-k)}{P(X+Y=n)}\\&=\frac{\frac{\lambda_1^k}{k!}e^{-\lambda_1}\cdot\frac{\lambda_2^{(n-k)}}{(n-k)!}e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}}\\&=\frac{n!}{k!(n-k)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-k}\end{aligned}$$

因此，在 $X+Y=n$ 的条件下，X 的条件分布为二项分布 $b\left(n,\frac{\lambda_1}{\lambda_1+\lambda_2}\right)$.

例题 10.3 设在一段时间内进入某一商店的顾客人数 X 服从泊松分布 $P(\lambda)$，每位顾客购买某商品的概率为 p，并且每位顾客是否购买该种物品相互独立。求进入商店的顾客购买这种物品的人数 Y 的分布列。

解 由题可知，

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

在进入商店的人数 $X=m$ 的条件下，购买某种物品的人数 Y 的条件分布为二项分布 $b(m,p)$，

即

$$P(Y=k|X=m)=\binom{m}{k}p^k(1-p)^{m-k},\quad k=0,1,\cdots,m.$$

由全概率公式有

$$\begin{aligned}P(Y=k)&=\sum_{m=k}^{+\infty}P(X=m)\cdot P(Y=k|X=m)\\&=\sum_{m=k}^{+\infty}\frac{\lambda^m}{m!}e^{-\lambda}\cdot\frac{m!}{k!(m-k)!}p^k(1-p)^{m-k}\\&=\frac{1}{k!}e^{-\lambda}(\lambda p)^k\cdot\sum_{m=k}^{+\infty}\frac{(\lambda(1-p))^{m-k}}{(m-k)!}\\&=\frac{1}{k!}(\lambda p)^k e^{-\lambda}\sum_{m=0}^{+\infty}\frac{(\lambda(1-p))^m}{m!}e^{-\lambda(1-p)}\cdot e^{\lambda(1-p)}\\&=\frac{(\lambda p)^k}{k!}e^{-\lambda p},\quad k=0,1,2,\cdots\end{aligned}$$

因此，$Y\sim P(\lambda p)$。

10.2 连续场合下的条件分布

设二维连续随机变量 (X,Y) 的联合密度为 $p(x,y)$，边际密度函数为 $p_X(x)$ 和 $p_Y(y)$。在离散随机变量场合，其条件分布函数为 $P(X\le x|Y=y)$，但是连续随机变量取某个值的概率为零，即 $P(Y=y)=0$。所以无法用条件概率直接计算 $P(X\le x|Y=y)$。一个自然的想法是，将 $P(X\le x|Y=y)$ 看成当 $h\to 0$ 时 $P(X\le x|y\le Y\le y+h)$ 的极限，即

$$\begin{aligned}P(X\le x|Y=y)&=\lim_{h\to 0}P(X\leqslant x|y\leqslant Y\leqslant y+h)\\&=\lim_{h\to 0}\frac{P(X\le x,y\le Y\le y+h)}{P(y\le Y\le y+h)}\\&=\lim_{h\to 0}\frac{\int_{-\infty}^{x}\int_{y}^{y+h}p(u,v)\mathrm{d}v\mathrm{d}u}{\int_{y}^{y+h}p(v)\mathrm{d}v}\\&=\lim_{h\to 0}\frac{\int_{-\infty}^{x}\left(\frac{1}{h}\int_{y}^{y+h}p(u,v)\mathrm{d}v\right)\mathrm{d}u}{\frac{1}{h}\int_{y}^{y+h}P_Y(v)\mathrm{d}v}\end{aligned}$$

当 $p_Y(y),p(x,y)$ 在 y 处连续时，由积分中值定理可得

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}\int_{y}^{y+h}p_Y(v)\mathrm{d}v&=\lim_{h\to 0}\frac{1}{h}\cdot(y+h-y)\cdot p_Y(y+\delta h)=p_Y(y)\\\lim_{h\to 0}\frac{1}{h}\int_{y}^{y+h}p(u,v)\mathrm{d}v&=\lim_{h\to 0}\frac{1}{h}(y+h-y)\cdot p(u,y+\delta h)=p(u,y)\end{aligned}$$

所以

$$P(X\le x|Y=y)=\int_{-\infty}^{x}\frac{p(u,y)}{p_Y(y)}\mathrm{d}u.$$

定义 10.3

对一切使 $p_Y(y)>0$ 的 y，给定 $Y=y$ 条件下 X 的条件分布函数和条件密度函数分别为

$$\begin{aligned} F(x|y) &= \int_{-\infty}^{x} \frac{p(u,y)}{p_Y(y)}\mathrm{d}u, \\ p(x|y) &= \frac{p(x,y)}{p_Y(y)}. \end{aligned}$$

♣

例题 10.4 设 (X,Y) 服从二维正态分布

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

由边际分布可知，X 服从正态分布 $N(\mu_1,\sigma_1^2)$，Y 服从正态分布 $N(\mu_2,\sigma_2^2)$。在给定 $Y=y$ 的条件下，X 的边际密度函数为

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p_Y(y)} \\ &= \frac{\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}}{\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{1}{2\sigma_2^2}(y-\mu_2)^2\right\}} \\ &= \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{\sigma_1^2(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{\rho^2(y-\mu_2)^2}{\sigma_2^2}\right)\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}}\exp\left\{-\frac{1}{2\sigma_1^2(1-\rho^2)}\left(x-\left(\mu_1+\rho\cdot\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)\right)^2\right\} \end{aligned}$$

因此，在 $Y=y$ 的条件下，X 的条件分布为

$$N\left(\mu_1+\frac{\rho\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2)\right).$$

注 在定义连续场合下的条件密度函数下，我们可以给出连续场合下的全概率公式及贝叶斯公式。因为

$$p_Y(y)=\frac{p(x,y)}{p_Y(y)}$$

所以，

$$p(x,y)=p_Y(y)p_Y(y).$$

于是，我们有

- 全概率公式

$$p_X(x)=\int_{-\infty}^{+\infty}p(x,y)\mathrm{d}y=\int_{-\infty}^{+\infty}p_Y(y)p(x|y)\mathrm{d}y.$$

- 贝叶斯公式

$$p(y|x)=\frac{p(x,y)}{p_X(x)}=\frac{p_Y(y)p(x|y)}{\int_{-\infty}^{+\infty}p_Y(y)p(x|y)\mathrm{d}y}.$$

例题 10.5 设 $X\sim N(\mu,\sigma_1^2)$ 且在给定 $X=x$ 的条件下，Y 的条件分布为 $N(x,\sigma_2^2)$，求 Y 的密

度函数 $p_Y(y)$。

解 由题可知，

$$\begin{aligned} p_X(x) &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu)^2\right\}, \\ p(y|x) &= \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{1}{2\sigma_2^2}(y-x)^2\right\}. \end{aligned}$$

所以，根据全概率公式可知，

$$\begin{aligned} P_Y(y) &= \int_{-\infty}^{+\infty} p_X(x)p(y|x)\mathrm{d}x \\ &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu)^2\right\}\cdot\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{1}{2\sigma_2^2}(y-x)^2\right\}\mathrm{d}x \\ &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}}\int_{-\infty}^{+\infty}\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}x^2-\frac{2\mu}{\sigma_1^2}x+\frac{\mu^2}{\sigma_1^2}\ +\frac{y^2}{\sigma_2^2}-\frac{2y}{\sigma_2^2}x+\frac{x^2}{\sigma_2^2}\right)\right\}\mathrm{d}x \\ &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}}\exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)\left(x-\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)^{-1}\left(\frac{\mu}{\sigma_1^2}+\frac{y}{\sigma_2^2}\right)\right)^2\right)\right\}\mathrm{d}x \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma_1^2+\sigma_2^2}}\exp\left\{-\left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)\left(\frac{\mu^2}{\sigma_1^4}+\frac{2\mu y}{\sigma_1^2\sigma_2^2}+\frac{y^2}{\sigma_2^4}\right)+\frac{y^2}{\sigma_2^2}+\frac{\mu^2}{\sigma_1^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\left(\sigma_1^2+\sigma_2^2\right)}}\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2+\sigma_2^2}y^2-\frac{2\mu y}{\sigma_1^2+\sigma_2^2}+\frac{\mu^2}{\sigma_1^2+\sigma_2^2}\right)\right\} \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left\{-\frac{1}{2\left(\sigma_1^2+\sigma_2^2\right)}(y-\mu)^2\right\} \end{aligned}$$

因此，

$$Y\sim N(\mu,\sigma_1^2+\sigma_2^2).$$

10.3 混合场合下的条件分布

例题 10.6 生活中，医生根据一些生理指标测量，如体温、血压等生化指标来进行医学诊断。

我们简化一下这个问题。令 A 是我们感兴趣的一个随机变量，而 $P(A)$ 是事件 A 发生的概率。令 Y 是一个连续型随机变量，并假定已知条件密度函数 $p(y|A)$ 和 $p(y|\overline{A})$。

问题 10.2 在给定 Y 取值为 y 时，事件 A 发生的条件概率 $P(A|Y=y)$ 是什么？如何计算？

和连续场合下的条件分布有同一个问题，因为 Y 是连续型的随机变量，所以事件 $\{Y=y\}$

发生的概率为零。这里我们考虑 $\{y \le Y \le y+\delta\}$，其中 $\delta > 0$。我们考虑

$$\begin{aligned} P(A|Y=y) &\approx P(A|y \le Y \le y+\delta) \\ &= \frac{P(A)P(y \le Y \le y+\delta|A)}{P(y \le Y \le y+\delta)} \\ &\approx \frac{P(A)p(y|A)\delta}{p(y)\delta} \\ &= \frac{P(A)p(y|A)}{p(y)} \end{aligned}$$

以上就是混合场景下条件分布的定义。由此，可以定义全概率公式和贝叶斯公式。

1. 全概率公式为

$$p_Y(y) = P(A)p(y|A) + P(\overline{A})p(y|\overline{A}).$$

2. 贝叶斯公式为

$$P(A|Y=y) = \frac{P(A)p(y|A)}{P(A)p(y|A) + P(\overline{A})p(y|\overline{A})}.$$

如果已知 $P(A|Y=y)$ 后，我们也可以推导出 $p(y|A)$，即

$$p(y|A) = \frac{p_Y(y)P(A|Y=y)}{P(A)} = \frac{p_Y(y)P(A|Y=y)}{\int_{-\infty}^{\infty} p_Y(y)P(A|Y=y)\mathrm{d}y}.$$

注 上述公式中将随机事件 A 推广到离散型随机变量。

10.4 条件数学期望

正如之前课程内容中介绍的，我们讨论过条件概率是符合概率的公理化定义。由此，根据条件概率，我们可以定义随机变量的**条件分布**。既然有分布，我们也可以定义这个分布的特征数。这里，我们首先介绍如何定义条件分布的数学期望——条件期望。

定义 10.4 (条件期望)

条件分布的数学期望（若存在）称为条件期望，即

$$E(X|Y=y) = \begin{cases} \sum_i x_i P\left(X = x_i|Y=y\right), & \text{离散场合}, \\ \int_{-\infty}^{+\infty} xP(x|y)dx, & \text{连续场合}. \end{cases}$$

♣

因为条件期望是一种期望，所以条件期望也满足期望的性质。

性质 $E\left(a_1X_1 + a_2X_2|Y=y\right) = a_1E\left(X_1|Y=y\right) + a_2E\left(X_2|Y=y\right)$;

笔记

这里从另一个角度来看条件期望。在给定 $Y=y$ 的条件下，X 的条件分布会因 y 的取值不同而不同，从而导致了 $E(X|Y=y)$ 亦是如此。所以，$E(X|Y=y)$ 可以看作 y 的函数。我们记 $g(y)=E(X|Y=y)$。

对于随机变量 Y，$g(Y)=E(X|Y)$ 是随机变量 Y 的函数。所以，条件期望 $E(X|Y)$ 可以看作随机变量 Y 的函数，通常仍是一个随机变量。这里，我们讨论一下，其期望 $E(E(X|Y))$ 是什么呢?

定理 10.1 (重期望公式) 设 (X,Y) 是二维随机变量且 $E(X)$ 存在，则

$$E(X)=E(E(X|Y)).$$

♡

证明 这里仅证明连续场合。

设二维随机变量 (X,Y) 的联合密度函数为 $p(x,y)$，记 $g(y)=E(X|Y=y)$，则 $g(X)=E(X|Y)$，由于 $p(x,y)=p(x|Y)\cdot p_Y(y)$，可得

$$\begin{aligned}
E(X) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xp(x,y)dxdy \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\cdot p(x|y)\cdot p_Y(y)dxdy \\
&= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} xp(x|y)dx\right)\cdot p_Y(y)dy \\
&= \int_{-\infty}^{+\infty} E(X|Y=y)\cdot p_Y(y)dy \\
&= \int_{-\infty}^{+\infty} g(y)\cdot p_Y(y)dy \\
&= E[g(Y)] \\
&= E(E(X|Y)).
\end{aligned}$$

类似地，我们可以定义条件方差，即 $Var(X|Y)$，即

$$\begin{aligned}
Var(X|Y) &= E\left((X-E(X|Y))^2|Y\right) \\
&= E(X^2|Y)-(E(X|Y))^2.
\end{aligned}$$

类似于重期望公式，我们也可以将 X 的方差拆解成两个部分，即

$$Var(X)=E(Var(X|Y))+Var(E(X|Y)),$$

其中，前者可以看成组内方差，后者可以看成组间方差。

例题 10.7 一矿工被困在有三个门的矿井里。第一个门通一坑道，沿此坑道走 3 小时可到达安全区；第二个门通一坑道，沿此坑道走 5 小时又回到原处；第三个门通一坑道，沿此坑道走 7 小时也回到原处。假定此矿工总是等可能地在三个门中选择一个，试求他平均要用多少时间才能到达安全区。

解 如果直接求 X 的分布，X 的可能取值为 $3, 5+3, 7+3, 5+5+3, 5+7+3, \cdots$ 这是很困难的。

这里，我们考虑另一种解法：Y 表示在矿井中选的门，即 $Y=i$ 表示选了第 i 个门。可知

$$P(Y=1)=P(Y=2)=P(Y=3)=\frac{1}{3}.$$

因为选了第一个门后，3 小时到达安全区，所以 $E(X|Y=1)=3$；因为选了第二个门后，5 小时回到原地，所以 $E(X|Y=2)=5+E(X)$；因为选了第三个门后，7 小时回到原地，所以 $E(X|Y=3)=7+E(X)$. 综上，

$$\begin{aligned}
E(X) &= E(E(X|Y)) \\
&= E(X|Y=1)\cdot\frac{1}{3}+E(X|Y=2)\cdot\frac{1}{3}+E(X|Y=3)\cdot\frac{1}{3} \\
&= 3\cdot\frac{1}{3}+(5+E(X))\cdot\frac{1}{3}+(7+E(X))\cdot\frac{1}{3} \\
&= 5+\frac{2}{3}E(X).
\end{aligned}$$

解得 $E(X)=15$.

例题 10.8 设 $X_1, X_2, \cdots$ 为一列独立同分布的随机变量，随机变量 N 只取正整数值且 N 与 X_n 独立，证明

1. $$E\left(\sum_{i=1}^{N} X_i\right)=E\left(X_1\right)\cdot E(N)$$

2. （课后自学）
$$Var\left(\sum_{i=1}^{N} X_i\right)=\operatorname{Var}\left(X_1\right)E(N)+\left(E\left(X_1\right)\right)^2\operatorname{Var}(N)$$

证明

1. 我们先考虑第一个问题，这里与之前介绍的期望的性质有一个明显的差异——这里考虑了随机变量个随机变量之和的期望，而之前考虑的是有限个随机变量之和的期望。

$$\begin{aligned}
E\left(\sum_{i=1}^{N} X_i\right) &= E\left(E\left(\sum_{i=1}^{N} X_i|N\right)\right) \\
&= \sum_{n=1}^{+\infty} E\left(\sum_{i=1}^{N} X_i|N=n\right)\cdot p_N(n) \\
&= \sum_{n=1}^{+\infty} E\left(\sum_{i=1}^{n} X_i|N=n\right) p_N(n) \\
&= \sum_{n=1}^{+\infty} E\left(\sum_{i=1}^{n} X_i\right)\cdot p_N(n) \\
&= \sum_{n=1}^{+\infty} n\cdot E\left(X_1\right)\cdot p_N(n) \\
&= E\left(X_1\right)\cdot E(N)
\end{aligned}$$

2. 随机变量个随机变量之和的方差的证明过程由学生课后自学。

$$\begin{aligned} Var\left(\sum_{i=1}^{N} X_i\right) &= E\left(\operatorname{var}\left(\sum_{i=1}^{N} X_i|N\right)\right) + \operatorname{Var}\left(E\left(\sum_{i=1}^{N} X_i|N\right)\right) \\ &= I_1 + I_2 \end{aligned}$$

其中

$$\begin{aligned} I_1 &= \sum_{n=1}^{+\infty} Var\left(\sum_{i=1}^{n} X_i|N=n\right) \cdot p_N(n) \\ &= \sum_{n=1}^{+\infty} Var\left(\sum_{i=1}^{n} X_i\right) p_N(n) \\ &= \sum_{n=1}^{+\infty} \sum_{i=1}^{n} Var\left(X_i\right) p_N(n) \\ &= \operatorname{Var}\left(X_1\right) \cdot E(N) \end{aligned}$$

$$\begin{aligned} I_2 &= E\left(E\left(\sum_{i=1}^{N} X_i|N\right)\right)^2 - E^2\left(E\left(\sum_{i=1}^{N} X_i|N\right)\right) \\ &= \sum_{n=1}^{+\infty}\left(E\left(\sum_{i=1}^{N} X_i|N=n\right)\right)^2 p_N(n) - \left(E\left(X_1\right) \cdot E(N)\right)^2 \\ &= \sum_{n=1}^{+\infty}\left(E\left(\sum_{i=1}^{n} X_i|N=n\right)\right)^2 p_N(n) - \left(E\left(X_1\right) \cdot E(N)\right)^2 \\ &= \sum_{n=1}^{+\infty}\left(n \cdot E\left(X_1\right)\right)^2 p_N(n) \\ &= \left(E\left(X_1\right)\right)^2 E\left(N^2\right) - \left(E\left(X_1\right)\right)^2 \cdot (E(N))^2 \\ &= \left(E\left(X_1\right)\right)^2 Var(N) \end{aligned}$$

第 11 讲 随机变量序列的收敛性

阅读章节

- Intro to Prob 5.3 5.5
- Prob & Stat 4.1

11.1 依概率收敛

定义 11.1 (数列的收敛性)

设一个数列 $\{a_1, a_2, \cdots, a_n, \cdots\}$。如果对于任意 $\varepsilon > 0$，存在一个正整数 n_0，当 $n \geq n_0$ 时，有

$$\lim_{n\to} a_n = a$$

则称数列 $\{a_n\}$ 是收敛的，其极限为 a。 ♣

问题 11.1 什么是随机变量序列的收敛性?

例题 11.1 在抛一枚均匀硬币（正面反面出现的概率相等）的场景下，考虑正面朝上的频率。

表 11.1: 10 次抛硬币的结果

第 i 次抛硬币	1	2	3	4	5	6	7	8	9	10
结果	反面	反面	正面	正面	反面	正面	正面	正面	反面	反面

笔记

结论 概率是频率的稳定值。

1. 频率 X_n 与概率 p 的绝对偏差 $|x_n - p|$ 将随 n 增大而呈现主键减小的趋势，但不能说它收敛于零；
2. 由于随机性，绝对偏差 $|x_n - p|$ 时大时小，虽然无法排除大偏差发生的可能性，但岁 n 不断增大，大偏差发生的可能性会越来越小。

由此，我们定义随机变量序列的收敛性。

定义 11.2

设 $\{X_n\}$ 为一随机变量序列，而 X 为一随机变量。如果对任意的 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left(|X_n - X| \geqslant \varepsilon\right) = 0$$

则称序列 $\{X_n\}$ 依概率收敛于 X, 记作 $X_n \xrightarrow{P} X$。♣

注

1. $P\left(|X_n - X| \geqslant \varepsilon\right) \to 0 \Leftrightarrow P\left(|X_n - X| < \varepsilon\right) \to 1$;
2. 特别地，$P(X = c) = 1$ 时, 则 $X_n \xrightarrow{P} c$。

性质 （依概率收敛的四则运算）

设 $\{X_n\}, \{Y_n\}$ 是两个随机变量序列，a, b 是两个常数。如果

$$X_n \xrightarrow{P} a, \quad Y_n \xrightarrow{P} b,$$

则有

1. $X_n \pm Y_n \overset{P}{\to} a \pm b$;
2. $X_n \times Y_n \overset{P}{\to} a \times b$;
3. $X_n / Y_n \overset{P}{\to} a/b \quad (b \neq 0)$。

例题 11.2 设随机变量 X_n 服从柯西分布，其密度函数为

$$p_n(x) = \frac{n}{\pi\left(1 + n^2 x^2\right)} \qquad -\infty < x < \infty.$$

试证 $X_n \overset{P}{\to} 0$。

证明 我们考虑事件 $\{|X_n - 0| \geq \varepsilon\}$ 发生的概率，即

$$\begin{aligned}
P\left(|X_n - 0| \geq \varepsilon\right) &= P\left(|X_n| \geqslant \varepsilon\right) \\
&= \int_{-\infty}^{-\varepsilon} p_n(x)dx + \int_{\varepsilon}^{+\infty} p_n(x)dx \\
&= \int_{-\infty}^{-\varepsilon} \frac{n}{\pi\left(1 + n^2x^2\right)}dx + \int_{\varepsilon}^{+\infty} \frac{n}{\pi\left(1 + n^2x^2\right)}dx \\
&= \frac{2}{\pi}\int_{\varepsilon}^{+\infty} \frac{1}{\pi\left(1 + n^2x^2\right)}d(nx) \\
&\overset{nx=\tan u}{=} \frac{2}{\pi}\int_{\arctan(n\varepsilon)}^{\frac{\pi}{2}} \frac{1}{1 + \tan^2 u} \cdot d\tan u \\
&= \frac{2}{\pi}\left(u|_{\arctan(n\varepsilon)}^{\frac{\pi}{2}}\right) \\
&= \frac{2}{\pi}\left(\frac{\pi}{2} - \arctan(n\varepsilon)\right) \\
&= 1 - \frac{2}{\pi} \cdot \arctan(n\varepsilon) \to 0
\end{aligned}$$

问题 11.2 比较一下，数列的收敛性与随机变量的收敛性。

问题 11.3 如果 $X_n \xrightarrow{P} a$, 那么 $E(X_n) \to a$ 成立吗?

接下来我们看下面一个例子。

例题 11.3 考虑一个随机变量，其分布列为

$$P(X_n = x) = \begin{cases} 1 - \dfrac{1}{n}, & x = 0 \\ \dfrac{1}{n}, & x = n^2 \\ 0, & \text{其他}. \end{cases}$$

解 由题可知，

$$P(|X_n| > \varepsilon) = P\left(X_n = n^2\right) = \frac{1}{n} \to 0.$$

因此，$X_n \xrightarrow{P} 0$。我们来计算一下 X_n 的数学期望，即

$$E(X_n) = 0 \cdot \left(1 - \frac{1}{n}\right) + n^2 \cdot \frac{1}{n} = n \to \infty.$$

因此，$E(X_n) \nrightarrow 0$。

11.2 以概率 1 收敛

定义 11.3

设 $\{X_n\}$ 是一列随机变量，如果

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

则称序列 $\{X_n\}$ 以概率 1 收敛于 X，记 $X_n \xrightarrow{a.s} X$。

♣

性质 $X_n \xrightarrow{a.s} X \Rightarrow X_n \xrightarrow{P} X$，反之不然。

例题 11.4 考虑一个离散时间的到达过程。我们假定到达的时刻属于正整数集 $1, 2, \cdots$。现将这个集合分割成若干个互不相交的集合，$I_k = \left\{2^k, 2^k + 1, \cdots, 2^{k+1} - 1\right\}, k = 0, 1, \cdots$。注意到集合 I_k 的元素个数为 2^k，随着 k 的增大而增大。具体来说，

$$\begin{aligned} I_0 &= \{1\} \\ I_1 &= \{2, 3\} \\ I_2 &= \{4, 5, 6, 7\} \\ &\vdots \end{aligned}$$

假设在每个区间 I_k 中只有唯一的一个到达时刻，且在区间内每个时刻到达是等可能的。同时假定在各个区间到达时刻是相互独立的。

记第 k 个区间 I_k 内的到达时刻为 n_k，则 n_k 是相互独立的随机变量序列，$k = 1, 2, \cdots$。定义一个新的随机变量序列 $\{Y_n\}$，如果时刻 n 到达了，则定义 $Y_n = 1$；否则，$Y_n = 0$。如果 $n \in I_k$，则 $P(Y_n = 1) = P(Y_n \neq 0) = \frac{1}{2^k}$。

接下来，我们来考虑随机变量序列 $\{Y_n\}$ 的收敛性。

1. 因为 I_k 互不相容，对于任意的 n，存在唯一 k，使得 $n \in I_k$。当 n 越大，则 k 也越大，即
$$\lim_{n\to+\infty} P(Y_n > \varepsilon) = \lim_{n\to+\infty} \frac{1}{2^k} = 0.$$
因此，$Y_n \xrightarrow{P} 0$。
2. 然而，每个区间 I_k 都有到达时刻，所以，取值为 1 的 Y_n 的个数是无穷多次，即对于任意 N，存在 $n > N$，使得
$$P(Y_n = 1) = 1$$
因此，$P(\lim_{n\to+\infty} Y_n = 0) \neq 1$，即 Y_n 不以概率 1 收敛到 0。

11.3 按分布收敛

问题 11.4 一个分布函数序列 $F_n(x)$ 的收敛性如何定义?

