概率论与数理统计讲义

唐宏岩

前言

讲义基于《概率论与数理统计》课程材料编写，初稿由助教曹子尧根据课堂板书笔记整理，之后由助教陈春晖完成第一轮内容补充. 本次呈现版本是在此基础上的更新，讲义将持续优化，旨在为同学们的课程学习提供相应的辅助与参考.

唐宏岩

2025 年 4 月

目录

第一部分

初等概率论

第一章　事件的概率

1.1　概率论发展简史

概率思想的萌芽最早可追溯至古埃及和古巴比伦时期，主要体现在骰子游戏和占卜等活动中. 古希腊罗马时期也有一些关于机会问题的思考，但都停留在经验层面.

在 17 世纪，帕斯卡 (Pascal) 和费马 (Fermat) 通过书信往来讨论了如何合理分配赌注的问题，通过初等数学推导出了概率计算的许多核心方法，虽然他们的研究尚未归纳成通用的定理，但这些思想为后来的数学家奠定了基础. 人们通常将此视为概率论作为一个学科分支诞生的标志.

在 18 世纪，概率论得到了进一步的发展，逐渐从零散的研究走向系统化. 雅各布·伯努利 (Jacob Bernoulli) 是这一时期的代表人物，他在其著作《猜度术》中提出了大数定律，首次建立概率与频率的数学联系，揭示了随机现象中规律性的数学本质.

19 世纪，拉普拉斯 (Laplace) 发展了一种系统化的分析方法，创立了分析概率论. 他的经典著作《概率的分析理论》给出了概率论的一些重要概念和定理，并将概率论广泛地应用于科学和社会领域.

到了 20 世纪，现代概率论的发展进入了一个全新的阶段. 柯尔莫戈罗夫 (Kolmogorov) 在 1933 年发表了《概率论的的基础》，创造性地运用测度论框架构建了概率论的公理化体系. 这一具有里程碑意义的工作使得概率论真正成为一门建立在严格数学基础之上的学科，为现代概率论的蓬勃发展奠定了坚实的理论基础，同时推动了随机过程、信息论等新领域的兴起.

1.2　随机试验与事件

定义 1.1. 概率论中的随机试验（简称试验）指的是符合下面两个特征的（真实或假想）过程：

1. 不能预先确知结果；
2. 可以预测所有可能的结果.

定义 1.2. 样本空间是指一个随机试验的所有可能结果构成的集合，常用符号 Ω 表示. 样本空间中的元素称为样本点.

样本空间的明确是对随机试验进行数学描述的第一步. 例如

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\},$$

其中每个 ω_i 都是试验的一个具体可能结果. 随机事件（简称事件）是样本空间的子集，它明确地界定了随机试验的某些可能结果. 一个事件发生，表示该事件所包含的某些元素被观测到.

需要注意的是，当样本空间中的元素个数不可数时，并不是样本空间的所有子集都可以称为随机事件. 这一限制将会在概率的公理化定义部分中进一步讨论.

定义 1.3. 事件是样本空间的一个良定义子集. 通常用大写字母（如 A、B 等）来表示事件.

1. 基本事件：基本事件是样本空间的单元素子集，即只包含一个试验结果的事件. 例如，在掷骰子的试验中，基本事件 $\{3\}$ 表示骰子结果为点数 3.
2. 全事件 Ω（必然事件）：全事件是样本空间本身，包括了试验的所有可能结果，因此它的发生是必然的.
3. 空事件 $\varnothing$（不可能事件）：空事件表示没有任何可能的试验结果，因此它的发生是不可能的.

1.3 事件的运算

在概率论中，事件本质上是集合（样本空间的子集），其运算可以通过集合运算来定义.

定义 1.4. 事件的基本运算：

1. 余 $A^c := \Omega \setminus A$，称为事件 A 的余事件（事件 A 不发生）.
2. 和 $A + B := A \cup B$（事件 A 或 B 至少有一个发生）.
3. 差 $A - B := A \setminus B$（事件 A 发生而 B 不发生）.
4. 积 $AB := A \cap B$（事件 A 和 B 同时发生）.

根据集合运算性质直接可得，$(A^c)^c = A$， $A + B = (A^c B^c)^c$，$AB = (A^c + B^c)^c$. 一般地，集合的 De Morgan 定律也同样适用于事件运算，相应的表达形式为：

$$\left(\sum_n A_n\right)^c = \prod_n A_n^c, \quad \left(\prod_n A_n\right)^c = \sum_n A_n^c.$$

定义 1.5. 两类特殊关系：

1. 当 $AB = \varnothing$ 时，称事件 A 和 B *互斥*（不能同时发生）.
2. 当 $A = B^c$ 时，称事件 A 和 B *对立*（非此即彼的关系）.

事件和集合运算的本质一致，因此也可以用 Venn 图可视化这些运算来帮助理解.

