概率统计 B

原著：陈家鼎、刘婉如、汪仁官

《概率统计讲义》（第三版，高等教育出版社）

2017 年 5 月 25 日

目录

课程介绍

课程介绍

- 掌握概率论和数理统计的基本数学知识。
- 训练用概率论和数理统计方法对实际问题进行数学建模的能力。
- 学会解决常见的统计分析问题。
- 是应用型很强的学科。

教材和参考书

- 教材：陈家鼎、刘婉如、汪仁官：《概率统计讲义》（第三版），高等教育出版社，2004.
- 何书元：《概率论》，北京大学出版社，2006。
- Sheldon M. Ross,《概率论基础教程》（A First Course in Probability），（第七版），人民邮电出版社。
- 陈家鼎、孙山泽、李东风、刘力平，《数理统计学讲义》第三版，高等教育出版社，2015 年。
- Robert V. Hogg and Allen T. Craig, Introduction to Mathematical Statistics(5th ed.), Prentice Hall, 1995.

概率论的内容

- 随机事件与概率；
- 随机变量与概率分布；
- 随机变量的数字特征；
- 随机向量；
- 随机过程。

数理统计的内容

- 统计的基本概念；
- 估计；
- 假设检验；
- 回归分析；
- 正交试验法；
- 统计决策和贝叶斯统计；
- 时间序列分析简介。

教学要求

- 认真预习；
- 完成作业；
- 自己学习一种统计数据分析软件，建议学习 R，见李东风主页。

第一章　随机事件与概率

1.1　随机事件及其概率

随机事件与概率

- **随机事件：** 在一定条件下，可能发生也可能不发生的事件。
- **例 1.1** 掷分币，结果“正面朝上”（记作 A）是随机事件。“正面朝下”（记作 B）也是随机事件。
- **例 1.2** 掷两枚分币。

  $A =$“两个都是正面朝上”

  $B =$“两个都是正面朝下”

  $C =$“一个正面朝上，一个正面朝下”
- **例 1.3** 10 件同类产品中有 8 个正品，2 个次品。任意抽取 3 个。

  $A =$“3 个都是正品”

  $B =$“3 个中至少一个是次品”

  $V =$“3 个都是次品”

  $U =$“3 个中至少有一个是正品”
- A, B 是随机事件；
- V 是“不可能事件”；
- U 是“必然事件”；
- 不可能事件和必然事件也看作是随机事件。

概率

- 事件是否发生无法预知，但是其可能性大小可以定量描述。
- 比如，投掷一枚均匀硬币，正面朝上和正面朝下可能性大小相同。
- 投掷两枚均匀硬币，同时为正面和同时为背面可能性大小相同；一个正面一个背面的可能性比都是正面的可能性大。
- **概率**用来定量描述随机事件发生可能性大小。$P(A)$。
- 概率有“频率定义”、“主观定义”、“公理化定义”。

频数

- 投掷一枚分币。条件组 S。
- 条件组 S 大量重复实现时，事件 A 发生的次数，称为**频数**。约占总试验次数的一半。

$$A \text{ 发生的频率} = \frac{\text{频数}}{\text{试验次数}}, \text{ 接近于} \frac{1}{2}$$

- 长期经验积累所得的、所谓某事件发生的可能性大小，就是这个“频率的稳定值”。
- 见 P.3 的多次投掷表格。次数越多，频率越接近 0.5。

概率的频率定义

- **定义 1.1** 在不变的一组条件 S 下，重复做 n 次实验。记 μ 是 n 次试验中事件 A 发生的次数。当试验的次数 n 很大时，如果频率 μ/n 稳定地在某一数值 p 的附近摆动，而且一般说来随着试验次数的增多，这种摆动的幅度越变越小，则称 A 为随机事件，并称数值 p 为随机事件 A 在条件组 S 下发生的**概率**, 记作

$$P(A) = p$$

- 数值 p 的大小是 A 在 S 下发生的可能性大小的数量刻画。例如 0.5 是掷一枚分币出现“正面朝上”的可能性的数量刻画。

- 定义简述：频率具有稳定性的事件叫做随机事件，频率的稳定值叫做该随机事件的概率。

- 随机事件简称事件。实际中遇到的事件一般都是随机事件。

- 频率 μ/n 取值在 $[0,1]$ 范围。所以概率

$$0 \leq P(A) \leq 1.$$

- 对不可能事件 V 和必然事件 U,

$$P(V) = 0,\ P(U) = 1.$$

- 概率的频率定义是近似值。许多测量值都是近似值，所以不必因为只能求得近似值而怀疑真实概率的存在。

概率的主观定义

- 不能重复或不能大量重复的事件如何定义概率？

- **定义 1.2** 一个事件的概率是人们根据已有的知识和经验对该事件发生可能性所给出的个人信念，这种信念用 $[0,1]$ 中的一个数来表示，可能性大的对应较大的数。

- 称为概率的主观定义。

- 例：企业家对产品畅销可能性的预测；医生对某特定病人手术成功的预测。

- 主观概率是当事人对事件作了详细考察并充分利用个人已有的经验形成的“个人信念”，而不是没有根据的乱说一通。也需要谨慎对待。

1.2 古典概型

古典概型

- 某些概率问题可以根据问题本身所具有的“对称性”，充分利用人类长期积累的关于对称性的实际经验，分析事件的本质，就可以直接计算其概率。
- 这是用数学模型求解概率的方法。
- 例如，投掷一枚分币，认为“正面朝上”和“正面朝下”概率相等（对称性），各为 0.5。
- **例 2.1** 盒中 5 个球，3 白 2 黑。从中任取一个。问：取到白球的概率？
- 直观看为 3/5。
- 把 5 个球编号，1—3 号为白球，4—5 号为黑球。
- 取到每个球的概率相同（对称性）。事件互相排斥，概率各为 1/5。把 3 个白球的概率加起来即可。
- **例 2.2** 盒中 5 个球，3 白 2 黑。从中任取两个。问：两个都是白球的概率？
- 这时不能直观得出概率。
- 把 5 个球编号，1—3 号为白球，4—5 号为黑球。
- 可能结果为：

$$\begin{array}{llll} 1+2 & 1+3 & 1+4 & 1+5 \\ 2+3 & 2+4 & 2+5 & \\ 3+4 & 3+5 & & \\ 4+5 & & & \end{array}$$

- 共 10 个可能结果，且发生的机会相同，互斥，除此之外无其它可能。
- 每个结果的概率为 1/10，其中有 $1+2, 1+3, 2+3$ 共 3 个结果为全白球。所以全是白球的概率为 3/10。

等概完备事件组

- **定义 2.1** 称一个事件组 $A_1, A_2, \ldots, A_n$ 为一个 **等概完备事件组**，如果它具有下列三条性质：

  (1) $A_1, A_2, \ldots, A_n$ 发生的机会相同（等可能性）；

  (2) 在任一次试验中，$A_1, A_2, \ldots, A_n$ 至少有一个发生（也就是所谓“除此之外，不可能有别的结果”）（完备性）；

  (3) 在任一次试验中，$A_1, A_2, \ldots, A_n$ 至多有一个发生（也就是所谓“它们是互相排斥的”）（互不相容性）。

- 等概完备事件组也称 **“等概基本事件组”**，其中任一事件 $A_i (i = 1, 2, \ldots, n)$ 称为**基本事件**。

- 例 1.1，投掷一枚分币，等概基本事件组 $n = 2$, 两个基本事件是“正面朝上”和“正面朝下”。

- 其它例子类似可求得等概基本事件组。

- 若 $A_1, A_2, \ldots, A_n$ 是一个等概基本事件组，事件 B 由其中的 m 个基本事件所构成，则

$$P(B) = \frac{m}{n}. \tag{2.1}$$

- **古典概型**就是用等概基本事件组和 (2.1) 来计算事件的概率的模型。

- **例 2.2（续）** 从三个白球和二个黑球中任取两个，共有 $C_5^2 = 10$ 种不同取法，出现机会相同。

- 每种取法为一个基本事件，构成等概完备事件组。

- 其中两个都是白球的事件为 3 个，概率等于 $m/n = 3/10$。

- **例 2.3** 100 件产品，有 5 件次品。任取 50 件。求无次品的概率。

- **解** 共有 C_{100}^{50} 个结果构成等概基本事件组。

- 事件 B: 任取 50 件其中无次品，包括多少个基本事件？

- 必须从 95 个正品中取出 50 件。取法有 C_{95}^{50} 种。
- $$\begin{aligned} P(B) =& C_{95}^{50}/C_{100}^{50} = \frac{95!/(50!45!)}{100!/(50!50!)} \\ =& \frac{50!/45!}{100!/95!} = \frac{50\cdot 49\cdot 48\cdot 47\cdot 46}{100\cdot 99\cdot 98\cdot 97\cdot 96} \\ =& \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{47}{99}\frac{46}{97} = \frac{1081}{38412} \approx 2.8\% \end{aligned}$$
- **例 2.4** 同例 2.3。问：恰好有 2 件次品（记为事件 A）的概率？
- 在基本事件组中，符合条件的事件，必须是从 5 个次品中任取 2 个，从 95 个正品中任取 48 个。
- 共有 $C_5^2 C_{95}^{48}$ 种取法。
- $$\begin{aligned} P(A) =& \frac{C_5^2 C_{95}^{48}}{C_{100}^{50}} \\ =& \frac{\frac{5!}{2!3!}\frac{95!}{47!48!}}{\frac{100!}{50!50!}} = 0.32 \end{aligned}$$
- **例 2.5** 设一批产品共 N 个，其中次品共 M 个。从中任取 n 个。问：恰好出现 m 个次品的概率？
- $0 \le m \le n,\ m \le M,\ n-m \le N-M$。
- 这是例 2.4 的一般化，所以

$$P(\text{恰好出现 } m \text{ 个次品}) = \frac{C_{N-M}^{n-m} C_M^m}{C_N^n} \tag{2.2}$$

- 记 $C_n^k = 0$，当 $k < 0$ 或 $k > n$ 时。
- 例 2.5 的产品分为两类：次品和正品。考虑分为多类的情形。
- **定理 2.1** 设有 N 个东西分成 k 类，其中第 i 类有 N_i 个东西（$i = 1, 2, \ldots, k$），$N_1 + N_2 + \ldots N_k = N$, 从这 N 个东西中任取 n 个，而 $n = m_1 + m_2 + \cdots + m_k$ $(0 \le m_i \le N_i, i = 1, 2, \ldots, k)$, 则事件 $A =$

“恰有 m_1 个属于第 1 类, 恰有 m_2 个属于第 2 类，••••••，恰有 m_k 个属于第 k 类”的概率为

$$P(A)=\frac{C_{N_1}^{m_1}C_{N_2}^{m_2}\cdots C_{N_k}^{m_k}}{C_N^n} \tag{2.3}$$

- 证明：板书。

1.3 事件的运算及概率的加法公式

1.3.1 事件的包含与相等

事件的包含与相等

- 如果事件 A 发生则事件 B 一定发生，就称事件 B 包含事件 A，记作

$$A\subset B \text{ 或} B\supset A.$$

- 如，投掷两枚硬币，A 表示“正好一个正面朝上”，B 表示“至少一个正面朝上”，则 $A\subset B$。
- 如果 $A\subset B$ 且 $B\subset A$，则称事件 A 与事件 B 相等，记作

$$A=B.$$

- 在概率的公理化定义中，事件等同于集合，事件的性质就是集合的性质。

1.3.2 事件的并与交

事件的并与交

- **定义 3.1** 事件“A 或 B”称为事件 A 与事件 B 的**并**，记作 $A\cup B$ 或 $A+B$；某次试验中 $A\cup B$ 发生，即“A 或 B”发生，意味着 A,B 中只要一个发生。
- 事件“A 且 B”称为事件 A 与 B 的**交**，记作 $A\cap B$ 或 AB 或 $A\cdot B$；$A\cap B$ 发生，即“A 且 B”发生，意味着 A 和 B 都发生。

- 例：投掷两枚硬币。A 表示“正好一个正面朝上”，B 表示“正要两个正面朝上, C 表示“至少一个正面朝上”，则

$$A \cup B = C, \quad AC = A, \quad BC = B,$$
$$AB = V(\text{不可能事件})$$

1.3.3 对立事件及事件的差

对立事件

- **定义 3.2** 事件“非 A”称为 A 的对立事件，记作 $\bar{A}$。
- 例如，投掷两枚硬币，“至少一个正面朝上”是“两个都是正面朝下”的对立事件。
- 对立事件是相互的：

$$\overline{(\bar{A})} = A$$

- 在一次试验中，A 和 $\bar{A}$ 互斥，且至少一个发生。即

$$\begin{aligned} A \cap \bar{A} &= V \\ A \cup \bar{A} &= U \end{aligned} \tag{3.1}$$

事件的差

- **定义 3.3** 事件 A 同事件 B 的差表示 A 发生而 B 不发生的时间，记作 $A \backslash B$。
-

$$A \backslash B = A \cap \bar{B} \tag{3.2}$$

- 事件及事件的运算用图形表示，见 P.12 图 1.1。

1.3.4 事件的运算规律

事件的运算规律

- 与集合运算规律相同。
- 并：

$$(1)A \cup B = B \cup A \text{ (" 并" 的交换律)}$$
$$(2)A \cup (B \cup C) = (A \cup B) \cup C \text{ (" 并" 的结合律)}$$
$$(3)A \cup A = A$$
$$(4)A \cup \bar{A} = U$$
$$(5)A \cup U = U$$
$$(6)A \cup V = A$$

- 交：

$$(7)A \cap B = B \cap A \text{ (" 交" 的交换律)}$$
$$(8)(AB)C = A(BC) \text{ (" 交" 的结合律)}$$
$$(9)A \cap A = A$$
$$(10)A \cap \bar{A} = V$$
$$(11)A \cap U = A$$
$$(12)A \cap V = V$$

- 分配律：

$$(13)A(B \cup C) = (AB) \cup (AC) \text{ (分配律)}$$
$$(14)A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \text{ (分配律)}$$

- 交或并的对立事件：

$$(15)\overline{A \cup B} = \bar{A} \cap \bar{B}$$
$$(16)\overline{A \cap B} = \bar{A} \cup \bar{B}$$

1.3.5 事件的互不相容性

事件的互不相容性

- **定义 3.4** 如果事件 A 与事件 B 不能都发生，即

$$AB = V \text{ (不可能事件)}$$

  则称 A 与 B 是互不相容的事件。

- 例：两枚分币，“正好一个正面朝上”与“两个都是正面朝上”互不相容。
- A 与 $\bar{A}$ 互不相容。
- 多个事件互不相容是指两两互不相容。
- 等概完全事件组定义中“互相排斥”也是两两互不相容的意思。

1.3.6 概率的加法公式

概率的加法公式 (1)

- 如果事件 A, B 互不相容，则

$$P(A \cup B) = P(A) + P(B) \tag{3.3}$$

- 其合理性和必要性可以用概率的频率定义解释。
- 推论：

$$P(A) + P(\bar{A}) = P(A \cup \bar{A}) = P(U) = 1$$

  从而得

$$P(A) = 1 - P(\bar{A}), \quad P(\bar{A}) = 1 - P(A) \tag{3.4}$$

- 这样，一个事件的概率难计算而其对立事件的概率容易计算时可用 (3.4) 计算。

- 概率的有限可加性：设 n 个事件 $A_1, A_2, \ldots, A_n$ 互不相容，则

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) \tag{3.5}$$

- 可以从 (3.3) 归纳证明。

概率的加法公式 (2)

- 对任意两个事件 A, B, 有

$$P(A \cup B) = P(A) + P(B) - P(AB) \tag{3.6}$$

- **证明** 易见 $A \cup B = A \cup (B\bar{A})$, 且

$$A \cap (B\bar{A}) = AB\bar{A} = B(A\bar{A}) = BV = V$$

所以

$$P(A \cup B) = P(A \cup (B\bar{A})) = P(A) + P(B\bar{A}) \tag{3.7}$$

又因

$$B = B \cap U = B \cap (A \cup \bar{A}) = (BA) \cup (B\bar{A})$$

且 BA 与 $B\bar{A}$ 互不相容，所以

$$P(B) = P(BA) + P(B\bar{A}), \quad P(B\bar{A}) = P(B) - P(AB)$$

代入 (3.7) 即得 (3.6)。证毕。

- **例 3.1** 袋中有红、黄、白球各一个，每次任取一个，有放回地抽取三次。
- 求：抽到的三个球中没有红球或没有黄球的概率。
- 记 $G =$“三个球都不是红球，$H =$“三个球都不是黄球”。要求 $P(G \cup H)$。
- 注意 G 和 H 不是互不相容。
- $P(G) = \frac{8}{27}$（共有 27 种可能结果，其中没有红球的结果是 $2 \times 2 \times 2$ 个），$P(H) = \frac{8}{27}$。

- $P(GH) = P$(三个球都是白球) $= \frac{1}{27}$(全是白球的结果只有一种)。
- $$P(G \cup H) = P(G) + P(H) - P(GH) = \frac{15}{27} = \frac{5}{9}.$$

无穷个事件的并和交

- **定义 3.5** 设 $A_1, A_2, \ldots$ 是一系列事件，事件 B 表示: 它的发生当且仅当 $A_1, A_2, \ldots$ 中至少一个发生，B 称为 $A_1, A_2, \ldots$ 的并（或者和），记作 $\bigcup_{k=1}^{\infty} A_k$(或 $\sum_{k=1}^{\infty} A_k$)，或 $A_1 \cup A_2 \cup \ldots$。
- $\bigcap_{k=1}^{\infty} A_k$ 表示这样的事件：当且仅当所有 $A_1, A_2, \ldots$ 都同时发生时此事件才发生。
- **例 3.2** 一射手向某目标连续射击，A_1 =｛第一次射击，命中｝，A_k =｛前 $k-1$ 次射击都未中，第 k 次射击命中｝$(k = 2, 3, \ldots)$。B =｛终于命中｝。则 $B = \bigcup_{k=1}^{\infty} A_k$。

概率的完全可加性

- 设 $A_1, A_2, \ldots$ 是一系列事件，如果 $A_1, A_2, \ldots$ 两两互不相容，则

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \tag{3.8}$$

- 由实践经验得出，不能证明。

1.4 集合与事件、概率的公理化定义

1.4.1 集合

集合

- 事件是一种特殊集合，所以事件的运算就是集合的运算。

- **定义 4.1** 一个集合是指具有确切含义的若干个东西的全体。
- 集合 $A, B, C, \ldots$。元素 $a, b, c, \ldots$。
- $a \in A$。
- $a \notin A$。
- 空集 $\emptyset$。

集合的例子

- **例 4.1** 全体正整数的集合。
- **例 4.2** 不大于 10 的正整数的集合。
- **例 4.3** 二维坐标平面上圆心在圆点的半径为 1 的圆（称为单位圆）内点的集合。
- **例 4.5** 红、黄、白三个球有放回抽取三次的结果集合。共 $3^3 = 27$ 个结果。

集合的关系

- 集合**相等**：两个集合的元素完全相同。记作 $A = B$。
- 集合**包含**关系 $A \subset B$: A 的元素都是 B 的元素。也记为 $B \supset A$。
- $A = B \Longleftrightarrow A \subset B$ 且$B \subset A$。

集合的运算

- **并集** $A \cup B$: 属于 A 或者属于 B 的元素的全体组成的集合。
- **交集** $A \cap B$: 既属于 A 也属于 B 的元素的全体组成的集合。
- 只讨论某个非空集合 Ω 的自己的关系，Ω 称为全集。

- **余集**:

$$A^c = \{x : x \in \Omega但x \notin A\}$$

- $(A^c)^c = A$.
- 集合运算可以用平面图形图示。

集合并的运算规则

1. $A \cup B = B \cup A$ (交换律)
2. $(A \cup B) \cup C = A \cup (B \cup C)$ (结合律)
3. $A \cup A = A$
4. $A \cup A^c = \Omega$
5. $A \cup \Omega = \Omega$
6. $A \cup \emptyset = A$

集合交的运算规则

1. $A \cap B = B \cap A$ 交换律
2. $(A \cap B) \cap C = A \cap (B \cap C)$ 结合律
3. $A \cap A = A$
4. $A \cap A^c = \emptyset$
5. $A \cap \Omega = A$
6. $A \cap \emptyset = \emptyset$

并与交的分配律

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

并、交、余的对偶律

1. $(A \cup B)^c = A^c \cap B^c$
2. $(A \cap B)^c = A^c \cup B^c$

1.4.2 事件与集合的关系

事件与集合的关系

- 事件是特殊的集合。事件的运算与集合的运算相同。
- 条件组 S 下的所有可能不同结果的集合记作 Ω，S 下的随机事件就是 Ω 的子集。
- Ω 是必然事件，$\emptyset$ 是不可能事件。
- $A^c = \Omega \backslash A$ 为 A 的对立事件 $\bar{A}$。
- A, B 互不相容即 $A \cap B = \emptyset$。
- **例 4.6** 投掷两枚分币 (条件 S)，所有可能结果为

$$\omega_1 = \text{“上，下”} \quad \omega_2 = \text{“上，上”}$$
$$\omega_3 = \text{“下，上”} \quad \omega_4 = \text{“下，下”}$$

- 全集 $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$。
- 事件 B 为“恰有一个正面朝上”, 则

$$B = \{\omega_1, \omega_3\}$$

- 事件 C 为“至少有一个正面朝上”，则

$$C = \{\omega_1, \omega_2, \omega_3\}$$

- 事件 A 为“两个正面都朝上”，则 $A = \{\omega_2\}$。
- $C = B \cup A$

1.4.3 概率的公理化定义介绍

概率的公理化定义介绍

- 概率的频率解释直观，但数学严密性不足。
- 概率的主观定义则不易被接受，数学严密性不足。
- 用集合论、测度论可以严格定义概率，对需要作公理化假设。
- 为柯尔莫戈罗夫 (Kolmogorov A. N., 1903-1987) 于 1933 年建立。

概率的公理化定义介绍

- 设 Ω 为一个非空集合，叫做**基本事件空间**。
- Ω 的一些子集组成的集合 $\mathcal{F}$ 叫做 σ **代数**，如果

$$(1)\Omega \in \mathscr{F}$$
$$(2)若A \in \mathscr{F}, 则A^c = \Omega - A \in \mathscr{F}$$
$$(3)若A_n \in \mathscr{F}(n = 1, 2, \dots), 则 \cup_{n=1}^{\infty} A_n \in \mathscr{F}$$

- 事件是 $\mathscr{F}$ 中的集合。
- $\mathscr{F}$ 上有定义的函数 $P = P(\cdot)$ 叫做**概率测度** (简称**概率**), 若

$$(1)P(A) \geq 0(\forall A \in \mathscr{F}) \tag{4.17}$$
$$(2)P(\Omega) = 1 \tag{4.18}$$
$$(3)若A_n \in \mathscr{F}(n = 1, 2, \dots), 且两两不相交, 则$$
$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \text{ (完全可加性)} \tag{4.19}$$

σ 代数的性质

- 附有 $\mathscr{F}, P$ 的 Ω 叫做概率空间。

- σ 代数是可以合理定义概率的事件的全体，有些情况下不是所有 Ω 的子集都可以合理定义概率。
- 若 Ω 为有限集或可数集则 $\mathscr{F}$ 通常取为 Ω 的所有子集的集合。
- $\mathscr{F}$ 关于基本集合运算封闭：有穷个或无穷个集合的并、交，两个集合的差。

概率的性质

1 $P(\emptyset) = 0$

2 若 $A \in \mathscr{F}$ 则 $P(A^c) = 1 - P(A)$

3 若 $A_1, \ldots, A_n$ 都属于 $\mathscr{F}$ 且两两不相交，则

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \text{ (有限可加性)} \tag{4.20}$$

4 若 $A \subset B$, $A, B \in \mathscr{F}$, 则 $P(A) \leq P(B)$ 且

$$P(B \backslash A) = P(B) - P(A)$$

5 若 $A_n \subset A_{n+1}, A_n \in \mathscr{F} (n = 1, 2, \ldots)$，则

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \text{ (单调上升事件的概率极限)}$$

6 若 $A_n \supset A_{n+1}, A_n \in \mathscr{F} (n = 1, 2, \ldots)$，则

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \text{ (单调下降事件的概率极限)}$$

1.5 条件概率、乘法公式、独立性

1.5.1 条件概率

条件概率

- 条件概率在条件 S 的基础上附加了条件，讨论则附加条件之后的概率。
- **例 5.1** 16 个球，6 个玻璃球，10 个木球。玻璃球中有 2 个红色，4 个蓝色；木球中有 3 个红色，7 个蓝色。从 16 个球中任取一个。
-

	玻璃	木质	
红	2	3	5
蓝	4	7	11
	6	10	16

- 记 $A=$“取到蓝球”, $B=$ “取到玻璃球”。$P(A)=\frac{11}{16}$, $P(B)=\frac{6}{16}$。
- 问：如果已知取到的是蓝球，则该球是玻璃球的概率？即事件 A 已经发生前提下事件 B 发生的概率，记作 $P(B|A)$。
- 可以用古典概型计算。蓝球共有 11 个，其中 4 个是玻璃球。

$$P(B|A)=\frac{4}{11}$$

- **定义 5.1** 如果 A,B 是条件组 S 下的两个随机事件，$P(A)\neq 0$，则称在 A 发生的前提下 B 发生的概率为 **条件概率**, 记作 $P(B|A)$。
- 注意这不是严格数学定义。
- **例 5.2** 5 个乒乓球，3 新 2 旧。每次取一个，无放回取两次。
- A=“第一次取到新球”；$B=$“第二次取到新球”。
- 求 $P(A)$, $P(B)$, $P(B|A)$。
- $P(A)=\frac{3}{5}$。
- $P(B)$ 用古典概型，可以把 5 个球编号，则两次抽取所有可能结果有 $5\times 4=20$ 种，其中第二次抽取到新球的结果数为 $3\times 2+2\times 3=12$ 种，$P(B)=\frac{12}{20}=\frac{3}{5}$，即抽签是公平的。
- 若 A 已经发生，则还剩 2 新 2 旧，于是第二次取到新球的概率为 $\frac{2}{4}=\frac{1}{2}$, 即 $P(B|A)=\frac{1}{2}$。

1.5.2 乘法公式

乘法公式

- 条件概率的等价定义为（多数教材这样定义条件概率）

$$P(B|A)=\frac{P(AB)}{P(A)} \tag{5.1}$$

- (5.1) 改写为

$$P(AB)=P(A)P(B|A) \tag{5.1'}$$

  称为概率的**乘法公式**。

- (5.1) 用来在已知 $P(A)$ 和 $P(AB)$ 时求条件概率；
- (5.1') 用来在已知 $P(A)$ 和 $P(B|A)$ 时求 $P(AB)$。
- 在古典概型下由定义 5.1 可以证明 (5.1)。
- 设条件组 S 下一个等概完备事件组有 n 个基本事件，A 由其中 m 个组成，B 由其中 l 个组成，AB 由 k 个组成。
- 则

$$\begin{aligned}P(B|A)=&\frac{\text{在 } A \text{ 发生的前提下 } B \text{ 中包含的基本事件数}}{\text{在 } A \text{ 发生的前提下的基本事件总数}}\\=&\frac{k}{m}=\frac{k/n}{m/n}=\frac{P(AB)}{P(A)}\end{aligned}$$

1.5.3 独立性

独立性

- **例 5.3** 5 个乒乓球，3 新 2 旧，每次取 1 个，有放回取 2 次。
- $A=$“第一次取到新球”
- $B=$“第二次取到新球”
- 显然 $P(B|A)=P(B)$，与 A 是否发生无关。

- 这时

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

- **定义 5.2** 称两个随机事件 A, B 是**相互独立**的, 如果

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

- 定义中不要求 $P(A) > 0$ 或 $P(B) > 0$。

独立性的直观解释

- 事件 A 是否发生不影响事件 B 的发生概率，事件 B 是否发生也不影响事件 A 的发生概率。
- 在 $P(A) \neq 0, P(B) \neq 0$ 时，独立等价于

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

  也等价于

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

- 即条件概率等于无条件概率为独立。
- **例 5.4** 甲、乙同时向一敌机炮击。
- 甲击中概率 0.6；乙击中概率 0.5。
- 求被击中的概率。
- **解：** 记 $A =$“甲击中”，$B ==$“乙击中”；$C =$“敌机被击中”。
- $P(C) = P(A \cup B) = P(A) + P(B) - P(AB)$。
- 可以认为 A, B 独立。

$$\begin{aligned} P(AB) =& P(A) \times P(B) = 0.6 \times 0.5 = 0.3 \\ P(C) =& 0.6 + 0.5 - 0.3 = 0.8 \end{aligned}$$

- 另解：

$$\begin{aligned} P(C) =& 1 - P(\bar{C}) = 1 - P(\overline{A \cup B}) \\ =& 1 - P(\bar{A} \cap \bar{B}) = 1 - P(\bar{A})P(\bar{B}) \\ =& 1 - (1 - 0.6)(1 - 0.5) = 0.8 \end{aligned}$$

- 把并的概率用交的概率来求解是常用的手法。
- 另解用到了：A, B 独立则 $\bar{A}, \bar{B}$ 也独立。

对立事件与独立

- **定理 5.1** 若四对时间 A, B, $A, \bar{B}$, $\bar{A}, B$, $\bar{A}, \bar{B}$ 中有一对独立，则另外三对也独立。
- 即这四对事件或者都独立，或者都不独立。
- **证明** 仅证明 A, B 独立 $\Longrightarrow A, \bar{B}$ 独立。

$$\begin{aligned} P(A) =& P(AU) = P(A \cap (B \cup \bar{B})) \\ =& P((AB) \cup (A\bar{B})) = P(AB) + P(A\bar{B}) \end{aligned}$$

于是

$$\begin{aligned} P(A\bar{B}) =& P(A) - P(AB) = P(A) - P(A)P(B) \\ =& P(A)[1 - P(B)] = P(A)P(\bar{B}) \end{aligned}$$

多个事件相互独立

- **定义 5.3** 称 A, B, C 是相互独立的，如果有

$$\begin{aligned} P(AB) &= P(A)P(B) \\ P(AC) &= P(A)P(C) \\ P(BC) &= P(B)P(C) \\ P(ABC) &= P(A)P(B)P(C) \end{aligned} \tag{5.3}$$

- **定义 5.4** 称 $A_1, A_2, \ldots, A_n$ 是相互独立的，如果对任意整数 $k(2 \leq k \leq n)$ 以及从 $1, 2, \ldots, n$ 中任意取出的 k 个 $i_1, i_2, \ldots, i_k$ 都满足

$$P(A_{i_1} A_{i_2} \ldots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \ldots P(A_{i_k}) \tag{5.4}$$

- 其中一个要求是

$$P(A_1 A_2 \ldots A_n) = P(A_1) P(A_2) \ldots P(A_n)$$

- **例 5.5** 某型号高射炮单发击中飞机概率为 0.6。若干门发射单发，欲以 99% 概率击中敌机。求高炮门数。

- **解：** 设需要 n 门，A_i 为“第 i 门高炮击中敌机”。

- $A =$“敌机被击中”。$A = A_1 \cup A_2 \cup \cdots \cup A_n$。

$$\begin{aligned} P(A) =& P(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - P(\bar{A}_1 \cap \bar{A}_2 \cap \cdots \cap \bar{A}_n) \\ =& 1 - P(\bar{A}_1) P(\bar{A}_2) \ldots P(\bar{A}_n) \text{ (独立性)} \\ =& 1 - (1 - 0.6)^n \geq 0.99 \\ n \geq& \frac{\log 0.01}{\log 0.4} = 5.026 \end{aligned}$$

- 需要 6 门高射炮。

- **例 5.6** 三个事件两两独立不能保证三个事件独立的例子。

- 均匀正四面体，四面涂红色、黄色、蓝色、红黄蓝混杂。

- 投掷一次，考察底面出现的颜色。

- $A =$“红色出现”，$B =$“黄色出现”，$C =$“蓝色出现”。

- 基本事件：$A_i =$“第 i 面在底面”，$i = 1, 2, 3, 4$, 构成等概基本事件组。

- $A = A_1 \cup A_4$, $B = A_2 \cup A_4$, $C = A_3 \cup A_4$。

- $P(A) = P(B) = P(C) = \frac{1}{2}$。

- $AB = AC = BC = A_4$，$P(AB) = P(AC) = P(BC) = \frac{1}{4}$, 按定义 A, B 相互独立，A, C 相互独立，B, C 相互独立。

- 但 $ABC = A_4$, $P(ABC) = \frac{1}{4} \neq P(A) P(B) P(C)$。

1.6 全概公式与逆概公式

1.6.1 全概公式

- **例 6.1** 5 个乒乓球，3 新 2 旧。每次取一个，无放回取两次。求第二次取到新球的概率。
-

$$\begin{aligned} A =&\text{“第一次取到新球”}\\ B =&\text{“第二次取到新球”}\\ B =&BA\cup B\bar{A} \qquad (6.1)\\ P(B) =&P(BA)+P(B\bar{A})\\ =&P(A)P(B|A)+P(\bar{A})P(B|\bar{A})\\ =&\frac{3}{5}\cdot\frac{2}{4}+\frac{2}{5}\cdot\frac{3}{4}=\frac{3}{5} \end{aligned}$$

- (6.1) 将复杂的事件（情况）分解为简单的事件（情况）。

全概公式

- **定理 6.1（全概公式）** 如果事件组 $A_1, A_2, \ldots, A_n$ 满足：
- (1) $A_1, A_2, \ldots, A_n$ 互不相容，且 $P(A_i) > 0 (i = 1, 2, \ldots, n)$。
- (2) $A_1 \cup A_2 \cup \cdots \cup A_n = U$(完备性)，
- 则对任一事件 B 皆有

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i). \qquad (6.2)$$

- **证明** $B = BU = BA_1 \cup BA_2 \cup \cdots \cup BA_n$,

$$\begin{aligned} P(B) =&P(BA_1)+P(BA_2)+\cdots+P(BA_n)\\ =&P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+\cdots+P(A_n)P(B|A_n). \end{aligned}$$

- 满足条件（1）和（2）的事件组 $A_1, A_2, \ldots, A_n$ 称为完备事件组。比等概完备事件组少了等概条件。

- 更一般的全概公式中的完备事件组可以包含可数个事件。
- 运用全概公式关键在于求完备事件组。
- **例 6.2** 甲、乙、丙三人射击敌机。击中概率:

  甲：0.4

  乙：0.5

  丙：0.7

- 若只有一人击中，飞机坠毁概率 0.2；若恰好二人击中，坠毁概率 0.6；三人全中，坠毁概率 1。
- 求飞机坠毁概率。
- **解** $B =$“飞机坠毁”, $A_0 =$“三人都不中”; $A_1 =$“恰好一人击中”; $A_2 =$“恰好二人击中”; $A_3 =$“三人都击中”。
- A_0, A_1, A_2, A_3 构成完备事件组。
- 已知 $P(B|A_0) = 0$, $P(B|A_1) = 0.2$, $P(B|A_2) = 0.6$, $P(B|A_3) = 1$。
-

$$\begin{aligned} P(A_0) =& P(\text{甲不中})P(\text{乙不中})P(\text{丙不中}) \\ =& (1-0.4)(1-0.5)(1-0.7) = 0.09 \end{aligned}$$

-

$$\begin{aligned} P(A_1) =& P(\text{甲中})P(\text{乙不中})P(\text{丙不中}) \\ &+ P(\text{甲不中})P(\text{乙中})P(\text{丙不中}) \\ &+ P(\text{甲不中})P(\text{乙不中})P(\text{丙中}) \\ =& 0.4 \times (1-0.5) \times (1-0.7) \\ &+ (1-0.4) \times 0.5 \times (1-0.7) \\ &+ (1-0.4) \times (1-0.5) \times 0.7 \\ =& 0.36 \end{aligned}$$

- $$\begin{aligned} P(A_2) =&P(\text{甲不中})P(\text{乙中})P(\text{丙中}) \\ &+ P(\text{甲中})P(\text{乙不中})P(\text{丙中}) \\ &+ P(\text{甲中})P(\text{乙中})P(\text{丙不中}) \\ =&(1-0.4)\times 0.5\times 0.7 \\ &+ 0.4\times(1-0.5)\times 0.7 \\ &+ 0.4\times 0.5\times(1-0.7) \\ =&0.41 \end{aligned}$$

- $$\begin{aligned} P(A_3) =&P(\text{甲中})P(\text{乙中})P(\text{丙中}) \\ =&0.4\times 0.5\times 0.7 = 0.14 \end{aligned}$$

- $$\begin{aligned} &P(\text{敌机坠毁}) = P(B) \\ =&\sum_{i=0}^{3} P(A_i)P(B|A_i) \\ =&0.09\times 0 + 0.36\times 0.2 + 0.41\times 0.6 + 0.14\times 1 \\ =&0.458 \end{aligned}$$

- **例 6.3** 设甲有赌本 M 元，乙有赌本 N 元（M, N 是正整数）。
- 每一局输赢为 1 元，没有和局。
- 每局甲胜概率为 $p(0 < p < 1)$。
- 问：甲输光的概率。
- **解** 记 $L = M + N$, $L \geq 2$。当 $L = 2$ 时 $M = N = 1$, 甲输光概率为 $1 - p$ (若第一局甲赢，则乙输光，赌局不能继续)。
- 只考虑 $L \geq 3$ 的情形。
- 问题扩充为：若甲乙共有赌本 L 元，甲有赌本 i 元，乙有赌本 $L - i$ 元，则甲输光的概率 p_i 是多少？（原问题求 p_M）

- 记 $A_i =$“甲有赌本 i 元而最后输光”（$i = 1, 2, \ldots, L-1$）, $B =$“甲赢了第一局”。
- 记 $q = 1 - p$。
- 当 $2 \le i \le L-2$ 时，得递推公式

$$\begin{aligned} p_i =& P(B)P(A_i|B) + P(\bar{B})P(A_i|\bar{B}) \\ =& pp_{i+1} + qp_{i-1} \end{aligned} \tag{6.3}$$

- $$\begin{aligned} p_1 =& pp_2 + q \\ p_{L-1} =& p \times 0 + qp_{L-2} \end{aligned}$$
- 记 $p_0 = 1$, $p_L = 0$, 则 (6.3) 对 $1 \le i \le L-1$ 成立。
- 由（6.3）

$$\begin{aligned} p_i =& pp_i + qp_i = pp_{i+1} + qp_{i-1} \\ p(p_{i+1} - p_i) =& q(p_i - p_{i-1}) \\ (p_{i+1} - p_i) =& \frac{q}{p}(p_i - p_{i-1}) = \cdots\cdots \\ =& \left(\frac{q}{p}\right)^i (p_1 - 1) \\ =& r^i(p_1 - 1), \quad (i = 1, 2, \ldots, L-1, \text{ 记} r = \frac{q}{p}) \\ p_{i+1} - p_1 =& \sum_{k=1}^{i} (p_{k+1} - p_k) \\ =& \sum_{k=1}^{i} r^k (p_1 - 1) \end{aligned} \tag{6.4}$$

- $$p_{i+1} - p_1 = \begin{cases} \dfrac{r - r^{i+1}}{1 - r}(p_1 - 1) & p \ne \frac{1}{2} \\ i(p_1 - 1) & p = \frac{1}{2} \end{cases}$$

- $i = L-1$ 时 $p_{i+1} = p_L = 0$,

$$\begin{aligned} p_1 &= \begin{cases} \dfrac{r - r^L}{1-r}(1-p_1) & p \neq \frac{1}{2} \\ i(1-p_1) & p = \frac{1}{2} \end{cases} \\ &= \begin{cases} \dfrac{r - r^L}{1-r^L} & p \neq \frac{1}{2} \\ 1 - \frac{1}{L} & p = \frac{1}{2} \end{cases} \end{aligned} \tag{6.5}$$

- 由（6.4），$p \neq \frac{1}{2}$ 时

$$\begin{aligned} p_i =& p_1 + \frac{r - r^i}{1-r}(p_1 - 1) \\ =& \frac{r^i - r^L}{1-r^L} \quad (2 \leq i \leq L-1) \end{aligned}$$

- $p = \frac{1}{2}$ 时

$$\begin{aligned} p_i =& p_1 + (i-1)(p_1 - 1) \\ =& 1 - \frac{1}{L} + (i-1)\left(-\frac{1}{L}\right) \\ =& 1 - \frac{i}{L} \quad (2 \leq i \leq L-1) \end{aligned}$$

- 甲输光的概率

$$p_M = \begin{cases} \frac{r^M - r^{M+N}}{1-r^{M+N}} & p \neq \frac{1}{2} \\ \frac{N}{M+N} & p = \frac{1}{2} \end{cases}$$

- 关键是根据第一局的输赢结果建立方程（6.3）。叫做“首步（首局）分析法”。

- **例 6.4** 在问卷调查时，某些敏感问题会遭到拒绝回答或谎报。

- 比如，要问卷调查运动员是否曾服用兴奋剂，直接问很难得到肯定回答。

- 设计如下两个问题：

- 问题 A：你的生日是否在 7 月 1 日之前（不含 7 月 1 日）?

- 问题 B：你是否服用过兴奋剂？
- 被调查者只需要回答其中一个问题，只需在只有“是”、“否”的答卷上选择其一。而回答哪一个是根据被调查者在其他人不能知道的情况下随机抽取一个颜色决定的。
- 若抽出白球，则回答问题 A；若抽出红球，则回答问题 B。红球比例 π 已知。
- 样本量较大（如 200 个受调查者）就可以统计, 估计服用兴奋剂比例 p。
- 设 n 张答卷，k 张答“是”，答“是”的比例 $\varphi = k/n$。
- 全概公式：

$$\begin{aligned} P(\text{回答“是”}) =& P(\text{抽到白球})P(\text{生日在 7 月 1 日前}|\text{抽到白球}) \\ &+ P(\text{抽到红球})P(\text{服用过兴奋剂}|\text{抽到红球}) \\ =& 0.5(1-\pi) + p\pi \\ \approx& \frac{k}{n} \\ p \approx& \frac{\frac{k}{n} - \frac{1-\pi}{2}}{\pi} \end{aligned}$$

- 例如，50 个球中有 30 个红球，$\pi = 0.6$。
- 某国 15 个项目 $n = 246$ 个运动员接受调查, 答“是”者 $k = 54$。
- $$p \approx 0.0325$$
- 即约 3.25% 服用兴奋剂。
- 思考：比例 π 的选取有何影响？

1.6.2 逆概公式

- **例 6.5（发报与接收）** 发报台分别以概率 0.6 和 0.4 发出信号“•”和“−”。

- 信号可能误码。正确接收与错误接收的概率如下表：

		接收	
		•	–
发出	•	0.8	0.2
	–	0.1	0.9

- 求当收到信号 “•”，发报台真的发出 “•” 的概率。
- 记 $A =$ “发出信号 ‘ • ’”，$B =$ “收到信号 ‘ • ’”。要求 $P(A|B)$。
-

$$
\begin{aligned}
P(A|B) =& \frac{P(AB)}{P(B)} \\
P(AB) =& P(A)P(B|A) = 0.6 \times 0.8 \\
P(B) =& P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) \\
P(\bar{A}) =& 0.4 \\
P(B|\bar{A}) =& 0.1 \\
P(A|B) =& \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \\
=& \frac{0.6 \times 0.8}{0.6 \times 0.8 + 0.4 \times 0.1} = 0.923
\end{aligned}
$$

逆概公式

- 逆概公式是全概公式的引申。
- 若有多个基本情况（事件）是完备事件组，则观测到一个结果后，可以逆推原来到底是哪一个情况。
- 如：接收到 “•” 后，可以知道原来发出的是 “•” 的概率为 0.923，即基本可以判断原来是发出的 “•”。
- **定理 6.2（逆概公式）** 设 $A_1, A_2, \ldots, A_n$ 为一完备事件组，则对任一事件 $B(P(B) \neq 0)$ 有

$$
P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \tag{6.6}
$$

- **证明** 联合条件概率定义及全概公式。
- 逆概公式也称为贝叶斯（Bayes）公式。
- **例 6.6（艾滋病检测）** 美国的艾滋病人比例保守估计 1000 分之一。
- 是否应该对新婚夫妇进行艾滋病毒血液检测？
- 血液检测方法各种结果及概率

		检验结果	
		报告阳性	报告阴性
实际情况	AIDS	真阳性 (0.95)	假阴性 (0.05)
	非 AIDS	假阳性 (0.01)	真阴性 (0.99)

- 如果报告阳性，则真正患病概率是多少？
- $A =$“受试者带有艾滋病毒”，$T =$“检测结果呈阳性”。
- 求 $P(A|T)$。
- $P(A) = 0.001$, $P(T|A) = 0.95$, $P(T|\bar{A}) = 0.01$。
- 由逆概公式

$$\begin{aligned} P(A|T) =& \frac{P(A)P(T|A)}{P(A)P(T|A) + P(\bar{A})P(T|\bar{A})} \\ =& \frac{0.001 \times 0.95}{0.001 \times 0.95 + 0.999 \times 0.01} = 0.087 \end{aligned}$$

- 即使检验报告阳性，真的患病的概率也只有 8.7%，所以全面的检验不太必要。
- 原因是 $P(A)$ 太小了。

$$\begin{aligned} P(A|T) =& \frac{0.95P(A)}{0.95P(A) + 0.01(1 - P(A))} \\ =& \frac{0.95}{0.94 + 0.01\frac{1}{P(A)}} \end{aligned}$$

是 $P(A)$ 的严格增函数。

1.7 独立试验序列概型

- **例 7.1** 独立重复掷 5 次分币。
- 求：恰有两次正面朝上的概率。
- **解** 古典概型。共有 $2^5 = 32$ 个等概基本事件。
- 其中恰有两次正面朝上的个数为 $C_5^2 = 10$。
- $p = \frac{10}{32}$。

$$p = C_5^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 \tag{7.1}$$

- （7.1）中，C_5^2 是事件对应的等概基本事件个数，每个基本事件的概率为 $\left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3$。(7.1) 可以用加法公式说明。
- 若分币不均匀，每一次“正面朝上”概率为 $\frac{2}{3}$, 则

$$P(\text{恰有两次正面朝上}) = C_5^2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^3$$

- 这已经不是古典概型，因为基本事件不等概。
- **例 7.2** 某人打靶，命中率为 0.7，独立重复射击 5 次。
- 求恰好命中 2 次的概率。
- 记 $p = 0.7, q = 1 - p = 0.3$。
- $P(\text{恰好命中 2 次}) = C_5^2 p^2 q^3$。
- $P(\text{恰好命中 3 次}) = C_5^3 p^3 q^2$。
- $P(\text{恰好命中 4 次}) = C_5^4 p^4 q^1$。
- $P(\text{恰好命中 5 次}) = C_5^5 p^5 q^0$。
- $P(\text{恰好命中 1 次}) = C_5^1 p^1 q^4$。
- $P(\text{恰好命中 0 次}) = C_5^0 p^0 q^5$。

独立试验序列概型

- **定理（独立试验序列概型）** 设单次试验中，事件 A 发生的概率为 $p(0<p<1)$, 则在 n 次重复试验中，

$$P(A\text{发生 } k \text{ 次}) = C_n^k p^k q^{n-k} \quad (q=1-p)$$
$$(k=0,1,2,\ldots,n)$$

- **证明** 在 n 次重复试验中，记 $B_1, B_2, \ldots, B_m$ 为构成事件“A 发生 k 次”的那些试验结果。
- （1）“A 发生 k 次”$= B_1 \cup B_2 \cup \cdots \cup B_m$, 互不相容；
- （2）$P(B_1) = P(B_2) = \cdots = P(B_m) = p^k q^{n-k}$;
- （3）$m = C_n^k$（从 n 次试验中选取 k 个成功试验的方法数）。
- 于是用加法公式（1）证明定理结论。
- **注意** “重复”蕴含两重含义：
- （1）每次试验的条件相同，从而事件 A 发生的概率（称为成功概率）不变；
- （2）各次试验的结果独立。
- 当然，这只是理想化假设，实际情况只要比较近似满足就可以了。
- **反例：** 已知 80 个产品中有 5 个次品，从中每次任取一个，无放回地取 20 次，求其中有 2 个次品的概率。
- 这个例子：（1）每次抽取的试验条件不同，不能直接认为每次的成功概率（这里是“取到次品”的事件概率）不变；
- （2）前后的抽取结果不是独立的。如果前 5 次抽取到的都是次品，则从第 6 次起只能抽取到正品。
- 所以不适用独立试验序列概型。
- 产品批量特别大时，“无放回”抽取与“有放回”抽取结果相似，可以用独立试验序列概型近似计算无放回抽样的概率。

- **例 7.3** 设每次射击打中目标的概率等于 0.001。如果射击 5000 次，求至少两次打中目标的概率。

- $p = 0.001, q = 0.999$。

$$\begin{aligned} P(\text{至少两次打中目标}) &= \sum_{k=2}^{5000} P(\text{恰有 } k \text{ 次打中目标}) \\ &= 1 - P(\text{都不中}) - p(\text{仅中一次}) \\ &= 1 - q^{5000} - 5000 \times pq^{4999} \approx 1 - 0.006721 - 0.03364 \\ &\approx 0.9596 \end{aligned}$$

成功 k 次概率近似公式一

- 当 n 很大同时 p 很小的时候，有近似公式

$$P(A \text{ 发生 } k \text{ 次}) \approx \frac{(np)^k}{k!} e^{-np} \tag{7.2}$$

  称为泊松分布近似，见 §2.2(P55)。

- 如，

$$\begin{aligned} P(\text{都不中}) &\approx e^{-5000\times 0.001} \approx 0.006738 \\ P(\text{仅中一次}) &\approx \frac{5000 \times 0.001}{1!} e^{-5000\times 0.001} \approx 0.03369 \\ P(\text{至少两次打中}) &\approx 1 - 0.006738 - 0.03369 \approx 0.9596 \end{aligned}$$

成功 k 次概率近似公式二

- 当 n 很大但 p 不是很小时，有第二近似公式

$$\begin{aligned} P(A \text{ 发生 } k \text{ 次}) &\approx \frac{1}{\sqrt{np(1-p)}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2} \\ x_k &= \frac{k - np}{\sqrt{np(1-p)}} \end{aligned}$$

- 参见 §4.6(P153) 的中心极限定理。

- **例 7.4** 设每次射击打中目标概率为 $\frac{1}{6}$。如果射击 6000 次，问：击中次数在 900 到 1100 之间的概率？
- 需要使用 §4.6 的中心极限定理。记 $n = 6000$, $p = \frac{1}{6}$,

$$\begin{aligned}
&P(\text{击中次数在 900 到 1100 之间})\\
=&P(900 - 0.5 < X < 1100 + 0.5)\\
=&P\left(\frac{900 - 0.5 - np}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{1100 + 0.5 - np}{\sqrt{np(1-p)}}\right)\\
=&\Phi\left(\frac{1100 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{900 - 0.5 - np}{\sqrt{np(1-p)}}\right)\\
\approx&0.99950
\end{aligned}$$

- **例 7.5（自由随机游动）** 设一质点在数轴上运动，在时刻 0 从原点出发。
- 每隔单位时间位置向右或向左移动一个单位，向右移动概率 $p(0 < p < 1)$, 向左移动概率 $q = 1 - p$。
- 问：质点在时刻 n 位于 K 的概率？（n 是正整数，K 是整数）
- 只考虑 K 为正整数情况，负整数和零的情况类似。
- 为了质点在时刻 n 位于 K, 必须且只需在前 n 步移动中向右移动的次数比向左移动的次数多 K。
- 设 x 表示向右移动的次数，y 表示向左移动的次数，则

$$x + y = n, \quad x - y = K$$

- $x = \frac{n+K}{2}$。
- 因 x, n, K 都是整数所以 n 与 K 有相同的奇偶性。当 n 与 K 奇偶性相反时概率为 0。
-

$$\begin{aligned}
&P(\text{质点在时刻 } n \text{ 位于 } K)\\
=&P(\text{质点在头 } n \text{ 次游动时有 } \frac{n+K}{2} \text{ 次向右，有 } \frac{n-K}{2} \text{ 次向左})\\
=&C_n^{\frac{n+K}{2}} p^{\frac{n+K}{2}} q^{\frac{n-K}{2}}
\end{aligned}$$

- 易见 $K \leq 0$ 时也成立。

第二章　随机变量与概率分布

2.1　随机变量

随机变量概念引入

- 一些事件的结果用变量值表述比较容易，如：
- 独立重复投掷分币 5 次，正面朝上的次数；
- 无放回抽取产品，抽取到的次品个数；
- 数轴上的随机游动，第 i 步移动到的点 K，等等。
- 这些结果值有随机性（取值不能预先确定，但发生的可能性大小可以预估）。

随机变量定义

- **例 1.1**　100 件产品中有 5 件次品，95 件正品。随机抽取 20 件，“抽取出的次品件数”在抽取前是一个随机数值，可能在 0，1，..., 5 中取值。
- 但是，每次抽取完毕以后必有一个确定的抽取结果，此抽取结果有一个确定的次品件数。
- **定义 1.1**　对于条件组 S 下的每一个可能结果 ω 都唯一地对应到一个实数值 $X(\omega)$, 则称实值变量 $X(\omega)$ 为一个随机变量，简记为 X。

- 随机变量实际是从结果到数值的一个函数（映射），以试验结果为自变量。
- **例 1.2** 盒中 5 个球，2 白 3 黑。从中随机抽取 3 个。“抽得的白球数”X 是一个随机变量。
- 把 5 个球编号，1，2，3 号为黑球，4，5 号为白球。
- 所有结果有 $C_5^2 = 10$ 种，与 X 值的对应关系可列表（板书）。
- X 只能取 0，1，2。“$X = 0$ “，“$X = 1$ “，“$X = 2$ “都是随机事件。
- 由古典概型：

$$
\begin{aligned}
P(X = 0) =& \frac{C_3^3}{C_5^3} = \frac{1}{10} \\
P(X = 1) =& \frac{C_3^2 C_2^1}{C_5^3} = \frac{6}{10} \\
P(X = 2) =& \frac{C_3^1 C_2^2}{C_5^3} = \frac{3}{10}
\end{aligned}
$$

- **例 1.3** 单次射击击中概率 0.8，射击 30 次，“击中目标的次数”X 是随机变量。
- 可以取值 $0, 1, 2, \ldots, 30$。
- “$X = 0$”，“$X = 1$”，“$X = 2$”，$\cdots\cdots$，“$X = 30$”都是随机事件。
- **例 1.4** 单次射击击中概率 0.8，连续射击直到第一次击中未知，所需的“射击次数”X 是随机变量。X 可以取 $1, 2, \ldots$。
- “$X = k(k = 1, 2, 3, \ldots)$”是随机事件。
- **例 1.5** 出租车 400 辆。每天每辆出租车故障概率 0.02。
- 一天内有故障的出租车辆数 X 是随机变量，取值在 $0, 1, 2, \ldots, 400$ 中。
- **例 1.6** 某公共汽车站每隔 5 分钟有一辆汽车通过。
- 一位乘客对于汽车通过的规律完全不知情，所以在任一时刻到达车站都是等可能的。
- 其候车时间 X 是随机变量，取值 $0 \leq X < 5$。

- “$X > 2$”, “$X \leq 3$”都是随机事件。
- 这里 X 的取值范围是一个区间内的所有实数值，取值是“连续的”。
- **例 1.7** 一门大炮瞄准某个地面目标射击，以目标为坐标原点建立坐标系，y 轴指向从大炮到目标的方向。
- 弹着点与目标的距离 ρ 是一个随机变量，$\rho > 0$。
- 弹着点的直角坐标 (X, Y) 的两个分量 X, Y 是随机变量。
- X, Y 连续取值，还可以取负值。
- 随机变量是重要的概率模型。
- 实际中的随机变量：工业生产中随机抽取的一件产品的质量指标（强度、硬度、光洁度、粘合力、纤度等），医学检验的测量值，问卷调查汇总的答卷选择比例，等等。
- 随机变量按取值范围分为两类：取有限个可能取值或可数个可能取值的，叫做**离散型**随机变量；可以在一个区间或若干个区间取值的，叫做**连续型**随机变量。虽然除此之外还可以有其它类型但本课程不考虑。

2.2 离散型随机变量

2.2.1 概率分布

概率分布

- 离散型随机变量取值范围是有限个值或可数个值。
- 设 X 可取的值为 $x_1, x_2, \ldots, x_k, \ldots$。
- 随机变量作为古典概型、独立试验序列概型等的推广，关键是它能用取值作为事件，并可以计算取值概率。
- 离散型随机变量 X 的每个取值的概率可列表如下

X	x_1	x_2	$\cdots$	x_k	$\cdots$
p	p_1	p_2	$\cdots$	p_k	$\cdots$

称为 X 的概率分布表。

- 概率分布表可简写为：

$$P(X = x_i) = p_i, \quad i = 1, 2, \dots \tag{2.1}$$

这是一个函数，自变量取值于 $x_i, i = 1, 2, \dots$，函数值取值在 $p_1, p_2, \dots$ 中。

- (2.1) 称为 X 的概率分布 (probability distribution), 或概率质量函数 (probability mass function，简记为 PMF)。

- 概率分布的直观理解：X 有一系列不同取值，在每个取值上的概率大小，总数为 1。

- 性质：

$$(1) p_k \geq 0 \quad (k = 1, 2, \dots)$$
$$(2) \sum_k p_k = 1$$

- 事件组 $\{X = x_k\}, k = 1, 2, \dots$ 构成完备事件组。

- **例 1.2（续）** X 为 3 黑 2 白共 5 个球中任取 3 个，结果白球个数。

- 其概率分布表为

X	0	1	2
p	0.1	0.6	0.3

- 概率质量函数（PMF）为

$$P(X = 0) = 0.1$$
$$P(X = 1) = 0.6$$
$$P(X = 2) = 0.3$$

2.2.2 两点分布

两点分布

- 设 $0 < p < 1$, 设随机变量 X 只能取 1，0，其分布为

$$\begin{aligned} P(X = 1) =& p \\ P(X = 0) =& 1 - p \end{aligned}$$

  称 X 服从**两点分布**，p 称为其**分布参数**。也叫做伯努利分布 (Bernoulli)。可记作 $\mathrm{b}(1, p)$ 分布。

- **例 2.1** 100 件产品中，95 正品，5 次品。随机抽取一件。

$$X = \begin{cases} 1 & \text{当取得正品} \\ 0 & \text{当取得次品} \end{cases}$$

- 则

$$\begin{aligned} P(X = 1) =& 0.95 \\ P(X = 0) =& 1 - 0.95 \end{aligned}$$

  $X \sim \mathrm{b}(1, p)$。

- 两点分布仅适用于描述只有两个结果的情况。

2.2.3 二项分布

二项分布

- 考虑独立试验序列概型问题。设每次试验只有“成功”和“失败”两种可能结果，成功概率为 $p(0 < p < 1)$, $q = 1 - p$，独立重复试验 n 次。
- 令 X 表示 n 次试验中成功的次数。则 X 的取值范围为 $\{0, 1, 2, \ldots, n\}$。
- 在 §1.7 中已经证明

$$P(X = k) = C_n^k p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n \tag{2.3}$$

- 如果随机变量 X 的分布为 (2.3), 则称 X 服从二项分布（参数为 n, p），记为 $X \sim \mathrm{B}(n, p)$。

- 二项分布所有概率之和等于 1：

$$1 = (p+q)^n = \sum_{k=0}^{n} C_n^k p^k q^{n-k} \quad \text{(二项式定理)}$$

- 两点分布就是 $n = 1$ 时的二项分布。

2.2.4 泊松分布

泊松分布

- 设 $\lambda > 0$, 若 X 分布为

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (X = 0, 1, 2, \dots) \tag{2.4}$$