一个很自然的想法是——对于任意的 $x \in R$，当 $n \to +\infty$ 时，有
$$F_n(x) \to F(x).$$

例题 11.5 我们考虑以下随机变量序列 $\{X_n\}$ 的分布列，即
$$P\left(X_n = \frac{1}{n}\right) = 1 \quad n = 1.2, \cdots.$$
于是，其分布函数为
$$F_n(x) = \begin{cases} 0, & x < \dfrac{1}{n}, \\ 1, & x \geqslant \dfrac{1}{n}. \end{cases}$$
那么，对于任意的 x，$F_n(x)$ 的极限函数为
$$g(x) = \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases}$$

然而，我们发现 $g(x)$ 并不满足右连续性，所以，$g(x)$ 不是一个随机变量的分布函数。由此，随机变量序列的分布函数的极限不是一个随机变量的分布函数，这是不合理的。

我们发现，随机变量 X_n 的分布函数 $F_n(x)$ 是在点 $x_0 = \frac{1}{n}$ 处有跳跃点。当 $n \to +\infty$ 时，跳跃点 $x_0 \to 0$。可以证明
$$F(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases}$$
是一个随机变量的分布函数。对于任意的 $x \in (-\infty, 0) \cup (0, +\infty)$，有 $\lim_{n\to+\infty} F_n(x) = F(x)$。但在 $x = 0$ 处，
$$\lim_{n\to\infty} F_n(0) = 0 \neq 1 = F_n(0).$$

注 在上面的例子中，分布函数的收敛性不成立的点 $x = 0$ 恰好是 $F(x)$ 的间断点（或跳跃点）。我们在定义分布函数的收敛性仅考虑 $F(x)$ 的连续点。

定义 11.4

设随机变量 $X, X_1, X_2, \cdots, X_n, \cdots$ 的分布函数分别为 $F(x), F_1(x), F_2(x), \cdots, F_n(x), \cdots$。若对 $F(x)$ 的任意连续点 x，都有

$$\lim_{n\to\infty} F_n(x) = F_{(x)}.$$

则称 $F_n(x)$ 弱收敛与 $F(x)$, 记作

$$F_n(x) \xrightarrow{\omega} F(x).$$

也称相应的随机变量序列 X_n 按分布收敛于 X,

$$X_n \xrightarrow{L} X.$$

♣

如果 $F(x)$ 是连续函数，那么弱收敛就点点收敛。

性质

1. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X$;
2. 如果 c 是一个常数，那么 $X_n \xrightarrow{P} c \Leftrightarrow X_n \xrightarrow{L} c$。

第 12 讲 大数定律

阅读章节

- ❑ Intro to Prob 5.2 5.5
- ❑ Prob & Stat 4.3

12.1 伯努利大数定律

通常说，**概率是频率的稳定值**。

考虑抛硬币这个实验，记第 i 次抛硬币的结果为 X_i。一般假定每次抛硬币的结果仅有两个：正面（感兴趣的）和反面（不感兴趣的）；而且每次的结果是相互独立的。所以，X_i 是服从伯努利分布或二点分布，即 $X_i \overset{iid}{\sim} b(1,p)$，其中 p 表示正面朝上的概率，即 $P(X_i = 1) = p$。

于是，在 n 次抛硬币的结果中，正面朝上的总次数为

$$S_n = \sum_{i=1}^{n} X_i,$$

在 n 次结果中，正面朝上的频率为

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

根据二项分布的可加性，可知

$$S_n \sim b(n,p).$$

那么，

$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= \frac{1}{n}E(S_n) = \frac{1}{n}\cdot(np) = p, \\ Var\left(\frac{S_n}{n}\right) &= \frac{1}{n^2}Var(S_n) = \frac{1}{n^2}\cdot(np(1-p)) = \frac{p(1-p)}{n}. \end{aligned}$$

我们发现这个频率的期望是概率，意味着频率是在概率附近周围波动的，而且频率的方差 $p(1-p)/n$，随着 n 不断增大而快速减小的。于是，一个很自然的问题：这个频率的“极限”是不是这个概率呢？由此，我们来看下边这个定理。

定理 12.1 (伯努利大数定律) 设 S_n 为 n 重伯努利试验中事件 A 发生的次数，p 为每次试验中 A 出现的概率，则对任意的 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1.$$

♡

证明 由于 $S_n \sim b(n,p)$ 且 $E(\frac{S_n}{n}) = p, Var(\frac{S_n}{n}) = \frac{p(1-p)}{n}$。

由切比雪夫不等式，可得

$$1 \geq P\left(\left|\frac{S_n}{n}-p\right|<\varepsilon\right) \geq 1-\frac{Var(\frac{S_n}{n})}{\varepsilon^2}=1-\frac{p(1-p)}{n\varepsilon^2} \to 1$$

因此

$$P\left(\left|\frac{S_n}{n}-p\right|<\varepsilon\right) \to 1.$$

注 随着 n 的增大，事件 A 发生的频率 $\frac{S_n}{n}$ 与其概率 p 的偏差 $\left|\frac{S_n}{n}-p\right|$ 大于预先给定的精度 ε 的可能性越来越小，这就是频率稳定于概率的含义。

12.2 大数定律的一般形式

我们从另一个视角看一下伯努利大数定律，所谓的频率是

$$\frac{S_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

而所谓的概率是

$$p=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i).$$

伯努利大数定律表明了对于 $\{X_n\}$ 这个随机变量序列，如果 X_i 是服从伯努利分布 $b(1,p)$，且每个随机变量是相互独立，那么对任意 $\varepsilon>0$，都有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|>\varepsilon\right)=1.$$

注意到的是，伯努利大数定律给了一个**很强**的条件——不仅约束了随机变量序列中每一个随机变量的分布类型，还约束了随机变量序列中每一个随机变量之间都是相互独立的。于是，我们可以给出以下这个定义。

定义 12.1 (弱大数定律)

设有一随机变量序列 $\{X_n\}$，如果其具有形如

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}E(X_i)\right|<\varepsilon\right)\to 1$$

的性质，则称该随机变量序列 X_n 服从（弱）大数定律。♣

我们自然有下面的一个问题：

问题 12.1 是否只有满足伯努利大数定律的条件，随机变量序列才能满足大数定律?

12.3 不同形式的大数定律

由此，我们可以给出以下不同形式的**大数定律**。它们的差异在于对于随机变量序列 $\{X_n\}$ 的假设条件不同。

1. 在伯努利大数定律中，要求每个随机变量是独立同分布，且该分布是伯努利分布。为了放松**独立同分布**这一条件，我们有以下这个定理。

定理 12.2 (切比雪夫大数定律)

设 $\{X_n\}$ 为一列两两不相关的随机变量序列。若每个 X_i 的方差存在，且有共同的上界，即

$$Var(X_i) \leq ci = 1, 2, \cdots.$$

则 $\{X_n\}$ 服从（弱）大数定律。

证明 因为 $\{X_n\}$ 两两不相关，故

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \leq \frac{c}{n}.$$

根据切比雪夫不等式可知，对任意的 $\varepsilon > 0$，有

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right) \geq 1 - \frac{Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2} \geq 1 - \frac{c}{n\varepsilon^2}.$$

于是，当 $n \to \infty$ 时，有

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right) = 1.$$

注 切比雪夫大数定律只要求 $\{X_n\}$ 互不相关，并不要求 X_n 是独立的，也不要求它们是同分布的。因此，如果 X_n 是独立同分布的随机变量序列且方差有限，则 $\{X_n\}$ 必定服从大数定律。这表明了伯努利大数定律是切比雪夫大数定律的一个特例。

2. 在切比雪夫大数定律证明的过程中，我们不难发现，其核心在于使用了切比雪夫不等式。而切比雪夫不等式本质上证明的是大偏差存在的概率可以被其方差所约束。因此，我们只需要要求这些随机变量的算术平均数的方差不能过大。

在切比雪夫大数定律中要求了随机变量序列中两两不相关，这样可以使得随机变量的算术平均数的方差可以由每一个随机变量的方差计算而得。切比雪夫大数定律中又约束了每一个随机变量的方差存在且存在共同上界。由此，解决了随机变量的算术平均数的方差不能过大的问题。

事实上，切比雪夫大数定律的条件仍是比较严苛的，因为随机变量序列中仍需要它们是两两不相关的。但，这个条件并不是必要的。于是，我们考虑以下这个定理。

定理 12.3 (马尔可夫大数定律)

对随机变量序列 $\{X_n\}$，若马尔可夫条件即

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \to 0$$

满足，则 $\{X_n\}$ 服从（弱）大数定律。

注 马尔可夫大数定律对 $\{X_n\}$ 已经没有同分布、独立性、不相关的假定。

证明 (本定理的证明过程由学生课后自行补充)

3. 之前我们介绍的大数定律都假定了随机变量序列 $\{X_n\}$ 的方差存在。而我们也知道，对于一个随机变量而言，如果它的方差是存在的，那么它的期望一定是存在的。但反之不然。在大数定律的一般形式中，并未涉及随机变量序列的方差。所以，在随机变量序列的方差不存在的情况下，大数定律仍可以存在。

定理 12.4 (辛钦大数定律)

设 $\{X_n\}$ 为一独立同分布的随机变量序列，若 X_i 的数学期望存在，则 $\{X_n\}$ 服从（弱）大数定律。♡

注

- 在辛钦大数定律中，可以放松对随机变量方差存在性的假定，但要求 $\{X_n\}$ 为独立同分布的随机变量序列。
- 因为 $\{X_n\}$ 是同分布的随机变量序列，所以
$$E(X) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$
因此，从另一角度来看辛钦大数定律，可以计算一个随机变量期望 $E(X)$ 的近似值：对随机变量 X 有 n 次独立重复地观测，记第 k 次观测值为 X_k。则 $X_1, X_2, \cdots, X_n$ 是相互独立的，且它们的分布应该与 X 的分布相同。于是，在 $E(X)$ 存在的条件下，按照辛钦大数定律，当 n 足够大时，可以把这些观测的平均值
$$\frac{1}{n}\sum_{i=1}^{n} X_i$$
作为 $E(X)$ 的近似值。
- 类似的方法也可以推广到计算随机变量的 k 阶原点矩的近似值。

12.4 强大数定律

除了弱大叔定律之外，在概率论中，还可以有“强大数定律”的概念。

定理 12.5 (强大数定律)

设 $\{X_n\}$ 为一列独立同分布的随机变量序列，其期望为 $E(X)$，则

$$P\left(\lim_{n\to\infty}\frac{\sum_{i=1}^{n}X_i}{n}=E(X)\right)=1$$

问题 12.2 大数定律的强弱差异在哪?

第 13 讲 中心极限定理

阅读章节

- Intro to Prob 5.4
- Prob & Stat 4.4

13.1 中心极限定理的动机

我们知道，对于 n 个独立同分布的随机变量 $X_i, i=1,2,\cdots,n$，如果这些随机变量均服从正态分布，其中均值为 $E(X_1)=\mu$，方差为 $Var(X_1)=\sigma^2<\infty$，那么，其和为

$$S_n=\sum_{i=1}^{n} X_i$$

仍服从正态分布，其期望为 μ，而方差为 $n\sigma^2$。

笔记

问题 13.1 如果 X_i 的分布并不是正态分布，那么结论会有什么变化？

这里我们从一个很简单的例子来推测这个问题的答案。

例题 13.1 如果 X_i 都是独立同分布的随机变量，其分布为均匀分布 $U(0,1)$，$i=1,2,\cdots$，那么我们考虑以下随机变量的密度函数。

1. $S_1=X_1$。显然，S_1 的密度函数为

$$p_1(s)=\begin{cases}1, & 0<s<1,\\ 0, & \text{其他}.\end{cases}$$

2. $S_2=X_1+X_2$。首先，S_2 的取值范围为 $(0,2)$。其次，利用卷积公式，可知 S_2 的密度函数为

$$\begin{aligned}p_2(s) &= \int_{-\infty}^{\infty} p_{X_1}(s-x_2)p_{X_2}(x_2)\mathrm{d}x_2\\ &= \int_{-\infty}^{\infty} I(0<s-x_2<1)I(0<x_2<1)\mathrm{d}x_2\\ &= \int_{-\infty}^{\infty} I(0<s-x_2<1, 0<x_2<1)\mathrm{d}x_2\\ &= \int_{-\infty}^{\infty} I(\max\{s-1,0\}<x_2<\min\{s,1\})\mathrm{d}x_2\end{aligned}$$

我们发现，随着 s 的变化，x_2 的取值范围是不同的。

- 当 $0<s<1$ 时，有 $\max\{s-1,0\}=0$ 且 $\min\{s,1\}=s$，那么 S_2 的密度函数为
$$p_2(s)=s.$$
- 当 $1<s<2$ 时，有 $\max\{s-1,0\}=s-1$ 且 $\min\{s,1\}=1$，那么 S_2 的密度函数为
$$p_2(s)=2-s.$$

综述，S_2 的密度函数为
$$p_2(s)=\begin{cases} s, & 0<s<1,\\ 2-s, & 1\le s<2,\\ 0, & \text{其他}.\end{cases}$$

3. $S_3=X_1+X_2+X_3$。首先，S_3 的取值范围为 $(0,3)$。其次，$S_3=S_2+X_3$，且 S_2 与 X_3 独立。根据卷积公式，可知，
$$\begin{aligned} p_3(s) &= \int_{-\infty}^{\infty} p_{S_2}(s-x_3)p_{X_3}(x_3)\mathrm{d}x_3\\ &= \int_{-\infty}^{\infty} p_{S_2}(s-x_3)I(0<x_3<1)\mathrm{d}x_3.\end{aligned}$$
类似于 S_2 的讨论，我们需要考虑以下两组方程组，即
$$\begin{cases} 0<s-x_3<1,\\ 0<x_3<1.\end{cases} \quad \text{或} \quad \begin{cases} 1<s-x_3<2,\\ 0<x_3<1.\end{cases}$$
于是，我们需要讨论 s 不同的取值。

- 当 $0<s<1$ 时，则 $0<s-x_3<1$ 且 $0<x_3<s$。于是，S_3 的密度函数为
$$p_{S_3}(s) = \int_{-\infty}^{\infty} p_{S_2}(s-x_3)p_{X_3}(x_3)\mathrm{d}x_3=\int_0^s (s-x_3)\mathrm{d}x_3 = sx_3-\frac{1}{2}x_3^2\Big|_0^s=\frac{1}{2}s^2.$$
- 当 $1<s<2$ 时，$0<s-x_3<2$。而 $p_{S_2}(s)$ 在 $(0,2)$ 上是分段的。于是，S_3 的密度函数为
$$\begin{aligned} p_{S_3}(s) &= \int_{-\infty}^{\infty} p_{S_2}(s-x_3)p_{X_3}(x_3)\mathrm{d}x_3\\ &= \int_{s-1}^{1}(s-x_3)\mathrm{d}x_3+\int_0^{s-1}(2-(s-x_2))\mathrm{d}x_2\\ &= sx_3-\frac{1}{2}x_3^2\Big|_{s-1}^{1}+(2-s)x_3+\frac{1}{2}x_3^2\Big|_0^{s-1}\\ &= \left(s-\frac{1}{2}\right)-\left(s(s-1)-\frac{1}{2}(s-1)^2\right)+(2-s)(s-1)+\frac{1}{2}(s-1)^2\\ &= -s^2+3s-\frac{3}{2}.\end{aligned}$$

- 当 $2<s<3$ 时，则 $1<s-x_3<2$，且 $s-1<x_3<1$。于是，S_3 的密度函数为

$$\begin{aligned} p_3(s) &= \int_{-\infty}^{\infty} p_{S_2}(s-x_3)p_{X_3}(x_3)\mathrm{d}x_3 \\ &= \int_{s-2}^{1}(2-(s-x_2))\mathrm{d}x_2 \\ &= (2-s)x_3+\frac{1}{2}x_3^2\Big|_{s-2}^{1} \\ &= (2-s)+\frac{1}{2}-(2-s)(s-2)-\frac{1}{2}(s-2)^2 \\ &= s^2-5s+6+\frac{1}{2}-\frac{1}{2}s^2+2s-2 \\ &= \frac{1}{2}s^2-3s+\frac{9}{2} \end{aligned}$$

综上，S_3 的密度函数为

$$p_3(s)=\begin{cases} \frac{1}{2}s^2, & 0<s<1, \\ -s^2+3s-\frac{3}{2}, & 1\le s<2, \\ \frac{1}{2}s^2-3s+\frac{9}{2}, & 2\le s<3. \\ 0, & \text{其他}. \end{cases}$$

4. S_4 的密度函数为

$$p_4(s)=\begin{cases} \frac{1}{6}s^3, & 0<s<1, \\ -\frac{1}{2}s^3+2s^2-2s-\frac{2}{3}, & 1\le s<2, \\ \frac{1}{2}s^3-4s^2+10s-\frac{22}{3}, & 2\le s<3, \\ -\frac{1}{6}s^3+2s^2-16s+\frac{32}{3}, & 2\le s<3. \\ 0, & \text{其他}. \end{cases}$$

笔记 （S_4 的密度函数留作学生课后推导）

注

1. 随着 k 的增加，S_k 的密度函数图像越来越光滑，且越来越接近正态分布曲线。
2. 当 k 很大时，S_k 的密度函数是 $k-1$ 次多项式，形式复杂。

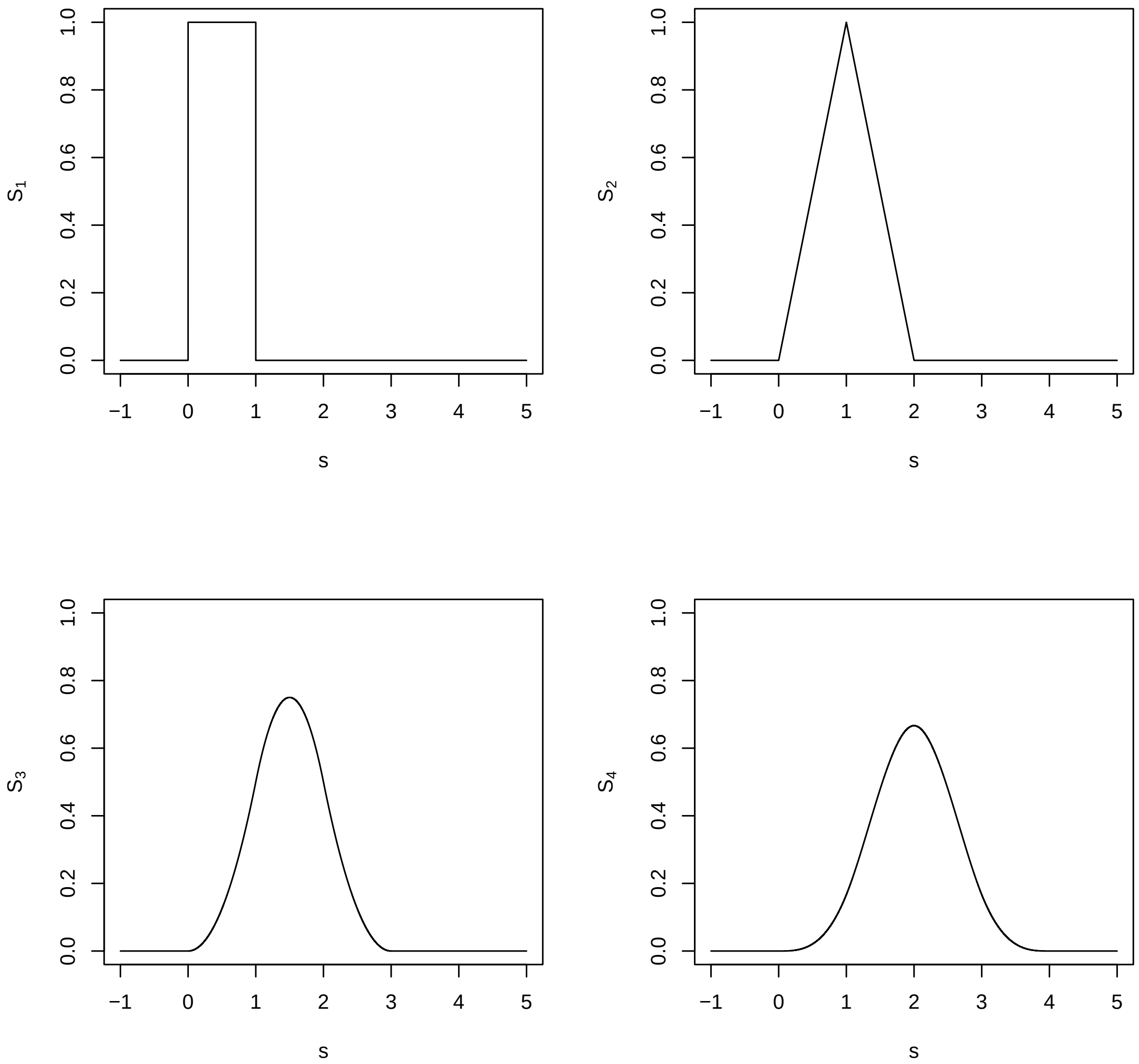

图 13.1: 均匀分布随机变量之和的密度函数

值得注意的是，当 k 增加时，S_k 的密度函数 $p_k(s)$ 的中心右移，且 $p_k(s)$ 的方差也增大。这意味着这个分布的中心趋向 ∞，其方差也趋向 ∞，分布极不稳定。因此，直接讨论 S_k 的分布是有困难的。

于是，在中心极限定理的研究中均对 S_k 进行标准化，即

$$S_k^* = \frac{S_k - E(S_k)}{\sqrt{Var(S_k)}}$$

可以证明

$$E(S_k^*) = 0$$

$$Var(S_k^*) = 1$$

一个很自然的问题是 S_k^* 的极限分布是否为标准正态分布 $N(0,1)$?