1.4 概率的几种解释

对于概率的概念，人们至今没有非常令人满意的定义，但可以从不同角度给出解释. 每种解释方式都有一定的适用场景和相应的局限，以下简单介绍三种常见的概率解释：

1. *古典解释*：基于等可能性的解释. 该解释认为，在一个随机试验中，每个可能的结果发生的机会是相等的. 例如，在掷一个均匀的六面骰子时，六个面可等同对待，所以假设每一面出现的概率相同是合理的，这种情况下概率被定义为事件发生的“有利结果”与“所有可能结果”之比. 古典解释在所有结果是等可能的前提下，提供了一种简单直观的概率计算方法.
2. *频率解释*：基于大量重复试验的解释. 频率学派认为，概率是为长期频率的稳定值，即在无限次重复独立试验的过程中，某个事件发生的频率会趋近于一个固定的数值，这个数值就是该事件的概率. 例如，在重复掷骰子的过程中，点数 6 出现的频率随着试验次数的增多会趋近于 $\frac{1}{6}$. 频率解释将概率视为一种客观规律.
3. *主观解释*：概率是一种对确信程度的度量，这也是 Bayes 学派采用的概率解释. 主观解释认为，概率反映的是我们对某个事件发生的确信程度或信心. 这种解释并不要求大量实验或等可能性假设，而是将概率视为一个个人判断的量度. 比如，在贝叶斯学派的框架下，概率代表了我们对于某一事件发生的信念强度，而这一信念可以根据新的信息进行更新.

不同的解释反映了人们对随机现象不同角度的理解. 在实际应用中，可以根据具体情况选择合适的解释方式. 例如，在统计实验中，频率解释通常更为适用，而在决策理论中，主观概率解释则显示出其独特优势，此外，在一些现代应用中人们也常采用混合框架.

1.5 概率的公理化定义

在概率论中，并非样本空间的所有子集都能作为随机事件. 为了保证概率定义的合理性且满足可测量性，子集族需要具备一定的数学结构. 用 2^Ω 表示 Ω 的所有子集构成的集合，通常称其为 Ω 的幂集.

定义 1.6. 称 $\mathscr{F} \subset 2^{\Omega}$ 为事件集类，若 $\mathscr{F}$ 满足以下条件：

1. $\Omega \in \mathscr{F}$；
2. 若 $A \in \mathscr{F}$，则 $A^c \in \mathscr{F}$（对余运算封闭）；
3. 若 $\{A_i\}_{i=1}^{\infty} \subset \mathscr{F}$，则 $\sum_{i=1}^{\infty} A_i \in \mathscr{F}$（对可列和封闭）.

由此可知，事件集类是包含样本空间 Ω 的 σ-代数，其元素称为随机事件. 根据 De Morgan 定律，事件集类对可列积也封闭，即 $\prod_{i=1}^{\infty} A_i \in \mathscr{F}$.

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

1. $\mathscr{F}_1 = 2^{\Omega}$：即样本空间的幂集，包含所有可能的子集，是最大的事件集类.
2. $\mathscr{F}_2 = \{\Omega, \varnothing\}$：最小的事件集类，只有全事件和空事件.
3. $\mathscr{F}_3 = \{\Omega, \varnothing, \{a, b\}, \{c, d\}\}$：该事件集类包含了全事件、空事件以及事件 $\{a, b\}$ 和 $\{c, d\}$，并且满足 σ-代数的封闭性条件.

借助事件集类的概念，可以定义概率函数. 概率函数为事件集类中的每个事件分配一个非负的实数值，用以描述事件发生的可能性.

定义 1.7 (Kolmogorov)**.** *概率函数* $P : \mathscr{F} \to \mathbb{R}$ 是满足以下公理的映射：

1. $P(A) \geq 0, \forall A \in \mathscr{F}$（非负性）.
2. $P(\Omega) = 1$（归一化）.
3. 若 $\{A_i\}_{i=1}^{\infty} \subset \mathscr{F}$，$A_i A_j = \varnothing, \forall i \neq j$，则 $P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$（可列可加性）.

此时，称三元组 $(\Omega, \mathscr{F}, P)$ 为一个*概率空间*.

根据公理化定义，可以推导出概率的一些基本性质，具体过程留给读者.

命题 1.1. 在概率空间 $(\Omega, \mathscr{F}, P)$ 中，有以下性质：

1. $P(A) \leq 1, \forall A \in \mathscr{F}$.
2. $P(\varnothing) = 0$.
3. $P(A) + P(A^c) = 1$.
4. 若 $\{A_i\}_{i=1}^{n} \subset \mathscr{F}$，$A_i A_j = \varnothing, \forall i \neq j$，则 $P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ （有限可加性）.
5. 若$A \subset B$，则$P(A) \leq P(B)$，称为“事件 A 蕴涵事件 B”.
6. 容斥恒等式:

$$\begin{aligned} P(A_1 + \cdots + A_n) = & \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq n} P(A_{i_1} A_{i_2}) \\ & + \cdots + (-1)^{r+1} \sum_{1 \leq i_1 < \cdots < i_r \leq n} P(A_{i_1} \cdots A_{i_r}) \\ & + \cdots + (-1)^{n+1} P(A_1 A_2 \cdots A_n). \end{aligned}$$

特别地，

$$P(A+B)=P(A)+P(B)-P(AB).$$

例 1.2 (配对问题). 有 n 个人，每人有一顶帽子. 将所有帽子收集到一起随机分配，考虑无人拿到自己帽子的概率.