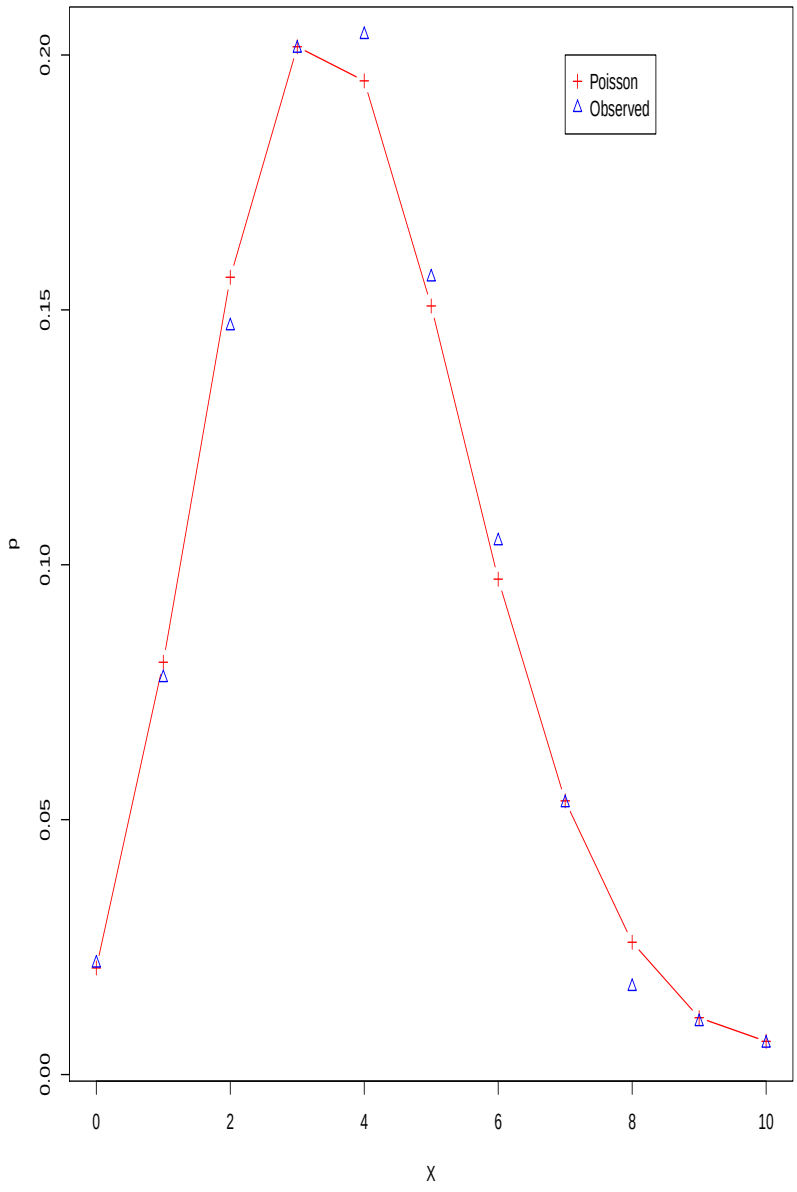
  则称 X 服从泊松分布 (参数为 λ)，记为 $X \sim \text{Poisson}(\lambda)$。

- **例 2.2** 放射性物质在某一段时间放射的粒子数 X 服从泊松分布。

- 每次观察 7.5 秒，共观察 2608 次。设 X 表示每次观察记录下的放射粒子个数。

- 观测结果的频率和 Poisson(3.87) 的概率很接近。

X	频数	频率	p_k
0	57	0.022	0.021
1	203	0.078	0.081
2	383	0.147	0.156
3	525	0.201	0.202
4	532	0.204	0.195
5	408	0.156	0.151
6	273	0.105	0.097
7	139	0.053	0.054
8	45	0.017	0.026
9	27	0.010	0.011
≥ 10	16	0.006	0.007
总计	2608	1.000	1.000



- 泊松分布还用于：
- 生物学、医学、工业、排队等问题。

- 如：容器内细菌数；铸件或布匹的疵点数；交换台的接入电话数；某路口经过的汽车数。
- 放射粒子数为何服从泊松分布？
- 体积为 V 的一块分割为 n 份相同体积 $\Delta V = \frac{V}{n}$ 的小块，假定：
- （1）每个特定的小块在 7.5 秒内放出两个以上 α 粒子的概率为 0（实际是很小到可忽略）；
- 小块放出一个 α 例子的概率为

$$p_n = \mu \Delta V = \mu \frac{V}{n}$$

- （2）各小块放出粒子与否相互独立。
- 则 7.5 秒内体积 V 的大块放射出 k 个粒子，可近似看作在 n 个独立的小块中共有 k 个小块放射出例子：

$$P(X = k) \approx C_n^k p_n^k q_n^{n-k} \quad (q_n = 1 - p_n)$$

- 令 $n \to \infty$ 可以逼近概率精确值：

$$P(X = k) = \lim_{n \to \infty} C_n^k p_n^k q_n^{n-k}$$

- 记 $\lambda = \mu V$, 则 $p_n = \frac{\lambda}{n}$。

$$\begin{aligned} C_n^k p_n^k q_n^{n-k} =& \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ =& \frac{1}{k!} \cdot \frac{n(n-1)\dots(n-k+1)}{n^k} \cdot \lambda^k \cdot \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \end{aligned}$$

- $n \to \infty$ 时, k 不变，第二个因子

$$\frac{n(n-1)\dots(n-k+1)}{n^k} = \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-(k-1)}{n} \to 1$$

- 第四个因子中

$$\begin{aligned} \left(1 - \frac{\lambda}{n}\right)^k \to& 1 \\ \left(1 - \frac{\lambda}{n}\right)^n =& \left[\left(1 - \frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\lambda} \to e^{-\lambda} \end{aligned}$$

- 于是

$$C_n^k p_n^k q_n^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad 当 n \to \infty \tag{2.5}$$

- 即

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

- 具体问题如果能符合类似于（1），（2）的条件，也可能会服从泊松分布。

用泊松分布近似二项分布

- **推论** 对二项分布 $\mathrm{B}(n,p)$，如果 $np \to \lambda > 0 (n \to \infty)$, 则

$$C_n^k p_n^k q_n^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad 当 n \to \infty$$

  即 p 很小而 n 较大时可以用泊松分布近似计算二项概率。

- $\sum_{k=0}^{\infty} p_k = 1$?
- 用指数函数的泰勒展开：

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

- 两边乘以 $e^{-\lambda}$ 即可。

2.2.5 超几何分布

超几何分布

- 设有 N 个同类产品，其中 M 个次品。从中任取 n 个（假定 $n \le N-M$）。则这 n 个中的次品数 X 是离散型随机变量，由第一章例 2.5

$$P(X=m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} \quad (m = 0, 1, 2, \dots, \min(M, n))$$

- 称 X 服从超几何分布（N, M, n 为参数）。

超几何分布与二项分布的关系

- 超几何分布是无放回抽样结果；二项分布可以看成有放回抽样结果。
- 当产品总数 N 很大时，两者分布近似相等。
- 设 $N \to \infty$ 时 $M/N \to p(n, m$ 不变$)$, 则

$$\frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = C_n^m p^m q^{n-m} \quad (N \to \infty)$$

证明

$$\begin{aligned}
&\frac{C_M^m C_{N-M}^{n-m}}{C_N^n} \\
=&\frac{M!}{(M-m)!m!} \cdot \frac{(N-M)!}{[N-M-(n-m)]!(n-m)!} \cdot \frac{(N-n)!n!}{N!} \\
=&\frac{n!}{m!(n-m)!} \cdot \left(\frac{M \cdot (M-1) \dots [M-(m-1)]}{N^m} \right) \cdot \\
&\left(\frac{[(N-M)] \cdot [(N-M)-1] \dots [(N-M)-(n-m-1)]}{N^{n-m}} \right) \cdot \\
&\left(\frac{N^n}{N \cdot (N-1) \cdots [N-(n-1)]} \right)
\end{aligned}$$

- 第二个因子趋于 p^m; 第三个因子趋于 $(1-p)^{n-m}$; 第四个因子趋于 1。

2.3 连续型随机变量

2.3.1 概率密度函数

连续型随机变量与概率密度函数

- 例：弹着点与目标间距离；等车时间。
- 随机变量取值不确定，但取每个值的可能性大小可以确定。

- 随机变量的优势：可以更准确地描述事件结果；可以更准确地描述事件概率（概率分布）。
- 一般地考虑 $\{a < X < b\}$ 这样的事件，对连续型随机变量不考虑 $\{X = a\}$ 这样的事件。
- **定义 3.1** 对于随机变量 X，如果存在非负可积函数 $p(x)(-\infty < x < \infty)$，使对任意 $a, b(a < b)$ 都有

$$P(a < X < b) = \int_a^b p(x)dx \tag{3.1}$$

则称 X 为**连续型随机变量**；称 $p(x)$ 为 X 的概率密度函数（probability density funcition, PDF），简称概率密度或密度。

概率密度的解释

- 若 $p(x)$ 在 x_0 处连续，则对很小的 δ,

$$P(X \in (x_0 - \delta, x_0 + \delta)) = \int_{x_0-\delta}^{x_0+\delta} p(x)dx \approx 2\delta p(x_0)$$

即 $p(x_0)$ 的大小代表了 X 在 x_0 附近取值的概率大小的一个比例。

- 图示。
- 概率本身还与邻域大小 2δ 有关，所以
- 概率密度不是概率！
- 概率密度非负，但不需要小于 1。

连续型随机变量单点概率为零

- 连续型随机变量至少在一个区间内可以取到任意实数值，所以取每个值的概率应该等于零。

- 对正整数 n,

$$\begin{aligned}\{X=a\} \subset & \left\{a-\frac{1}{n}<X<a+\frac{1}{n}\right\} \\ P(X=a) \leq & P\left(a-\frac{1}{n}<X<a+\frac{1}{n}\right) \\ = & \int_{a-\frac{1}{n}}^{a+\frac{1}{n}} p(x)dx \to 0 \quad (n \to \infty)\end{aligned}$$

所以

$$P(X=a) =0$$

概率密度函数积分等于 1

- 按定积分定义

$$\int_{-\infty}^{\infty} p(x)\,dx = \lim_{a\to\infty} \int_{-a}^{a} p(x)\,dx$$

- 对正整数 n, 事件 $A_n = \{-n < X < n\}$ 构成单调递增事件列：

$$\begin{aligned}A_n \subset & A_{n+1}, n=1,2,\ldots \\ \bigcup_{n=1}^{\infty} A_n = & \{-\infty < X < \infty\} \\ & P(-\infty < X < \infty) = 1\end{aligned}$$

由概率的单调极限性质

$$\begin{aligned}1 = & P(-\infty < X < \infty) = P\left(\bigcup_{n=1}^{\infty} A_n\right) \\ = & \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \int_{-n}^{n} p(x)dx = \int_{-\infty}^{\infty} p(x)dx\end{aligned}$$

概率密度性质

- 概率密度为非负可积函数，在 $(-\infty,\infty)$ 积分等于 1。

- 实际中的非离散型的随机变量一般是连续型的，而且 $p(x)$ 至多有有限多个间断点，在其它地方连续。
- 概率密度改变单个点的数值仍为 X 的密度。
- 但是若 $p_1(x), p_2(x)$ 都是 X 的密度且都在 x_0 连续，则 $p_1(x_0) = p_2(x_0)$。

2.3.2 均匀分布

均匀分布

- 若 X 有概率密度

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{当} x \in [a,b] \\ 0 & \text{其它} \end{cases}$$

则称 X 服从 $[a,b]$ 区间上的**均匀分布** (uniform distribution)，记为 $X \sim \mathrm{U}[a,b]$。

- 对 $a \le c < d \le b$,

$$P(c < X < d) = \int_c^d p(x) = \frac{d-c}{b-a}$$

- X 取值于 $[a,b]$ 中任一区间的概率与该区间长度成正比。
- 概率密度是常数，说明 X 取 $[a,b]$ 内任何一点附近的值的可能性大小都是相同的，所以叫“均匀”分布。
- 例：§2.1 例 1.6 等车时间服从均匀分布。

2.3.3 指数分布

指数分布

- 若 X 有概率密度

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{当} x \ge 0 \\ 0 & \text{当} x < 0 \end{cases} \quad (\lambda > 0)$$

则称 X 服从**指数分布** (exponential distribution)（参数为 λ），记作 $X \sim \mathrm{E}(\lambda)$。

- 对 $0 \le a < b$ 有

$$\begin{aligned} P(a<X<b) =& \lambda \int_a^b e^{-\lambda x} dx = \int_{\lambda a}^{\lambda b} e^{-t} dt \\ =& e^{-\lambda a} - e^{-\lambda b} \\ P(X>a) =& \lambda \int_a^{\infty} e^{-\lambda x} dx = e^{-\lambda a} \\ P(X<b) =& \lambda \int_0^b e^{-\lambda x} dx = 1 - e^{-\lambda b} \\ \int_{-\infty}^{\infty} p(x) dx =& \int_0^{\infty} p(x) dx = 1 \end{aligned}$$

2.3.4 正态分布

正态分布

- 若 X 有概率密度

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \quad (-\infty < x < \infty) \qquad (3.4)$$
$$(-\infty < \mu < \infty, \sigma > 0)$$

则称 X 服从正态分布 $\mathrm{N}(\mu, \sigma^2)$（normal distribution, 或 Gaussian distribution）, 简记为 $X \sim \mathrm{N}(\mu, \sigma^2)$。

- 概率密度图形演示。μ, σ^2 变化时曲线的变化。
- $p(x)$ 曲线呈钟形，最大值点在 $x = \mu$，关于 $x = \mu$ 轴对称；
- 在 $x = \mu \pm \sigma$ 处有拐点（二阶导数等于零的点，是曲线由凹变凸或由凸变凹的点）。
- 当 $x \to \pm\infty$ 时曲线以 x 轴为渐近线。
- μ 决定曲线中心位置；σ 越大，曲线越平缓，σ 越小，曲线越陡峭。

正态分布密度性质

- N(0,1) 叫做标准正态分布，其分布密度为

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

- 用微积分的二重积分极坐标变换可以证明 $\int_{-\infty}^{\infty}\phi(x)dx=1$。

$$I=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{-\frac{x^2+y^2}{2}}dxdy=\left(\int_{-\infty}^{\infty}e^{-\frac{x^2}{2}}dx\right)^2$$

作极坐标变换，令 $x=r\cos\varphi, y=r\sin\varphi$,

$$\begin{vmatrix}\frac{dx}{dr} & \frac{dy}{dr}\\ \frac{dx}{d\varphi} & \frac{dy}{d\varphi}\end{vmatrix}=\begin{vmatrix}\cos\varphi & \sin\varphi\\ -r\sin\varphi & r\cos\varphi\end{vmatrix}=r$$

$$I=\int_0^{\infty}\int_0^{2\pi}e^{-\frac{r^2}{2}}rdrd\varphi=2\pi$$

- 一般的 $\mathrm{N}(\mu,\sigma^2)$ 密度可以写成

$$\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$

- 由 $\int_{-\infty}^{\infty}\phi(x)dx=1$, 作变换 $\frac{x-\mu}{\sigma}=t$ 可以证明 $\mathrm{N}(\mu,\sigma^2)$ 密度积分等于 1。

- 测量误差和许多质量指标，如一批产品的长度、强度等，可以看作或近似看作服从正态分布。

- 正态分布在实际中最为常用，以至于一些不服从正态分布的数据也用正态分布来处理，这也是不合适的。

正态随机变量在一个区间的取值概率

- 设 $X\sim\mathrm{N}(0,1)$, $a<b$。记

$$\Phi(x)=\int_{-\infty}^{x}\phi(t)dt$$

则

$$P(a<X<b)=\int_a^b \phi(t)dt=\Phi(b)-\Phi(a)$$

(演示)

- 由 $\phi(x)$ 的对称性

$$\Phi(-x)=1-\Phi(x),\quad x\in(-\infty,\infty)$$

(图示)

- 函数 $\Phi(x)(x>0)$ 已制成表格，见 P.431 附表 1。
- 现在一般用统计软件计算，在 R 软件中为 `pnorm(x)`。
- 如：

$$\begin{aligned}P(1<X<2)=&\Phi(2)-\Phi(1)=0.9773-0.8413=0.1360\\ P(-1<X<1)=&\Phi(1)-\Phi(-1)=\Phi(1)-[1-\Phi(1)]\\ =&2\Phi(1)-1=2\times 0.8413-1=0.6826\end{aligned}$$

- 对 $N(\mu,\sigma^2)$,

$$\begin{aligned}P(a<X<b)=&\int_a^b \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx\\ =&\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}}\phi(t)dt\quad (\text{令}t=\frac{x-\mu}{\sigma})\\ =&\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\\ P(X<b)=&\Phi\left(\frac{b-\mu}{\sigma}\right)\\ P(X>a)=&1-\Phi\left(\frac{a-\mu}{\sigma}\right)\end{aligned}$$

- **例 3.3** 设 $X\sim \mathrm{N}(2,0.3^2)$, 求 $P(X>2.4)$。
- **解**

$$P(X>2.4)=1-\Phi\left(\frac{2.4-2}{0.3}\right)=1-\Phi(1.33)$$

- 附表中 $\Phi(1.32) = 0.90658, \Phi(1.35) = 0.91149$。
- 线性插值公式：

$$\begin{aligned} f(x) \approx & f(a) + \frac{f(b) - f(a)}{b - a} \cdot (x - a) \\ \Phi(1.33) \approx & 0.90658 + \frac{0.91149 - 0.90658}{1.35 - 1.32} \cdot (1.33 - 1.32) \\ = & 0.9082167 \end{aligned}$$

- $P(X > 2.4) = 1 - 0.9082167 = 0.0918$。
- 用 R 软件直接计算：`1 - pnorm(2.4, 2, 0.3)`, 结果为 0.09121122。

正态分布的经验规则

- 对 $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{aligned} P(X \in (\mu - \sigma, \mu + \sigma)) = & 0.6827 \\ P(X \in (\mu - 2\sigma, \mu + 2\sigma)) = & 0.9545 \\ P(X \in (\mu - 3\sigma, \mu + 3\sigma)) = & 0.9973 \end{aligned}$$

- 基本在 $(\mu - 2\sigma, \mu + 2\sigma)$ 内取值（超过 95%）。
- 几乎不在 $(\mu - 3\sigma, \mu + 3\sigma)$ 之外取值（不到千分之三）。

2.3.5 伽玛分布

伽玛分布

- 若 X 有概率密度

$$p(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \le 0 \end{cases} (\alpha > 0, \beta > 0) \tag{3.6}$$

则称 X 服从伽玛分布 (gamma distribution)，简记为 $X \sim \Gamma(\alpha, \beta)$。

- 其中

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

- $\Gamma(1)=1,\Gamma(\frac{1}{2})=\sqrt{\pi}$。
- $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)(\alpha>0)$；$n!=\Gamma(n+1)$。
- $\Gamma(1,\beta)$ 是指数分布 $\mathrm{Exp}(\beta)$。
- $\Gamma(\alpha,1)$ 叫做标准伽玛分布。
- $\Gamma(\frac{n}{2},\frac{1}{2})$ 为 $\chi^2(n)$ 分布。
- 密度演示。

2.3.6 威布尔分布

威布尔分布

- 若 X 有概率密度

$$p(x)=\begin{cases}m\frac{x^{m-1}}{\eta^m}\exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}, & x>0\\ 0, & x\le 0\end{cases}$$

  则称 X 服从威布尔分布 (Weibull distribution), 简记为 $X\sim\mathrm{W}(m,\eta)$, 其中 $m>0,\eta>0$, m 称为形状参数 (shape parameter)，η 称为尺度参数 (scale parameter)。

- 许多机电产品的寿命服从威布尔分布。
- $\mathrm{W}(1,\eta)$ 是参数为 $\frac{1}{\eta}$ 的指数分布。
- 威布尔分布和指数分布在工业产品的寿命与可靠性研究中有广泛应用。
- 密度演示。

2.4 分布函数与随机变量函数的分布

2.4.1 分布函数

分布函数

- 分布密度不利于直接计算概率。比如计算连续型随机变量 X 的概率 $P(a < X < b)$ 需要对密度积分。
- 引入“分布函数”。
- **定义 4.1** 设 X 是一个随机变量，称函数

$$F(x) = P(X \le x) \quad (-\infty < x < \infty) \tag{4.1}$$

  为 X 的分布函数 (distribution function, 或 cumulative distribution function, CDF)。
- 任何一个随机变量都有分布函数。

分布函数的性质

(1) $0 \le F(x) \le 1$ $(-\infty < x < \infty)$;

(2) $F(x)$ 是 x 的单调递增函数;

(3) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$;

(4) $F(x)$ 是 x 右连续函数;

(5) $P(a < X \le b) = F(b) - F(a)$。

(6) 若 X 为连续型，则 $P(a < X < b) = F(b) - F(a)$。

- 性质 (1), (2) 由定义和 $A \subset B \Longrightarrow P(A) \le P(B)$ 可得。
- 性质 (3), (4) 需要概率的公理化定义中的性质：完全可加性，单调事件的概率极限。

- 性质 (5):

$$\{X \le b\} = \{X \le a\} \cup \{a < X \le b\}$$

且不相容所以

$$\begin{aligned} P(X \le b) =& P(X \le a) + P(a < X \le b) \\ F(b) =& F(a) + P(a < X \le b) \end{aligned}$$

- 随机变量 X 的分布函数可记作 $F_X(x)$, 随机变量 Y 的分布函数可记作 $F_Y(x)$ 或 $F_Y(y)$ (注意函数的自变量符号的选用不影响函数本身)。

- **例 4.1** 设 $X \sim \mathrm{b}(1,p)$，$q = 1-p$, 则

$$F(x) = \begin{cases} 0 & x < 0 \\ q & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

- 作图并说明 $F(x)$ 的分段表达式。

连续型随机变量的分布函数

- 设 X 是连续型随机变量，有密度 $p(x)$, 分布函数 $F(x)$，则

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(t)dt \tag{4.2}$$

- $F(x)$ 是 $p(x)$ 的可变上限的定积分，也是 $p(x)$ 的一个原函数。

- (1) $F(x)$ 是 $x \in (-\infty, \infty)$ 的连续函数;

- (2) 对于 $p(x)$ 的连续点 x_0 而言有

$$F'(x_0) = p(x_0)$$

(注意密度函数可以修改单个点的函数值而仍为原随机变量密度)

- 若 $p(x)$ 只有至多有限个间断点，则对非间断点的 x

$$p(x) = F'(x) \tag{4.3}$$

- **例 4.2** 已使用了 t 小时的电子管在以后的 Δt 小时内损坏概率为 $\lambda\Delta t + o(\Delta t)$, $\lambda > 0$ 不依赖于 t。
- 电子管寿命为 0 的概率是 0。求电子管在 T 小时内损坏的概率。
- 令 X 为电子管的寿命，对于成批电子管 X 是一个随机变量，设分布函数为 $F(x)$。求 $P(X \leq T)$。
- “已使用了 t 小时的电子管在以后的 Δt 小时内损坏的概率”：

$$P(t < X \leq t + \Delta t | X > t) = \lambda\Delta t + o(\Delta t)$$

- 左边等于

$$\frac{P(t < X \leq t + \Delta t)}{P(X > t)} = \frac{F(t + \Delta t) - F(t)}{1 - F(t)}$$

- 于是

$$\begin{aligned}
\frac{F(t + \Delta t) - F(t)}{1 - F(t)} =& \lambda\Delta t + o(\Delta t) \\
\frac{F(t + \Delta t) - F(t)}{\Delta t} =& [1 - F(t)] \left[\lambda + \frac{o(\Delta t)}{\Delta t}\right]
\end{aligned}$$

- 令 $\Delta t \to 0$

$$\begin{aligned}
F'(t) =& \lambda[1 - F(t)] \\
\frac{F'(t)}{1 - F(t)} =& \lambda \\
\frac{d}{dt}\log[1 - F(t)] =& -\lambda \\
F(t) =& 1 - c_1 e^{-\lambda t}
\end{aligned}$$

- 联系 $F(0) = 0$ 有

$$\begin{aligned}
F(t) =& 1 - e^{-\lambda t}, \quad (t > 0) \\
F(t) =& 0, \quad (t \leq 0)
\end{aligned}$$

- 电子管子 T 小时内损坏的概率为

$$P(X \leq T) = F(T) = 1 - e^{-\lambda T}$$

- 分布密度为

$$p(t) = F'(t) = \begin{cases} \lambda e^{-\lambda t} & \text{当} t > 0 \\ 0 & \text{当} t \leq 0 \end{cases}$$

- 即 X 服从参数为 λ 的指数分布。

2.4.2 随机变量函数的分布

随机变量函数

- 设 $f(x)$ 是一个函数，随机变量 $f(X)$ 是随机变量 Y, 当 $X = x$ 时，Y 取值 $y = f(x)$。记作 Y = f(X)
- 相当于复合函数 $Y = f(X) = f(X(\omega))$。
- 如：设 X 是分子的速率，Y 是分子的动能，则

$$Y = \frac{1}{2}mX^2 \quad (m \text{ 为分子的质量})$$

- 已知 X 的分布后如何确定 $Y = f(X)$ 的分布？

离散型随机变量的函数

- 若 X 是离散型随机变量，取值 $x_1, x_2, \ldots, x_k, \ldots$, $P(x_k) = p_k (k = 1, 2, \ldots)$,
- 则 $Y = f(X)$ 取值 $f(x_1), f(x_2), \ldots, f(x_k), \ldots$, 若各 $f(x_k), k = 1, 2, \ldots$ 互不相同则

$$P(Y = f(x_k)) = p_k \qquad (*)$$

(*) 即为 Y 的概率分布。

- 若 $f(x_k), k = 1, 2, \ldots$ 有重复值，设所有互不相等的值为 $y_1, y_2, \ldots$, 则

$$P(Y = y_k) = \sum_{f(x_j)=y_k} p_j, \quad k = 1, 2, \ldots$$

- **例 4.3** X 概率分布为

X	0	1	2	3	4	5
$P(X = x_i)$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{2}{9}$	$\frac{1}{9}$

- 则 $Y = 2X + 1$ 的概率分布为:

Y	1	3	5	7	9	11
$P(Y = y_i)$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{2}{9}$	$\frac{1}{9}$

- **例 4.4** X 同例 4.3, 但 $Y = (X - 2)^2$。

- X 概率分布：

X	0	1	2	3	4	5
$P(X = x_i)$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{2}{9}$	$\frac{1}{9}$

- 对应到每个（可重复）$f(x_i)$ 的概率：

Y	4	1	0	1	4	9
$P(Y = f(x_i))$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{2}{9}$	$\frac{1}{9}$

- 合并：

Y	0	1	4	9
$P(Y = y_i)$	$\frac{1}{3}$	$\frac{1}{6} + \frac{1}{12}$	$\frac{1}{12} + \frac{2}{9}$	$\frac{1}{9}$

连续型随机变量函数

- **例 4.5** $X \sim \mathrm{N}(\mu, \sigma^2)$, $Y = \frac{X-\mu}{\sigma}$。

- 设 $Y \sim F_Y(y)$,

$$\begin{aligned} F_Y(y) =& P(Y \le y) \\ =& P(\frac{X - \mu}{\sigma} \le y) \\ =& P(X \le \mu + \sigma y) \\ =& F_X(\mu + \sigma y) \end{aligned}$$

- 求导得

$$p_Y(y) = p_X(\mu + \sigma y)\sigma$$

- 而

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

- 代入得

$$p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[(\mu+\sigma y)-\mu]^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

- 即 $Y \sim \mathrm{N}(0,1)$。

求连续型随机变量函数的密度的一般方法

- 证明用到：
- 若连续型随机变量的密度在除去几个点之外连续，则 $F'(x) = p(x)$ 对除去几个点后；
- **定理（习题七第 16 题）** 若 $F'(x)$ 连续，则 X 是连续型随机变量，$F'(x)$ 为密度。
- 以上通过计算 Y 的分布函数并用 X 的分布函数表示，然后求导得到密度函数的方法是一般性的。
- **例 4.6** $X \sim \mathrm{N}(\mu, \sigma^2)$, $Y = a + bX$(a, b 为常数，$b \neq 0$)。
- 设 $b > 0$,

$$\begin{aligned} F_Y(y) =& P(Y \le y) = P(a + bX \le y) \\ =& P\left(X \le \frac{y-a}{b}\right) = F_X\left(\frac{y-a}{b}\right) \end{aligned}$$

求导得

$$p_Y(y) = p_X(\frac{y-a}{b})\frac{1}{b}$$

- 当 $b<0$ 时，

$$\begin{aligned} F_Y(y) =& P(Y \le y) = P(a+bX \le y) \\ =& P\left(X \ge \frac{y-a}{b}\right) = 1 - F_X\left(\frac{y-a}{b}\right) \end{aligned}$$

求导得

$$p_Y(y) = -\, p_X(\frac{y-a}{b})\frac{1}{b}$$

- 于是

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

- 即 $Y \sim \mathrm{N}(a+b\mu, b^2\sigma^2)$。
- 正态分布经线性变换（斜率不为零）仍服从正态分布。
- 特别地，正态分布 $X \sim \mathrm{N}(\mu, \sigma^2)$ 可标准化：

$$Y = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$$

- 反之，若 $X \sim \mathrm{N}(0,1)$, $Y = \mu + \sigma X$, 则 $Y \sim \mathrm{N}(\mu, \sigma^2)$。
- **例 4.7** 圆片直径服从 $\mathrm{U}[5,6]$。求圆片面积概率分布。
- $Y = \frac{1}{4}\pi X^2$。易见 Y 取值范围是 $[\frac{\pi}{4}5^2, \frac{\pi}{4}6^2]$。
- 对于均匀分布 $\mathrm{U}[a,b]$, 其密度为

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{其它} \end{cases}$$

- 所以分布函数为

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

- 这里 $a=5, b=6$。当 $y \in [\frac{\pi}{4}5^2, \frac{\pi}{4}6^2]$ 时

$$\begin{aligned} F_Y(y) =& P(\frac{\pi}{4}X^2 \le y) = P(|X| \le \sqrt{\frac{4y}{\pi}}) \\ =& P(X \le \sqrt{\frac{4y}{\pi}}) = \sqrt{\frac{4y}{\pi}} - 5 \\ p_Y(y) =& F'(y) = \frac{1}{\sqrt{\pi y}} \end{aligned}$$

- 所以

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{\pi y}} & \text{当} \frac{25}{4}\pi \le y \le 9\pi \\ 0 & \text{其它} \end{cases}$$

- **定理 (习题七第 17 题)** 如果随机变量 X 的分布函数 $F(x)$ 满足以下条件：

- (1) $F(x)$ 连续；

- (2) 存在 $x_1 < x_2 < \cdots < x_n (n \ge 1)$, 在区间 $(-\infty, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n), (x_n, \infty)$ 上 $F'(x)$ 存在且连续。

- 令

$$f(x) = \begin{cases} F'(x) & \text{当 } F'(x) \text{ 存在时} \\ 0 & \text{当 } F'(x) \text{ 不存在时} \end{cases}$$

- 则 $f(x)$ 是 X 的分布密度。

- **例 4.8** 设 X 的密度函数为 $p_X(x)$, 仅有至多有限个不连续点，函数 $f(x)$ 导数 $f'(x)$ 连续且处处大于零。求 $Y = f(X)$ 的密度 $p_Y(y)$。

- **解** 这时 $f(x)$ 是严格单调递增连续函数，设其值域为 $(A, B)(-\infty \le A < B \le \infty)$, 反函数为 $g(y)$, 在 (A, B) 有定义。

- $g'(y)$ 存在：

$$g'(y) = \frac{1}{f'(g(y))}$$

- $\forall y \in (A, B)$,

$$\begin{aligned}\{f(X) \le y\} =& \{X \le g(y)\} \\ F_Y(y) =& P(Y \le y) = P(f(X) \le y) \\ =& P(X \le g(y)) = F_X(g(y)) \\ p_Y(y) =& F_Y'(y) = p_X(g(y))g'(y) = p_X(g(y))\frac{1}{f'(g(y))}\end{aligned}$$

- 对 $y \le A$, $F_Y(y) = P(Y \le A) = 0$, $p_Y(y) = F_Y'(y) = 0$;
- 对 $y \ge B$, $F_Y(y) = P(Y \le B) = 1$, $p_Y(y) = F_Y'(y) = 0$;
- 所以

$$F_Y(y) = \begin{cases} 0 & y \le A \\ F_X(g(y)) & y \in (A, B) \\ 1 & y \ge B \end{cases}$$

$$p_Y(y) = \begin{cases} p_X(g(y))\frac{1}{f'(g(y))} & y \in (A, B) \\ 0 & y \notin (A, B) \end{cases}$$

- 如果 $f(x)$ 是连续可导严格单调下降函数（$f'(x) < 0$），则推导中有一步改变：

$$P(f(X) \le y) = P(X \ge g(y)) = 1 - F_X(g(y))$$

- 于是

$$F_Y(y) = \begin{cases} 0 & y \le A \\ 1 - F_X(g(y)) & y \in (A, B) \\ 1 & y \ge B \end{cases}$$

$$p_Y(y) = \begin{cases} -p_X(g(y))\frac{1}{f'(g(y))} & y \in (A, B) \\ 0 & y \notin (A, B) \end{cases}$$

- 如果 $f(x)$ 不是连续可导且严格单调函数，则这样的方法不适用。
- **例 4.9** 设随机变量 $X \sim F(x)$ 且 $F(x)$ 是连续函数，则 $Y = F(X) \sim \mathrm{U}[0, 1]$。

- $F(x)$ 的值域为 $[0,1]$ 所以 Y 取值于 $[0,1]$。
- 对 $y \in (0,1)$，必有 x_0 使 $F(x_0) = y$。
- 这样的 x_0 不一定唯一，取

$$x_0 = \sup\{x : F(x) \le y\}$$

  应有 $x_0 \in (-\infty, \infty)$, 否则与 $\lim_{x\to\infty} F(x) = 1$ 矛盾；
- 且 $F(x_0) = y$,

$$F(x) \le y \Leftrightarrow x \le x_0$$

- 则

$$F_Y(y) = P(Y \le y) = P(X \le x_0) = F(x_0) = y$$

- 还要证明 $F_Y(y) = 0(y \le 0)$, $F_Y(y) = 1(y \ge 1)$。
- 第二条显然 (必然事件)；第一条只要证 $F_Y(0) = 0$。
- 因为 $F(x)$ 连续，设

$$a = \sup\{x : F(x) = 0\}$$

- 若 a 不存在 (当 $F(x)$ 处处为正时)，则 $\{F(X) \le 0\}$ 是不可能事件。
- 若 a 存在，由 $F(x)$ 连续知 $F(a) = 0$, 于是

$$\begin{aligned} F(x) \le 0 \Longleftrightarrow & x \le a \\ F_Y(0) = & P(F(X) \le 0) \\ = & P(X \le a) = F(a) = 0 \end{aligned}$$

- **例 4.10** 设函数 $F(x)$ 具有下列性质：
- (1) $0 \le F(x) \le 1$, $\forall x \in (-\infty, \infty)$;

- (2) $F(x)$ 是 x 的单调递增函数；
- (3) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$;
- (4) $F(x)$ 是右连续函数；
- 令

$$g(y) = \inf\{x : F(x) \geq y\} \quad (0 < y < 1) \qquad (*)$$

- 若 $U \sim \mathrm{U}(0,1)$, 则 $X = g(U) \sim F(x)$。
- (*) 保证 $\forall y \in (0,1)$

$$F(x) \geq y \Longleftrightarrow x \geq g(y)$$

- 于是

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

第三章　随机变量的数字特征

3.1　离散型随机变量的期望

随机变量的数字特征

- 随机变量的概率函数（PMF）、密度函数（PDF）、分布函数是对随机变量分布的最完整刻画。
- 但实际中分布难以描述和确定。
- 容易计算分布的一些数字特征：
  - 中心位置特征；
  - 分散程度特征。
- 期望是重要的概率论研究工具。

3.1.1　期望

期望

- **例**　设某种彩票发行了 10 万张，每张售 1 元，其中 10 张有奖，各奖励 5000 元，其它无奖。
- 问：买一张彩票，平均盈利多少？
- 设 X 为盈利的随机变量。则 $X = 10000 - 1$ 或 -1。

X	4999	-1
p	$\frac{10}{100000}$	$1 - \frac{10}{100000}$

- 平均盈利为

$$EX = 4999 \cdot \frac{10}{100000} - 1 \cdot (1 - \frac{10}{100000}) = -0.5(\text{元})$$

- 期望是一种加权平均，加权为每个值的取值概率。这与大量重复试验时的结果是一致的。
- 比如，彩票例子中，如果用 4999 与 -1 直接算术平均得 2499 元，显然是荒谬的。原因是取 4999 的概率远比取 -1 的概率小得多。

放射性粒子求平均粒子数

X	频数	频率	p_k
0	57	0.022	0.021
1	203	0.078	0.081
2	383	0.147	0.156
3	525	0.201	0.202
4	532	0.204	0.195
5	408	0.156	0.151
6	273	0.105	0.097
7	139	0.053	0.054
8	45	0.017	0.026
9	27	0.010	0.011
≥ 10	16	0.006	0.007
总计	2608	1.000	1.000

- 平均粒子数的合理公式：

$$\begin{aligned} &\frac{1}{2608}(0 \times 57 + 1 \times 203 + 2 \times 383 + \cdots + 10 \times 16) \\ =&\sum_k k\hat{p}_k \quad (\hat{p}_k\text{为取 k 的百分比}) \end{aligned}$$

期望定义

- **定义 1.1** 设离散型随机变量 X 的概率分布为

$$P(X = x_k) = p_k, k = 1, 2, \ldots$$

则称

$$\sum_k x_k p_k$$

为随机变量 X 的**期望**（或 **数学期望**），记作 EX 或 $E(X)$。

- EX 是随机变量取值的加权平均，按照概率的频率定义，这更符合多次重复观测后随机变量的平均表现。也叫做 X 的**均值**, 或 X 的分布的均值。

3.1.2 几个常用分布的期望

二点分布的期望

- X 分布

X	1	0
概率	p	1-p

-

$$EX = 1 \times p + 0 \times (1-p) = p$$

二项分布的期望

- $X \sim \mathrm{B}(n,p)$,

$$P(X=k) = C_n^k p^k q^{n-k}, \quad k = 0,1,\ldots,n$$

- 则

$$\begin{aligned}EX=&\sum_{k=0}^{n}kP(X=k)=\sum_{k=1}^{n}kC_n^kp^kq^{n-k}\\=&\sum_{k=1}^{n}\frac{kn!}{k!(n-k)!}p^kq^{n-k}\\=&\sum_{k=1}^{n}\frac{np\cdot(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}q^{(n-1)-(k-1)}\\=&np\sum_{k'=0}^{n-1}\frac{(n-1)!}{k'![(n-1)-k']!}p^{k'}q^{(n-1)-k'}\\=&np\sum_{k'=0}^{n-1}C_{n-1}^{k'}p^{k'}q^{(n-1)-k'}=np\end{aligned}$$

泊松分布的期望

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

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

- 则

$$\begin{aligned}EX=&\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}\\=&e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\\=&e^{-\lambda}\lambda\sum_{k'=0}^{\infty}\frac{\lambda^{k'}}{k'!}\\=&e^{-\lambda}\lambda e^{\lambda}=\lambda\end{aligned}$$

超几何分布的期望

- 设 X 服从参数为 N,M,n 的超几何分布 $(n\le N-M)$：

$$P(X=m)=\frac{C_M^mC_{N-M}^{n-m}}{C_N^n}\quad m=0,1,2,\min(M,n)$$

- 超几何分布期望

$$
\begin{aligned}
EX =& \sum_{m=0}^{\min(M,n)} m \cdot P(X=m) \\
=& \sum_{m=1}^{\min(M,n)} m \frac{M!}{m!(M-m)!} \\
& \cdot \frac{(N-M)!}{(n-m)![(N-M)-(n-m)]!} \cdot \frac{n!(N-n)!}{N!} \\
=& \sum_{m=1}^{\min(M,n)} \frac{M \cdot (M-1)!}{(m-1)![(M-1)-(m-1)]!} \\
& \cdot \frac{[(N-1)-(M-1)]!}{[(n-1)-(m-1)]![(N-1)-(M-1)-((n-1)-(m-1))]!} \\
& \cdot \frac{n \cdot (n-1)![(N-1)-(n-1)]!}{N \cdot (N-1)!} \\
=& \frac{nM}{N} \sum_{m'=0}^{\min(M-1,n-1)} \frac{C_{M-1}^{m'} C_{(N-1)-(M-1)}^{(n-1)-m'}}{C_{N-1}^{n-1}} \\
=& \frac{nM}{N}
\end{aligned}
$$

3.2 连续型随机变量的期望

连续型随机变量的期望

- **定义 2.1** 设连续型随机变量的密度为 $p(x)$，称

$$\int_{-\infty}^{\infty} xp(x)dx \tag{2.1}$$

为 X 的期望（或均值），记作 EX(或 $E(X)$)。

- 要求 (2.1) 的积分收敛才有期望。

- 解释：连续型随机变量的 $p(x)$ 是一个比例系数，不是概率。

- 把 X 的取值范围划分为区间 $(x_i, x_{i+1}], i = 1, 2, \ldots,$ 当 $p(x)$ 连续，每

个区间都很短时近似计算 X 的平均为

$$\begin{aligned} EX &\approx \sum_i x_i P(x_i < X \le x_{i+1}) \\ &\approx \sum_i \int_{x_i}^{x_{i+1}} xp(x)dx \\ &= \int_{-\infty}^{\infty} xp(x)dx \end{aligned}$$

均匀分布的期望

- $X \sim \mathrm{U}(a,b)$:

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{其它} \end{cases}$$

-

$$\begin{aligned} EX = \int_{-\infty}^{\infty} xp(x)dx &= \int_a^b x \cdot \frac{1}{b-a} dx \\ = \frac{1}{b-a} \left.\frac{x^2}{2}\right|_a^b &= \frac{1}{2} \cdot \frac{b^2-a^2}{b-a} \\ = \frac{a+b}{2} & \end{aligned}$$

- EX 是区间 $[a,b]$ 的中点。与“均匀”分布意义相符。

指数分布的期望

- $X \sim \mathrm{E}(\lambda)$:

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{其它} \end{cases} (\lambda > 0)$$

- 期望

$$\begin{aligned} EX =& \int_{-\infty}^{\infty} xp(x)dx = \lambda \int_{0}^{\infty} xe^{-\lambda x}dx \\ =& \frac{1}{\lambda}\int_{0}^{\infty} te^{-t}dt \quad (\text{令} t = \lambda x) \\ =& \frac{1}{\lambda}\left[(-te^{-t})\big|_{0}^{\infty} + \int_{0}^{\infty} e^{-t}dt\right] \\ =& \frac{1}{\lambda} \end{aligned}$$

- 所以 $\frac{1}{\lambda}$ 是指数分布的期望。

正态分布

- $X \sim \mathrm{N}(\mu, \sigma^2)$:

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

- 期望

$$\begin{aligned} EX =& \int_{-\infty}^{\infty} xp(x)dx \\ =& \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx \\ =& \sigma\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma} + \frac{\mu}{\sigma}\right)\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}d\frac{x-\mu}{\sigma} \\ =& \sigma\int_{-\infty}^{\infty} t \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \sigma \cdot \frac{\mu}{\sigma} \cdot \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\ =& \sigma \cdot 0 + \mu = \mu \end{aligned}$$

- 正态分布密度以 $x = \mu$ 为对称轴，且 μ 为均值。
- 其它分布如果也有对称轴 $x = c$ 且期望存在则其期望也等于 c。

伽玛分布

- $X \sim \Gamma(\alpha, \beta)$:

$$p(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases} \quad (\alpha > 0, \beta > 0)$$

- 期望

$$\begin{aligned} EX =& \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\beta x}\, dx \\ =& \frac{1}{\Gamma(\alpha)\beta} \int_0^\infty t^\alpha e^{-t} dt \quad (\text{令} t = \beta x) \\ =& \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\beta} = \frac{\alpha}{\beta} \end{aligned}$$

3.3 期望的简单性质及随机变量函数的期望公式

3.3.1 期望的简单性质

期望的简单性质

- 对常数 c, k, b 和随机变量 X, 有

$$\begin{aligned} &(1) \quad E(c) = c; \\ &(2) \quad E(kX) = kE(X); \\ &(3) \quad E(X+b) = E(X) + b; \\ &(4) \quad E(kX+b) = kE(X) + b. \end{aligned} \tag{3.1}$$

证明 (1)

- $E(c) = c$?

- 常数 c 可以看成离散随机变量，取 c 的概率为 1，按期望公式有

$$E(X) = c \times 1 = c$$

证明 (2)

- $E(kX) = kE(X)$?
- 当 $k = 0$ 时显然成立。
- 当 $k \neq 0$ 时，若 X 为离散分布：

$$P(X = x_i) = p_i, \quad i = 1, 2, \ldots$$

- 则 $Y = kX$ 的分布为

$$P(Y = kx_i) = p_i, \quad i = 1, 2, \ldots$$

- 按定义

$$E(Y) = \sum_i (kx_i)p_i = k\sum_i x_i p_i = kE(X)$$

- 若 $k \neq 0$，X 为连续型，密度 $p(x)$，则 $Y = kX$ 有密度

$$p_Y(y) = \frac{1}{|k|} p\left(\frac{y}{k}\right)$$

- 事实上, $k > 0$ 时 $\forall c < d$,

$$\begin{aligned}\int_c^d \frac{1}{k} p\left(\frac{y}{k}\right) dy = &\int_{\frac{c}{k}}^{\frac{d}{k}} p(t)\, dt \quad （令\ t = \frac{y}{k}）\\ =&P\left(\frac{c}{k} < X < \frac{d}{k}\right) = P(c < kX < d)\\ =&P(c < Y < d)\end{aligned}$$

  即 Y 的密度为 $\frac{1}{k} p\left(\frac{y}{k}\right)$。

- $k < 0$ 类似可证。

- 按期望定义

$$\begin{aligned}E(Y)=&\int_{-\infty}^{\infty}yp_Y(y)dy\\=&\frac{1}{|k|}\int_{-\infty}^{\infty}yp\left(\frac{y}{k}\right)dy\\=&\int_{-\infty}^{\infty}ktp(t)dt\quad(\text{令}t=\frac{y}{k})\\=&k\int_{-\infty}^{\infty}tp(t)dt\\=&kE(X)\end{aligned}$$

证明（3）

- $E(X+b)=E(X)+b$?
- 只对 $X\sim p(x)$ 证明。
- $Y=X+b\sim p(y-b)$。事实上，$\forall c<d$,

$$\begin{aligned}\int_c^d p(y-b)\,dy=&\int_{c-b}^{d-b}p(t)\,dt\quad(\text{令 }t=y-b)\\=&P(c-b<X<d-b)=P(c<X+b<d)\\=&P(c<Y<b).\end{aligned}$$

- 期望

$$\begin{aligned}E(X+b)=&E(Y)=\int_{-\infty}^{\infty}yp_Y(y)dy=\int_{-\infty}^{\infty}yp(y-b)dy\\=&\int_{-\infty}^{\infty}(t+b)p(t)dt\quad(\text{令}t=y-b)\\=&\int_{-\infty}^{\infty}tp(t)dt+b\int_{-\infty}^{\infty}p(t)dt\\=&E(X)+b\end{aligned}$$

证明（4）

- 由（2）、（3）得：

$$\begin{aligned} E(kX+b) =& E(kX)+b \quad \text{(性质 3)} \\ =& kE(X)+b \quad \text{(性质 2)} \end{aligned}$$

- 性质（4）包含了（2）和（3），叫做期望的**线性性质**。

3.3.2 随机变量函数的期望公式

随机变量函数的期望公式

- 对 $X \sim p(x)$, $Y = f(X)$,

$$E[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx \tag{3.2}$$

(要求右边绝对收敛)

- 若 $P(X = x_i) = p_i, i = 1, 2, \ldots,$

$$E[f(X)] = \sum_i f(x_i)p_i \tag{3.3}$$

(要求右边的级数绝对收敛)

- 这样的公式免去了求 $Y = f(X)$ 的分布的过程。
- **例 3.1** $X \sim \mathrm{N}(0, 1)$, 求 $E(X^2)$。
- 用两种办法。
- **解法 1** 用公式 (3.2)。

$$\begin{aligned} E(X^2) =& \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ =& -\int_{-\infty}^{\infty} x d\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) \\ =& -\left[x\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ =& 0 + 1 = 1 \end{aligned}$$

- （注意 $\lim_{x\to\pm\infty} xe^{-\frac{x^2}{2}} = 0$）
- **解法 2** 用分布函数法先求 $Y = f(X)$ 的分布密度再用期望定义。
- 显然

$$\begin{aligned} F_Y(y) =& P(X^2 \le y) = P(|X| \le \sqrt{y}) \\ =& P(-\sqrt{y} \le X \le \sqrt{y}) \\ =& F(\sqrt{y}) - F(-\sqrt{y}) \quad (y \ge 0) \end{aligned}$$

  ($Y < 0$ 是不可能事件)

- $Y = X^2$ 的密度

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

  (这是一个 Gamma$(\frac{1}{2}, \frac{1}{2})$ 分布密度)

- 于是

$$\begin{aligned} E(X^2) =& E(Y) = \int_0^\infty y p_Y(y) dy \\ =& \int_0^\infty \frac{\sqrt{y}}{\sqrt{2\pi}} e^{-\frac{y}{2}} dy \\ =& \int_0^\infty \frac{t}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} 2t dt \quad (t = \sqrt{y}, y = t^2) \\ =& -\int_0^\infty \frac{1}{\sqrt{2\pi}} 2t d(e^{-\frac{t^2}{2}}) \\ =& 0 + 2\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ =& \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1 \end{aligned}$$

- 利用了公式 (3.2) 的解法更简便。

- **例 3.2** $X \sim \mathrm{U}[0, 2\pi]$, 求 $E(\sin X)$。
- **解**

$$\begin{aligned} E(\sin X) &= \int_{-\infty}^{\infty} \sin(x) p_X(x) dx \\ &= \int_0^{2\pi} \sin(x) \frac{1}{2\pi} dx = 0 \end{aligned}$$

期望线性性质证明

- 期望的线性性质

$$E(kX + b) = kE(X) + b$$

是 $f(x) = kx + b$ 时公式 (3.2),(3.3) 的特例：

$$\begin{aligned} E(kX + b) =& \int_{-\infty}^{\infty} (kx + b) p(x) dx \\ =& k \int_{-\infty}^{\infty} x p(x) dx + b \int_{-\infty}^{\infty} p(x) dx \\ =& kE(X) + b \end{aligned}$$

3.4 方差及其简单性质

3.4.1 方差的概念

方差

- 随机变量的数字特征可以用数字刻画最重要的分布特征。
- 分布特征中最重要的特征是位置特征和分散程度（分布宽窄）特征。
- 方差（标准差）用来刻画分散程度。
- 例：甲乙两个女生小合唱队的身高：
- 甲队：1.60, 1.62, 1.59, 1.60, 1.59;

- 乙队：1.80, 1.60, 1.50, 1.50, 1.60.
- 平均都是 1.60，但是甲队很整齐，乙队站在一起很杂乱。
- 数据波动程度也是反映客观现象的一种指标。
- 如：产品的某种特性（如强度）波动大，说明生产不稳定。
- 如：生物的某种特性（如血压）波动大，表示病态。
- 所以对于数据，除了关心均值，还要研究其波动程度（分散程度）。

数据的方差

- 对于给定的一批数据 $x_1, x_2, \ldots, x_n$, 用

$$\frac{1}{n-1}[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2] \tag{4.1}$$

  来刻画这批数据的分散程度（波动），其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$。
- 当所有 x_i 为常数时，(4.1) 等于零。
- 各 x_i 差别越大，则与平均值差别越大，(4.1) 越大。

随机变量的方差

- 仿照数据的方差提出：
- **定义 4.1** 设离散型随机变量的概率分布为

$$P(X=x_k)=p_k,\ k=1,2,\ldots$$

  则称

$$\sum_k [x_k-E(X)]^2 p_k \tag{4.2}$$

  为 X 的**方差**，记作 $D(X)$ 或 $\mathrm{Var}(X)$。

- **定义 4.2** 设连续型随机变量 X 的密度是 $p(x)$，则称

$$\int_{-\infty}^{\infty}[x-E(X)]^2p(x)dx \tag{4.2'}$$

为 X 的方差，记作 $D(X)$ 或 $\mathrm{Var}(X)$。

- 为了方差存在，期望 $E(X)$ 必须先存在；
- 对离散型，(4.2) 的级数若非有限项，则级数收敛时方差才存在；
- 对连续型，(4.2') 的积分收敛时方差才存在；
- 方差总是非负的：

$$D(X)\geq 0$$

方差等价定义

- **方差等价定义**

$$D(X)=E[X-E(X)]^2 \tag{4.3}$$

按随机变量函数的期望公式，(4.3) 与 (4.2) 和 (4.2') 一致。

- 随机变量的方差，也称为其分布的方差：
- 注意：期望、方差只由分布决定，两个随机变量服从相同的分布（参数也要相同），则其期望、方差必定相同。

方差的恒等式

- 方差有如下恒等式

$$D(X)=E(X^2)-[E(X)]^2 \tag{4.4}$$

- **证明**

$$\begin{aligned} D(X) =& E[X - E(X)]^2 \\ =& E\left\{X^2 - 2E(X) \cdot X + [E(X)]^2\right\} \\ =& E(X^2) - E[2E(X) \cdot X] + E[E(X)]^2 \\ =& E(X^2) - 2E(X) \cdot E(X) + [E(X)]^2 \\ =& E(X^2) - [E(X)]^2 \end{aligned}$$

- 这个证明用到了第四章定理：

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

- 方差恒等式（4.4）的另一证明：以 $X \sim p(x)$ 为例，

$$\begin{aligned} D(X) =& \int_{-\infty}^{\infty} [x - E(X)]^2 p(x) dx \\ =& \int_{-\infty}^{\infty} \left\{x^2 - 2E(X) \cdot x + [E(X)]^2\right\} dx \\ =& \int_{-\infty}^{\infty} x^2 p(x) dx - 2E(X) \int_{-\infty}^{\infty} x p(x) dx \\ & + [E(X)]^2 \int_{-\infty}^{\infty} p(x) dx \\ =& E(X^2) - 2E(X) \cdot E(X) + [E(X)]^2 \cdot 1 \\ =& E(X^2) - [E(X)]^2 \end{aligned}$$

3.4.2 常用分布的方差

二点分布的方差

- 对二点分布

$$\begin{aligned} E(X) =& p \\ E(X^2) =& 1^2 \cdot p + 0^2 \cdot q = p \\ D(X) =& E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p) \end{aligned}$$

二项分布的方差

- 对二项分布

$$
\begin{aligned}
E(X) =& np \\
E(X^2) =& \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} \\
=& \sum_{k=1}^{n} [k(k-1)+k] \frac{n!}{k!(n-k)!} p^k q^{n-k} \\
=& \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!} p^k q^{n-k} + E(X) \\
=& n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)![(n-2)-(k-2)]!} \\
& \cdot p^{k-2} q^{(n-2)-(k-2)} + E(X) \\
=& n(n-1)p^2 \sum_{k'=0}^{n-2} C_{n-2}^{k'} p^{k'} q^{(n-2)-k'} + E(X) \\
=& n(n-1)p^2 + np = n^2 p^2 + np(1-p) \\
D(X) =& E(X^2) - [E(X)]^2 = np(1-p)
\end{aligned}
$$

泊松分布的方差

- 对泊松分布

$$\begin{aligned}
E(X) =& \lambda \\
E(X^2) =& \sum_{k=1}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \cdot \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
=& \sum_{k=1}^{\infty} [(k-1)+1] \cdot \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
=& \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
=& \lambda^2 \sum_{k'=0}^{\infty} \frac{\lambda^{k'}}{k'!} e^{-\lambda} + \lambda \sum_{k''=0}^{\infty} \frac{\lambda^{k''}}{k''!} e^{-\lambda} = \lambda^2 + \lambda \\
D(X) =& E(X^2) - [E(X)]^2 = \lambda
\end{aligned}$$

均匀分布的方差

- 若 $X \sim \mathrm{U}(a,b)$, 则

$$\begin{aligned}
E(X) =& \frac{a+b}{2} \\
E(X^2) =& \int_a^b x^2 \cdot \frac{1}{b-a} dx \\
=& \frac{1}{b-a} \left. \frac{x^3}{3} \right|_a^b \\
=& \frac{1}{3}(a^2 + ab + b^2) \\
D(X) =& E(X^2) - [E(X)]^2 \\
=& \frac{1}{3}(a^2 + ab + b^2) - \frac{1}{4}(a^2 + 2ab + b^2) \\
=& \frac{1}{12}(b-a)^2
\end{aligned}$$

指数分布的方差

- 设 $X \sim \mathrm{E}(\lambda)$, 则

$$\begin{aligned} E(X) =& \frac{1}{\lambda} \\ E(X^2) =& \int_0^\infty x^2 \cdot \lambda e^{-\lambda x} dx \\ =& \frac{1}{\lambda^2} \int_0^\infty t^2 e^{-t} dt \\ =& \frac{1}{\lambda^2} \Gamma(3) = \frac{1}{\lambda^2} 2! = \frac{2}{\lambda^2} \\ D(X) =& E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2} \end{aligned}$$

正态分布的方差

- 设 $X \sim \mathrm{N}(\mu, \sigma^2)$，则

$$\begin{aligned} D(X) =& E(X-\mu)^2 \\ =& \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\} dx \\ =& \int_{-\infty}^{\infty} \sigma^2 t^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\{\frac{1}{2} t^2\} dt \quad (\text{令} t = \frac{x-\mu}{\sigma}) \\ =& \sigma^2 \int_{-\infty}^{\infty} t^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt \\ =& \sigma^2 \quad (\text{见例 3.1}) \end{aligned}$$

- 所以正态分布的两个参数 μ, σ^2 分别是分布的期望和方差。

伽玛分布的方差

- 设 $X \sim \Gamma(\alpha, \beta)$:

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

- 则

$$\begin{aligned}
E(X) =& \frac{\alpha}{\beta} \\
E(X^2) =& \int_0^{\infty} x^2 \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\
=& \frac{1}{\beta^2} \int_0^{\infty} \frac{(\beta x)^{\alpha+1}}{\Gamma(\alpha)} e^{-(\beta x)} d(\beta x) \\
=& \frac{1}{\beta^2} \int_0^{\infty} \frac{t^{\alpha+1}}{\Gamma(\alpha)} e^{-t} dt \\
=& \frac{1}{\beta^2} \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{\beta^2} \\
D(X) =& \frac{(\alpha+1)\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}
\end{aligned}$$

3.4.3 方差的简单性质

方差的简单性质

- 当 k, b, c 为常数时

$$\begin{aligned}
&(1) \quad D(c) = 0; \\
&(2) \quad D(kX) = k^2 D(X); \\
&(3) \quad D(X+b) = D(X); \\
&(4) \quad D(kX+b) = k^2 D(X).
\end{aligned} \tag{4.5}$$

- **证明** （1）

$$\begin{aligned}
E(c) =& c \\
E(c^2) =& c^2 \\
D(X) =& c^2 - c^2 = 0
\end{aligned}$$

- (2)

$$\begin{aligned}
E(kX) =& kE(X) \\
E[(kX)^2] =& E(k^2 X^2) = k^2 E(X^2) \\
D(kX) =& k^2 E(X^2) - [kE(X)]^2 \\
=& k^2 \{E(X^2) - [E(X)]^2\} = k^2 D(X)
\end{aligned}$$

- （3）

$$\begin{aligned} E(X+b) =& E(X)+b \\ E(X+b)^2 =& E(X^2+2bX+b^2) = E(X^2)+2bE(X)+b^2 \\ D(X) =& E(X^2)+2bE(X)+b^2 \\ & -\{[E(X)]^2+2bE(X)+b^2\} \\ =& E(X^2)-[E(X)]^2 = D(X) \end{aligned}$$

- (4) 是（2）、（3）的推论:

$$D(kX+b) = D(kX) = k^2 D(X)$$

3.5 其它

3.5.1 切比雪夫不等式

切比雪夫不等式

- **定理 5.1** 设随机变量 X 存在均值 $E(X)$ 与方差 $D(X)$，则有

$$P\left(|X-E(X)| \geq \varepsilon\right) \leq \frac{D(X)}{\varepsilon^2} \quad (\varepsilon > 0) \tag{5.1}$$

- **证明** （仅对连续型情形）

$$\begin{aligned} D(X) =& \int_{-\infty}^{\infty} [x-E(X)]^2 p(x)dx \\ \geq& \int_{-\infty}^{E(X)-\varepsilon} [x-E(X)]^2 p(x)dx \\ & + \int_{E(X)+\varepsilon}^{\infty} [x-E(X)]^2 p(x)dx \\ \geq& \varepsilon^2 \int_{-\infty}^{E(X)-\varepsilon} p(x)dx + \varepsilon^2 \int_{E(X)+\varepsilon}^{\infty} p(x)dx \\ =& \varepsilon^2 P(X \leq E(X)-\varepsilon) + \varepsilon^2 P(X \geq E(X)+\varepsilon) \\ =& \varepsilon^2 P(|X-E(X)| \geq \varepsilon) \end{aligned}$$

- 得证。
- 在切比雪夫不等式中取 $\varepsilon = k\sqrt{D(X)}$, 则

$$P(|X - E(X)| \geq k\sqrt{D(X)}) \leq \frac{1}{k^2}$$

其中 $\sqrt{D(X)}$ 叫做 X 的标准差。

- 特别地，$k = 3$ 时

$$P(|X - E(X)| \geq 3\sqrt{D(X)}) \leq \frac{1}{9}$$

- 对比正态分布，这个比例是小于千分之三。
- 从切比雪夫不等式可知 $D(X)$ 越小则 X 取值远离其均值的概率越小，所以反映了分布的分散程度。

3.5.2 原点矩与中心矩

原点矩与中心矩

- 称

$$E(X^k) \quad (k = 1, 2, \dots)$$

为 X 的 k 阶**原点矩**，记为 ν_k;