问题 13.2 中心极限定理本身研究的是**在什么条件下**，随机变量之和的极限分布是正态分布。

13.2 独立同分布下的中心极限定理

定理 13.1 (林德伯格-莱维（Lindeberg-Levy）中心极限定理)

设 $\{X_n\}$ 是独立同分布的随机变量序列，且 $E(X_i)=\mu, Var(X_i)=\sigma^2>0$ 存在，若记

$$S_n^*=\frac{\sum_{i=1}^n X_i-n\mu}{\sigma\sqrt{n}},$$

则对任意实数 y，有

$$\lim_{n\to\infty}P(S_n^*\le s)=\Phi(s)$$

注

1. 这个定理的证明需要利用特征函数，学生可以参考《概率论与数理统计教程》第 212 页。
2. 林德伯格-莱维中心极限定理表示独立同分布，且二阶矩存在的随机变量序列是满足中心极限定理。

以下，我们讨论一种特定的分布——二项分布。

定理 13.2 (棣莫弗-拉普拉斯（de Moivre-Laplace）中心极限定理)

设 n 重伯努利实验中，事件 A 在每次实验中出现的概率为 $p(0<p<1)$，记 S_n 为 n 次实验中实验 A 出现的次数，且记

$$S_n^*=\frac{S_n-np}{\sqrt{np(1-p)}}$$

则对任意实数 y，有

$$\lim_{n\to\infty}P(S_n^*\le s)=\Phi(s).$$

注

1. 二项分布的近似分布，可以采用泊松分布，也可以使用正态分布。一般来说，在 p 比较小时，用泊松分布近似较好；而在 $np>5$ 和 $n(1-p)>5$ 时，用正态分布比较好。
2. 因为二项分布是离散分布，而正态分布是连续分布，在近似时通常可以做一些修正来提高精度。对任意 $s_1<s_2$，$P(s_1\le S_n\le s_2)$ 的修正方案为

$$P(s_1-0.5\le S_n\le s_2+0.5)=\Phi\left(\frac{s_2+0.5-np}{\sqrt{np(1-p)}}\right)-\Phi\left(\frac{s_1-0.5-np}{\sqrt{np(1-p)}}\right).$$

例题 13.2 若 $S_n\sim b(25,0.4)$，那么 $P(5\le S_n\le 15)=0.9774$。我们可以计算

$$\begin{aligned}\text{修正后：}P(5\le S_n\le 15)&=\Phi\left(\frac{15+0.5-10}{\sqrt{25*0.4*0.6}}\right)-\Phi\left(\frac{5-0.5-10}{\sqrt{25*0.4*0.6}}\right)=0.9753,\\ \text{未修正：}P(5\le S_n\le 15)&=\Phi\left(\frac{15-10}{\sqrt{25*0.4*0.6}}\right)-\Phi\left(\frac{5-10}{\sqrt{25*0.4*0.6}}\right)=0.9588.\end{aligned}$$

同时，这种修正方案也可以适用于计算 $P(S_n=k)$ 的情况。

例题 13.3 若 $S_n \sim b(25, 0.4)$，那么 $P(S_n = 10) = 0.1612$。根据修正方案，我们可以计算

$$\Phi\left(\frac{10+0.5-10}{\sqrt{25*0.4*0.6}}\right) - \Phi\left(\frac{10-0.5-10}{\sqrt{25*0.4*0.6}}\right) = 0.1617.$$

13.3 棣莫弗-拉普拉斯中心极限定理的应用案例

回顾一下棣莫弗-拉普拉斯中心极限定理，我们知道

$$P(S_n^* \leq s) = \Phi(s).$$

注意到，$\Phi(s)$ 本质上是一个概率值。对于给定概率 $\Phi(s)$ 时，s 可以认为是相应的分位数。而二项分布中由两个参数 n 和 p，其中 n 表示实验次数，也通常被认为是样本量（这个概念将会在第 14 讲中给出）。

于是，在棣莫弗-拉普拉斯中心极限定理的应用，我们要解决以下三个问题：

1. 求分位数；
2. 求概率；
3. 求样本量。

13.3.1 求分位数

例题 13.4 某奶茶店共售卖两种奶茶：珍珠奶茶和纯奶茶。其差异在于是否放珍珠。假设每位顾客点任何一种奶茶是等可能的，且点单是相互不受影响。已知每杯珍珠奶茶中需放入 5 克珍珠。请问，至少要准备多少克珍珠，才可以有 95% 的可能性，接受 100 杯奶茶的点单？

解 设第 i 位顾客点了珍珠奶茶为 X_i，$X_i \sim b(1, p_0), p_0 = 0.5$，即

$$X_i = \begin{cases} 1, & \text{如果第 i 位顾客点了珍珠奶茶} \\ 0, & \text{如果第 i 位顾客点了纯奶茶} \end{cases}$$

所以，100 位顾客中共有

$$\sum_{i=1}^{100} X_i \sim b(100, p_0)$$

点了珍珠奶茶。设我们拟准备 s 克珍珠。于是，s 克珍珠能够满足 100 位顾客的点单，即

$$\begin{aligned} 95\% &\leq P\left(5 \times \sum_{i=1}^{100} X_i \leq s\right) \\ &= P\left(\sum_{i=1}^{100} X_i \leq \frac{s}{5}\right) \\ &= P\left(\frac{\sum_{i=1}^{100} X_i - 50}{\sqrt{25}} \leq \frac{\frac{s}{5}+0.5-50}{\sqrt{25}}\right) \\ &= \Phi\left(\frac{s}{25} - 9.9\right) \end{aligned}$$

可以解得

$$s = (\Phi^{-1}(0.95) + 0.99) \times 25 = 288.62 \approx 290.$$

13.3.2 求概率

例题 13.5 已知我们预定了 290 克珍珠，共 58 杯珍珠奶茶。某公司一次性订购 100 杯奶茶，该公司的员工更加偏好珍珠奶茶。假设每一杯珍珠奶茶的点单可能性提高到了 60%，请问 290 克是否能够满足这 100 杯奶茶的订单？

解 由于 $X_i \sim b(1, p_0)$，这里 $p_0 = 0.6$。我们考虑以下概率

$$P\left(\sum_{i=1}^{100} X_i \leq 58\right).$$

这个概率可以直接利用二项分布来计算，即

$$P\left(\sum_{i=1}^{100} X_i \leq 58\right) = \sum_{k=0}^{58} 0.6^{58} 0.4^{100-58} = 0.3775.$$

也可以用正态分布来近似计算，即

$$\begin{aligned} P\left(\sum_{i=1}^{100} x_i \leq 58\right) &= P\left(\frac{\sum_{i=1}^{100} x_i - 60}{\sqrt{24}} \leq \frac{58 + 05 - 60}{\sqrt{24}}\right) \\ &= 0.3797. \end{aligned}$$

13.3.3 求样本量

例题 13.6 结合上述两个例子，我们发现顾客的奶茶偏好对备料十分重要。于是，我们需要调研一下顾客的奶茶偏好。假定顾客的真实奶茶偏好为 p。理解为本奶茶店接单中每杯奶茶是珍珠奶茶的概率是 p。我们通常可以利用历史订单得到概率 p 的估计值 $\hat{p}$。请问至少要看多少历史订单数量（多少杯奶茶），要保证有 95% 的把握使得估计值 $\hat{p}$ 与真实的 p 之间的差异不超过 1%?

解 令 X_i 表示第 i 杯奶茶是珍珠奶茶，即

$$X_i = \begin{cases}1, & \text{如果第 } i \text{ 杯奶茶是珍珠奶茶;} 0, \quad \text{其他.}\end{cases}$$

于是，$X_i \sim b(1, p)$，p 表示珍珠奶茶的点单概率。假设共需要看 n 杯奶茶的历史订单信息，则

$$S_n = \sum_{i=1}^{n} X_i$$

表示 n 杯奶茶中的珍珠奶茶的数量，即 $S_n \sim b(n, p)$。根据大数定律 (LLN)，当 n 很大时，

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} p.$$

根据中心极限定理（CLT），

$$\begin{aligned} 95\% &\leq P\left(\left|\frac{S_n}{n}-p\right| \leqslant 0.01\right) \\ &= = P\left(\left|\frac{S_n/n-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq \frac{0.01}{\sqrt{\frac{p(1-p)}{n}}}\right) \\ &= 2\Phi\left(0.01\sqrt{\frac{n}{p(1-p)}}\right)-1, \end{aligned}$$

于是，

$$\Phi\left(0.01\sqrt{\frac{n}{p(1-p)}}\right) \geq 0.975$$

我们可以解出

$$\frac{n}{p(1-p)} \geq \left(\frac{\Phi^{-1}(0.975)}{0.01}\right)^2.$$

注意到

$$p(1-p) \leqslant \frac{1}{2} \times \frac{1}{2} = 0.25$$

所以，

$$n \geq \left(\frac{\Phi^{-1}(0.975)}{0.01}\right)^2 \times 0.25 = 9604.$$

因此，至少要看 9604 杯奶茶的历史订单数量，要保证有 95% 的把握使得估计值 $\hat{p}$ 与真实的 p 之间的差异不超过 1%。

13.4 其他条件下的中心极限定理

除了满足独立同分布的随机变量序列之外，其他类型的随机变量序列也可以满足中心极限定理。以下两个定理供学生课后自学。

定义 13.1 (林德伯格条件)

假设一个随机变量序列 $\{X_n\}$，其期望 $E(X_i)=\mu_i$ 且方差 $Var(X_i)=\sigma_i^2$，$i=1,2,\cdots$。令 $S_n=\sum_{i=1}^n X_i$，且 $B_n=\sigma(S_n)=\sqrt{Var(S_n)}$。如果对于任意 $\tau>0$，有

$$\lim_{n\to\infty}\frac{1}{\tau^2 B_n}\sum_{i=1}^n\int_{|x-\mu_i|>\tau B_n}(x-\mu_i)^2 p_i(x)\mathrm{d}x = 0.$$

那么称该随机变量序列满足林德伯格条件。♣

注 林德伯格条件保证了 $S_n^*=B_n^{-1}\sum_{i=1}^n(X_i-\mu_i)$ 中各加项“均匀地小”。

定理 13.3 (林德伯格中心极限定理)

设独立随机变量序列 $\{X_n\}$ 满足林德伯格条件，则对任意的 x，有

$$\lim_{n\to\infty} P\left(\frac{1}{B_n}\sum_{i=1}^{n}(X_i-\mu_i)\leq x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}\mathrm{d}t.$$

♡

定理 13.4 (李雅普诺夫中心极限定理)

设 $\{X_n\}$ 为独立随机变量序列，若存在 $\delta>0$，满足

$$\lim_{n\to\infty}\frac{1}{B_n^{2+\delta}}\sum_{i=1}^{n}E\left(|X_i-\mu_i|^{2+\delta}\right)=0,$$

则对任意的 x，有

$$\lim_{n\to\infty} P\left(\frac{1}{B_n}\sum_{i=1}^{n}(X_i-\mu_i)\leq x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}\mathrm{d}t.$$

♡

第 14 讲 数理统计绪论

阅读章节

- Prob & Stat 5.1

问题 14.1 当下，没有人能够忽略数据的价值。在本课程中，我们需要思考以下三个问题：

1. 数据是什么？
2. 如何分析数据？
3. 得到怎样的结论？

我们来看看以下三个例子。

例题 14.1（总统选举） 若有一个"乌托邦"的国家，通过民意来选举总统。目前有两位总统候选人：A 和 B。假定每位公民只能从这两位候选人中进行投票。在唱票前，没有人能够确定谁能够当选。于是，我们可以考虑候选人 A 能够当选，来构造为一个随机变量。也就是说，A 是否能够当选可以看作一个伯努利分布的随机变量 $b(1,p)$。而每一位在乌托邦具有投票权的公民的投票结果是**个体**，而全体公民的投票可以看作一个**总体**。为了研究总体，我们需要通过观测到的数据来了解这个总体。

在唱票之前，我们可以采访部分投票公民，从而得知他们的投票结果。假设共采访了 n 位公民，得到了他们的投票结果。可以看作 n 次独立的观测：$x_1, x_2, \cdots, x_n$。在统计学中，称 $x_1, x_2, \cdots, x_n$ 为**样本**，也是我们所观测到的数据。除了**调查**之外，我们还可以通过**实验**来获得样本。

问题 14.2 怎样的样本才能够帮助我们研究总体？

例题 14.2（总统选举（续））假定在乌托邦里有两个城镇：T_1, T_2，两个城镇里各有 10 个百姓。通过唱票结果发现：因为候选人 A 来自于城镇 T_1，他们中有 7 名会为候选人 A 投票，剩下的投票给了候选人 B；而候选人 B 来自于城镇 T_2，他们中 9 人会为 B 投票，只有 1 人投票给了 A。如果我们只调研城镇 T_1 的公民，很自然会认为 A 当选，但实际结果相反。

由此我们知道，只有当样本具有**代表性**，基于样本对总体的推断才有意义。

一般来说，样本是通过总体经过简单抽样而获得的，也就是说，样本之间是独立同分布的，其分布与总体的分布相同。通常来说，样本具有二重性：可观测性和随机性。这两种性质是**同等重要的**。在数据分析中，需要得到具体的结果并进行合理的结论，主要依赖于样本的可观测性；而如何进行分析，需要事先确定数据分析的方法，这主要依赖于样本的随机性。

于是，在数理统计中，对于样本 x，它既可以看作随机变量，可以看作观测值。初学者需要注意这个问题。偶尔在推理过程中借助于 X 表示随机变量来辅助学生理解。

例题 14.3（测量） 每天早上起床我们可以多次测量自己的身高。如果使用精度高的测量仪器来进行数据采集，可以得到一系列数据：166.01 厘米，165.98 厘米，165.89 厘米，166.11 厘米，165.96 厘米，$\cdots$。假设我们测量了 100 万次，那么共有 100 万个数据。于是，我们可以得到一个分布。这里我们可以认为自己的身高是一个总体，其有一个均值，表示我们真实的

高度，而每次测量是存在一些偏差。所以，我们定义所测量的数据为

$$x = \mu + \varepsilon$$

这个式子是所有测量的最简单结构式，放之四海皆准。其中，x 表示观测值（可以观测到的），μ 表示真值（永远未知），ε 表示误差（需要假定）。

为了认识 μ，我们需要对误差进行一些假定，才能对 μ 做出合理的推断。

假定一：$\varepsilon \sim N(0, \sigma^2)$，其中 σ^2 是未知的。在这个假定中我们利用高斯分布（正态分布）$N(\mu, \sigma^2)$ 来刻画测量数据。高斯分布由两个参数唯一确定，其中，μ 是我们感兴趣的参数，称为目标参数；而 σ^2 是我们不感兴趣的参数，称为讨厌参数（或冗余参数）。

假定二：$\varepsilon \sim N(0, \sigma_0^2)$，其中 σ_0^2 是已知的。

假定三：ε 服从一个以 0 为中心的对称分布，即 x 服从一个以 μ 为中心的对称分布。这里我们并未对数据的分布给出一个具体的假定。与假设一和假设二中参数总体的假定不同，假设三中是非参数总体的假定。

例题 14.4 1979 年 4 月 17 日，日本《朝日新闻》曾刊登过这样的一条消息：美国人喜欢购买日本索尼工厂生产的彩电，而不愿购买设在美国加州的索尼工厂生产的彩电。而美国本土生产的电视机出厂合格率为 100%，日本产的合格率只有 99.73%. 这究竟是什么原因呢？美国一家咨询公司采用统计抽样的方法，对此进行了专题调查分析。结果发现，两地生产彩电质量特征的概率分布不同。

在美国的工厂中，采用门柱法来进行产品质量管理；而在日本的工厂中，采用 $3-\sigma$ 质量管理策略：将产品划分优、良和合格三类，严格控制每一类产品在市场的占比。基于 $3-\sigma$ 质量管理策略，日本产品的优良率明显高于美国本土生产的产品。所以更受到美国消费者的青睐。

笔记

第 15 讲 经验分布函数、统计量及其分布

阅读章节

- Prob & Stat 5.2，5.3

15.1 经验分布函数

笔记

问题 15.1 基于样本 $x_1, x_2, \cdots, x_n$，我们能否定义数据的分布函数?

定义 15.1 (经验分布函数)

设 $x_1, x_2, \cdots, x_n$ 是取自总体分布函数 $F(x)$ 的样本，若将样本的观测值由小到大进行排列，记为 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，则称 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 为有序样本，用有序样本定义函数

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}, \\ k/n, & \text{当 } x_{(k)} \le x < x_{(k+1)}, k = 1, 2, \cdots, n-1, \\ 1, & \text{当 } x \ge x_{(n)}. \end{cases}$$

则 $F_n(x)$ 是样本的经验分布函数。♣

笔记 $F_n(x)$ 是一个分布函数。

注

1. 对固定的 x，$F_n(x)$ 是样本中事件 $\{x_i \le x\}$ 发生的频率。
2. 当 n 固定时，$F_n(x)$ 是样本的函数，而样本是随机变量，所以 $F_n(x)$ 也是一个随机变量。

若对任意给定的实数 x，定义

$$I_i(x)=\begin{cases}1, & x_i\leq x,\\ 0, & x_i>x.\end{cases}$$

则经验分布函数的定义可以看出，

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_i(x).$$

注意到 $I_i(x)$ 是独立同分布的随机变量，其共同分布为 $b(1,F(x))$.

3. 由伯努利大数定律可知，只要 n 充分大，$F_n(x)$ 依概率收敛于 $F(x)$.

比较一下分布函数 $F(x)$ 和经验分布函数 $F_n(x)$ 的区别：

1. $F_n(x)$ 不是 $F(x)$
2. 但 $F_n(x)$ 可代表 $F(x)$
3. $F_n(x)$ 可观测到，而 $F(x)$ 不可观测到

定理 15.1 (格利文科定理)

设 $x_1,x_2,\cdots,x_n$ 是取自总体分布函数为 $F(x)$ 的样本，$F_n(x)$ 是其经验分布函数，当 n 充分大时，$F_n(x)$ 能充分逼近 $F(x)$.

注 格利文科定理保证了经典统计中一切统计推断都以样本为依据。

定义 15.2 (统计量)

设 $x_1,x_2,\cdots,x_n$ 为取自某总体的样本，若样本函数

$$T=T(x_1,x_2,\cdots,x_n)$$

中不含任何未知参数，则称 T 为统计量。而统计量的分布称为抽样分布。

15.2 样本矩

假定在样本 $x_1,x_2,\cdots,x_n$ 中有 k 个不同的取值 $a_1,a_2,\cdots,a_k$。每一个样本 x_i 来自于一个均值为 μ 和方差为 σ^2 分布的随机变量。

数据	a_1	a_2	$\cdots$	a_k
频数	m_1	m_2	$\cdots$	m_k
频率	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\cdots$	$\frac{m_k}{n}$

其中，$\sum_{i=1}^{k}m_i=n,\sum_{i=1}^{k}a_im_i=\sum_{i=1}^{n}x_i$。

因为 $F_n(x)$ 是一个某个随机变量的分布函数，假定该随机变量为 X，所以，

$$\begin{aligned}E_{F_n}(X) &= \sum_{i=1}^{k}a_i\cdot\frac{m_i}{n}=\frac{1}{n}\sum_{i=1}^{k}a_im_i=\frac{1}{n}\sum_{i=1}^{n}x_i\\ E_{F_n}\left(X^l\right) &= \sum_{i=1}^{k}a_i^l\frac{m_i}{n}=\frac{1}{n}\sum_{i=1}^{k}a_i^lm_i=\frac{1}{n}\sum_{i=1}^{n}x_i^l\end{aligned}$$

$$\begin{aligned}
Var_{F_n}(X) &= \sum_{i=1}^{k}(a_i - E_{F_n}(X))^2 \cdot \frac{m_i}{n} \\
&= \sum_{i=1}^{k}(a_i - \bar{x})^2 \cdot \frac{m_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{k}(a_i - \bar{x})^2 \cdot m_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2
\end{aligned}$$

注 我们通常记 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，而 $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$。

定义 15.3

设 $x_1, x_2, \cdots, x_n$ 是样本，k 为正整数，则统计量

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

称为样本 k 阶原点矩；统计量

$$b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

称为样本 k 阶中心矩。♣

注 样本一阶原点矩是样本均值；样本二阶中心矩是样本方差。

15.2.1 样本均值

样本均值 $\bar{x}$ 是最简单的统计量。首先考察其期望和方差。我们可以证明

$$\begin{aligned}
E(\bar{x}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \mu \\
Var(\bar{x}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{\sigma^2}{n}.
\end{aligned}$$

其次，我们考虑 $\bar{x}$ 的分布。

1. 精确分布：如果 $x_1, x_2, \ldots, x_n$ 均来自正态分布 $N(\mu, \sigma^2)$，那么 $\bar{x}$ 的分布是 $N(\mu, \sigma^2/n)$。
2. 近似分布 1（渐近分布）：如果 $x_1, x_2, \ldots, x_n$ 均来自未知分布 $\Pi(\mu, \sigma^2)$，那么，根据 CTL，
$$\frac{\bar{x} - E(\bar{x})}{\sqrt{Var(\bar{x})}} \xrightarrow{L} N(0, 1).$$
3. 近似分布 2（蒙特卡洛分布）

15.2.2 样本方差和样本标准差

样本方差 s_n^2 也是一种常用的统计量。为了考虑 s_n^2 的期望，我们可以计算偏差平方和 $\sum_{i=1}^n (x_i-\bar{x})^2$ 的期望，即

$$\begin{aligned}
E\left(\sum_{i=1}^n (x_i-\bar{x})^2\right) &= E\left(\sum_{i=1}^n \left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\right)\\
&= E\left(\sum_{i=1}^n x_i^2-2n\bar{x}^2+n\bar{x}^2\right)\\
&= E\left(\sum_{i=1}^n x_i^2-n\bar{x}^2\right)\\
&= \left(\sum_{i=1}^n E\left(x_i^2\right)-nE\left(\bar{x}^2\right)\right)\\
&= \left[\sum_{i=1}^n \left(Var\left(x_i\right)+E^2\left(x_i\right)\right)-n\left(E^2(\bar{x})+Var(\bar{x})\right)\right]\\
&= \cdot\left(n\cdot\left(\sigma^2+\mu^2\right)-n\left(\mu^2+\frac{\sigma^2}{n}\right)\right)\\
&= (n-1)\sigma^2
\end{aligned}$$

因此，我们发现 $E(s_n^2)=(n-1)\sigma^2/n\neq\sigma^2$。在实际中，我们常采用

$$\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2$$

作为样本方差。而定义样本标准差为 $s=\sqrt{s^2}$。

15.2.3 样本偏度和样本峰度

定义 15.4

设 $x_1,x_2,\cdots,x_n$ 是样本，则称统计量

$$\hat{\beta}_s=\frac{b_3}{b_2^{3/2}}$$

为样本偏度；称统计量

$$\hat{\beta}_k=\frac{b_4}{b_2^2}-3$$

为样本峰度。 ♣

问题 15.2 如何得到样本偏度 $\hat{\beta}_s$ 的分布?

即便数据来自于标准正态分布，其样本偏度的分布仍是很难解决的问题，这里我们介绍如何用蒙特卡洛方法来得到抽样分布。具体步骤如下：

1. 生成 n 个来自于 $N(0,1)$ 的随机数；
2. 计算这 n 个样本的样本偏度 $\hat{\beta}_s$；

3. 重复第 1 和第 2 步 M 次。

于是，可以得到 M 个 $\hat{\beta}_s$ 的观测值，从而可以得到 $\hat{\beta}_s$ 的蒙特卡洛分布。**注** 如何保留蒙特卡洛分布的信息？

15.3 次序统计量

例题 15.1 我们共有 n 个学生参加本学期的期中考试，记为 $x_1, x_2, \cdots, x_n$。常常会对学生的考试成绩进行排名，学生的考试成绩可以按从小到大进行排序，于是可以得到一组有序样本 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$。我们记 $y_i = x_{(i)}$。易知，$y_1, y_2, \cdots, y_n$ 既不独立也不同分布。

定义 15.5

设 $x_1, x_2, \cdots, x_n$ 是来自于总体 X 的样本，$x_{(i)}$ 称为样本的第 i 个次序统计量，它表示将样本观测值从小到大排序后得到的第 i 个观测值，其中 $x_{(1)} = \min\{x_1, x_2, \cdots, x_n\}$ 称为该样本的最小次序统计量，$x_{(n)} = \max\{x_1, x_2, \cdots, x_n\}$ 称为该样本的最大次序统计量，$(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ 称为该样本的次序统计量。♣

问题 15.3

1. 如何求 $x_{(k)}$ 的分布？
2. 如何求 $x_{(n)} - x_{(1)}$ 的分布？

定理 15.2

设总体 X 的密度函数为 $p(x)$，分布函数为 $F(x)$，$x_1, x_2, \cdots, x_n$ 为样本，则第 k 个次序统计量 $x_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(1-F(x))^{n-k}p(x).$$

♡

证明 因为我们需要将 $x_{(k)}$ 看成一个随机变量，所以记 $X_{(k)}$ 为第 k 个次序统计量。首先考虑 $X_{(k)}$ 的分布函数 $F_k(x) = P(X_{(k)} \le x)$。注意到事件

$$\{X_{(k)} \le x\} \Leftrightarrow \{x_1, x_2, \cdots, x_n \text{中至少有 } k \text{ 个} \le x\} \Leftrightarrow \cup_{j=k}^{n} \{x_1, x_2, \cdots, x_n \text{中恰好有 } j \text{ 个} \le x\}$$

所以，

$$\begin{aligned} F_k(x) &= \sum_{j=k}^{n} P\left(x_1, x_2, \cdots, x_n \text{中恰好有 } j \text{ 个} \le x\right) \\ &= \sum_{j=k}^{n} C_n^j (F(x))^j (1-F(x))^{n-j} \end{aligned}$$

根据引理 2.1，我们可知

$$F_k(x) = \int_0^{F(x)} \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \mathrm{d}x.$$

于是，定理得证。

引理 15.1

对于 $0<p<1$，有

$$\sum_{j=k}^{n} C_n^j p^j (1-p)^{n-j} = \int_0^p \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \mathrm{d}x$$

证明 令

$$g(p) = \sum_{j=k}^{n} C_n^j p^j (1-p)^{n-j}.$$

于是，关于 $g(p)$ 对 p 求导，即

$$\begin{aligned}\frac{\partial}{\partial p} g(p) &= \frac{\partial}{\partial p}\left(\sum_{j=k}^{n} C_n^j p^j (1-p)^{n-j}\right) \\ &= \sum_{j=k}^{n}\left(\frac{\partial}{\partial p} p^j (1-p)^{n-j}\right)\end{aligned}$$

注意到，第 k 项为

$$C_n^k \left(k p^{k-1}(1-p)^{n-k} - (n-k)p^k(1-p)^{n-k-1}\right),$$

而第 $k+1$ 项为

$$C_n^{k+1}\left((k+1)p^k(1-p)^{n-k-1} - (n-k-1)p^{k+1}(1-p)^{n-k-2}\right).$$

同时，

$$C_n^k p^k (n-k)(1-p)^{n-k-1} = C_n^{k+1}(k+1)p^k(1-p)^{n-k-1}.$$

所以，进行前后消除，我们有

$$\frac{\partial}{\partial p} g(p) = \frac{n!}{(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k}.$$

我们可以验证其初始值相等即可。

笔记

定理 15.3

对于统计量 $(x_{(i)}, x_{(j)})(i<j)$ 得联合分布密度函数为

$$\begin{aligned} p_{ij}(y,z) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ &\quad (F(y))^{i-1}(F(z)-F(y))^{j-i-1}(1-F(z))^{n-j}p(y)p(z), y \le z. \end{aligned}$$

笔记

例题 15.2 设总体分布为 $U(0,1)$，$x_1, x_2, \cdots, x_n$ 为其样本，则

1. $x_{(k)}$ 的分布是 $Be(k, n-k+1)$.

   因为 x_i 的分布函数为

$$F(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < 1 \\ 1, & x \ge 1. \end{cases}$$

   所以，$x_{(k)}$ 的密度函数为

$$\begin{aligned} p_k(x) &= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{(n-k+1)-1}, 0 < x < 1. \end{aligned}$$

2. $(x_{(k)}, x_{(s)})$ 的联合密度函数为

$$p_{k,s}(x,y) = \frac{n!}{(k-1)!(s-k-1)!(n-s)!} x^{k-1}(y-x)^{s-k-1}(1-y)^{n-s}, 0 \le x \le y \le 1.$$

3. 若 $Y = x_{(s)} - x_{(k)}$。令 $U = x_{(k)}$。则 Y 的密度函数为

$$\begin{aligned}
p_Y(y) &= \int_0^{1-y} p(y,u)\mathrm{d}u \\
&= \int_0^{1-y} \frac{n!}{(k-1)!(s-k-1)!(n-s)!} u^{k-1} y^{s-k-1} (1-u-y)^{n-s} \mathrm{d}u \\
&\overset{u=(1-y)v}{=} \frac{n!}{(k-1)!(s-k-1)!(n-s)!} \\
&\quad \cdot \int_0^1 (1-y)^{k-1} v^{k-1} y^{s-k-1} (1-y)^{n-s} (1-v)^{n-s} (1-y)\mathrm{d}v \\
&= y^{s-k-1}(1-y)^{n-s+k} \int_0^1 \frac{n!}{(k-1)!(s-k-1)!(n-s)!} v^{k-1}(1-v)^{n-s}\mathrm{d}v \\
&= y^{s-k-1}(1-y)^{n-s+k} \cdot \frac{n!}{(k-1)!(s-k-1)!(n-s)!} \cdot \frac{(k-1)!(n-s)!}{(n+k-s)!} \\
&\quad \int_0^1 \frac{\Gamma(n+k-s+1)}{\Gamma(k)\Gamma(n-s+1)} v^{k-1}(1-v)^{(n-s+1)-1}\mathrm{d}v \\
&= \frac{\Gamma(n+1)}{\Gamma(s-k)\Gamma(n-s+k+1)} y^{(s-k)-1}(1-y)^{(n-s+k+1)-1}
\end{aligned}$$

因此，Y 的分布是 $Be(s-k, n-s+k+1)$。

注 均匀分布的次序统计量是贝塔分布的来源之一。

注

1. 经验分布函数是次序统计量的函数，即

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{x_i \le x\}} = \frac{1}{n}\sum_{i=1}^n I\left\{X_{(i)} \le x\right\}.$$

2. 样本分位数也是基于次序统计量而定义。

15.3.1 样本分位数

定义 15.6

$m_{0.5}$ 定义如下：

$$m_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\text{为奇数}, \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n\text{为偶数}. \end{cases}$$

称 $m_{0.5}$ 为中位数。

♣

定义 15.7

m_p 定义如下：

$$m_p = \begin{cases} x_{([np+1])} & \text{若 } np \text{ 不为整数}, \\ \frac{1}{2}\left(x_{([np])} + x_{([np+1])}\right) & \text{若 } np \text{ 为整数}. \end{cases}$$

称 m_p 为样本 p 分位数。

♣

对于样本分位数，我们也有相应的渐近分布，如下定理，供学生进行选修。

定理 15.4

设总体密度函数为 $p(x)$，x_p 为其 p 分位数，$p(x)$ 在 x_p 处连续且 $p(x_p)>0$，则当 $n\to\infty$ 时，样本 p 分位数 m_p 的渐近分布为

$$m_p \dot{\sim} N\left(x_p, \frac{p(1-p)}{np^2(x_p)}\right).$$

特别地，对于样本中位数，当 $n\to\infty$ 时近似地有

$$m_{0.5} \dot{\sim} N\left(x_{0.5}, \frac{1}{4np^2(x_{0.5})}\right).$$

♡

第 16 讲 三大抽样分布

阅读章节

- ❑ Prob & Stat 5.4

在本课程中，大部分的推断是基于总体分布是正态分布这一假定。本讲中，我们重点介绍正态分布的样本方差的分布。首先我们介绍三种“新”的分布，它们是常见抽样分布。与之前学习思路不同，这些分布是通过我们已经学过的分布构造出来的，所以，我们应关注这些分布的构造方式。

问题 16.1 对于正态分布随机变量 X 而言，如果 X 的分布是 $N(\mu,\sigma^2)$，其中，$\mu=E(X),\sigma^2=Var(X)$。那么，经过标准化，

$$\frac{X-\mu}{\sqrt{\sigma^2}}\sim N(0,1).$$

而我们经过观测，得到一组样本：$x_1,x_2,\cdots,x_n\overset{iid}{\sim}N(\mu,\sigma^2)$。我们可以计算样本均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$ 和样本方差 $s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\bar{x})^2$。很自然有一个问题

$$\frac{x_1-\bar{x}}{\sqrt{s^2}}$$

的分布是什么？仍是正态分布吗？

16.1 三种分布的构造方式

16.1.1 卡方分布

定理 16.1

设 $X_1,X_2,\cdots,X_n$ 独立同分布于标准正态分布 $N(0,1)$，则称

$$\chi^2=X_1^2+X_2^2+\cdots+X_n^2=\sum_{i=1}^n X_i^2$$

的分布为自由度为 n 的 χ^2 分布，记为 $\chi^2\sim\chi^2(n)$.

注

1. 若 $X\sim N(0,1)$，则 $X^2\sim Ga(1/2,1/2)$；
2. 伽马分布具有可加性，可加性；
3. $E(\chi^2)=n,Var(\chi^2)=2n$.

16.1.2 F 分布

定义 16.1 (F 分布)

设随机变量 $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$，且 X_1 与 X_2 独立，则称

$$F = \frac{X_1/m}{X_2/n}$$

的分布是自由度为 m 与 n 的 F 分布，记为 $F(m,n)$。其中，m 为分子自由度，n 为分母自由度。♣

注

1. F 的密度函数为
$$p_F(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}.$$
2. F 分布是一个偏态分布。
3. 若 $F \sim F(m,n)$，则 $1/F \sim F(n,m)$；
4. 令 $F_\alpha(m,n)$ 为自由度为 m 和 n 的 F 分布分位数。于是，
$$\alpha = P(1/F \leq F_\alpha(n,m)) = P(F \geq \frac{1}{F_\alpha(n,m)}).$$
从而
$$P\left(F \leq \frac{1}{F_\alpha(n,m)}\right) = 1-\alpha.$$
因此，
$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}.$$

16.1.3 t 分布

定义 16.2 (t 分布)

设随机变量 X_1 与 X_2 独立且 $X_1 \sim N(0,1), X_2 \sim \chi^2(n)$，则称

$$t = \frac{X_1}{\sqrt{X_2/n}}$$

的分布为自由度为 n 的 t 分布，记为 $t \sim t(n)$。♣

注

1. t 的密度函数为
$$p_t(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < y < \infty.$$
2. t 分布的密度函数的图像也是偶函数，与标准正态分布的密度函数形状类似。特别当 n 比较小时，尾部概率比标准正态分布的大一些。

16.2 正态分布下样本方差 s^2 的分布

定理 16.2

设 $x_1, x_2, \cdots, x_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，其样本均值和样本方差分别为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad 和 \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

则有

1. $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;
2. $\frac{(n-1)s^2}{\sigma^2} = \frac{\sum(x_i-\bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$;
3. $\bar{x}$ 与 s^2 独立。

推论 16.1

设 $x_1, x_2, \cdots, x_n$ 是来自正态分布 $N(\mu, \sigma^2)$ 的一组样本，$\bar{x}$ 和 s^2 分别是该样本的样本均值和样本方差，则有

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t(n-1)$$

推论 16.2

设 $x_1, x_2, \cdots, x_m$ 是来自 $N(\mu_1, \sigma_1^2)$ 的样本，$y_1, y_2, \cdots, y_n$ 是来自 $N(\mu_2, \sigma_2^2)$ 的样本，且两样本相互独立，记

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 \quad 和 \quad s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

其中，$\bar{x} = m^{-1}\sum_{i=1}^{m} x_i$ 和 $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$，则有

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(m-1, n-1).$$

特别地，当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时，并记

$$s_w^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{m+n-2}$$

则

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2}{s_w\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

笔记

接下来，我们给出定理的证明。

证明 我们可以构建一个 n 维随机向量

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)' \sim N(\mu\mathbf{1}_n, \sigma^2 I_n),$$

其中 $\mathbf{1}_n = (1, 1, \cdots, 1)'$ 一个 $n \times 1$ 向量，且元素均为 1；I_n 是单位矩阵，即其主对角线上的元素为 1，而非主对角线上的元素为 0。于是，$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$ 的密度函数为

$$p(\boldsymbol{x}) = (2\pi)^{-n/2}|\sigma^2 I_n|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \mu\mathbf{1}_n)'(\sigma^2 I_n)^{-1}(\boldsymbol{x} - \mu\mathbf{1}_n)\right\}.$$

接下来，我们构造一个特别的正交矩阵 A，即

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\times1}} & -\frac{1}{\sqrt{2\times1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\times2}} & \frac{1}{\sqrt{3\times2}} & -\frac{2}{\sqrt{3\times2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n\times(n-1)}} & \frac{1}{\sqrt{n\times(n-1)}} & \frac{1}{\sqrt{n\times(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n\times(n-1)}} \end{pmatrix}$$

并令 $\boldsymbol{y} = A\boldsymbol{x}$。构造 A 的目的是为了将 $\bar{x}$ 和 $x_2 - \bar{x}, \cdots, x_n - \bar{x}$ 构成一些新的随机变量，从而可以通过变量变换法得到我们想要求的统计量的分布。这个 A 的构造方法不为一，这里提供了一个相对容易的证明方案。基于矩阵 A，我们发现

1. $\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i = y_1$，即 $\bar{x} = \frac{1}{\sqrt{n}}y_1$；
2. $\sum_{i=1}^n y_i = \boldsymbol{y}'\boldsymbol{y} = (A\boldsymbol{x})'(A\boldsymbol{x}) = \boldsymbol{x}'A'A\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^n x_i$。

于是，

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = y_1^2 + \sum_{i=2}^n y_i = n(\bar{x})^2 + \sum_{i=2}^n y_i$$

则

$$\sum_{i=2}^n y_i = \sum_{i=1}^n x_i - n(\bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s^2.$$

所以，$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 的密度函数为

$$\begin{aligned} p(\boldsymbol{y}) &= (2\pi)^{-n/2}|\sigma^2 I_n|^{-1/2} \exp\left\{-\frac{1}{2}(A^{-1}\boldsymbol{y} - \mu\mathbf{1}_n)'(\sigma^2 I_n)^{-1}(A^{-1}\boldsymbol{y} - \mu\mathbf{1}_n)\right\} \cdot ||A|| \\ &= (2\pi)^{-n/2}|\sigma^2 I_n|^{-1/(2\sigma^2)} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - A\mu\mathbf{1}_n)'A(I_n)^{-1}A'(\boldsymbol{y} - A\mu\mathbf{1}_n)\right\} \\ &= (2\pi)^{-n/2}|\sigma^2 I_n|^{-1/(2\sigma^2)} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - A\mu\mathbf{1}_n)'(AA')^{-1}(\boldsymbol{y} - A\mu\mathbf{1}_n)\right\} \end{aligned}$$

这表明了 $\boldsymbol{y}$ 仍服从 n 维正态分布，其均值向量为 $A\mu\mathbf{1}_n$，而方差-协方差矩阵为 $\sigma^2 AA'$。这里我们具体地讨论一下均值向量

$$A \cdot \mu\mathbf{1}_n = \mu A\mathbf{1}_n = \begin{pmatrix} \sqrt{n}\mu \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

而方差-协方差矩阵为

$$\sigma^2 AA'A' = \sigma^2 I_n = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

这表明了 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 中每个随机变量均是独立的，且 $y_1 \sim N(\sqrt{n}\mu, \sigma^2)$，而 $y_i^2 \sim N(0, \sigma^2), i = 2, 3, \cdots, n$。所以，样本均值

$$\bar{x} = \frac{1}{\sqrt{n}} y_1 \sim N(\mu, \sigma^2),$$

而样本方差

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=2}^n y_i^2}{\sigma^2} \sim \chi^2(n-1),$$

同时，$\bar{x}$ 仅由 y_1 构造而得，s^2 由 $y_2, \cdots, y_n$ 构造而得，于是，$\bar{x}$ 和 s^2 是独立的。

注 表16.1中列出了在正态总体下常见统计量的抽样分布及其分位数的记号。

表 16.1: 几种抽样分布及其分位数

分布	α 分位数
$N(0,1)$	z_α
$\chi^2(n)$	$\chi^2_\alpha(n)$
$F(m,n)$	$F_\alpha(m,n)$
$t(n)$	$t_\alpha(n)$

第 17 讲 充分统计量

问题 17.1 我们知道，统计量是样本的函数。我们为什么要构造统计量呢?

笔记

17.1 充分统计量的定义

我们考虑以下这个简单的例子。

例题 17.1 30 秒口算是小学数学课课前一项测试。在表中是四位同学的 20 题的结果。

题号	学生 1	学生 2	学生 3	学生 4
1	$\surd$	$\times$	$\surd$	$\times$
2	$\times$	$\times$	$\surd$	$\times$
3	$\surd$	$\surd$	$\surd$	$\surd$
4	$\surd$	$\surd$	$\surd$	$\surd$
5	$\surd$	$\surd$	$\times$	$\surd$
6	$\surd$	$\surd$	$\surd$	$\surd$
7	$\surd$	$\surd$	$\surd$	$\surd$
8	$\times$	$\surd$	$\surd$	$\surd$
9	$\surd$	$\surd$	$\times$	$\surd$
10	$\surd$	$\surd$	$\surd$	$\surd$
11	$\surd$	$\surd$	$\surd$	$\surd$
12	$\surd$	$\surd$	$\surd$	$\times$
13	$\surd$	$\times$	$\surd$	$\surd$
14	$\times$	$\surd$	$\surd$	$\surd$
15	$\surd$	$\surd$	$\times$	$\surd$
16	$\surd$	$\surd$	$\surd$	$\surd$
17	$\times$	$\surd$	$\surd$	$\times$
18	$\surd$	$\surd$	$\surd$	$\surd$
19	$\surd$	$\surd$	$\surd$	$\surd$
20	$\surd$	$\times$	$\times$	$\surd$
总分	16	16	16	16

从数理统计中我们重新来考虑这个例子。我们记 $x_{i,j}$ 为第 j 个学生在第 i 道题目的结果，若正确，$x_{i,j}$ 取值为 1；否则为 0，$i=1,2,\cdots,20, j=1,2,3,4$。以学生 1 为例，可观测到一组样本 $\{1,0,1,1,1,1,1,0,1,1,1,1,1,0,1,1,0,1,1,1\}$。以学生 2 为例，可以观测到另一组样本 $\{0,0,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,0\}$。这里的 1 或者 0 都是 $x_{i,j}$ 具体的取值。也就是

说，对于学生 j，可以得到一组样本 $\{x_{1,j}, x_{2,j}, \cdots, x_{20,j}\}$，而总分

$$\sum_{i=1}^{20} x_{i,j}$$

是样本的统计量。显然，在这组样本中，我们可以知道统计量的信息（根据样本我们能够计算统计量），但是根据统计量，我们无法得知样本的所有信息（根据总分，我们无法反推出学生 j 在每道题是否回答正确的结果）。因此，在统计量对样本加工或者简化的过程中，信息被丢失了。

问题 17.2 在构建统计量时，应该有要求。怎样的要求是合理的？

笔记

在以参数模型为代表的概率模型中，我们知道，要计算一个事件发生的概率，是依赖于概率模型中的参数的。例如：对于一个正态分布随机变量 $X \sim N(\mu, 1)$，我们要激素其大于 3 的概率。如果 $\mu = 0$，即 $X \sim N(0,1)$，那么 $\{X > 3\}$ 的概率接近于零；而如果 $\mu = 3$，即 $X \sim N(3,1)$，那么 $\{X > 3\}$ 的概率等于 0.5。

由此，这些未知参数包含的就是“有用信息”。如果我们希望统计量不损失信息，本质上就要求了统计量囊括了未知参数的一切信息。

例题 17.2 对于每一个学生，其以一定概率能够回答正确口算题，假定其总体分布为 $b(1, p)$，并所观测到的数据可以看作独立同分布的样本。这里假定了所有学生的正确概率都是 p，认为学生之间是同质的，忽略了学生个体差异。于是，样本的符号也进一步简化，记为

$$x_1, x_2, \cdots, x_n$$

其中，在问题里 n 取 20。这里用 n 表示只是为了更一般的情况。

对于样本来说，其联合概率（质量）函数为

$$\begin{aligned}
P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) &= p_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \\
&= \prod_{i=1}^{n} p_{X_i}(x_i) \\
&= \prod_{i=1}^{n} P(X_i = x_i) \\
&= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.
\end{aligned}$$

这是样本所有的信息，可以看出其依赖于未知参数 p。

接下来，我们来考虑统计量

$$T=\sum_{i=1}^{n} X_i$$

的分布。根据所学习到的知识，T 的分布为 $b(n,p)$，其概率（质量）函数为

$$P(T=t)=p_T(t)=C_n^t p^t(1-p)^{n-t}.$$

然后，在已知统计量的分布条件下，我们来考虑样本的剩余的分布，这里我们用条件概率（质量）函数来表示，即

$$\begin{aligned}
P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t) &= \frac{P((X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T=t)}{P(T=t)} \\
&= \begin{cases} \dfrac{1}{C_n^t}, & \text{如果 } t=\displaystyle\sum_{i=1}^{n} x_i. \\ 0, & \text{如果 } t\neq\displaystyle\sum_{i=1}^{n} x_i. \end{cases}
\end{aligned}$$

我们发现这个条件分布竟然与未知参数 p 无关，这意味着，这个统计量已经囊括了样本中所有有用的信息，也就是数理统计里所定义的“充分的”。

定义 17.1

设 $X\sim F(x;\theta), x_1,\cdots,x_n$ 是来自某个总体的样本，总体分布函数为 $F(x;\theta)$，统计量 $T=T(x_1,x_2,\cdots,x_n)$。如果在给定 T 的取值后，$x_1,x_2,\cdots,x_n$ 的条件分布与 θ 无关，则称 T 为 θ 的充分统计量。 ♣

接下来，我们来看另一个例子。

例题 17.3 设 $x_1,x_2,\cdots,x_n$ 是来自正态分布 $N(\mu,1)$ 的样本，$T=\bar{x}$，则 T 是充分的。

证明 我们可知，$x_1,x_2,\cdots,x_n$ 的联合密度函数为

$$\begin{aligned}
p(x_1,x_2,\cdots,x_n;\mu) &= \prod_{i=1}^{n} p(x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i-\mu)^2\right\}.
\end{aligned}$$

因为统计量

$$T=\bar{x}\sim N(\mu,\frac{1}{n}),$$

所以，T 的密度函数为

$$p_T(t)=\frac{1}{\sqrt{2\pi/n}}\exp\left\{-\frac{1}{2\sigma^2/n}(t-\mu)^2\right\}.$$

根据定义来判断一个统计量是否是充分的，需要考虑一个条件概率（密度）函数，即

$$p(x_1,x_2,\cdots,x_n|t)=\frac{p(x_1,x_2,\cdots,x_n,t)}{p_T(t)}$$

我们注意到，$p(x_1,x_2,\cdots,x_n,t)$ 这是一个退化分布。一旦确定 t 之后，不是所有的样本都是自

由的，x_n 可以改写成 $nt-(x_1+\cdots+x_{n-1})$。于是，

$$
\begin{aligned}
& p(x_1,x_2,\cdots,x_n,t) \\
= & \ p(x_1,x_2,\cdots,x_{n-1},t) \\
= & \ (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n-1}(x_i-\mu)^2+\left(nt-\sum_{i=1}^{n-1}x_i-\mu\right)^2\right)\right\} \\
= & \ (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n-1}x_i^2-2\mu(nt-x_n)+(n-1)\mu^2+\left(nt-\sum_{i=1}^{n-1}x_i\right)^2-2(x_n\mu)+\mu^2\right)\right\} \\
= & \ (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\left(n\mu^2-2n\mu t+\sum_{i=1}^{n}x_i^2\right)\right\}
\end{aligned}
$$

因此，条件概率函数为

$$
\begin{aligned}
p(x_1,x_2,\cdots,x_n|t) &= \frac{p(x_1,x_2,\cdots,x_n,t)}{p(t)} \\
&= \frac{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\left(n\mu^2-2n\mu t+\sum_{i=1}^{n}x_i^2\right)\right\}}{(2\pi/n)^{-1/2}\exp\left\{-\frac{1}{2/n}(t-\mu)^2\right\}} \\
&= \frac{(2\pi)^{-n/2}}{(2\pi/n)^{-1/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}x_i^2+\frac{n}{2}t^2\right\}
\end{aligned}
$$

这与参数 μ 无关。由此得证。

17.2 因子分解定理

定理 17.1 (因子分解定理) 设总体概率函数为 $f(x;\theta)$，$x_1,\cdots,x_n$ 是样本，则 $T=T(x_1,x_2,\cdots,x_n)$ 为充分统计量的充分必要条件是：存在两个函数 $g(t,\theta)$ 和 $h(x_1,x_2,\cdots,x_n)$ 使得对任意的 θ 和任一组观测值 $x_1,x_2,\cdots,x_n$，有

$$f(x_1,x_2,\cdots,x_n;\theta)=g(T(x_1,x_2,\cdots,x_n),\theta)h(x_1,x_2,\cdots,x_n),$$

其中，$g(t,\theta)$ 是通过统计量 T 的取值而依赖于样本的。♡

证明 本定理的一般性结果的证明过程超过了本课程的内容。以下仅考虑离散随机比纳凉的证明。此时概率函数为

$$f(x_1,x_2,\cdots,x_n;)=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n;\theta).$$

首先证明必要性。假定 T 是充分统计量，所以

$$P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t)$$

与 θ 无关。我们将其记为 $h(x_1,x_2,\cdots,x_n)$。令 $A(t)=\{x_1,x_2,\cdots,x_n|T(x_1,x_2,\cdots,x_n)=t\}$。

当样本 $(x_1,x_2,\cdots,x_n)\in A(t)$ 时有

$$\{T=t\}\subset\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\},$$

故

$$\begin{aligned}P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n) &= P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T=t;\theta)\\ &= P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t)P(T=t;\theta)\\ &= h(x_1,x_2,\cdots,x_n)g(t,\theta).\end{aligned}$$

其中 $g(t,\theta)=P(T=t;\theta)$，而 $h(x_1,x_2,\cdots,x_n)=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t)$ 与 θ 无关。因此，必要性得证。

其次证明充分性。由于

$$\begin{aligned}P(T=t;\theta) &= \sum_{(x_1,x_2,\cdots,x_n)\in A(t)} P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n;\theta)\\ &= \sum_{(x_1,x_2,\cdots,x_n)\in A(t)} g(t,\theta)h(x_1,x_2,\cdots,x_n).\end{aligned}$$