- 称

$$E[X - E(X)]^k \quad (k = 1, 2, \dots)$$

为 X 的 k 阶**中心矩**, 记为 μ_k。

- 均值为 ν_1, 方差为 μ_2。

3.5.3 分位数与中位数

分位数与中位数

- 若 X 的分布函数 $F(x)$ 严格单调上升连续，则其存在反函数 $q(p), p \in (0,1)$, 使得

$$F(q(p)) = p, \quad \forall p \in (0,1)$$

称 $q(p)$ 为 X 的**分位数函数**，$x_p = q(p)$ 叫做 X 的 p **分位数**。

- 一般地，对 $p \in (0,1)$, 称 x_p 是随机变量 X 的 p 分位数，若

$$P(X < x_p) \le p \le P(X \le x_p) \tag{5.4}$$

- (5.4) 也写作

$$P(X \le x_p) \ge p, \quad P(X \ge x_p) \ge 1 - p \tag{5.4'}$$

- 二分之一分位数叫做中位数。
- 分位数必存在，不一定唯一。有一种唯一化的定义为

$$x_p = \inf\{x : F(X) \ge p\} \quad (\forall p \in (0,1))$$

例：二点分布的分位数

- 二点分布 b(1,p_0), 记 $q_0 = 1 - p_0$：

$$F(x) = \begin{cases} 0 & x < 0 \\ q_0 & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

- 分位数

$$x_p = \begin{cases} 0 & p \in (0, q_0) \\ [0,1] & p = q_0 \\ 1 & p \in (q_0, 1) \end{cases}$$

第四章　随机向量

4.1　随机向量的联合分布与边缘分布

随机向量

- 需要研究多个随机变量的情况，如：
- 弹着点横纵坐标 (X, Y);
- 炼钢厂每炉钢的硬度、含碳量、含硫量。
- 变量之间有联系。
- **定义 0.1** 称 n 个随机变量 $X_1, X_2, \ldots, X_n$ 的整体 $\boldsymbol{\xi} = (X_1, X_2, \ldots, X_n)$ 为 n 维随机向量。
- “维数”是分量的个数。比如，弹着点坐标 (X, Y) 是坐标平面的随机点。
- 着重讨论二维随机向量。

4.1.1　二维离散型随机向量

二维离散型随机向量

- **定义 1.1** 如果二维随机向量 $\boldsymbol{\xi} = (X, Y)$ 可能取的值只有有限个或可数个，则称 $\boldsymbol{\xi}$ 为离散型随机向量。
- 注意二维随机向量取值在平面（二维空间）中。

- 若 $\boldsymbol{\xi} = (X, Y)$ 是离散型，则两个分量 X, Y 都是离散型；反之亦然。
- 设 X 的取值范围是 $\{x_i, i = 1, 2, \dots\}$, Y 的取值范围是 $\{y_j, j = 1, 2, \dots\}$, 则 (X, Y) 的取值范围是 $\{(x_i, y_j) : i = 1, 2, \dots, j = 1, 2, \dots\}$ (其中有些组合可能是不可能事件)。
- 二维随机变量 $\boldsymbol{\xi} = (X, Y)$ 的概率分布：

$$P((X, Y) = (x_i, y_j)) = p_{ij}, \quad i = 1, 2, \dots, j = 1, 2, \dots \tag{1.1}$$

也称为 (X, Y) 的**联合分布**。

- 二维概率分布表也可以排列成

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

二维概率分布性质

- 性质：

$$(1) p_{ij} \geq 0 \quad (i = 1, 2, \dots, j = 1, 2, \dots)$$
$$(2) \sum_i \sum_j p_{ij} = 1$$

- (1) 显然。
- (2)：所有的 $\{(X, Y) = (x_i, y_j)\}$ 事件构成完备事件组，其并集为必然事件，由概率的完全可加性即得。

- **例 1.1** 设二维随机变量 (X,Y) 仅取 5 个不同点：

$$(1,1)\ (1.2,1)\ (1.4,1.5)\ (1,1.3)\ (0.9,1.2)$$

且取每个点的概率相等（$\frac{1}{5}$）。

- 联合分布为

$$\begin{aligned}P((X,Y)=(1,1))=&\frac{1}{5}\\P((X,Y)=(1,2.1))=&\frac{1}{5}\\P((X,Y)=(1.4,1.5))=&\frac{1}{5}\\P((X,Y)=(1,1.3))=&\frac{1}{5}\\P((X,Y)=(0.9,1.2))=&\frac{1}{5}\end{aligned}$$

- 概率分布表为

$X\setminus Y$	1	1.2	1.3	1.5
0.9	0	$\frac{1}{5}$	0	0
1	$\frac{1}{5}$	0	$\frac{1}{5}$	0
1.2	$\frac{1}{5}$	0	0	0
1.4	0	0	0	$\frac{1}{5}$

- **例 1.2** 设 (X,Y) 的联合分布为

$$\begin{aligned}&P((X,Y)=(k_1,k_2))\\=&\frac{n!}{k_1!k_2!(n-k_1-k_2)!}p_1^{k_1}p_2^{k_2}(1-p_1-p_2)^{n-k_1-k_2},\\&k_1=0,1,\ldots,n,k_2=0,1,\ldots,n,k_1+k_2\le n\end{aligned}\tag{1.3}$$

其中 n 是给定的正整数；$0<p_1<1$, $0<p_2<1$, $p_1+p_2<1$。

- 称为三项分布（参数 $n;p_1,p_2$）。

- 来验证三项分布的所有概率之和等于 1。

$$\begin{aligned}
&\sum_{k_1=0}^{n}\sum_{k_2=0}^{n-k_1} P((X,Y)=(k_1,k_2)) \\
=&\sum_{k_1=0}^{n}\sum_{k_2=0}^{n-k_1} \frac{n!}{k_1!k_2!(n-k_1-k_2)!}p_1^{k_1}p_2^{k_2}(1-p_1-p_2)^{n-k_1-k_2} \\
=&\sum_{k_1=0}^{n} \frac{n!}{k_1!(n-k_1)!}p_1^{k_1}\cdot \\
&\left[\sum_{k_2=0}^{n-k_1} \frac{(n-k_1)!}{k_2!((n-k_1)-k_2)!}p_2^{k_2}(1-p_1-p_2)^{(n-k_1)-k_2}\right] \\
=&\sum_{k_1=0}^{n} \frac{n!}{k_1!(n-k_1)!}p_1^{k_1}\cdot[p_2+(1-p_1-p_2)]^{n-k_1} \text{ (二项式定理)} \\
=&\sum_{k_1=0}^{n} \frac{n!}{k_1!(n-k_1)!}p_1^{k_1}(1-p_1)^{n-k_1}=1 \\
&\text{(二项式定理，或二项分布概率之和)}
\end{aligned}$$

- **例 1.3（三项分布的实例）** 一大批粉笔，60% 白色，25% 黄色，15% 红色。随机地、顺序地取出 6 支。问这 6 支中恰好有 3 支白色、1 支黄色、2 支红色的概率。

- **解** 用 (白，白，白，黄，红，红) 表示顺序抽到的 6 只的结果。

- 大批量可以认为各次抽取独立且抽到各颜色概率不变：

$$\begin{aligned}
&P(\text{白，白，白，黄，红，红}) \\
=&P(\text{白})P(\text{白})P(\text{白})P(\text{黄})P(\text{红})P(\text{红}) \\
=&(0.6)^3(0.25)^1(0.15)^2
\end{aligned}$$

- 这只是“6 支中恰有 3 支白色、1 支黄色、2 支红色”事件其中一种可能，还有其它可能且每个与上面概率相等。

- 可能个数为

$$m=\frac{6!}{3!1!2!}=60$$

- 这 6 支粉笔打乱次序后重排方式有 6! 个。

- 其中的 3 支白色的次序无关，所以不关心 3 支白色粉笔具体次序情况下重排方式个数为 6!/(3!)。
- 1 支黄色粉笔次序无关，重排方式个数为 6!/(3!1!)。
- 2 支红色粉笔次序无关，重排方式个数为 6!/(3!1!2!)。
- 这就是“6 支中恰有 3 支白色、1 支黄色、2 支红色”事件，只关心个排队位置的颜色而不关心同颜色粉笔位置，总的组合个数。
- 于是

$$\begin{aligned}&P(6\text{ 支中恰有 }3\text{ 支白色、}1\text{ 支黄色、}2\text{ 支红色})\\=&60\cdot(0.6)^3(0.25)^1(0.15)^2=0.0729\end{aligned}$$

- 令

$$\begin{aligned}X=&6\text{ 支中白粉笔的个数}\\Y=&6\text{ 支中黄粉笔的个数}\end{aligned}$$

- 则事件“6 支中恰有 3 支白色、1 支黄色、2 支红色”即

$$\{X=3,Y=1\}=\{(X,Y)=(3,1)\}$$

- 上面推导得

$$P((X,Y)=(3,1))=\frac{6!}{3!1!2!}(0.6)^3(0.25)^1(0.15)^2$$

- 一般地，对于 $0\le k_1\le 6, 0\le k_2\le 6, k_1+k_2\le 6$ 有

$$\begin{aligned}&P(6\text{ 支中恰有 }k_1\text{ 支白、}k_2\text{ 支黄、}6-k_1-k_2\text{ 支红})\\=&P((X,Y)=(k_1,k_2))\\=&\frac{6!}{k_1!k_2!(6-k_1-k_2)!}(0.6)^{k_1}(0.25)^{k_2}(0.15)^{6-k_1-k_2}\end{aligned}$$

  是参数为 $n=6,p_1=0.6,p_2=0.25$ 的三项分布。

4.1.2 边缘分布及其与联合分布的关系

联合分布与边缘分布

- 二维随机变量 (X,Y) 的分布也称为**联合分布**，分量 X 的概率分布称为 (X,Y) 的关于 X 的边缘分布；分量 Y 的概率分布称为 (X,Y) 的关于 Y 的边缘分布。

- 联合分布决定边缘分布。

二维离散联合分布决定边缘分布

- 设

$$P((X,Y)=(x_i,y_j))=p_{ij}\ (i=1,2,\dots,j=1,2,\dots)$$

- 则

$$\begin{aligned} P(X=x_i) &= \sum_j P(X=x_i,Y=y_j) \quad \text{(全概公式)} \\ &= \sum_j p((X,Y)=(x_i,y_j)) \\ &= \sum_j p_{ij} \quad \text{(这是 } X \text{ 的边缘分布)} \qquad (1.4) \\ P(Y=y_j) &= \sum_i p_{ij} \quad \text{(这是 } Y \text{ 的边缘分布)} \qquad (1.4') \end{aligned}$$

- **例 1.4** 对立 1.1 的概率分布表，分别按行求和和按列求和，就得到了 X 和 Y 的边缘分布：

$X \setminus Y$	1	1.2	1.3	1.5	
0.9	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$
1	$\frac{1}{5}$	0	$\frac{1}{5}$	0	$\frac{2}{5}$
1.2	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$
1.4	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$
	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	

- **例 1.5** 设 (X,Y) 服从参数为 n,p_1,p_2 的三项分布：

$$\begin{aligned}&P((X,Y)=(k_1,k_2))\\=&\frac{n!}{k_1!k_2!(n-k_1-k_2)!}p_1^{k_1}p_2^{k_2}(1-p_1-p_2)^{n-k_1-k_2},\\&k_1=0,1,\ldots,n,k_2=0,1,\ldots,n,k_1+k_2\le n\end{aligned}$$

- 则 X 的边缘分布为

$$\begin{aligned}&P(X=k_1)\\=&\sum_{k_2=0}^{n-k_1}P((X,Y)=(k_1,k_2))\\=&\sum_{k_2=0}^{n-k_1}\frac{n!}{k_1!k_2!(n-k_1-k_2)!}p_1^{k_1}p_2^{k_2}(1-p_1-p_2)^{n-k_1-k_2}\\=&\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}\\&\cdot\sum_{k_2=0}^{n-k_1}\frac{(n-k_1)!}{k_2![(n-k_1)-k_2]!}p_2^{k_2}(1-p_1-p_2)^{(n-k_1)-k_2}\\=&\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}[p_2+(1-p_1-p_2)]^{n-k_1}\\&\text{(二项式定理)}\\=&\frac{n!}{k_1!(n-k_1)!}p_1^{k_1}(1-p_1)^{n-k_1}\sim\mathrm{B}(n,p_1)\end{aligned}$$

- 类似地，Y 的边缘分布为 $\mathrm{B}(n, p_2)$。
- 虽然有三种抽取结果，但是如果只关心是否第一种结果，就变成了二项分布。

4.1.3 二维连续型随机向量的分布密度

- **定义 1.2** 对于二维随机向量 $\boldsymbol{\xi} = (X, Y)$, 如果存在非负函数 $p(x, y)(-\infty < x < \infty, -\infty < y < \infty)$, 使对任意 $a < b, c < d$ 及 $D = \{(x, y) : a < x < b, c < y < d\}$ 有

$$P((X, Y) \in D) = \iint_D p(x, y)dxdy \tag{1.5}$$

则称随机向量 $\xi = (X, Y)$ 为连续型的，并称 $p(x, y)$ 为 $\boldsymbol{\xi}$ 的分布密度，也称 $p(x, y)$ 为 (X, Y) 的**联合分布密度** (简称联合密度)。

- 连续型随机向量属于更一般的平面子集 D 的概率为

$$P((X, Y) \in D) = \iint_D p(x, y)dxdy \tag{1.6}$$

但集合 D 的要求涉及到勒贝格积分，这里不做讨论。一般的开集、并集及其有限运算都符合条件。

关于联合密度

- 联合密度不是概率，其在平面点 (x, y) 的小邻域的积分才是概率；
- 类似于物理学中质量面密度的概念；
- $p(x, y)$ 是一个全平面上有定义的二元非负函数。实际中使用的二元密度一般在全平面连续，或者除去个别几条线之后是连续的。

- (1.6) 中的集合 D 可以是全平面，所以

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy = 1$$

- 概率的计算转化为二重积分，$P((X,Y)\in D)$ 的概率是以 D 为底面、以密度函数曲面为顶面的曲顶柱体的体积。

- **例 1.6** 设 (X,Y) 的联合密度为

$$p(x,y) = \begin{cases} Ce^{-(x+y)} & x\geq 0, y\geq 0 \\ 0 & \text{其它} \end{cases}$$

- (1) 求常数 C; (2) 求 $P(0<X<1, 0<Y<1)$。

- **解** (1)

$$\begin{aligned} 1 =& \int_0^{\infty}\int_0^{\infty} p(x,y)dxdy \\ =& C\int_0^{\infty}\int_0^{\infty} e^{-(x+y)}dxdy \\ =& C\int_0^{\infty} e^{-x}dx \int_0^{\infty} e^{-y}dy = C \end{aligned}$$

- (2) 记 $D=\{(x,y): 0<x<1, 0<y<1\}$, 则

$$\begin{aligned} & P(0<X<1, 0<Y<1) = P((X,Y)\in D) \\ =& \iint_D p(x,y)dxdy \\ =& \int_0^1\int_0^1 e^{-(x+y)}dxdy \\ =& \int_0^1 e^{-x}dx \int_0^1 e^{-y}dy \\ =& (1-e^{-1})^2 \end{aligned}$$

连续型二维分布的边缘分布

- **定义 1.3** 对于随机向量 (X,Y), 作为其分量的随机变量 X（或 Y）的密度函数 $p_X(x)$ (或 $p_Y(y)$)，称为 (X,Y) 的关于 X（或 Y）的**边缘分布密度**。

- 连续型二维随机向量的分量一定是连续型随机变量。
- **定理 1.1** 若 (X,Y) 的联合密度是 $p(x,y)$, 则

$$\begin{aligned} p_1(x) &= \int_{-\infty}^{\infty} p(x,y)dy \\ p_2(y) &= \int_{-\infty}^{\infty} p(x,y)dx \end{aligned} \tag{1.7}$$

  分别是 X,Y 的分布密度。
- **证明** 对任意 $a<b$ 有

$$P(a<X<b) = P(a<X<b, -\infty<Y<\infty)$$

- 令

$$D = \{(x,y) : a<x<b, -\infty<y<\infty\}$$

- 由 (1.6) 有

$$\begin{aligned} P(a<X<b) &= \iint p(x,y)dxdy \\ &= \int_a^b \left[\int_{-\infty}^{\infty} p(x,y)dy\right] dx \\ &= \int_a^b p_1(x)dx \end{aligned}$$

- 其中 $p_1(x)$ 是非负可积函数，由密度定义知 $p_1(x)$ 是 X 的密度函数。$p_2(y)$ 类似。

二维均匀分布

- **定义 1.4** 设 G 是平面上面积为 $a(0<a<\infty)$ 的区域，称 (X,Y) 服从 G 上的均匀分布，若 $P((X,Y)\in G)=1$, 且 (X,Y) 取值属于 G 中任意部分 A(A 是 G 的子区域) 的概率与 A 的面积成正比。
- 联合密度为

$$p(x,y) = \begin{cases} \frac{1}{a} & (x,y)\in G \\ 0 & (x,y)\notin G \end{cases}$$

- 二维均匀分布可以用“几何概型”来计算概率。
- 设平面区域 G 为 $y=x^2$ 和 $y=x$ 所夹的有限区域上的均匀分布。
- 求联合密度和边缘密度。
- **解** G 的面积为

$$a=\int_0^1 x dx-\int_0^1 x^2 dx=\frac{1}{6}$$

- 联合密度为

$$p(x,y)=\begin{cases}6 & (x,y)\in G\\ 0 & (x,y)\notin G\end{cases}$$

- 边缘密度

$$\begin{aligned}p_X(x)&=\int_{-\infty}^{\infty}p(x,y)dy\\&=\int_{x^2}^{x}6dy=6(x-x^2)\quad(x\in[0,1])\\p_Y(y)&=\int_{-\infty}^{\infty}p(x,y)dx\\&=\int_{y}^{\sqrt{y}}6dx=6(\sqrt{y}-y)\quad(y\in[0,1])\end{aligned}$$

- $p_X(x)=0, x\notin[0,1]$, $p_Y(y)=0, y\notin[0,1]$。
- 注意边缘分布 X,Y 都可以取遍 $[0,1]$ 内任一个值，但联合分布 (X,Y) 则不能取遍 $[0,1]\times[0,1]$ 内任一个值。

4.1.4 随机变量的独立性

随机变量的独立性

- 随机变量的独立性，就是关于随机变量的事件的独立性。
- **定义 1.5** 设 X,Y 是两个随机变量，如果对任意 $a<b$, $c<d$，事件 $\{a<X<b\}$ 与事件 $\{c<Y<d\}$ 相互独立，则称 X 与 Y 是**相互独立**的，简称独立。

- **定理 1.2** 设 X,Y 分别有分布密度 $p_X(x)$, $p_Y(y)$, 则 X 与 Y 相互独立的充分必要条件是：二元函数

$$p_X(x)p_Y(y) \tag{1.8}$$

是随机向量 (X,Y) 的联合密度。

- **证明** “充分性”。设 $p_X(x)p_Y(y)$ 是 (X,Y) 的联合密度，则

$$\begin{aligned}
&P(a<X<b, c<Y<d)\\
=&\iint\limits_{\substack{a<x<b\\ c<y<d}} p_X(x)p_Y(y)dxdy\\
=&\int_a^b p_X(x)dx\cdot\int_c^d p_Y(y)dy\\
=&P(a<X<b)\cdot P(c<Y<d)
\end{aligned}$$

按独立性定义 X 与 Y 相互独立。

- “必要性”设 X 与 Y 相互独立，令 $D=\{(x,y): a<x<b, c<y<d\}$，则

$$\begin{aligned}
&P((X,Y)\in D)=P(a<X<b, c<Y<d)\\
=&P(a<X<b)\cdot P(c<Y<d)\quad (\text{独立性})\\
=&\int_a^b p_X(x)dx\cdot\int_c^d p_Y(y)dy\\
=&\iint\limits_D [p_X(x)p_Y(y)]dxdy
\end{aligned}$$

- 由联合密度定义知 $p_X(x)p_Y(y)$ 是 (X,Y) 的联合密度。

离散型随机变量的独立性

- **定理 1.3** 设 X 可能取的值是 $x_1, x_2, \ldots$(有限个或可列个), Y 可能取的值是 $y_1, y_2, \ldots$(有限个或可列个), 则 X 与 Y 相互独立的充分必要条件是：对一切 i, j 成立

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) \tag{1.9}$$

- 证明："充分性"：任给定 $a < b, c < d$, 令

$$A = \{x_i : a < x_i < b\}, \quad B = \{y_j : c < y_j < d\}$$

则在 (1.9) 条件下

$$\begin{aligned}
&P(a < X < b, c < Y < d) = P(X \in A, Y \in B) \\
=&\sum_{x_i \in A} \sum_{y_j \in B} P(X = x_i, Y = y_j) \\
=&\left[\sum_{x_i \in A} P(X = x_i)\right] \cdot \left[\sum_{y_j \in B} P(Y = y_j)\right] \\
=&P(a < X < b) \cdot P(c < Y < d)
\end{aligned}$$

- "必要性"：对正整数 n 有

$$\begin{aligned}
&P(X = x_i, Y = y_j) \\
=&P\left(\bigcap_n \left\{X \in (x_i - \frac{1}{n}, x_i + \frac{1}{n}), Y \in (y_j - \frac{1}{n}, y_j + \frac{1}{n})\right\}\right) \\
=&\lim_{n\to\infty} P\left(X \in (x_i - \frac{1}{n}, x_i + \frac{1}{n}), Y \in (y_j - \frac{1}{n}, y_j + \frac{1}{n})\right) \\
=&\lim_{n\to\infty} \left\{P(X \in (x_i - \frac{1}{n}, x_i + \frac{1}{n})) \cdot P(Y \in (y_j - \frac{1}{n}, y_j + \frac{1}{n}))\right\} \\
=&\lim_{n\to\infty} P\left(X \in \left(x_i - \frac{1}{n}, x_i + \frac{1}{n}\right)\right) \cdot \\
&\lim_{n\to\infty} P\left(Y \in \left(y_j - \frac{1}{n}, y_j + \frac{1}{n}\right)\right) \\
=&P\left(\bigcap_n \left\{X \in \left(x_i - \frac{1}{n}, x_i + \frac{1}{n}\right)\right\}\right) \cdot \\
&P\left(\bigcap_n \left\{Y \in \left(y_j - \frac{1}{n}, y_j + \frac{1}{n}\right)\right\}\right) \\
=&P(X = x_i) \cdot P(Y = y_j)
\end{aligned}$$

联合密度与边缘密度的关系

- 联合密度决定边缘密度；
- 边缘密度一般不能决定联合密度；
- 分量独立时边缘密度乘积就是联合密度。
- **例 1.8** 设 $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$, 且 X_1 与 X_2 相互独立，求 (X_1, X_2) 的联合密度。
- **解** X_1, X_2 的密度分别为

$$X_1 \sim p_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right\}$$
$$X_2 \sim p_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right\}$$

- 独立条件下 (X_1, X_2) 的联合密度为

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

4.1.5 二维正态分布

二维正态分布

- 二维正态分布是最常见最重要的多维分布。
- **定义 1.6** 称 $\boldsymbol{\xi} = (X, Y)$ 服从二维正态分布，如果其密度联合密度为

$$\begin{aligned} & p(x,y) \\ =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\cdot\right. \\ & \left.\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\} \end{aligned} \tag{1.10}$$

- 其中 $\mu_1, \mu_2, \sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1$ 是 5 个参数。
- $p(x, y)$ 称为二维正态密度。

二维正态分布的边缘密度

$$
\begin{aligned}
&p_X(x) = \int_{-\infty}^{\infty} p(x,y)dy \\
=&\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}\cdot \\
&\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right\} dy \\
=&\frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}\cdot \\
&\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[t^2 - 2\rho\frac{x-\mu_1}{\sigma_1}t\right]\right\} dt \ \left(\text{令} t = \frac{y-\mu_2}{\sigma_2}\right) \\
=&\frac{1}{2\pi\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}\cdot \\
&\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(t-\rho\frac{x-\mu_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]\right\} dt \\
=&\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \rho^2\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]\right\}\cdot \\
&\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(t-\rho\frac{x-\mu_1}{\sigma_1}\right)^2\right\} dt \\
=&\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\} \sim \mathrm{N}(\mu_1, \sigma_1^2)
\end{aligned}
$$

- 类似有 $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$。
- 二维正态分布的两个分量都服从一元（一维）正态分布。
- 二维正态密度的 5 个参数中的 $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ 分别是其边缘分布的均值和方差参数。
- 上面的推导还说明了 $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy = 1$ (只要先对 y 积分得到 $p_X(x)$ 对 x 的积分)

二维正态分布分量独立的充要条件

- **性质** 若 (X,Y) 服从二维正态分布（参数为 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$），则 X 与 Y 相互独立 $\Longleftrightarrow \rho=0$。
- “充分性”：设 $\rho=0$, 则 $p(x,y)$ 中 $\sqrt{1-\rho^2}=1$, exp 的方括号内第二项消失，而 $p_X(x)$ 和 $p_Y(y)$ 分别是 $\mathrm{N}(\mu_1,\sigma_1^2)$ 和 $\mathrm{N}(\mu_2,\sigma_2^2)$ 密度，显然有

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

- “必要性”：若 X,Y 独立则 $p(x,y)=p_X(x)p_Y(y)$,

$$\begin{aligned}&p_X(x)p_Y(y)\\=&\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}\end{aligned}\tag{1.11}$$

于是 (1.10) 与 (1.11) 都是 (X,Y) 的联合密度。由于这两个密度函数都是连续的，它们应当处处相等，特别地应有

$$p(\mu_1,\mu_2)=p_X(\mu_1)p_Y(\mu_2)$$

即

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}=\frac{1}{2\pi\sigma_1\sigma_2}$$

所以 $\rho=0$。

二维随机变量的分布函数

- **定义 1.7** 设 (X,Y) 是二维随机变量，称函数

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

为它的**分布函数**或**联合分布函数**。

- 若 $\boldsymbol{\xi}=(X,Y)$ 的分布函数有二阶连续偏导数，则 $\frac{\partial^2F(x,y)}{\partial x\partial y}$ 就是 $\boldsymbol{\xi}$ 的分布密度。

4.2 两个随机变量的函数的分布

两个随机变量的函数的分布

- 一个随机变量 X 的函数 $Y = f(X)$ 的分布可以用分布函数法。
- 设 (X, Y) 是随机向量，联合密度为 $p(x, y)$, 求 $Z = f(X, Y)$ 的密度。

4.2.1 和的分布

和的分布

- $Z = X + Y$。
- 用分布函数法。

$$\begin{aligned} F_Z(z) =& P(Z \le z) = P(X + Y \le z) \\ =& P((X, Y) \in D) \end{aligned}$$

其中 D 是平面区域

$$D = \{(x, y) : y \le z - x\}$$

- 积分：

$$\begin{aligned} P(Z \le z) &= \iint_D p(x, y)dxdy \\ &= \iint_{y \le z-x} p(x, y)dxdy \qquad (2.1) \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p(x, y)dy \\ &\quad （二重积分化为累次积分） \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} p(x, u - x)du \\ &\quad （作变换 $u = x + y$, 则 $y = u - x$） \\ &= \int_{-\infty}^{z} du \int_{-\infty}^{\infty} p(x, u - x)dx \ （积分交换次序） \end{aligned}$$

- 即

$$\begin{aligned} F_Z(z) =& P(Z \le z) \\ =& \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} p(x, u-x) dx \right] du \end{aligned}$$

- 这说明

$$p_Z(z) = \int_{-\infty}^{\infty} p(x, z-x) dx \tag{2.2}$$

是 Z 的密度。

- **注** 若 $F(x)$ 是随机变量 X 的分布函数，非负可积函数 $p(x)$ 满足

$$F(x) = \int_{-\infty}^{x} p(u) du, \quad \forall x \in (-\infty, \infty)$$

则 X 为连续型分布，$p(x)$ 是它的分布密度。

- **例 2.1** 设 X 与 Y 相互独立，服从相同的分布 $\mathrm{N}(\mu, \sigma^2)$，求 $X+Y$ 的分布密度。

- **解** (X, Y) 的联合密度为

$$p(x, y) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2} \left[(x-\mu)^2 + (y-\mu)^2 \right] \right\}$$

- 由 (2.2)

$$
\begin{aligned}
& p_Z(z)=\int_{-\infty}^{\infty} p(x, z-x) dx \\
=& \int_{-\infty}^{\infty} \frac{1}{2 \pi \sigma^2} \exp \left\{-\frac{1}{2 \sigma^2}\left[(x-\mu)^2+(z-x-\mu)^2\right]\right\} dx \\
=& \frac{1}{2 \pi \sigma^2} \int_{-\infty}^{\infty} \exp \left\{-\frac{1}{2 \sigma^2}\left[t^2+(z-t-2 \mu)^2\right]\right\} dt \\
& (\text{ 令 } t=x-\mu) \\
=& \frac{1}{2 \pi \sigma^2} \int_{-\infty}^{\infty} \exp \left\{-\frac{1}{2 \sigma^2}\left[2 t^2-2(z-2 \mu) t+(z-2 \mu)^2\right]\right\} dt \\
=& \frac{1}{\sqrt{2 \pi}(\sqrt{2} \sigma)} \exp \left\{-\frac{1}{2} \frac{(x-2 \mu)^2}{2 \sigma^2}\right\} dt \\
& \int_{-\infty}^{\infty} \frac{1}{\sqrt{2 \pi} \frac{\sigma}{\sqrt{2}}} \exp \left\{-\frac{1}{2 \frac{\sigma^2}{2}}\left(t-\frac{x-2 \mu}{2}\right)^2\right\} dt \\
=& \frac{1}{\sqrt{2 \pi}(\sqrt{2} \sigma)} \exp \left\{-\frac{1}{2} \frac{(x-2 \mu)^2}{2 \sigma^2}\right\} \sim \mathrm{N}(2 \mu, 2 \sigma^2)
\end{aligned}
$$

4.2.2 两个例子

一般方法

- 设 (X, Y) 联合密度为 $p(x, y)$, $Z=f(X, Y)$，
- (1) 求 Z 的分布函数

$$P(f(X, Y) \leq z)$$

- (2) 对 $p(x, y)$ 积分

$$P(f(X, Y) \leq z)=\iint_{f(x, y) \leq z} p(x, y) dx dy$$

并进行积分变换，最终化为

$$F_Z(z)=\int_{-\infty}^{z} p_Z(u) du$$

的形式，或 $F_Z(z)$ 可导的形式。

- **例 2.2** 设 X, Y 独立同 $\mathrm{N}(0,1)$ 分布。求 $Z=\sqrt{X^2+Y^2}$ 的密度。
- **解**

$$\begin{aligned} F_Z(z) =& P(\sqrt{X^2+Y^2} \le z) \\ =& \iint\limits_{\sqrt{x^2+y^2}\le z} \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x^2+y^2)\right\} dxdy \quad (z>0) \end{aligned}$$

- 作极坐标变换：

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases} \quad (r \ge 0, 0 \le \theta \le 2\pi)$$

- 有

$$\begin{aligned} F_z(z) &= \int_0^{2\pi} d\theta \int_0^z \frac{1}{2\pi} e^{-\frac{1}{2}r^2} r dr \\ &= \int_0^z r e^{-\frac{1}{2}r^2} dr \quad (z>0) \end{aligned}$$

- 对 $z \le 0$, $F_Z(z)=0$。
- 于是

$$p_Z(z) = \begin{cases} z e^{-\frac{z^2}{2}} & z>0 \\ 0 & z \le 0 \end{cases}$$

- 称为瑞利 (Rayleigh) 分布。
- **例 2.3** 设 X, Y 独立同分布，共同的密度函数为 $p(\cdot)$，分布函数为 $F(\cdot)$。求 $Z=\max(X,Y)$ 的密度函数。
- **解**

$$\begin{aligned} F_Z(z) =& P(Z \le z) = P(\max(X,Y) \le z) \\ =& P(X \le z, Y \le z) \\ =& P(X \le z) \cdot P(Y \le z) \text{ (独立性)} \\ =& F(z) \cdot F(z) = F^2(z) \end{aligned}$$

- 于是 Z 的密度为

$$p_Z(z) = [F^2(z)]' = 2F(z) \cdot F'(z) = 2F(z)p(z)$$

4.2.3 二维变换后的密度

二维变换后的密度

- 已知 (X, Y) 的联合密度，作变换

$$\begin{aligned} U &= f(X, Y) \\ V &= g(X, Y) \end{aligned}$$

  则 (U, V) 联合密度？

- **定理 2.1** 设 (X, Y) 有联合密度 $p(x, y)$, 且区域 A(可以是全平面) 满足 $P((X, Y) \in A) = 1$。

  又函数 $f(x, y), g(x, y)$ 满足：

- (1) 对任意实数 u, v，方程组

$$\begin{cases} f(x, y) = u \\ g(x, y) = v \end{cases} \tag{2.4}$$

  在 A 中至多有一个解 $x = x(u, v), y = y(u, v)$;

- (2) f, g 在 A 中有连续偏导数；

- (3) 雅可比行列式 $\frac{\partial(f,g)}{\partial(x,y)}$ 在 A 中处处不等于 0。

- 设 $U = f(X, Y), V = g(X, Y)$，

$$G = \{(u, v) : \text{方程组 (2.4) 在 } A \text{ 中有解}\}$$

  则

$$q(u, v) = \begin{cases} p(x(u, v), y(u, v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right| & \text{当}(u, v) \in G \\ 0 & \text{当}(u, v) \notin G \end{cases}$$

  是 (U, V) 的联合密度。

- **证** 给定 $a<b, c<d$, 设 $D=\{(u,v): a<u<b, c<v<d\}$, $D^*=\{(x,y):(f(x,y),g(x,y))\in D\}$, 则 $(f(x,y),g(x,y))$ 是 $D^*\cap A$ 到 $D\cap G$ 上的一一映射，其逆映射为 $(x(u,v),y(u,v))$。由重积分的变数替换公式知

$$\iint\limits_{D^*\cap A} p(x,y)dxdy=\iint\limits_{D\cap G} p(x(u,v),y(u,v))\left|\frac{\partial(x,y)}{\partial(u,v)}\right|dudv$$

- 于是

$$\begin{aligned}&P((U,V)\in D)=P((f(X,Y),g(X,Y))\in D)\\ =&P((X,Y)\in D^*)=P((X,Y)\in D^*\cap A)=\iint_{D^*\cap A}p(x,y)dxdy\\ =&\iint\limits_{D\cap G} p(x(u,v),y(u,v))\left|\frac{\partial(x,y)}{\partial(u,v)}\right|dudv=\iint\limits_{D} q(u,v)dudv\end{aligned}$$

- 即 $q(u,v)$ 是 (U,V) 的联合密度。

- **例 2.4** 设 X,Y 相互独立，都服从 $[0,1]$ 上的均匀分布。

$$\begin{aligned}U=&\sqrt{-2\ln X}\cos 2\pi Y\\ V=&\sqrt{-2\ln X}\sin 2\pi Y\end{aligned}$$

求 (U,V) 的联合密度。

- **解** 用定理 2.1。

$$\begin{aligned}f(x,y)=&\sqrt{-2\ln x}\cos 2\pi y\\ g(x,y)=&\sqrt{-2\ln x}\sin 2\pi y\\ A=&\{(x,y):0<x<1,0<y<1\text{但}y\neq\frac{1}{4},\frac{2}{4},\frac{3}{4}\}\end{aligned}$$

- 则

$$G=\{(u,v):u\neq 0, v\neq 0\}$$

(u,v) 是极径为 $\sqrt{-2\ln x}$、极角为 $2\pi y$ 的极坐标对应的直角坐标，所以给定 $(u,v)\neq(0,0)$ 后 (x,y) 唯一确定,

$$\begin{aligned}x=&x(u,v)=e^{-\frac{1}{2}(u^2+v^2)}\\ y=&y(u,v)=\frac{1}{2\pi}\operatorname{atan2}(u,v)\end{aligned}$$

其中 $\operatorname{atan2}(u,v)$ 表示求直角坐标 (u,v) 对应的极角的函数，只要 $(u,v)\neq(0,0)$ 它就是存在唯一取值于 $[0,2\pi)$ 的, 为了简单起见我们在 A 的定义中还去掉了两个直角坐标系坐标轴。

- 对 $y(u,v)$

$$\begin{aligned} y(u,v) =& \frac{1}{2\pi}\operatorname{atan2}(u,v) \\ =& \begin{cases} \frac{1}{2\pi}\arctan\frac{v}{u} & u>0, v>0 \\ 1+\frac{1}{2\pi}\arctan\frac{v}{u} & u>0, v<0 \\ \frac{1}{2}+\frac{1}{2\pi}\arctan\frac{v}{u} & u<0 \end{cases} \end{aligned}$$

- 所以

$$\begin{aligned} \frac{\partial(x,y)}{\partial(u,v)} =& \begin{vmatrix} -ue^{-\frac{1}{2}(u^2+v^2)} & -ve^{-\frac{1}{2}(u^2+v^2)} \\ -\frac{1}{2\pi}\frac{v}{u^2+v^2} & \frac{1}{2\pi}\frac{u}{u^2+v^2} \end{vmatrix} \\ =& -\frac{1}{2\pi}e^{-\frac{1}{2}(u^2+v^2)} \end{aligned}$$

- 而 (X,Y) 的联合密度为

$$p(x,y)=\begin{cases} 1 & 0\leq x\leq 1, 0\leq y\leq 1 \\ 0 & \text{其它} \end{cases}$$

- 由定理 2.1 知 (U,V) 的联合密度为

$$q(u,v)=\begin{cases} \frac{1}{2\pi}e^{-\frac{1}{2}(u^2+v^2)} & u\neq 0, v\neq 0 \\ 0 & u=0\text{或}v=0 \end{cases}$$

- 随机向量的联合密度在若干条曲线上改变值后仍为该随机向量的联合密度。

- 所以 (U,V) 的联合密度也可写成

$$\varphi(u,v)=\frac{1}{2\pi}e^{-\frac{1}{2}(u^2+v^2)}$$

- 这是参数为 $(0,0,1,1,0)$ 的二元正态分布密度，所以 U,V 相互独立，分别服从标准正态分布。

- 设 X,Y 相互独立，都服从 $\mathrm{N}(0,1)$，

$$\begin{aligned} X =& R\cos\Theta \\ Y =& R\sin\Theta \\ & (R\geq 0, 0\leq\Theta<2\pi) \end{aligned}$$

  求 (R,Θ) 的联合密度与边缘密度。

- **解** 用定理 2.1。取

$$A=\{(x,y): x\neq 0, y\neq 0\}$$

  并改变密度函数 $q(\cdot,\cdot)$ 在个别点上的值，可以求得 (R,Θ) 的联合密度为

$$\varphi(r,\theta)=\begin{cases} \frac{1}{2\pi}re^{-\frac{1}{2}r^2} & r>0, 0<\theta<2\pi \\ 0 & \text{其它} \end{cases}$$

- 取

$$G=\{(r,\theta): r>0, 0<\theta<2\pi\text{且}\theta\neq\frac{\pi}{2},\pi,\frac{3\pi}{2}\}$$

- (R,Θ) 是 (X,Y) 的极坐标, 于是

$$\frac{\partial(x,y)}{\partial(r,\theta)}=\begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix}=r$$

- 所以 (R,Θ) 的联合密度为

$$f(x(r,\theta),y(r,\theta))\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right|=\frac{1}{2\pi}e^{-\frac{1}{2}r^2}\cdot r$$

- 令

$$f(r)=\begin{cases} re^{-\frac{1}{2}r^2} & r>0 \\ 0 & r\le 0 \end{cases}$$

$$g(\theta)=\begin{cases} \frac{1}{2\pi} & 0<\theta<2\pi \\ 0 & \text{其它} \end{cases}$$

- 则 $\varphi(r,\theta)=f(r)\cdot g(\theta)$，$R,\Theta$ 相互独立，分别以 $f(r)$ 和 $g(\theta)$ 为边缘密度（瑞利分布和均匀分布）。

- **例 2.6** 设 R 与 Θ 相互独立，都服从 $(0,1)$ 上的均匀分布，r_1 和 r_2 为常数 $(0\le r_1<r_2)$，

$$X=r_2\sqrt{\frac{r_1^2}{r_2^2}+\left(1-\frac{r_1^2}{r_2^2}\right)R}\cos(2\pi\Theta)$$

$$Y=r_2\sqrt{\frac{r_1^2}{r_2^2}+\left(1-\frac{r_1^2}{r_2^2}\right)R}\sin(2\pi\Theta)$$

则 (X,Y) 服从环

$$D=\{(x,y):r_1^2\le x^2+y^2\le r_2^2\}$$

上的均匀分布。

- **证** 令

$$x=f(r,\theta)=r_2\sqrt{\frac{r_1^2}{r_2^2}+\left(1-\frac{r_1^2}{r_2^2}\right)r}\cos(2\pi\theta)$$

$$y=g(r,\theta)=r_2\sqrt{\frac{r_1^2}{r_2^2}+\left(1-\frac{r_1^2}{r_2^2}\right)r}\sin(2\pi\theta)$$

- 则

$$r=r(x,y)=\frac{x^2+y^2-r_1^2}{r_2^2-r_1^2}$$

$$\theta=\theta(x,y)=\frac{1}{2\pi}=\operatorname{atan2}(x,y)$$

- 雅可比行列式

$$J = \frac{\partial(r,\theta)}{\partial(x,y)} = \frac{1}{\pi(r_2^2 - r_1^2)}$$

- (R,Θ) 的取值区域为 $A = \{(r,\theta) : 0 < r < 1, 0 < \theta < 1\}$, 对应的 (X,Y) 的取值区域为环 D。由定理 2.1, (X,Y) 的联合密度为

$$\begin{cases} \frac{1}{\pi(r_2^2 - r_1^2)} & (x,y) \in D \\ 0 & \text{其它} \end{cases}$$

- 即 (X,Y) 服从圆环 D 上的均匀分布。

4.3 随机向量的数字特征

随机向量的数字特征

- 对一个随机变量，我们讨论了其线性变换的期望和方差的性质。
- 对两个随机变量的函数，其期望如何计算？两个随机变量的线性组合的期望和方差有什么性质？

4.3.1 两个随机变量的函数的均值公式

两个随机变量的函数的均值公式

- 设随机向量 (X,Y) 有密度 $p(x,y)$, 对两个随机变量的函数 $Z = f(X,Y)$，也有

$$E(Z) = E[f(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p(x,y)dxdy$$

- **例 3.1** 设 X,Y 独立同分布，共同分布是 $\mathrm{N}(0,1)$，求

$$E\sqrt{X^2 + Y^2}$$

- **解法 1** 用公式 (3.1)

$$\begin{aligned}
&E\sqrt{X^2+Y^2}\\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sqrt{x^2+y^2}\cdot\frac{1}{2\pi}\cdot e^{-\frac{1}{2}(x^2+y^2)}dxdy\\
=&\int_0^{2\pi}d\theta\int_0^{\infty}r\cdot\frac{1}{2\pi}\cdot e^{-\frac{1}{2}r^2}rdr\ (\text{作极坐标变换})\\
=&\int_0^{\infty}r^2e^{-\frac{1}{2}r^2}dr=\frac{\sqrt{2\pi}}{2}\int_{-\infty}^{\infty}x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx\\
=&\frac{\sqrt{2\pi}}{2}
\end{aligned}$$

- **解法 2** 由 §2, $\sqrt{X^2+Y^2}$ 服从瑞利分布，其密度为

$$p(z)=\begin{cases}ze^{-\frac{1}{2}z^2} & z>0\\ 0 & z\le 0\end{cases}$$

- 所以

$$\begin{aligned}
&E\sqrt{X^2+Y^2}\\
=&\int_{-\infty}^{\infty}zp(z)dz\\
=&\int_0^{\infty}z^2e^{-\frac{1}{2}z^2}dz\\
=&\frac{\sqrt{2\pi}}{2}
\end{aligned}$$

- 解法 1 不需要求 Z 的分布密度。

4.3.2 均值与方差的性质

分量的均值与方差

- 设 (X,Y) 的联合密度为 $p(x,y)$，分量的边缘密度为 $p_X(x)$ 和 $p_Y(y)$，

则

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x p_X(x) dx \\
E(Y) &= \int_{-\infty}^{\infty} y p_Y(y) dy \\
D(X) &= \int_{-\infty}^{\infty} [x - E(X)]^2 p_X(x) dx \\
D(Y) &= \int_{-\infty}^{\infty} [y - E(Y)]^2 p_Y(y) dy
\end{aligned}
$$

- 由随机向量函数的期望公式 (3.1) 又有

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x p(x,y) dx dy \\
E(Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y p(x,y) dx dy \\
D(X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - E(X)]^2 p(x,y) dx dy \\
D(Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [y - E(Y)]^2 p(x,y) dx dy
\end{aligned}
$$

随机变量和的期望和方差

- 对随机变量 X, Y 有

$$
\begin{aligned}
E(X+Y) =& E(X) + E(Y) && (3.3) \\
D(X+Y) =& D(X) + D(Y) \\
&+ 2E\{[X - E(X)][Y - E(Y)]\} && (3.4)
\end{aligned}
$$

- 当 X, Y 独立时有

$$
\begin{aligned}
E(X \cdot Y) =& E(X) \cdot E(Y) && (3.5) \\
D(X+Y) =& D(X) + D(Y) && (3.6) \\
E\{[X - E(X)][Y - E(Y)]\} =& 0
\end{aligned}
$$

- **证** 对 (3.3), 由 (3.1) 式 (这需要 (X,Y) 为连续型，但结论对一般随机变量成立)

$$\begin{aligned}E(X+Y)=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)p(x,y)dxdy\\=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xp(x,y)dxdy\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yp(x,y)dxdy\\=&E(X)+E(Y)\end{aligned}$$

- 对 (3.4),

$$\begin{aligned}&D(X+Y)=E[(X+Y)-E(X+Y)]^2\\=&E\{[X-E(X)]+[Y-E(Y)]\}^2\\=&E\{[X-E(X)]^2+[Y-E(Y)]^2+2[X-E(X)][Y-E(Y)]\}\\=&E[X-E(X)]^2+E[Y-E(Y)]^2+2E\{[X-E(X)][Y-E(Y)]\}\\&\text{(用 (3.3) 式)}\\=&D(X)+D(Y)+2E\{[X-E(X)][Y-E(Y)]\}\end{aligned}$$

- 对 (3.5), 因 X,Y 相互独立，有

$$p(x,y)=p_X(x)\cdot p_Y(y)$$

- 由 (3.1) 知

$$\begin{aligned}&E(X\cdot Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyp(x,y)dxdy\\=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyp_X(x)p_Y(y)dxdy\\=&\left[\int_{-\infty}^{\infty}xp_X(x)dx\right]\cdot\left[\int_{-\infty}^{\infty}yp_Y(y)dy\right]\\=&E(X)\cdot E(Y)\end{aligned}$$

- 对 (3.6), 当 X,Y 相互独立时用 (3.5)

$$\begin{aligned}
&E\{[X-E(X)][Y-E(Y)]\}\\
=&E\{XY-XE(Y)-YE(X)+E(X)\cdot E(Y)\}\\
=&E(XY)-E(X)\cdot E(Y)-E(Y)\cdot E(X)+E(X)\cdot E(Y)\\
&\text{(用 (3.3) 及期望的线性性质)}\\
=&E(XY)-E(X)\cdot E(Y)\\
=&0 \quad \text{(用 (3.5) 式)}
\end{aligned}$$

- 从而由 (3.4) 即得 (3.6) 式。

4.3.3 协方差

协方差

- **定义 3.1** 称向量 $(E(X),E(Y))$ 为随机向量 (X,Y) 的均值，称数值 $E\{[X-E(X)][Y-E(Y)]\}$ 为 X,Y 的**协方差**, 记为 $\mathrm{Cov}(X,Y)$ 或 σ_{XY}。
- 由公式 (3.1)

$$\begin{aligned}
&\mathrm{Cov}(X,Y)=\sigma_{XY}\\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x-E(X)][y-E(Y)]p(x,y)dxdy
\end{aligned}$$

- $D(X)$, $D(Y)$ 也可以记成 σ_{XX}, σ_{YY}。
- 当 X,Y 相互独立时 $\mathrm{Cov}(X,Y)=0$。
- 若 $\mathrm{Cov}(X,Y)=0$，则称 X,Y **不相关** 。
- 独立必不相关；不相关不一定独立。
- **例 3.2** 设 (X,Y) 的联合密度为

$$p(x,y)=\begin{cases}\frac{1}{\pi} & x^2+y^2\le 1\\ 0 & \text{其它}\end{cases}$$

(单位圆内的二维均匀分布), 求 $\sigma_{XX},\sigma_{YY},\sigma_{XY}$。

- **解** 先求 $E(X), E(Y)$。

$$\begin{aligned} E(X) =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)dxdy = \iint\limits_{x^2+y^2\le 1} x\cdot\frac{1}{\pi}dxdy \\ =& \frac{1}{\pi}\int_{-1}^{1} dy \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} xdx = 0 \end{aligned}$$

这是因为内层积分被积函数 x 是奇函数，在关于 0 对称的区间 $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$ 积分等于 0。

- 类似地，$E(Y) = 0$。

- 对 σ_{XX}:

$$\begin{aligned} \sigma_{XX} =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x-E(X)]^2 p(x,y)dxdy \\ =& \frac{1}{\pi}\iint\limits_{x^2+y^2\le 1} x^2 dxdy \\ =& \frac{1}{\pi}\int_0^{2\pi} d\theta \int_0^1 r^2\cos^2\theta\cdot r dr \text{ (作极坐标变换)} \\ =& \frac{1}{\pi}\left(\int_0^{2\pi}\cos^2\theta d\theta\right)\cdot\left(\int_0^1 r^3 dr\right) \\ =& \frac{1}{\pi}\cdot\pi\cdot\frac{1}{4} = \frac{1}{4} \end{aligned}$$

- 同理 $\sigma_{YY} = \frac{1}{4}$。

- 对 σ_{XY}:

$$\begin{aligned} \sigma_{XY} =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [x-E(X)][y-E(Y)]p(x,y)dxdy \\ =& \frac{1}{\pi}\iint\limits_{x^2+y^2\le 1} xydxdy \\ =& \frac{1}{\pi}\int_{-1}^{1} ydy\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} xdx \\ =& 0 \end{aligned}$$

其中内层积分是在关于 0 对称的区间 $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$ 上的奇函数 x 的积分。

- 这个联合分布的协方差 $\sigma_{XY}=0(X$ 和 Y 不相关), 但是 X 和 Y 不独立：

- X 和 Y 单独都可以取 $[-1,1]$ 内的任意值，但是给定 $X=x$ 后，Y 只能在 $[-\sqrt{1-x^2},\sqrt{1-x^2}]$ 内取值。

- 比如，考虑 $A=[\frac{\sqrt{2}}{2},1]$, $B=[\frac{\sqrt{2}}{2},1]$, 则 $P(X\in A)>0, P(Y\in B)>0$，但

$$P(X\in A,Y\in B)=0\neq P(X\in A)\cdot P(Y\in B)$$

- 可以作为不相关不一定独立的例子。

- **例 3.3** 设 (X,Y) 服从二维正态分布，密度函数为

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

  求 σ_{XY}。

- 前面已说明 $E(X)=\mu_1$, $E(Y)=\mu_2$, $D(X)=\sigma_1^2$, $D(Y)=\sigma_2^2$。

- **解**

$$
\begin{aligned}
\sigma_{XY} =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x-E(X)][y-E(Y)]p(x,y)dxdy \\
& (\text{作变量替换 } u=\frac{x-\mu_1}{\sigma_1},\, v=\frac{y-\mu_2}{\sigma_2}) \\
=& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sigma_1 u\cdot\sigma_2 v\cdot \\
& \exp\left\{-\frac{1}{2(1-\rho^2)}(u^2-2\rho uv+v^2)\right\}\sigma_1\sigma_2 dudv \\
=& \frac{\sigma_1\sigma_2}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}vdv\int_{-\infty}^{\infty}u\cdot \\
& \exp\left\{-\frac{1}{2(1-\rho^2)}\left[(u-\rho v)^2-\rho^2v^2+v^2\right]\right\}du \\
=& \frac{\sigma_1\sigma_2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ve^{-\frac{v^2}{2}}dv\int_{-\infty}^{\infty}u\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left\{-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right\}du \\
=& \frac{\sigma_1\sigma_2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}ve^{-\frac{v^2}{2}}dv\cdot\rho v \qquad (\mathrm{N}(\rho v,1-\rho^2)\text{ 的期望}) \\
=& \rho\sigma_1\sigma_2\int_{-\infty}^{\infty}v^2\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}dv \\
=& \rho\sigma_1\sigma_2 \qquad (\mathrm{N}(0,1)\text{ 的方差})
\end{aligned}
$$

- 二元正态分布中参数 ρ 的意义:

$$\rho=\frac{\sigma_{XY}}{\sigma_1\sigma_2}=\frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}}$$

- §4.1 证明了，对二元正态分布，分量独立的充分必要条件是 $\rho=0$。

- 所以二元正态分布分量独立的充分必要条件是 $\sigma_{XY}=0$, 即对二元正态分布，独立与不相关是等价的。

- 对于一般的二元分布，不相关不能推出独立。

4.3.4 相关系数

相关系数

- **定义 3.2** 称

$$\frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}}$$

为 X,Y 的**相关系数**（要求分母不等于 0），记作 ρ_{XY} 或 ρ。即

$$\rho_{XY}=\frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}} \tag{3.9}$$

- 对二元正态分布，参数 ρ 是两个分量的相关系数。

相关系数的性质

- 相关系数满足

$$|\rho|\le 1 \tag{3.10}$$

- **证** 对于任意实数 λ，有

$$\begin{aligned}
&D(Y-\lambda X)\\
=&E[(Y-\lambda X)-E(Y-\lambda X)]^2\\
=&E\{[Y-E(Y)]-\lambda[X-E(X)]\}^2\\
=&E[Y-E(Y)]^2+\lambda^2E[X-E(X)]^2\\
&-2\lambda E\{[X-E(X)][Y-E(Y)]\}\\
=&\sigma_{YY}+\lambda^2\sigma_{XX}-2\lambda\sigma_{XY}\\
=&\sigma_{XX}\lambda^2-2\sigma_{XY}\lambda+\sigma_{YY}
\end{aligned} \tag{3.11}$$

- 取 $\lambda=b\triangleq\frac{\sigma_{XY}}{\sigma_{XX}}$, 则

$$\begin{aligned}
D(Y-bX)=&\sigma_{YY}\left(b^2\frac{\sigma_{XX}}{\sigma_{YY}}-2b\frac{\sigma_{XY}}{\sigma_{YY}}+1\right)\\
=&\sigma_{YY}\left(\frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}-2\frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}+1\right)\\
=&\sigma_{YY}\left(1-\rho^2\right)
\end{aligned}$$

- 因方差总是非负所以

$$1-\rho^2\ge 0,\quad |\rho|\le 1$$

- 另外由证明可见 $|\rho| = 1$ 当且仅当 $D(Y - bX) = 0$，而随机变量方差等于 0，则此随机变量等于某个常数 a 的概率为 1。所以 $|\rho| = 1$ 当且仅当存在常数 a 使得

$$P(Y - bX = a) = 1, \quad \text{即} P(Y = a + bX) = 1$$

相关系数的意义

- 相关系数 $\rho = \rho_{XY}$ 刻画了 X, Y 间线性关系的近似程度。
- $|\rho|$ 越接近于 1, X 与 Y 越近似地有线性关系。
- 但是，ρ 不一定能反映非线性的关系。
- 比如, $X \sim \mathrm{N}(0, 1)$，$Y = X^2$，则 Y 与 X 有密切的非线性关系，但是 $\rho_{XY} = 0$。

4.3.5 线性预测与相关系数

线性预测

- 预测问题是统计学的重要研究课题。
- 用随机变量 X 的函数去预测随机变量 Y，最简单的函数就是 X 的线性函数 $a + bX$。
- 设 $\sigma_{XX} > 0, \sigma_{YY} > 0$，如何选取 a, b 使得 $a + bX$ 与 Y 最接近？
- 什么是“最接近”?
- 用

$$Q = Q(a, b) = E[Y - (a + bX)]^2$$

  作为 Y 与 $a + bX$ 的“距离”，衡量预测的接近程度，叫做预测的**均方误差**。

- 求 a, b 使 $Q(a, b)$ 最小。

- $Q(a,b)$ 关于 a,b 是二次多项式，可以用配方的办法或令偏导数等于零的办法来求最小值点。
- 配方法：

$$\begin{aligned} Q(a,b) =& E[Y-(a+bX)]^2 \\ =& E\big\{\,[(Y-E(Y))-b(X-E(X))] \\ &+[E(Y)-bE(X)-a]\,\big\}^2 \end{aligned} \tag{*}$$

- 记

$$Z = (Y-E(Y))-b(X-E(X))$$

易见

$$E(Z) = 0$$

- 并注意 $E(Y)-bE(X)-a$ 是非随机的，
- 这样 (*) 展开时交叉项为

$$\begin{aligned} & 2E\big\{Z\,[E(Y)-bE(X)-a]\,\big\} \\ =& 2E(Z)\cdot[E(Y)-bE(X)-a]=0 \end{aligned}$$

- 于是 (*) 式展开为

$$\begin{aligned} & Q(a,b) \\ =& E\,[(Y-E(Y))-b(X-E(X))]^2 \;+\; [E(Y)-bE(X)-a]^2 \\ =& E[Y-E(Y)]^2+b^2E[X-E(X)]^2 \\ & -2bE[(X-E(X))(Y-E(Y))]+[E(Y)-bE(X)-a]^2 \\ =& \sigma_{YY}+b^2\sigma_{XX}-2b\sigma_{XY}+[E(Y)-bE(X)-a]^2 \\ =& \sigma_{XX}\left(b-\frac{\sigma_{XY}}{\sigma_{XX}}\right)^2+[E(Y)-bE(X)-a]^2+\left(\sigma_{YY}-\frac{\sigma_{XY}^2}{\sigma_{XX}}\right) \end{aligned}$$

- 最小化 $Q(a,b)$，第一项最小应取

$$b = \frac{\sigma_{XY}}{\sigma_{XX}}$$

- 第二项最小只要

$$a = E(Y)-bE(X)$$

- $Q(a,b)$ 的最小值点为

$$\begin{aligned} b^* =& \frac{\sigma_{XY}}{\sigma_{XX}} = \rho\sqrt{\frac{\sigma_{YY}}{\sigma_{XX}}} \\ a^* =& E(Y) - b^* E(X) \end{aligned}$$

- 最小值为

$$Q(a^*, b^*) = \sigma_{YY} - \frac{\sigma_{XY}^2}{\sigma_{XX}} = \sigma_{YY}(1-\rho^2)$$

- 只要 $\rho \neq 0$，预测误差就小于 σ_{YY}，σ_{YY} 是用 $a + BX \equiv E(Y)$ 来预报 Y 的均方误差。
- $|\rho|$ 越接近于 1，预测误差越小。

4.4 关于 n 维随机向量

n 维随机向量

- n 维随机向量的有关结论都与二维时类似。

4.4.1 联合密度与边缘密度

联合密度与边缘密度

- **定义 4.1** 对于 n 维随机向量 $\boldsymbol{\xi} = (X_1, X_2, \ldots, X_n)$, 如果存在非负函数 $p(x_1, x_2, \ldots, x_n)$，使对于任意 n 维长方体

$$\begin{aligned} D =& \big\{(x_1, x_2, \ldots, x_n) : a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, \\ & a_n < x_n < b_n\big\} \end{aligned}$$

均成立

$$P(\boldsymbol{\xi} \in D) = \underbrace{\iint}_{D} \cdots \int p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n \tag{4.1}$$

则称 $\boldsymbol{\xi}=(X_1,X_2,\ldots,X_n)$ 是连续型的，并称 $p(x_1,x_2,\ldots,x_n)$ 为 $\boldsymbol{\xi}$ 的分布密度，或称联合分布密度（简称联合密度）。

- 对于 n 维空间中相当一般的集合 D 仍成立

$$P(\boldsymbol{\xi}\in D)=\underset{D}{\iint}\cdots\int p(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n \tag{4.2}$$

- 称 $(X_1,X_2,\ldots,X_n)$ 的一部分分量所构成的向量（如 (X_1,X_2)）的分布密度为边缘密度。

- 每个分量 X_i 的分布密度 $p_i(x_i)$ 也是 $(X_1,X_2,\ldots,X_n)$ 的边缘密度，称为单个密度。

- 联合密度决定边缘密度。

- 求某几个分量的边缘密度，只要从联合密度中把其它分量的自变量积分掉，如

$$\begin{aligned} p_1(x_1)&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}p(x_1,x_2,\ldots,x_n)dx_2dx_3\ldots dx_n \\ p_{12}(x_1,x_2)&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}p(x_1,x_2,\ldots,x_n)dx_3dx_4\ldots dx_n \end{aligned} \tag{4.3}$$

n 维正态分布

- **定义 4.2** 称随机向量 $\boldsymbol{\xi}=(X_1,X_2,\ldots,X_n)$ 服从 n 维正态分布 (也称 $\boldsymbol{\xi}$ 是 n 维正态随机向量), 如果它有分布密度

$$\begin{aligned} &p(x_1,x_2,\ldots,x_n) \\ =&\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big\} \end{aligned} \tag{4.4}$$

其中 $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)^T$, $\boldsymbol{\mu}=(\mu_1,\mu_2,\ldots,\mu_n)$ 是非随机的常数向量，$\Sigma=(\sigma_{ij})_{n\times n}$ 是 n 阶正定常数矩阵。

- 记 $\boldsymbol{\xi}\sim \mathrm{N}(\boldsymbol{\mu},\Sigma)$。

4.4.2 独立性

独立性

- **定义 4.3** 设 $X_1, X_2, \ldots, X_n$ 是 n 个随机变量，如果对任意 $a_i < b_i (i = 1, 2, \ldots, n)$, 事件 $\{a_1 < X_1 < b_1\}$, $\{a_2 < X_2 < b_2\}$, ..., $\{a_n < X_n < b_n\}$ 相互独立，则称 $X_1, X_2, \ldots, X_n$ 是**相互独立的**。
- **定理 4.1** 设 $X_1, X_2, \ldots, X_n$ 的分布密度分别是 $p_1(x_1)$, $p_2(x_2)$, ..., $p_n(x_n)$, 则 $X_1, X_2, \ldots, X_n$ 相互独立的充分必要条件是

  $$p_1(x_1)p_2(x_2)\ldots p_n(x_n)$$

  是 $(X_1, X_2, \ldots, X_n)$ 的联合分布密度。

4.4.3 n 个随机变量的函数的分布

n 个随机变量的函数的分布

- 设 $Y = f(X_1, X_2, \ldots, X_n)$，求 Y 的分布函数

$$\begin{aligned} F_Y(y) =& P(f(X_1, X_2, \ldots, X_n) \leq y) \\ =& \int\cdots\int_{f(x_1,x_2,\ldots,x_n)\leq y} p(x_1, x_2, \ldots, x_n)dx_1dx_2\ldots dx_n \end{aligned} \tag{4.5}$$

4.4.4 数字特征

数字特征

- 随机向量函数的期望

$$\begin{aligned} & E[f(X_1, X_2, \ldots, X_n)] \\ =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) \\ & \cdot p(x_1, x_2, \ldots, x_n)dx_1dx_2\ldots dx_n \end{aligned}$$

  (要求右端的积分绝对收敛)

- 离散型随机向量积分变成求和。
- 称 $(E(X_1), E(X_2), \ldots, E(X_n))$ 为随机向量 $(X_1, X_2, \ldots, X_n)$ 的期望（均值）。

- 期望的性质

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

- 当 $X_1, X_2, \ldots, X_n$ 相互独立时，

$$\begin{aligned} E(X_1 X_2 \ldots X_n) &= E(X_1)E(X_2)\ldots E(X_n) \qquad (4.8)\\ D(X_1 + X_2 + \cdots + X_n) &= D(X_1) + D(X_2) + \cdots + D(X_n) \end{aligned}$$

- 这三条都可以由两个随机变量的相应结论递推证明。
- 关于独立性：设 $X_1, X_2, \ldots, X_m$ 与 $Y_1, Y_2, \ldots, Y_n$ 相互独立，则 $f(X_1, X_2, \ldots, X_m)$ 与 $g(Y_1, Y_2, \ldots, Y_n)$ 相互独立。

协方差阵

- 记

$$\begin{aligned} \sigma_{ij} &= E[(X_i - E(X_i))(X_j - E(X_j))] \\ &\quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n) \qquad (4.9) \end{aligned}$$

  则 $\sigma_{ii} = D(X_i)$, $i \neq j$ 时 σ_{ij} 是 X_i 与 X_j 的协方差。
- 称矩阵

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

  为 $(X_1, X_2, \ldots, X_n)$ 的**协差阵** (或**协方差阵，方差阵**)，记为 Σ。
- Σ 是实对称矩阵，是非负定阵。

Σ 非负定的证明

- 对任意实数 $a_1, a_2, \ldots, a_n$

$$\begin{aligned} D(\sum_{i=1}^{n} a_i X_i) =& E\left[\sum_{i=1}^{n} a_i (X_i - E(X_i))\right]^2 \\ =& E\left[\sum_{i=1}^{n} a_i (X_i - E(X_i)) \cdot \sum_{j=1}^{n} a_j (X_j - E(X_j))\right] \\ =& \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j E[(X_i - E(X_i))(X_j - E(X_j))] \\ =& \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma_{ij} \end{aligned}$$

- 由方差非负及非负定的定义即知 Σ 非负定。

相关系数与相关阵

- 记

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} \quad (i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n)$$

则 $i \neq j$ 时 ρ_{ij} 是 X_i 与 X_j 的相关系数。而

$$\rho_{ii} \equiv 1 \quad (i = 1, 2, \ldots, n)$$

- 称矩阵

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$$

为 $(X_1, X_2, \ldots, X_n)$ 的**相关阵** (或相关系数阵), 记为 R。R 也是实对称非负定阵。

- 如果把 $X_1, X_2, \ldots, X_n$ 分别标准化得到 $Y_1, Y_2, \ldots, Y_n$,则 R 是 $(Y_1, Y_2, \ldots, Y_n)$ 的协方差阵。
- 记

$$C = \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{nn}}} \end{pmatrix}$$

- 则有

$$R = C\Sigma C$$

n 维分布函数

- **定义 4.4** 设 $\boldsymbol{\xi} = (X_1, X_2, \ldots, X_n)$ 是 n 维随机向量，称 n 元函数

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$$

  为 $\boldsymbol{\xi}$ 的分布函数（也称为 $X_1, X_2, \ldots, X_n$ 的联合分布函数)。
- 如果 $\boldsymbol{\xi}$ 有联合密度 $p(x_1, x_2, \ldots, x_n)$ 则

$$\begin{aligned} &F(x_1, x_2, \ldots, x_n) \\ =&\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_n} p(u_1, u_2, \ldots, u_n)du_1du_2\ldots du_n \end{aligned}$$

4.5 条件分布与条件期望

4.5.1 条件分布

条件概率与随机变量

- 随机变量可以用来表示事件的不同结果，如：

  - 成败型试验 n 次独立重复成功次数；
  - 测量两地之间距离误差大小；
  - 电子产品从启用到失效的时间，等等。

- 条件概率是在事件 A 发生的条件下，事件 B 发生的概率。

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

- 条件概率可以针对两个随机变量的相关事件。

例：条件概率与随机变量

- **例 5.1** 一射手进行射击，单发击中目标的概率为 $p(0 < p < 1)$，射击进行到击中目标两次为止。
- 设 X 表示第一次击中目标需要的次数；Y 表示到击中两次为止所用的射击次数。
- 求条件概率：已知射手用了 10 次完成任务，问射手于第 5 次首次击中的概率？
- 更一般地，已知射手用了 10 次完成任务，问射手于第 m 次首次击中的概率？

$$P(X = m|Y = 10) = \frac{P(X = m, Y = 10)}{P(Y = 10)}, \quad m = 1, 2, \ldots, 9$$

- 作为 m 的函数，这是一个“条件分布”。

例：连续型随机变量的条件概率

- 设 (X, Y) 服从单位圆内的二维均匀分布。
- 问：当 $X = \frac{3}{5}$ 时，$Y \in [a, b]$ 的概率？
- 注意：$P(X = \frac{3}{5}) = 0$，但上述条件概率是有意义的。

- 已知 $X=\frac{3}{5}$ 后，Y 取值于 $(-\frac{4}{5},\frac{4}{5})$，且均匀取值。对 $-\frac{4}{5}\le a<b\le\frac{4}{5}$，

$$P(Y\in(a,b)|X=\frac{4}{5})=\frac{b-a}{2\cdot\frac{4}{5}}$$

- 在 $X=\frac{4}{5}$ 条件下，Y 服从 $(-\frac{4}{5},\frac{4}{5})$ 上的“条件分布”。

条件分布

- 设 X,Y 是两个随机变量，给定实数 y，如果 $P(Y=y)>0$，则称 x 的函数

$$P(X\le x|Y=y)$$

为 $Y=y$ 条件下 X 的**条件分布函数**，记为 $F_{X|Y}(x|y)$。

$$F_{X|Y}(x|y)=\frac{P(X\le x,Y=y)}{P(Y=y)}$$

- 但是当 $P(Y=y)=0$ 时，上述的条件概率也是有意义的。
- 用极限来定义这样的条件概率。

- **定义 5.1** 设对任意 $\varepsilon>0$,

$$P(y-\varepsilon<Y\le y+\varepsilon)>0.$$

若极限

$$\lim_{\varepsilon\to0}P(X\le x|y-\varepsilon<Y\le y+\varepsilon)$$

存在，则称此极限为 $Y=y$ 的条件下 X 的**条件分布函数**，记作 $P(X\le x|Y=y)$ 或 $F_{X|Y}(x|y)$，即

$$F_{X|Y}(x|y)=\lim_{\varepsilon\to0}P(X\le x|y-\varepsilon<Y\le y+\varepsilon)\tag{5.2}$$

- 条件分布函数也是分布函数：单调上升右连续，$-\infty$ 极限为 0，$+\infty$ 极限为 1。
- 条件分布与联合分布有关，被联合分布决定。

离散型的条件分布

- 设 (X,Y) 是二维离散型随机向量，

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

  其中 $P(Y=y_j)>0, j=1,2,\dots$。

- 则在 $Y=y_j$ 的条件下 X 的条件分布为

$$\begin{aligned}P(X=x_i|Y=y_j)=&\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}\\=&\frac{p_{ij}}{\sum_k p_{kj}}\quad (i=1,2,\dots)\end{aligned}$$

- **例 5.1(续)** 来求给定 $Y=n$ 后 X 的条件分布。

- (X,Y) 的联合分布为

$$\begin{aligned}&P(X=m,Y=n)\\=&P(\text{第 }1\sim\text{ 第 }m-1\text{ 次未击中，第 }m\text{ 次击中，}\\&\text{第 }m+1\sim\text{ 第 }n-1\text{ 次未击中，第 }n\text{ 次击中})\\=&q^{m-1}pq^{n-m-1}p\\=&p^2q^{n-2},\quad (n=2,3,\dots,m=1,2,\dots,n-1)\end{aligned}$$

- 其中 $q=1-p$。

- 注意 X,Y 不独立，因为其概率大于零的区域不是矩形的。

- 于是 X 的边缘分布为

$$\begin{aligned} P(X=m) &= \sum_{n=m+1}^{\infty} P(X=m, Y=n) \\ &= \sum_{n=m+1}^{\infty} p^2 q^{n-2} = p^2 q^{m-1} \sum_{n=m+1}^{\infty} q^{n-m-1} \\ &= p^2 q^{m-1} \sum_{k=0}^{\infty} q^k \qquad (k = n-m-1) \\ &= p^2 q^{m-1} \frac{1}{1-q} = p q^{m-1} \quad (m = 1, 2, \dots) \end{aligned}$$

- Y 的边缘分布为

$$\begin{aligned} P(Y=n) &= \sum_{m=1}^{n-1} P(X=m, Y=n) \\ &= \sum_{m=1}^{n-1} p^2 q^{n-2} \\ &= (n-1) p^2 q^{n-2} \quad (n = 2, 3, \dots) \end{aligned}$$

- 已知 $Y=n$ 时 X 的条件分布为

$$P(X=m|Y=n) = \begin{cases} \frac{1}{n-1} & n \geq 2, m = 1, 2, \dots, n-1 \\ 0 & \text{其它} \end{cases}$$

  这是取值于 $\{1, 2, \dots, n-1\}$ 的“离散均匀分布”。

- 类似地，已知 $X=m$ 时 Y 的条件分布为

$$P(Y=n|X=m) = \begin{cases} p q^{n-m-1} & n = m+1, m+2, \dots \\ 0 & \text{其它} \end{cases}$$

连续型条件分布

- 设随机向量 (X,Y) 有联合分布函数 $F(x,y)$，联合密度 $p(x,y)$。
- 对 $p(x,y)$ 加一些条件（实际中通常可以满足）后给出连续型条件分布的表达式。
- 联合分布函数 $F(x,y)$ 可表示为

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

- Y 的边缘分布密度为

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

分布函数为

$$F_Y(y)=\int_{-\infty}^{y}p_Y(u)du$$

- 于是在 $Y=y$ 的条件下，若 $p_Y(y)>0$, X 条件分布函数为

$$\begin{aligned}
&F_{X|Y}(x|y)\\
=&\lim_{\varepsilon\to 0}P(X\le x|y-\varepsilon<Y\le y+\varepsilon)\\
=&\lim_{\varepsilon\to 0}\frac{P(X\le x,\ y-\varepsilon<Y\le y+\varepsilon)}{P(y-\varepsilon<Y\le y+\varepsilon)}\\
=&\lim_{\varepsilon\to 0}\frac{F(x,y+\varepsilon)-F(x,y-\varepsilon)}{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)}\\
=&\frac{\lim\limits_{\varepsilon\to 0}\frac{F(x,y+\varepsilon)-F(x,y-\varepsilon)}{2\varepsilon}}{\lim\limits_{\varepsilon\to 0}\frac{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)}{2\varepsilon}} \qquad (*)\\
&(\text{ 若分子和分母两个极限存在})
\end{aligned}$$

- 若 $p_Y(u)$ 在 $u=y$ 处连续，则

$$\frac{dF_Y(y)}{dy}=p_Y(y)$$

从而

$$\lim_{\varepsilon\to 0}\frac{F_Y(y+\varepsilon)-F_Y(y-\varepsilon)}{2\varepsilon}=p_Y(y)$$

- 即这时 (*) 的分母为 $p_Y(y)$。

- 若 $\int_{-\infty}^{x}p(u,v)du$ 在 $v=y$ 处连续，则

$$\begin{aligned}\frac{\partial F(x,y)}{\partial y}&=\frac{\partial}{\partial y}\int_{-\infty}^{y}\int_{-\infty}^{x}p(u,v)dudv\\&=\int_{-\infty}^{x}p(u,y)du\end{aligned}$$

- 从而 (*) 的分子

$$\lim_{\varepsilon\to 0}\frac{F(x,y+\varepsilon)-F(x,y-\varepsilon)}{2\varepsilon}=\int_{-\infty}^{x}p(u,y)du$$

- 于是

$$\begin{aligned}F_{X|Y}(x|y)=&\frac{\int_{-\infty}^{x}p(u,y)du}{p_Y(y)}\\=&\int_{-\infty}^{x}\frac{p(u,y)}{p_Y(y)}du\qquad(\text{当 }p_Y(y)>0\text{ 时})\end{aligned}$$

- 称

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

为 $Y=y$ 的条件下 X 的**条件分布密度**，记作 $p_{X|Y}(x|y)$，即

$$p_{X|Y}(x|y)=\frac{p(x,y)}{p_Y(y)}\tag{5.4}$$

- 条件分布密度公式与离散型的条件概率分布公式类似。
- 要求 $p_Y(y) > 0$。
- **例 5.2** 设 (X, Y) 服从二维正态分布，联合密度为

$$\begin{aligned} & p(x,y) \\ =& \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \cdot \right. \\ & \left. \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right\} \end{aligned} \tag{1.10}$$

- X 的边缘分布为 $\mathrm{N}(\mu_1, \sigma_1^2)$，所以

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\Big\{ -\frac{1}{2\sigma_1^2}(x-\mu_1)^2 \Big\}$$

- 于是给定 $X = x$ 时 Y 的条件分布密度为

$$\begin{aligned} & p_{Y|X}(y|x) \\ =& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_2} \exp\big\{ y \text{ 的一元二次多项式} \big\} \end{aligned}$$

- 推导：记 $z = \frac{x-\mu_1}{\sigma_1}$, 则

$$\begin{aligned} & -\frac{1}{2(1-\rho^2)} \cdot \left[z^2 - 2\rho z \frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] + \frac{1}{2} z^2 \\ =& -\frac{1}{2(1-\rho^2)} \cdot \left[\frac{1}{\sigma_2^2} y^2 - 2\frac{\mu_2}{\sigma_2^2} y - 2\frac{\rho z}{\sigma_2} y \right] + \text{Const.} \\ =& -\frac{1}{2\sigma_2^2(1-\rho^2)} \cdot \big[y^2 - 2(\mu_2 + \rho z \sigma_2) y \big] + \text{Const.} \\ =& -\frac{1}{2\sigma_2^2(1-\rho^2)} \cdot \big[y - (\mu_2 + \rho\sigma_2 z) \big]^2 + \text{Const.} \end{aligned}$$

- 其中 Const. 表示与 y 无关的项。

- 所以

$$\begin{aligned} & P_{Y|X}(y|x) \\ =& \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_2}\cdot \\ & \exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\cdot[y-(\mu_2+\rho\sigma_2 z)]^2\right\} \end{aligned}$$

- 即

$$Y|X=x \sim \mathrm{N}\left(\mu_2+\rho\sigma_2\frac{x-\mu_1}{\sigma_1},\ \sigma_2^2(1-\rho^2)\right)$$

- 类似地，$Y=y$ 时 X 的条件分布为

$$X|Y=y \sim \mathrm{N}\left(\mu_1+\rho\sigma_1\frac{y-\mu_2}{\sigma_2},\ \sigma_1^2(1-\rho^2)\right)$$

4.5.2 条件期望

条件期望

- 条件分布也是分布，期望由分布决定，条件分布决定的期望叫做条件期望。
- **定义 5.2** 设 X,Y 是两个随机变量，有联合密度 $p(x,y)$, 设 $Y=y$ 的条件下 X 有条件分布密度 $p_{X|Y}(x|y)$，则

$$\int_{-\infty}^{\infty} x p_{X|Y}(x|y)dx$$

  叫做 $Y=y$ 的条件下 X 的**条件期望**, 记作 $E(X|Y=y)$。
- 要求上面的积分绝对收敛。
- 对离散型条件分布类似定义条件期望。

- 由连续型随机向量的条件密度公式 (5.4) 知

$$E(X|Y=y)=\frac{1}{p_Y(y)}\int_{-\infty}^{\infty}xp(x,y)dx \tag{5.5}$$

- $E(X|Y=y)$ 的意义：在 $Y=y$ 的条件下，X 取值的平均大小。

- $E(X|Y=y)$ 是 y 的函数，记为 $g(y)$。把 $g(Y)$ 记为 $E(X|Y)$，这是随机变量 Y 的函数。

- 性质：

$$\begin{aligned}E[E(X|Y)]=&E(g(Y))\\=&\int\limits_{\{y:p_Y(y)>0\}}g(y)p_Y(y)dy\\=&\int\limits_{\{y:p_Y(y)>0\}}E(X|Y=y)p_Y(y)dy\\=&E(X)\end{aligned} \tag{5.6}$$

- 公式 (5.6) 的一个用处是：有时 $E(X)$ 不易求，但是 $E(X|Y=y)$ 容易求，$E[g(Y)]=E[E(X|Y)]$ 也容易求，就可以用 (5.6) 求 $E(X)$。

- 可以看成全概公式的推广。

- **证** 当 $p_Y(y)=0$ 时

$$\int_{-\infty}^{\infty}xp(x,y)=0. \tag{*}$$

- 事实上，这时对任意 $A>0$,

$$\begin{aligned}\left|\int_{-A}^{A} xp(x,y)dx\right| \le & A\int_{-A}^{A} p(x,y)dx\\ \le & A\int_{-\infty}^{\infty} p(x,y)dx = Ap_Y(y) = 0\end{aligned}$$

  令 $A\to\infty$ 即可知 $p_Y(y)=0$ 时 (*) 式成立。

- 于是由 (5.5)

$$\begin{aligned}&\int\limits_{\{y:p_Y(y)>0\}} E(X|Y=y)p_Y(y)dy\\ =&\int\limits_{\{y:p_Y(y)>0\}}\left[\int_{-\infty}^{\infty} xp(x,y)dx\right]dy\\ =&\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} xp(x,y)dx\right]dy\\ =&\int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} p(x,y)dy\right]dx\\ =&\int_{-\infty}^{\infty} xp_X(x)dx = E(X)\end{aligned}$$

- 于是 (5.6) 成立。

- **例 5.3** 设 $U_1,U_2,\ldots$ 是独立同分布的随机变量列，共同分布为 $\mathrm{U}(0,1)$。

$$N=\min\left\{n:\sum_{i=1}^{n}U_i>1\right\}$$

  求 $E(N)$。

- 对任意 $x\in[0,1]$，令

$$\begin{aligned}N(x)=&\min\left\{n:\sum_{i=1}^{n}U_i>x\right\}\\ m(x)=&E[N(x)]\end{aligned}$$

- 由 (5.6) 有

$$m(x) = E\{E[N(x)|U_1]\} = \int_0^1 E[N(x)|U_1 = y]dy$$

- 这里 $N(x)$ 为离散分布而 U_1 为连续型分布，但

$$E[N(x)] = E\{E[N(x)|U_1]\}$$

仍成立。

- 其中

$$E[N(x)|U_1 = y] = \begin{cases} 1 & y > x \\ 1 + m(x-y) & y \le x \end{cases}$$

- 于是

$$\begin{aligned} m(x) &= \int_x^1 dy + \int_0^x [1 + m(x-y)]dy \\ &= 1 + \int_0^x m(u)du \end{aligned}$$

- 有微分方程

$$m'(x) = m(x)$$

- 于是

$$m(x) = ke^x$$

- 而 $m(0) = 1$, 所以 $k = 1$，$m(x) = e^x$。
- $E(N) = m(1) = e$。

- **例 5.4** 设某工厂每月电力供应服从 U(10, 30)（单位：万度）。电力需求服从 U(10, 20)。

- 如果电力足够，每 1 万度电可创造 30 万元利润。

- 如果电力不足，则不足部分另外解决，不足部分电力每 1 万度只能创造 10 万元利润。

- 求此工厂每月平均利润。

- **解** 设电力需求为 X(万度)，供应为 Y(万度)，工厂每月的利润为 R(万元)。

- 可以认为 X, Y 相互独立。

- 利润为

$$R = \begin{cases} 30X & X \le Y \\ 30Y + 10(X - Y) & X > Y \end{cases}$$

- 当 $20 \le y \le 30$ 时，一定有 $X \le y$，

$$\begin{aligned} E(R|Y = y) =& E(30X) = 30E(X) \\ =& 30 \cdot \frac{10 + 20}{2} = 450(\text{万度}) \end{aligned}$$

- 当 $10 \le y < 20$ 时，

$$
\begin{aligned}
& E(R|Y=y) \\
=& E[RI_{\{X\le y\}}|Y=y] + E[RI_{\{X>y\}}|Y=y] \\
=& E[30XI_{\{X\le y\}}|Y=y] + E[(30y+10(X-y))I_{\{X>y\}}|Y=y] \\
=& E[30XI_{\{X\le y\}}] + E[(30y+10(X-y))I_{\{X>y\}}] \\
=& \int_{10}^{y} 30x\cdot\frac{1}{10}dx + \int_{y}^{20}(30y+10(x-y))\cdot\frac{1}{10}dx \\
=& 50+40y-y^2
\end{aligned}
$$

- 由 (5.6)

$$
\begin{aligned}
E(R) =& E[E(R|Y)] \\
=& \int_{10}^{30} E(R|Y=y)p_Y(y)dy \\
=& \int_{10}^{20}(50+40y-y^2)\cdot\frac{1}{20}dy + \int_{20}^{30}450\cdot\frac{1}{20}dy \\
\approx& 433(\text{万元})
\end{aligned}
$$

- 即工厂每月的平均利润约为 433 万元。

随机向量的条件分布和条件期望

- 设 $\boldsymbol{X}=(X_1,X_2,\ldots,X_m)$ 和 $\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)$ 是两个随机向量。
- 设 $\forall\varepsilon>0$,

$$
\begin{aligned}
& P(y_1-\varepsilon<Y_1\le y_1+\varepsilon, y_2-\varepsilon<Y_2\le y_2+\varepsilon,\ldots, \\
& \quad y_n-\varepsilon<Y_n\le y_n+\varepsilon)>0
\end{aligned}
$$

- 称

$$\begin{aligned}&F_{\boldsymbol{X}|\boldsymbol{Y}}(x_1,x_2,\ldots,x_m|y_1,y_2,\ldots,y_n)\\=&\lim_{\varepsilon\to 0}P(X_1\le x_1,X_2\le x_2,\ldots,X_m\le x_m|\\&y_1-\varepsilon<Y_1\le y_1+\varepsilon,y_2-\varepsilon<Y_2\le y_2+\varepsilon,\ldots,\\&y_n-\varepsilon<Y_n\le y_n+\varepsilon)\end{aligned}$$

(若极限存在) 为 $\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)$ 条件下 $\boldsymbol{X}$ 的条件分布函数。

- 如果 $(\boldsymbol{X},\boldsymbol{Y})$ 有联合密度 $p(x_1,x_2,\ldots,x_m,y_1,y_2,\ldots,y_n)$, 在相当广泛的条件下

$$\begin{aligned}&F_{\boldsymbol{X}|\boldsymbol{Y}}(x_1,x_2,\ldots,x_m|y_1,y_2,\ldots,y_n)\\=&\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_m}\frac{p(u_1,u_2,\ldots,u_m,y_1,y_2,\ldots,y_n)}{p_Y(y_1,y_2,\ldots,y_n)}\\&du_1du_2\ldots du_m\end{aligned}$$

- 称上式中的被积函数为 $\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)$ 的条件下 $\boldsymbol{X}$ 的条件分布密度。

- 在 $\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)$ 条件下 X_i 的条件分布密度是上述条件分布密度的一个边缘密度，为

$$\begin{aligned}p_i(u_i|\boldsymbol{y})=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{p(u_1,u_2,\ldots,u_m,y_1,y_2,\ldots,y_n)}{p_Y(y_1,y_2,\ldots,y_n)}\\&du_1du_2\ldots u_{i-1}u_{i+1}\ldots du_n\\&(\text{当 } p_Y(y_1,y_2,\ldots,y_n)>0 \text{ 时})\end{aligned}$$

- 在 $\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)$ 条件下 X_i 的条件期望为

$$\begin{aligned}&E(X_i|Y=(y_1,y_2,\ldots,y_n))\\=&\int_{-\infty}^{\infty}up_i(u|\boldsymbol{y})du\quad(i=1,2,\ldots,m)\end{aligned}$$

- 定义

$$\begin{aligned}&E(\boldsymbol{X}|\boldsymbol{Y}=(y_1,y_2,\ldots,y_n))\\=&(E(X_1|\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)),E(X_2|\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)),\\&\ldots,E(X_m|\boldsymbol{Y}=(y_1,y_2,\ldots,y_n)))\end{aligned}$$

4.5.3 最佳预测与条件期望

最佳预测

- 考虑用 X 的观测值去预测 Y 的值的问题。
- 求函数 $\psi(\cdot)$ 使得 $\psi(X)$ 与 Y 最接近。
- 什么是“最接近”?
- 一个准则是“均方误差最小准则”，求 $\psi(\cdot)$ 使

$$E[Y-\psi(X)]^2$$

最小。

最佳预测与条件期望

- **定理 5.1** 设 (X,Y) 有联合密度 $p(x,y)$, $E(Y^2)$ 存在，令

$$\phi(x)=\begin{cases}E(Y|X=x) & \text{当}p_X(x)>0\\0 & \text{当}p_X(x)=0\end{cases}$$

记 $E(Y|X)=\phi(X)$。则

$$E[Y-E(Y|X)]^2=\min_{\psi}E[Y-\psi(X)]^2 \tag{5.12}$$

- (X,Y) 为离散型随机向量时也有同样结果。
- 即用给定 X 后的 Y 的条件期望去预测 Y，在所有的 X 的函数作的预测中均方误差最小。

- **证** 不妨设 $E[\psi(X)]^2$ 存在。易知

$$\begin{aligned}&E[Y-\psi(X)]^2\\=&E\{[Y-\phi(X)]+[\phi(X)-\psi(X)]\}^2\\=&E[Y-\phi(X)]^2+E[\phi(X)-\psi(X)]^2\\&+2E\{[Y-\phi(X)][\phi(X)-\psi(X)]\}\end{aligned}$$

- 来证其中交叉项等于 0。只要证

$$E\{Y[\phi(X)-\psi(X)]\}=E\{\phi(X)[\phi(X)-\psi(X)]\} \tag{5.13}$$

- 对 x 使得 $p_X(x)=0$，有 $\int_{-\infty}^{\infty}p(x,y)dy=0$, 从而

$$\int_{-\infty}^{\infty}yp(x,y)dy=0=\phi(x)p_X(x)$$

- 对 x 使得 $p_X(x)>0$，有

$$\begin{aligned}&\phi(x)=E(Y|X=x)\\=&\int_{-\infty}^{\infty}y\frac{p(x,y)}{p_X(x)}dy\end{aligned}$$

所以

$$\int_{-\infty}^{\infty}yp(x,y)dy=\phi(x)p_X(x)$$

- 总之，对任意 x 都有

$$\int_{-\infty}^{\infty}yp(x,y)dy=\phi(x)p_X(x)$$

- 于是

$$\begin{aligned}
&E\left\{Y\left[\phi(X)-\psi(X)\right]\right\}\\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}y[\phi(x)-\psi(x)]p(x,y)dxdy\\
=&\int_{-\infty}^{\infty}[\phi(x)-\psi(x)]\int_{-\infty}^{\infty}yp(x,y)dydx\\
=&\int_{-\infty}^{\infty}[\phi(x)-\psi(x)]\phi(x)p_X(x)dx\\
=&E\left\{\phi(X)\left[\phi(X)-\psi(X)\right]\right\}
\end{aligned}$$

- 即交叉项等于 0，

$$E[Y-\psi(X)]^2\geq E[Y-\phi(X)]^2$$

- **例 5.5** 设 (X,Y) 服从二元正态分布，参数为 $(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$。
- 由例 5.2 知 $Y|X=x$ 服从 $\mathrm{N}\left(\mu_2+\rho\sigma_2\frac{x-\mu_1}{\sigma_1},\sigma_2^2(1-\rho^2)\right)$。
- 于是

$$E(Y|X=x)=\int_{-\infty}^{\infty}yp_{Y|X}(y|x)dy=\mu_2+\rho\sigma_2\frac{x-\mu_1}{\sigma_1}$$

- 所以用

$$\phi(X)=\mu_2+\rho\sigma_2\frac{X-\mu_1}{\sigma_1}$$

  预测 Y，在用 X 的函数作的预测中均方误差最小。

多元最佳预测

- 设 $X_1,X_2,\ldots,X_m,Y$ 是 $m+1$ 个随机变量，求函数 $\psi(x_1,x_2,\ldots,x_m)$ 使得用

$$\psi(X_1,X_2,\ldots,X_m)$$

  去预测 Y 的均方误差最小。

- 解为

$$\phi(x_1, x_2, \ldots, x_m) = E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$$

- 即

$$E\left[Y - \phi(X_1, X_2, \ldots, X_m)\right]^2 = \min_{\psi} E\left[Y - \psi(X_1, X_2, \ldots, X_m)\right]^2$$

- 记

$$\phi(X_1, X_2, \ldots, X_n) = E(Y|X_1, X_2, \ldots, X_n)$$

4.6 大数定律和中心极限定理

4.6.1 大数定律

大数定律和中心极限定理

- 大数定律是概率的频率定义的理论基础。
- 中心极限定理是现代统计推断中一个重要基础理论。
- **定义 6.1** 称随机变量列 $X_1, X_2, \ldots, X_n, \ldots$ 是相互独立的，如果对任何 $n \geq 1$，$X_1, X_2, \ldots, X_n$ 是相互独立的，此时，如果所有的 X_i 又有相同的分布函数，则称 $X_1, X_2, \ldots, X_n, \ldots$ 是独立同分布的随机变量序列，记为 iid。

大数定律

- **定理 6.1(大数定律)** 设 $X_1, X_2, \ldots, X_n, \ldots$ 是独立同分布的随机变量列，且 $E(X_1)$, $D(X_1)$ 存在，则对任何 $\varepsilon > 0$，有

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - E(X_1)\right| \geq \varepsilon\right) = 0 \tag{6.1}$$

其中

$$S_n = X_1 + X_2 + \cdots + X_n$$

- $\frac{S_n}{n}$ 是 n 个观测的算术平均值，定理指出，重复观测个数 n 充分大时，算术平均值无限逼近理论期望值。

- **证** 利用切比雪夫不等式

$$P\left(\left|\frac{S_n}{n}-E(X_1)\right|\geq\varepsilon\right)\leq\frac{1}{\varepsilon^2}D\left(\frac{S_n}{n}\right)$$

- 而

$$\begin{aligned}D\left(\frac{S_n}{n}\right)=&\frac{1}{n^2}D(S_n)\\=&\frac{1}{n^2}D(X_1+X_2+\cdots+X_n)\\=&\frac{1}{n^2}D(X_1)+D(X_2)+\cdots+D(X_n)\\=&\frac{D(X_1)}{n}\end{aligned}$$

- 所以

$$P\left(\left|\frac{S_n}{n}-E(X_1)\right|\geq\varepsilon\right)\leq\frac{D(X_1)}{n\varepsilon^2}\to 0\quad(n\to\infty)$$

强大数定律

- 更多数学讨论可以证明：只要 $E(X_1)$ 存在，不管 $D(X_1)$ 是否存在，则（6.1）就成立，而且成立比 (6.1) 更强的结论

$$P\left(\lim_{n\to\infty}\frac{S_n}{n}=E(X_1)\right)=1 \tag{6.2}$$

- 符合 (6.1) 的同分布的随机变量列 $X_1,X_2,\ldots,X_n,\ldots$ 叫做服从**大数定律**（或**弱大数定律**）。

- 符合 (6.2) 的同分布的随机变量列 $X_1,X_2,\ldots,X_n,\ldots$ 叫做服从**强大数定律**。

- (6.1) 那样的概率极限叫做**依概率收敛**。

- 设 $\xi_n, n = 1, 2, \ldots$ 为随机变量序列，ξ 为随机变量（可以为常数），若 $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|\xi_n - \xi| > \varepsilon\right) = 0$$

则称 ξ_n **依概率收敛**到 ξ，记作

$$\xi_n \xrightarrow{\Pr} \xi, \quad (n \to \infty)$$

- (6.2) 那样的概率极限叫做**以概率 1 收敛**，或 **a.s. 收敛**。

- 若

$$P\left(\lim_{n\to\infty} \xi_n = \xi\right) = 1$$

则称 ξ_n 以**概率 1 收敛**到 ξ，记作

$$\lim_{n\to\infty} \xi_n = \xi, \text{ a.s.}$$

概率的频率定义的理论依据

- **例 6.1** 设条件 S 下事件 A 的概率是 p。将条件 S 独立地重复 n 次，设 A 出现的次数是 μ。

- 令

$$X_i = \begin{cases} 1 & \text{当第 } i \text{ 次重复条件 } S \text{ 时 } A \text{ 出现} \\ 0 & \text{当第 } i \text{ 次重复条件 } S \text{ 时 } A \text{ 不出现} \end{cases}$$

- 则 $S_n = X_1 + X_2 + \cdots + X_n = \mu$, $E(X_1) = P(X_1 = 1) = p$。

- 由 (6.1),

$$\lim_{n\to\infty} P\left(\left|\frac{\mu}{n} - p\right| \geq \varepsilon\right) = 0$$

- 即独立重复次数 n 增加时 A 发生的频率 $\frac{\mu}{n}$ 与概率 p 可以任意地接近。

4.6.2 中心极限定理

中心极限定理

- **定理 6.2(中心极限定理)** 设 $X_1, X_2, \ldots, X_n, \ldots$ 是独立同分布随机变量列，而且 $E(X_1), D(X_1)$ 存在，$D(X_1) > 0$，则对一切实数 $a < b$，有

$$\lim_{n\to\infty} P\left(a < \frac{S_n - nE(X_1)}{\sqrt{nD(X_1)}} < b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \int_a^b \varphi(u)du$$

- 其中 $\varphi(\cdot)$ 是标准正态分布密度。

- 记

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

  (6.3) 也可以写成

$$\lim_{n\to\infty} P\left(a < \frac{\bar{X}_n - E(X_1)}{\sqrt{D(X_1)/n}} < b\right) = \int_a^b \varphi(u)du$$

- 中心极限定理说明当 n 很大时，X_n 的近似分布为

$$\mathrm{N}\left(E(X_1), \frac{D(X_1)}{n}\right)$$

  而不管 X_1 原来是什么分布。

- 演示。

4.6.3 一般情形下的大数定律和中心极限定理

更一般随机变量列

- 考虑不一定独立，不一定同分布的随机变量列。
- 设 $X_1, X_2, \ldots, X_n, \ldots$ 为随机变量列，

$$S_n = \sum_{i=1}^{n} X_i \quad (n \geq 1)$$

- (6.1) 式改为

$$\lim_{n\to\infty} P\left(\left|\frac{S_n - E(S_n)}{n}\right| \geq \varepsilon\right) = 0 \quad (\forall \varepsilon > 0) \tag{6.4}$$

成立时称序列 $\{X_n, n \geq 1\}$ 服从大数定律。

- (6.2) 式改为

$$P\left(\lim_{n\to\infty} \frac{S_n - E(S_n)}{n} = 0\right) = 1 \tag{6.5}$$

成立时称序列 $\{X_n, n \geq 1\}$ 服从强大数定律。

更一般的大数定律

- (6.3) 式改为

$$\lim_{n\to\infty} P\left(a < \frac{S_n - E(S_n)}{\sqrt{D(S_n)}} < b\right) = \int_a^b \varphi(u)du \tag{6.6}$$

（6.6）成立时，称中心极限定理对序列序列 $\{X_n, n \geq 1\}$ 成立。

- 若 $X_1, X_2, \ldots$ 是相互独立的随机变量列，方差 $D(X_n)$, $n \geq 1$ 有界，由切比雪夫不等式易知大数定律 (6.4) 成立。
- 比 $D(X_n), n \geq 1$ 有界稍弱的条件：
- **定理 6.3(Kolmogorov A N)** 设 $X_1, X_2, \ldots$ 是相互独立的随机变量列，若

$$\sum_{n=1}^{\infty} \frac{D(X_n)}{n^2} < \infty$$

则该序列服从强大数定律。

Liapunov 中心极限定理

- **定理 6.4(Liapunov, A M)** 设 $X_1, X_2, \dots$ 是相互独立的随机变量列，$\sigma_i^2 = D(X_i)$ 存在 $(i \geq 1)$，

$$B_n = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{\frac{1}{2}} \quad (n \geq 1)$$

满足条件

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{i=1}^{n} E|X_i - E(X_i)|^3 = 0 \tag{6.7}$$

则对一切 $a < b$ 有

$$\lim_{n\to\infty} P\left(a < \frac{S_n - E(S_n)}{B_n} < b \right) = \int_a^b \varphi(u)du$$

- 定理 6.4 表明在相当一般的条件下独立随机变量和近似服从正态分布。

4.6.4 中心极限定理的例子

二项分布的中心极限定理

- 对独立重复试验，设成功概率为 p，

$$X_i = \begin{cases} 1 & \text{第 } i \text{ 次试验成功} \\ 0 & \text{第 } i \text{ 次试验失败} \end{cases} \quad (i = 1, 2, \dots)$$

则 $X_1, X_2, \dots$ 独立同两点分布 $\mathrm{b}(1, p)$。

- $S_n = X_1 + X_2 + \cdots + X_n$ 为 n 次独立重复试验中总成功次数，服从 $\mathrm{B}(n, p)$ 分布。

- 由中心极限定理，对整数 $0 \le k_1 \le k_2 \le n$，

$$\begin{aligned} & P(k_1 \le S_n \le k_2) \\ =& P(k_1 - 0.5 < S_n < k_2 + 0.5) \\ =& P\left(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) \\ \approx& \Phi\left(\frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

- **例 6.2** 某座桥，最大负荷重量服从 $\mathrm{N}(300, 40)$(吨)。可能同时多辆车在桥上。车的平均重量为 5 吨，方差为 2。
- 问：为了保证桥不被压塌的概率不小于 0.99997, 最多允许多少辆车在桥上？
- **解** 用 Y 表示桥的最大负荷重量，若有 M 辆车在桥上，第 i 辆车重量为 $X_i (i = 1, 2, \ldots, M)$。
- M 辆车总重量为

$$S_n = X_1 + X_2 + \cdots + X_n$$

- 可以认为 $Y, X_1, X_2, \ldots, X_M$ 相互独立，

$$\begin{aligned} E(X_i) =& 5 \\ D(X_i) =& 2 \quad (i = 1, 2, \ldots, M) \end{aligned}$$

- 求最大的使得

$$R = P(S_M < Y) \ge 0.99997$$

  的 M。
- M 相当大，所以 S_M 近似服从

$$\mathrm{N}(M\mu_1, M\sigma_1^2)$$

  ($\mu_1 = 5$, $\sigma_1^2 = 2$)

- S_M 与 Y 独立，都服从（近似服从）正态分布，所以 $Z = S_M - Y$ 也近似服从正态分布

$$\mathrm{N}(M\mu_1 - 300, M\sigma_1^2 + 40)$$

- 为使 $P(Z < 0) \geq 0.99997$, 即

$$P(Z < 0) = \Phi\left(\frac{0 - (M\mu_1 - 300)}{\sqrt{M\sigma_1^2 + 40}}\right) \geq 0.99997$$

- 只要

$$\frac{0 - (M\mu_1 - 300)}{\sqrt{M\sigma_1^2 + 40}} \geq \Phi^{-1}(0.99997) = 4$$

- 即

$$\frac{-(5M - 300)}{\sqrt{2M + 40}} \geq 4$$

- 即

$$300 - 5M \geq 4\sqrt{2M + 40}$$

- 等价于

$$\begin{cases} 300 - 5M \geq 0 \\ 25M^2 - 3032M + 89360 \geq 0 \end{cases}$$

- 即

$$\begin{cases} M \leq 60 \\ M \leq 50.5 \text{ 或} M \geq 70.78 \end{cases}$$

- 即 $M \leq 50.5$, 所以最多允许 50 辆车同时在桥上。

第五章　统计估值

5.1　总体与样本

随机变量模型与数据

- 随机变量是刻画随机结果及随机结果的取值分布情况的数学模型。
- 在实际问题中，假定某个随机结果的模型为随机变量 X，需要求 X 的分布（分布函数、分布密度或概率函数），至少要求其数字特征（期望、方差等）。
- 还可能需要回答 X 的有关问题。
- 这些需要数据的支持。
- 钢铁厂一天生产了 10000 根 16Mn 型钢筋。强度小于 52kg/mm^2 的算次品。
- 如何求这批产品的次品率 p?
- 检验所有 10000 根不可能：时间、费用、或破坏。
- 概率统计模型：设随机取一根，结果用随机变量 X 表示：

$$X=\begin{cases}1 & \text{是次品}\\ 0 & \text{不是次品}\end{cases}$$

$$P(X=1)=p$$

- 抽取少量样品来估计 p。

- 灯泡厂生产灯泡，寿命是随机的。求一批灯泡的平均寿命和寿命的分布情况。
- 不能全部检验：破坏性。
- 设随机抽取一只灯泡的寿命为随机变量 X。设 $X \sim \mathrm{Exp}(\lambda)$。求 λ 可以回答平均寿命以及寿命分布问题。

随机抽样法

- 从要研究的对象的全体中抽取一小部分来进行观察和研究，从而对整体进行推断。
- 重要意义：普查方法经常不可行，因为人力、物力、时间限制，或破坏性试验。
- **例 1.1（续）** 从 10000 根中抽取 50 根，对这 50 根进行检验。用 50 根的次品率作为所有全体的次品率的估计。
- 为什么这样是科学的？
- 随机抽样法包括：
  - 如何抽样，抽多少，怎样抽取；
  - 得到抽样结果（一批数据）后如何进行数据分析，进行统计推断。

总体

- 把所研究的对象的全体称为**总体**。
- 把总体中每一个基本单位称为**个体**。
- 主要关心每个个体的某一特性值（即数量指标，如钢筋的强度、灯泡的寿命）机器在总体中的分布情况（如钢筋强度在 $50\mathrm{kg/mm^2}$ 到 $60\mathrm{kg/mm^2}$ 之间的在 10000 根钢筋中所占的比例，灯泡寿命在 1000 小时到 2000 小时之间的占全天生产的灯泡的比例）。

- 要考察总体中个体特性值的分布规律，可以将个体特性值看成一个随机变量 X，代表从总体中随机抽取一个个体的特性值，其概率分布就可以体现总体中个体特性值的分布规律。
- 因为只关心总体的某个特性值，所以把总体认作是其特性值的随机变量 X。

样本

- 在一个总体（如 10000 根钢筋，或 10000 根钢筋的强度）X 中，抽取 n 个个体 $X_1, X_2, \ldots, X_n$, 这 n 个个体 $X_1, X_2, \ldots, X_n$ 称为总体 X 的一个容量为 n 的 **样本** (或叫**子样**)，也称 n 为**样本量**。
- 由于 $X_1, X_2, \ldots, X_n$ 是从总体 X 随机抽取出来的可能结果，可以看成是 n 个随机变量。
- 在一次抽取之后，又可以看成是 n 个具体的数值，称为**样本值**, 在使用这个意义时记为小写的 $x_1, x_2, \ldots, x_n$。
- 有时大小写也会混用。

样本的代表性

- 样本值应该对总体具有代表性。
- 如果样本 $X_1, X_2, \ldots, X_n$ 是相互独立的而且与总体 X 具有相同的概率分布，则称其为**简单随机样本**。
- 有放回逐次随机抽样法可以得到简单随机样本。当总体个数很大时，无放回地逐次随机抽样也可以认为是得到简单随机样本。
- 对总体 X 进行多次独立的重复观测，可以认为是得到简单随机样本。

总体与样本的数学模型

- 总体就是一个随机变量 X，我们关心其分布。

- 样本就是 n 个相互独立且与 X 有相同概率分布的随机变量 $X_1, X_2, \ldots, X_n$。
- 每一次具体抽样，所得的样本的值就是这 n 个随机变量的值（样本值），用 $x_1, x_2, \ldots, x_n$ 表示。
- **定义 1.1** 称随机变量 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的容量为 n 的 (简单随机) 样本，如果 $X_1, X_2, \ldots, X_n$ 相互独立，而且每个 X_i 与 X 有相同的概率分布。这时，若 X 有分布密度 $p(x)$，则称 $X_1, X_2, \ldots, X_n$ 是来自总体 $p(x)$ 的样本。
- **定理 1.1** 若 $X_1, X_2, \ldots, X_n$ 是来自总体 $p(x)$ 的样本，则 $(X_1, X_2, \ldots, X_n)$ 有联合密度

$$p(x_1)p(x_2)\ldots p(x_n).$$

5.2 分布函数与分布密度的估计

5.2.1 分布函数和分位数估计

经验分布函数

- 描述随机变量分布，可以用分布函数、密度函数或概率函数。
- 给定样本值 $x_1, x_2, \ldots, x_n$，如何估计分布函数 $F(x)$?
- 注意到

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

联系概率的频率含义以及简单随机样本可以看成是独立重复试验结果，用 $x_1, x_2, \ldots, x_n$ 中小于等于 x 的比例估计概率 $F(x)$。

- **定义 2.1** 设 $x_1, x_2, \ldots, x_n$ 是 X 的样本，称 x 的函数

$$F_n(x) = \frac{\nu_n}{n}$$

为 X 的经验分布函数，其中 ν_n 表示 $x_1, x_2, \ldots, x_n$ 中小于等于 x 的个数。

次序统计量

- 将样本值 $x_1, x_2, \ldots, x_n$ 从小到大排列后记为

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

  $x_{(i)}$ 叫做样本的第 i 个次序统计量 $(i = 1, 2, \ldots, n)$。

经验分布函数的阶梯函数表示

- 经验分布函数是一个阶梯函数：
- 若 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 两两不同，易见

$$F_n(x) = \begin{cases} 0 & x < x_{(1)} \\ \frac{k}{n} & x_{(k)} \le x < x_{(k+1)} \quad (k = 1, 2, \ldots, n-1) \\ 1 & x \ge x_{(n)} \end{cases}$$

  这是仅在每个 $x_{(i)}$ 处向上跳跃 $\frac{1}{n}$ 的阶梯函数，每一段在左端点连续，右端点不连续。

- 如果某个 $x_{(j)}$ 有 m 个相同的样本值，则 $F_n(x)$ 在 $x_{(j)}$ 处向上跳跃 $\frac{m}{n}$。

经验分布函数的 (强) 相合性

- 根据大数定律和强大数定律，对固定的 x，只要 n 相当大，$F_n(x)$ 与 $F(x)$ 很接近。
- 给定 x 后可以定义新的随机变量

$$Y_i = \begin{cases} 1 & X_i \le x \\ 0 & X_i > x \end{cases} \quad (i = 1, 2, \ldots, n)$$

- 则

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- $Y_1, Y_2, \ldots, Y_n$ 相互独立同 $\mathrm{b}(1, F(x))$ 二点分布, $E(Y_i) = F(x)$ $(i = 1, 2, \ldots, n)$。有

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = F(x)\right) = 1$$

经验分布函数的一致强相合性

- 对所有 x，当 $n \to \infty$ 时，$F_n(x)$ 一致地逼近 $F(x)$:
- **定理 (Glivenko-Cantelli)** 设

$$D_n = \sup_x |F_n(x) - F(x)|$$

则

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1$$

分位数估计

- 当 $F(x)$ 严格单调上升且连续时，反函数 $F^{-1}(p)$ $(p \in (0, 1))$ 为分位数函数。
- 一般地，$F(x)$ 的 p 分位数为满足

$$P(X \le x_p) \ge p, \quad P(X \ge x_p) \ge 1 - p$$

的数 x_p。

- 对样本值 $x_1, x_2, \ldots, x_n$,从小到大排列为次序统计量 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, 令

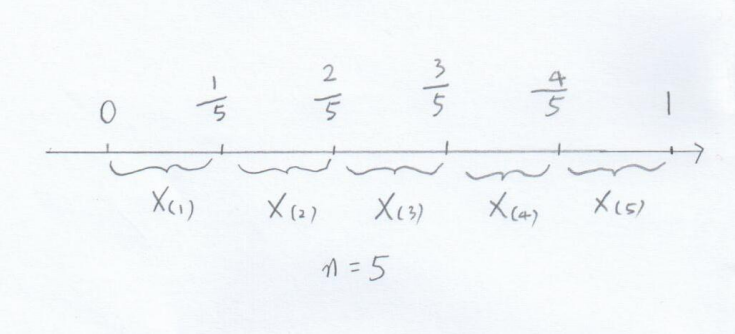
$$r = [pn] + 1$$

用 $x_{(r)}$ 估计 x_p。

- 当 $F(x) = p$ 至多有一个根时，x_p 存在唯一，

$$P\left(\lim_{n\to\infty} x_{(r)} = x_p\right) = 1$$

- 例如，$n=5$，则 $x_{(1)}, x_{(2)}, \ldots, x_{(5)}$ 代表的 p 范围如下图：



- **例 2.1** 自动装罐头的净重随机，额定 345 克。

- 从生产线随机抽取 10 个罐头：

$$344, 336, 345, 342, 340, 338, 344, 343, 344, 343$$

- 要求估计分布函数 $F(x)$ 和中位数。

- **解** 可以认为是简单随机样本。用经验分布函数 $F_n(x)$ 估计 $F(x)$。

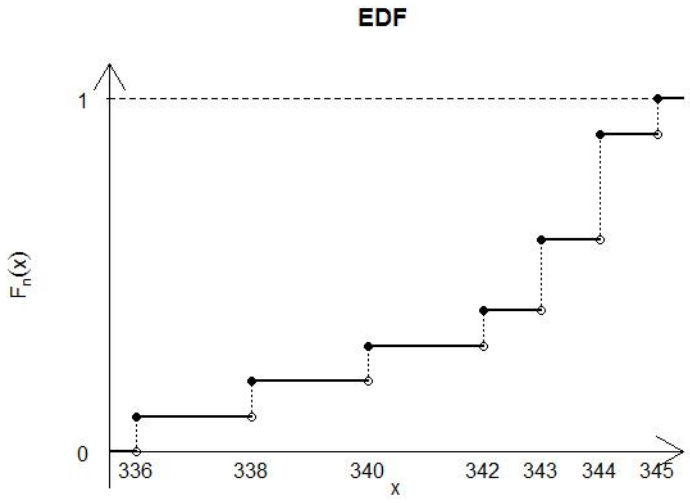
- 样本值从小到大排列得

$$336, 338, 340, 342, 343, 343, 344, 344, 344, 345$$

- 经验分布函数为

$$F_n(x)=\begin{cases}0 & x<336\\ \frac{1}{10} & 336\le x<338 \quad (\text{跳跃 } \frac{1}{10})\\ \frac{2}{10} & 338\le x<340 \quad (\text{跳跃 } \frac{1}{10})\\ \frac{3}{10} & 340\le x<342 \quad (\text{跳跃 } \frac{1}{10})\\ \frac{4}{10} & 342\le x<343 \quad (\text{跳跃 } \frac{1}{10})\\ \frac{6}{10} & 343\le x<344 \quad (\text{跳跃 } \frac{2}{10})\\ \frac{9}{10} & 344\le x<345 \quad (\text{跳跃 } \frac{3}{10})\\ 1 & x\ge 345 \quad (\text{跳跃 } \frac{1}{10})\end{cases}$$

- 经验分布函数图：



- 中位数估计：用次序统计量中间一个（奇数个时）或中间两个的平均（偶数个时），或者直接用 $x_{[0.5n]+1}$。即 $(x_{(5)}+x_{(6)})/2=(343+343)/2=343$, 或 $x_{([0.5\times 10]+1)}=x_{(6)}=343$。

5.2.2 直方图法

分布密度估计

- 对于连续型总体（随机变量），分布函数不如分布密度直观。高维时分布函数更不实用。
- 分布密度估计有直方图法、核估计法、最近邻估计法等。

直方图法

- 直方图法用阶梯函数估计密度函数。
- 把样本 $x_1, x_2, \ldots, x_n$ 从小到大排列为次序统计量 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 后，把数轴分成 m 个小区间，在每个小区间中用

$$\frac{\text{落入小区间的样本点个数}}{n} \cdot \frac{1}{\text{小区间长度}}$$

  估计该小区间的密度值 $p(x)$。

直方图估计的理论依据

- 设 $x_1, x_2, \ldots, x_n$ 为来自密度为 $p(x)$ 的总体的样本。
- 用 $R_n(a, b)$ 表示落入区间 $(a, b]$ 的样本点个数。
- 若 $(a, b]$ 很短，可认为 $x \in (a, b]$ 时 $p(x)$ 近似为常数，于是

$$P(a < X \leq b) = \int_a^b p(x)dx \approx p(x)(b - a)$$

- 用频率 $R_n(a, b)/n$ 估计概率 $P(a < X \leq b)$，有

$$\begin{aligned} \frac{R_n(a, b)}{n} \approx & p(x)(b - a) \\ p(x) \approx & \frac{R_n(a, b)}{n} \cdot \frac{1}{b - a}, \quad x \in (a, b] \end{aligned}$$

直方图法

- 为了估计密度函数 $p(x)$，设

  $$t_0 < t_1 < \cdots < t_m$$

  是 $m+1$ 个实数，通常假定 $t_i - t_{i-1} \equiv h > 0$ $(i = 1, 2, \ldots, m)$。

- 令

  $$p_n(x) = \begin{cases} \frac{R_n(t_{i-1}, t_i)}{n} \frac{1}{h} & \text{当} x \in (t_{i-1}, t_i],\ (i = 1, 2, \ldots, m) \\ 0 & \text{当} x \le t_0 \text{ 或} x > t_m \end{cases}$$

- 用 $p_n(x)$ 作为 $p(x)$ 的估计，这就是**直方图估计法**。

- 在第 i 个小区间 $(t_{i-1}, t_i]$ 用

  $$p_n(x) = \frac{R_n(t_{i-1}, t_i)}{n} \frac{1}{h}$$

  估计 $p(x)$，其图像是以底边为 h、高为 $p_n(x)$ 的矩形，面积为

  $$\begin{aligned} p_n(x) \cdot h =& \frac{R_n(t_{i-1}, t_i)}{n} \\ \approx& P(t_{i-1} < X \le t_i) = \int_{t_{i-1}}^{t_i} p(x)dx \\ \approx& p(x) \cdot h \end{aligned}$$

作直方图步骤

- 步骤一、对样本值 $x_1, x_2, \ldots, x_n$ 进行分组。找到最小值 $x_{(1)}$ 和最大值 $x_{(n)}$。

- 取 a 比 $x_{(1)}$ 略小，b 比 $x_{(n)}$ 略大，取适当分组数 m 把区间 $(a, b]$ m 等分，分点为

  $$t_i = a + i\frac{b-a}{m} \quad (i = 0, 1, \ldots, m)$$

记每组的区间长度为

$$h = \frac{b-a}{m}$$

- m 的取法：n 小的时候 m 也较小，n 很大时 m 可以随之增大也可以不再增大。应尽可能使得多数小区间中包含有样本的值。一般取 a, b, m 使得各分点的小数位数比观测值的小数位数多一位，这样不会有样本值落在小区间端点上。
- m 的建议公式之一：

$$m \approx 1 + 3.322 \lg n$$

- m 的建议取法二：

n	m
< 50	$5 \sim 6$
$50 \sim 100$	$6 \sim 10$
$100 \sim 250$	$7 \sim 12$
> 250	$10 \sim 20$

- 步骤二、决定了分点后，用唱票的方法数出样本值落入第 i 组 $(t_{i-1}, t_i]$ 中的个数，记为 $\nu_i (i = 1, 2, \ldots, m)$。
- 计算样本值落入各组的频率

$$f_i = \frac{\nu_n}{n} \quad (i = 1, 2, \ldots, m)$$

- 步骤三、做直方图。画坐标系，x 轴范围为 (a, b)，y 轴范围为 $[0, \max\limits_i f_i]$。
- 对 $i = 1, 2, \ldots, m$，以 x 轴的区间 $[t_{i-1}, t_i]$ 为底，以 f_i/h 为高做矩形框。这一系列矩形叫做直方图。
- 每个竖着的长方形的面积，近似代表 X 取值落入“底边”的概率。

- 某区间上 $p(x)$ 下的面积为 X 落入该区间的概率，直方图用频率估计概率，从而估计 $p(x)$。（示意图）
- **例 2.2** $n = 120$。
- $x_{(1)} = 0.64$, $x_{(n)} = 0.95$, $x_{(n)} - x_{(1)} = 0.31$。
- 把距离 0.31 略增大为 0.32 就容易分解因数。可取 $m = 16$, $h = 0.02$，$a = x_{(1)} - 0.005 = 0.635$, $b = x_{(n)} + 0.005 = 0.955$，各分点千分位都有 0.005，没有样本点落在区间端点上。
- 算出各区间端点，用唱票法统计出 $n_i, i = 1, 2, \ldots, m$。
- 作图。

直方图估计的相合性

- 若密度函数 $p(x)$ 在 $(-\infty, \infty)$ 上一致连续，对某个 $\delta > 0$，

$$\int_{-\infty}^{\infty} |x|^{\delta} p(x) dx$$

收敛，又小区间长度 h_n 满足

$$\lim_{n\to\infty} h_n = 0$$
$$h_n \geq \frac{(\ln n)^2}{n}$$

- 则有

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |p_n(x) - p(x)|\right) = 1.$$

(一致强相合)

5.2.3 核估计和最近邻估计介绍

Rosenblatt 估计

- 为了估计 $p(x)$，若 $p(x)$ 连续，可根据

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

- 其中 $F(x)$ 可以用经验分布函数 $F_n(x)$ 估计。有

$$\hat{p}_n(x) = \frac{1}{2h}\left[F_n(x+h) - F_n(x-h)\right], \quad x \in (-\infty, \infty)$$

- 这叫做 Rosenblatt 密度估计。

- 注意到

$$\begin{aligned} F_n(x+h) - F_n(x-h) =& \frac{R_n(x-h, x+h)}{n} \\ \hat{p}_n(x) =& \frac{R_n(x-h, x+h)}{n}\frac{1}{2h} \end{aligned}$$

- 所以 Rosenblatt 估计和直方图估计类似。

- 但是，直方图估计中小区间 $(t_{i-1}, t_i]$ 是固定的，而这里的小区间 $(x-h, x+h]$ 随自变量 x 而变化。

- 记

$$K_0(x) = \frac{1}{2} I_{[-1,1]}(x)$$

则

$$\begin{aligned} K_0\left(\frac{x - x_i}{h}\right) =& \frac{1}{2} I_{[x-h, x+h]}(x_i) \\ R_n(x-h, x+h) =& 2\sum_{i=1}^{n} K_0\left(\frac{x - x_i}{h}\right) \\ \hat{p}_n(x) =& \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{x - x_i}{h}\right) \end{aligned}$$