对任意 $(x_1,x_2,\cdots,x_n)$ 和 t，满足 $(x_1,x_2,\cdots,x_n)\in A(t)$ 有

$$\begin{aligned}P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t) &= \frac{P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T=t;\theta)}{P(T=t;\theta)}\\ &= \frac{P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n;\theta)}{P(T=t;\theta)}\\ &= \frac{g(t,\theta)h(x_1,x_2,\cdots,x_n)}{g(t,\theta)\sum_{(x_1,x_2,\cdots,x_n)\in A(t)} h(x_1,x_2,\cdots,x_n)}\\ &= \frac{h(x_1,x_2,\cdots,x_n)}{\sum_{(x_1,x_2,\cdots,x_n)\in A(t)} h(x_1,x_2,\cdots,x_n)}.\end{aligned}$$

该分布与 θ 无关，这证明了充分性。

注 这里 T 可以是一维的，也可以是多维的。

例题 17.4 设 $x_1,x_2,\cdots,x_n$ 是来自于 $b(1,\theta)$ 的样本。于是，$x_1,x_2,\cdots,x_n$ 的概率函数为

$$P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

令 $t=\sum_{i=1}^n x_i$，$g(t,\theta)=(1-\theta)^n\left(\theta/(1-\theta)\right)^t$ 且 $h(x_1,x_2,\cdots,x_n)=1$。根据因子分解定理，t 是 θ 的充分统计量。

例题 17.5 设 $x_1,x_2,\cdots,x_n$ 是来自于 $N(\mu,\sigma^2)$ 的样本。

1. 若 $\sigma^2=\sigma_0^2$，其中 σ_0^2 已知。$x_1,x_2,\cdots,x_n$ 的概率函数为

$$\begin{aligned}p(x_1,x_2,\cdots,x_n) &= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\mu)^2\right\}\\ &= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n x_i^2-2n\bar{x}\mu+n\mu^2\right)\right\}\\ &= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n x_i^2+\frac{n}{\sigma_0^2}\bar{x}\mu-\frac{n}{2\sigma_0^2}\mu^2\right\}\end{aligned}$$

令 $t=\bar{x}$, $g(t,\mu)=\exp\{n\bar{x}\mu/\sigma_0^2-n\mu^2/(2\sigma_0^2)\}$, 和 $h(x_1,x_2,\cdots,x_n)=(2\pi\sigma_0^2)^{-n/2}\exp\{-\sum_{i=1}^n x_i^2/(2\sigma_0^2)$
根据因子分解定理，$t=\bar{x}$ 是充分统计量。

2. 若 $\mu=0$。

$$t=\sum_{i=1}^n x_i^2$$

是充分统计量。

笔记

3. 令 $\theta=(\mu,\sigma_0^2)'$。

$$\boldsymbol{t}=(t_1,t_2)'=(\bar{x},\sum_{i=1}^n x_i^2)'$$

是充分统计量。

笔记

定理 17.2

设统计量 T 是充分统计量，而 S 也是一个统计量。如果 T 可以表示为 S 的函数，即 $T=\varphi(S)$，那么 S 也是充分统计量。

例题 17.6 总体分布为指数型分布族，即其概率函数为

$$p(x;\theta)=C(\theta)\exp\{g(\theta)t(x)\}h(x).$$

如果 $x_1,x_2,\cdots,x_n$ 是样本，则 $\sum_{i=1}^n x_i$ 是充分统计量。

注 $\sum_{i=1}^n x_i$ 不仅是充分统计量，还是充分完备统计量。

第 18 讲 频率学派的常见点估计方法

阅读章节

- ❑ Prob & Stat 6.2 6.3

从本讲开始，我们介绍统计学中最关键的两项内容——估计和检验。有以下的四个观点：

- 点估计是灵魂；
- 区间估计是工具；
- 假设检验是核心；
- 统计量是基础。

在参数模型中，假定总体分布 $X \sim p(x;\theta), \theta \in \Theta$。这里 θ 是总体分布中的未知参数，可以是一维的，也可以是多维的。

例题 18.1 考虑以下两个总体分布：

1. 若 $X \sim P(\lambda)$。这里总体分布存在一个泊松分布族，即
$$\mathcal{P} = \left\{ \frac{\lambda^x}{x!} e^{-\lambda} : \lambda > 0 \right\},$$
其中参数是 $\theta = \lambda, \Theta = R^+$。一旦 $\lambda = \lambda_0$ 确定后，总体分布就唯一确定。
2. 若 $X \sim N(\mu, \sigma^2)$。这里总体分布存在一个正态分布族，即
$$\mathcal{P} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} : \mu \in R, \sigma^2 > 0 \right\},$$
其中参数是 $\theta = (\mu, \sigma^2)', \Theta = R \times R^+$。一旦 $(\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ 确定后，总体分布就唯一确定。

实际中，参数模型中的参数都是未知的，而我们需要通过样本 $x_1, x_2, \cdots, x_n$ 来推断 θ（或其函数 $g(\theta)$）。

定义 18.1 (估计)

设 $x_1, x_2, \cdots, x_n$ 是来自总体的一个样本，称用于估计未知参数 θ 的统计量
$$\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$$
为 θ 的估计量，或称为 θ 的点估计，简称估计。

♣

这里我们讲介绍三种频率学派的点估计思想：替换、拟合、似然。

18.1 替换

18.1.1 矩法

替换的本质思想是：利用经验分布函数 $F_n(x)$ 来替换总体分布函数 $F(x)$。这里因为我们假定了参数模型，由参数唯一能够决定分布函数，于是记为 $F_\theta(x)$。比较一下这两个“分布”函数是不同的。

- $F_n(x)$：基于样本可计算而得，是完全已知的；
- $F_\theta(x)$：一定存在未知的信息，完全未知或部分未知；

替换原理，是有 K · 皮尔逊教授于 1900 年提出的，称为**矩法**。简单可以概括为

1. 用样本矩替换总体矩（既可以是原点矩，也可以是中心矩）；
2. 用样本矩的函数去替换相应的总体矩的函数；

这个替换原理也可以应用于分布未知的场合，对参数作出估计，即

1. 用样本均值 $\bar{x}$ 估计总体均值 $E(X)$；
2. 用样本方差 s^2 估计总体方差 $Var(X)$；
3. 用事件 A 出现的频率估计事件 A 发生的概率；
4. 用样本的 p 分位数估计总体的 p 分位数。

18.1.2 矩估计

基于替换原理，以下我们给出得到矩估计的具体步骤：

1. 设总体具有已知的概率函数 $p(x;\theta_1,\theta_2,\cdots,\theta_k),(\theta_1,\cdots,\theta_k)\in\Theta$ 是未知参数（向量）；
2. 得到样本 $x_1,x_2,\cdots,x_n$；
3. 假设总体分布的 k 阶原点矩 μ_k 存在（对所有的 $0<j\le k$，各低阶矩 μ_j 均存在）；
4. 若假设 $\theta_1,\theta_2,\cdots,\theta_k$ 能表示成总体矩 $\mu_1,\mu_2,\cdots,\mu_k$ 的函数，即
$$\theta_j=\theta_j(\mu_1,\mu_2,\cdots,\mu_k)$$
则可给出诸 θ_j 的矩估计
$$\hat{\theta}_j=\theta_j(a_1,a_2,\cdots,a_k),j=1,2,\cdots,k$$
其中，$a_j=\frac{1}{n}\sum_{i=1}^n x_i^j$，是 j 阶样本原点矩，$j=1,2,\cdots,k$.
5. 另外，我们要估计参数的函数 $\eta=\eta(\theta_1,\theta_2,\cdots,\theta_k)$，则 η 的矩估计为
$$\hat{\eta}=\eta(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k).$$

例题 18.2 设总体为指数分布，即

$$p(x;\lambda)=\lambda e^{-\lambda},x\ge 0.$$

$x_1,x_2,\cdots,x_n$ 是样本。

1. 首先，注意到 $E(X)=1/\lambda$，即 $\lambda=1/E(X)$。故 λ 的矩估计为
$$\hat{\lambda}=\frac{1}{\bar{x}}.$$

2. 其次，由于 $Var(X) = 1/\lambda^2$，即 $\lambda = 1/\sqrt{Var(X)}$。故 λ 的矩估计为

$$\hat{\lambda} = \frac{1}{s}.$$

注 这表明了矩估计不唯一，但通常应该尽量采用低阶矩给出未知参数的估计。

例题 18.3 设一个实验有三种可能结果，其发生的概率分别为

$$p_1 = \theta^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = (1-\theta)^2$$

现在估了 n 次试验，观测到三种结果发生的次数分别为 n_1, n_2, n_3。如何估计 θ?

解 设该实验的三种结果分别是 a_1, a_2, a_3。于是，分布列如表18.1。为估计 θ，我们可以构建三个矩估计，即

表 18.1: 三种结果的实验

结果	a_1	a_2	a_3
概率	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$
频率	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\frac{n_3}{n}$

1. 可以用 n_1/n 来估计 θ^2，则

$$\hat{\theta}_1 = \sqrt{\frac{n_1}{n}}.$$

2. 可以用 n_3/n 来估计 $(1-\theta)^2$，则

$$\hat{\theta}_2 = 1 - \sqrt{\frac{n_3}{n}}.$$

3. 可以用 $n_1/n + n_2/(2n)$ 来估计 θ，则

$$\hat{\theta}_3 = \frac{2n_1 + n_2}{2n}.$$

18.2 拟合

最小二乘估计是利用拟合思想来构造的估计方法。在后续的课程中，我们将详细地介绍最小二乘估计。这里我们介绍一个简单的例子。

例题 18.4 总体分布 $X \sim N(\mu, 1)$，而 $x_1, x_2, \cdots, x_n$ 是样本。我们想要估计 μ。因为 μ 是总体分布的期望，每一个样本 $x_1, x_2, \cdots, x_n$ 都在 μ 附近波动。我们想要找到一个实数 c 与这些样本最**接近**。于是，我们需要定义一个损失函数

$$l(c) = \sum_{i=1}^{n} g(|x_i - c|) = \sum_{i=1}^{n} |x_i - c|^k = \begin{cases} \sum_{i=1}^{n} |x_i - c|^2, \text{如果 } k = 2; \\ \sum_{i=1}^{n} |x_i - c|, \text{如果 } k = 1. \end{cases}$$

拟合思想促使我们所得到的估计为

$$\hat{\mu} = \arg\min_{c} l(c).$$

当 $k=2$ 时，则损失函数定义为 $l(c)=\sum_{i=1}^{n}(x_i-c)^2$。我们需要求 $l(c)$ 的最小值点，则对 $l(c)$ 求导，即

$$\frac{\partial l(c)}{\partial c}=-2\sum_{i=1}^{n}(x_i-c).$$

令 $\frac{\partial l(c)}{\partial c}=0$。所以，

$$c=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$$

因此，我们通常称 $\bar{x}$ 为 μ 的最小二乘估计。

问题 18.1 当 $k=1$ 时，则损失函数定义为 $l(c)=\sum_{i=1}^{n}x_i-c$。通过最小化 $l(c)$ 可以得到最小一乘估计。这个最小一乘估计是什么？

18.3 似然

问题 18.2 考虑有两个盒子各有 100 个球，记为 A 盒和 B 盒。A 盒中有 99 个白球，1 个黑球；B 盒中有 1 个白球，99 个黑球。我们从某个盒子中抽出了一个球，发现这个球是白球。请问：我们是从哪个盒子中抽出球的？

注 似然思想的本质是“以成败论英雄”。

首先介绍什么是似然函数。

定义 18.2 (似然函数)

设总体分布的概率分布列或概率密度函数为 $p(x;\theta)$，其中 θ 是未知参数。而 $x_1,x_2,\cdots,x_n$ 是样本。$x_1,x_2,\cdots,x_n$ 的联合分布列或联合密度函数为

$$p(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta).$$

对于未知参数 θ 而言，称 $p(x_1,x_2,\cdots,x_n;\theta)$ 为似然函数，记为 $L(\theta)$. ♣

其次，介绍什么是最大似然估计。

定义 18.3 (最大似然估计)

设 θ 是待估参数，$L(\theta)$ 是似然函数。如果统计量 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 满足

$$L(\hat{\theta})=\sup_{\theta\in\overline{\Theta}}L(\theta).$$

那么称 $\hat{\theta}$ 是 θ 的最大似然估计，记为 MLE(maximum likelihood estimate)。 ♣

注 这里 $\overline{\Theta}$ 包括 θ 的定义域 Θ 及其边界。

最后，最大似然估计的一般步骤：

1. 求对数似然函数，即 $l(\theta)=\ln L(\theta)$.

2. 求对数似然函数的驻点 $\hat{\theta}$，即 $\hat{\theta}$ 满足

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

3. 验证 $\hat{\theta}$ 是 $L(\theta)$ 或 $l(\theta)$ 的最大值点。

例题 18.5 设总体分布 $X \sim b(1, \theta)$，而 $x_1, x_2, \cdots, x_n$ 是样本。于是，$x_1, x_2, \cdots, x_n$ 的联合分布列为

$$p(x_1, x_2, \cdots, x_n) = (\theta)^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

这也是似然函数，记为 $L(\theta)$。然后，对数似然函数为

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^n x_i \ln(\theta) + (n - \sum_{i=1}^n x_i)\ln(1-\theta).$$

对 $l(\theta)$ 关于 θ 求导，即

$$0 = \frac{\partial l(\theta)}{\partial \theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}.$$

由此可得

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}.$$

我们还可以验证

$$\left.\frac{\partial^2 l(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -\frac{\sum_{i=1}^n x_i}{\theta^2} - \left.\frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2}\right|_{\theta=\hat{\theta}} < 0$$

所以，$\hat{\theta}$ 是 $l(\theta)$ 最大值点。因此，θ 的最大似然估计为 $\bar{x}$.

例题 18.6 设该实验的三种结果分别是 a_1, a_2, a_3。于是，分布列如表18.2。这里我们介绍 θ 的最大似然估计。

表 18.2: 三种结果的实验

结果	a_1	a_2	a_3
概率	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$
频率	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\frac{n_3}{n}$

首先，似然函数为

$$L(\theta) = \frac{n!}{n_1!n_2!n_3!}(\theta^2)^{n_1}(2\theta(1-\theta))^{n_2}((1-\theta)^2)^{n_3} = \frac{n!}{n_1!n_2!n_3!}2^{n_2}\theta^{2n_1+n_2}(1-\theta)^{n_2+2n_3}.$$

而对数似然函数为

$$l(\theta) = \ln L(\theta) = \ln\left(\frac{n!}{n_1!n_2!n_3!}2^{n_2}\right) + (2n_1+n_2)\ln(\theta) + (n_2+2n_3)\ln(1-\theta).$$

然后对 $l(\theta)$ 关于 θ 求导，即

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{2n_1+n_2}{\theta} - \frac{n_2+2n_3}{1-\theta}.$$

可以解得

$$\hat{\theta} = \frac{2n_1+n_2}{2n_1+n_2+n_2+2n_3} = \frac{2n_1+n_2}{2n}.$$

注 “求导”是求最大似然估计最常用的方法，但并不是在所有场合求导都有效。

例题 18.7 设总体分布 $U(0,\theta)$，而 $x_1,x_2,\cdots,x_n$ 是样本。θ 的似然函数为

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}\frac{1}{\theta}I(0<x_i<\theta) \\ &= \frac{1}{\theta^n}I(0<x_1,x_2,\cdots,x_n<\theta) \\ &= \frac{1}{\theta^n}I(0<x_{(1)}\leq x_{(n)}<\theta) \end{aligned}$$

要使得 $L(\theta)$ 达到最大。对于示性函数取值应为 1，而 $1/(\theta^n)$ 是 θ 的减函数。所以，θ 的取值应尽可能小，而 θ 要大于 $x_{(n)}$，所以 $\Theta=(x_{(n)},\infty)$，而 $\overline{\Theta}=[x_{(n)},\infty)$。因此，$\theta$ 的最大似然估计为

$$\hat{\theta}=x_{(n)}.$$

例题 18.8 设总体分布 $X\sim N(\mu,\sigma^2)$，$\theta=(\mu,\sigma^2)$ 是二维参数。而 $x_1,x_2,\cdots,x_n$ 是样本。于是，似然函数为

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}p(x_i)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

而其对数似然函数为

$$l(\mu,\sigma^2)=\ln L(\mu,\sigma^2)=-(n/2)\ln(2\pi)-(n/2)\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

对 $l(\mu,\sigma^2)$ 分别关于 μ 和 σ^2 求导，即

$$\begin{aligned} \frac{\partial l(\mu,\sigma^2)}{\partial\mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \\ \frac{\partial l(\mu,\sigma^2)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0 \end{aligned}$$

由此解得

$$\begin{cases} \hat{\mu}=\bar{x}, \\ \hat{\sigma}^2=s_n^2. \end{cases}$$

问题 18.3 如何求 σ 的最大似然估计?

性质 最大似然估计的不变性 如果 $\hat{\theta}$ 是 θ 的最大似然估计，则对任一函数 $g(\theta)$。$g(\hat{\theta})$ 是其最大似然估计。

18.3.1 EM 算法

最大似然估计是一种非常有效的参数估计方法，但是当分布里有多余参数或数据为截尾或缺失时，求最大似然估计就变得困难。EM 算法由 Dempster 等人于 1977 年提出的，最初可以解决缺失数据场景下的最大似然估计求解问题。EM 算法的核心想法是分两步骤：E 步求期望；M 步求最值。这里我们仅用一个例子来演示一下如何利用 EM 算法来求最大似然估计。

例题 18.9 设一次试验可能有四个结果，其发生的概率分别为 $\frac{1}{2}-\frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1+\theta}{4}, \frac{\theta}{4}$，其中 $\theta \in (0,1)$。现进行了 197 次试验，四种结果发生的次数分别为 75，18，70，34。求 θ 的最大似然估计。

解 设 y_1, y_2, y_3, y_4 表示四种结果发生的次数。这个总体分布是一个多项分布，故其似然函数为

$$\begin{aligned} L(y_1,y_2,y_3,y_4;\theta) &= \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}-\frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1+\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2-\theta)^{y_1}(1-\theta)^{y_2}(1+\theta)^{y_3}\theta^{y_4}. \end{aligned}$$

其对数似然函数为

$$l(\theta) \propto y_1\ln(2-\theta) + y_2\ln(1-\theta) + y_3\ln(1+\theta) + y_4\ln\theta.$$

我们可以对 $l(\theta)$ 关于 θ 求导，即

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{y_1}{2-\theta} - \frac{y_2}{1-\theta} + \frac{y_3}{1+\theta} + \frac{y_4}{\theta}.$$

方程 $\frac{\partial l(\theta)}{\partial \theta}=0$ 的求解是相对复杂的。于是，这里介绍 EM 算法的一般形式。我们引入两个潜变量 z_1, z_2。假设第一种结果可以分为两部分，其发生的概率分别为 $\frac{1-\theta}{4}$ 和 $\frac{1}{4}$，令 z_1 和 y_1-z_1 分别表示落入这两部分的次数。再假设第三种结果分为两部分，其发生的概率分别为 $\frac{\theta}{4}$ 和 $\frac{1}{4}$，令 z_2 和 y_3-z_2 分别表示落入这两部分的次数。这里 z_1, z_2 是我们虚构的，它们具体的数值是不可观测的。

数据 $(y_1,y_2,y_3,y_4,z_1,z_2)$，称为完全数据，而数据 (y_1,y_2,y_3,y_4)，称为不完全数据。考虑我们可以观测到完全数据，其似然函数为

$$\begin{aligned} & L(\theta; y_1,y_2,y_3,y_4,z_1,z_2) \\ = & \frac{n!}{z_1!(y_1-z_1)!y_2!z_2!(y_3-z_2)!y_4!}\left(\frac{1}{4}\right)^{y_1-z_1}\left(\frac{1-\theta}{4}\right)^{z_1+y_2}\left(\frac{1}{4}\right)^{y_3-z_2}\left(\frac{\theta}{4}\right)^{z_2+y_4} \\ \propto & (1-\theta)^{z_1+y_2}(\theta)^{z_2+y_4}. \end{aligned}$$

其对数似然函数为

$$l(\theta; y_1,y_2,y_3,y_4,z_1,z_2) = (z_1+y_2)\ln(1-\theta) + (z_2+y_4)\ln\theta.$$

基于此，可以解得其最大似然估计为

$$\hat{\theta} = \frac{z_2+y_4}{z_1+y_2+z_2+y_4}.$$

遗憾的是，我们并没有观测到 z_1, z_2。但是我们知道，当 y_1,y_2,y_3,y_4 及 θ 已知时，

$$z_1 \sim b\left(y_1, \frac{1-\theta}{2-\theta}\right), \quad z_2 \sim b\left(y_3, \frac{\theta}{1+\theta}\right).$$

EM 算法具体如下：

- E 步：在观测数据 y_1,y_2,y_3,y_4 和已知 $\theta=\theta^{(i)}$ 的条件下，我们求基于完全数据的对数似然函数的期望

$$\begin{aligned} Q(\theta|y_1,y_2,y_3,y_4,\theta^{(i)}) &= E_{z_1,z_2} l(\theta; y_1,y_2,y_3,y_4,z_1,z_2) \\ &= (E_{\theta^{(i)}}(z_1|y_1)+y_2)\ln(1-\theta) + (E_{\theta^{(i)}}(z_2|y_3)+y_4)\ln(\theta) \\ &= \left(\frac{1-\theta^{(i)}}{2-\theta^{(i)}}y_1+y_2\right)\ln(1-\theta) + \left(\frac{\theta^{(i)}}{1+\theta^{(i)}}y_3+y_4\right)\ln\theta. \end{aligned}$$

- M 步：求 $Q(\theta|y_1,y_2,y_3,y_4,\theta^{(i)})$ 关于 θ 的最大值 $\theta^{(i+1)}$，即
$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|y_1,y_2,y_3,y_4,\theta^{(i)})$$
也就是说，对 $Q(\theta|y_1,y_2,y_3,y_4,\theta^{(i)})$ 关于 θ 求导，即
$$\frac{\frac{\theta^{(i)}}{1+\theta^{(i)}}y_3 + y_4}{\theta^{(i+1)}} - \frac{\frac{1-\theta^{(i)}}{2-\theta^{(i)}}y_1 + y_2}{1-\theta^{(i+1)}} = 0$$
由此可得
$$\theta^{(i+1)} = \frac{\frac{\theta^{(i)}}{1+\theta^{(i)}}y_3 + y_4}{\frac{\theta^{(i)}}{1+\theta^{(i)}}y_3 + y_4 + \frac{1-\theta^{(i)}}{2-\theta^{(i)}}y_1 + y_2}.$$

问题 18.4 EM 算法是一个迭代算法。我们仍有以下三个问题：
- EM 算法在什么条件下是收敛?
- EM 算法如何才能停止?
- 如何选择 EM 算法的初始值?

第 19 讲 频率学派中点估计的常见评价方法

阅读章节

- Prob & Stat 6.1 6.2 6.4

问题 19.1 第 18 讲我们介绍了不同的点估计方法。一个很自然的问题是怎样的估计是一个**好**的点估计?

我们一般认为一个好的点估计,应该有一些合理性的要求。那么什么是合理性的要求(性质)?