- $K_0(\cdot)$ 叫做一个“核函数”，$\hat{p}_n(x) \cdot 2h$ 是 x 邻域内的样本值百分比。

核密度估计

- Rosenblatt 估计采用了 x 邻域 $[x-h, x+h]$ 内的样本点数，x 邻域内样本点越多，密度估计越大。
- 推广：采纳样本点时，不一刀切，而是离 x 越近的样本点加权越大。
- **定义 2.2** 设 $K(x)$ 是非负函数且 $\int_{-\infty}^{\infty} K(x)dx = 1$, 则称 $K(x)$ 是核函数。此时称

$$\tilde{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

  为 $p(x)$ 的核估计。
- 核函数一般选为偶函数，且在正半轴单调下降（类似于正态分布曲线）。

常用核函数

-

$$\begin{aligned}
K_0(x) =& \frac{1}{2} I_{[-1,1]}(x) \\
K_1(x) =& \left(1-|x|^3\right)^3 I_{[-1,1]}(x) \\
K_2(x) =& \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\
K_3(x) =& \frac{1}{\pi\left(1+x^2\right)} \\
K_4(x) =& \frac{1}{2\pi}\left(\frac{\sin\frac{x}{2}}{\frac{x}{2}}\right)^2
\end{aligned}$$

- 演示。

核估计的相合性

- 当 n 无限增大且 $h = h_n$ 无限减小时，核估计 $\tilde{p}(x)$ 与密度 $p(x)$ 无限接近。

- 若 $p(x)$ 在 $(-\infty,\infty)$ 上一致连续，且

$$\lim_{n\to\infty} h_n = 0$$
$$\sum_{n=1}^{\infty} \exp\left\{-rnh_n^2\right\} < \infty \quad (\forall r>0)$$

又核函数 $K(x)$ 为有界变差函数，则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |\tilde{p}_n(x)-p(x)| = 0\right) = 1.$$

(一致强相合)

- 在核函数和 h 选取合适时核估计比直方图估计精度更高。

最近邻估计

- 核估计是 x 附近的样本点越多则密度估计值越大。
- 最近邻估计是固定 x 附近需要有的样本点数，令邻域区间长度可变。
- 取自然数 $K(n)$（n 为样本量），令

$$a_n(x) = \min\{t : t>0, R_n(x-t, x+t) \geq K(n)\}$$
$$p_n^*(x) = \frac{K(n)}{n}\frac{1}{2a_n(x)}$$

最近邻估计的相合性

- 适当条件下 $n\to\infty$ 时 $p_n^*(x)$ 与 $p(x)$ 可以任意接近。
- 若 $p(x)$ 在 $(-\infty,\infty)$ 上一致连续，且

$$\lim_{n\to\infty} \frac{K(n)}{n} = 0 \qquad \lim_{n\to\infty} \frac{K(n)}{\ln n} = \infty$$

则

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |p_n^*(x)-p(x)| = 0\right) = 1.$$

(一致强相合)

5.3 最大似然估计

参数方法与非参数方法

- 经验分布函数、直方图估计、核密度估计、最近邻密度估计都不需要假定总体来自那一种分布，称为**非参数**统计方法。
- 但是，直接估计一个函数需要的信息量很大。
- 如果已知总体分布类型，只是分布参数未知，则可以只估计分布参数，然后总体分布就知道了。这种方法叫做**参数**统计方法。
- 例如，设产品指标服从正态分布 $\mathrm{N}(\mu,\sigma^2)$，但 μ,σ^2 未知，就可以由样本估计 μ,σ^2，代入密度函数中作为分布密度估计。
- 又如，产品寿命常服从威布尔分布或对数正态分布，可以从样本中估计分布参数后得到总体分布的估计。

参数估计问题

- 设总体 X 的密度函数或概率函数为 $p(x;\theta_1,\theta_2,\ldots,\theta_m)$,其中 $\theta_1,\theta_2,\ldots,\theta_m$ 是未知参数。
- 若 X 的样本值为

$$x_1,x_2,\ldots,x_n$$

- 问：如何估计参数 $\theta_1,\theta_2,\ldots,\theta_m$?
- 估计方法有很多，常用的有最大似然估计法和矩估计法。

似然函数

- 给定样本值 $x_1,x_2,\ldots,x_n$ 后，令

$$\begin{aligned}&L_n(x_1,x_2,\ldots,x_n;\theta_1,\theta_2,\ldots,\theta_m)\\=&\prod_{i=1}^{n}p(x_i;\theta_1,\theta_2,\ldots,\theta_m)\end{aligned}$$

把样本值 $x_1, x_2, \ldots, x_n$ 看作常数，这样 $L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$ 是参数 $\theta_1, \theta_2, \ldots, \theta_m$ 的函数，叫做样本 $x_1, x_2, \ldots, x_n$ 的**似然函数**。

- 似然函数如果看成自变量 $x_1, x_2, \ldots, x_n$ 的函数，实际是样本 $(X_1, X_2, \ldots, X_n)$ 看成随机变量时的联合密度函数。

最大似然估计

- **定义 3.1** 如果 $L_n(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$ 在 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 达到最大值，则称 $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ 为参数 $(\theta_1, \theta_2, \ldots, \theta_m)$ 的**最大似然估计** (MLE)。

最大似然估计直观解释例子

- **例** 假设两个人一同出去打猎，两人只允许开一枪。
- 甲击中概率为 0.9，乙击中概率为 0.2。
- 二人返回时，带回一头猎物。
- 众人会认为是谁的开枪？
- 两人射击结果用随机变量 X 表示：

$$X = \begin{cases} 1 & \text{击中} \\ 0 & \text{未击中} \end{cases}$$

- $P(X = 1)$ 可能为 $\{0.9, 0.2\}$ 两种取值（对应甲开枪和乙开枪）。
- 在甲开枪的情况下，得到已发生的“命中”结果可能性大于乙开枪的情况下，得到已发生的“命中”结果可能性。
- 所以推测是甲开的枪。
- 在结果（样本值）给定时，似然函数代表参数取不同值时，观测到已发生的结果的可能性大小。

最大似然估计直观解释例子 II

- 假定一个盒子里有许多黑球和白球，比例 3:1，但不知道黑球多还是白球多。
- 则随机抽取一个球，得到黑球的概率或者是 $\frac{1}{4}$, 或者是 $\frac{3}{4}$。
- 如果有放回地从盒子里抽 3 个球，则黑球数目服从二项分布：

$$P(X=x) = C_3^x p^x (1-p)^{3-x}, \quad x = 0,1,2,3$$
$$(p = \frac{1}{4}, \frac{3}{4} \text{为抽到黑球的概率})$$

- 根据抽取到的黑球数判断到底是黑球多还是白球多。
- 直观看，如果 $x = 0, 1$，则猜白球多；如果 $x = 2, 3$，则猜黑球多。

- 样本值固定后，取不同参数值的似然函数大小代表该种参数下观测到已发生的情况的可能性大小。
- 我们取使得已发生的情况出现的可能性最大的参数作为估计值。
- 分别计算 $p = \frac{1}{4}, \frac{3}{4}$ 的似然函数值：

x	0	1	2	3
$p = \frac{1}{4}$ 时 $P(X=x)$ 的值	$\mathbf{\frac{27}{64}}$	$\mathbf{\frac{27}{64}}$	$\frac{9}{64}$	$\frac{1}{64}$
$p = \frac{3}{4}$ 时 $P(X=x)$ 的值	$\frac{1}{64}$	$\frac{9}{64}$	$\mathbf{\frac{27}{64}}$	$\mathbf{\frac{27}{64}}$

- 于是 $x = 0, 1$ 时应取 $\hat{p} = \frac{1}{4}$（猜白球多）, $x = 2, 3$ 时应取 $\hat{p} = \frac{3}{4}$（猜黑球多）。

最大似然估计求法

- 把似然函数简记为 L_n。L_n 与 $\ln L_n$ 同时达到最大值，可以求 $\ln L_n$ 的最大值点。

- 当 $\ln L_n$ 的一阶偏导数存在时，最大值点处一阶偏导数都等于零：

$$\begin{cases} \frac{\partial \ln L_n}{\partial \theta_1} = 0 \\ \frac{\partial \ln L_n}{\partial \theta_2} = 0 \\ \cdots\cdots \\ \frac{\partial \ln L_n}{\partial \theta_m} = 0 \end{cases} \tag{3.1}$$

  这个关于参数 $(\theta_1, \theta_2, \ldots, \theta_m)$ 的方程组叫做似然方程组。

- 注意一阶偏导存在条件下似然方程组成立，但似然方程组的解不能保证为最大值点。

最大似然估计的优良性

- 在相当一般的条件下：
- 相合性：n 充分大时最大似然估计结果与参数真值之间可以无限接近。
- 有效性：一定意义下没有比最大似然估计更精确的估计。
- 渐近正态性：n 充分大时最大似然估计近似服从正态分布。

指数分布参数的最大似然估计

- 密度

$$p(x;\lambda) = \lambda e^{-\lambda x},\ x > 0, \lambda > 0$$

- 样本 $x_1, x_2, \ldots, x_n$ 的似然函数和对数似然函数

$$\begin{aligned} L_n(x_1, x_2, \ldots, x_n; \lambda) =& \lambda^n \exp\left\{-\lambda \sum_1^n x_i\right\} \\ \ln L_n =& n \ln \lambda - \lambda \sum_1^n x_i \end{aligned}$$

- 似然方程

$$\frac{d\ln L_n}{d\lambda} = \frac{n}{\lambda} - \sum_1^n x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_1^n x_i} = \frac{1}{\bar{x}}$$

此 $\hat{\lambda}$ 确实是 $\ln L_n$ 的最大值点，是 λ 的最大似然估计。

- **例 3.1** 一直某种电子设备的使用寿命服从指数分布 $\mathrm{Exp}(\lambda)$。今随机抽取 18 台，测得寿命数据如下（单位：小时）：

$$16, 29, 50, 68, 100, 130, 140$$

$$270, 280, 240, 410, 450, 520, 620$$

$$190, 210, 800, 1100$$

- 求 λ 的估计。

- **解** 用最大似然估计。$\bar{x} = 318$，

$$\hat{\lambda} = \frac{1}{318} \approx 0.03144$$

是 λ 的估计值。

正态分布参数的最大似然估计

- $\mathrm{N}(\mu, \sigma^2)$ 的密度函数为 (记 $\delta = \sigma^2$)

$$p(x; \mu, \delta) = \frac{1}{\sqrt{2\pi\delta}} \exp\left\{-\frac{1}{2\delta}(x-\mu)^2\right\}$$

- 样本 $x_1, x_2, \ldots, x_n$ 的似然函数与对数似然函数为

$$\begin{aligned} & L_n(x_1, x_2, \ldots, x_n;\ \mu, \delta) \\ =& (2\pi)^{-\frac{n}{2}} \delta^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\delta}\sum_{i=1}^n (x_i-\mu)^2\right\} \end{aligned}$$

$$\begin{aligned} & \ln L_n \\ =& -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\delta - \frac{1}{2\delta}\sum_{i=1}^n (x_i-\mu)^2 \end{aligned}$$

- 似然方程组为

$$\begin{cases} \dfrac{\partial \ln L_n}{\partial \mu} = \dfrac{1}{\delta}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ \dfrac{\partial \ln L_n}{\partial \delta} = -\dfrac{n}{2\delta} + \dfrac{1}{2\delta^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases}$$

- 解得

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

- $(\hat{\mu}, \hat{\delta})$ 确实是 $\ln L_n$ 的最大值点，所以是 (μ, σ^2) 的最大似然估计。

威布尔分布参数的最大似然估计

- Weibull(m, η) 分布密度为

$$p(x; m, \eta) = \frac{m}{\eta^m} x^{m-1} \exp\left\{-\left(\frac{x}{\eta}\right)^m\right\},$$
$$x > 0;\ m > 0, \eta > 0$$

- 似然函数和对数似然函数分别为

$$\begin{aligned} & L_n(x_1, x_2, \ldots, x_n; m, \eta) \\ =& m^n \eta^{-mn} \left(\prod_{i=1}^{n} x_i\right)^{m-1} \exp\left\{-\frac{1}{\eta^m}\sum_{i=1}^{n} x_i^m\right\} \\ & \ln L_n \\ =& n\ln m - nm\ln\eta + (m-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\eta^m}\sum_{i=1}^{n} x_i^m \end{aligned}$$

- 似然方程组为

$$\begin{aligned}\frac{\partial \ln L_n}{\partial m} =& \frac{n}{m} - n\ln\eta + \sum_{i=1}^{n}\ln x_i \\ & - \frac{1}{\eta^m}\sum_{i=1}^{n} x_i^m \ln x_i + \frac{\ln\eta}{\eta^m}\sum_{i=1}^{n} x_i^m = 0 \qquad (3.2a)\end{aligned}$$

$$\frac{\partial \ln L_n}{\partial \eta} = -\frac{nm}{\eta} + \frac{m}{\eta^{m+1}}\sum_{i=1}^{n} x_i^m = 0 \qquad (3.2b)$$

- 由 (3.2b) 得

$$\eta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^m\right)^{\frac{1}{m}} \qquad (3.3)$$

- 代入 (3.2a) 可得

$$\frac{1}{m} + \frac{1}{n}\sum_{i=1}^{n}\ln x_i - \frac{\sum_{i=1}^{n} x_i^m \ln x_i}{\sum_{i=1}^{n} x_i^m} = 0 \qquad (3.4)$$

- 当 $n \geq 2$, $x_1, x_2, \ldots, x_n$ 不完全相等时，方程 (3.4) 恰有一个根 $\hat{m}$，代回 (3.3) 中得

$$\hat{\eta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{m}}\right)^{\frac{1}{\hat{m}}}$$

- 可以证明 $(\hat{m}, \hat{\eta})$ 是似然方程组的解，也是对数似然函数的最大值点，所以是参数 (m, η) 的最大似然估计。

- 方程 (3.4) 一定有唯一解且方程左边是 m 的严格单调减函数，可以用二分法求解 (计算机数值算法求解)。

- **例 3.2** 轴承的寿命一般服从威布尔分布。$n = 20$ 的样本数据如下 (单位：小时)：

$$153, 223, 313, 373, 378, 385, 424,$$
$$232, 452, 452, 547, 561, 634, 699,$$
$$759, 859, 1000, 1132, 1152, 1466$$

  估计形状参数 m 和刻度参数 η。

- **解** 用最大似然估计法，解方程 (3.4) 得 m 的最大似然估计 $\hat{m} = 1.9$。再利用 (3.5) 可得 η 的最大似然估计 $\hat{\eta} = 685$。

均匀分布参数的最大似然估计

- 参数似然函数不可导时，可以具体研究似然函数求最大值点。

- 均匀分布 $\mathrm{U}(a, b)$ 的密度函数为

$$p(x; a, b) = \frac{1}{b-a} I_{[a,b]}(x)$$

  ($a < b$ 是未知参数)

- 似然函数

$$\begin{aligned} L_n =& \frac{1}{(b-a)^n} \prod_{i=1}^{n} I_{[a,b]}(x_i) \\ =& \begin{cases} \frac{1}{(b-a)^n} & \text{当} x_{(1)} \geq a \text{且} x_{(n)} \leq b \\ 0 & \text{其它} \end{cases} \end{aligned}$$

- 似然函数最大，首先要求不为零，即 $a \leq x_{(1)}, b \geq x_{(n)}$。在此条件下 $b - a$ 最小，应取 a 最大, b 最小，所以

$$\hat{a} = x_{(1)} \qquad\qquad \hat{b} = x_{(n)}$$

5.4 期望与方差的点估计

数字特征的估计

- 直接估计分布函数、分布密度、概率函数要求数据很多，参数最大似然估计有时比较复杂。
- 如果只是需要估计期望、方差等数字特征，则比较容易。

5.4.1 期望的点估计

期望的点估计

- 例 1.1 中钢筋次品率估计可以看成是二点分布总体 X 的期望 $E(X)$ 的估计问题。
- 一般地，对总体 X 的样本 $x_1, x_2, \ldots, x_n$，估计 $E(X)$ 可以用样本平均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 这个估计是好的，而且样本量 n 越大，估计越精确。

统计量和抽样分布

- 把样本 $X_1, X_2, \ldots, X_n$ 看成随机变量，样本平均值

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

  也是随机变量。
- 若函数 $\psi(x_1, x_2, \ldots, x_n)$ 是不依赖于未知参数的函数，样本的函数 $Y = \psi(X_1, X_2, \ldots, X_n)$ 叫做样本**统计量**。
- 统计量是随机变量，其分布叫做**抽样分布**。
- 比如 $\bar{X}$ 是统计量。

样本平均值的无偏性

- 虽然用 $\bar{X}$ 估计 $E(X)$ 有时比 $E(X)$ 大，有时比 $E(X)$ 小，但平均来说是等于 $E(X)$ 的：
- **定理 4.1** 设 $E(X)$ 存在，则

$$E(\bar{X}) = E(X)$$

- **证**

$$\begin{aligned} E(\bar{X}) =& E\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] \\ =& \frac{1}{n} E\left(X_1 + X_2 + \cdots + X_n\right) \\ =& \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right] \\ =& \frac{1}{n} \cdot nE(X) = E(X) \end{aligned}$$

- 称这样的估计为**无偏**估计。

有效性

- 记

$$\bar{X}_n = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$

- 按常理，$\bar{X}_2$ 应该比 $\bar{X}_1$ 更好。
- 在无偏的条件下，

$$D(\bar{X}_n) = E\left[\bar{X}_n - E(\bar{X}_n)\right]^2 = E\left[\bar{X}_n - E(X)\right]^2$$

  代表了估计误差大小。$D(\bar{X}_n)$ 越小则估计约精确。

- 抽样分布方差越小，称统计量越有效。

- **定理 4.2** 设 X 的期望、方差都存在，则

$$D(\bar{X}_n) = \frac{D(X)}{n}$$

- **证**

$$\begin{aligned} D(\bar{X}_n) =& D\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] \\ =& \frac{1}{n^2} D\left(X_1 + X_2 + \cdots + X_n\right) \\ =& \frac{1}{n^2}\left[D(X_1) + D(X_2) + \cdots + D(X_n)\right] \\ =& \frac{1}{n^2}\left[nD(X)\right] = \frac{D(X)}{n} \end{aligned}$$

- 定理说明样本量越大，估计越精确。

说明

- 为什么 n 越大，估计越精确？
- 利用切比雪夫不等式：

$$P\left\{\left|\bar{X} - E(\bar{X})\right| < \varepsilon\right\} \geq 1 - \frac{D(\bar{X})}{\varepsilon^2}$$

- 其中 $E(\bar{X}) = E(X)$, $D(\bar{X}) = \frac{D(X)}{n}$, 有

$$P\left\{\left|\bar{X} - E(X)\right| < \varepsilon\right\} \geq 1 - \frac{D(X)}{n\varepsilon^2} \tag{4.1}$$

- 从而

$$\lim_{n\to\infty} P\left\{\left|\bar{X} - E(X)\right| < \varepsilon\right\} = 1 \tag{4.2}$$

即 n 充分大时可以有充分大把握保证 $\left|\bar{X} - E(X)\right| < \varepsilon$，即 $\bar{X} \approx E(X)$。

关于估计的优良性

- 估计的优良性包括无偏性、有效性、相合性等。
- 设 X 的分布密度为 $p(x;\theta)$，其中 $\theta=(\theta_1,\theta_2,\ldots,\theta_m)\in\Theta$, Θ 是 m 维空间 R^m 中的某个集合。（当 X 为离散型时可做类似讨论）。
- 设 $g(\theta)$ 是参数 θ 的函数，$X_1,X_2,\ldots,X_n$ 是 X 的样本。
- **定义 4.1** 称样本的函数 $\varphi(X_1,X_2,\ldots,X_n)$ 为 $g(\theta)$ 的**估计量**。
- 估计量是统计量，只要给定样本值就可以计算出数值，计算不能依赖于参数 θ。
- 如何选择估计量？

无偏性

- 由于 $X_1,X_2,\ldots,X_n$ 的联合密度与 θ 有关，所以 $\varphi(X_1,X_2,\ldots,X_n)$ 的期望与 θ 有关，为此显式地记为

$$E_\theta\left[\varphi(X_1,X_2,\ldots,X_n)\right]$$

- **定义 4.2** 称 $\varphi(X_1,X_2,\ldots,X_n)$ 是 $g(\theta)$ 的无偏估计，若

$$E_\theta\left[\varphi(X_1,X_2,\ldots,X_n)\right]=g(\theta)\quad(\forall\theta\in\Theta)$$

有效性

- **定义 4.3** 若 $\varphi_1(X_1,X_2,\ldots,X_n)$ 和 $\varphi_2(X_1,X_2,\ldots,X_n)$ 都是 $g(\theta)$ 的估计量，满足

$$\begin{aligned}&E_\theta\left[\varphi_1(X_1,X_2,\ldots,X_n)-g(\theta)\right]^2\\ \leq&E_\theta\left[\varphi_2(X_1,X_2,\ldots,X_n)-g(\theta)\right]^2\quad(\forall\theta\in\Theta)\end{aligned}$$

  且存在 θ_0 使上式中小于号成立，则称 φ_1 比 φ_2 **有效**。
- 比如，$\bar{X}_k$ 和 $\bar{X}_{k-1}$ 都是 $E(X)$ 的无偏估计，但 $\bar{X}_k$ 比 $\bar{X}_{k-1}$ 有效（设 $D(X)>0$）。

最小方差无偏估计

- **定义 4.4** 如果 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的无偏估计量，而且对于 $g(\theta)$ 的任一无偏估计量 $\psi(X_1, X_2, \ldots, X_n)$ 都有

$$D\left(\varphi(X_1, X_2, \ldots, X_n)\right) \leq D\left(\psi(X_1, X_2, \ldots, X_n)\right) \quad (\forall \theta \in \Theta)$$

  则称 $\varphi(X_1, X_2, \ldots, X_n)$ 为 $g(\theta)$ 的最小方差无偏估计量。

相合性

- **定义** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的相合估计，若对任意 $\varepsilon > 0$，

$$\lim_{n\to\infty} P\left(|\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| > \varepsilon\right) = 0.$$

- **定义** 称 $\varphi(X_1, X_2, \ldots, X_n)$ 是 $g(\theta)$ 的强相合估计，若

$$P\left(\lim_{n\to\infty} |\varphi(X_1, X_2, \ldots, X_n) - g(\theta)| = 0\right) = 1.$$

5.4.2 方差的点估计

方差的点估计

- 设 $X_1, X_2, \ldots, X_n$ 为 X 的样本，为估计 $D(X)$，使用

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  称为**样本方差**。

- 显然，如果所有样本值都相等，样本值波动最小，$S^2 = 0$。样本值之间差异越大，S^2 越大。

- 为什么除以 $n-1$ 而不是除以 n？

样本方差的无偏性

- **定理 4.3** 设 X 的方差存在，则

$$E(S^2) = D(X)$$

- **证**

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 =& \sum_{i=1}^{n}\left(X_i^2 - 2\bar{X}\cdot X_i + \bar{X}^2\right)\\ =& \sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot\sum_{i=1}^{n} X_i + n\bar{X}^2\\ =& \sum_{i=1}^{n} X_i^2 - 2\bar{X}\cdot n\bar{X} + n\bar{X}^2\\ =& \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \end{aligned} \tag{4.3}$$

- 于是

$$\begin{aligned}E(S^2) =& E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]\\ =& \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\\ =& \frac{1}{n-1}\sum_{i=1}^{n} E(X_i^2) - \frac{n}{n-1}E\left(\bar{X}^2\right)\\ =& \frac{n}{n-1}\left[E\left(X^2\right) - E\left(\bar{X}^2\right)\right]\end{aligned}$$

- 其中

$$\begin{aligned}E\left(X^2\right) =& D(X) + [E(X)]^2\\ E\left(\bar{X}^2\right) =& D\left(\bar{X}\right) + \left[E\left(\bar{X}\right)\right]^2\\ =& \frac{D(X)}{n} + [E(X)]^2\end{aligned}$$

- 所以

$$E(S^2) = \frac{n}{n-1}\left[D(X) - \frac{1}{n}D(X)\right] = D(X)$$

- S^2 是 $D(X)$ 的无偏估计。如果用

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

估计 $D(X)$，则

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}D(X) < D(X)$$

- 但是，如果知道了所有总体的值 $x_1, x_2, \ldots, x_N$，则应该使用 $\frac{1}{N}$ 来计算总体方差。

- 注：知道所有总体时，可以认为分布为所有总体值上的离散均匀分布，这时 $E(X) = \bar{X}$, $D(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$。

5.4.3 矩估计法

矩估计法

- 如果总体参数可以从各阶矩表示，则估计了各阶矩后可以估计参数。

- 设随机变量 X 的分布密度是 $p(x;\theta_1,\theta_2,\ldots,\theta_m)$，$\nu_k$ 是 X 的 k 阶矩 $(k = 1, 2, \ldots)$，ν_k 是 $\theta_1, \theta_2, \ldots, \theta_m$ 的函数：

$$\nu_k = E\left(X^k\right) = g_k(\theta_1, \theta_2, \ldots, \theta_m)$$

- 设 $\nu_1, \nu_2, \ldots, \nu_m$ 已知，如果从方程组

$$\begin{cases} g_1(\theta_1, \theta_2, \ldots, \theta_m) = \nu_1 \\ g_2(\theta_1, \theta_2, \ldots, \theta_m) = \nu_2 \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ g_m(\theta_1, \theta_2, \ldots, \theta_m) = \nu_m \end{cases}$$

可以求出

$$\begin{cases} \theta_1 = f_1(\nu_1, \nu_2, \ldots, \nu_m) \\ \theta_2 = f_2(\nu_1, \nu_2, \ldots, \nu_m) \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ \theta_m = f_m(\nu_1, \nu_2, \ldots, \nu_m) \end{cases}$$

- 设 $x_1, x_2, \ldots, x_n$ 是 X 的样本值，用

$$\hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (k = 1, 2, \ldots, m)$$

来估计 ν_k；

- 用

$$\hat{\theta}_k = f_k(\hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_m)$$

估计 $\theta_k (k = 1, 2, \ldots, m)$。

- 这种估计未知参数的方法叫做**矩估计法**。

- **例 4.2** 设 $X \sim \mathrm{N}(\mu, \sigma^2)$，$x_1, x_2, \ldots, x_n$ 是样本，求 μ, σ^2 的矩估计。

-

$$\begin{aligned} \nu_1 &= E(X) = \mu \\ \nu_2 &= E(X^2) = \sigma^2 + \mu^2 \end{aligned}$$

- 反解得

$$\begin{cases} \mu = \nu_1 \\ \sigma^2 = \nu_2 - \nu_1^2 \end{cases}$$

- 估计 ν_1, ν_2 为

$$\hat{\nu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
$$\hat{\nu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- 估计 μ, σ^2 为

$$\hat{\mu} = \hat{\nu}_1 = \bar{x}$$
$$\hat{\sigma}^2 = \hat{\nu}_2 - \hat{\nu}_1^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 与最大似然估计相同。

- 其它分布不一定。样本量较大时，最大似然估计一般更精确。

- **例 4.3** $X \sim \mathrm{U}(0, \theta)$，$\theta$ 为未知参数。

- 最大似然估计 $\hat{\theta} = x_{(n)}$。

- 因为 $\nu_1 = E(X) = \frac{\theta}{2}$，所以 θ 矩估计为

$$\tilde{\theta} = 2\hat{\nu}_1 = 2\bar{x} = \frac{2}{n}\sum_{i=1}^{n} x_i$$

- **例 4.4** 台风可以引起内陆降雨。以下 $n = 36$ 个观测值为 24 小时降

雨量实际观测数据：

$$
\begin{aligned}
&31.00, 2.82, 3.95, 4.02, 9.50, 4.50, 11.40,\\
&10.71, 6.31, 4.95, 5.64, 5.51, 13.40, 9.72,\\
&6.47, 10.16, 4.21, 11.60, 4.75, 6.85, 6.25,\\
&3.42, 11.80, 0.80, 3.69, 3.10, 22.22, 7.43,\\
&5.00, 4.58, 4.46, 8.00, 3.73, 3.50, 6.20, 0.67
\end{aligned}
$$

- 降雨量一般服从伽玛分布。密度为

$$
p(x;\alpha,\beta)=\begin{cases}\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & x\le 0\end{cases}
$$

- 用矩法估计 α,β。

$$
\begin{aligned}
\nu_1=&\frac{\alpha}{\beta}\\
\nu_2=&\frac{\alpha(\alpha+1)}{\beta^2}
\end{aligned}
$$

- $\hat{\nu}_1=7.29$, $\hat{\nu}_2=85.59$。
- 解方程组

$$
\begin{cases}
\dfrac{\alpha}{\beta}=&7.29\\
\dfrac{\alpha(\alpha+1)}{\beta^2}=&85.59
\end{cases}
$$

- 得 $\hat{\alpha}=1.64$, $\hat{\beta}=0.22$。

5.5 期望的置信区间

置信区间

- 前面找到了期望 $E(X)$ 和方差 $D(X)$ 的估计量，这种估计量又称为**点估计**，因为它们是用一个数值来估计未知的参数或数字特征的。
- 我们有时还希望了解估计的准确程度，这时应该用一个可能取值的范围（区间）来估计未知参数和数字特征。
- 将讨论正态总体的区间估计：

  (1) 已知方差，对 $E(X)$ 进行区间估计；

  (2) 未知方差，对 $E(X)$ 进行区间估计；

  (3) 方差 $D(X)$ 的区间估计在下一节讨论。

方差已知时期望的置信区间

- 设总体 X 服从 $\mathrm{N}(\mu, \sigma^2)$ 分布。
- 则 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 也是正态分布随机变量，分布为

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) . = \mathrm{N}\left(E(X), \frac{D(X)}{n}\right)$$

- (参见 P144 习题十四第 9 题及 P422 附录二定理 5 的系)。
- 于是

$$\eta = \frac{\bar{X} - E(X)}{\sqrt{\frac{D(X)}{n}}} \sim \mathrm{N}(0, 1)$$

- 根据正态分布的经验规则，

$$P(|\eta| \leq 1.96) = 0.95$$

- 即

$$P\left(|\bar{X} - E(X)| \leq 1.96\sqrt{\frac{D(X)}{n}}\right) = 0.95 \tag{5.1}$$

- 从 (5.1) 式看出，有 95% 的把握保证

$$|E(X) - \bar{X}| \leq 1.96\sqrt{\frac{D(X)}{n}}$$

即

$$\bar{X} - 1.96\sqrt{\frac{D(X)}{n}} \leq E(X) \leq \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}$$

- 即随机区间

$$\left[\bar{X} - 1.96\sqrt{\frac{D(X)}{n}}, \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}\right] \tag{5.2}$$

以很大的概率包含 $E(X)$。

- 这样的区间叫做**置信区间**。
- 称这样的置信区间的**置信水平**或**置信度**为 0.95。
- 如果做 100 次抽样（每次抽 n 个样品），则从平均的意义讲，算出的 $\bar{x}$ 值有 95 次，使得区间 (5.2) 包含真值 μ。(演示)
- 注意：在计算出置信区间后，我们不能说 $E(X)$ 属于这个区间的概率是 95%。因为一个计算出来的区间或者包含 $E(X)$, 或者不包含 $E(X)$。
- 样本量 n 越大，置信区间越短。
- 置信度越高，计算的置信区间越长。
- **例 5.1** 滚珠直径 X 服从正态分布。随机抽取 $n = 6$ 个，测量值（单位: mm）：

$$14.70,\ 15.21,\ 14.90,\ 14.91,\ 15.32,\ 15.32$$

- 估计直径的平均值。
- 如果知道 X 的方差为 0.05, 求平均直径的置信区间。

- **解**

$$\begin{aligned}\bar{x} =& \frac{1}{6}(14.70 + 15.21 + 14.90 + 14.91 + 15.32 + 15.32)\\ =& 15.06(\text{mm})\end{aligned}$$

为 $E(X)$ 的估计。

- 为计算 $E(X) = \mu$ 的置信区间，计算半径

$$1.96\sqrt{\frac{D(X)}{n}} = 1.96 \times \sqrt{\frac{0.05}{6}} = 0.18$$

- $E(X)$ 的置信区间为

$$[15.06 - 0.18, 15.06 + 0.18] = [14.88, 15.24]$$

非正态分布的情形

- 如果 X 不是服从正态分布，根据中心极限定理，当 n 充分大时

$$\eta = \frac{\bar{X} - E(X)}{\sqrt{\frac{D(X)}{n}}}$$

近似服从标准正态分布。

- 所以 $E(X)$ 的置信度为 95% 的置信区间仍可用公式

$$\left[\bar{X} - 1.96\sqrt{\frac{D(X)}{n}}, \bar{X} + 1.96\sqrt{\frac{D(X)}{n}}\right]$$

计算。

方差未知时均值的置信区间

- 如果 $D(X)$ 未知时求 $E(X)$ 的置信区间，想到的是用样本方差 S^2 代替 (5.2) 中的 $D(X)$。

- 但是，这时

$$T = \frac{\bar{X} - E(X)}{\sqrt{\frac{S^2}{n}}}$$

不再服从标准正态分布，不能利用正态分布经验规则。

- 需要推导 T 的分布。

t 分布

- 当总体 $X \sim \mathrm{N}(\mu, \sigma^2)$，$X_1, X_2, \ldots, X_n$ 是 X 的样本的时候，可以证明 T 的分布密度为

$$p_{n-1}(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} \tag{5.4}$$

- T 的分布只依赖于样本量 n 而与总体参数 μ, σ^2 无关。

- **定义 5.1** 如果随机变量 Y 的分布密度为

$$p_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \cdot \sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \tag{5.6}$$

则称 Y 服从 n 个自由度的 t 分布，记为 $Y \sim \mathrm{t}(n)$。

- 统计量 T 服从 $\mathrm{t}(n-1)$ 分布。

- $\mathrm{t}(n)$ 密度为偶函数，形状与标准正态分布类似，但两个尾部比正态分布厚。

- 可以证明

$$\lim_{n\to\infty} p_n(t) = \phi(t), \quad t \in (-\infty, \infty)$$

- 可以找到 $\lambda > 0$ 使得

$$\int_{-\lambda}^{\lambda} p_{n-1}(t)dt = 0.95$$

- 即

$$\begin{aligned} P(|T| \le \lambda) &= 0.95 \\ P\left(\bar{X} - \lambda\sqrt{\frac{S^2}{n}} \le E(X) \le \bar{X} + \lambda\sqrt{\frac{S^2}{n}}\right) &= 0.95 \end{aligned}$$

- 于是 $E(X)$ 的置信度为 95% 的置信区间为

$$\left[\bar{X} - \lambda\sqrt{\frac{S^2}{n}},\ \bar{X} + \lambda\sqrt{\frac{S^2}{n}}\right]$$

- λ 叫做 t 分布双侧 0.05 临界值，在 P432 附表 2 中列有不同自由度和不同置信度的对应值。

- **例 5.2** 用某仪器间接测量温度，重复测量 5 次，得到的结果如下（单位：°C）

$$1250,\ 1265,\ 1245,\ 1260,\ 1275$$

  假设仪器没有系统偏差，求真值的范围。

- **解** 用 μ 表示温度真值，X 表示测量值。X 通常服从正态分布。有 $n = 5$ 的样本。

- μ 的置信区间为

$$\left[\bar{x} - \lambda\sqrt{\frac{S^2}{n}}, \bar{x} + \lambda\sqrt{\frac{S^2}{n}}\right]$$

- 计算得 $\bar{x} = 1259$, $S^2 = \frac{570}{4}$，自由度为 $n - 1 = 4$，查 t 分布临界值表 ($\alpha = 0.05$) 得 $\lambda = 2.776$，半径为

$$\lambda\sqrt{\frac{S^2}{n}} = 2.776 \times \sqrt{\frac{570}{4 \times 5}} \approx 14.8$$

- 置信区间为

$$[1259 - 14.8, 1259 + 14.8] = [1244.2, 1273.8]$$

- **例 5.3** 对飞机的飞行速度进行 15 次独立试验，测得飞机的最大飞行速度 ($\mathrm{m} \cdot \mathrm{s}^{-1}$) 如下：

$$422.2,\ 418.7,\ 425.6,\ 420.3,\ 425.8$$
$$423.1,\ 431.5,\ 428.2,\ 438.3,\ 434.0$$
$$412.3,\ 417.2,\ 413.5,\ 441.3,\ 423.7$$

  根据长期经验，可以认为最大飞行速度服从正态分布。求最大飞行速度期望的置信区间。

- **解** 用 X 表示最大飞行速度。$D(X)$ 未知，求 $E(X)$ 的置信区间。
- 这里 $\bar{x} = 425.047$, $S^2 = \frac{1006.34}{14}$。
- 自由度 $n - 1 = 14$，查表得 $\lambda = 2.145$。
- 半径

$$\lambda\sqrt{\frac{S^2}{n}} = 2.145\sqrt{\frac{1006.34}{14 \times 15}} = 4.696$$

- 置信区间为

$$[425.047 - 4.696, 425.047 + 4.696] = [420.35, 429.74]$$

方差未知时求期望置信区间的步骤

- 由样本值 $x_1, x_2, \ldots, x_n$ 计算出 $\bar{x}, S^2$。
- 查 t 分布临界值表，自由度为 $n - 1$，α 为 1− 置信度，得临界值 λ。

- 计算半径

$$d = \lambda\sqrt{\frac{S^2}{n}}$$

- 得到 $E(X)$ 的置信度为 $1-\alpha$ 的置信区间为

$$[\bar{x} - d, \bar{x} + d]$$

5.6 方差的置信区间

方差的置信区间

- 希望对方差 $D(X)$ 给出区间估计。方差本身也是一个重要指标。
- **例 6.1** 某自动车床加工零件，抽查 16 个零件，测得长度如下（单位: mm）:

$$12.15,\ 12.12,\ 12.01,\ 12.08,\ 12.09,\ 12.16$$
$$12.03,\ 12.01,\ 12.06,\ 12.13,\ 12.07,\ 12.11$$
$$12.08,\ 12.01,\ 12.03,\ 12.06$$

- 估计方差。$\bar{x} = 12.075$, $S^2 = 0.00244$。
- 如何给出方差的区间估计？

正态分布总体方差的置信区间

- 设总体 $X \sim \mathrm{N}(\mu, \sigma^2)$,$X_1, X_2, \ldots, X_n$ 是 X 的样本。由样本值 $x_1, x_2, \ldots, x_n$ 给出 σ^2 的置信区间。
- 已知 S^2 是 σ^2 的无偏估计，但不知道 S^2 与 σ 的具体偏离情况。
- 来求 $\frac{S^2}{\sigma^2}$ 的分布。
- 可以证明 $\eta = \frac{(n-1)S^2}{\sigma^2}$ 的分布密度为

$$p(u) = \begin{cases} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} u^{\frac{n-3}{2}} e^{-\frac{u}{2}} & u > 0 \\ 0 & u \le 0 \end{cases} \tag{6.1}$$

- $n \geq 3$ 时图形手绘示意。

卡方分布

- η 的分布叫做卡方分布，是一种特殊的伽玛分布。
- **定义 6.1** 如果随机变量 Y 的分布密度函数为

$$k_n(u) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} u^{\frac{n}{2}-1} e^{-\frac{u}{2}} & u > 0 \\ 0 & u \leq 0 \end{cases}$$

  则称 Y 服从 n 个自由度的**卡方分布**，记作 $Y \sim \chi^2(n)$。
- 易见 $\chi^2(n)$ 分布是 $\text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$ 分布。
- $\eta = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$。

正态分布总体方差置信区间（续）

- η 分布已知，可以取 λ_1, λ_2 $(0 < \lambda_1 < \lambda_2)$，使得

$$P(\lambda_1 \leq \eta \leq \lambda_2) = 0.95 \tag{6.2}$$

- 一般选

$$\int_0^{\lambda_1} p(u)du = 0.025 \tag{6.3}$$

$$\int_{\lambda_2}^{\infty} p(u)du = 0.025 \tag{6.4}$$

- λ_1 和 λ_2 可以从 P433 的附表 3 查到（附表三给出的是卡方分布的右侧分位数）。

- 查出 λ_1, λ_2 后，以 95% 把握保证如下不等式：

$$\lambda_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq \lambda_2$$

- 即

$$\frac{(n-1)S^2}{\lambda_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\lambda_1}$$

- 即

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\lambda_2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\lambda_1}$$

  这就是 σ^2 的置信度为 0.95 的置信区间。

- 为了得到标准差 σ 的置信区间，只要把 σ^2 的置信区间端点开平方根。

- **例 6.1（续）** 这里

$$\sum_{n=1}^{n} (x_i - \bar{x})^2 = 0.0366$$

  $n = 16$, 查 15 个自由度的卡方分布临界值得 $\lambda_1 = 6.26$, $\lambda_2 = 27.5$，

$$\begin{aligned} &\left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\lambda_2}, \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\lambda_1}\right] \\ =&\left[\frac{0.0366}{27.5}, \frac{0.0366}{6.26}\right] \\ =&[0.0013, 0.0058] \end{aligned}$$

- σ 的置信区间为 $[0.036, 0.076]$。

5.7 寻求置信区间和置信限的一般方法

寻求置信区间和置信限的一般方法

- 概念：置信区间；置信水平（置信度）；置信系数。单侧的置信限。
- 方法：枢轴量方法；统计量方法；假设检验接受域方法。

第六章　假设检验

6.1　问题的提法

假设检验的问题

- 估计是统计学的一个重要问题，本章讨论另一个重要问题：“假设检验”。
- 本节用一些例子给出假设检验的典型问题。
- **例 1.1**　某厂有一批产品，共 200 件，须经检验合格才能出厂。按国家标准，次品率不得超过 1%，今在其中抽取 5 件，发现这 5 件中含有次品。
- 问这批产品能否出厂？
- 直观看不能出厂，但理由何在？
- 设这批产品的次品率为 p。
- 问题化为：如何根据抽样的结果来判断不等式“$p \leq 0.01$”是否成立？
- **例 1.2**　用某仪器间接测量温度，重复 5 次，所得数据如下：

  $$1250,\ 1265,\ 1245,\ 1260,\ 1275$$

  而用别的精确办法测得温度为 1277(可以看作温度的真值)。
- 试问此仪器间接的温度有无系统偏差？
- 用 X 表示这个仪器测得的数值，X 是随机变量。

- 得到的 5 个数据值是 X 的一个样本。
- 问题化为：如何判断等式“$E(X) = 1277$”成立与否？
- **例 1.3** 某工厂近 5 年来发生了 63 次事故，这些事故在工作日的分布如下：

星期	一	二	三	四	五	六
次数	9	10	11	8	13	12

- 问：事故发生是否与星期几有关？
- 用 X 表示这样的随机变量：若事故发生在星期 i，则 $X = i$。
- X 的可能取值集合为 $\{1, 2, \ldots, 6\}$（星期日是该厂厂休日）。
- 问题化为：如何判断

$$P(X = i) \equiv \frac{1}{6} \quad (i = 1, 2, \ldots, 6)$$

是否成立？

- **例 1.4** 在针织品的漂白工艺过程中，要考察温度对针织品的断裂强力（主要质量指标）的影响。
- 为了比较 70°C 与 80°C 的影响有无差别，在这两个温度下，分别重复做了 8 次试验，得数据如下（单位：千克力）

70°C时的强力：20.5, 18.8, 19.8, 20.9,
21.5, 19.5, 21.0, 21.2
80°C时的强力：17.7, 20.3, 20.0, 18.8,
19.0, 20.1, 20.2, 19.1

- 从试验数据看，两种温度下的强度有无区别？

- 用 X 表示 70°C 下的强力，Y 表示 80°C 下的强力，问题变成：

- 如何判断等式

$$E(X) = E(Y)$$

成立与否？

- 还可以进一步问等式

$$D(X) = D(Y)$$

成立与否？

- 比较新型吹风机与原吹风机。原吹风机是成功，但面临竞争压力。
- 新型吹风机成本减少了 15%，如果产品可靠性也不比原产品差则可以上市取代原产品。
- 否则不采纳新产品，因为产品有 1 年保质期，保质期内损坏需要免费更换。
- 公司进行了可靠性试验，将新产品和原产品各取 250 件在模拟一年使用的条件下进行试验。
- 发现新产品中有 11 个失效，而原产品中有 20 个失效。
- 问：新产品的可靠性是否不比原产品的差？
- 用 p_1 表示新产品的失效率，p_2 表示原产品的失效率。
- 问题化为：判断

$$p_1 \leq p_2$$

是否成立？

- 怎样根据一个随机变量的样本值，判断该随机变量是否服从正态分布 $\mathrm{N}(\mu, \sigma^2)$?
- 更一般地，如何根据样本的特性去判断随机变量是否以给定的函数 $F(x)$ 为其分布函数？

假设检验问题

- 例 1.1—例 1.6 代表了一类广泛的统计学问题。
- 共同点是要从样本值出发去判断一个“看法”是否成立。
- 例 1.1 的“看法”是“次品率 $p \leq 0.01$”;
- 例 1.2 的看法是“$E(X) = 1277$”;
- 例 1.3 的看法是“$P(X = i) \equiv \frac{1}{6}\ (i = 1, 2, \ldots, 6)$”;
- 例 1.4 的看法是“$E(X) = E(Y)$”;
- 例 1.5 的看法是“$p_1 \leq p_2$”;
- 例 1.6 的看法是“X 的分布函数是 $F(x)$”。

- “看法”又叫“假设”。这些例子就是所谓**假设检验**问题。
- 本章介绍一些常用的检验方法，判断所关心的“假设”是否成立。
- 例 1.1、例 1.2 和例 1.3 中的“假设”都是关于一个随机变量的参数的判断，这叫做**一个总体的参数检验问题**。
- 例 1.6 也是一个总体的检验问题，但不是参数检验，而是概率分布的检验问题。
- 例 1.4 和例 1.5 的“假设”是关于两个随机变量的判断，叫做**二总体的检验问题**。

假设检验的思想

- 假设检验的思想是“带概率的反证法”。
- 以例 1.1 为例说明。

- 要检验假设“$p \leq 0.01$”。
- 在假设成立的条件下进行分析。如果假设成立，则总体中至多有 2 件次品。
- 任抽取 5 件产品，先来求这 5 件中“无次品”的概率。

- 用古典概型：

$$\begin{aligned} P(\text{无次品}) = & \begin{cases} \frac{C_{198}^5}{C_{200}^5} & \text{当 200 件中有 2 件次品时} \\ \frac{C_{199}^5}{C_{200}^5} & \text{当 200 件中有 1 件次品时} \\ \frac{C_{200}^5}{C_{200}^5} & \text{当 200 件中没有次品时} \end{cases} \\ \geq & \frac{C_{198}^5}{C_{200}^5} = \frac{198 \times 197 \times \cdots \times 194}{200 \times 199 \times \cdots \times 196} \approx 0.9505 \\ \geq & 0.95 \end{aligned}$$

- 于是

$$P(\text{任取的 5 件中有次品}) \leq 1 - 0.95 = 0.05$$

- 计算说明：如果次品率真的 ≤ 0.01，那么抽取 5 件样品，出现次品的机会是很少的，平均在 100 次抽样中，出现不到 5 次。
- 如果 $p \leq 0.01$ 成立，则在一次抽样中，人们实际上很少遇见出现次品的情况。
- 现在的实际情况是遇到了次品，这是“不合理”的。
- 产生这种不合理现象的根源在于假设 $p \leq 0.01$；
- 因此假设“$p \leq 0.01$”是不能接受的。
- 所以按照国家标准，这批产品不能出厂。

概率性质的反证法

- （1）反证法的思想。为了检验一个假设（如“$p \leq 0.01$”），先假定这个假设是成立的，如果实际从样本中观察到的情况在这个假设下是不合理的，就认为原来的假设是不正确的，**拒绝**原来的假设。
- 如果样本没有不合理现象，就**不能拒绝**原来的假设。
- （2）这不是纯粹的反证法。这里的“不合理”，不是形式逻辑中绝对的矛盾，而是认为小概率事件在一次观察中基本不可能发生。
- 但原假设成立的情况下小概率还是有可能发生的，一旦出现这种情况，我们拒绝原假设就是错误的。
- 所以在观察到小概率事件后拒绝原假设是有可能犯错的，只不过这种可能性比较小而且可以控制。
- 在原假设下概率小到什么程度算是“小概率事件”? 通常取界限为 0.05，有时也取为 0.01 或 0.10。

本章内容

- §2 先讲一个正态总体的检验问题。
- §3 介绍假设检验的一般概念和数学描述，如功效、p 值等。
- §4 讲两个正态总体的检验问题。
- §5 介绍比率的假设检验（一个总体和两个总体）。
- §6 为概率分布的检验。

6.2 一个正态总体的假设检验

一个正态总体的假设检验

- 设总体 $X \sim \mathrm{N}(\mu, \sigma^2)$, 关于一个正态总体有四种假设检验问题：

- 已知方差 σ^2, 检验假设 $H_0 : \mu = \mu_0$;
- 未知方差 σ^2, 检验假设 $H_0 : \mu = \mu_0$;
- 未知期望 μ，检验假设 $H_0 : \sigma^2 = \sigma_0^2$;
- 未知期望 μ，检验假设 $H_0 : \sigma^2 \leq \sigma_0^2$。

已知方差检验均值

- **例 2.1** 某车间生产铜丝，主要质量指标是折断力大小。
- 用 X 表示该车间生产的铜丝的折断力。
- 由以往经验，可以认为 X 服从正态分布，期望为 570 千克力，标准差是 8 千克力。
- 换了一批原料后，认为方差不会有什么变化，但需要检验折断力是否和原来一样。
- 即：已知方差 $\sigma^2 = 8^2$, 检验假设

$$H_0 : \ \mu = 570$$

- 抽取了 10 个样品，测得折断力值为（单位：千克力）：

$$578, 572, 570, 568, 572, 570, 570, 570, 572, 596, 584$$

- 如何检验 H_0?
- 用概率性质的反证法。考虑在 H_0 成立的假设下，观测到的样本是否有不合理现象。
- 在 H_0 下，$X \sim \mathrm{N}(570, 8^2)$。
- 设 X 的一个样本为 $X_1, X_2, \ldots, X_n$（看成随机变量），则

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i \sim \mathrm{N}(570, \frac{8^2}{10})$$

$$U = \frac{\bar{X} - 570}{\sqrt{8^2/10}} \sim \mathrm{N}(0, 1)$$

(正态分布的线性组合仍服从正态分布)

- 由正态分布的经验规则，

$$P(|U| > 1.96) = 0.05$$

所以在 H_0 成立时 $\{|U| > 1.96\}$ 是一个小概率事件。

- 计算得 $\bar{x} = 575.2, U = 2.05$，说明小概率事件发生了，认为是不合理的，故此拒绝 H_0，习惯上说“折断力的大小和原来**有显著差异**”。

- **例 2.2** 根据历史经验和资料分析，某砖瓦厂所生产的砖的“抗断强度”X 服从正态分布，方差 $\sigma^2 = 1.21$。

- 从一批产品中抽取 $n = 6$ 的样本，测得抗断强度（单位：$\mathrm{kg \cdot cm^{-2}}$）：

$$32.56,\ 29.66,\ 31.64,\ 30.00,\ 31.87,\ 31.03$$

- 问：这批砖的平均抗断强度可否认为是 32.80?

- **解** 待检验的假设是 $H_0 : \mu = 32.50$。

- 计算统计量 U:

$$U = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{31.13 - 32.50}{\sqrt{\frac{1.21}{6}}} = -3.05$$

- 于是 $|U| = 3.05 > 1.96$，小概率事件发生。拒绝 H_0，这批砖的平均抗断强度不能认为是 32.50 (或称这批砖的平均抗断强度与 32.50 **有显著差异**)。

检验水平

- 以 H_0 下计算的概率小于 0.05 作为拒绝 H_0 的标准，这样 $\alpha = 0.05$ 叫做**检验水平**。

- 有些问题中需要把检验水平取得更小一些，如取 $\alpha = 0.01$。这时查表得 H_0 下

$$P(|U| > 2.58) = 0.01$$

- 在例 2.1 中，$U = 2.05$，不超过 2.58，在 0.01 水平下“小概率事件”没有发生，假设 H_0 与数据是**相容**的，简称 H_0 是相容的。
- 这与 $\alpha = 0.05$ 时结论不同，可见检验的结果与水平 α 的选择有关。
- 检验水平 α 的直观意义：把 H_0 成立时概率不超过 α 的事件当作一次观察时不会发生的“小概率事件”。

第一类错误

- 对于一般的 α（无论如何可以假设 $\alpha < 0.5$），取**临界值** λ 使得

$$P(|U| > \lambda) = \alpha$$

  从样本中计算统计量 U 的值，当 $|U| > \lambda$ 时就拒绝 H_0。
- 这样下结论不能绝对不犯错误（但是从部分（样本）推断整体（总体）本来就不能保证绝对正确，管中窥豹和盲人摸象是典型例子）。
- 即使 H_0 成立，仍有 α 的概率我们会拒绝 H_0，这种错误叫做**第一类错误**，检验水平 α 是犯第一类错误的概率的上界。

第二类错误

- 那么，是不是第一类错误越小越好？
- 不是。比如，我们取临界值 λ 接近于无穷大，这时 α 几乎等于零。
- 但是，这样检验相当于不论 H_0 成立与否都不拒绝 H_0，于是当 H_0 不成立时一定会犯错误。
- 当 H_0 不成立时，如果按照检验规则，数据与 H_0 相容，不能拒绝 H_0，就犯了**第二类错误**。
- 用 β 表示第二类错误的概率。
- 两类错误概率越小越好，但两者互相矛盾。
- 经典的统计假设检验做法是固定第一类错误概率的水平 α，然后尽可能设法减小第二类错误概率，如设计更好的检验方案，增大样本量。

未知方差时期望的检验

- 设总体为 $\mathrm{N}(\mu,\sigma^2)$, μ 和 σ^2 都未知，检验 $H_0:\mu=\mu_0(\mu_0$ 已知)。
- 例如，例 1.2，测量值 X 服从正态分布，μ 和 σ^2 都未知，$\mu_0=1277$，检验 $H_0:\mu=1277$。
- 如果 σ^2 已知，检验使用的统计量为

$$U=\frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

- 在 σ^2 未知时，在 U 的公式中用 σ^2 的估计量 S^2 (样本方差) 代替 σ^2 得到

$$T=\frac{\bar{X}-\mu_0}{\sqrt{\frac{S^2}{n}}}$$

- 由 §5.5(P151) 知 H_0 成立时 $T\sim\mathrm{t}(n-1)$ 分布。

- 对例 1.2，$n=5$，查自由度为 4 的 t 分布表得

$$P(|T|>2.776)=0.05$$

- 计算得 $\bar{x}=1259$, $S^2=142.5$，$T=-3.37$, $|T|>2.776$，小概率事件发生，所以拒绝 H_0，认为间接测温的平均值不等于精确测量值，即间接测温有系统误差。

正态总体方差未知时关于期望的检验步骤

(1) 提出带检验的假设 $H_0:\mu=\mu_0(\mu_0$ 已知)。

(2) 根据样本值 $x_1, x_2, \ldots, x_n$ 计算统计量

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

的值。

(3) 对于检验水平 α，自由度 $n-1$，查 t 分布临界值表（附表 2），得临界值 λ。

(4) 若 $|T| > \lambda$，拒绝 H_0；否则 H_0 相容（此时常接受 H_0）。

两类错误的比较

- 在经典的统计假设检验中，第一类错误由检验水平 α 控制，第二类错误只能尽量设法减小但不可控制。
- 所以，设计检验 H_0 时，尽可能取 H_0 使得 H_0 代表原来的、标准的情况或做法，从而接受 H_0 即使犯错（发生第二类错误）也没有严重后果，而拒绝 H_0 则需要更为谨慎（可能发生第一类错误）。
- 这样，一旦拒绝 H_0，我们是比较有把握结论是可信的。
- 实践中，人们一般希望得到拒绝 H_0 的结论。
- 这样也会发生报告偏差：不同的研究人员对同一问题做了试验，接受 H_0 的时候就不报告了，拒绝 H_0 就报告出来。
- **例 2.3** 根据长期资料分析，知道某种钢生产出的钢筋的强度服从正态分布。今随机抽取 6 根钢筋进行强度试验，测得强度为（单位：$\mathrm{kg \cdot mm^{-2}}$）：

  $$48.5,\ 49.0,\ 53.5,\ 49.5,\ 56.0,\ 52.5$$

- 问：能否认为该种钢生产的钢筋的平均强度为 52.0?

- **解** 用 X 表示钢筋强度，$X \sim \mathrm{N}(\mu, \sigma^2)$.
- （1）要检验的假设是 $H_0 : \mu = 52.0$。
- （2）计算统计量 T 的值。$\bar{x} = 51.5, S^2 = 8.9,$

$$T = \frac{51.5 - 52.0}{\sqrt{8.9/6}} = -0.411$$

- （3）查附表 2，$\alpha = 0.05$，自由度 $n - 1 = 5$，得 $\lambda = 2.571$。
- （4）下判断。现在 $|T| = 0.411 < 2.571$，故 H_0 是相容的，不能否定钢筋的平均强度为 $52.0\mathrm{kg} \cdot \mathrm{mm}^{-2}$。

假设检验与置信区间

- $H_0 : \mu = \mu_0$ 的检验与 μ 的置信区间有密切联系。
- $\mu = E(X)$ 的置信水平为 $1 - \alpha$ 的置信区间是满足

$$\left| \frac{\bar{x} - \mu}{\sqrt{S^2/n}} \right| \leq \lambda$$

  的 μ 的集合。
- 检验 $H_0 : \mu = \mu_0$ 的规则是：当且仅当

$$\left| \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \right| \leq \lambda$$

  时不拒绝 H_0。
- 所以检验法也可以说是：当且仅当 μ_0 属于 μ 的置信水平为 $1 - \alpha$ 的置信区间时不拒绝 H_0。
- 反过来以可以由检验法构造置信区间。

方差的双边检验

- 对正态总体，设期望和方差未知，检验：

$$H_0 : \sigma^2 = \sigma_0^2$$

(σ_0^2 已知)

- **例 2.4** 某车间生产铜丝，生产一直比较稳定。
- 今从产品中随机抽出 10 根检查折断力，得数据如下（单位：千克力）：

$$578, 572, 570, 568, 572, 570, 570, 570, 572, 596, 584$$

- 问：是否可相信该车间的铜丝的折断力的方差为 64?

- 用 X 表示铜丝的折断力，已知 $X \sim \mathrm{N}(\mu, \sigma^2)$，要根据样本检验

$$H_0 : \sigma^2 = \sigma_0^2 \qquad (\sigma_0^2 = 64)$$

- 自然想到用 S^2 与 σ_0^2 比较。如果 S^2/σ_0^2 很大或很小，则应拒绝 H_0。如果 S^2/σ_0^2 的值接近于 1，应该接受 H_0。
- 取统计量

$$W = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

在 W 很大或很小时拒绝 H_0。

- 由 §5.6, 当 H_0 成立时 $W \sim \chi^2(n-1)$。查 χ^2 分布临界值表找到 λ_1, λ_2 使得

$$P(W < \lambda_1) = 0.025$$
$$P(W > \lambda_2) = 0.025$$

- 则 H_0 成立时 $\{W < \lambda_1 \text{或} W > \lambda_2\}$ 是小概率事件。

- 对例 2.4，从样本中计算得 $\bar{x} = 575.2$, $S^2 = 75.73$, 故

$$W = \frac{9 \times 75.73}{64} = 10.65$$

- 查 χ^2 分布表，自由度 $n - 1 = 9$，得 $\lambda_1 = 2.70$, $\lambda_2 = 19.0$。现在

$$\lambda_1 < W < \lambda_2$$

- 所以 H_0 是相容的，可以相信铜丝折断力方差为 64。

方差的单侧检验

- 设正态总体的期望和方差都未知，检验

$$H_0 : \sigma^2 \leq \sigma_0^2 \qquad (\sigma_0^2\text{已知})$$

- 这种检验也是常用的。H_0 代表生产的加工精度没有变差，如果拒绝了 H_0，就说明加工精度变差了，要检查原因。

- 设总体 $X \sim \mathrm{N}(\mu, \sigma^2)$，$X_1, X_2, \ldots, X_n$ 为样本。

- 自然想到如果 S^2/σ_0^2 远大于 1 则应该拒绝 H_0。

- 仍采用统计量

$$W = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

  当 W 很大时拒绝 H_0。

- 在 $H_0 : \sigma^2 \leq \sigma_0^2$ 的情况下，W 的分布依赖于未知的 σ^2。

- 用不等式推导中的放大法。

- 令

$$Y = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{\sigma_0^2}{\sigma^2}W$$

- 则 H_0 成立时 $Y \geq W$。
- $Y \sim \chi^2(n-1)$，设 λ 是 $\chi^2(n-1)$ 分布右侧 α 临界值，即

$$P(Y > \lambda) = \alpha$$

- 则 $Y > \lambda$ 是小概率事件。
- 但是 Y 不是统计量，其计算公式涉及未知参数 σ^2。
- 然而，在 H_0 下

$$P(W > \lambda) \leq P(\frac{\sigma_0^2}{\sigma^2}W > \lambda) = P(Y > \lambda) = \alpha$$

- 也就是说，H_0 下 $\{W > \lambda\}$ 也是小概率事件，而 W 是统计量。
- 从样本值 $x_1, x_2, \ldots, x_n$ 中计算得到 W，如果 $W > \lambda$，就应该拒绝 H_0。

方差检验步骤

- 检验步骤：
- (1) 提出待检验的假设 $H_0 : \sigma^2 = \sigma_0^2$(或 $H_0 : \sigma^2 \leq \sigma_0^2$)。
- (2) 计算统计量

$$W = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

- (3) 查 χ^2 分布临界值表（附表 3），自由度为 $n-1$，得 λ_1,λ_2 (或 λ) 满足

$$P(\chi^2 < \lambda_1) = (\chi^2 > \lambda_2) = \frac{\alpha}{2}$$

$$(\text{或} P(\chi^2 > \lambda) = \alpha)$$

- (4) 比较 W 与 λ_1, λ_2（或 λ），作出判断。

单边检验假设选取

- 对于单边的问题，假设可以是问题本身或者问题的反面。
- 一般取假设为原来的、标准的情况或做法，这样接受假设即使犯错后果也不严重。
- 如果希望回答可信，应该把希望得到的回答的反面作为假设（拒绝了 H_0 时的结论是比较可靠的）。
- **例 2.5** 已知罐头番茄汁中，维生素 C 含量服从正态分布。
- 按照规定，维生素 C 含量不得少于 21 毫克。
- 现从一批罐头中抽了 17 罐，算得维生素 C 含量统计量为 $\bar{x} = 23$, $S^2 = 3.98^2$。
- 为这批罐头的维生素 C 的含量是否合格？
- **解** 设这批罐头中维生素 C 含量 $X \sim \mathrm{N}(\mu, \sigma^2)$。
- 我们希望作出合格的判断时比较有把握，所以设 H_0 为合格的反面：

$$H_0 : \mu < 21$$

- 参考 $H_0 : \mu = \mu_0$ 时的做法，考虑统计量

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

  当 T 很大时拒绝 H_0。
- 在 H_0 成立时 T 的分布依赖于未知参数 μ，取

$$Y = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

- 则 H_0 成立时 $Y > T$。

- 取 λ 使

$$P(Y > \lambda) = \alpha$$

- 则 H_0 成立时

$$P(T > \lambda) \leq P(Y > \lambda) = \alpha$$

所以 $\{T > \lambda\}$ 是小概率事件。

- 为了从附表 2 中查 λ，注意自由度为 $n-1$，另外

$$P(Y > \lambda) = \alpha \Longleftrightarrow P(|Y| > \lambda) = 2\alpha$$

所以取 $\alpha = 0.05$ 时只要从附表 2 中查 0.10 对应的 λ。

- 对例 2.5，查附表 2，自由度 $n-1 = 16$，0.10 对应的临界值为 $\lambda = 1.746$。
- 计算得

$$T = \frac{\bar{X} - 21}{\sqrt{\frac{S^2}{n}}} = 2.07 > 1.746$$

- 所以小概率事件发生了，拒绝 H_0，认为该批罐头合格。

非正态总体的均值假设检验

- 对于非正态的总体，当 $n \to \infty$ 时根据中心极限定理以及相关的概率极限理论可以证明在 $H_0 : \mu = \mu_0$ 下 T 统计量近似标准正态分布。
- 所以 $\alpha = 0.05$ 时可以用 $\{|T| > 1.96\}$ 作为拒绝 $H_0 : \mu = \mu_0$ 的标准。
- 为了使得这种近似有意义，样本量 n 一般不小于 30，最好是 50 以上，或 100 以上。

6.3 假设检验的某些概念和数学描述

假设检验的某些概念和数学描述

- 检验法；
- 两类错误；
- 功效函数；
- 临界值与 p 值；
- 假设检验与置信区间的联系。

6.3.1 检验法与功效函数

零假设和对立假设

- 要检验的“假设”记作 H_0，叫做**零假设**或**原假设**。
- H_0 是关于随机变量 X(总体) 的一个“看法”。数学描述：
- 设 X 的分布函数为 $F(x,\theta)$，其中 $\theta \in \Theta$，这里 Θ 是实数（或向量、或其它符号）组成的已知集合。
- “看法”H_0 表示成：$\theta \in \Theta_0$，这里 Θ_0 是 Θ 的非空真子集。
- 把 $\theta \in \Theta - \Theta_0$ 叫做**对立假设**或 **备择假设**，记作 H_a。
- 例如，在例 2.1 中，考虑铜丝折断力的总体，要检验的假设是 $\mu = 570$。

- 这里

$$
\begin{aligned}
X \sim & \mathrm{N}(\mu, 8^2) \\
\theta = & \mu \\
\Theta = & (-\infty, \infty) \\
\Theta_0 = & \{570\} \\
\Theta - \Theta_0 = & (-\infty, 570) \cup (570, \infty) \\
H_0 : & \theta \in \Theta_0 \\
H_a : & \theta \in \Theta - \Theta_0
\end{aligned}
$$

检验法

- 检验法就是给出一个规则，对给定的样本值 $x_1, x_2, \ldots, x_n$，进行明确表态：接受假设 H_0 还是拒绝假设 H_0。
- 设 S 是所有可能的样本值 $(x_1, x_2, \ldots, x_n)(n$ 固定) 组成的集合（样本空间）。很多情况下 $S = R^n$。
- 检验法就是指空间 S 的一个划分：$S = S_1 \cup S_2$，$S_1 \cap S_2 = \emptyset$。当 $(x_1, x_2, \ldots, x_n) \in S_1$ 时，接受假设 H_0；当 $(x_1, x_2, \ldots, x_n) \in S_2$ 时，拒绝 H_0。
- S_1 叫做**接受域**, S_2 叫做**否定域**或**拒绝域**。
- 否定域与检验法互相唯一确定。

假设检验的两类错误

- 在解决假设检验的问题时, 无论作出否定还是接受原假设 H_0 的决定, 都有可能犯错误.
- 我们称否定 H_0 时犯的错误为第一类错误, 接受 H_0 时犯的错误为第二类错误.

		检验结果	
		H_0	H_1
真实情况	H_0	✓	X(第 I 类)
	H_1	X(第 II 类)	✓

- 假设检验一般控制第一类错误在检验水平 α 以下，所以否定 H_0 时结论比较可靠。
- 如果承认 H_0，可能犯第二类错误，错误概率可能会比较大。
- 要同时减小两类错误是不可能的，一种错误减小另一种错误必然增大。

功效函数

- 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本，W 是否定域，当事件 $\{(X_1, X_2, \ldots, X_n) \in W\}$ 发生时拒绝零假设 H_0。
- 设 $X \sim F(x, \theta)$，此事件的概率记作

$$M_W(\theta) = P_\theta((X_1, X_2, \ldots, X_n) \in W), \quad \theta \in \Theta$$

  称作否定域 W(或对应的检验法) 的**功效函数**。
- 功效函数在两种情况下的意义:

		意义	记号
真实情况	$H_0(\theta \in \Theta_0)$	第 I 类错误概率	$\alpha_W(\theta)$
	$H_1(\theta \in \Theta_1)$	$1 -$ 第二类错误概率	$1 - \beta_W(\theta)$

检验水平

- 在 H_0 成立的时候，功效函数越小越好；
- 当 H_1 成立的时候，功效函数越大越好。
- 功效函数图示。

- **定义 3.1** 给定 $\alpha(0<\alpha<1)$，如果

$$\alpha_W(\theta)\le\alpha,\quad \forall\theta\in\Theta_0$$

则称 W 是**检验水平** (或**显著性水平**) 为 α 的否定域（检验法）。

- 如果

$$\sup_{\theta\in\Theta_0}\alpha_W(\theta)=\alpha$$

则称 α 为 W 的精确检验水平。

检验法讨论

- 统计学中假设检验一般都是固定一个检验水平 α，找到 H_1 下第二类错误尽可能小（功效尽可能大）的检验法进行检验。
- 这样，如果最后结论是拒绝 H_0，可能犯的是第 I 类错误，此错误概率收到检验水平的控制，所以结论是比较可信的。
- 如果最后结论是接受 H_0，可能犯的是第 II 类错误，此错误概率虽然已经尽可能控制但不像第 I 类错误那样有明确界限。
- 所以接受 H_0 的结论是一种“维持原状、不改变原来结论”之类的做法，一般不能把接受 H_0 作为一个新的发现使用。
- 这和“带概率的反证法”是一致的，反证法不能推出矛盾，并不意味着假设一定成立。

6.3.2 临界值和 p 值

临界值

- 否定域 W 一般用一个直观上有明确意义的统计量（称为**检验统计量**）$\varphi(X_1,X_2,\dots,X_n)$ 来确定。
- 单边情形的否定域：

$$W=\{(x_1,x_2,\dots,x_n):\ \varphi(x_1,x_2,\dots,x_n)>\lambda\}\tag{3.1}$$

- 双边情形的否定域：

$$W = \big\{(x_1, x_2, \ldots, x_n) : \varphi(x_1, x_2, \ldots, x_n) < \lambda_1 \text{或} \\ \varphi(x_1, x_2, \ldots, x_n) > \lambda_2\big\} \tag{3.2}$$

- (3.1) 中 λ 叫做单边情形的临界值，(3.2) 中 λ_1 和 λ_2 叫做双边情形的临界值。
- 临界值根据检验水平确定。

单侧临界值确定

- 为了使得 H_0 成立时第一类错误概率不超过检验水平 α，应找 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) = \alpha \tag{3.3}$$

- 如果检验统计量服从连续型分布，这样的 λ 存在。
- 如果检验统计量服从离散型分布，这样的 λ 不一定存在，这时找 λ 使得

$$\begin{aligned} &\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) \le \alpha \\ &< \sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \lambda) \end{aligned} \tag{3.4}$$

  (这是水平为 α 的最小的 λ)
- 注意：第一类错误概率不是越小越好！所以 λ 不是越大越好。

双边情形的临界值

- 对于双边情形，应找 λ_1 和 λ_2 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) < \lambda_1) = \frac{\alpha}{2} \tag{3.5}$$

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda_2) = \frac{\alpha}{2} \tag{3.6}$$

- 如果检验统计量服从连续型分布，这样的 λ_1 和 λ_2 存在。

- 如果检验统计量服从离散型分布，这样的 λ_1 和 λ_2 不一定存在，这时找 λ_1 和 λ_2 使得

$$\begin{aligned} &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)<\lambda_1)\le\frac{\alpha}{2}\\ <&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\le\lambda_1) \qquad (3.7)\\ &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)>\lambda_2)\le\frac{\alpha}{2}\\ <&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\ge\lambda_2) \qquad (3.8)\end{aligned}$$

- 根据检验水平确定临界值从而获得否定域的方法，称为**临界值方法**。

- **例 3.1** 设 $X\sim \mathrm{N}(\mu,\sigma^2)$，未知 σ，检验假设

$$H_0:\mu\le\mu_0 \longleftrightarrow H_a:\mu>\mu_0$$

- 令 $\theta=(\mu,\sigma^2)$，及

$$\begin{aligned}\Theta=&\{(\mu,\sigma^2):\mu\in(-\infty,\infty),\sigma^2>0\}\\ \Theta_0=&\{(\mu,\sigma^2):\mu\le\mu_0,\sigma^2>0\}\\ \Theta_1=&\Theta-\Theta_0=\{(\mu,\sigma^2):\mu>\mu_0,\sigma^2>0\}\end{aligned}$$

- 检验可表示为

$$H_0:\theta\in\Theta_0 \longleftrightarrow H_a:\theta\in\Theta_1$$

- 设样本为 $X_1,X_2,\ldots,X_n$，取检验统计量

$$\varphi(X_1,X_2,\ldots,X_n)=\frac{\bar{X}-\mu_0}{\sqrt{S^2/n}}$$

- φ 值越大，数据越倾向于否定 H_0 而接受 H_1。取单边否定域

$$\begin{aligned} W =& \{(x_1, x_2, \ldots, x_n) : \varphi(x_1, x_2, \ldots, x_n) > \lambda\} \\ =& \left\{(x_1, x_2, \ldots, x_n) : \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} > \lambda\right\} \end{aligned}$$

- 对给定检验水平 α，临界值 λ 应满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) = \alpha$$

- 当 $\mu \leq \mu_0$ 时，

$$\frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

- 于是

$$\sup_{\theta \in \Theta_0} P_\theta \left(\frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} > \lambda\right) = \sup_{\theta \in \Theta_0} P_\theta \left(\frac{\bar{X} - \mu}{\sqrt{S^2/n}} > \lambda\right)$$

- 其中 $\frac{\bar{X} - \mu}{\sqrt{S^2/n}}$ 对任意 (μ, σ^2) 都服从 $\mathrm{t}(n-1)$ 分布，所以可取 λ 为 $\mathrm{t}(n-1)$ 分布的 $1-\alpha$ 分位数，则上式右边为 α。

- 否定域求得。

p 值方法

- p 值方法是确定否定域的另一方法，p 值可以直观表示数据中否定 H_0 的倾向的强烈程度。

- 以单边否定域 (3.1) 为例:

$$W = \{(x_1, x_2, \ldots, x_n) : \ \varphi(x_1, x_2, \ldots, x_n) > \lambda\} \tag{3.1}$$

- 设 $x_1, x_2, \ldots, x_n$ 是样本值，$X_1, X_2, \ldots, X_n$ 是总体分布参数为 θ 时的样本。

- 定义

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ = & \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n) \right) \end{aligned}$$

- $p(x_1, x_2, \ldots, x_n)$ 是 H_0 成立的条件下统计量 $\varphi(X_1, X_2, \ldots, X_n)$ 取到和观测到的统计量值 $\varphi(x_1, x_2, \ldots, x_n)$ 一样大或更大的概率的最大值。
- 这个概率是数据取值倾向于否定 H_0 的情况的概率。
- 如果 H_0 成立，这个概率应该很小。
- 注意：

  (1) $p(x_1, x_2, \ldots, x_n)$ 取值于 $[0, 1]$;

  (2) $p(X_1, X_2, \ldots, X_n)$ 是统计量。

p 值与否定域

- 用 p 值可以表示否定域：
- **引理 3.1** 设对给定的 $\alpha \in (0, 1)$，恰有一个 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) = \alpha \tag{3.10}$$

  则

$$\varphi(x_1, x_2, \ldots, x_n) > \lambda \Longleftrightarrow p(x_1, x_2, \ldots, x_n) < \alpha$$

- 即：当且仅当 p 值小于检验水平 α 时否定零假设 H_0:

$$W = \{(x_1, x_2, \ldots, x_n) : p(x_1, x_2, \ldots, x_n) < \alpha\}$$

- 这种确定否定域的方法叫做 **p 值方法**。
- 这个否定域的精确检验水平为 α。

p 值特点

- p 值越小，从数据看否定 H_0 的倾向越强烈。
- p 值大小代表了数据与 H_0 的相容程度，当 p 值小于 α 时就不相容了。
- p 值总是在 $[0,1]$ 取值的。
- p 值是能够否定 H_0 可取的最小的检验水平 α_0，取比 p 值更小的检验水平 $\alpha < \alpha_0$ 就不能否定 H_0 了。

引理 3.1 的证明

- 设 $p(x_1, x_2, \ldots, x_n) < \alpha$, 即

$$\sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)\right) < \alpha \qquad (*)$$

- 而 λ 满足

$$\sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) > \lambda\right) = \alpha$$

- 所以一定有

$$\varphi(x_1, x_2, \ldots, x_n) > \lambda$$

- 这是因为如果

$$\varphi(x_1, x_2, \ldots, x_n) \leq \lambda$$

则

$$\begin{aligned} &\varphi(X_1, X_2, \ldots, X_n) > \lambda \\ \Longrightarrow &\varphi(X_1, X_2, \ldots, X_n) > \varphi(x_1, x_2, \ldots, x_n) \\ \Longrightarrow &\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n) \end{aligned}$$

从而

$$\sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)\right) \geq \alpha$$

- 反过来，设 $\varphi(x_1, x_2, \ldots, x_n) > \lambda$, 则有 $\varepsilon > 0$ 使得

$$\varphi(x_1, x_2, \ldots, x_n) - \varepsilon > \lambda$$

- 于是

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ = & \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)\right) \\ \leq & \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) > \varphi(x_1, x_2, \ldots, x_n) - \varepsilon\right) \\ \leq & \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) > \lambda\right) = \alpha \qquad (**) \end{aligned}$$

- 且 (**) 式的小于等于号必为严格小于号，否则令

$$\lambda' = \varphi(x_1, x_2, \ldots, x_n) - \varepsilon$$

  则 $\lambda' > \lambda$ 也是

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) = \alpha$$

  的解，与引理 3.1 题设矛盾。

- 即：$\varphi(x_1, x_2, \ldots, x_n) > \lambda$ 时有

$$p(x_1, x_2, \ldots, x_n) < \alpha$$

- 引理 3.1 证毕。

精确水平不等于 α 的情况

- 给定 $\alpha \in (0, 1)$ 不一定有临界值 λ 使得检验的精确水平为 α。

- 这时求满足 (3.4) 的 λ。
- p 值定义不变。
- **引理 3.2** 设对给定的 $\alpha \in (0,1)$，有 λ 满足

$$\begin{aligned} &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) \le \alpha \\ &< \sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \lambda) \end{aligned} \tag{3.11}$$

  则

$$\varphi(x_1, x_2, \ldots, x_n) > \lambda \Longleftrightarrow p(x_1, x_2, \ldots, x_n) \le \alpha$$

- **引理 3.2 的证明** 设 $\varphi(x_1, x_2, \ldots, x_n) > \lambda$，则

$$\begin{aligned} &P_\theta\big(\varphi(X_1, X_2, \ldots, X_n) \ge \varphi(x_1, x_2, \ldots, x_n)\big) \\ \le &P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) \end{aligned}$$

- 两边取 $\sup_{\theta\in\Theta_0}$，有

$$\begin{aligned} &p(x_1, x_2, \ldots, x_n) \\ = &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \varphi(x_1, x_2, \ldots, x_n)) \\ \le &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda) \le \alpha \quad (\text{由 (3.11)}) \end{aligned}$$

- 反过来，如果 $\varphi(x_1, x_2, \ldots, x_n) \le \lambda$, 则

$$\begin{aligned} &P_\theta\big(\varphi(X_1, X_2, \ldots, X_n) \ge \varphi(x_1, x_2, \ldots, x_n)\big) \\ \ge &P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \lambda) \end{aligned}$$

- 两边取 $\sup_{\theta\in\Theta_0}$，有

$$\begin{aligned} &p(x_1, x_2, \ldots, x_n) \\ = &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \varphi(x_1, x_2, \ldots, x_n)) \\ \ge &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \ge \lambda) > \alpha \quad (\text{由 (3.11)}) \end{aligned}$$

- 因此 $p(x_1, x_2, \ldots, x_n) \le \alpha$ 时一定有 $\varphi(x_1, x_2, \ldots, x_n) > \lambda$。
- 引理 3.2 证毕。

- 在引理 3.2 条件下，若 λ 满足 (3.11)(不一定唯一)，则否定域

$$W = \{(x_1, x_2, \ldots, x_n) : \varphi(x_1, x_2, \ldots, x_n) > \lambda\}$$

  的精确检验水平不超过 α，是一个检验水平为 α 的否定域。
- 样本值落入否定域 W 的充分必要条件是 p 值小于等于 α，与引理 3.1 的做法只有微小的的差别（引理 3.1 要求 p 值严格小于 α）。
- 这也叫做 p 值方法。
- **例 3.2** 设 $X \sim \mathrm{N}(\mu, \sigma^2)$。未知 σ，检验 $H_0 : \mu \le \mu_0$。
- 否定域为

$$\begin{aligned} W =& \{(x_1, x_2, \ldots, x_n) : \varphi(x_1, x_2, \ldots, x_n) > \lambda_0\} \\ \varphi(x_1, x_2, \ldots, x_n) =& \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \end{aligned}$$

  λ_0 为 $\mathrm{t}(n-1)$ 的 $1-\alpha$ 分位数。
- 计算 p 值（是统计量）

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \ge \varphi(x_1, x_2, \ldots, x_n)\right) \end{aligned}$$

- 这里 $\theta = (\mu, \sigma^2)$, $\Theta_0 = \{(\mu, \sigma^2) : \mu \le \mu_0, \sigma^2 > 0\}$。

- 当 $\mu \le \mu_0$ 时

$$\frac{\bar{x} - \mu}{\sqrt{S^2/n}} \ge \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}}$$

- 所以

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& \sup_{\theta \in \Theta_0} P_\theta \left(\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \geq \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \right) \\ =& P_{(\mu_0, \sigma^2)} \left(\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \geq \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \right) \\ =& P\left(T \geq \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \right) \end{aligned} \tag{*}$$

  (其中 T 表示 $\mathrm{t}(n-1)$ 分布随机变量)

- 由 (*) 可见（或由引理 3.1 可知）

$$p(x_1, x_2, \ldots, x_n) < \alpha \Longleftrightarrow \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} > \lambda$$

- 例如，为了检验 $H_0 : \mu \leq 25$，设某样本 $n = 64$，$\bar{x} = 25.9$, $S^2 = 17.3$，则

$$\begin{aligned} \varphi(x_1, x_2, \ldots, x_n) =& \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = 1.731 \\ p(x_1, x_2, \ldots, x_n) =& P(T \geq 1.731) = 0.044 < 0.05 \end{aligned}$$

  (这个 p 值可以在 R 软件中用 `1 - pt(1.731, 63)` 计算)

- 在 $\alpha = 0.05$ 检验水平下应拒绝 H_0。

双边否定域的 p 值

- 考虑否定域 (3.2):

$$\begin{aligned} W = \big\{ & (x_1, x_2, \ldots, x_n) : \ \varphi(x_1, x_2, \ldots, x_n) < \lambda_1 \text{或} \\ & \varphi(x_1, x_2, \ldots, x_n) > \lambda_2 \big\} \end{aligned} \tag{3.2}$$

- 不去考虑由 α 确定 λ_1 和 λ_2 的问题，直接定义 p 值。
- 根据“p 值是能够拒绝 H_0 的最小的可取检验水平”，可以找到特定的 λ_0，$\lambda_1 \leq \lambda_0 < \lambda_2$。

- 当 $\varphi(x_1, x_2, \ldots, x_n) \leq \lambda_0$ 时，令

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& \min\left\{ \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \leq \varphi(x_1, x_2, \ldots, x_n)\right),\ 1 \right\} \end{aligned} \tag{3.12}$$

- 当 $\varphi(x_1, x_2, \ldots, x_n) > \lambda_0$ 时，令

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& \min\left\{ \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)\right),\ 1 \right\} \end{aligned} \tag{3.13}$$

- **定义 3.3** 由 (3.12) 和 (3.13) 定义的 $p(x_1, x_2, \ldots, x_n)$ 叫做双边情形下样本值 $(x_1, x_2, \ldots, x_n)$ 的 p 值。
- **引理 3.3** 设对给定的 $\alpha \in (0, 1)$，有唯一的 λ_1 和 λ_2 满足

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) < \lambda_1) = \frac{\alpha}{2} \tag{3.14}$$

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda_2) = \frac{\alpha}{2} \tag{3.15}$$

则

$$\begin{aligned} & \varphi(x_1, x_2, \ldots, x_n) < \lambda_1 \text{ 或} \varphi(x_1, x_2, \ldots, x_n) > \lambda_2 \\ \Longleftrightarrow & p(x_1, x_2, \ldots, x_n) < \alpha. \end{aligned}$$

- 否定域为 p 值小于 α。

- **引理 3.4** 设对给定的 $\alpha \in (0,1)$，有 λ_1 和 λ_2 满足

$$\begin{aligned} &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)<\lambda_1) \le \frac{\alpha}{2} \\ <&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\le\lambda_1) \qquad (3.16) \\ &\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)>\lambda_2) \le \frac{\alpha}{2} \\ <&\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\ge\lambda_2) \qquad (3.17) \end{aligned}$$

  则

$$\begin{aligned} &\varphi(x_1,x_2,\ldots,x_n)<\lambda_1 \text{ 或} \varphi(x_1,x_2,\ldots,x_n)>\lambda_2 \\ \Longleftrightarrow &p(x_1,x_2,\ldots,x_n)\le\alpha. \end{aligned}$$

- 与引理 3.3 的差别是否定域为 p 值小于等于 α。

- **引理 3.3 的证明** 设 $\varphi(x_1,x_2,\ldots,x_n)<\lambda_1$, 则 $\varphi(x_1,x_2,\ldots,x_n)<\lambda_0$（只要 $\lambda_0\in[\lambda_1,\lambda_2)$）。

- 而且存在 $\varepsilon>0$ 使得 $\varphi(x_1,x_2,\ldots,x_n)<\lambda_1-\varepsilon$。

- 由 (3.12)

$$\begin{aligned} &p(x_1,x_2,\ldots,x_n) \\ \le&2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)\le\varphi(x_1,x_2,\ldots,x_n)) \\ \le&2\sup_{\theta\in\Theta_0} P_\theta(\varphi(X_1,X_2,\ldots,X_n)<\lambda_1-\varepsilon)<\alpha \\ &\text{(由 (3.17) 对 } \lambda_1 \text{ 的要求及唯一性)} \end{aligned}$$

- 若 $\varphi(x_1,x_2,\ldots,x_n)>\lambda_2$, 则 $\varphi(x_1,x_2,\ldots,x_n)>\lambda_0$（只要 $\lambda_0\in[\lambda_1,\lambda_2)$）。

- 而且存在 $\varepsilon>0$ 使得 $\varphi(x_1,x_2,\ldots,x_n)>\lambda_2+\varepsilon$。

- 由 (3.13) 得

$$\begin{aligned}
&p(x_1, x_2, \ldots, x_n) \\
\leq& 2 \sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)) \\
\leq& 2 \sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) > \lambda_2 + \varepsilon) < \alpha \\
&\text{(由 (3.17) 对 } \lambda_2 \text{ 的要求及唯一性)}
\end{aligned}$$

- 于是我们证明了引理 3.3 的 "$\Longrightarrow$"。

- 反过来，设 $p(x_1, x_2, \ldots, x_n) < \alpha$。

- 若 $\varphi(x_1, x_2, \ldots, x_n) \leq \lambda_0$, 则

$$\begin{aligned}
&p(x_1, x_2, \ldots, x_n) \\
=& 2 \sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \leq \varphi(x_1, x_2, \ldots, x_n)) \\
&\text{(这是因为 } p(x_1, x_2, \ldots, x_n) < \alpha \text{ 时 (3.12) 不能取 1)} \\
<& \alpha
\end{aligned}$$

- 由 (3.14) 中 λ_1 的唯一性用反证法可知这时

$$\varphi(x_1, x_2, \ldots, x_n) < \lambda_1$$

- 若 $\varphi(x_1, x_2, \ldots, x_n) > \lambda_0$, 则

$$\begin{aligned}
&p(x_1, x_2, \ldots, x_n) \\
=& 2 \sup_{\theta \in \Theta_0} P_\theta(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n)) \\
&\text{(这是因为 } p(x_1, x_2, \ldots, x_n) < \alpha \text{ 时 (3.13) 不能取 1)} \\
<& \alpha
\end{aligned}$$

- 由 (3.15) 中 λ_2 的唯一性用反证法可知这时

$$\varphi(x_1, x_2, \ldots, x_n) > \lambda_2$$

- 这样，引理 3.3 的必要性和充分性都证明了。

- 从引理 3.3 证明看出，在取定了检验水平 α 以后，λ_0 可以取区间 $[\lambda_1, \lambda_2)$ 中的任何值。
- 这样当可以拒绝 H_0 时，p 值定义不受 λ_0 选取的影响。当 H_0 相容时，(3.12) 和 (3.13) 两个 p 值定义受到 λ_0 选取影响但都会做出相容的判断。
- 如果在引理 3.3 条件下定义

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& 2 \min \left\{ \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \leq \varphi(x_1, x_2, \ldots, x_n) \right), \right. \\ & \left. \sup_{\theta \in \Theta_0} P_\theta \left(\varphi(X_1, X_2, \ldots, X_n) \geq \varphi(x_1, x_2, \ldots, x_n) \right) \right\} \end{aligned}$$

  则引理 3.3 仍成立，证明类似，但不再需要 λ_0。
- 引理 3.4 证明类似。

- 引理 3.3 和引理 3.4 给出了双边检验问题的 p 只方法。
- 在引理 3.3 条件下，否定域为 p 值小于 α。
- 在引理 3.4 条件下，否定域为 p 值小于等于 α。
- 两种情况的检验水平都不超过 α。
- 优点：适用于任何检验水平；p 值大小给出了拒绝零假设的强烈程度。

- 现代的统计软件中假设检验的结果一般都给出 p 值。
- **例 3.3** 设 $X \sim \mathrm{N}(\mu, \sigma^2)$，$\mu, \sigma^2$ 均未知。
- 检验

$$H_0 : \sigma^2 = \sigma_0^2 \longleftrightarrow H_a : \sigma^2 \neq \sigma_0^2$$

- 用检验统计量

$$\varphi(X_1, X_2, \ldots, X_n) = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- 当 φ 太大或太小时拒绝 H_0。
- 使用双边否定域

$$\begin{aligned} W = \{&(x_1, x_2, \ldots, x_n) : \varphi(x_1, x_2, \ldots, x_n) < \lambda_1 \\ &\text{或} \varphi(x_1, x_2, \ldots, x_n) > \lambda_2\} \end{aligned}$$

- 从直观看 $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ 可以作为 σ^2 的估计。
- H_0 成立时 $\frac{1}{n-1}\varphi(x_1, x_2, \ldots, x_n)$ 应该与 1 相差不大，$\varphi(x_1, x_2, \ldots, x_n)$ 应该与 $n-1$ 相差不大。
- λ_1 应该小于 $n-1$，λ_2 应该大于 $n-1$。取 $\lambda_0 = n-1$，则 $\lambda_1 < n-1 < \lambda_2$。
- $\varphi(x_1, x_2, \ldots, x_n) \leq n-1 \Leftrightarrow S^2 \leq \sigma_0^2$.
- 当 $S^2 \leq \sigma_0^2$ 时定义

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ =& \min \left\{ 2 \sup_{\substack{\mu \in (-\infty, \infty) \\ \sigma^2 = \sigma_0^2}} P_{(\mu, \sigma^2)}(\varphi(X_1, X_2, \ldots, X_n) \leq \varphi_0), 1 \right\} \\ =& \min \{2P(\xi \leq \varphi_0), 1\} \end{aligned}$$

- 其中 $\varphi_0 = \varphi(x_1, x_2, \ldots, x_n)$, ξ 为 $\chi^2(n-1)$ 随机变量。
- 当 $S^2 > \sigma_0^2$ 时类似有

$$p(x_1, x_2, \ldots, x_n) = \min\{2P(\xi \geq \varphi_0), 1\}$$

- 回到例 2.4 的 10 个铜丝折断力数据，要检验 $\sigma^2 = 64$。
- $S^2 = 75.7 > 64$。
- $\varphi_0 = \frac{1}{\sigma_0^2}\sum_{i=1}^{10}(x_1 - \bar{x})^2 = 10.65.$
- $P(\xi \geq \varphi_0) = 0.30 (\xi \sim \chi^2(9))$。
- p 值为 0.60。
- 只要检验水平 α 不大于 0.60，H_0 都是相容的。
- 但是注意：假设检验应该预先确定检验水平，而不能看到 p 值后再选检验水平。

6.3.3 假设检验与置信区间的联系

假设检验与置信区间的联系

- 在 §6.2.2 讲关于 $\mu = \mu_0$ 的检验时，我们讨论了检验的接受域与 μ 的置信区间的联系。
- 这种联系是一般性的。
- 设 X 的分布函数为 $F(x, \theta)$，θ 是未知参数, $\theta \in \Theta$, $X_1, X_2, \ldots, X_n$ 是 X 的样本。
- 对任何 $\theta_0 \in \Theta$, 考虑假设问题

$$H_0 : \theta = \theta_0 \longleftrightarrow H_a : \theta \neq \theta_0$$

- 设 $A(\theta_0)$ 是 H_0 的检验水平为 α 的接受域（其补集是 H_0 的检验水平为 α 的否定域）。

- 即当且仅当 $(x_1, X_2, \ldots, x_n) \in A(\theta_0)$ 时接受 $H_0 : \theta = \theta_0$，且

$$\begin{aligned} P_{\theta_0}((X_1, X_2, \ldots, X_n) \notin A(\theta_0)) \leq & \alpha \\ P_{\theta_0}((X_1, X_2, \ldots, X_n) \in A(\theta_0)) \geq & 1 - \alpha \end{aligned} \qquad (*)$$

- 令

$$S(x_1, x_2, \ldots, x_n) = \{\theta : (x_1, x_2, \ldots, x_n) \in A(\theta)\} \qquad (3.18)$$

- 则

$$(x_1, x_2, \ldots, x_n) \in A(\theta_0) \Longleftrightarrow \theta_0 \in S(x_1, x_2, \ldots, x_n) \qquad (**)$$

- 由 (*) 和 (**)，得

$$P_{\theta_0}(\theta_0 \in S(X_1, X_2, \ldots, X_n)) \geq 1 - \alpha$$

- 其中的 θ_0 是任意的，所以

$$P_\theta(\theta \in S(x_1, x_2, \ldots, x_n)) \geq 1 - \alpha \quad \forall \theta \in \Theta$$

- 可见，θ 的集合 $S(x_1, x_2, \ldots, x_n)$ 如果是个区间，则它是 θ 的置信水平为 $1 - \alpha$ 的置信区间。
- 我们可以用 $\theta = \theta_0$ 的检验接受域来构造 θ 的置信区间。
- 反过来，如果统计量 $\underline{\theta}(X_1, X_2, \ldots, X_n)$ 和 $\overline{\theta}(X_1, X_2, \ldots, X_n)$ 使得

$$P_\theta\left(\theta \in (\underline{\theta}, \overline{\theta})\right) \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

  则 $\theta_0 \in (\underline{\theta}, \overline{\theta})$ 可以作为 $H_0 : \theta = \theta_0$ 的检验水平为 α 的接受域。

- **例 3.4** 设 $X \sim \mathrm{N}(\theta, 1)$，$\theta \in (-\infty, \infty)$。

- $(X_1, X_2, \ldots, X_n)$ 是 X 的样本。
- 对任何 θ_0，考虑检验问题

$$H_0 : \theta = \theta_0 \longleftrightarrow H_a : \theta \neq \theta_0$$

- 从 §1 知可用接受域

$$A(\theta_0) = \{(x_1, x_2, \ldots, x_n) : |\bar{x} - \theta_0| \leq c\}$$

- 其中

$$c = \sqrt{\frac{1}{n}} z_{1-\frac{\alpha}{2}}$$

- z_p 表示标准正态分布的 p 分位数：

$$\begin{aligned} z_p =& \Phi^{-1}(p) \\ \Phi(p) =& p \end{aligned}$$

- 这时检验的精确水平为 α。
- 有

$$P_{\theta_0}(|\bar{X} - \theta_0| \leq c) = 1 - \alpha$$

- 因 θ_0 任意，所以

$$P_\theta(|\bar{X} - \theta| \leq c) = 1 - \alpha, \quad \forall \theta \in (-\infty, \infty)$$

- 于是

$$P_\theta(\bar{X} - c \leq \theta \leq \bar{X} + c) = 1 - \alpha, \quad \forall \theta \in (-\infty, \infty)$$

- 从接受域得到了 θ 的置信水平为 $1 - \alpha$ 的置信区间。
- 反过来，当且仅当 θ_0 属于这个置信区间时接受 $H_0 : \theta = \theta_0$，检验水平为 α。

6.4 两个正态总体的假设检验

两个正态总体的假设检验

- 设 $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, 且 X, Y 相互独立。
- (1) 未知 σ_1^2, σ_2^2，但知道 $\sigma_1^2 = \sigma_2^2$，检验假设

$$H_0 : \mu_1 = \mu_2$$

- (2) 未知 μ_1, μ_2，检验假设

$$H_0 : \sigma_1^2 = \sigma_2^2$$

- (3) 未知 μ_1, μ_2，检验假设

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$

- (4) 未知 σ_1^2, σ_2^2，但知道 $\sigma_1^2 \neq \sigma_2^2$，检验假设

$$H_0 : \mu_1 = \mu_2$$

6.4.1 独立两样本 t 检验

独立两样本 t 检验

- 在已知两个独立正态总体方差相等时检验两个总体的期望是否相等，使用“独立两样本 t 检验”。
- 在实际问题中，用来比较独立的两组的同一属性平均来说有无显著差异。
- **例 4.1(即例 1.4)** 在针织品的漂白工艺过程中，要考察温度对针织品的断裂强力（主要质量指标）的影响。

- 为了比较 70°C 与 80°C 的影响有无差别，在这两个温度下，分别重复做了 8 次试验，得数据如下（单位：千克力）

$$\begin{aligned} 70^\circ\text{C时的强力：}&20.5,\ 18.8,\ 19.8,\ 20.9,\\ &21.5,\ 19.5,\ 21.0,\ 21.2\\ 80^\circ\text{C时的强力：}&17.7,\ 20.3,\ 20.0,\ 18.8,\\ &19.0,\ 20.1,\ 20.2,\ 19.1 \end{aligned}$$

- 从试验数据看，两种温度下的强度有无区别？

- 用 X, Y 分别表示 70°C 与 80°C 下的断裂强力，试验结果按常识判断是独立的。
- 根据过去的经验，可以认为 X, Y 分别服从正态分布且方差相等。
- 即 $X \sim \mathrm{N}(\mu_1, \sigma^2)$，$Y \sim \mathrm{N}(\mu_2, \sigma^2)$，要检验 $H_0 : E(X) = E(Y)$，即 $H_0 : \mu_1 = \mu_2$。
- 自然想到比较两个样本的平均值。70°C 的样本的平均强力为 20.4 千克力，80°C 的样本的平均强力为 19.4 千克力，70°C 的平均值高出 1 千克力。

- 能否就此断言 70°C 下的总体期望值 $E(X)$ 与 80°C 的总体期望值 $E(Y)$ 不同？
- 要注意的是，我们只是在一组样本中观测到了样本平均值有 1 千克力的差距。换一组样本，差距可能就变了。
- 样本得到的两个样本均值的差即使在 $\mu_1 = \mu_2$ 时也一般不会等于零。
- 如果这个差距很大，我们可以比较有把握地说两个总体期望不同。
- 如果差距很小，我们可以认为两个总体期望相同。

- 问题是，如何找这个临界值？
- 思路是找到问题中的随机性的分布，然后在 $\mu_1 = \mu_2$ 情况下产生较大差距为小概率事件时否定 H_0。

两样本 t 检验的推导

- 设 $x_1, x_2, \ldots, x_n$ 是来自 X 的样本值，$y_1, y_2, \ldots, y_n$ 是来自 Y 的样本值。
- 设法研究样本平均值的差

$$\bar{x} - \bar{y}$$

  的分布，当此差超出了 $\mu_1 = \mu_2$ 时随机变化的正常范围时否定 $H_0 : \mu_1 = \mu_2$。
- 还是“带概率的反证法”的思想。先假设 $\mu_1 = \mu_2$，看是否有小概率事件发生。
- 在假设 $\mu_1 = \mu_2$ 的条件下，$\bar{X} - \bar{Y}$ 的分布仍与未知的 σ^2 有关。其方差等于

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}$$

- 于是

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{2\sigma^2}{n}}} \sim \mathrm{N}(0, 1)$$

- 这不是统计量。估计 σ^2 为

$$\hat{\sigma}^2 = \frac{1}{2n-2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]$$

  称为合并方差估计。

- 在 $H_0 : \mu_1 = \mu_2$ 成立时统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{2\hat{\sigma}^2}{n}}} \sim \mathrm{t}(2n-2) \tag{4.2}$$

- 查 $\mathrm{t}(2n-2)$ 分布临界值表可得 λ 满足

$$P(|T| > \lambda) = \alpha$$

- 从样本中计算统计量 T 的值，当 $|T| > \lambda$ 时拒绝 $H_0 : \mu_1 = \mu_2$，当 $|T| \leq \lambda$ 时 H_0 相容。

- 称为**两样本 t 检验**。

例 4.1（续）

- 把两样本 t 检验一般步骤应用于例 4.1。

- (1) 提出待检验的假设:

$$H_0 : E(X) = E(Y) \longleftrightarrow H_a : E(X) \neq E(Y)$$

- (2) 计算 t 统计量的值。

$$\begin{aligned}
&n = 8 \quad \bar{x} = 20.4 \quad \bar{y} = 19.4 \\
&\sum_{i=1}^{n}(x_i - \bar{x})^2 = 6.20 \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 5.80 \\
&\hat{\sigma}^2 = \frac{1}{2 \times 8 - 2}(6.20 + 5.80) = 0.8571 \\
&T = \frac{20.4 - 19.4}{\sqrt{\frac{2*0.8571}{8}}} = 2.16
\end{aligned}$$

- (3) 查 t 分布表，自由度是 $2n-2=14$，取 $\alpha=0.05$，得 $\lambda=2.145$。
- (4) 下结论：$|T|=2.16>\lambda$ 所以否定零假设，认为 70°C 下的总体期望值 $E(X)$ 与 80°C 的总体期望值 $E(Y)$ 不等，而且是 70°C 下的强力更大。
- 第 (3)、(4) 步也可以计算 p 值:

$$p(x_1,x_2,\ldots,x_n)=P(|\xi|\geq|2.16|)=0.0486<\alpha$$

其中 ξ 为服从 $\mathrm{t}(2n-2)$ 分布的随机变量。

两样本的样本量不等情况

- 设 $X_1,X_2,\ldots,X_{n_1}$ 是来自 $\mathrm{N}(\mu_1,\sigma^2)$ 的样本，$Y_1,Y_2,\ldots,Y_{n_2}$ 是来自 $\mathrm{N}(\mu_2,\sigma^2)$ 的样本。
- 这时定义 σ^2 的估计为 (合并方差估计)

$$\hat{\sigma}^2=\frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2\right]$$

- 检验统计量为

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{\sigma}^2}}$$

- 在 $H_0:\mu_1=\mu_2$ 下 $T\sim\mathrm{t}(n_1+n_2-2)$, 使用、$\mathrm{t}(n_1+n_2-2)$ 分布临界值，当且仅当 $|T|>\lambda$ 时拒绝 H_0。
- 这个检验叫做**两样本 t 检验**，也叫平均数的显著性鉴定。

- 如果拒绝了 $H_0 : \mu_1 = \mu_2$，一般称（在 α 水平下）两个总体的平均数**有显著 (性) 差异**。
- **例 4.2** 研究口服避孕药对妇女血压影响。
- 对某公司工作的 35 岁至 39 岁的非怀孕妇女，用抽查方法收集到如下数据。
- 有 8 人使用口服避孕药，其收缩压平均值为 132.86（mmHg），标准差为 15.35。
- 有 21 人未使用，收缩压平均值为 127.44(mmHg)，标准差为 18.23。
- 问：这两种血压的平均值的差异是否显著？

- **解** 假设使用口服避孕药的妇女的收缩压总体为 $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$，假设不使用的妇女的收缩压总体为 $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$，并假定 $\sigma_1^2 = \sigma_2^2 = \sigma^2$。
- 计算得

$$\begin{aligned}\hat{\sigma}^2 =& \frac{1}{8+21-2}((8-1)\times 15.35^2 + (21-1)\times 18.23^2)\\ =& 294.95\end{aligned}$$

- T 统计量为

$$T = \frac{132.86 - 127.44}{\sqrt{\left(\frac{1}{8}+\frac{1}{21}\right)294.95}} = 0.760$$

- 查 27 个自由度的 t 分布的 $\alpha = 0.05$ 临界值得 $\lambda = 2.052$，$|T| < \lambda$ 所以 H_0 相容，两个平均值无显著差异。

成对数据的比较

- 有些实际问题中的数据是成对的，比如同一个人两次测量同一指标。

- 这时，两个变量 X 和 Y 一般是不独立的，不能使用上述的两样本 t 检验。
- 设 (X,Y) 的样本为 $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$。
- 考虑新的总体 $Z=X-Y$，样本为 $Z_i=X_i-Y_i, i=1,2,\ldots,n$。
- 如果 (X,Y) 服从联合正态分布则 Z 也服从正态分布。
- 要检验 $H_0: E(X)=E(Y)$，只要检验等价的假设 $E(Z)=0$。
- 问题化为单样本 t 检验问题。
- **例 4.3** 为了鉴定两种工艺方法生产的产品某性能指标有无显著差异，对于 9 批材料分别用两种工艺进行生产，得到该指标的 9 对数据如下：

0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
0.10	0.21	0.52	0.32	0.78	0.59	0.68	0.77	0.89

- 问：根据以上数据，能否说两种工艺生产的产品性能指标有显著差异？（检验水平 $\alpha=0.05$）

- 计算 9 对数据的差：

0.10	0.09	-0.12	0.18	-0.18	0.11	0.12	0.13	0.11

- 设差的总体为 $Z\sim \mathrm{N}(\mu,\sigma^2)$，检验 $H_0:\mu=0$。
- 用 §2 中的 t 检验，计算得

$$T=\frac{\bar{x}}{\sqrt{S^2/9}}=1.467$$

- 查 t(8) 分布临界值表得 $\alpha=0.05$ 对应临界值 $\lambda=2.306$。
- 或计算 p 值:

$$p=P(|\xi|>1.467)=0.19$$

- 现在 $|T|<\lambda$（p 值小于 α），所以 H_0 相容，两种工艺方法生产的产品性能指标无显著差异。
- 也称这两种工艺方法对产品该性能指标**无显著影响**。

方差的双边检验

- 设 $X_1, X_2, \ldots, X_{n_1}$ 是来自总体 $\mathrm{N}(\mu_1, \sigma_1^2)$ 的样本，$Y_1, Y_2, \ldots, Y_{n_2}$ 是来自总体 $\mathrm{N}(\mu_2, \sigma_2^2)$ 的样本，两个总体独立。
- 要求检验

$$H_0 : \sigma_1^2 = \sigma_2^2$$

- 用“带概率的反证法”。先假设 H_0 成立。
- 要比较 σ_1^2 和 σ_2^2，想到比较其估计量：

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$
$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

- 取检验统计量

$$F = \frac{S_1^2}{S_2^2}$$

  当 F 远大于 1 或远小于 1 时拒绝 H_0。
- 如果 F 的分布不依赖于未知参数，则可以取 λ_1 和 λ_2 使得

$$P(F < \lambda_1) = \frac{\alpha}{2}, \quad P(F > \lambda_2) = \frac{\alpha}{2}$$

  当 $F < \lambda_1$ 或 $F > \lambda_2$ 时拒绝 H_0 （这样第 I 类错误概率等于 α，是水平 α 的检验法）

F 分布

- 如果随机变量 Z 有如下分布密度

$$
\begin{aligned}
&f_{n_1,n_2}(u)\\
&=\begin{cases}
\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}u^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}u\right)^{-\frac{n_1+n_2}{2}}, & u>0\\
0 & u\le 0
\end{cases}
\end{aligned}
$$

  则称 Z 服从自由度为 n_1, n_2 的 F 分布，这里 n_1, n_2 是两个正整数，分别称为第一自由度和第二自由度（或分子自由度和分母自由度），记 $Z \sim \mathrm{F}(n_1, n_2)$。

- 上面的统计量 F 在 $H_0 : \sigma_1^2 = \sigma_2^2$ 成立时服从 $\mathrm{F}(n_1 - 1, n_2 - 1)$ 分布。

F 分布的临界值

- F 分布的临界值在第 434–439 页附表 4–附表 6 中，只给出了右侧临界值：

$$P(F > \lambda) = \alpha$$

- F 分布有性质：

$$F \sim \mathrm{F}(n_1 - 1, n_2 - 1) \Longrightarrow \frac{1}{F} \sim \mathrm{F}(n_2 - 1, n_1 - 1)$$

- 所以对于 $H_0 : \sigma_1^2 = \sigma_2^2$ 的检验，为求临界值 λ_1 和 λ_2，先查自由度为 $(n_1 - 1, n_2 - 1)$ 的水平 $\alpha/2$ 的临界值 λ_2。

- 再查自由度为 (n_2-1, n_1-1) 的水平 $\alpha/2$ 的临界值 λ_1'，并令 $\lambda_1 = 1/\lambda_1'$。

- **例 4.1(续)** 考虑 70°C 和 80°C 下强力的方差的比较。

- 这里 $n_1 = n_2 = 8$，$s_1^2 = 6.20/7$, $s_2^2 = 5.80/7$，

$$F = \frac{s_1^2}{s_2^2} = \frac{6.20}{5.80} = 1.07$$

- 查 F 分布表，取 $\alpha = 0.05$，查自由度 $(7, 7)$ 的 0.025 临界值得

$$\lambda_2 = 4.99$$

  对于 λ_1，这里自由度交换后仍为 $(7, 7)$，得

$$\lambda_1 = \frac{1}{4.99} = 0.200$$

- 现在 $\lambda_1 < F < \lambda_2$ 所以 H_0 相容，在 0.05 水平下不能否认 70°C 下和 80°C 下有相同的方差（或称：70°C 下和 80°C 下的方差**无显著差异**）。

- **例 4.4（例 4.2 续）** 检验使用口服避孕药和不使用的妇女的血压的方差是否相等：

$$H_0 : \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_a : \sigma_1^2 \neq \sigma_2^2$$

- 这里 $n_1 = 8$, $n_2 = 21$, $S_1^2 = (15.35)^2$, $S_2^2 = (18.23)^2$，

$$F = \frac{S_1^2}{S_2^2} = 0.709$$

- 查自由度 $(7, 20)$ 的 0.025 水平 F 分布临界值得 $\lambda_2 = 3.01$, 查自由度 $(20, 7)$ 的 0.025 水平 F 分布临界值得 $\lambda_1 = 1/4.42 = 0.226$。

- 现在 $\lambda_1 < F < \lambda_2$，所以 H_0 相容，在 0.05 水平下不能否认两组人的血压有相同的方差（或称：两组人的方差**无显著差异**）。

关于查表

- 上例中 λ_2 在附表 5 中直接查到。而 $1/\lambda_1$ 需要找自由度 $(20, 7)$ 情况下的临界值，表中 $n_2 = 7$ 的行存在，但 $n_1 = 20$ 的列不存在。

- 怎么处理？

- 简化的做法是找最近的表格值：最近的自由度是 $(24, 7)$，表格值为 4.42。

- 更精细一点的做法是用线性插值近似：要求解的临界值和 $(12, 7)$ 自由度下的 4.67 以及 $(24, 7)$ 自由度下的 4.42 最接近，按照自由度接近程度进行插值：

$$1/\lambda_1 = 4.67 + \frac{4.42 - 4.67}{24 - 12}(20 - 12) = 4.50$$

- 如果两个自由度都不在表格中，这种方法要两次插值，而且精度也只是略有提高。

- 用统计软件计算更容易。如在 R 中

```
qf(0.975, 20, 7)
```

返回 4.46674。

6.4.2 两总体方差单边检验

两总体方差单边检验

- 考虑未知 μ_1, μ_2，$H_0 : \sigma_1^2 \leq \sigma_2^2$ 检验问题。

- **例 4.5** 有两台车床生产同一种型号的滚珠。可认为直径分别服从正态分布。

- 从这两台车床生产的产品中分别抽取 8 个和 9 个，测得滚珠直径如下（毫米）：

甲车床:15.0, 14.5, 15.2, 15.5, 14.8, 15.1, 15.2, 14.8

乙车床:15.2, 15.0, 14.8, 15.2, 15.0, 15.0, 14.8, 15.1, 14.8

- 问：乙车床产品直径的方差是否比甲车床的小？

- **解** 用 X, Y 分别表示甲、乙两车床的产品的直径。
- 设 $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$，$X, Y$ 独立。
- 问题是想判断 $\sigma_1^2 > \sigma_2^2$ 是否成立，把它作为对立假设：

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \longleftrightarrow H_a : \sigma_1^2 > \sigma_2^2$$

- 这样在拒绝 H_0 后就可使断言乙车床的方差较小。

方差单边检验的推导

- 设 $X_1, X_2, \ldots, X_{n_1}$ 是来自总体 X 的样本，$Y_1, Y_2, \ldots, Y_{n_2}$ 是来自总体 Y 的样本，$X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$，$X, Y$ 独立。
- 要检验

$$H_0 : \sigma_1^2 \leq \sigma_2^2 \longleftrightarrow H_a : \sigma_1^2 > \sigma_2^2$$

- H_0 相当于 $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$，用统计量

$$F = \frac{S_1^2}{S_2^2}$$

  当 F 超过 1 且很大时拒绝 H_0。
- F 在 H_0 下的分布依赖于 σ_1^2, σ_2^2 的值。
- 令

$$\tilde{F} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

  则 $\tilde{F} \sim \mathrm{F}(n_1 - 1, n_2 - 1)$, 且在 $H_0 : \sigma_1^2 \leq \sigma_2^2$ 下

$$\tilde{F} = \frac{\sigma_2^2}{\sigma_1^2} F \geq F$$

- 取 λ 使

$$P(\tilde{F} > \lambda) = \alpha$$

则

$$\begin{aligned} F > \lambda \Longrightarrow & \tilde{F} > \lambda \\ P(F > \lambda) \leq & P(\tilde{F} > \lambda) = \alpha \end{aligned}$$

- 于是

$$W = \{F > \lambda\}$$

是 H_0 的水平 α 的否定域。

- 回到例 4.5。
- 第一步：提出待检验的假设 $H_0 : \sigma_1^2 \leq \sigma_2^2$。
- 第二步：计算统计量 $F = S_1^2/S_2^2$ 的值。

$$\begin{aligned} & n_1 = 8, \quad n_2 = 9 \\ & S_1^2 = 0.09554, \quad S_2^2 = 0.02611 \\ & F = S_1^2/S_2^2 = 3.66 \end{aligned}$$

- 查 $\alpha = 0.05$ 的 $\mathrm{F}(7, 8)$ 分布右侧临界值得 $\lambda = 3.50$。
- 现在 $F = 3.66 > 3.50$，在 0.05 水平下否定零假设，认为乙车床产品直径的方差显著小于甲车床产品直径的方差。
- **例 4.6** 赈灾捐赠的男女差异。随机抽查了 25 个男士，平均捐赠 12.40 美元，标准差 2.50 美元；随机抽查了 25 个女士，平均捐赠 8.90 美元，标准差 1.34 美元。
- 问：男士捐赠额的方差是否大于女士捐赠额的方差？

- **解** 设一个男士的捐赠额 X 服从 $\mathrm{N}(\mu_1,\sigma_1^2)$，一个女士的捐赠额 Y 服从 $\mathrm{N}(\mu_2,\sigma_2^2)$，问题 $\sigma_1^2>\sigma_2^2$ 作为对立假设：

$$H_0:\sigma_1^2\le\sigma_2^2 \longleftrightarrow H_a:\sigma_1^2>\sigma_2^2$$

- 计算 F 统计量得 $F=(2.50)^2/(1.34)^2=3.48$，查自由度 $(24,24)$ 的 F 分布水平 0.01 的临界值得 $\lambda=2.66$。$F=3.48>2.66$ 所以拒绝 H_0，在 0.01 水平下男士捐赠额的方差显著大于女士捐赠额的方差。

6.4.3 方差不等时均值的比较

方差不等时均值的比较

- 如果未知 σ_1^2,σ_2^2 但知道 $\sigma_1^2\ne\sigma_2^2$，检验 $H_0:\mu_1=\mu_2$，这个问题称为 Behrens–Fisher 问题。
- 可以建立一个统计量分布在 H_0 下近似服从 t 分布的检验法。
- 设 $X_1,X_2,\dots,X_{n_1}$ 是来自总体 X 的样本，$Y_1,Y_2,\dots,Y_{n_2}$ 是来自总体 Y 的样本，$X\sim\mathrm{N}(\mu_1,\sigma_1^2)$, $Y\sim\mathrm{N}(\mu_2,\sigma_2^2)$，$\sigma_1^2\ne\sigma_2^2$, X,Y 独立。

- 令

$$\bar X=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i \qquad \bar Y=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$$
$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar X)^2 \qquad S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar Y)^2$$

- 易知

$$\bar X-\bar Y\sim\mathrm{N}\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$
$$\frac{\bar X-\bar Y-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim\mathrm{N}(0,1)$$

- 在 $H_0 : \mu_1 = \mu_2$ 下

$$\xi = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0, 1)$$

- 在 $|\xi|$ 太大时应该拒绝 H_0。但是 ξ 不是统计量。
- 用估计量代替 σ_1^2, σ_2^2，得统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- 在 H_0 下 T 的精确分布复杂且依赖于 σ_1^2/σ_2^2 的值，但 T 近似服从 $\mathrm{t}(m^*)$:

$$m^* = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2} \tag{4.8}$$

- 查表时可取 m^* 为最近的整数。查 t 分布临界值表找到 λ 使 $P(|T| > \lambda) = \alpha$，当且仅当 $|T| > \lambda$ 时拒绝 $H_0 : \mu_1 = \mu_2$。
- 类似可解决当 $\sigma_1^2 \neq \sigma_2^2$ 时 $H_0 : \mu_1 \leq \mu_2$ 的检验问题。
- **例 4.7** 研究父亲患心脏病的家庭中子女的胆固醇水平是否偏高的问题。
- 随机调查了 100 个 2 到 14 岁的孩子（父亲死于心脏病），其胆固醇水平平均值为 207.3，标准差为 35.6;
- 另外随机调查了父亲无心脏病史的 74 个 2 至 14 岁的孩子，其胆固醇水平平均值为 193.4，标准差为 17.3。
- 问：前者的胆固醇水平的平均值与后者的胆固醇水平的平均值是否有显著差异？

- **解** 设父亲死于心脏病的孩子的胆固醇水平 $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, 父亲无心脏病史的孩子的胆固醇水平 $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$，其中 $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 都未知，要检验

$$H_0 : \mu_1 = \mu_2 \longleftrightarrow H_a : \mu_1 \neq \mu_2$$

- 首先判断 $\sigma_1^2 = \sigma_2^2$ 是否相等。设 $\alpha = 0.05$，计算得 $F = 4.23$，查表得 $\lambda_1 = 0.6548$, $\lambda_2 = 1.5491$，$F = 4.23 > \lambda_2$，所以方差有显著差异。

- 计算 T 统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = 3.40$$

- 计算得 $m^* = 151.4$，$m = 151$，查 t(151) 的双侧 0.05 水平临界值得 $\lambda = 1.980$(使用 t(120) 的临界值)。现在 $|T| = 3.40 > \lambda$ 所以拒绝 H_0，认为在 0.05 水平下胆固醇水平的平均值有显著差异，父亲死于心脏病的孩子的胆固醇水平更高。

t 分布与 F 分布的关系

- 设 $X \sim \mathrm{t}(n)$，则 $Y = X^2 \sim \mathrm{F}(1, n)$。

- **证** 设 $X \sim f(x)$，则

$$\begin{aligned} F_Y(y) =& P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ =& F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ p_Y(y) =& F_Y'(y) = p_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + p_X(\sqrt{y}) \frac{1}{2\sqrt{y}} \\ =& \frac{1}{2\sqrt{y}} [p_X(\sqrt{y}) + p_X(-\sqrt{y})], \quad y > 0 \end{aligned}$$

- $\mathrm{t}(n)$ 分布的分布密度为

$$p_X(x) = \frac{\Gamma\left(\frac{n+1}{n}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

- 所以

$$\begin{aligned} p_X(\sqrt{y}) =& p_X(-\sqrt{y}) = \frac{\Gamma\left(\frac{n+1}{n}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y}{n}\right)^{-\frac{n+1}{2}} \\ p_Y(y) =& \frac{1}{\sqrt{y}}\frac{\Gamma\left(\frac{n+1}{n}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y}{n}\right)^{-\frac{n+1}{2}} \\ =& \frac{\Gamma\left(\frac{1+n}{n}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{n}\right)^{\frac{1}{2}} y^{\frac{1}{2}-1}\left(1+\frac{y}{n}\right)^{-\frac{n+1}{2}} \\ \sim& \mathrm{F}(1,n) \end{aligned}$$

- 于是

$$P(|X|>\lambda) = P(Y>\lambda^2)$$

6.5 比率的假设检验

比率的假设检验

- 设 $X\sim \mathrm{b}(1,p)$, $0<p<1$ 是未知参数，p 就是“比率”，如成功率、失败率、有效率等。
- 本节包括：
- 单总体比率的小样本单边、双边假设检验；
- 单总体比率的大样本单边、双边假设检验；
- 两总体比率的大样本单边、双边假设检验。