笔记

19.1 无偏性

注 无偏性是最常见的性质,它是重要的,但不是必要的。

定义 19.1 (无偏性)

设 θ 是我们待估计的参数,而 $\hat{\theta}$ 是 θ 的一个点估计。如果 $\hat{\theta}$ 满足

$$E_\theta(\hat{\theta}) = \theta, \theta \in \Theta.$$

则称 $\hat{\theta}$ 是 θ 的无偏估计 (unbiased estimate, U.E.)。

注 与无偏估计相对,我们统一地称不具有无偏估计的估计是有偏估计。

例题 19.1 若总体分布为一个未知分布,其分布函数记为 $F(x)$。设其期望为 μ,即 $E(X) = \mu$,方差为 σ^2,即 $Var(X) = \sigma^2$。现有样本 $x_1, x_2, \cdots, x_n$。

1. 通常,样本均值 $\bar{x}$ 是总体均值 μ 的一个估计。我们可以计算 $\bar{x}$ 的期望,即

$$\begin{aligned} E(\bar{x}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\ &= \mu. \end{aligned}$$

于是,$\bar{x}$ 是 μ 的无偏估计。

2. 通常，样本方差 $s^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ 来估计总体方差。我们可以计算 s_n^2 的期望，即
$$E(s_n^2)=\frac{n-1}{n}\sigma^2\neq\sigma^2.$$
于是，s_n^2 不是 σ^2 的无偏估计。$E\left(S_n^2\right)=\frac{n-1}{n}\sigma^2$　S_n^2 不是 σ^2 的 U.E.
3. 易于证明，$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 是 σ^2 的无偏估计。

笔记

注 回看样本方差 s_n^2，这个估计量的期望为
$$\frac{n-1}{n}\sigma^2.$$
当 n 越大，该期望越接近 σ^2。当样本量充分大时，s_n^2 可以近似地看作 σ^2。我们称 s_n^2 是 σ^2 的**渐近无偏**估计。

问题 19.2 总体标准差 σ 的无偏估计会是怎样的？样本标准差 s 是 σ 的无偏估计吗？

例题 19.2 考虑总体分布为 $N(\mu,\sigma^2)$，μ 和 σ^2 均是待估参数。现有样本 $x_1,x_2,\cdots,x_n$。已知样本方差 s^2 为
$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$
而样本标准差 $s=\sqrt{s^2}$。为计算 s 的期望，我们先来回顾一下 s^2 的分布。我们知道在正态分布假定下，
$$\frac{(n-1)s^2}{\sigma^2}\sim\chi^{n-1}=Ga((n-1)/2,1/2)$$
因为 $s=(s^2)^{1/2}$，我们来看待一般的伽马分布 $Y\sim Ga(\alpha,\gamma)$ 的 k 阶矩的期望。
$$\begin{aligned}E(Y^k)&=\int_0^\infty y^k\frac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp\{-\lambda y\}\mathrm{d}y\\&=\int_0^\infty\frac{\lambda^\alpha}{\Gamma(\alpha)}y^{k+\alpha-1}\exp\{-\lambda y\}\mathrm{d}y\\&=\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}}\int_0^\infty\frac{\lambda^{\alpha+k}}{\Gamma(\alpha+k)}y^{k+\alpha-1}\exp\{-\lambda y\}\mathrm{d}y\\&=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\cdot\lambda^{-k}.\end{aligned}$$
于是，
$$E(s)=E\left((s^2)^{1/2}\right)=\frac{\Gamma\left(\frac{n-1}{2}+\frac{1}{2}\right)}{\Gamma(\frac{n-1}{2})}\cdot\left(\frac{1}{2}\right)^{-1/2}$$
即
$$\frac{\sqrt{n-1}}{\sigma}E(s)=\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma(\frac{n-1}{2})}\cdot\left(\frac{1}{2}\right)^{-1/2}$$

因此，

$$E(s)=\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sigma=\frac{\sigma}{c_n}.$$

注

1. s 是 σ 的有偏估计；
2. s 是 σ 的渐近无偏估计；
3. $c_n s$ 是 σ 的无偏估计，但一般不修偏。

值得注意的是，不是所有的参数都有无偏估计。以下介绍一个例子，供同学们课后阅读。

例题 19.3 设总体为二点分布 $b(1,p),0<p<1$。$x_1,x_2,\cdots,x_n$ 是样本，令参数为 $\theta=1/p$。以下说明 θ 是不可估的。

首先，$T=x_1+x_2+\cdots+x_n$ 是充分统计量，则 $T\sim b(n,p)$。若有一个 $\hat{\theta}=\hat{\theta}(t)$ 是 θ 的无偏估计，则有

$$E(\hat{\theta})=\sum_{i=1}^{n}\binom{n}{i}\hat{\theta}(i)p^i(1-p)^{n-i}=\frac{1}{p}$$

也就是说，

$$\sum_{i=1}^{n}\binom{n}{i}\hat{\theta}(i)p^{i+1}(1-p)^{n-i}-1=0,0<p<1.$$

这是 p 的 $n+1$ 次方程，最多有 $n+1$ 个实根，要使它对 $(0,1)$ 中所有的 p 都成立是不可能的，故参数 $\theta=1/p$ 是不可估的。

其次，若有某个 $h(x_1,x_2,\cdots,x_n)$ 是 θ 的无偏估计，则令 $\tilde{\theta}=E(h(x_1,x_2,\cdots,x_n)|T)$。由重期望公式可知，

$$E(\tilde{\theta})=E(E(h(x_1,x_2,\cdots,x_n)|T))=E(h(x_1,x_2,\cdots,x_n)=\theta.$$

这说明 $\tilde{\theta}(T)$ 是 θ 的无偏估计。因此这是不可能的。

19.2 有效性

问题 19.3 对于同一个参数 θ，无偏估计仍有很多，如何在无偏估计中进行选择？

定义 19.2

设 $\hat{\theta}_1,\hat{\theta}_2$ 是 θ 的两个无偏估计。如果对任意的 $\theta\in\Theta$ 有

$$Var(\hat{\theta}_1)\le Var(\hat{\theta}_2),$$

且至少有一个 $\theta\in\Theta$ 使得上述不等号严格成立，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。

注 方差越小的无偏估计越有效。

例题 19.4 设 $x_1,x_2,\cdots,x_n$ 是来自均匀总体 $U(0,\theta)$ 的样本，现有两个无偏估计。

1. 估计为 $\hat{\theta}_1 = \frac{n+1}{n}x_{(n)}$。因为 $x_{(n)}/\theta \sim Be(n,1)$，所以 $E(x_{(n)}) = \frac{n}{n+1}\theta$，而
$$Var(x_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2.$$
于是，$\hat{\theta}_1$ 是 θ 的无偏估计，其方差为
$$Var(\hat{\theta}_1) = \frac{(n+1)^2}{n^2}\frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{\theta^2}{n(n+2)}.$$
2. 估计为 $\hat{\theta}_2 = 2\bar{x}$。可以计算
$$E(\hat{\theta}_2) = 2E(\bar{x}) = 2\cdot\frac{\theta}{2} = \theta.$$
而
$$Var(\hat{\theta}_2) = 4Var(\bar{x}) = \frac{4}{n}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

根据比较可知，当 $n>1$ 时，$\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。

注 我们发现，$\hat{\theta}_1$ 是 $x_{(n)}$ 的函数，而 $x_{(n)}$ 是 θ 的充分统计量。

19.3 相合性

相合性是参数估计的必要条件。

定义 19.3

设 $\theta \in \Theta$ 为未知参数，$\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$ 是 θ 的一个估计，n 为样本容量。若对任何一个 $\varepsilon > 0$，有
$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| \geqslant \varepsilon\right) = 0$$
则称 $\hat{\theta}_n$ 为参数 θ 的相合估计。♣

注 $\hat{\theta}_n$ 依概率收敛于 θ。

例题 19.5 设 $x_1, x_2, \cdots$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本序列。由 CLT 可知，

1. $\bar{x}$ 是 μ 的相合估计；
2. s_n^2 是 σ^2 的相合估计；
3. s^2 是 σ^2 的相合估计。

定理 19.1

设 $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$ 是 θ 的一个估计，若
$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta, \quad \lim_{n\to\infty} Var(\hat{\theta}_n) = 0.$$
则 $\hat{\theta}$ 是 θ 的相合估计。♡

证明 我们考虑
$$\left\{|\hat{\theta}_n - \theta| \geq \varepsilon\right\}$$

其中，

$$\hat{\theta}_n - \theta = \left(\hat{\theta}_n - E(\hat{\theta}_n)\right) + \left(E(\hat{\theta}_n) - \theta\right).$$

于是，

$$|\hat{\theta}_n - \theta| \leq \left|\hat{\theta}_n - E(\hat{\theta}_n)\right| + \left|E(\hat{\theta}_n) - \theta\right|.$$

如果 $\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| < \varepsilon/2$ 且 $\left|E(\hat{\theta}_n) - \theta\right| < \varepsilon/2$，那么 $\left|\hat{\theta}_n - \theta\right| < \varepsilon$。考虑其逆否命题，如果 $\left|\hat{\theta}_n - \theta\right| \geq \varepsilon$, 那么

$$\left\{\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| \geq \varepsilon/2\right\} \cup \left\{\left|E(\hat{\theta}_n) - \theta\right| \geq \varepsilon/2\right\}.$$

于是，当 $n \to \infty$ 时，

$$\begin{aligned} P\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) &\leq P\left(\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| \geq \varepsilon/2\right) + P\left(\left|E(\hat{\theta}_n) - \theta\right| \geq \varepsilon/2\right) \\ &\leq \frac{4}{\varepsilon^2} Var(\hat{\theta}_n) + 0 \to 0 \end{aligned}$$

其中，第二个不等式满足因为切比雪夫不等式，即对任意的 $\varepsilon > 0$，有

$$P\left(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq \frac{\varepsilon}{2}\right) \leq \frac{4}{\varepsilon^2} Var(\hat{\theta}_n).$$

因此，$\hat{\theta}_n$ 是 θ 的相合估计。

例题 19.6 设 $x_1, x_2, \cdots, x_n$ 是来自于均匀总体 $U(0, \theta)$ 的样本。接下来我们考虑 θ 的估计 $\hat{\theta}_1 = x_{(n)}$。因为 $x_{(n)} \sim Be(n, 1)$，所以，当 $n \to \infty$ 时，

$$\begin{aligned} E(\hat{\theta}_1) &= \frac{n}{n+1}\theta \to \theta \\ Var(\hat{\theta}_1) &= \frac{n}{(n+1)^2(n+2)}\theta \to 0. \end{aligned}$$

由此可证，$x_{(n)}$ 是 θ 的相合估计。

定理 19.2

若 $\hat{\theta}_{n1}, \hat{\theta}_{n2}, \cdots, \hat{\theta}_{nk}$ 分别是 $\theta_1, \theta_2, \cdots, \theta_k$ 的相合估计, $\eta = g(\theta_1, \theta_2, \cdots, \theta_k)$ 是 $\theta_1, \theta_2, \cdots, \theta_k$ 的连续函数，则 $\hat{\eta}_n = g(\hat{\theta}_{n1}, \hat{\theta}_{n2}, \cdots, \hat{\theta}_{nk})$ 是 η 的相合估计。

注 矩估计和最大似然估计均具有相合性。

19.4 渐近正态性

定义 19.4

若 $\hat{\theta}_n$ 是 θ 的相合估计。若存在趋于 0 的非负常数序列 $\sigma_n(\theta)$，使得

$$\frac{\hat{\theta} - n\theta}{\sigma_n(\theta)}$$

按分布收敛于标准正态分布。称 $\hat{\theta}_n$ 是渐近正态的，又称 $\hat{\theta}_n$ 服从渐近正态分布 $N(\theta, \sigma_n^2(\theta))$，记为

$$\hat{\theta}_n \sim AN(\theta, \sigma_n^2(\theta)).$$

其中称 $\sigma_n^2(\theta)$ 为 $\hat{\theta}_n$ 的渐近方差。

定理 19.3

设总体 X 有概率函数 $p(x;\theta), \theta \in \Theta$，$\Theta$ 为非退化区间，假定

1. 对任意 x，偏导数 $\frac{\partial \ln p}{\partial \theta}, \frac{\partial^2 \ln p}{\partial \theta^2}, \frac{\partial^3 \ln p}{\partial \theta^3}$ 对所有 $\theta \in \Theta$ 都存在；
2. 对任意 $\theta \in \Theta$，有
$$\left|\frac{\partial \ln p}{\partial \theta}\right| < F_1(x), \quad \left|\frac{\partial^2 \ln p}{\partial \theta^2}\right| < F_2(x), \quad \left|\frac{\partial^3 \ln p}{\partial \theta^3}\right| < F_3(x),$$
其中函数 $F_1(x), F_2(x), F_3(x)$ 满足
$$\int_{-\infty}^{\infty} F_1(x)\mathrm{d}x < \infty, \int_{-\infty}^{\infty} F_2(x)\mathrm{d}x < \infty, \sup_{\theta\in\Theta}\int_{-\infty}^{\infty} F_3(x)p(x;\theta)\mathrm{d}x < \infty.$$
3. Fisher 信息量为
$$I(\theta) = \int_{-\infty}^{\infty}\left(\frac{\partial \ln p}{\partial \theta}\right)^2 p(x;\theta)\mathrm{d}x$$
对任意 $\theta \in \Theta$，$0 < I(\theta) < \infty$。

若 $x_1, x_2, \cdots, x_n$ 是来自该总体的样本，则存在未知参数 θ 的最大似然估计 $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$，且 $\hat{\theta}_n$ 具有相合性和渐近正态性，即

$$\hat{\theta}_n \sim AN\left(\theta, \frac{1}{nI(\theta)}\right).$$

例题 19.7 设总体分布为泊松分布 $P(\lambda)$，可以证明 λ 的矩估计和最大似然估计为

$$\hat{\lambda}_n = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

根据 CLT 可知，

$$\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda/n}} \xrightarrow{L} N(0,1).$$

所以，$\hat{\lambda}_n$ 是渐近正态的，即

$$\hat{\lambda}_n \sim AN(\lambda, \lambda/n).$$

总体分布的分布列为

$$p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

其对数为

$$\ln p(x;\lambda) = x\ln\lambda - \ln x! - \lambda.$$

其导数为

$$\frac{\partial \ln p}{\partial \lambda} = \frac{x}{\lambda} - 1$$

所以，Fisher 信息量为

$$\begin{aligned} I(\lambda) &= \int_{-\infty}^{\infty} \left(\frac{\partial \ln p}{\partial \lambda} \right)^2 p(x;\lambda) \mathrm{d}x \\ &= \sum_{x=0}^{\infty} \left(\frac{x}{\lambda} - 1 \right)^2 p(x;\lambda) \\ &= \frac{E(X^2)}{\lambda^2} - \frac{2E(X)}{\lambda} + 1 \\ &= \frac{\lambda + \lambda^2}{\lambda^2} - 1 \\ &= \frac{1}{\lambda}. \end{aligned}$$

所以，$\hat{\theta}_n$ 服从渐近正态分布，即

$$\hat{\theta}_n \sim AN(\lambda, \lambda/n).$$

19.5 均方误差

相合性和渐近正态性都是大样本场合下评价估计好坏的连个指标。在样本量不是很大时，我们需要新的评价指标。

定义 19.5

若 $\hat{\theta}$ 是参数 θ 的一个估计。称

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

为 $\hat{\theta}$ 的均方误差（Mean Squared Error, MSE）。 ♣

注

1. 均方误差是评价点估计的最一般的标准。希望点估计的均方误差越小越好。
2. 值得注意的是，

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= E(\hat{\theta} - E(\hat{\theta}))^2 + 2E(\hat{\theta} - E(\hat{\theta})) \cdot (E(\hat{\theta}) - \theta) + E(E(\hat{\theta}) - \theta)^2 \\ &= Var(\hat{\theta}) + bias^2(\hat{\theta}) \end{aligned}$$

其中，交叉项 $E\left((\hat{\theta} - E(\hat{\theta})) \cdot (E(\hat{\theta}) - \theta) \right) = 0$。上式表明了均方误差由点估计的方差与偏差平方两部分组成的。

问题 19.4 对于任意 $\theta \in \Theta$，我们能否找到一致最小均方误差估计 $\hat{\theta}$，即

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta})?$$

笔记

注 在**无偏估计类**中，存在一致最小均方误差估计，此时一致最小均方误差估计为一致最小方差无偏估计。

19.6 充分性原则

定理 19.4 (Rao-Blackwell 定理)

设总体概率函数是 $p(x;\theta), x_1, x_2, \cdots, x_n$ 是其样本，$T = T(x_1, x_2, \cdots, x_n)$ 是 θ 的充分统计量。则对 θ 的任一无偏估计 $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$，令 $\tilde{\theta} = E(\hat{\theta}|T)$，有 $\tilde{\theta}$ 也是 θ 的无偏估计，且

$$Var(\tilde{\theta}) \leq Var(\hat{\theta}).$$

证明 第一，欲证明 $\tilde{\theta}$ 是 θ 的无偏估计，即

$$E(\tilde{\theta}) = E(E(\hat{\theta}|T)) = E(\hat{\theta}) = \theta,$$

其中，第二个等号成立是由于重期望公式。第二，考察 $\tilde{\theta}$ 的方差，即

$$\begin{aligned} Var(\hat{\theta}) &= E(\hat{\theta} - E\hat{\theta})^2 \\ &= E(\hat{\theta} - \tilde{\theta} + \tilde{\theta} - E\tilde{\theta})^2, E(\hat{\theta}) = E(\tilde{\theta}) \\ &= E(\hat{\theta} - \tilde{\theta})^2 + E(\tilde{\theta} - E(\tilde{\theta}))^2 + 2E((\hat{\theta} - \tilde{\theta}) \cdot (\tilde{\theta} - E(\tilde{\theta}))) \\ &= E(\hat{\theta} - \tilde{\theta})^2 + Var(\tilde{\theta}) \end{aligned}$$

其中交叉项为

$$\begin{aligned} E((\hat{\theta} - \tilde{\theta}) \cdot (\tilde{\theta} - E(\tilde{\theta})) &= E(E((\hat{\theta} - \tilde{\theta}) \cdot (\tilde{\theta} - E(\tilde{\theta}) \mid T)) \\ &= E((\tilde{\theta} - E(\tilde{\theta}))E((\hat{\theta} - \tilde{\theta}) \mid T)) \\ &= 0 \end{aligned}$$

注 Rao-Blackwell 定理表明，对于任何无偏估计，如果其不是充分统计量的函数，那么将其对充分统计量求条件期望可以得到一个新的无偏估计，而且该估计的方差比原来的估计方差要小。

例题 19.8 总体分布为泊松分布 $P(\lambda)$，其中参数 $\lambda > 0$。先有两个样本 x_1, x_2。x_1, x_2 的联合分布列为

$$p(x_1, x_2; \lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!}e^{-\lambda} = \lambda^{(x_1+x_2)}e^{-2\lambda}\frac{1}{x_1!x_2!}.$$

令 $T(x_1, x_2) = x_1 + x_2$, $g(t, \lambda) = \lambda^t e^{-2\lambda}$ 和 $h(x_1, x_2) = \frac{1}{x_1!x_2!}$。根据因子分解定理，有 $T(x_1, x_2) = x_1 + x_2$ 是充分的。而 $T(x_1, x_2) \sim P(2\lambda)$。

因为 $E(x_1) = \lambda$，所以 $\hat{\lambda} = x_1$ 是 λ 的无偏估计。令 $\tilde{\lambda} = E(\hat{\lambda}|T)$。则

$$\begin{aligned}
P(X_1 = x_1 | X_1 + X_2 = n) &= \frac{P(X_1 = x_1, X_1 + X_2 = n)}{P(X_1 + X_2 = n)} \\
&= \frac{\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{n-x_1}}{(n-x_1)!}e^{-\lambda}}{\frac{(2\lambda)^n}{n!}e^{-2\lambda}} \\
&= \frac{\frac{\lambda^n}{x_1!(n-x_1)!}}{\frac{(2\lambda)^n}{n!}} \\
&= \frac{n!}{x_1!(n-x_1)!}(1/2)^{x_1}(1/2)^{n-x_1}
\end{aligned}$$

且

$$E(X_1 | X_1 + X_2 = x_1 + x_2) = \frac{x_1 + x_2}{2}.$$

我们分别可以计算

$$Var(\hat{\lambda}) = \lambda > Var(\tilde(\lambda)) = \frac{\lambda}{2}, \lambda > 0.$$

所以，$\tilde{\lambda}$ 的方差更小。

第 20 讲 区间估计

阅读章节

- Prob & Stat 6.6

20.1 区间估计的概念

设 $x_1, x_2, \cdots, x_n$ 是样本。我们想要找到两个统计量 $\hat{\theta}_L = \hat{\theta}_L(x_1, \cdots, x_n)$ 和 $\hat{\theta}_u = \hat{\theta}_u(x_1, \cdots, x_n)$，$\hat{\theta}_L < \hat{\theta}_u$。于是，所构造的一个区间 $[\hat{\theta}_L, \hat{\theta}_u]$ 为 θ 的一个区间估计。

问题 20.1 一个合适的区间估计应该有什么要求?

笔记

因为样本具有随机性，所以，$\left[\hat{\theta}_L, \hat{\theta}_u\right]$ 是一个随机区间。但待估参数 θ 是一个未知常数。我们通常要求区间 $\left[\hat{\theta}_L, \hat{\theta}_u\right]$ 盖住 θ 的概率

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_u) = P\left(\left\{\hat{\theta}_L \leq \theta\right\} \cap \left\{\hat{\theta}_u \leq \theta\right\}\right)$$

尽可能大。

定义 20.1

设 θ 是总体的一个参数，其参数空间为 Θ，$x_1, x_2, \cdots, x_n$ 是来自该总体的样本，对给定的一个 $\alpha(0 < \alpha < 1)$，假设有两个统计量 $\hat{\theta}_L = \hat{\theta}_L(x_1, x_2, \cdots, x_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(x_1, x_2, \cdots, x_n)$，若对任意的 $\theta \in \Theta$，有

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_u) \geq 1 - \alpha$$

则称随机区间 $\left[\hat{\theta}_L, \hat{\theta}_u\right]$ 为 θ 的置信水平为 $1-\alpha$ 的置信区间，或简称 $\left[\hat{\theta}_L, \hat{\theta}_u\right]$ 是 θ 的 $1-\alpha$ 置信区间，$\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别称为 θ 的（双侧）置信下限和置信上限。♣

注

1. 若对给定的 $\alpha(0 < \alpha < 1)$，对任意的 $\theta \in \Theta$，有
$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_u) = 1 - \alpha$$
则称 $\left[\hat{\theta}_L, \hat{\theta}_u\right]$ 是 θ 的 $1 - \alpha$ 同等置信区间;
2. 若对给定的 $\alpha(0 < \alpha < 1)$，对任意的 $\theta \in \Theta$，有
$$P(\hat{\theta}_L \leq \theta) \geq 1 - \alpha$$

则称 $\hat{\theta}_L$ 为 θ 的置信水平为 $1-\alpha$ 的（单侧）置信下限；

3. 若对给定的 $\alpha(0<\alpha<1)$，对任意的 $\theta \in \Theta$，有
$$P(\hat{\theta}_L \leq \theta) = 1-\alpha$$
则称 $\hat{\theta}_L$ 为 θ 的置信水平为 $1-\alpha$ 的（单侧）同等置信下限；
4. 若对给定的 $\alpha(0<\alpha<1)$，对任意的 $\theta \in \Theta$，有
$$P(\theta \leq \hat{\theta}_U) \geq 1-\alpha$$
则称 $\hat{\theta}_U$ 为 θ 的置信水平为 $1-\alpha$ 的（单侧）置信上限；
5. 若对给定的 $\alpha(0<\alpha<1)$，对任意的 $\theta \in \Theta$，有
$$P(\theta \leq \hat{\theta}_U) = 1-\alpha$$
则称 $\hat{\theta}_U$ 为 θ 的置信水平为 $1-\alpha$ 的（单侧）同等置信上限；

20.2 枢轴量法

问题 20.2 如何求 $1-\alpha$ 置信区间?