6.5.1 单总体比率检验的大样本方法

大样本情况下单个比例的假设检验

- 对单个比例 p，设 S 是 n 个独立抽样中成功的个数，则 $S \sim \mathrm{B}(n,p)$，当 n 很大时（一般要求成功和失败个数都超过 5 个），根据中心极限定理，S 近似服从正态分布 $\mathrm{N}(np, np(1-p))$。
- 令 $\hat{p} = S/n$ 为样本中成功比例。则

$$\eta = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

近似服从标准正态分布。

- 当 n 很大时另一近似为

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

近似服从标准正态分布。

双侧检验

- 对双侧问题

$$H_0 :\ p = p_0 \longleftrightarrow H_1 :\ p \neq p_0$$

在 H_0 下

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

近似服从标准正态分布。

- 给定检验水平 α，取否定域为

$$W = \left\{ \frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}} \geq z_{1-\frac{\alpha}{2}} \right\} \tag{6.7}$$

- 其中 $z_{1-\frac{\alpha}{2}}$ 是标准正态分布双侧 α 分位数。

右侧检验

- 对右侧检验问题

$$H_0:\ p \le p_0 \longleftrightarrow H_1:\ p > p_0$$

- 取否定域为

$$W = \left\{ \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \ge z_{1-\alpha} \right\} \tag{6.8}$$

- 其中 $z_{1-\alpha}$ 是标准正态分布右侧 α 分位数。

左侧检验

- 对左侧检验问题

$$H_0:\ p \ge p_0 \longleftrightarrow H_1:\ p < p_0$$

- 取否定域为

$$W = \left\{ \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_{\alpha} \right\} \tag{6.9}$$

- 其中 z_{α} 是标准正态分布左侧 α 分位数。
- 这样的比例检验方法称为**正态逼近法**。

例

- 收藏家一年中购入了 98 幅名画，经鉴定其中 26 幅是赝品。能否认为该收藏家的鉴定准确率大于等于 75%?
- **解答：** 准确率 p 的估计为 $\hat{p} = (98-26)/98 = 0.7347$，低于 75%。
- 取单侧检验的方向为

$$H_0: p \ge 0.75 \longleftrightarrow H_1: p < 0.75$$

- 用正态近似法，取检验水平 $\alpha = 0.05$, $n = 98$, 否定域为

$$\left\{ \frac{\hat{p} - 0.75}{\sqrt{\hat{p}(1-\hat{p})/n}} < -1.645 \right\}$$

- 计算得

$$\frac{\hat{p} - 0.75}{\sqrt{\hat{p}(1-\hat{p})/n}} = -0.3432$$

未落入否定域，在 0.05 水平下没有充分证据拒绝鉴定准确率大于等于 75% 的假设，但是也不能认为有证据表明准确率大于等于 75%。

- 注意，如果我们取假设的方向为右侧问题

$$H_0 : p \leq 0.75 \longleftrightarrow H_1 : p > 0.75$$

- 0.05 水平否定域为

$$\left\{ \frac{\hat{p} - 0.75}{\sqrt{\hat{p}(1-\hat{p})/n}} > 1.645 \right\}$$

- 现在统计量

$$\frac{\hat{p} - 0.75}{\sqrt{\hat{p}(1-\hat{p})/n}} = -0.3432$$

也没有落入否定域，在 0.05 水平下不能拒绝鉴定准确率小于等于 75% 的假设。

- 这是通常的假设检验方法的局限性：当不能拒绝 H_0 时，最后的结论往往是不确定的。

6.5.2 单总体比率检验的小样本方法

单总体比率右侧假设检验小样本方法

- 设总体 $X \sim \mathrm{b}(1, p)$，$X_1, X_2, \ldots, X_n$ 为样本。

- 检验单边假设问题

$$H_0 : p \leq p_0 \longleftrightarrow H_a : p > p_0$$

- 用统计量

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

当 $\hat{p}$ 超过 p_0 很多时拒绝 H_0。

- 改用统计量

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

当 S 很大时拒绝 H_0。

- 否定域是

$$W = \{S \geq c\}$$

- 临界值 c 取为满足

$$\sup_{p \leq p_0} P_p(S \geq c) \leq \alpha \tag{5.1}$$

的最小整数。

- $S \sim B(n,p)$。有恒等式

$$\begin{aligned} P_p(S \geq k) =& \sum_{i=k}^{n} C_n^i p^i (1-p)^{n-i} \\ =& \frac{n!}{(k-1)!(n-k)!} \int_0^p u^{k-1}(1-u)^{n-k} du \end{aligned}$$

- $P_p(S \geq c)$ 作为 p 的函数是增函数，所以 (5.1) 化为

$$P_{p_0}(S \geq c) \leq \alpha \tag{5.2}$$

- 求临界值 c 比较麻烦，我们用 p 值来表示否定域。
- 设样本值为 $x_1, x_2, \ldots, x_n$，记 $S_0 = \sum_{i=1}^{n} x_i$，则 p 值为

$$\begin{aligned} & p(x_1, x_2, \ldots, x_n) \\ = & \sup_{p \le p_0} P_p(S \ge S_0) = P_{p_0}(S \ge S_0) \\ = & \sum_{i=S_0}^{n} C_n^i p_0^i (1-p_0)^{n-i} \end{aligned}$$

- 当且仅当 p 值小于等于 α：

$$\sum_{i=S_0}^{n} C_n^i p_0^i (1-p_0)^{n-i} \le \alpha \tag{5.3}$$

  时拒绝 H_0。
- 统计软件中可以计算这样的 p 值。

- (5.3) 左边是 p 的严格单调递增函数，为了判断 (5.3) 是否成立，还可以解方程

$$\sum_{i=S_0}^{n} C_n^i p_0^i (1-p_0)^{n-i} = \alpha$$

  得唯一解 $p_\alpha(S_0)(S_0 \ge 1$ 时)。
- $S_0 = 0$ 时规定 $p_\alpha(0) = 0$。
- 这样 $S_0 \ge c$ 当且仅当 $p_0 \le p_\alpha(S_0)$。
- 设 $F_b(n_1, n_2)$ 表示 $\mathrm{F}(n_1, n_2)$ 分布的 b 分位数，则

$$\begin{aligned} & p_\alpha(S_0) \\ = & \left\{ 1 + \frac{n - S_0 + 1}{S_0} F_{1-\alpha}(2(n - S_0 + 1), 2S_0) \right\}^{-1} \end{aligned} \tag{5.5}$$

- 原药物有效率为 0.80，制药公司声称新药有效率高于 0.80，且药价更低。
- 收集了临床数据，使用新药的病人中随机抽查了 30 人，其中 27 人有效。
- 问：能否认为新药的有效率高于 0.80?
- **解** 用 X 表示新药的效果，$X = 1$ 表示有效，$X = 0$ 表示无效，$p = P(X = 1)$。
- $n = 30$, $S_0 = \sum_{i=1}^{n} x_i = 27$。
- 取检验水平 $\alpha = 0.05$，从 (5.5) 得

$$\begin{aligned} p_\alpha(S_0) =& p_{0.05}(27) = \left\{1 + \frac{4}{27} F_{0.95}(8, 54)\right\}^{-1} \\ =& \left(1 + \frac{4}{27} \times 2.13\right)^{-1} = 0.76 < p_0 = 0.80 \end{aligned}$$

- 所以不能拒绝 $H_0 : p \le p_0$，没有理由说新药比原来药物有更高的有效率。
- 用 R 软件可以计算上面公式中的 p 值:

```
> 1 - pbinom(26, 30, 0.8)
[1] 0.1227108
```

结果为 0.1227，超过 α，结果不显著。

- R 软件还可以直接做小样本的比例假设检验，如：

```
> binom.test(27, 30, p=0.8, alternative="greater")
        Exact binomial test

data:  27 and 30
number of successes = 27, number of trials = 30,
  p-value =0.1227
```

```
alternative hypothesis: true probability of success
  is greater than 0.8
95 percent confidence interval:
 0.7614021 1.0000000
sample estimates:
probability of success
                  0.9
```

- 如果 30 人中有 28 人有效，则 p 值为 0.0442，在 0.05 水平下显著。
- 这个问题结果不显著，但是不能断言新药不如老药，有可能增大样本量后可以得到显著结果。

单总体比率左侧假设检验

- 再来考虑单总体假设检验

$$H_0 : p \geq p_0 \longleftrightarrow H_a : p < p_0$$

- 否定域为

$$W = \{S \leq c\}$$

- c 是满足

$$\sup_{p \geq p_0} P_p(S \leq c) \leq \alpha \tag{5.6}$$

  的最大整数。

- $P_p(S \leq c)$ 是 p 的严格单调递减函数。(5.6) 化为

$$P_{p_0}(S \leq c) \leq \alpha \tag{5.7}$$

- 用 p 值方法。p 值为

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) =& P_{p_0}(S \le S_0) \\ =& \sum_{i=0}^{S_0} C_n^i p_0^i (1-p_0)^{n-i} \end{aligned}$$

- 当且仅当 p 值小于等于 α 时拒绝 H_0。
- 也可以解方程

$$\sum_{i=0}^{S_0} C_n^i p_0^i (1-p_0)^{n-i} = \alpha \tag{5.9}$$

得 $\tilde{p}_\alpha(S_0)$，当且仅当 $\tilde{p}_\alpha(S_0) \le p_0$ 时拒绝 H_0。

单总体比率的双侧检验

- 设 $X_1, X_2, \ldots, X_n$ 是 $X \sim \mathrm{B}(n,p)$ 的样本，$S = \sum_{i=1}^n X_i$。
- 检验

$$H_0 : p = p_0 \longleftrightarrow H_a : p \ne p_0$$

- 当 S 太大或太小时拒绝 H_0。否定域为

$$W = \{S \le c_1 \text{或} S \ge c_2\}$$

- 其中 c_1, c_2 取为最大的整数 c_1 和最小的整数 c_2 满足

$$P_{p_0}(S \le c_1) \le \frac{\alpha}{2}, \quad P_{p_0}(S \ge c_2) \le \frac{\alpha}{2}$$

- 用 p 值方法。设样本值为 $x_1, x_2, \ldots, x_n$，$S_0 = \sum_{i=1}^n x_i$。

- p 值为

$$\begin{aligned} &p(x_1, x_2, \ldots, x_n) \\ =&2\min\{P_{p_0}(S \le S_0), P_{p_0}(S \ge S_0)\} \end{aligned}$$

- 当且仅当 p 值小于等于 α 时拒绝 H_0。

6.5.3 两总体比率比较

两总体比率比较

- 设两个总体 X 与 Y 相互独立，$X \sim \mathrm{b}(1, p_1)$, $Y \sim \mathrm{b}(1, p_2)$, 样本分别为 $X_1, X_2, \ldots, X_{n_1}$, $Y_1, Y_2, \ldots, Y_{n_2}$。
- 考虑如下三个假设检验问题：

$$\begin{aligned} H_0 : p_1 \le p_2 &\longleftrightarrow H_a : p_1 > p_2 \\ H_0 : p_1 \ge p_2 &\longleftrightarrow H_a : p_1 < p_2 \\ H_0 : p_1 = p_2 &\longleftrightarrow H_a : p_1 \ne p_2 \end{aligned}$$

- 令

$$S_1 = \sum_{i=1}^{n_1} X_i \qquad\qquad S_2 = \sum_{i=1}^{n_2} Y_i$$

- 则

$$\hat{p}_1 = \frac{S_1}{n_1} \qquad\qquad \hat{p}_2 = \frac{S_2}{n_2}$$

  分别是 p_1 和 p_2 的估计。
- 当 $\hat{p}_1$ 远大于 $\hat{p}_2$ 时拒绝 $H_0 : p_1 \le p_2$;
- 当 $\hat{p}_1$ 远小于 $\hat{p}_2$ 时拒绝 $H_0 : p_1 \ge p_2$;
- 当 $\hat{p}_1$ 和 $\hat{p}_2$ 相差很多时拒绝 $H_0 : p_1 = p_2$。
- 使用两种方法：大样本情形的正态近似方法和 Fisher 精确检验方法。

正态近似法

- 由注意到 $\hat{p}_1$ 和 $\hat{p}_2$ 分别是两个样本的样本平均值，所以

$$D(\hat{p}_1) = \frac{D(X)}{n_1} = \frac{p_1(1-p_1)}{n_1} \qquad D(\hat{p}_2) = \frac{D(Y)}{n_2} = \frac{p_2(1-p_2)}{n_2}$$

- 令

$$\xi = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{5.11}$$

$$\eta = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{5.12}$$

$$\zeta = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} \tag{5.13}$$

- 其中

$$\hat{p} = \frac{1}{n_1 + n_2}(n_1\hat{p}_1 + n_2\hat{p}_2) = \frac{S_1 + S_2}{n_1 + n_2}$$

- 当 n_1 和 n_2 相当大（一般要求 $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$）时，$\xi$ 近似服从标准正态分布。

- 有

$$P(\xi > z_{1-\alpha}) \approx \alpha$$

两总体比率右侧检验的正态近似

- 在 $H_0 : p_1 \leq p_2$ 成立时，$\xi \geq \eta$，

$$P(\eta > z_{1-\alpha}) \leq P(\xi > z_{1-\alpha}) \approx \alpha$$

- 所以

$$H_0 : p_1 \leq p_2 \longleftrightarrow H_a : p_1 > p_2$$

的否定域可取为

$$W = \{\eta > z_{1-\alpha}\}$$

两总体比率左侧检验的正态近似

- 类似地，对

$$H_0 : p_1 \geq p_2 \longleftrightarrow H_a : p_1 < p_2$$

- 在 H_0 下 $\xi \leq \eta$。
- 从而

$$P(\eta < z_\alpha) \leq P(\xi < z_\alpha) \approx \alpha$$

- 否定域可取为

$$W = \{\eta < z_\alpha\}$$

两总体比率双侧检验的正态近似

- 考虑检验问题

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 \neq p_2$$

- 当 $H_0 : p_1 = p_2$ 成立时，只要 n_1, n_2 相当大（一般要求 $n_1\hat{p}_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2(1-\hat{p}_2) \geq 5$），统计量 ζ 近似服从标准正态分布，

$$P(|\zeta| > z_{1-\frac{\alpha}{2}}) \approx \alpha$$

- 否定域可取为

$$W = \{|\zeta| > z_{1-\frac{\alpha}{2}}\}$$

- **例 5.2** 研究口服避孕药对年龄在 40 至 44 岁的妇女的心脏的影响。
- 5000 个使用口服避孕药的妇女中三年内出现心肌梗死的有 13 人；
- 在 10000 个不服用口服避孕药的妇女中出现心肌梗死的有 7 人。
- 问：两组人的心肌梗死比率是否有显著差异？

- **解** 用 p_1 表示使用口服避孕药的妇女中三年内出现心肌梗死的比率，p_2 表示不使用口服避孕药的妇女中三年内出现心肌梗死的比率。
- 要检验

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 \neq p_2$$

- 统计量

$$\begin{aligned}\hat{p}_1 =& \frac{13}{5000} = 0.0026\\ \hat{p}_2 =& \frac{7}{10000} = 0.0007\\ \hat{p} =& \frac{13+7}{5000+10000} = 0.00133\end{aligned}$$

- 近似条件：$n_1\hat{p}_1(1-\hat{p}_1) = 6.66 \geq 5$, $n_2\hat{p}_2(1-\hat{p}_2) = 6.70 \geq 5$，可以用统计量 ζ。
- $\zeta = 3.01$，水平取为 $\alpha = 0.01$，否定域为 $W = \{|\zeta| > z_{1-\frac{0.01}{2}}\} = \{|\zeta| > 2.58\}$。
- $|\zeta| > 2.58$，拒绝 H_0，在 0.01 水平下两组的心肌梗死比率有显著差异。
- 或称口服避孕药对心肌梗死比率有显著影响。

两样本比率比较的 Fisher 精确检验

- Fisher 精确检验计算精确 p 值，不要求大样本。
- 令 $S_1 = \sum_{i=1}^{n_1} X_i$, $S_2 = \sum_{i=1}^{n_2} Y_i$。
- 令

$$S_1^0 = \sum_{i=1}^{n_1} x_i,\ S_2^0 = \sum_{i=1}^{n_2} y_i,\ t = S_1^0 + S_2^0 \tag{5.14}$$

- 考虑检验问题

$$H_0 : p_1 \le p_2 \longleftrightarrow H_a : p_1 > p_2$$

- p 值为

$$p_1(S_1^0) = \sum_{i \ge S_1^0} p(i) \tag{5.15}$$

- 其中

$$p(i) = \frac{\binom{n_1}{i}\binom{n_2}{t-i}}{\binom{n_1+n_2}{t}} \quad (i = 0, 1, \dots)$$

- 当且仅当 p 值小于等于 α 时拒绝 $H_0 : p_1 \le p_2$。

- 对于假设

$$H_0 : p_1 \ge p_2 \longleftrightarrow H_a : p_1 < p_2$$

- p 值为

$$p_1(S_1^0) = \sum_{i \le S_1^0} p(i) \tag{5.17}$$

- 对于双边假设

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 \neq p_2$$

- p 值的一种算法为

$$p_3(S_1^0) = 2 \min \left\{ \sum_{i \le S_1^0} p(i), \ \sum_{i \ge S_1^0} p(i) \right\} \tag{5.18}$$

- 计算 $p(i)$ 的递推公式

$$p(i+1) = p(i) \frac{(n_1 - i)(t - i)}{(i+1)(n_2 - t + i + 1)} \tag{5.19}$$

- 对于双边假设

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 \neq p_2$$

- 双侧 p 值的另一种算法为 (R 函数 fisher.test() 中使用此公式）

$$p_4(S_1^0) = \sum_{i : p(i) \le p(S_1^0)} p(i)$$

即给定 n_1, n_2, t 情况下的所有四个表中 $p(i)$ 值等于当前表的概率以及比当前表概率更小的概率之和。

- **例 5.3** 某公安局有两个专案组，在过去一年内一组接手 25 件人命案，侦破了 23 件；二组接手 35 件人命案，侦破了 30 件。

- 问：两个组的侦破能力有无差别？

- **解** 比较两个组的侦破率。

- 设两个组的侦破率分别为 p_1, p_2。要检验 $H_0 : p_1 = p_2$。

- 用 Fisher 精确检验。
- $n_1 = 25$, $n_2 = 35$, $S_1^0 = 23$, $S_2^0 = 30$, $t = 53$。
- p 值为

$$p_3(S_1^0) = 2\min\left\{\sum_{i=0}^{23} p(i),\ \sum_{i=23}^{25} p(i)\right\}$$

- 其中

$$\begin{aligned} p(23) =& 0.252 \\ p(24) =& p(23)\frac{2\times 30}{24\times 6} = 0.105 \\ p(25) =& p(24)\frac{1\times 29}{25\times 7} = 0.017 \end{aligned}$$

- 于是

$$\begin{aligned} \sum_{i=23}^{25} p(i) =& 0.374 \\ \sum_{i=0}^{23} p(i) =& 1 - \sum_{i=23}^{25} p(i) + p(23) = 0.878 \\ p_3(S_1^0) =& 2\times 0.374 = 0.748 > 0.05 \end{aligned}$$

- 在 0.05 水平下不应拒绝 $H_0 : p_1 = p_2$。
- 两个专案组在破案能力上没有显著差异。

6.6 总体的分布函数的假设检验

总体分布函数的假设检验

- 假设检验的参数方法先假定总体服从某种带有未知参数的分布（常用正态分布），然后回答针对总体参数的问题。
- 还可以不假定总体分布类型，直接回答分布有关的问题，如总体是否来自正态分布。
- 如何判断一个总体 X 是否分布函数为 $F(x)$?
- 有时候从学科知识可以建模得到，如前面放射性粒子数服从泊松分布的模型推导。
- 很多情况下只能从观测数据判断。
- 一般先作直方图（对连续型总体），推测可能的分布类型，再进行检验。

拟合优度卡方检验

- 检验

$$H_0 : X\text{的分布函数为}F(x)$$

- 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本。
- 类似于画直方图，在数轴上取 m 个点：$t_1, t_2, \ldots, t_m$（$t_1 < t_2 < \cdots < t_m$），把数轴 $(-\infty, +\infty)$ 分成 $m+1$ 段：

$$(-\infty, t_1],\ (t_1, t_2],\ \ldots,\ (t_{m-1}, t_m],\ (t_m, +\infty)$$

- 用 ν_i 表示 $X_1, X_2, \ldots, X_n$ 中落入第 i 段的个数（$i = 1, 2, \ldots, m+1$）。这里 ν_i 是频数，$\frac{\nu_i}{n}$ 是频率。

- 在 H_0 下 X 落入第 i 段的概率 p_i 可计算：

$$\begin{aligned} p_1 =& P(X \le t_1) = F(t_1) \\ p_i =& P(t_{i-1} < X \le t_i) = F(t_i) - F(t_{i-1}), \\ & i = 2, 3, \ldots, m \\ p_{m+1} =& P(X > t_m) = 1 - F(t_m) \end{aligned}$$

- 根据概率与频率的关系（大数定律），如果 H_0 成立，频率 $\frac{\nu_i}{n}$ 应该接近于概率 p_i。
- 定义

$$V=\sum_{i=1}^{m+1}\left(\frac{\nu_i}{n}-p_i\right)^2\cdot\frac{n}{p_i}=\sum_{i=1}^{m+1}\frac{(\nu_i-np_i)^2}{np_i}$$

- 当 V 很大时拒绝 H_0。
- 其中多乘的因子 $\frac{n}{p_i}$ 的作用是放大 p_i 很小的项的作用，分布密度曲线的形状恰好是更多由两侧 (对应 p_i 很小) 的形状决定，如果不乘这个因子则 p_i 很小的项作用被严重低估。

- 另一解释：记

$$\xi_i=\frac{\nu_i}{n}-p_i$$

- 则 H_0 成立时 $\nu_i\sim\mathrm{B}(n,p_i)$，

$$E(\xi_i)=0,\quad D(\xi_i)=\frac{p_i(1-p_i)}{n}$$

- 为标准化 $\left(\frac{\nu_i}{n}-p_i\right)^2$, 应该用

$$\frac{\left(\frac{\nu_i}{n}-p_i\right)^2}{\frac{p_i(1-p_i)}{n}}$$

- 但其中的 p_i 一般都比较小 (尤其是样本量很大时)，$1-p_i\approx1$，就变成了

$$\frac{\left(\frac{\nu_i}{n}-p_i\right)^2}{\frac{p_i}{n}}=\left(\frac{\nu_i}{n}-p_i\right)^2\frac{n}{p_i}$$

- 拟合优度卡方检验的否定域为

$$W=\{V>\lambda\}$$

- 要找临界值 λ 使 W 的检验水平为 α。
- 需要 V 在 H_0 下的分布。在 H_0 成立的条件下，样本量较大时 V 近似服从 $\chi^2(m)$。
- 查 $\chi^2(m)$ 的右侧 α 水平临界值就可以得到 λ。
- 这种检验叫做**拟合优度（卡方）检验**，或分布的**卡方检验**。
- 当零假设下的分布函数包含从样本中估计的未知参数时，V 的自由度改为 m 减去估计的未知参数的个数。
- **例 6.1** 某车间生产滚珠，随机抽取了 50 个产品，直径数据（mm）：

  15.0 15.8 15.2 15.1 15.9 14.7 14.8 15.5 15.6
  15.3 15.1 15.3 15.0 15.6 15.7 14.8 14.5 14.2
  14.9 14.9 15.2 15.0 15.3 15.6 15.1 14.9 14.2
  14.6 15.8 15.2 15.9 15.2 15.0 14.9 14.8 14.5
  15.1 15.5 15.5 15.1 15.1 15.0 15.3 14.7 14.5
  15.5 15.0 14.7 14.6 14.2

- 直方图：演示。
- 计算得 $\bar{x} = 15.1$, $S^2 = (0.4325)^2$。
- 问：滚珠直径是否服从 $\mathrm{N}(15.1, (0.4325)^2)$?

- **解** 分组与直方图法类似。找到样本值中最小与最大值，取比最小数略小的 a，比最大数略大的 b，将区间 $[a, b]$ 做 $m+1$ 等分，得分点

$$a < t_1 < t_2 < \cdots < t_m < b$$

- m 的个数选取参考直方图方法。
- 这 50 个数据最小 14.2，最大 15.9，取 $a = 14.05$, $b = 16.15$, $b - a = 2.1 = 0.3 * 7$, $m = 6$，分成 7 段。
- 在 $F(x)$ 为 $\mathrm{N}(15.1, (0.4325)^2)$ 分布函数时求得各个 p_i。

- 为求 $F(x)$，用

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
$$\Phi(-x) = 1 - \Phi(x) \quad (x > 0)$$

(μ, σ^2 是正态分布的期望和方差)

- 计算得

i	p_i	np_i	ν_i	$(np_i - \nu_i)^2$	$\frac{(np_i-\nu_i)^2}{np_i}$
1	0.0414	2.070	3	0.8649	0.4178
2	0.1077	5.385	5	0.1482	0.0275
3	0.2154	10.770	10	0.5925	0.0551
4	0.2710	13.550	16	6.0025	0.4430
5	0.2154	10.770	8	7.6729	0.7124
6	0.1077	5.385	6	0.3782	0.0702
7	0.0414	2.070	2	0.0049	0.0024

- 检验统计量

$$V = \sum_{i=1}^{7} \frac{(\nu_i - np_i)^2}{np_i} = 1.7284$$

- 自由度用 $7 - 1 - 2$（额外扣除两个估计的未知参数的自由度）。取 $\alpha = 0.05$，$\lambda = 9.49$。

- $V < \lambda$，H_0 相容，在 0.05 水平下不拒绝总体服从正态分布的假设。

- 注意：如果最后的结论是不拒绝 H_0，可能会有较大的第二类错误概率。

- 有可能会以威布尔分布作为零假设，检验不拒绝；再以对数正态作为零假设，检验还是不拒绝。

- 但是，只要不拒绝零假设，就可以说明数据与该分布差距不大。

离散分布的卡方检验

- 连续型分布用卡方检验需要分组，离散型分布不需要分组。
- 概率特别小的组可以合并。
- 设 X 的分布是

$$P(X = a_i) = p_i, \quad (i = 1, 2, \ldots, m+1)$$

  $x_1, x_2, \ldots, x_n$ 是样本值。

- 取统计量

$$V = \sum_{i=1}^{m+1} \frac{(\nu_i - np_i)^2}{np_i} \tag{6.1}$$

- 其中 ν_i 表示 n 个样品中 a_i 出现的频数。
- 在 H_0 下 V 近似服从 $\chi^2(m)$; 如果分布中含有从样本值估计的未知参数，则自由度要扣除未知参数个数。
- **例 6.2（例 1.3 续）** 某工厂近 5 年来发生了 63 次事故，这些事故在工作日的分布如下：

星期	一	二	三	四	五	六
次数	9	10	11	8	13	12

- 问：事故发生是否与星期几有关？
- **解** 用 X 表示这样的随机变量：若事故发生在星期 i，则 $X = i$。
- X 的可能取值集合为 $\{1, 2, \ldots, 6\}$（星期日是该厂厂休日）。
- 检验

$$H_0 : P(X = i) \equiv \frac{1}{6} \quad (i = 1, 2, \ldots, 6)$$

- 使用统计量 (6.1)。$m = 5$, H_0 成立时 $p_i = 1/6, i = 1, 2, \ldots, 6$。
- 计算得 $V = 1.67$。
- 查 5 个自由度卡方分布右侧 $\alpha = 0.05$ 临界值得 $\lambda = 11.07$。
- $V < \lambda$，H_0 相容，不能认为出事故与星期几有关。

第七章　回归分析方法

7.1　一元线性回归

回归分析方法

- 回归分析方法是数理统计的重要工具，是处理多个变量之间**相关关系**的一种数学方法。
- **函数关系**: 确定性关系。如自由落体运动

$$s = \frac{1}{2}gt^2 \quad (0 \le t \le T)$$

- **相关关系**是给定了 x 的值后并不能确定 y 的值，但 y 的值与 x 的值有关。
- 即使是确定性关系的变量，其测量值因为含有误差所以也有不确定性。
- 回归分析可以建立变量间的关系的数学表达式（经验公式），并可判断这样的公式的有效性，以及如何利用所得到的经验公式去达到预测、控制等目的。

7.1.1　经验公式与最小二乘法

一元线性回归

- 在一元线性回归分析中，考察随机变量 Y 与一个普通变量 x(非随机)之间的联系。

- 数据成对观测：

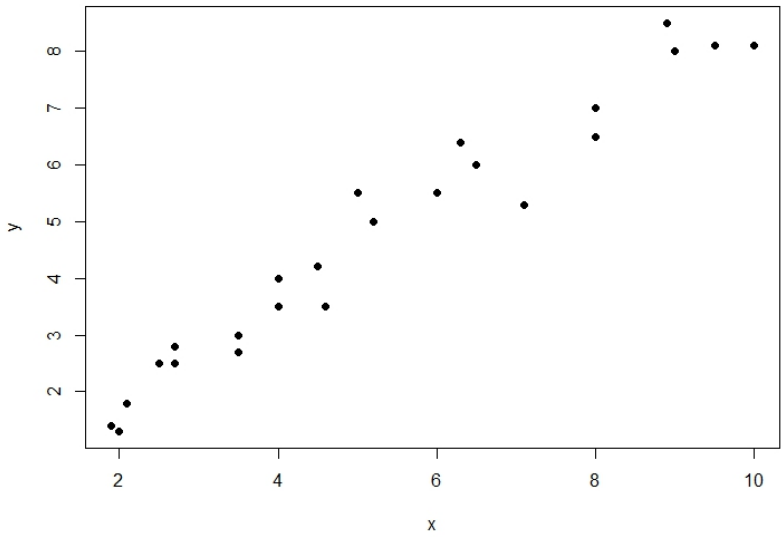
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- **例 1.1** 某种合成纤维的强度与其拉伸倍数有关。
- 有 24 个纤维样品的强度与相应的拉伸倍数的实测记录。
- 设拉伸倍数为 x，强度为 Y，希望根据观测数据找出 x 和 Y 的关系式。
- 数据（部分）如

$$(1.9, 1.4), (2.0, 1.3), (2.1, 1.8), \ldots, (9.5, 8.1), (10.0, 8.1)$$

- 一种直观的考察方式是**散点图**（演示）。

纤维强度对拉伸倍数的散点图



- 散点图中每个点以 x 为横坐标，以 y 为纵坐标。
- 从散点图看散点围绕在一条直线周围，有

$$\hat{y} = a + bx \tag{1.1}$$

  其中 $\hat{y}$ 表示建立 Y 与 x 的关系后用 x 对 Y 做的预测。
- 于是借助于散点图确定了经验公式的形式。只需要确定 (1.1) 中的 a 和 b。
- b 叫做**回归系数**，关系式 $\hat{y} = a + bx$ 叫做**回归方程**。
- **线性**: x 每增加 1，y 的变化量是恒定的。
- **非线性**: x 每增加 1，y 的变化量不是恒定的。

求解回归直线

- 要找一条直线与散点图中所有点尽可能最接近。
- 直接作图过于粗略，且无法推广到多个自变量的情形。
- 定义距离：用

$$[y_i - (a + bx_i)]^2$$

  衡量点 (x_i, y_i) 到直线 $\hat{y} = a + bx$ 的距离。
- 这是两个纵坐标的距离平方，不是点到直线的垂直距离。
- **理由**: 需要衡量的是对 Y 的预测精度。

最小二乘原则

- 平方和

$$Q(a,b)=\sum_{i=1}^{n}[y_i-(a+bx_i)]^2 \tag{1.2}$$

  衡量直线 $\hat{y}=a+bx$ 与所有散点的距离远近。

- 求回归直线问题化为：找两个数 $\hat{a},\hat{b}$，使得二元函数 $Q(a,b)$ 在 $a=\hat{a},b=\hat{b}$ 处达到最小。

- 这种方法叫做**最小二乘法**。

最小二乘求解的微分法

- 为了求 $Q(a,b)$ 的最小值点，求解其一阶偏导数都等于零的方程：

$$\frac{\partial Q}{\partial a}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]=0 \tag{1.3}$$

$$\frac{\partial Q}{\partial b}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]\cdot x_i=0 \tag{1.4}$$

- 由 (1.3) 解得

$$\begin{aligned} na&=\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_i \\ a&=\bar{y}-b\bar{x} \end{aligned} \tag{1.5}$$

- 由 (1.4) 得

$$\sum_{i=1}^{n}x_i[y_i-a-bx_i]=0$$

- 把 (1.5) 代入上式可得

$$\begin{aligned}
&\sum_{i=1}^{n} x_i[y_i - \bar{y} - b(x_i - \bar{x})] = 0 \\
&b\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) \\
&b\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned} \tag{1.6}$$

- 代入 (1.5) 可得 $\hat{a} = \bar{y} - \hat{b}\bar{x}$。

- 当 $x_1, x_2, \ldots, x_n$ 不全相等时有解。
- 可以证明这样用微分法求得的 $(\hat{a}, \hat{b})$ 是 $Q(a, b)$ 的最小值点。
- 事实上，二阶导数矩阵，即海色阵为

$$H = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2\right) \end{pmatrix}$$

  易见其主子式都大于零，H 正定，$Q(a, b)$ 的唯一的一阶偏导数等于零的点一定是全局最小值点。

最小二乘解的配方法

- 拆分平方与交叉项：

$$\begin{aligned}
Q(a, b) =& \sum_{i=1}^{n} \{(y_i - \bar{y}) + [\bar{y} - (a + b\bar{x})] - b(x_i - \bar{x})\}^2 \\
=& \sum_{i=1}^{n} (y_i - \bar{y})^2 + n[\bar{y} - (a + b\bar{x})]^2 \\
&+ b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2b\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\end{aligned}$$

- 其中 $[\bar{y}-(a+b\bar{x})]$ 是常数，所以它与 $x_i-\bar{x}$ 和 $y_i-\bar{x}$ 的交叉项为零。
- 记

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$
$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- 则

$$\begin{aligned}Q(a,b)=&l_{yy}+n[\bar{y}-(a+b\bar{x})]^2+l_{xx}b^2-2l_{xy}b\\=&l_{yy}+n[\bar{y}-(a+b\bar{x})]^2+l_{xx}\left(b-\frac{l_{xy}}{l_{xx}}\right)^2-\frac{l_{xy}^2}{l_{xx}}\\\geq&l_{yy}-\frac{l_{xy}^2}{l_{xx}}\end{aligned}$$

- 等于号成立当且仅当

$$\begin{aligned}b=&\frac{l_{xy}}{l_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\a=&\bar{y}-b\bar{x}\end{aligned}$$

- $Q(a,b)$ 的最小值为

$$\begin{aligned}Q(\hat{a},\hat{b})=&l_{yy}-\frac{l_{xy}^2}{l_{xx}}\\=&l_{yy}-\hat{b}l_{xy}\\=&\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{b}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\end{aligned}$$

- 得到了 $\hat{a},\hat{b}$ 就确定了回归的经验公式，确定了回归直线。
- 演示（Flash）。
- 易见点 $(\bar{x},\bar{y})$ 落在回归直线上：

$$\bar{y}=\hat{a}+\hat{b}\bar{x}$$

- 回归一般使用统计软件计算。比如在 R 中

```
lm1 <- lm(y ~ x)
summary(lm1)
plot(lm1)
```

  可以计算并显示回归结果、画回归诊断图形。

- 对于例 1.1 的 24 个点，计算得 $\hat{b} = 0.859, \hat{a} = 0.15$，纤维强度 (Y) 与拉伸倍数 (x) 的经验公式为

$$\hat{y} = 0.15 + 0.859x$$

- 经验公式也叫**回归方程**, 相应的直线叫做**回归直线**。

- 回归系数 b 的含义：拉伸倍数 (x) 每增加一个单位，强度 (Y) 平均增加 0.859 个单位。

非线性关系线性化

- 某些非线性关系可以通过变换转化为线性关系。

- **例 1.2** 彩色显影中，染料光学密度 Y 与析出银的光学密度 x 有如下类型的关系

$$Y \approx Ae^{-B/x}, \quad B > 0$$

- 这不是线性关系。两边取对数得

$$\ln Y \approx \ln A - B\frac{1}{x}$$

- 令

$$Y^* = \ln Y \qquad\qquad x^* = \frac{1}{x}$$

- 则

$$Y^* \approx \ln A - Bx^*$$

为线性关系。

- 从 n 组数据 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 得到变换的数据 $(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)$。
- 对变换后的数据建立线性回归方程

$$\hat{y}^* = \hat{a} + \hat{b}x^*$$

- 反变换得

$$\hat{A} = e^{\hat{a}} \qquad \hat{B} = -b$$

- 则有

$$\hat{Y} = \hat{A}e^{-\hat{B}/x}$$

- **例 1.3** 炼钢钢包随使用次数增加而容积增大。
- 测量了 13 组这样的数据（部分）：

$$(2, 106.42), (3, 108.20), (4, 109.58), \ldots, (19, 111.20)$$

- 画出了散点图（演示）。用双曲线

$$\frac{1}{y} \approx a + b\frac{1}{x}$$

- 令 $x^* = 1/x, y^* = 1/y$，化为线性模型

$$y^* \approx a + bx^*$$

- 解得 $\hat{=}0.008967$, $\hat{b} = 0.0008292$, 经验公式为

$$\frac{1}{\hat{y}} = 0.008967 + 0.0008292\frac{1}{x}$$

7.1.2 平方和分解公式与线性相关关系

线性相关性

- 只要数据中 $x_1, x_2, \ldots, x_n$ 不全相等，最小二乘法存在唯一解，总可以得到经验公式

$$\hat{y} = \hat{a} + \hat{b}x$$

- 所以，经验公式并不都能反映实际情况。
- 需要判别 x 与 Y 之间是否真的具有线性相关关系：Y 是否随着 x 增大而线性地增大（或者线性地减小）。

平方和分解公式

- 对于任意 n 组数据 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，只要 $x_1, x_2, \ldots, x_n$ 不全相等，就有

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{1.7}$$

- 其中 $\bar{y}$ 是 $y_1, y_2, \ldots, y_n$ 的平均值，

$$\hat{y}_i = \hat{a} + \hat{b}x_i \quad (i = 1, 2, \ldots, n)$$

- **证**

$$\begin{aligned}&\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2\\ =&\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\end{aligned}$$

- 注意

$$\begin{aligned}\hat{y}_i =&\hat{a} + \hat{b}x_i = \bar{y} - \hat{b}\bar{x} + \hat{b}x_i\\ =&\bar{y} + \hat{b}(x_i - \bar{x})\end{aligned}$$

- 所以交叉项

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})\\
=&\sum_{i=1}^{n}[y_i-\bar{y}-\hat{b}(x_i-\bar{x})]\hat{b}(x_i-\bar{x})\\
=&\hat{b}\left\{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-\hat{b}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}\\
=&0
\end{aligned}$$

- 于是

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i-\bar{y})^2\\
=&\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2
\end{aligned}$$

- 即平方和分解公式 (1.7) 成立。

平方和分解公式的解释

- (1.7) 式坐标的 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 是因变量的离差（偏差）平方和（Corrected Sum of Squares），描述了因变量的分散程度，是我们要用模型解释的目标。记为 l_{yy}。
- 考虑分解的第一项 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$。这是用模型得到的因变量**拟合值** $\hat{y}_i$ 与实际因变量值 y_i 之间的差距的一个度量，是最小二乘法最后得到的最小化的目标函数值。这个平方和越小，说明模型与实际数据越相符。
- 记

$$\hat{\varepsilon}_i=y_i-\hat{y}_i\quad(i=1,2,\ldots,n)$$

称为**残差** (residual)。记

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

称 Q 为**残差平方和** (Error Sum of Squares)。

- 分解的第二项：

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- 易见

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i =& \frac{1}{n}\sum_{i=1}^{n} (\hat{a} + \hat{b}x_i) \\ =& \hat{a} + \hat{b}\bar{x} = \bar{y} \end{aligned}$$

- 所以 U 是拟合值 $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ 的离差平方和。
- U 受什么因素影响呢？

-

$$\begin{aligned} U =& \sum_{i=1}^{n} (\hat{a} + \hat{b}x_i - \bar{y})^2 \\ =& \sum_{i=1}^{n} (\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2 \\ =& \hat{b}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ =& \hat{b}^2 l_{xx} \end{aligned}$$

- l_{xx} 是自变量的离差平方和。所以 U 代表了自变量对因变量的变化的解释，在分解中 U 越大，残差平方和 Q 越小，模型对数据拟合越好。
- U 叫做回归平方和。

- 所以，平方和分解公式把因变量的变差（离差平方和）分解为两部分：

$$l_{yy} = Q + U$$

- U 来源于自变量 x 的分散程度，通过线性关系影响了因变量 Y 造成 Y 的变差，是模型可以解释的部分；
- Q 是模型拟合的误差的度量，是自变量和模型不能解释的部分。
- 分解中，U 越大，Q 越小，模型越准确描述自变量和因变量之间的线性相关关系。
- 反之，如果 Q 很大，则自变量和因变量之间没有线性相关关系。
- 取统计量

$$F = \frac{U}{Q/(n-2)} \tag{1.9}$$

则当 F 相当大时，表明 x 对 Y 的线性影响越强，两者有线性相关性。否则没有线性相关性。

7.1.3 数学模型与相关性检验

- F 值多大才认为线性相关性成立？这是一个假设检验问题。
- 需要对模型进行进一步细化。
- 设数据满足如下结构

$$\begin{aligned} Y_1 =& a + bx_1 + \varepsilon_1 \\ Y_2 =& a + bx_2 + \varepsilon_2 \\ & \cdots\cdots\cdots \\ Y_n =& a + bx_n + \varepsilon_n \end{aligned} \tag{1.10}$$

- 其中 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 是独立同分布随机变量列，共同分布为 $\mathrm{N}(0, \sigma^2)$(σ^2 未知)。
- 这样的模型是单总体模型、两总体模型的进一步推广。它等价于

$$Y_i \sim \mathrm{N}(a + bx_i, \sigma^2), \quad i = 1, 2, \ldots, n$$

且相互独立。这给出了 n 维随机向量 $(Y_1, Y_2, \ldots, Y_n)$ 的联合分布。

- 在承认了 (1.10) 模型基础上，x 与 Y 之间有无线性相关关系的问题变成假设

$$H_0 : b = 0$$

- 当 H_0 成立时，模型 (1.10) 退化为

$$Y_i = a + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

模型中不含自变量 x，所以这时 x 与 Y 没有线性相关关系。

- 当 H_0 不成立时，Y 与 x 有线性相关关系。

相关性检验

- 当 H_0 成立时，统计量 F 服从 $\mathrm{F}(1, n-2)$ 分布。
- 对检验水平 α，取 $\mathrm{F}(1, n-2)$ 分布的右侧 α 临界值 λ，H_0 的否定域为

$$W = \{F > \lambda\}$$

- 从样本中计算统计量 F 的值，当 $F > \lambda$ 时拒绝 H_0，认为 x 与 Y 之间存在线性相关性（有**显著的线性相关性**）；
- 当 $F \leq \lambda$ 时 H_0 相容，认为 x 与 Y 之间没有显著的线性相关性。

随机误差方差估计

- 可以证明

$$\frac{1}{\sigma^2}Q \sim \chi^2(n-2)$$

- 从而

$$E\left(\frac{1}{\sigma^2}Q\right) = n-2$$
$$E\left(\frac{1}{n-2}Q\right) = \sigma^2$$

- 所以

$$\hat{s}^2 \triangleq \frac{1}{n-2}Q = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  是 σ^2 的无偏估计。

（样本）相关系数

- 设 U, V 是两个随机变量，其相关系数定义为

$$\rho = \frac{\mathrm{Cov}(U,V)}{\sqrt{D(U)D(V)}}$$

- 若 (U,V) 有样本 $(u_1,v_1),(u_2,v_2),\ldots,(u_n,v_n)$，则可计算样本相关系数

$$R = \frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i-\bar{u})^2\sum_{i=1}^{n}(v_i-\bar{v})^2}}$$

- 在线性回归中虽然 x 是非随机的变量，但也可以定义样本相关系数为

$$R = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- 当 $|R|$ 相当大时拒绝 $H_0: b=0$。

复相关系数平方

- 易见

$$\begin{aligned} R^2 =& \frac{l_{xy}^2}{l_{xx}l_{yy}} = \frac{l_{xy}^2}{l_{xx}^2} \cdot \frac{l_{xx}}{l_{yy}} \\ =& \frac{\hat{b}^2 l_{xx}}{l_{yy}} = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}} \end{aligned}$$

- 一般地，定义

$$R^2 = \frac{U}{l_{yy}} = 1 - \frac{Q}{l_{yy}},$$

  称 R^2 为复相关系数平方，这个定义在多元回归时照样适用。

- R^2 取值于 $[0,1]$，代表了回归平方和在总平方和中的比例，R^2 越接近于 1，回归模型对数据拟合得越好。
- 当 $R^2 = 1$ 时，$Q = 0$，所有的 n 个数据点

$$(x_i, y_i), \quad i = 1, 2, \ldots, n$$

  都落在直线

$$\hat{y} = \hat{a} + \hat{b}x$$

  上。

- $R^2 = 1$ 的情况一般只出现在确定性关系中。

F 统计量与 R^2

- 检验 $H_0 : b = 0$ 用的 F 统计量与 R^2 一一对应：

$$\begin{aligned} F =& \frac{U}{Q/(n-2)} = (n-2)\frac{U}{l_{yy} - U} \\ =& (n-2)\frac{R^2}{1-R^2} = \frac{n-2}{\frac{1}{R^2} - 1} \end{aligned}$$

- 两者为一一对应的严格单调递增关系。

两个平方和的计算公式

- 按定义，

$$U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 但是，在有了 l_{yy}, l_{xx}，$l_{xy}, \hat{b}$ 后可简单计算为

$$U = \hat{b}^2 l_{xx} = \hat{b} l_{xy} \tag{1.11}$$
$$Q = l_{yy} - U \tag{1.12}$$

- **例 1.4** 炼钢基本是氧化脱碳的过程，原来碳含量越高，需要的冶炼时间越长。

- 有某平炉 34 炉的熔毕碳 (x) 与精炼时间 (y) 的记录如下（部分）：

$$(180, 200), (104, 100), \ldots, (143, 160)$$

- 散点图见演示。

- 计算过程：见 R 程序演示。

- 主要结果

$$\hat{a} = -23.20, \qquad \hat{b} = 1.270$$
$$F = 145.0 \qquad R^2 = 0.8192$$

- F(1,32) 右侧 0.01 分位数为 $\lambda = 4.15$，$F > \lambda$，可以认为 x, Y 之间存在线性相关关系，或：直线回归是显著的。

7.1.4 预报与控制

回归模型的作用

- 揭示变量之间的数量关系；
- 预报；
- 控制。

预报

- 设

$$Y = a + bx + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- 由数据 $(x_i, y_i), i = 1, 2, \ldots, n$ 得到参数最小二乘估计 $\hat{a}, \hat{b}$ 和误差方差估计 s^2。
- 对新的自变量值 x_0，设

$$Y_0 = a + bx_0 + \varepsilon_0$$

- 用

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

  预报 Y_0 的值。

- 还需要衡量预报精度。
- 若 $\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$，则

$$t \triangleq \frac{Y_0 - \hat{y}_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim \mathrm{t}(n-2)$$

- 为了求 Y_0 的**预报区间**, 设 λ 为 $\mathrm{t}(n-2)$ 分布的双侧 α 临界值，由

$$P\left(\left|\frac{Y_0-\hat{y}_0}{s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\right|\le\lambda\right)=1-\alpha \tag{1.13}$$

- 得 Y_0 的置信度 $1-\alpha$ 的预报区间为

$$\hat{y}_0\pm\lambda s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}} \tag{1.14}$$

- 演示：熔毕碳与精炼时间的预报区间，连线为曲线，但只有单点意义。

- x_0 离 $\bar{x}$ 越远，预报区间长度越长。
- 注意：回归模型的应用范围不能超出原数据的范围。
- 作为 (1.14) 的近似，当 n 较大且 x_0 离 $\bar{x}$ 不远的时候，

$$\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\approx 1$$

  所以预测区间近似为

$$[\hat{y}_0-\lambda s,\hat{y}_0+\lambda s]$$

- 当 n 较大时 λ 可以用标准正态分布的双侧 α 临界值，如 $\alpha=0.05$ 时用 $\lambda=1.96$。
- 误差标准差估计 s 越小，预报区间越短，预报越精确。

控制问题

- 控制问题是：要求控制 Y 在区间 $[A,B]$ 内，如何选取 x 的值？
- 办法是要求 (1.14) 得到的上下限都在 $[A,B]$ 内，反解符合要求 x_0 的区间。

回归诊断和残差分析

- 即使线性相关性检验否定了 $H_0 : b = 0$，也并不说明模型就是合适的。
- 常见问题包括：
- 缺少重要自变量；
- 有非线性相关；
- 误差项方差非恒定；
- 误差项存在序列相关；
- 自变量严重共线（多元回归中）；
- 数据有异常值或强影响点。
- 可以用残差散点图等进行回归诊断。

残差分析

- 残差

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- 令

$$\begin{aligned} h_i =& \frac{1}{n} + \frac{(x_i - \bar{x})^2}{l_{xx}} \\ s =& \sqrt{\frac{Q}{n-2}} \\ r_i =& \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} \end{aligned}$$

- 则在 (1.10) 模型成立时 $r_1, r_2, \ldots, r_n$ 近似相互独立，且近似服从标准正态分布。

- 有

$$P(|r_i| > 2) \approx 0.05$$

- 当 n 比较大时，$r_i, i = 1, 2, \ldots, n$ 应该只有约 $[0.05n]$ 个绝对值大于 2。
- 这可以用来检验模型关于误差项的假设是否成立，以及发现异常值点。

7.2 多元线性回归

多元线性回归

- 考虑多个自变量与因变量的关系。
- 要解决的问题与一元回归相同。
- 解决方法类似。

7.2.1 模型

模型

- 设因变量 Y 与自变量 $x_1, x_2, \ldots, x_k$ 有关系式

$$Y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

- 其中自变量 $x_1, x_2, \ldots, x_k$ 是非随机的变量，ε 是随机项。
- 有 n 组数据

$$\begin{array}{l} (y_1; x_{11}, x_{12}, \ldots, x_{1k}) \\ (y_2; x_{21}, x_{22}, \ldots, x_{2k}) \\ \cdots\cdots\cdots \\ (y_n; x_{n1}, x_{n2}, \ldots, x_{nk}) \end{array} \tag{2.1}$$

- 假定数据满足

$$\begin{cases} Y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_k x_{1k} + \varepsilon_1 \\ Y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_k x_{2k} + \varepsilon_2 \\ \quad \cdots\cdots\cdots \\ Y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_k x_{nk} + \varepsilon_n \end{cases} \tag{2.2}$$

  这里 Y_t 写成大写是为了强调在模型中它是随机变量。

- 其中 $b_0, b_1, \ldots, b_k$ 是待估参数，$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 相互独立且服从相同的 $\mathrm{N}(0, \sigma^2)$ 分布。

说明

- “多元”是指自变量有多个，但因变量还是只有一个。另外，自变量是非随机的普通变量，因变量是随机变量。

- (2.1) 中的各个 y_t 是数据值，(2.2) 中大写的 Y_t 看作随机变量。把 y_t 看作 Y_t 的观测值。

- (2.2) 表示 Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系。对于某些非线性关系，可以通过变换转化为线性。比如一元多项式回归

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k$$

  只要记 $x_1 = x, x_2 = x^2, \ldots, x_k = x^k$ 就变成自变量为 $x_1, x_2, \ldots, x_k$ 的多元线性回归

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

7.2.2 最小二乘估计与正规方程

- 为估计未知参数，最小化误差平方和

$$\begin{aligned}&Q(b_0,b_1,\ldots,b_k)\\=&\sum_{t=1}^{n}[y_t-(b_0+b_1x_{t1}+b_2x_{t2}+\cdots+b_kx_{tk})]^2\end{aligned}$$

- 使 $Q(b_0,b_1,\ldots,b_k)$ 达到最小值的点 $\hat{b}_0,\hat{b}_1,\ldots,\hat{b}_k$ 称为参数 $b_0,b_1,\ldots,b_k$ 的**最小二乘估计**。

- 记

$$\begin{aligned}\bar{y}=&\frac{1}{n}\sum_{t=1}^{n}y_t,\\\bar{x}_i=&\frac{1}{n}\sum_{t=1}^{n}x_{ti},\ i=1,2,\ldots,k\\l_{ij}=&l_{ji}=\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j),\ i,j=1,2,\ldots,k\\l_{iy}=&\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(y_t-\bar{y}),\ i=1,2,\ldots,k\end{aligned}$$

- 则 $\hat{b}_1,\hat{b}_2,\ldots,\hat{b}_k$ 为如下 n 阶线性方程组的解：

$$\begin{pmatrix}l_{11}&l_{12}&\cdots&l_{1k}\\l_{21}&l_{22}&\cdots&l_{2k}\\\cdots&\cdots&\cdots&\\l_{k1}&l_{k2}&\cdots&l_{kk}\end{pmatrix}\begin{pmatrix}b_1\\b_2\\\vdots\\b_k\end{pmatrix}=\begin{pmatrix}l_{1y}\\l_{2y}\\\vdots\\l_{ky}\end{pmatrix}$$

- 而

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2-\cdots-b_k\bar{x}_k$$

- 这 $k+1$ 个方程组成的方程组叫做**正规方程**。

- 可以证明，最小二乘估计一定存在，而且 $b_0, b_1, \ldots, b_k$ 是最小二乘估计的充分必要条件为满足正规方程。

- 当如下矩阵

$$\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

为满秩矩阵（要求 $n > k+1$，满秩指列满秩）时正规方程的解唯一，所以最小二乘估计唯一。

7.2.3 平方和分解公式与 σ^2 的无偏估计

平方和分解公式

- 平方和分解公式：

$$\begin{aligned} l_{yy} =& Q + U \\ l_{yy} =& \sum_{t=1}^{n} (y_t - \bar{y})^2 \\ Q =& \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \quad \text{(残差平方和)} \\ U =& \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 \quad \text{(回归平方和)} \\ =& \hat{b}_1 l_{1y} + \hat{b}_2 l_{2y} + \cdots + \hat{b}_k l_{ky} \\ \hat{y}_t =& \hat{b}_0 + \hat{b}_1 x_{t1} + \hat{b}_2 x_{t2} + \cdots + \hat{b}_k x_{tk},\ t = 1, 2, \ldots, n \end{aligned}$$

σ^2 的无偏估计

- $Q/\sigma^2 \sim \chi^2(n-k-1)$，所以

$$E(Q/\sigma^2) = n-k-1$$
$$E\left(\frac{1}{n-k-1}Q\right) = \sigma^2$$

- 记

$$\hat{\sigma}^2 = \frac{1}{n-k-1}Q$$

  $\hat{\sigma}^2$ 为 σ^2 的无偏估计。

- 有时记为 s^2。

7.2.4 相关性检验

相关性检验

- 最小二乘估计总存在，所以不管 Y 和 $x_1, x_2, \ldots, x_k$ 之间有没有线性相关关系总能建立回归方程。
- 必须检验线性相关关系是否成立。
- 转化为：

$$H_0 : b_1 = b_2 = \cdots = b_k = 0$$

  的检验。

- 当 H_0 成立时，模型中不出现自变量 $x_1, x_2, \ldots, x_k$，所以没有线性相关关系。
- 当 H_0 不成立时，Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系。

- 检验统计量为

$$F = \frac{U/k}{Q/(n-k-1)}$$

- 在 H_0 下 $F \sim \mathrm{F}(k, n-k-1)$。
- 给定检验水平 α 后查 $\mathrm{F}(k, n-k-1)$ 的临界值表得 λ。
- 计算 F 的值后，当且仅当 $F > \lambda$ 时拒绝 H_0，认为 Y 与 $x_1, x_2, \ldots, x_k$ 有线性相关关系，也称**回归方程显著**。
- 若 F 的值为 v，可以计算检验的 p 值

$$p = P(F > v)$$

其中 F 为服从 $\mathrm{F}(k, n-k-1)$ 分布的随机变量，当且仅当 p 值小于 α 时拒绝 H_0。

7.2.5 偏回归平方和与因素主次的判别

因素主次的判别

- 多元回归时，即使能否定 $H_0 : b_1 = b_2 = \cdots = b_k = 0$，仍然有可能一部分自变量与 Y 没有线性相关关系。
- 或者，虽然某自变量 x_i 与 Y 有线性相关关系，但是其它自变量能够代表它，所以 x_i 也不需要出现在模型中。
- 另外，即使部分自变量都是在模型中有意义的，也会有因素主次之分。

偏回归平方和

- 在平方和分解中，回归平方和 U 代表了所有 k 个自变量的作用。
- 为了研究某个自变量的贡献，不妨考虑 x_k 的作用。
- 从原来的数据中建立 Y 对 $x_1, x_2, \ldots, x_{k-1}$ 的回归，得到一个回归平方和 $U_{(k)}$，一定有 $U_{(k)} \leq U$。
- 称

$$u_k = U - U_{(k)}$$

为 $x_1, x_2, \ldots, x_k$ 中 x_k 的**偏回归平方和**。

- 类似可以定义每个自变量的偏回归平方和 $u_i, i = 1, 2, \ldots, k$。
- 注意偏回归平方和都是在一个变量集合的前提下讨论的。

偏回归平方和的计算

- u_i 的计算不需要真的重新拟合回归模型，而是有公式

$$u_i = \frac{\hat{b}_i^2}{c_{ii}}$$

其中 c_{ii} 为

$$L = (l_{ij})_{k \times k}$$

的逆矩阵的第 i 个主对角线元素。

- 为了检验 $H_0 : b_i = 0$，可以用

$$F_i = \frac{u_i}{s^2}$$

在 H_0 下 $F_i \sim \mathrm{F}(1, n-k-1)$。

单个自变量的显著性

- 设 F_i 的值为 v，则

$$p = P(F > v)$$

(其中 F 为 $\mathrm{F}(1, n-k-1)$ 分布随机变量) 是检验的 p 值。

- 当 p 值小于 0.05 时称变量 x_i 是显著的。
- 当 p 值小于 0.01 时称变量 x_i 是高度显著的。
- 当 F_i 的值很小时，应该从回归方程中剔除自变量 x_i。
- 注意：当 x_i 不显著时，可能有两种原因：
  - x_i 对 Y 没有线性的影响；

  - x_i 对 Y 有线性的影响，但存在其它的自变量能够代替 x_i 对 Y 施加相同的影响。

- 即使回归方程显著，所有自变量显著，也不能断言模型就是符合实际的，还可能有各种模型设定错误或缺陷（类似一元回归时所述）。

多元回归的计算

- 各统计软件都可以很容易地计算多元回归。
- 比如，在 R 软件中输入了自变量 `x1`，`x2` 和因变量 `y` 后，只要用

```
lm1 <- lm(y ~ x1 + x2)
summary(lm1)
plot(lm1)
```

  就可以得到回归结果并绘制回归诊断图形。

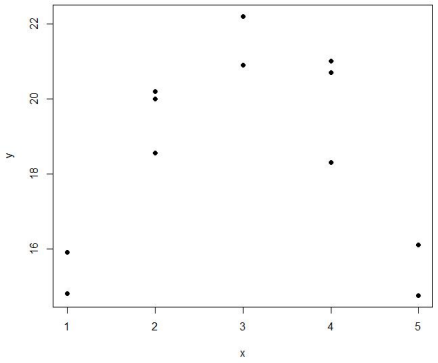
7.2.6 多元回归的例子

- **例 2.1（广告策略）** 研究广告费用 x 与纯利润 y 之间的关系，以确定最佳的广告策略。
- 数据：

x	1	1	2	2	2	3
y	14.80	15.90	20.20	20.00	18.55	22.20
x	3	4	4	4	5	5
y	20.90	21.00	18.30	20.70	16.10	14.75

- 试找出 y 与 x 的相关关系是并确定最优的广告策略。

- 画出散点图：



- 可以看出 y 与 x 不是线性关系。

- 最简单的非线性关系是一元二次多项式，设

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

其中 ε 是随机项。

- 若令 $x_1 = x, x_2 = x^2$，则方程化为

$$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$$

- 但是，在多项式回归时为了避免共线性问题，令

$$x_1 = x \qquad\qquad x_2 = (x-3)^2$$

- 用统计软件计算得

$$\begin{aligned}
\hat{y} =& 21.26 + 0.07045x - 1.504(x-3)^2 \\
=& 7.627 - 9.094x - 1.504x^2 \\
s =& 0.9788 \\
F =& 35.08, \quad \text{p 值} < 0.0001
\end{aligned}$$

- 为了求 $\hat{y}$ 的最大值（纯利润最大值），求导得

$$x = \frac{-9.093}{2 \times (-1.504)} = 3.02$$

时达到最大值。

- **例 2.2（生理节律模型）** 为了测定一个人在 24 小时内的生理节律（例如血压如何随时间变化），一些学者提出了如下模型：

$$f(t) = M + A\cos(\omega t + \phi)$$

- 其中 M 是基准值，A 是振幅，ϕ 是想为，ω 是角频率（周期 $T = 2\pi/\omega$）。
- 问：设有观测值

$$y_j = f(t_j) + \varepsilon_j,\ j = 1, 2, \ldots, n$$

这里 t_j 是第 j 个观测时刻，$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 是相互独立的随机项，$\varepsilon_j \sim \mathrm{N}(0, \sigma^2)(\sigma^2$ 未知)。

- 如何估计 $M, A, \phi(0 \leq \phi < 2\pi)$?