定义 20.2

若 $G(x_1, x_2, \cdots, x_n, \theta)$ 是样本和待估参数 θ 的函数，而 G 的分布不依赖于未知参数，则称 G 为枢轴量。 ♣

以下我们介绍一种构造 $1-\alpha$ 置信区间的通法——枢轴量法。

1. 构造一个枢轴量 G；
2. 适当地选择两个常数 c 和 d，使对给定的 $\alpha(0<\alpha<1)$，有
$$P(c \leq G \leq d) = 1-\alpha.$$
如果 G 是一个离散型分布，取大于等于号。
3. 假如能将 $c \leq G \leq d$ 进行不等式变形化为 $\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$，则有
$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha.$$
4. 由此，$\left[\hat{\theta}_L, \hat{\theta}_U\right]$ 是 θ 的 $1-\alpha$ 同等置信区间。

例题 20.1 若总体分布为 $N(\mu, \sigma_0^2)$，其中 μ 是待估参数，而 σ_0^2 是已知的。$x_1, x_2, \cdots, x_n$ 是样本。于是，我们介绍如何利用枢轴量法来构造 μ 的区间估计。因为 μ 的点估计为
$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
我们知道，$\bar{x}$ 的分布是 $N(\mu, \sigma_0^2/n)$。对其进行标准化，
$$G = \frac{\bar{x}-\mu}{\sqrt{\sigma_0^2/n}} \sim N(0,1).$$
于是，G 就是我们所构造的枢轴量。于是，存在两个常数 c_1 和 c_2，满足
$$P(c_1 \leq G \leq c_2) = 1-\alpha.$$

虽然 (c_1,c_2) 的取法有无数种，但是最简单的取法为 $c_1=z_{\alpha/2}$ 和 $c_2=z_{1-\alpha/2}$。所以，

$$z_{\alpha/2}\le\frac{\bar{x}-\mu}{\sqrt{\sigma_0^2/n}}\le z_{1-\alpha/2},$$

可以变形为

$$\bar{x}-\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2}\le\mu\le\bar{x}-\frac{\sigma_0}{\sqrt{n}}u_{\alpha/2}=\bar{x}+\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2}$$

因此，μ 的 $1-\alpha$ 同等置信区间为

$$\left[\bar{x}-\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2},\bar{x}+\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2}\right].$$

注 区间 $\left[\bar{x}-\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2},\bar{x}+\frac{\sigma_0}{\sqrt{n}}u_{1-\alpha/2}\right]$ 是最短的。

例题 20.2 若总体分布为 $N(\mu,\sigma^2)$，其中 μ 是待估参数，而 σ^2 也是未知的。$x_1,x_2,\cdots,x_n$ 是样本。于是，我们介绍如何利用枢轴量法来构造 μ 的区间估计。因为 μ 的点估计为

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

我们知道，$\bar{x}$ 的分布是 $N(\mu,\sigma^2/n)$。对其进行标准化，

$$\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}\sim N(0,1).$$

除了待估计的参数之外，还有一个未知参数 σ^2，称其为冗余参数。对于冗余参数 σ^2，我们用其估计代替参数，即

$$G=\frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}}=\frac{\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)\hat{\sigma}^2}{\sigma^2}/(n-1)}}\sim t(n-1).$$

于是，G 就是我们所构造的枢轴量。于是，存在两个常数 c_1 和 c_2，满足

$$P(c_1\le G\le c_2)=1-\alpha.$$

取 $c_1=t_{\alpha/2}(n-1)$ 和 $c_2=t_{1-\alpha/2}(n-1)$。所以，

$$t_{\alpha/2}(n-1)\le\frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}}\le t_{1-\alpha/2}(n-1),$$

可以变形为

$$\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}t_{1-\alpha/2}(n-1)\le\mu\le\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}t_{1-\alpha/2}(n-1)$$

因此，μ 的 $1-\alpha$ 同等置信区间为

$$\left[\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}t_{1-\alpha/2}(n-1),\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}t_{1-\alpha/2}(n-1)\right].$$

问题 20.3 置信水平为 $1-\alpha$ 指的是什么?

笔记 在《概率论与数理统计教程》第 300 页例 6.6.1 中，我们取 $\alpha=0.1$ 以及 $n=10$ 的情形。假定真实的分布为正态分布 $N(\mu,\sigma^2)$，其中 $\mu=15$，方差 $\sigma^2=4$。我们从这个正态分布里抽一组样本量为 10 的样本，记为 $x_{m,1},x_{m,2},\cdots,x_{m,10}$。基于样本，我们可以计算样本均值 $\bar{x}_m=\frac{1}{n}\sum_{i=1}^{n}x_{m,i}$，样本方差 $s_m^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{m,i}-\bar{x}_m)^2$。同时，根据置信水平为 $1-\alpha=0.9$

和样本量 $n=10$，来确定分位数

$$t_{1-\alpha/2}(n-1)=t_{0.95}(9)=1.8331.$$

根据这组样本，我们可以构造一个置信区间。重复这个行为 100 次，于是可以得到 100 个区间。对这 100 个区间进行统计，可以发现其中 91 个区间是包含真值 $\mu=15$ 的，而这个比例 91% 与我们所构造的置信水平差不多。

这就是置信水平的一种理解。

例题 20.3 若总体分布为 $U(0,\theta)$，其中 θ 是待估参数。现有 $x_1,x_2,\cdots,x_n$ 是样本。我们想要得到 θ 的 $1-\alpha$ 置信区间。

解 我们知道 θ 的点估计为 $x_{(n)}$，而且我们知道

$$\frac{x_{(n)}}{\theta}\sim Be(n,1).$$

于是，$x_{(n)}/\theta$ 就是我们所构造的枢轴量。我们希望找到两个常数 c_1 和 c_2 使得

$$P(c_1\leq\frac{x_{(n)}}{\theta}\leq c_2)=1-\alpha.$$

一旦可以确定 c_1 和 c_2，我们可以得到 θ 的 $1-\alpha$ 置信区间为

$$\left[\frac{x_{(n)}}{c_2},\frac{x_{(n)}}{c_1}\right].$$

显然满足条件的 (c_1,c_2) 个数是无穷的。这里我们考虑最短的置信区间。因为我们知道，对于一个随机变量 $X\sim Be(n,1)$，其密度函数为

$$p(x)=nx^{n-1},0<x<1.$$

于是，

$$P(X\leq c)=\int_0^c nx^{n-1}\mathrm{d}x=c^n.$$

我们令 $c_1^n=\alpha_1$，$c_2^n=1-(\alpha-\alpha_1)$。我们要求 (c_1,c_2) 使得

$$\frac{1}{c_1}-\frac{1}{c_2}$$

达到最小。所以，我们构造了一个函数

$$l(\alpha_1)=\alpha_1^{-1/n}-(1-\alpha+\alpha_1)^{-1/n},0<\alpha_1<\alpha$$

我们可以证明 $l(\alpha_1)$ 是一个单调递减的函数，则 $l(\alpha_1)$ 在 $\alpha_1=\alpha$ 处取到最小值。于是，

$$c_1=\alpha^{1/n},\quad c_2=1.$$

因此，θ 的 $1-\alpha$ 置信区间为

$$\left[x_{(n)},x_{(n)}\cdot\alpha^{-1/n}\right].$$

例题 20.4 若总体分布为 $N(\mu,\sigma^2)$，其中 σ^2 是待估参数，而 μ 也是未知的。$x_1,x_2,\cdots,x_n$ 是样

本。这里我们考虑 σ^2 的 $1-\alpha$ 置信区间。首先考虑 σ^2 的点估计为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

其分布为

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1).$$

很自然地我们想要找到两个常数 c_1 和 c_2 使得

$$P(c_1 \leq \frac{(n-1)s^2}{\sigma^2} \leq c_2) = 1-\alpha$$

当然，我们可以类似于例题 1.3 的做法求解最短的区间，但是我们也可以取

$$c_1 = \chi^2_{\alpha/2}(n-1), \quad c_2 = \chi^2_{1-\alpha/2}(n-1).$$

虽然这个区间不是最短的，但是它是最简单的取法。于是，σ^2 的 $1-\alpha$ 置信区间为

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right].$$

这个区间被称为等尾置信区间。

以上介绍例子大多都是在正态总体假定下的置信区间估计的构造方法。这些方法还可以应用于大样本的情形中。

例题 20.5 若总体分布为 $b(1,\theta)$，其中 θ 是待估参数。$x_1, x_2, \cdots, x_n$ 是样本。当样本量 n 比较大时，我们想要得到 θ 的 $1-\alpha$ 置信区间。

首先考虑 θ 的点估计，即

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

因为样本量 n 比较大，这里我们可以考虑 $\bar{x}$ 的渐近分布，即

$$\frac{\bar{x}-\theta}{\sqrt{\theta(1-\theta)/n}} \sim AN(0,1).$$

因此，$\frac{\bar{x}-\theta}{\sqrt{\theta(1-\theta)/n}}$ 就是我们所构造的（渐近）枢轴量。类似于正态总体下的置信区间，我们取 $\pm z_{1-\alpha/2}$，有

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{x}-\theta}{\sqrt{\theta(1-\theta)/n}} \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

于是，我们需要从上述等式的左边将 θ 反解出来。

1. 第一种想法：通过解一个一元二次不等式，从而得到 θ 的置信区间。具体来说，

$$\frac{|\bar{x}-\theta|}{\sqrt{\theta(1-\theta)/n}} \leq z_{1-\alpha/2}$$

2. 第二种想法：实际上我们注意到分母中 $\theta(1-\theta)/n$ 是 $\bar{x}$ 也可以看作一种冗余参数，很自然的想法是我们用估计来代替。在样本量大的情形下，$\bar{x}$ 可以近似 θ，所以
$$\frac{\theta(1-\theta)}{n} \approx \frac{\bar{x}(1-\bar{x})}{n}.$$
这样可以很容易反解出 θ，从而得到 θ 的 $1-\alpha$ 置信区间
$$\left[\bar{x}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right].$$

例题 20.6 现有两个独立总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$。设 $x_1,x_2,\cdots,x_m$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本，而 $y_1,y_2,\cdots,y_n$ 是来自 $N(\mu_2,\sigma_2^2)$ 的样本。考虑不同 θ 以及它们的 $1-\alpha$ 置信区间。

1. $\theta=\mu_1-\mu_2$。
   (a). 若 σ_1^2 和 σ_2^2 已知。由于 θ 的点估计为 $\hat{\theta}=\bar{x}-\bar{y}$ 其分布为
$$\hat{\theta}=\bar{x}-\bar{y}\sim N\left(\theta,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right).$$
   于是，枢轴量为
$$G=\frac{\hat{\theta}-\theta}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1).$$
   所以，θ 的 $1-\alpha$ 置信区间为
$$\left[(\bar{x}-\bar{y})-\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}z_{1-\alpha},(\bar{x}-\bar{y})+\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}z_{1-\alpha}\right]$$
   (b). 若 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知。令
$$s_w^2=\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}.$$
   则 $(m+n-2)s_w^2/\sigma^2$ 的分布为 $\chi^2(m+n-2)$。于是，我们可以构造枢轴量
$$\frac{\hat{\theta}-\theta}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$
   所以，θ 的 $1-\alpha$ 置信区间为
$$\bar{x}-\bar{y}\pm s_w\sqrt{\frac{1}{m}+\frac{1}{n}}t_{1-\alpha/2}(m+n-2)$$
2. $\theta=\sigma_1^2/\sigma_2^2$。我们考虑 θ 的点估计为 s_1^2/s_2^2，则
$$\frac{s_1^2}{s_2^2}/\theta\sim F(m-1,n-1).$$
   所以，θ 的 $1-\alpha$ 置信区间为
$$\left[\frac{s_1^2}{s_2^2}/F_{1-\alpha/2}(m-1,n-1),\frac{s_1^2}{s_2^2}/F_{\alpha/2}(m-1,n-1)\right].$$

第 21 讲 贝叶斯学派的估计方法

阅读章节

- ❑ Intro to Prob 8.1
- ❑ Prob & Stat 6.5

问题 21.1 某工厂生产了一种产品，想要知道该产品的合格率 θ。为了估计参数 θ，我们从工厂生产的产品中进行了抽样，样本为 $x_1, x_2, \cdots, x_n$。一个很自然的估计为

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

这个估计既是矩估计，又是最大似然估计。

如果这一天抽出的样本均是合格品，那么 θ 的估计为 $\hat{\theta} = 1$，我们能够打保票推断合格率为 100% 吗？尤其抽取的样本容量 n 比较小时，很有可能抽取的样本都是合格品。

仅仅通过数据来进行推断，这个估计可能是不恰当的。

本讲中，我们将介绍贝叶斯学派的估计方法。与频率学派不同，贝叶斯学派是一类新的统计思想，可以概括为以下四句话：

1. 待估参数 θ 是一个随机变量；
2. 在得到样本前，对 θ 有一个认识。这个认识可用一个分布来描述，记为 $\pi(\theta)$，称为先验分布（Prior Distribution）或验前分布。
3. 样本是用来调整对 θ 的认识。调整的方法要用 Bayes 公式。调整后的认识，也用一个分布来刻画，该分布称为后验分布（Posterior distribution），记为 $\pi(\theta \mid x)$。
4. 所有的统计推断是基于后验分布进行。

21.1 点估计

例题 21.1 回顾工厂生产产品的例子。现有样本 $x_1, x_2, \cdots, x_n$。样本分布为 $x_i|\theta \sim b(1,\theta), i = 1,2,\cdots,n$。需要对 θ 有一个认识，即认为 θ 有一个先验分布。那么这个先验分布怎样才是合适的？

- 倘若有 m 天的历史数据 $(n_1, x_1), \cdots, (n_m, x_m)$，可以得到 θ 的一个分布 $\pi(\theta)$。假定 $\pi(\theta) = Be(a,b)$，这里两个超参数 a 和 b 可以通过矩估计来获得。
- 倘若没有历史数据，我们可以对先验分布做一个"同等无知"的假定，即 $\pi(\theta) = Be(1,1) = U(0,1)$。

由此可以得到先验分布 $Be(a,b)$。接下来，如何求出 θ 的后验分布？利用贝叶斯公式来得到后验分布。

设样本分布为 $p(x_1, x_2, \cdots, x_n|\theta)$。于是 $(x_1, x_2, \cdots, x_n, \theta)$ 的联合概率函数为

$$h(x_1, x_2, \cdots, x_n, \theta) = \pi(\theta)p(x_1, x_2, \cdots, x_n|\theta)$$

令

$$m(x_1,x_2,\cdots,x_n)=\int h(x_1,x_2,\cdots,x_n,\theta)\mathrm{d}\theta=\int \pi(\theta)p(x_1,x_2,\cdots,x_n|\theta)\mathrm{d}\theta.$$

于是，

$$\begin{aligned}
p(\theta|x_1,x_2,\cdots,x_n) &= \frac{h(x_1,x_2,\cdots,x_n,\theta)}{m(x_1,x_2,\cdots,x_n)}\\
&= \frac{\pi(\theta)p(x_1,x_2,\cdots,x_n|\theta)}{\int \pi(\theta)p(x_1,x_2,\cdots,x_n|\theta)\mathrm{d}\theta}\\
&= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\mathrm{d}\theta}\\
&= \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a+\sum_{i=1}^n x_i-1}(1-\theta)^{b+n-\sum_{i=1}^n x_i-1}}{\int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a+\sum_{i=1}^n x_i-1}(1-\theta)^{b+n-\sum_{i=1}^n x_i-1}\mathrm{d}\theta}\\
&= \frac{\Gamma(a+b+n)}{\Gamma(a+\sum_{i=1}^n x_i)\Gamma(b+n-\sum_{i=1}^n x_i)}\theta^{a+\sum_{i=1}^n x_i-1}(1-\theta)^{b+n-\sum_{i=1}^n x_i-1}
\end{aligned}$$

因此，$\theta|x_1,x_2,\cdots,x_n\sim Be(a+\sum_{i=1}^n x_i, b+n-\sum_{i=1}^n x_i)$，这就是 θ 的后验分布。基于后验分布，如何得到点估计？

我们取后验分布的均值，即

$$\hat{\theta}_{\text{Bayes}}=\frac{a+\sum_{i=1}^n x_i}{a+b+n}=\frac{a+b}{a+b+n}\cdot\frac{a}{a+b}+\frac{n}{a+b+n}\cdot\frac{\sum_{i=1}^n x_i}{n}.$$

例题 21.2 样本 $x_1,x_2,\cdots,x_n$ 来自于正态分布 $N(\mu,\sigma_0^2)$。μ 的先验分布为 $N(\mu_0,\tau^2)$。求 μ 的贝叶斯估计。

解 μ 的先验分布为

$$\pi(\mu)=\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}.$$

而样本 $(x_1,x_2,\cdots,x_n)$ 的联合密度函数为

$$\begin{aligned}
p(x_1,\cdots,x_n|\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma_0^2}(x_i-\mu)^2\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\mu)^2\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\bar{x}+\bar{x}-\mu)^2\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n\left((x_i-\bar{x})^2+(\bar{x}-\mu)^2\right)\right\}\\
&= (2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\bar{x})^2\right\}\cdot\exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x}-\mu)^2\right\}.
\end{aligned}$$

于是，μ 的后验分布为

$$
\begin{aligned}
\pi(\mu|x_1,x_2,\cdots,x_n) &= \frac{\pi(\mu)p(x_1,x_2,\cdots,x_n|\mu)}{m(x_1,x_2,\cdots,x_n)}\\
&\propto \pi(\mu)p(x_1,x_2,\cdots,x_n|\mu)\\
&\propto \exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\}\cdot\exp\left\{-\frac{n}{2\sigma_0^2}(\mu-\bar{x})^2\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\left(\frac{1}{\tau^2}+\frac{1}{\sigma_0^2/n}\right)\mu^2-2\left(\frac{\mu_0}{\tau^2}+\frac{\bar{x}}{\sigma_0^2/n}\right)\mu\right)\right\}\\
&= \exp\left\{-\frac{1}{2}\left(A\mu^2-2B\mu\right)\right\}.
\end{aligned}
$$

其中，$A=\left(\frac{1}{\tau^2}+\frac{1}{\sigma_0^2/n}\right)$ 和 $B=\left(\frac{\mu_0}{\tau^2}+\frac{\bar{x}}{\sigma_0^2/n}\right)$。这里利用配方公式，我们可知，

$$
A\mu^2-2B\mu=A\left(\mu-\frac{B}{A}\right)^2-\frac{B^2}{A}.
$$

从分布核中可以发现，μ 的后验分布是正态分布 $N\left(\frac{B}{A},A^{-1}\right)$，即

$$
N\left(\frac{\frac{\mu_0}{\tau^2}+\frac{\bar{x}}{\sigma_0^2/n}}{\frac{1}{\tau^2}+\frac{1}{\sigma_0^2}},\frac{1}{\frac{1}{\tau^2}+\frac{1}{\sigma_0^2/n}}\right).
$$

由于 μ 的贝叶斯估计是 μ 后验分布的期望，即

$$
\hat{\mu}_{\text{Bayes}}=\frac{\frac{\mu_0}{\tau^2}+\frac{\bar{x}}{\sigma_0^2/n}}{\frac{1}{\tau^2}+\frac{1}{\sigma_0^2/n}}=\frac{1/\tau^2}{1/\tau^2+1/(\sigma_0^2/n)}\mu_0+\frac{1/(\sigma_0^2)}{1/\tau^2+1/(\sigma_0^2/n)}\bar{x}
$$

注

1. 在正态分布的例子中，均值 μ 的贝叶斯估计本质上是先验分布均值 μ_0 和样本均值 $\bar{x}$ 的加权平均数，其权重本质上是该分布的方差的倒数，也称为该分布的精度。
2. 所求的后验分布与先验分布的分布都是正态分布，这是贝叶斯估计里一种特殊且具有一定普遍性的现象。

定义 21.1

若后验分布 $\pi(\theta\mid x_1,\cdots,x_n)$ 与先验分布属于同一分布族，则称该分布族是 θ 的共轭先验分布（族）。♣

这里我们针对贝叶斯估计做一些更为深入的讨论。

1. **充分性原则体现在贝叶斯估计之中**。
   由于贝叶斯估计是根据参数的后验分布求得的，而参数的后验分布为

$$
\begin{aligned}
\pi(\theta|x_1,x_2,\cdots,x_n) &\propto \pi(\theta)\cdot p(x_1,x_2,\cdots,x_n|\theta)\\
&= \pi(\theta)\cdot g(t(x_1,x_2,\cdots,x_n),\theta)h(x_1,x_2,\cdots,x_n)\\
&\propto \pi(\theta)\cdot g(t(x_1,x_2,\cdots,x_n),\theta)
\end{aligned}
$$

   其中第二个等式成立是根据因子分解定理，$t(x_1,x_2,\cdots,x_n)$ 是 θ 的充分统计量。
2. **贝叶斯估计是一种整体最优的估计方法**。
   对于 θ 的任意一种估计 $\hat{\theta}(x_1,x_2,\cdots,x_n)$，我们之前介绍过可以通过均方误差来进行评

价，即

$$\begin{aligned}\mathrm{MSE}_\theta(\hat\theta) &= E_{x_1,x_2,\cdots,x_n}(\hat\theta-\theta)^2\\ &= \int_{x_n\in R}\cdots\int_{x_1\in R}(\hat\theta-\theta)^2 p(x_1,x_2,\cdots,x_n|\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n\\ &= \int_{\boldsymbol{x}\in R^n}(\hat\theta-\theta)^2 p(\boldsymbol{x}|\theta)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

这里 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)'$。我们之前在介绍均方误差时，参数 θ 认为是一个未知常数，但在贝叶斯学派中，θ 认为是一个随机变量，所以 $MSE(\hat\theta)$ 也被认为是一个随机变量，其依赖于 θ。我们可以整体来比较估计的均方误差，于是计算均方误差的期望，即

$$\begin{aligned}E_\theta(\mathrm{MSE}_\theta(\hat\theta)) &= \int_{\theta\in\Theta}\int_{\boldsymbol{x}\in R^n}\left(\int(\hat\theta-\theta)^2 p(\boldsymbol{x}|\theta)\mathrm{d}\boldsymbol{x}\right)\cdot\pi(\theta)\mathrm{d}\theta\\ &= \int_{\theta\in\Theta}\int_{\boldsymbol{x}\in R^n}(\hat\theta-\theta)^2 p(\boldsymbol{x}|\theta)\cdot\pi(\theta)\mathrm{d}\boldsymbol{x}\mathrm{d}\theta\\ &= \int_{\theta\in\Theta}\int_{\boldsymbol{x}\in R^n}(\hat\theta-\theta)^2\pi(\theta|\boldsymbol{x})m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\theta\\ &= \int_{\boldsymbol{x}\in R^n}\int_{\theta\in\Theta}(\hat\theta-\theta)^2\pi(\theta|\boldsymbol{x})\mathrm{d}\theta m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\end{aligned}$$

令 $\int_{\theta\in\Theta}(\hat\theta-\theta)^2\pi(\theta|\boldsymbol{x})\mathrm{d}\theta$ 为 $R(\hat\theta|\boldsymbol{x})$，称其为条件风险函数。对于给定的样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)'$，我们所求的贝叶斯估计是使得条件风险函数最小的估计，即

$$\hat\theta_{\text{Bayes}}=\arg\min_{\hat\theta}R(\hat\theta|\boldsymbol{x}).$$

我们对 $R(\hat\theta|\boldsymbol{x})$ 关于 $\hat\theta$ 求导，即

$$\begin{aligned}\frac{\partial R(\hat\theta|\boldsymbol{x})}{\partial\hat\theta} &= \frac{\partial}{\partial\hat\theta}\int_{\theta\in\Theta}(\hat\theta-\theta)^2\pi(\theta|\boldsymbol{x})\mathrm{d}\theta\\ &= \int_{\theta\in\Theta}\frac{\partial}{\partial\hat\theta}\left((\hat\theta-\theta)^2\pi(\theta|\boldsymbol{x})\right)\mathrm{d}\theta\\ &= \int_{\theta\in\Theta}\left(2(\hat\theta-\theta)\pi(\theta|\boldsymbol{x})\right)\mathrm{d}\theta\end{aligned}$$

其中第二个等式成立是有条件的。如果感兴趣的同学可以参考华东师范大学数学系编写的《数学分析（下册)》第 187 页定理 19.3。令

$$\int_{\theta\in\Theta}\left(2(\hat\theta-\theta)\pi(\theta|\boldsymbol{x})\right)\mathrm{d}\theta=0$$

即

$$\hat\theta=\int_{\theta\in\Theta}\theta\pi(\theta|\boldsymbol{x})\mathrm{d}\theta=E(\theta|\boldsymbol{x})$$

这就是我们定义的贝叶斯估计。

参考书目

对贝叶斯学派感兴趣的同学，这里提供了一些贝叶斯统计方面的学习资料列表。

1. Gelman A, Carlin J B, Stern H S, et al. Bayesian data analysis[M]. CRC press, 2013.

2. B 站：《贝叶斯统计梅长林西安交通大学》系列视频.
3. 茆诗松，汤银才. 贝叶斯统计. 中国统计出版社.

第 22 讲 假设检验

阅读章节

- Prob & Stat 7.1, 7.2

例题 22.1 （女士品茶）

在 20 世纪初期，英国人在午后会吃下午茶。在下午茶中，奶茶是一种常见的饮品。简单来说，奶茶是由茶和奶混合而成的。某位女士声称，她能够品尝得出这杯奶茶是先加奶，还是先加茶。前者记为 “MT”，后者记为 “TM”。为了确定这位女士的确具有这种 “分辨能力”，R. A. Fisher 在这个问题中提出了一种实验方案：他请侍者制作了 10 杯奶茶，可能是 MT，也可能是 TM，请这位女士进行品尝，并让女士对每一杯奶茶进行甄别。令 X 表示该女士甄别正确的数量。如何度量出女士是否具有真实的分辨能力?

这里，X 可以看作一个二项分布 $b(10,p)$ 随机变量，用 p 来度量这位女士是否具有分辨能力。考虑一下两种情况：

- 如果女士没有甄别能力，甄别正确和甄别错误时等可能的，那么 $p=0.5$。
- 如果女士有甄别能力，当然她仍可能犯错，那么 $0.5<p\leq 1$。

如果女士在 10 杯的判断中都甄别正确，那么这个事件发生的可能性为

$$P(X=10)=p^{10}.$$

如何来看待这个概率值?