- **解** 模型是非线性的，设法转换为线性。
- 易见

$$y_j = M + A\cos\phi \cdot \cos(\omega t_j) - A\sin\phi \cdot \sin(\omega t_j) + \varepsilon_j$$

- 记

$$\begin{aligned} x_j &= \cos(\omega t_j), & z_j &= \sin(\omega t_j) \\ \beta &= A\cos\phi, & \gamma &= -A\sin\phi \end{aligned}$$

- 则

$$y_j = M + \beta x_j + \gamma z_j + \varepsilon_j,\ (j = 1, 2, \ldots, n)$$

化为线性模型。

- 计算正规方程中各项：

$$l_{11}=\sum_{j=1}^{n}(x_j-\bar{x})^2, \qquad l_{22}=\sum_{j=1}^{n}(z_j-\bar{z})^2$$
$$l_{12}=\sum_{j=1}^{n}(x_j-\bar{x})(z_j-\bar{z})$$
$$l_{1y}=\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y}), \qquad l_{2y}=\sum_{j=1}^{n}(z_j-\bar{z})(y_j-\bar{y})$$

- 解正规方程得

$$\hat{\beta}=\frac{l_{22}l_{1y}-l_{12}l_{2y}}{l_{11}l_{22}-l_{12}^2}, \qquad \hat{\gamma}=\frac{-l_{12}l_{1y}+l_{11}l_{2y}}{l_{11}l_{22}-l_{12}^2}$$
$$M=\bar{y}-\hat{\beta}\bar{x}-\hat{\gamma}\bar{z}$$

- 反推得到原始模型参数估计

$$\hat{A}=\sqrt{\hat{\beta}^2+\hat{\gamma}^2} \qquad \hat{\phi}=\mathrm{Arg}(\hat{\beta}-i\hat{\gamma})$$

(这里 i 表示虚数单位，$\hat{\phi}$ 是平面直角坐标系中坐标为 $(\hat{\beta},\hat{\gamma})$ 的点的辐角)

- 检验 y 与 t 是否有指定的非线性关系，可检验 $H_0:A=0$，等同于检验

$$H_0:\beta=\gamma=0$$

- 仍使用统计量

$$F=\frac{U/2}{Q/(n-3)}$$

取 $\mathrm{F}(2,n-3)$ 的右侧 α 水平临界值 λ，当且仅当 $F>\lambda$ 时拒绝 H_0，认为回归方程显著。

- 实际中 t_j 一般是等间隔的，

$$t_j = \frac{j-1}{n},\ j = 1, 2, \ldots, n$$

且 $\omega = 2\pi$(周期为 1)，常用 $n = 24$ 或 $n = 12$。

- 这时公式可以化简：

$$\begin{aligned}
\sum_{j=1}^{n} x_j &= \sum_{j=1}^{n} \cos(\omega t_j) = 0 \\
\sum_{j=1}^{n} z_j &= \sum_{j=1}^{n} \sin(\omega t_j) = 0 \\
\sum_{j=1}^{n} x_j z_j &= \sum_{j=1}^{n} \cos(\omega t_j)\sin(\omega t_j) = 0 \\
\sum_{j=1}^{n} x_j^2 &= \sum_{k=0}^{n-1} \frac{1+\cos(2k\theta)}{2} = \frac{n}{2} \quad \left(\theta = \frac{2\pi}{n}\right) \\
\sum_{j=1}^{n} z_j^2 &= \frac{n}{2}
\end{aligned}$$

- 于是得

$$\begin{aligned}
\hat{M} &= \bar{y} \\
\hat{\beta} &= \frac{1}{n}\sum_{j=1}^{n} x_j y_j \\
\hat{\gamma} &= \frac{1}{n}\sum_{j=1}^{n} z_j y_j \\
F &= \frac{n\hat{A}^2/2}{Q/(n-3)}
\end{aligned}$$

7.3 逻辑斯蒂 (Logistic) 回归

二值因变量的问题

- 经典线性回归分析中因变量和自变量都是连续取值的。
- 实际工作中经常需要处理因变量为二分类值的情况。
- 比如，x 表示一个家庭年收入，$Y=1$ 表示该家庭在某段时间购买某种耐用消费品（如汽车），$Y=0$ 表示不购买。
- 研究 $P(Y=1)$ 与 x 的关系。
- 更一般地，若随机变量 Y 只取值 0 或 1，有若干个变量 $x_1, x_2, \ldots, x_k$ 影响 Y 的取值，关心 $p=P(Y=1)$ 如何依赖于 $x_1, x_2, \ldots, x_k$。

优比和 logit 函数

- 对 $0<p<1$，有 $\frac{p}{1-p} \in (0,\infty)$ 为 p 的严格单调递增函数，$\frac{p}{1-p}$ 叫做发生比或优比 (odds ratio)。
- 定义函数

$$\text{logit}(p) = \ln \frac{p}{1-p},\ 0<p<1$$

则 $\text{logit}(p) \in (-\infty,\infty)$ 是 p 的严格单调递增函数, 叫做 logit 函数。

逻辑斯蒂回归模型

- 设因变量和自变量间的关系为

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \tag{3.1}$$

其中 $p=P(Y=1)$，$\beta_0, \beta_1, \ldots, \beta_k$ 是常数，这时称二分类变量 Y 与自变量 $x_1, x_2, \ldots, x_k$ 的关系符合逻辑斯蒂回归模型。

- 易见 (3.1) 等同于

$$P(Y=1|x_1,x_2,\ldots,x_k)=\frac{\exp(\beta_0+\sum_{i=1}^k\beta_i x_i)}{1+\exp(\beta_0+\sum_{i=1}^k\beta_i x_i)}$$

逻辑斯蒂回归参数估计

- 模型 (3.1) 中的常数 $\beta_0,\beta_1,\ldots,\beta_k$ 通常是未知的，需要从数据中估计。这个模型中没有方差项。
- 下面只考虑 $k=1$，即只有一个自变量的情形，用 x 表示 x_1。
- (3.1) 化为

$$\ln\frac{p}{1-p}=\beta_0+\beta_1 x \tag{3.2}$$

- 令 $p(x)=P(Y=1|x)$，则

$$p(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \tag{3.3}$$

- 参数估计可以用最大似然法和最小二乘法。

最大似然估计

- 设数据为 $(x_i,y_i),i=1,2,\ldots,n$。
- 则

$$P(Y=y_i|x_i)=[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}$$

- 观测值 $(x_i,y_i),i=1,2,\ldots,n$ 对应的似然函数为

$$L(\beta_0,\beta_1)=\prod_{i=1}^n[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i}$$

- 对数似然函数为

$$\ln L(\beta_0,\beta_1)=\sum_{i=1}^n y_i(\beta_0+\beta_1 x_i)-\sum_{i=1}^n\ln(1+e^{\beta_0+\beta_1 x_i})$$

- 令一阶偏导数都等于零的似然方程组

$$\sum_{i=1}^{n}\left(y_i-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)=0$$
$$\sum_{i=1}^{n}\left(y_i-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)x_i=0$$

- 若 $(\hat{\beta}_0,\hat{\beta}_1)$ 是似然方程组的根且 $x_1,x_2,\ldots,x_n$ 不全相等，则似然方程组的根是惟一的，而且 $(\hat{\beta}_0,\hat{\beta}_1)$ 是 $L(\beta_0,\beta_1)$ 的最大值点从而是模型参数的最大似然估计。

- 可以证明 $\ln L(\beta_0,\beta_1)$ 是二元严格凹函数。

- 似然方程组有时无解，如所有 y_i 都等于 1 时。

加权最小二乘估计

- 数据有特殊要求。

- 设 $x=x_i$ 时共有 n_i 次观测，n_i 较大，其中事件 $\{Y=1\}$ 发生了 γ_i 次 $(i=1,2,\ldots,m)$（$x_1,x_2,\ldots,x_m$ 两两不同）。

- 用

$$z_i=\ln\frac{\gamma_i+0.5}{n_i-\gamma_i+0.5} \tag{3.4}$$

  作为 $\ln\frac{p(x_i)}{1-p(x_i)}$ 的估计值 $(i=1,2,\ldots,m)$。

- 令

$$\nu_i=\frac{(n_i+1)(n_i+2))}{n_i(\gamma_i+1)(n_i-\gamma_i+1)}\ (i=1,2,\ldots,m) \tag{3.5}$$
$$\tilde{Q}(\beta_0,\beta_1)=\sum_{i=1}^{m}\frac{1}{\nu_i}(z_i-\beta_0-\beta_1 x_i)^2$$

- 使 $\tilde{Q}(\beta_0, \beta_1)$ 达到最小值的 $\tilde{\beta}_0, \tilde{\beta}_1$ 称为 β_0, β_1 的加权最小二乘估计。
- 可以证明加权最小二乘估计存在且惟一。
- 令两个一阶偏导数都等于零的方程组

$$\beta_0 \sum_{i=1}^{m} \frac{1}{\nu_i} + \beta_1 \sum_{i=1}^{m} \frac{x_i}{\nu_i} = \sum_{i=1}^{m} \frac{z_i}{\nu_i}$$
$$\beta_0 \sum_{i=1}^{m} \frac{x_i}{\nu_i} + \beta_1 \sum_{i=1}^{m} \frac{x_i^2}{\nu_i} = \sum_{i=1}^{m} \frac{x_i z_i}{\nu_i}$$

- 记

$$l_1 = \sum_{i=1}^{m} \frac{1}{\nu_i}, \qquad l_2 = \sum_{i=1}^{m} \frac{x_i}{\nu_i}$$
$$l_3 = \sum_{i=1}^{m} \frac{x_i^2}{\nu_i} \qquad l_4 = \sum_{i=1}^{m} \frac{x_i z_i}{\nu_i}$$
$$l_5 = \sum_{i=1}^{m} \frac{z_i}{\nu_i}$$

- 则

$$\tilde{\beta}_0 = \frac{l_5 l_3 - l_2 l_4}{l_1 l_3 - l_2^2} \tag{3.6}$$
$$\tilde{\beta}_1 = \frac{l_1 l_4 - l_2 l_5}{l_1 l_3 - l_2^2} \tag{3.7}$$

加权最小二乘法的理由

- 应该用 $\frac{\gamma_i}{n_i - \gamma_i}$ 作为 $\frac{p(x_i)}{1-p(x_i)}$ 的估计，为避免分子和分母出现零，做连续型修正变成 $\frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$。

- 可以证明，

$$z_i = \ln \frac{\gamma_i + 0.5}{n_i - \gamma_i + 0.5}$$

近似服从正态分布

$$\mathrm{N}\left(\ln \frac{p(x_i)}{1-p(x_i)}, \frac{1}{n_i p(x_i)[1-p(x_i)]}\right)$$

- 利用 (3.2)，有

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, m$$

其中 ε 近似服从 $\mathrm{N}(0, \nu_i)$。

- 令

$$\tilde{\varepsilon}_i = \frac{1}{\sqrt{\nu_i}} \varepsilon_i$$

- 则

$$\frac{1}{\sqrt{\nu_i}} z_i = \frac{1}{\sqrt{\nu_i}} (\beta_0 + \beta_1 x_i) + \tilde{\varepsilon}_i, i = 1, 2, \ldots, n$$

其中 $\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_n$ 的方差相等，仿照最小二乘法思想令

$$\sum_{i=1}^{m} \left[\frac{1}{\sqrt{\nu_i}} z_i - \frac{1}{\sqrt{\nu_i}} (\beta_0 + \beta_1 x_i)\right]^2$$

达到最小，即 $\tilde{Q}(\beta_0, \beta_1)$ 达到最小。

- **例 3.1（社会调查）** 一个人在家是否害怕生人来？
- 研究人的文化程度对此问题的影响。
- 因变量 $Y = 1$ 表示害怕，0 表示不害怕。
- 自变量 x 是文化程度，$x_1 = 0$ 表示文盲，$x_2 = 1$ 表示小学，$x_3 = 2$ 表示中学，$x_4 = 3$ 表示大专以上。

- 根据一项社会调查有如下数据：

自变量 (x)	不害怕人数	害怕人数
0	11	7
1	45	32
2	664	422
3	168	72

- 用逻辑斯蒂回归模型分析。用 $p(x)$ 表示文化程度为 x 的人害怕生人的概率。设模型

$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x$$

- 用加权最小二乘法估计 β_0, β_1。
- 计算得 $z_1 = -0.3847$, $z_2 = -0.3269$, $z_3 = -0.4515$, $z_4 = -0.8425$, $\nu_1 = 0.2199$, $\nu_2 = 0.0527$, $\nu_3 = 0.00387$, $\nu_4 = 0.0197$。
- 用 (3.6) 和 (3.7) 得 $\tilde{\beta}_0 = 0.013$, $\beta_1 = -0.25$。
- 回归方程为

$$\ln \frac{p(x)}{1-p(x)} \approx 0.013 - 0.25x$$

- 可见文化程度越高，害怕生人的概率越低。

在统计软件中计算逻辑斯蒂回归

- 一般用统计软件计算逻辑斯蒂回归。
- 如上例的 R 程序：

```
x <- 0:3
n1 <- c(11, 45, 664, 168)
n0 <- c(7, 32, 422, 72)
```

```
y <- cbind(n1, n0)
glm1 <- glm(y ~ x, family=binomial)
print(summary(glm1))
```

第八章　正交试验法

试验设计

- 在工业领域，为了改进产品、工艺，经常需要进行试验。
- 改善产品指标常涉及多个因素。
- 那些因素的影响最重要，影响是怎样的？
- 经常进行试验。
- 全面试验只适用于一两个因素的情况。
- 多因素试验使用正交试验法，可以节约费用和时间，快速实现质量改进目标。

8.1　正交表

正交表 $L_8(2^7)$

试验号	列号						
	1	2	3	4	5	6	7
1	1	1	1	2	2	1	2
2	2	1	2	2	1	1	1
3	1	2	2	2	2	2	1
4	2	2	1	2	1	2	2
5	1	1	2	1	1	2	2
6	2	1	1	1	2	2	1
7	1	2	1	1	1	1	1
8	2	2	2	1	2	1	2

$L_8(2^7)$ 的特点

- 有 8 行，7 列，只包含数字 1 和 2。
- 每直列恰有 4 个“1”和 4 个“2”;
- 任意两个直列，行组合的 8 对数字中，恰好 (1,1), (1,2), (2,1), (2,2) 各出现 2 次。即任意两个直列的搭配是均衡的。

正交表 $L_9(3^4)$

试验号	列号			
	1	2	3	4
1	1	1	2	2
2	2	1	1	1
3	3	1	2	3
4	1	2	2	1
5	2	2	3	3
6	3	2	1	2
7	1	3	1	3
8	2	3	2	2
9	3	3	3	1

$L_9(3^4)$ 的特点

- 有 9 行，3 列，只包含数字 1，2，3。
- 每直列中，数字“1”，“2”，“3”出现的次数相同，都是 3 次；
- 任意两个直列，同行的数字的搭配，两两搭配恰有 9 种可能情况，每种情况恰好出现 1 次，即任意两直列搭配均衡。
- 正交表记号的含义：$L_8(2^7)$ 中，L 是表格的意思，8 为行数，2 为数字个数，7 为列数。

8.2 几个实例

8.2.1 2,4–二硝基苯肼的工艺改进

2,4–二硝基苯肼的工艺改进

- 1. 试验目的与考核指标：试剂产品。采用了新工艺后产率低 (45%)，考核指标是产率与外观（颜色）。
- 2. 制定因素位级表：分析找出最有可能的影响因素。要考察乙醇用量、水合肼用量、反应温度、反应时间、水合肼纯度和搅拌速度六种因素。每个因素取两种不同的值（叫做**位级**或因素水平）。综合成因素位级表：

因素	乙醇用量 A(mL)	水合肼用量 B	温度 C	时间 D (h)	水合肼纯度 E	搅拌速度 F
位级 1	200	2 倍	回流	4	精品	中快
位级 2	0	1.2 倍	60°C	2	粗品	快速

- 3. 确定试验方案。6 个因素所有位级完全搭配需要 $2^6 = 64$ 次试验。用 $L_8(2^7)$ 正交表安排试验方案。
- (1) 因素顺序上列。
- (2) 位级对号入座。

- (3) 列出试验条件。
- 方案见下表（包括试验结果）。从全面试验的 64 次中选取了有代表性的 8 次试验：
- 每个因素的每个位级都试验了 4 次；
- 两个因素的位级两两搭配都试验了 2 次。
- 严格按照试验方案进行试验得到结果。试验次序没有关系。

表 8.2

试验计划							试验结果	
因素 / 试验列号	乙醇用量 A	水合肼用量 B	温度 C	时间 D	水合肼纯度 E	搅拌速度 F	产率 (%)	颜色
试验号	1	2	3	4	5	6		
1	1(200 mL)	1(2 倍)	1(回流)	2(2 h)	2(粗品)	1(中快)	56	桔黄
2	2(0 mL)	1	2(60°C)	2	1(精品)	1	65	紫色,桔黄
3	1	2(1.2 倍)	2	2	2	2(快)	54	桔黄
4	2	2	1	2	1	2	43	桔黄
5	1	1	2	1(4 h)	1	2	63	桔黄
6	2	1	1	1	2	2	60	桔黄
7	1	2	1	1	1	1	42	紫色,桔黄
8	2	2	2	1	2	1	42	桔黄
Ⅰ = 位级 1 四次产率之和	215	244	201	207	213	205	Ⅰ + Ⅱ = 425 = 总和	
Ⅱ = 位级 2 四次产率之和	210	181	224	218	212	220		
极差 R = Ⅰ, Ⅱ中, 大数 − 小数	5	63	23	11	1	15		

- 4. 试验结果的分析。结果见上表。
- (1) 直接看。第 2 号试验条件产率 65% 最高。其次是第 5 号试验产率 63%。这是实际试验结果，比较可靠。发现第 2 号和第 7 号试验出现紫色不合格产品，分析认为可能与加料速度有关，在第二轮试验中考察。
- (2) 算一算。以上的最好结果只在 8 次试验中比较，有可能有其它组合更好。
- 对每个因素，分别计算两个位级的总产率 I 和 II，并计算极差 R（为 I 和 II 的差的绝对值）。或计算两个位级的平均指标以及平均指标的差。

- 对每个因素，找到总指标（或平均指标）最大的位级。

- 对所有因素，极差越大，因素越重要。第 2 列水合肼用量最重要，理论量的 2 倍比理论量的 1.2 倍明显地提高产率，可在下轮试验中进一步考察。

- 第 3，6，4 列的反应温度、搅拌速度、反应时间是中等重要的因素，生产中可以采用它们的好位级（C2, F2 和 D2）。

- 第 1，5 列的乙醇用量和水合肼纯度 R 很小，是次要因素，采用有利于节约的不加乙醇 (A2) 和粗品水合肼（E2）。

- 5. 第二批正交试验。

- 目标是弄清出现颜色不合格的原因并进一步提高收率。

- （1）指定因素位级表。

因素	水合肼用量	时间	加料速度
位级 1	1.7 倍	2h	快
位级 2	2.3 倍	4h	慢

  其中反应时间再次试验是因为工程人员对能否用 2 小时代替 4 小时很重视。

- （2）利用正交表确定试验方案。只有三个因素恰好可以用 $L_4(2^3)$ 表。

表 8.3

试验计划				试验结果	
因素 / 试验号	水合肼用量 A	时间 B	加料速度 C	产率 (%)	颜色
	1	2	3		
1	1(1.7 倍)	1(2 h)	1(快)	62	不合格
2	2(2.3 倍)	1	2(慢)	86	合 格
3	1	2(4 h)	2	70	合 格
4	2	2	1	70	不合格
Ⅰ=位级 1 二次产率之和	132	148	132		
Ⅱ=位级 2 二次产率之和	156	140	156	Ⅰ+Ⅱ=288=总和	
极差 R = Ⅰ,Ⅱ中,大数 − 小数	24	8	24		

- （3）试验结果分析。
- 关于颜色，“快速加料”的第 1，4 号都不合格，慢速加料都合格。其它两个因素的每个位级则都出现了合格和不合格。说明只有加料速度影响颜色。
- 关于产率，从“直接看”和“算一算”，都是第 2 号试验条件结果最好。产率从开始的 45% 提高到了 86%。
- 投产效果。通过正交试验法，决定用下列工艺投产：用工业 2，4–二硝基氯代苯与粗品水合肼在乙醇溶剂中合成；水合肼用量为理论值的 2.3 倍，反应时间为 2 小时，温度掌握在 60 ～ 70°C 之间，采用慢速加料与快速搅拌。
- 效果是：平均产率超过 80%，从未出现紫色外形，质量达到出口标准。
- 总之，这是一个较优的方案，可以达到优质、高产、低消耗的目的。

8.2.2 晶体退火工艺的改进

晶体退火工艺的改进

- 1. 试验目的与考核指标。

- 检查癌细胞用到一种碘化钠晶体 $\phi 40$，要求应力越小越好，希望不超过 2 度。

- 退火工艺是影响质量的重要环节。

- 其它指标已经合格，只有应力未能低于 7 度，希望找到降低应力的工艺条件。

- 考核指标是应力（度）。

- 2. 挑因素、选位级，制定因素位级表

- 考察升温速度、恒温温度、恒温时间和降温速度共四个因素，每个因素取三个位级。

- 因素位级表如下：

因素	升温速度 A ($^\circ$C·h^{-1})	恒温温度 B $^\circ$C	恒温时间 C (h)	降温速度 D
位级 1	30	600	6	1.5A 电流降温
位级 2	50	450	2	1.7A 电流降温
位级 3	100	500	4	15°C ·h^{-1}

- 3. 确定试验方案。用 $L_9(3^4)$ 正交表。

- 试验方案和结果：

表 8.4

· 324 ·

试验号 \ 因素 列号	升温速度 A	恒温温度 B	恒温时间 C	降温速度 D	应力
	1	2	3	4	(度)
1	1(30 ℃·h⁻¹)	1(600 ℃)	3(4 h)	2(1.7A 电流降温)	6
2	2(50 ℃·h⁻¹)	1	1(6 h)	1(1.5A 电流降温)	7
3	3(100 ℃·h⁻¹)	1	2(2 h)	3(15 ℃·h⁻¹)	15
4	1	2(450 ℃)	2	1	8
5	2	2	3	3	0.5
6	3	2	1	2	7
7	1	3(500 ℃)	1	3	1
8	2	3	2	2	6
9	3	3	3	1	13
Ⅰ = 位级 1 三次应力之和	15	28	15	28	
Ⅱ = 位级 2 三次应力之和	13.5	15.5	29	19	
Ⅲ = 位级 3 三次应力之和	35	20	19.5	16.5	Ⅰ + Ⅱ + Ⅲ = 63.5 度 = 总和
极差 R = Ⅰ, Ⅱ, Ⅲ中, 最大数 − 最小数	21.5	12.5	14	11.5	

- 4. 试验结果分析。

- （1）直接看。第 5 号试验的 0.5 度结果最好，第 7 号试验的 1 度次之。

- （2）算一算。依因素重要程度：

- 升温速度 A，A2 的 50°C·h^{-1} 最好，这是原工艺条件。

- 恒温时间 C，C1 的 6 小时最好，这是原工艺条件。

- 恒温温度 B，B2 的 450°C 最好，原来需要 600°C。

- 降温速度 D，D3 的等速降温最好，原来认为等速降温不好。

- 把四个因素的最好位级搭配 A2,B2,C1,D3 搭配起来叫做全体配合中关于应力的可能好配合。

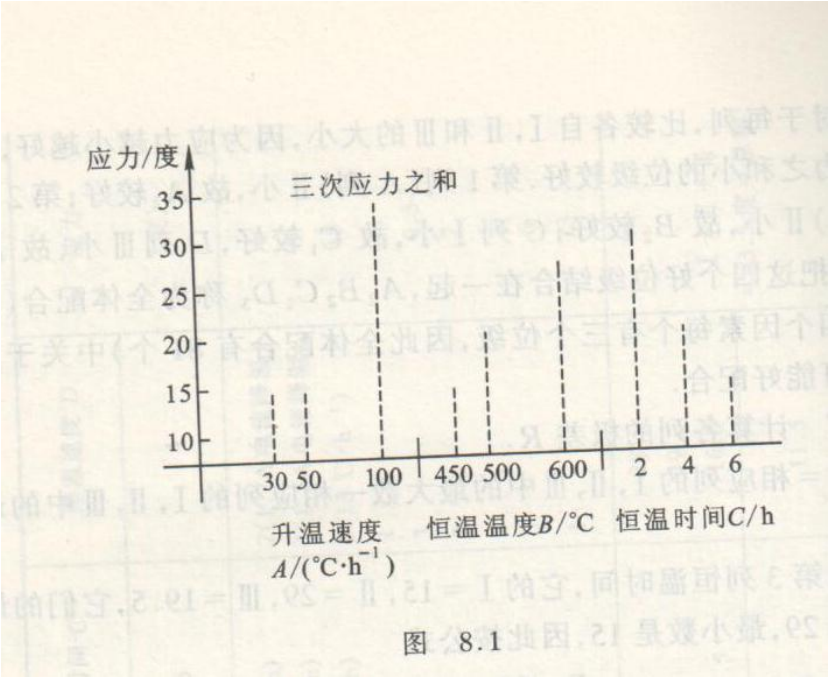


图 8.1

- （3）画趋势图。对于数量型位级的三位级因素，画出用量与实验结果之和的关系图。
- （4）可能好配合与大范围可能好配合。
- i) 通过“算一算得到 A2,B2,C1,D3 为 81 个配合中的可能好配合。
- ii) 可能好位级与大范围可能好配合。从趋势图中看出，恒温温度越低越好，恒温时间越长越好。
- 原来选的恒温温度都太高了，这是一个重要发现。
- 但恒温时间太长不利于节约电力和提高工效，所以恒温时间固定在 4 小时。
- 于是，恒温温度 B 的好位级是 B2=450°C, 但有可能降到 B4=400°C 更好；
- 恒温时间为节约成本、提高功效取为 C3=4h；
- 升温速度维持原值 A2=50°C h^{-1}；

- 降温速度好位级是等速降温 D3=15°C $\cdot h^{-1}$。比较发现原来说等速降温不好是因为原来用的恒温温度太高所以等速降温不好，这种因素之间互相影响的情况叫做**交互作用**。

- 5. 第二批正交试验

- 主要研究降低恒温温度的影响。

- （1）制定因素位级表。

- 升温速度，升温快的 A3=100°C $\cdot h^{-1}$，第 3，6，9 号试验应力都很差，不能再用；升温满的 A1=30°C $\cdot h^{-1}$，升温时间过长，而且也不是最好位级，不再使用。所以此次升温速度固定在好位级 50°C $\cdot h^{-1}$ 不再考虑。

- 其它三个因素，以大范围的可能好配合 B4, C3, D3 为主，参考“直接看”的好位级，各取两个位级。

- 因素位级表：

因素	恒温温度 (°C)	恒温时间 C (h)	降温速度 D (°C $\cdot h^{-1}$
位级 1	450	3	15
位级 2	400	5	25

- (2) 用正交表 $L_4(2^3)$ 确定试验方案。

- 方案及结果：

表 8.5

试验号 \ 因素列号	1 恒温温度	2 恒温时间	3 等速降温	应力（度）
1	1(450℃)	1(3 h)	1(15℃·h^{-1})	0
2	2(400℃)	1	2(25℃·h^{-1})	0.2
3	1	2(5 h)	2	0.4
4	2	2	1	0
Ⅰ	0.4	0.2	0	Ⅰ + Ⅱ = 0.6
Ⅱ	0.2	0.4	0.6	
R	0.2	0.2	0.6	

- (3) 试验结果分析。
- 从试验结果看都已经基本消除了应力，可以采用对生产最节省、效率最高的方案。
- 正交试验考虑的因素多，正交试验方案相当于撒了一个网眼均匀的大网，能够捕获“大鱼”。
- 三个位级时可以进一步考虑数量型因素的趋势影响。
- 再次试验使用网眼更密的小网获得了更理想的结果。

8.2.3 VC 的配方试验

VC 的配方试验

- 1. 试验目的与考核指标。
- 提高 VC 的氧化率、降低成本。考核指标是氧化率。
- 2. 确定试验方案
- （1）因素位级表

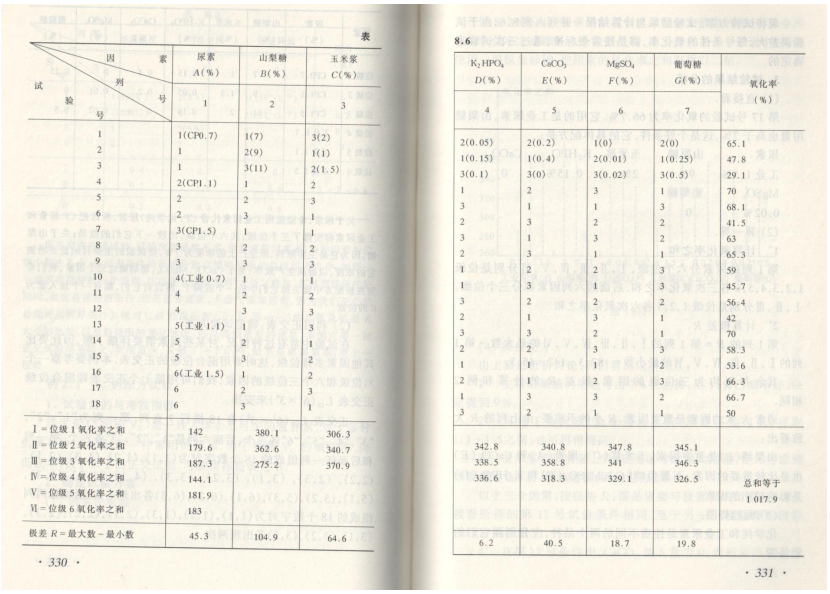
因素	尿素（%）	山梨糖（%）	玉米浆（%）	K_2HPO_4（%）	$CaCO_3$（%）	$MgSO_4$（%）	葡萄糖（%）
位级 1	CP0.7	7	1	0.15	0.4	0	0.25
位级 2	CP1.1	9	1.5	0.05	0.2	0.01	0
位级 3	CP1.5	11	2	0.10	0	0.02	0.5
位级 4	工业 0.7						
位级 5	工业 1.1						
位级 6	工业 1.5						

- 其中尿素希望能用工业尿素代替化学纯尿素。山梨糖的原生产浓度为7%，考察增加后能否提高生产效率。$CaCO_3$, $MgSO_4$，葡萄糖希望能去掉一个或两个。

- （2）利用正交表，确定试验方案。

- 用混合位级的正交表，一个六位级因素和六个三位级因素，用 $L_{18}(6^1 \times 3^6)$ 来安排。

- 试验方案和结果如下表。



表

试验号 \ 因素 列号	尿素 A(%)	山梨糖 B(%)	玉米浆 C(%)
	1	2	3
1	1(CP0.7)	1(7)	3(2)
2	1	2(9)	1(1)
3	1	3(11)	2(1.5)
4	2(CP1.1)	1	2
5	2	2	3
6	2	3	1
7	3(CP1.5)	1	1
8	3	2	2
9	3	3	3
10	4(工业 0.7)	1	1
11	4	2	2
12	4	3	3
13	5(工业 1.1)	1	3
14	5	2	1
15	5	3	2
16	6(工业 1.5)	1	2
17	6	2	3
18	6	3	1
Ⅰ = 位级 1 氧化率之和	142	380.1	306.3
Ⅱ = 位级 2 氧化率之和	179.6	362.6	340.7
Ⅲ = 位级 3 氧化率之和	187.3	275.2	370.9
Ⅳ = 位级 4 氧化率之和	144.1		
Ⅴ = 位级 5 氧化率之和	181.9		
Ⅵ = 位级 6 氧化率之和	183		
极差 R = 最大数 − 最小数	45.3	104.9	64.6

· 330 ·

8.6

K_2HPO_4 D(%)	$CaCO_3$ E(%)	$MgSO_4$ F(%)	葡萄糖 G(%)	氧化率 (%)
4	5	6	7	
2(0.05)	2(0.2)	1(0)	2(0)	65.1
1(0.15)	1(0.4)	2(0.01)	1(0.25)	47.8
3(0.1)	3(0)	3(0.02)	3(0.5)	29.1
1	2	3	1	70
3	1	1	3	68.1
2	3	2	2	41.5
3	1	3	2	63
2	3	1	1	65.3
1	2	2	3	59
1	3	1	3	45.7
3	2	2	2	56.4
2	1	3	1	42
3	3	2	1	70
2	2	3	3	58.3
1	1	1	2	53.6
2	1	2	3	66.3
1	3	3	2	66.7
3	2	1	1	50
342.8	340.8	347.8	345.1	
338.5	358.8	341	346.3	
336.6	318.3	329.1	326.5	总和等于 1 017.9
6.2	40.5	18.7	19.8	

· 331 ·

- 3. 试验结果的分析
- （1）直接看。第 17 号结果氧化率 66.7 比较高，用了工业尿素，山梨糖用量超过 7%，是好条件。
- 具体配方为：
- 尿素：工业 1.5%;
- 山梨糖：9%;
- 玉米浆：2%;
- K_2HPO_4: 0.15%;
- $CaCO_3$: 0;
- $MgSO_4$: 0.02%;
- 葡萄糖：0。
- (2) 算一算。每个因素计算各个位级对应的指标之和和极差。
- 从极差大小看，山梨糖（B）是重要因素，玉米浆 (C)、尿素 (A) 和 $CaCO_3$(E) 也是比较重要的因素。

- 葡萄糖 (G)、$MgSO_4$（F）和 K_2HPO_4 是影响较小的因素。
- (3) 画趋势图。

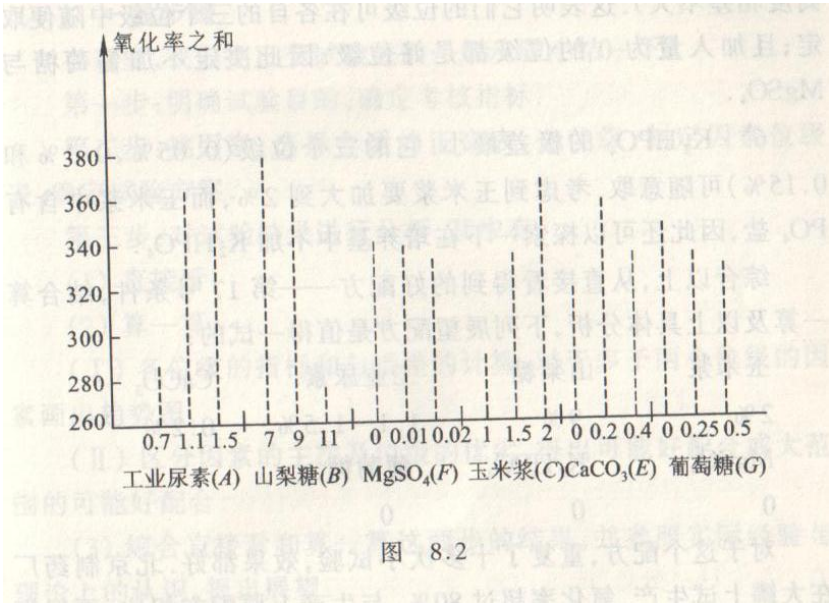


图 8.2

- 尿素只考虑工业尿素，而且因为求和个数不同所以尿素的趋势图数值乘了 2 倍。
- (4) 展望配方。
- i) 主料山梨糖的用量可以由 7% 往上提，但 11% 已知不好，决定提到 9%。
- ii) 可用工业尿素，由于图形上升，用量不能过低，可取 1.1 ∼ 1.5 之间。
- iii) 玉米浆图形明显商城，还可以考虑加大用量，但又考虑到浓度超过 2% 后溶液过稠，因此定位 2%。
- 以上三个因素都是重要与较重要的因素，而且与直接看所得的第 17 号试验条件相同。
- iv) 在第 17 号条件中 $CaCO_3$ 的加入量为 0，而趋势图显示中间值 0.2% 最好，讨论认为算一算的结果值得重视。

- v) 葡萄糖和 $MgSO_4$ 的极差很小（趋势图上三个位级的差别不大），可以随意选取，而且加入量为 0 的位级都是好位级，决定不加葡萄糖和 $MgSO_4$。
- vi) K_2HPO_4 的极差最小，三个位级可以随意取，而玉米浆的浓度加大后考虑到玉米浆中含有 PO_4 盐，考虑尝试不加 K_2HPO_4。
- 展望配方如下：
- 玉米浆: 2%;
- 山梨糖：9%;
- 工业尿素：$1.1 \sim 1.5\%$；
- $CaCO_3$: 0.2%;
- $MgSO_4$: 0;
- K_2HPO_4: 0;
- 葡萄糖：0。
- 用这个展望配方重复了十多次小试验，效果都好。
- 工业化生产氧化度超过 80%，与原配方相比，逼近提高了主料山梨糖的浓度，还减少了三种成分，达到了节约成本、简化工艺和提高生产效率的目的。

$L_{18}(6^1 \times 3^6)$ 正交表特点

- 不仅可以安排众多的 3 位级因素，还可以安排一个 6 位级因素。
- 用它安排的试验既照顾了一般又突出了重点。
- 当试验难度大，试验次数受很大限制，要考察的因素较多时用该表合适。

8.3 小结

小结—一般步骤

- 1. 明确试验目的，确定考核指标。
- 2. 挑因素，选择合适的正交表，选位级，制定因素位级表，确定试验方案。
- 3. 试验结果分析。包括
- （1）直接看；
- （2）算一算，计算各位级的指标和与极差，对于两个位级的数量型因素画趋势图。
- 区分因素的主次和位级的优劣，得出可能好配合或大范围的可能好配合。
- （3）综合直接看和算一算这两步的结果，并参考实际经验与理论上的认识，提出展望。

小结—关于挑选因素

- 要考察的因素的分类：
- （1）不能测量也不能定性了解的因素，不能选入。
- （2）能测量但是不能控制的因素，不能选入；能近似控制则可以选入但要记录真实试验值。
- （3）可控因素。由试验工作者决定，但注意：漏掉重要因素会大大降低试验效果；正交试验不怕因素多；有时增加一两个因素不会增加试验次数。
- 所以倾向于多考虑因素，除了实现能肯定作用很小的因素不选入之外其它因素应尽可能都选。

小结—选择合用的正交表及其它

- (1) 位级个数的确定。有些是不可变更的，有些数量型的则需要认为选取并配合正交表。
- (2) 正交表的选择：要考察因素的个数，一批允许做试验的次数，有无重点因素要详细考察，位级个数 (位级个数和正交表选择需要同时考虑)。
- (3) 位级用量的选取。确定范围后二位级可以取范围的三等分点，三位级可取范围的四等分点，或根据实际情况定。
- (4) 制定因素位级表。位级号码与实际因素数值一般打乱次序对应。
- (5) 确定试验方案。

均衡分散性和整齐可比性

- 正交表任意两列，搭配个数总是一样多，叫做正交性。
- 这保证了试验条件均衡地分散在配合完全的位级组合之中，因而代表性强，容易出现好条件，这叫做**均衡分散性**。
- 对于各列因素，在每个位级的指标和中，任意一个其它因素的不同位级的出现次数是一样多的，最大限度地排除了其它因素的干扰，因而能很有效第进行比较，为我们提供有参考价值的展望。这叫做**整齐可比性**。
- 正交试验法效率高的原因主要在于这两种特性。

第九章　统计决策与贝叶斯统计大意

9.1　统计决策问题概述

统计决策的要素

- 统计决策由四个要素组成。
- 设随机变量 X 的分布函数是 $F(x,\theta)$，θ 是未知参数，$\theta \in \Theta$, Θ 叫做**参数空间**。
- 设 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 为 X 的样本。
- 设 A 是某项实际工作中可能采取的各种行动所组成的非空集合，叫做**行动空间**。
- $L(\theta, a)$ 是二元函数 $(\theta \in \Theta, a \in A)$，表示参数是 θ 时采取行动 a 引起的损失，叫做**损失函数**。
- $(\Theta, \mathbf{X}, A, L(\theta, a))$ 叫做统计决策的四个要素。

统计决策

- **定义 1.1** 称样本空间（及样本所有可能值组成的集合）到行动空间 A 的映射 $\delta = \delta(x_1, x_2, \ldots, x_n)$ 为**决策函数**，简称**决策**。

- **定义 1.2** 设 $\delta = \delta(x_1, x_2, \ldots, x_n)$ 是一个决策，称平均损失

$$R(\theta, \delta) = E\left[L(\theta, \delta(X_1, X_2, \ldots, X_n))\right]$$

为 δ 的**风险**。

- 希望找到对所有 $\theta \in \Theta$ 风险一致小的决策，但一般不存在。

minimax 决策

- **定义 1.3** 称决策 $\delta = \delta(x_1, x_2, \ldots, x_n)$ 是容许的，如果不存在另一决策 $\tilde{\delta}$ 使得

$$R(\theta, \tilde{\delta}) \leq R(\theta, \delta) \quad \forall \theta \in \Theta$$

且对至少一个 θ 成立严格不等式。

- **定义 1.4** 称决策 δ^* 是 minimax 决策，若对一切决策 δ 成立

$$\sup_{\theta} R(\theta, \delta^*) \leq \sup_{\theta} R(\theta, \delta)$$

- minimax 是一种保守的决策，目的是避免可能最大损失。
- 如果能了解哪些 θ 可能性大，哪些 θ 可能性小，就可以计算关于 θ 的平均损失，使这样的平均损失最小。

9.2 什么是贝叶斯统计

贝叶斯统计

- 在贝叶斯统计中，设总体 $X \sim p(x|\theta)$，参数 θ 也看成随机的，有**先验分布** $\pi(\theta)$。
- 有了样本 X 后，得条件分布

$$p(\theta|X) = \frac{\pi(\theta)p(x|\theta)}{p(x)}$$

叫做参数的**后验分布**。其中

$$p(x) = \int p(x|\theta)\pi(\theta)d\theta$$

- 先验分布也可以有参数，叫做**超参数**。

贝叶斯统计的优点

- 可以利用先验知识。
- 很多问题比经典统计叙述更简单。
- 可以提出任意复杂的模型，没有解析解时主要使用计算机模拟方法得到后验分布进行推断。

第十章　随机过程初步

10.1　随机过程的概念

随机过程介绍

- **定义 1.1** 给定参数集 $T \subset (-\infty, \infty)$，如果对每个 $t \in T$，对应一个随机变量 X_t，则称随机变量族 $\{X_t, t \in T\}$ 为**随机过程**。
- **例 1.1** 用 X_t 表示某电话机从时刻 0 开始到时刻 t 为止所接到的呼唤次数，则 $\{X_t, t \in [0, \infty\}$ 是随机过程。
- **例 1.2** 对晶体管热噪声电压进行测量，每隔一微秒测一次。测量时刻记作 $1, 2, \ldots$，在时刻 t 的测量值记作 X_t，则 $\{X_t, t = 1, 2, \ldots\}$ 是随机过程。
- **例 1.3** 布朗运动。X_t 是花粉颗粒在 t 时刻所在位置的横坐标，则 $\{X_t, t \in [0, \infty)\}$ 是随机过程。
- **例 1.4** 一根面纱中，X_t 表示 t 时刻纺出的纱的横截面直径。则 $\{X_t, t \in [0, \infty)\}$ 是随机过程。
- 考察随机现象如何随着时间而变，就会遇到随机过程。
- 用 E 表示各 X_t 所可能取的值所组成的集合，E 叫做**状态空间**（或**相空间**）。如果 $X_t = x$，就称过程 $\{X_t, t \in T\}$ 在时刻 t 处于状态 x。

- 当 T 是一个有限集或可列集时，$\{X_t, t \in T\}$ 叫做**离散时间的随机过程（随机序列）**。最常见的情况是 $T = \{0, 1, 2, \dots\}$ 或 $T = \{\dots, -1, 0, 1, \dots\}$。
- 当 T 是一个区间（可以是无穷区间）时，$\{X_t, t \in T\}$ 叫做**连续时间的随机过程**。最常见的情况是 $T = [0, \infty)$ 或 $T = (-\infty, \infty)$。

- 给定 T 中 n 个数 $t_1, t_2, \dots, t_n$，记 $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ 的分布函数为 $F_{t_1 t_2 \dots t_n}(x_1, x_2, \dots, x_n)$，这种分布函数的全体

$$\{F_{t_1 t_2 \dots t_n}(x_1, x_2, \dots, x_n) : n = 1, 2, \dots, t_1, t_2, \dots, t_n \in T\}$$

  叫做 $\{X_t, t \in T\}$ 的 **有限维分布函数族**, 它描述了随机过程的概率特性。

- 随机过程也可以看成是二维函数：

$$X_t = X_t(\omega), \quad t \in T, \omega \in \Omega$$

  其中 Ω 是条件组 S 下所有可能结果的集合。给定 $\omega \in \Omega$ 后，$X_t(\omega)$ 是 t 的函数，叫做随机过程的一个**实现**，或**现实**, 或**轨道**。

10.2 独立增量过程

独立增量过程

- **定义 2.1** 称 $\{X_t, t \in T\}$ 为**独立增量过程**，如果对任何 $t_1 < t_2 < \cdots < t_n (t_i \in T, i = 1, 2, \dots, n)$，随机变量

$$X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \dots, X_{t_n} - X_{t_{n-1}}$$

  是相互独立的。

- 如果对独立增量过程 $\{X_t, t \in T\}$，对任意 $\tau > 0$ 都有 $X_{t+\tau} - X_t$ 的分布函数只依赖于 τ 而不依赖于 t，则称 $\{X_t, t \in T\}$ 为**时齐的独立增量过程**。

- 若 $\{X_t, t \in T\}$ 是独立增量过程，Y 是随机变量，则 $\{X_t + Y, t \in T\}$ 也是独立增量过程。所以独立增量过程可以假定 $X_0 \equiv 0$(当 $0 \in T$ 时)。

- **例 2.1** 设 $X_1, X_2, \dots$ 是相互独立的随机变量列，$S_n = X_1 + X_2 + \cdots + X_n (n = 1, 2, \dots)$，则 $\{S_n, n = 1, 2, \dots\}$ 是独立增量过程。若各 $X_1, X_2, \dots$ 还是同分布的，则 $\{S_n, n = 1, 2, \dots\}$ 是时齐的独立增量过程。

泊松过程

- **定义 2.2** 称 $\{X_t, t \geq 0\}$ 是泊松过程，若它是独立增量的，而且 X_t 取值是非负整数，增量 $X_t - X_s (0 \leq s < t)$ 服从泊松分布

$$P(X_t - X_s = k) = e^{-\lambda(t-s)} \frac{[\lambda(t-s)]^k}{k!}, \ (k = 0, 1, \dots)$$

  其中 λ 是与 t, s 无关的正常数。

- **定理 2.1** 设 $\{X_t, t \geq 0\}$ 是取非负整数值的时齐的独立增量过程，满足

$$\begin{aligned} &P(X_0 = 0) = 1 \\ &P(X_{t+\Delta t} - X_t = 1) = \lambda \Delta t + o(\Delta t) \quad (\Delta t \to 0+) \\ &P(X_{t+\Delta t} - X_t \geq 2) = o(\Delta t) \end{aligned}$$

  (这里 $\lambda > 0$)。则 $\{X_t, t \geq 0\}$ 就是泊松过程。

维纳过程

- **定义 2.3** 称独立增量过程 $\{X_t, t \geq 0\}$ 是维纳过程，如果对任何 $s < t$，

$$X_t - X_s \sim \mathrm{N}(0, (t-s)\sigma^2)$$

  这里 σ 是与 s, t 无关的正常数。

10.3 马尔可夫过程

马尔可夫过程

- 设 $\{X_t, t \in T\}$ 是一个随机过程，状态空间是 E，我们可以把这个随机过程看成某系统的“状态”的演变过程。“$X_t = x$”表示该系统在时刻 t 处于状态 x。
- **定义 3.1** 称 $\{X_t, t \in T\}$ 是马尔可夫过程，如果对于 T 中任何 n 个数 $t_1 < t_2 < \cdots < t_n$，E 中任何 n 个状态 $x_1, x_2, \ldots, x_n$ 及任何实数 x 均成立

$$\begin{aligned} &P(X_{t_n} \leq x | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, x_{t_{n-1}} = x_{n-1}) \\ =&P(X_{t_n} \leq x | x_{t_{n-1}} = x_{n-1}) \end{aligned} \tag{3.1}$$

- 即马尔可夫过程的特征是，如已知“现在：$X_{t_{n-1}} = x_{n-1}$”，则“将来: $X_{t_n} \leq x$”不依赖于“过去：$X_{t_1} = x_1, X_{t_2} = x_2, \ldots, x_{t_{n-2}} = x_{n-2}$”。
- 这表达了过程的“无后效性”。
- (3.1) 所表达的性质称为**马氏性**。马尔可夫过程简称**马氏过程**。

马尔可夫链

- 这里对 $T = \{0, 1, 2, \ldots\}$，E 为至多可列集的情形作初步介绍。
- **定义 3.2** 设 $\{X_n, n = 0, 1, 2, \ldots\}$ 是随机序列，状态空间 E 至多可列，若对任何 $i_0, i_1, \ldots, i_n \in E$，只要

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \neq 0$$

就成立

$$\begin{aligned} &P(X_n = i_n | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\ =&P(X_n = i_n | X_{n-1} = i_{n-1}) \end{aligned}$$

则称 $\{X_n, n = 0, 1, 2, \ldots\}$ 为**马尔可夫链**，简称**马氏链**。

- 马氏链是一种特殊的马氏过程。

转移概率

- 条件概率 $P(X_t = j|X_s = i)$ 称为**转移概率**，记作 $p_{ij}(s,t)$(这里 $s \le t$)。
- **定理 3.1** 设 $\{X_n, n = 0,1,2,\dots\}$ 是马氏链，则对 $s < t < u$，有

$$p_{ij}(s,u) = \sum_{k\in E} p_{ik}(s,t)p_{kj}(t,u) \tag{3.3}$$

- (3.3) 叫做 Chapman-Kolmogorov 方程。

时齐马氏链

- 若任意固定 i,j 后，$p_{ij}(s,t) = p_{ij}(s+\tau, t+\tau)$ （对一切 $\tau \ge 0$），则称马氏链 $\{X_n, n = 0,1,2,\dots\}$ 是**时齐**的，也叫齐次的。以下只讨论时齐马氏链。
- 对时齐马氏链记

$$p_{ij} = P(X_{t+1} = j|X_t = i)$$

  称矩阵 $P = (p_{ij}, i,j \in E)$ 为**一步转移概率矩阵**。

- **定理 3.2**

$$\begin{aligned} p_{ij} &\ge 0 \\ \sum_{j\in E} p_{ij} &= 1 \quad (\forall i \in E) \end{aligned}$$

- 只要一步转移概率 $P = (p_{ij})$ 知道了，则马氏链的转移概率特性就完全确定了。
- 记

$$p_{ij}^{(n)} = P(X_{s+n} = j|X_s = i)$$

  表示 n 步转移概率。

- 矩阵 $(p_{ij}^{(n)}, i, j \in E) = P^n$。
- **例 3.1（自由随机游动）** 某质点在整数点集 $\{\dots, -1, 0, 1, \dots\}$ 上随机游动。设开始时指点在位置 0，以后每经过一个单位时间按下列概率规则改变一次位置：
- 如果它在某时刻位于点 i，则它以概率 $p(0 < p < 1)$ 转移到 $i+1$，以概率 $1-p$ 转移到 $i-1$。
- 用 X_n 表示质点在时刻 n 所在的位置，则 $\{X_n, n = 0, 1, 2, \dots\}$ 就是一个马氏链，转移概率为

$$p_{ij} = \begin{cases} p & j = i+1 \\ 1-p & j = i-1 \\ 0 & j = i \text{或} j < i-1 \text{或} j > i+1 \end{cases}$$

- 质点在整数点集 $\{0, 1, 2, \dots\}$ 上随机游动。
- 转移规律是：
- 若在某时刻处于位置 $i > 0$，则在下一步以概率 $p(0 < p < 1)$ 转移到 $i+1$，以概率 $1-p$ 转移到 $i-1$。
- 若某时刻处于位置 0，则下一步仍停留在位置 0。
- 如果开始时质点位于 $i_0 (i_0 > 0)$，在时刻 n 时位置记为 X_n，则 $\{X_n, n = 0, 1, 2, \dots\}$ 是一个马氏链，其一步转移概率矩阵为

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

- **例 3.3（Ehrenfest 模型）** 考察带有 $m+1$ 个状态（记为 $0, 1, \dots, m$）的系统的转移问题。
- 其转移规律为：
- 若系统在某时刻处于状态 $i (1 \le i \le m-1)$，则在下一步以概率 $\frac{i}{m}$ 转移到状态 $i-1$，以概率 $1 - \frac{i}{m}$ 转移到状态 $i+1$；

- 若某时刻处于状态 0，则下一步转移到状态 1；
- 若某时刻处于状态 m，则下一步转移到状态 $m-1$。
- 设开始时系统的状态是 X_0，时刻 n 的状态为 X_n，则 $\{X_n, n = 0, 1, 2, \dots\}$ 是马氏链。
- 这个马氏链的状态空间是有限集。

马氏链理论中的问题

- 1. 状态的性质怎样？是否有些状态经常出现？
- 2. 转移概率 $p_{ij}^{(n)}$ 当 $n \to \infty$ 时是否有极限？如果有极限，极限是什么？
- 3. 设马氏链 $\{X_n, n = 0, 1, 2, \dots\}$ 的一步转移概率矩阵 $P = (p_{ij}, i, j \in E)$,什么条件下各 X_n 有相同的概率分布?什么条件下序列 $f(X_0), f(X_1), \dots, f(X_n), \dots$ 符合强大数律，即

$$P(\lim_n \frac{1}{n} \sum_{i=0}^{n-1} |f(X_i) - Ef(X_i)| = 0) = 1$$

- **定义 3.3** 称状态 i 是马氏链 $\{X_n, n = 0, 1, 2, \dots\}$ 的**常返状态**, 若

$$P(\text{存在 } n > s \text{ 使} X_n = i | X_s = i) = 1$$

否则称 i 为**非常返状态**。

- **定义 3.4** 称马氏链 $\{X_n, n = 0, 1, 2, \dots\}$ 是**不可约**的，如果对任意两个状态 i, j，都有

$$P(\text{存在 } n > s \text{ 使} X_n = j | X_s = i) > 0$$

常返的性质

- 如果 i 是常返状态，则从 i 出发将来无穷多次出现 i 的概率等于 1。
- 状态 i 常返的充要条件是

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$$

- 如果状态 i 是常返的，又存在 $n > s$ 使

$$P(X_n = j | X_s = i) > 0$$

则 j 也是常返的。

n 步转移概率性质

- n 步转移概率 $p_{ij}^{(n)}$ 的性质：
- （1）$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} p_{ij}^{(n)} = \pi_{ij}$ 永远存在。
- （2）如果状态空间 E 是有限集，存在 n_0 使

$$p_{ij}^{(n_0)} > 0 \quad (\forall i, j \in E)$$

则 $\lim_n p_{ij}^{(n)} = \pi_j$ 存在 (π_j 与 i 无关)，而且 $\{\pi_j, j \in E\}$ 是下列方程组

$$x_j = \sum_{i \in E} x_i p_{ij} \quad (\forall j \in E)$$

的满足条件 $p_j > 0, \sum_{j\in E} x_j = 1$ 的唯一解。

平稳分布

- 若 X_0 的概率分布 $\{p_i, i \in E\}$ 满足

$$p_i = \sum_{k \in E} p_k p_{ki} \quad (\forall i \in E) \tag{3.4}$$

时每个 X_n 与 X_0 有相同的概率分布。

- 如果 (3.4) 成立且马氏链是不可约，则强大数律成立：

$$P(\lim_n \frac{1}{n}\sum_{i=0}^{n-1}|f(X_i)-Ef(X_i)|=0)=1$$

其中 $f(\cdot)$ 是任意有界函数。

10.4 平稳过程

平稳过程

- 以下恒设参数集 T 具有性质：若 $s,t\in T$ 则 $s+t\in T$。
- **定义 4.1** 称随机过程 $\{X_t,t\in T\}$ 是**严平稳过程**，若对任何 $n=1,2,\ldots$，$t_1,\ldots,t_n,\tau\in T$ 及实数 $x_1,x_2,\ldots,x_n$，都成立

$$\begin{aligned}&P(X_{t_1+\tau}\le x_1,X_{t_2+\tau}\le x_2,\ldots,X_{t_n+\tau}\le x_n)\\=&P(X_{t_1}\le x_1,X_{t_2}\le x_2,\ldots,X_{t_n}\le x_n)\end{aligned}$$

- 即有限维分布函数随着时间的推移而不变。

宽平稳

- 复值随机变量：设 U,V 是实值随机变量，则 $Z=U+iV$ 为复值随机变量，定义 $EZ=EU+iEV$。
- **定义 4.2** 称随机过程 $\{X_t,t\in T\}$ 为宽平稳过程，如果它满足

$$\begin{aligned}&(1)E|X_t|^2\text{存在且有限}\quad(t\in T)\\&(2)E(X_t)\equiv C\quad(t\in T)\\&(3)E[(X_t-C)\overline{(X_{t+\tau}-C)}]\text{只依赖于 }\tau\text{，与 }t\text{ 无关。}\end{aligned}$$

- **定义 4.3** 称函数

$$B(\tau)=E\left\{[X_t-E(X_t)]\overline{[X_{t+\tau}-E(X_{t+\tau})]}\right\}$$

为宽平稳过程的**自协方差函数**。

白噪声

- **定义** 若 $\{X_t, t = \ldots, -1, 0, 1, \ldots\}$ 满足

$$(1)E(X_t) = 0, \quad (t \in T)$$
$$(2)E(X_t^2) = \sigma^2, \quad (t \in T)$$
$$(3)E(X_s X_t) = 0, \quad (s \neq t, s \in T, t \in T)$$

  则称 $\{X_t, t \in T\}$ 为白噪声序列。白噪声序列是平稳过程。

高斯过程

- **定义 4.4** 称 $\{X_t, t \in T\}$ 为**高斯过程**，如果对 T 中任意 n 个不同的数 $t_1, t_2, \ldots, t_n$，$(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ 是随机向量。
- **定理 4.1** 设 $\{X_t, t \in T\}$ 是高斯过程，则为了它是严平稳的，必须且只需它是宽平稳的。

自协方差函数性质

- $|B(\tau)| \leq B(0)$;
- $B(-\tau) = \overline{B(\tau)}$。

均方连续

- 考虑 $T = (-\infty, +\infty)$ 和 $T = [0, +\infty)$ 的情况。
- **定义 4.5** 称随机过程 $\{X_t, t \in T\}$ 是均方连续的，如果对任意 $t \in T$，有

$$\lim_{k \to 0} E|X_{t+k} - X_t|^2 = 0.$$

- 如果 $\{X_t, t \in T\}$ 是宽平稳过程，则它均方连续的充要条件是自协方差函数 $B(\tau)$ 是连续的。

谱函数和谱密度

- **定理 4.2** 设 $B(\tau)$ 是均方连续的宽平稳过程的自协方差函数，则存在唯一的右连续不减函数 $F(\lambda)$ 使得

$$\begin{aligned}\lim_{\lambda\to-\infty} F(\lambda) =&0\\ B(\tau) =&\int_{-\infty}^{+\infty} e^{i\tau\lambda}dF(\lambda) \quad (\forall\tau)\end{aligned} \tag{4.2}$$

- (4.2) 叫做自协方差函数的谱展式，$F(\lambda)$ 叫做过程的**谱函数**。
- 如果存在非负函数 $f(\lambda)$ 使

$$F(x) = \int_{-\infty}^{x} f(\lambda)d\lambda$$

  则称 $f(\lambda)$ 为过程的**谱密度**。
- 如果 $\int_{-\infty}^{\infty} |B(\tau)|d\tau$ 存在有限则谱密度存在，且

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ix\tau}B(\tau)d\tau$$

平稳序列

- 当 $T = \{\dots, -1, 0, 1, \dots\}$ 或 $T = \{0, 1, 2, \dots\}$ 时宽平稳过程叫做宽平稳序列。
- 任意宽平稳序列 $\{X_n, n = \dots, -1, 0, 1, \dots\}$ 有谱展式

$$B(\tau) = \int_{-\pi}^{\pi} e^{i\tau\lambda}dF(\lambda)$$

  其中 $F(\lambda)$ 是不减的右连续函数，$F(-\pi) = 0$。
- 如果级数 $\sum_{n=-\infty}^{+\infty} |B(n)|$ 收敛，则存在非负函数 $f(\lambda)$(**谱密度**) 使得

$$\begin{aligned}F(x) =&\int_{-\pi}^{x} f(\lambda)d\lambda \quad (\forall x \in [-\pi, \pi]\\ f(\lambda) =&\frac{1}{2\pi}\sum_{k=-\infty}^{\infty} B(k)e^{-ik\lambda}\end{aligned}$$