- 解释一:这个女士是无这种能力的,则 $p=0.5$。但 $P(X=10)=0.5^{10}=(1024)^{-1}\approx 0.001$，说明：这位女士今天超常发挥。
- 解释二：这个女士是有这种能力的，则 $0.5<p\leq 1$。这个情况下，$P(X=10)=p^{10}$，说明：这个女士今天是正常发挥。

对于这两种解释，我们通常认为：无法否决解释一，但更倾向于接受解释二。这是因为解释一出现的概率太小了。

注 假设检验的根本思想是：在任何假设下，小概率事件通常是不可能发生的；一旦这个小概率事件发生，则所提出的假设不应该被接受。

笔记

22.1 假设检验的基本概念

这里我们先给出假设检验的一些基础概念。

定义 22.1

设有来自某一个参数分布族 $\{F(x,\theta)|\theta\in\Theta\}$ 的样本 $x_1,x_2,\cdots,x_n$，其中 Θ 为参数空间。设 $\Theta_0\subset\Theta$ 且 $\Theta_0\neq\emptyset$。则称命题

$$H_0:\theta\in\Theta_0$$

为一个假设，或原假设，或零假设 (null hypothesis)。若有另一个 $\Theta_1(\Theta_1\subset\Theta,\Theta_1\cap\Theta_0=\emptyset)$，则称命题

$$H_1:\theta\in\Theta_1$$

为 H_0 的对立假设或备择假设（alternative hypothesis）。于是，我们感兴趣的一对假设就是

$$H_0:\theta\in\Theta_0\quad \text{vs}\quad H_1:\theta\in\Theta_1$$

其中，vs 是 versus 的缩写，即表示 H_0 对 H_1 的假设检验问题。♣

注

1. Θ_1 的常见一种情况是 $\Theta_1=\Theta-\Theta_0$；
2. 如果 Θ_0 仅包含一个点，则称其为简单（simple）原假设；否则称为复杂（composite）或复合原假设。备择假设也会有简单和复杂的差别。
3. 当原假设为简单原假设，即 $H_0:\theta=\theta_0$，此时，备择假设通常有三种：
   - $H_1^{'}:\theta\neq\theta_0$；
   - $H_1^{''}:\theta<\theta_0$；
   - $H_1^{'''}:\theta>\theta_0$；

   称 H_0 vs $H_1^{'}$ 为双侧假设或双边假设；称 H_0 vs $H_1^{''}$ 或 H_0 vs $H_1^{'''}$ 为单侧假设或单边假设。

例题 22.2 在女士品茶的例子中，我们提出的假设为

$$H_0:p=0.5\quad \text{vs}\quad H_1:p>0.5.$$

原假设时简单的，备择假设是复杂的；而且这是一个单边假设检验问题。

在提出一个假设检验问题（给出一对假设）之后，我们需要给出一个具体的判断规则，称这个判断规则为该假设的一个检验或者检验法则。

定义 22.2

给定样本 $x_1,x_2,\cdots,x_n$。对于其样本空间 Ω，我们可以确定一组划分

$$W\cap\overline{W}=\emptyset,\quad W\cup\overline{W}=\Omega.$$

当 $(x_1,x_2,\cdots,x_n)\in W$ 时，拒绝 H_0；否则接受 H_0。于是，称 W 为该检验的拒绝域，而称 $\overline{W}$ 为接受域。♣

注

1. 一旦拒绝域确定了，检验的判断准则也就确定了。

因为我们的检验是基于样本而确定的，所以，对于任何的检验而言，我们所作出的判断都可能犯错。检验有常见的两类错误，如表22.1所示。同时，我们需要定义这两类错误发生概率，分别为

表 22.1: 检验的两种错误

		真实情况	
		H_0 为真	H_1 为真
判断结果	判 H_0 为真	正确	犯第二类错误
	判 H_1 为真	犯第一类错误	正确

- 第一类错误发生的概率为 $\alpha(\theta) = P_\theta((x_1, x_2, \cdots, x_n) \in W), \theta \in \Theta_0$.
- 第二类错误发生的概率为 $\beta(\theta) = P_\theta((x_1, x_2, \cdots, x_n) \in \overline{W}), \theta \in \Theta_1$.

注 可以类比"无罪假定"，第一类错误也称为"错判"，而第二类错误也称为"漏判"。

问题 22.1 事实上，第一类错误发生概率和第二类错误发生概率无法同时降低，那么哪一种错误值得我们在意呢?

为统一地表示第一类错误发生的概率和第二类错误发生的概率，我们引入势函数或功效函数 (power function)。

定义 22.3

设检验问题

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

的拒绝域为 W，则样本 $(x_1, x_2, \cdots, x_n)$ 落在拒绝域 W 内的概率称为该检验的势函数，记为

$$g(\theta) = P_\theta((x_1, x_2, \cdots, x_n) \in W), \theta \in \Theta = \Theta_0 \cup \Theta_1.$$

♣

注 势函数

$$g(\theta) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0 \\ 1 - \beta(\theta), & \theta \in \Theta_1 \end{cases}$$

笔记

问题 22.2 坊间流传：码农的平均年龄不大于 35 岁。针对这个传言，你觉得需要怎么验证？

笔记

例题 22.3 设 $x_1, x_2, \cdots, x_n$ 是来自正态分布 $N(\mu, \sigma_0^2)$ 的样本，其中 σ_0^2 是已知的。我们需要检验以下假设

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_0 : \mu > \mu_0.$$

在上述假设中，我们想要知道总体均值 μ 是否大于给定的值 μ_0。因为样本均值 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 是总体均值 μ 的一个合理估计，所以我们需要依据样本均值来推断哪种假设更可能成立。以下有两种情况：

- 第一种情况：$\bar{x} \le \mu_0$，于是，我们没有任何证据来支持备择假设。
- 第二种情况：$\bar{x} > \mu_0$，于是，我们能够直接下结论——支持备择假设吗？

在第二种情况下，我们仍无法直接接受备择假设。这是因为如果只是大一点，这可能是由于样本的随机性造成的。于是，我们需要考虑一个问题：样本均值 $\bar{x}$ 多大，才能接受备择假设。也就是说，我们构造一个拒绝域为

$$\{\bar{x} \ge c\}$$

其中，c 指的是临界值，怎么来确定这个 c 是另一个问题。

为了确定 c，我们需要计算第一类错误发生的概率，即

$$\begin{aligned} P(\bar{x} > c | H_0) &= P\left(\frac{\bar{x} - \mu_0}{\sqrt{\sigma_0^2/n}} > \frac{c - \mu_0}{\sqrt{\sigma_0^2/n}} | H_0\right) \\ &= 1 - \Phi\left(\frac{c - \mu}{\sqrt{\sigma_0^2/n}}\right) \end{aligned}$$

其中第二个等式成立是因为 $\bar{x} \sim N(\mu, \sigma_0^2)$ 且在 H_0 成立时，$\mu = \mu_0$。

记

$$\alpha_{\mu_0}(c) = 1 - \Phi\left(\frac{c - \mu_0}{\sqrt{\sigma_0^2/n}}\right).$$

我们发现 $\alpha_{\mu_0}(c)$ 是关于 c 的减函数。为控制住第一类错误率，我们事先给出一个很小的值 $\alpha > 0$，这个值 α 被称为显著性水平，也就是说，$\alpha_{\mu_0}(c) \le \alpha$。我们可以通过这个不等式将临界值 c 反解出来，即

$$c = \mu_0 + z_{1-\alpha}\sqrt{\sigma_0^2/n}.$$

因此，我们所构造的拒绝域为

$$\begin{aligned} W &= \{(x_1,x_2,\cdots,x_n): \bar{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\sigma_0^2/n}\} \\ &= \left\{(x_1,x_2,\cdots,x_n): \frac{\bar{x}-\mu_0}{\sqrt{\sigma_0^2/n}} \geq z_{1-\alpha}\right\}. \end{aligned}$$

注 这个检验也被称为 z 检验，其中

$$z = \frac{\bar{x}-\mu_0}{\sqrt{\sigma_0^2/n}}$$

是 z 检验的检验统计量。

根据这个例子，我们进一步讨论第一类错误发生的概率 $\alpha(\mu)$ 和第二类错误发生的概率 $\beta(\mu)$ 是无法同时减小。

例题 22.4 （例题 1.3 续）因为我们所构造的拒绝域为 $W = \{(x_1,x_2,\cdots,x_n): \bar{x} \geq c\}$，所以接受域为 $\overline{W} = \{(x_1,x_2,\cdots,x_n): \bar{x} < c\}$. 于是，对于 $\mu > \mu_0$，第二类错误率为

$$\beta_\mu(c) = P_\mu(\bar{x} < c|H_1) = \Phi\left(\frac{c-\mu}{\sqrt{\sigma_0^2/n}}\right),$$

其中，在 H_1 成立时，$\bar{x} \sim N(\mu,\sigma_0^2)$。我们发现 $\beta_\mu(c)$ 是关于 c 的增函数。由此，

- $\alpha_{\mu_0}(c)$ 减小 $\Rightarrow c$ 增加 $\Rightarrow \beta_\mu(c)$ 增加。
- $\beta_\mu(c)$ 减小 $\Rightarrow c$ 减小 $\Rightarrow \alpha_{\mu_0}(c)$ 增加。

综上，第一类错误发生的概率和第二类错误发生的概率不可同时减小。

定义 22.4 (显著性检验)

对检验问题

$$H_0: \theta \in \Theta_0 \quad \text{vs} \quad H_1: \theta \in \Theta_1,$$

如果一个检验满足对任意的 $\theta \in \Theta_0$，都有

$$g(\theta) = \alpha(\theta) \leq \alpha$$

则称该检验是显著性水平为 α 的显著性检验，简称水平为 α 的检验。♣

22.2 单个总体正态分布下的假设检验问题

22.2.1 方差已知时，μ 的检验

例题 22.5 设 $x_1,x_2,\cdots,x_n$ 是来自正态分布 $N(\mu,\sigma_0^2)$ 的样本，其中 σ_0^2 是已知的。我们需要检验以下假设

$$H_0: \mu \leq \mu_0 \quad \text{vs} \quad H_0: \mu > \mu_0.$$

求显著性水平为 α 的检验。

解 因为备择假设依旧为 $\mu > \mu_0$，所以，我们构造拒绝域为

$$W = \{\bar{x} \geq c\}.$$

考虑第一类错误发生的概率为

$$\begin{aligned}\alpha_\mu(c) &= P_\mu(\bar{x} \geq c|H_0) \\ &= 1-\Phi\left(\frac{c-\mu}{\sqrt{\sigma_0^2/n}}\right), \mu \leq \mu_0.\end{aligned}$$

注意到 $\alpha_\mu(c)$ 是关于 μ 的增函数。我们要求对任意 $\mu \leq \mu_0$，有

$$\alpha_\mu(c) \leq \alpha.$$

只需要 $\max_{\mu\leq\mu_0} \alpha_\mu(c) = \alpha$，等价于 $\alpha_{\mu_0}(c) \leq \alpha$。于是可以解得

$$c = \mu_0 + z_{1-\alpha}\sqrt{\sigma_0^2/n}.$$

由此，所构造的拒绝域仍为

$$W = \{(x_1, x_2, \cdots, x_n) : \bar{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\sigma_0^2/n}\}.$$

注 假设 $H_0 : \mu = \mu_0 \quad \text{vs} \quad H_0 : \mu > \mu_0$ 和 $H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_0 : \mu > \mu_0$ 我们构造的水平为 α 的显著性检验是相同的，拒绝域均为 $W = \left\{\bar{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\sigma_0^2/n}\right\}$。

22.2.2 方差未知时，μ 的检验

例题 22.6 设 $x_1, x_2, \cdots, x_n$ 是来自正态分布 $N(\mu, \sigma)$ 的样本，其中 σ 是未知的。我们需要检验以下假设

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_0 : \mu > \mu_0.$$

求显著性水平为 α 的检验。

解 我们构造的拒绝域为

$$W = \{\bar{x} \geq c\}.$$

考虑第一类错误发生的概率为

$$\begin{aligned}\alpha_{\mu_0}(c) &= P_\mu(\bar{x} \geq c|H_0) \\ &= P_\mu\left(\frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}} \geq \frac{c-\mu}{\sqrt{\hat{\sigma}^2/n}}|H_0\right) \\ &= 1 - P\left(\frac{\bar{x}-\mu_0}{\sqrt{\hat{\sigma}^2/n}} \leq \frac{c-\mu_0}{\sqrt{\hat{\sigma}^2/n}}\right)\end{aligned}$$

其中，$\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$ 且检验统计量为

$$\frac{\bar{x}-\mu_0}{\sqrt{\hat{\sigma}^2/n}} \sim t(n-1).$$

为了 $\alpha_{\mu_0}(c) = \alpha$，可解得

$$c = \mu_0 + t_{1-\alpha}(n-1)\sqrt{\hat{\sigma}^2/n}.$$

由此，所构造的拒绝域仍为

$$W = \{(x_1, x_2, \cdots, x_n) : \bar{x} \geq \mu_0 + t_{1-\alpha}(n-1)\sqrt{\hat{\sigma}^2/n}\}.$$

注 这个检验是单样本的 t 检验，这是因为检验统计量服从 t 分布。

上述介绍的检验都是单边检验，以下我们给一个双边检验的例子。

例题 22.7 设 $x_1, x_2, \cdots, x_n$ 是来自正态分布 $N(\mu, \sigma)$ 的样本，其中 σ 是未知的。我们需要检验以下假设

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_0 : \mu \neq \mu_0.$$

求显著性水平为 α 的检验。

解 我们构造的拒绝域为

$$W = \{\bar{x} \leq c_1\} \cup \{\bar{x} \geq c_2\}.$$

考虑第一类错误发生的概率为

$$\begin{aligned}
\alpha_{\mu_0}(c) &= P_{\mu_0}\left(\{\bar{x} \leq c_1\} \cup \{\bar{x} \geq c_2\}\right) \\
&= P_{\mu_0}\left(\{\bar{x} \leq c_1\} | H_0\right) + P\left(\{\bar{x} \geq c_2\}\right) \\
&= P_{\mu_0}\left(\frac{\bar{x} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \leq \frac{c_1 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right) + P_{\mu_0}\left(\frac{\bar{x} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \geq \frac{c_2 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right) \\
&= P\left(t \leq \frac{c_1 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right) + 1 - P\left(t \leq \frac{c_2 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right).
\end{aligned}$$

为了 $\alpha_{\mu_0}(c) = \alpha$，我们将 α 拆成两部分，即

$$\begin{cases}
P\left(t \leq \dfrac{c_1 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right) = \alpha/2 \\
1 - P\left(t \leq \dfrac{c_2 - \mu_0}{\sqrt{\hat{\sigma}^2/n}}\right) = \alpha/2
\end{cases}$$

于是解得

$$\begin{aligned}
c_1 &= \mu_0 - t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n} \\
c_2 &= \mu_0 + t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}.
\end{aligned}$$

由此，所构造的拒绝域仍为

$$W = \{(x_1, x_2, \cdots, x_n) : \bar{x} \leq \mu_0 - t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n} \text{ 或} \bar{x} \geq \mu_0 + t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}\}.$$

22.2.3 假设检验与区间估计之间的关系

在例题 1.7 中，我们构造的拒绝域为

$$W = \{\bar{x} \leq \mu_0 - t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n} \text{ 或} \bar{x} \geq \mu_0 + t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}\}.$$

则接受域为

$$\overline{W} = \left\{\mu_0 - t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n} \leq \bar{x} \leq \mu_0 + t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}\right\}.$$

根据上述区间，并进行变换，可以解得

$$\begin{cases}
\mu_0 \leq \bar{x} + t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n} \\
\mu_0 \geq \bar{x} - t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}
\end{cases}$$

从 μ_0 的角度来看，μ_0 的区间为

$$[\bar{x}-t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n},\bar{x}+t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}].$$

例题 22.8 设 $x_1,x_2,\cdots,x_n$ 来自于一个正态分布 $N(\mu,\sigma^2)$，其中 σ^2 未知。求 μ 的置信水平为 $1-\alpha$ 的置信区间。

解 因为 μ 的点估计为 $\hat{\mu}=\bar{x}$，其分布为 $N(\mu,\sigma^2/n)$，而 $\hat{\sigma}^2=(n-1)^{-1}\sum_{i=1}^n(x_i-\bar{x})^2$。所构造的枢轴量为

$$\frac{\bar{x}-\mu}{\sqrt{\hat{\sigma}^2/n}}\sim t(n-1).$$

所以，μ 的置信水平为 $1-\alpha$ 的置信区间为

$$[\bar{x}-t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n},\bar{x}+t_{1-\alpha/2}(n-1)\sqrt{\hat{\sigma}^2/n}]$$

注 区间估计也可以解决假设检验问题。以显著性水平为 α 的双边检验问题为例，如果所构造的置信水平为 $1-\alpha$ 置信区间能够盖住 θ_0，那么我们可以通过变换确定样本落在接受域 $\overline{W}$ 中，从而接受原假设。

22.3 两个总体正态分布下的假设检验问题

22.3.1 方差已知时，均值差的检验

例题 22.9 设 $x_1,x_2,\cdots,x_m$ 来自正态总体 $N(\mu,\sigma_1^2)$ 的样本，$y_1,y_2,\cdots,y_n$ 来自正态总体 $N(\mu,\sigma_2^2)$ 的样本。两样本相互独立。我们想要检验

$$H_0:\mu_1\le\mu_2+l \quad \text{vs} \quad H_1:\mu_1>\mu_2+l.$$

如果 σ_1^2,σ_2^2 已知，那么求水平为 α 的显著性检验。

解 令 $\theta=\mu_1-\mu_2-l$。原本的假设等价于

$$H_0:\theta\le 0 \quad \text{vs} \quad H_1:\theta>0.$$

通常我们对 θ 的点估计为

$$\hat{\theta}=\bar{x}-\bar{y}-l.$$

在原假设 H_0 下，取 $\theta=0$，有

$$\bar{x}-\bar{y}-l\sim N\left(0,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right).$$

所以，经标准化后，可得检验统计量为

$$\frac{\bar{x}-\bar{y}-l}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}.$$

因此，水平为 α 的显著性检验的拒绝域为

$$W = \left\{ \frac{\bar{x} - \bar{y} - l}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \geq z_{1-\alpha} \right\}.$$

22.3.2 方差未知时，均值差的检验

例题 22.10 设 $x_1, x_2, \cdots, x_m$ 来自正态总体 $N(\mu, \sigma_1^2)$ 的样本，$y_1, y_2, \cdots, y_n$ 来自正态总体 $N(\mu, \sigma_2^2)$ 的样本。两样本相互独立。我们想要检验

$$H_0 : \mu_1 \leq \mu_2 + l \quad \text{vs} \quad H_1 : \mu_1 > \mu_2 + l.$$

如果 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知，那么求水平为 α 的显著性检验。

解 令 $\theta = \mu_1 - \mu_2 - l$。原本的假设等价于

$$H_0 : \theta \leq 0 \quad \text{vs} \quad H_1 : \theta > 0.$$

通常我们对 θ 的点估计为

$$\hat{\theta} = \bar{x} - \bar{y} - l.$$

在原假设 H_0 下，取 $\theta = 0$，有

$$\bar{x} - \bar{y} - l \sim N\left(0, \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)\right).$$

因为 σ^2 是冗余参数，这里我们用和方差进行估计，即

$$s_w^2 = \frac{(m-1)\sum_{i=1}^{m}(x_i - \bar{x})^2 + (n-1)\sum_{i=1}^{n}(y_i - \bar{b})^2}{m+n-2}.$$

所以，经标准化后，可得检验统计量为

$$\frac{\bar{x} - \bar{y} - l}{s_w\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

因此，水平为 α 的显著性检验的拒绝域为

$$W = \left\{ \frac{\bar{x} - \bar{y} - l}{s_w\sqrt{\frac{1}{m} + \frac{1}{n}}} \geq t_{1-\alpha}(m+n-2) \right\}.$$

注 这个检验是两样本独立 t 检验。

22.3.3 成对数据的检验

例题 22.11 为了比较两种谷物种子的优劣，特选取 10 块土质不全相同的工地，并将每块分为面积相同的两部分，分别种植这两种种子，施肥与田间管理在 20 块小块土地上都是一样，表22.2是各小块上的单位产量。假定单位产量服从等方差的正态分布，试问：两种种子的平均单位产量在显著性水平 $\alpha = 0.05$ 上有无差异差异？

解 假定 $x_1, x_2, \cdots, x_n$ 来自正态分布 $N(\mu_1, \sigma^2)$ 和 $y_1, y_2, \cdots, y_n$ 来自正态分布 $N(\mu_2, \sigma^2)$ 且两

表 22.2: 成对数的数据

土地	1	2	3	4	5	6	7	8	9	10
种子 1 (x)	23	35	29	42	39	29	37	34	35	28
种子 2 (y)	30	39	35	40	38	34	36	33	41	31
差 $d = x - y$	-7	-4	-6	2	1	-5	1	1	-6	-3

样本独立。我们需要检验的问题为

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2.$$

首先，我们考虑两样本独立 t 检验。由于检验统计量为

$$t_1 = \frac{\bar{x} - \bar{y}}{s_w\sqrt{1/n + 1/n}} = -1.1937$$

其中，$\bar{x} = 33.1$，$\bar{y} = 35.7$，$s_1^2 = 33.2111$，$s_y^2 = 14.2333$，$s_w^2 = 23.7222$。且 t 分布的 $1 - \alpha/2$ 分位数为

$$t_{0.975}(18) = 2.1009 > |t_1|$$

所以，无法拒绝原假设，即认为两种种子的单位产量平均值没有显著差别。

但是重新看一看数据，我们发现在同一地块上不同种子的产量的差，大于零的数的绝对值比较小，但小于零的数的绝对值比较大，这让我们认为这两种种子的产量是有差异的，但与两样本独立 t 检验的结论不一致。这是因为土地之间的差异性比较大，而在比较种子时，选取了同一个地块上分别验证了两种种子的产量。基于此，我们计算其差，得到 $d_i, i = 1, 2, \cdots, n$。由于两种种子的产量均服从正态分布，其差仍服从正态分布 $N(\mu_d, \sigma_d^2)$，σ_d^2 未知。我们需要检验的问题也转换为

$$H_0 : \mu_d = 0 \quad \text{vs} \quad H_1 : \mu_d \neq 0$$

因为方差未知，我们采用单样本 t 检验。检验统计量为

$$t_2 = \frac{\bar{d}}{\sqrt{s_d^2/n}} = -2.3475,$$

而 t 分布的 $1 - \alpha/2$ 分位数为

$$t_{0.975}(9) = 2.2622 < |t_2|.$$

因此，在显著性水平为 0.05 时，我们发现种子 y 的产量与种子 x 的产量是不同的，且种子 y 的产量更高。

22.4 p 值

在假设检验中，实际问题中需要考虑如何选取显著性水平 α 更为合适？这里我们利用一个简单的假设检验来比较在不同显著性水平的取值下的检验。

例题 22.12 （例题 1.4 续）

正态分布 $N(\mu,\sigma_0^2)$ 下，σ_0^2 已知，我们需要检验

$$H_0:\mu=\mu_0 \quad \text{vs} \quad H_1:\mu>\mu_0$$

基于样本 $x_1,x_2,\cdots,x_n$，我们所构造的拒绝域为

$$\{\bar{x}\geq\mu_0+z_{1-\alpha}\sqrt{\sigma_0^2/n}\}=\left\{\frac{\bar{x}-\mu_0}{\sqrt{\sigma_0^2/n}}\geq z_{1-\alpha}\right\}.$$

如果根据真实的数据，我们可以计算检验统计量 $u=2.25$，那么在不同的显著性水平 α 下，我们得到的结论是不同的，如表22.3。

表 22.3: 不同显著性水平下拒绝域的对比

α	$z_{1-\alpha}$	结论
10^{-4}	3.72	接受
10^{-3}	3.09	接受
0.01	2.33	接受
0.025	1.96	拒绝
0.05	1.64	拒绝
0.1	1.28	拒绝

我们发现存在一个 $\tilde{\alpha}$ 值，恰好使得 $z_{1-\tilde{\alpha}}=2.25$，自然 $\tilde{\alpha}=\Phi(-2.25)$，这就是我们定义的 p 值。

定义 22.5

在一个假设检验问题中，利用样本观测值能够作出拒绝原假设的最小显著性水平称为检验的 p 值。♣

注 有了 p 值之后，我们不需要预先设置显著性水平 α 的具体取值，直接反馈 p 值即可。

根据 p 值，也可以做出假设检验的结论：

- 如果 $p\leq\alpha$，则在显著性水平 α 下拒绝 H_0；
- 如果 $p>\alpha$，则在显著性水平 α 下接受 H_0。