分析：设事件 A_i 为“第 i 个人拿到自己帽子”，则 $P(A_i)=\dfrac{1}{n}$，而

$$\sum_{1\le i_1<\cdots<i_r\le n} P(A_{i_1}A_{i_2}\cdots A_{i_r})=\frac{(n-r)!}{n!}\binom{n}{r}=\frac{1}{r!},$$

利用容斥恒等式，至少一人拿到自己帽子的概率为

$$P(A_1+\cdots+A_n)=1-\frac{1}{2!}+\frac{1}{3!}-\frac{1}{4!}+\cdots+(-1)^{r+1}\frac{1}{r!}+\cdots+(-1)^{n+1}\frac{1}{n!},$$

进而所求概率为

$$P_n=1-P(A_1+\cdots+A_n)=1-\left(1-\frac{1}{2!}+\cdots+(-1)^{n+1}\frac{1}{n!}\right).$$

当 n 趋向无穷时，P_n 收敛于一个极限值 $\dfrac{1}{e}$.

思考：恰有 k 个人拿到自己帽子的概率是多少？

1.6 条件概率

条件概率描述了在已知某事件发生的前提下，另一个事件发生的概率. 通过条件概率的定义，可以重新评估事件在缩小的样本空间中的概率.

定义 1.8. 若 $P(B)>0$，定义条件概率为

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

条件概率 $P(A|B)$ 表示已知事件 B 发生的前提下，事件 A 发生的概率. 计算条件概率时，通常有两种方法：

1. 在缩小的样本空间中直接分析：样本空间限制在事件 B中，在此基础上考虑事件 A 的概率；
2. 利用定义公式计算：即直接使用 $P(A|B)=\dfrac{P(AB)}{P(B)}$.

利用条件概率可得到重要的概率乘法法则

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

这一法则表明，两个事件同时发生的概率可以通过一个事件的条件概率与作为条件的另一个事件的概率计算得到.

例 1.3. 掷一个均匀六面骰子，设样本空间为 $\Omega = \{1,2,3,4,5,6\}$，事件 $A = \{2,3,4,5\}$，事件 $B = \{1,3,5\}$. 易知，$P(A) = \dfrac{4}{6}$，$P(B) = \dfrac{3}{6}$，$P(AB) = \dfrac{2}{6}$. 根据条件概率的定义，计算可得：

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{2}{3}.$$

例 1.4. 一个袋子中有 8 个红球和 4 个白球，若无放回地连续取出两个球，考虑两个都是红球的概率. 记 R_i 表示第 i 个球是红球的事件（$i = 1,2$）. 利用组合数计算：

$$P(R_1R_2) = \frac{\binom{8}{2}}{\binom{12}{2}}.$$

还利用条件概率计算：

$$P(R_1R_2) = P(R_1) \cdot P(R_2|R_1) = \frac{8}{12} \times \frac{7}{11}.$$

两种方法的计算结果一致，条件概率方法通过分步分析，简化了计算过程.

对于多个事件乘积的概率，可以利用条件概率的乘法法则递归计算：

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

例 1.5 (配对问题续). 有 n 个人，每人有一顶帽子. 将所有帽子收集到一起随机分配. 事件 $A_{i_1}, A_{i_2}, \ldots, A_{i_r}$ 分别表示第 $i_1, i_2, \ldots, i_r$ 个人都拿到自己的帽子. 根据条件概率的乘法法则，可以计算乘积事件的概率：

$$\begin{aligned} P(A_{i_1}A_{i_2} \cdots A_{i_r}) &= P(A_{i_1}) \cdot P(A_{i_2}|A_{i_1}) \cdot \cdots \cdot P(A_{i_r}|A_{i_1} \cdots A_{i_{r-1}}) \\ &= \frac{1}{n} \cdot \frac{1}{n-1} \cdot \cdots \cdot \frac{1}{n-(r-1)} \\ &= \frac{(n-r)!}{n!}. \end{aligned}$$

需要注意的是，前面定义的条件概率确实是概率，即给定一个事件 B（$P(B) > 0$），条件概率 $P(\cdot|B)$ 定义了一个新的概率函数.

命题 1.2. 给定事件 B（$P(B) > 0$），条件概率 $P(\cdot|B) : \mathscr{F} \to \mathbb{R}$ 是一个概率函数，从而 $(\Omega, \mathscr{F}, P(\cdot|B))$ 是一个概率空间.

对于上述命题的证明，只需验证 $P(\cdot|B)$ 满足概率定义的三条公理即可，过程留给读者.

如果将 $P(A)$ 称为事件 A 的“先验概率”，则 $P(A|B)$ 是在观察到事件 B 发生后重新评估的“后验概率”. 注意，观测到事件 A 已经发生后，并不能说 $P(A)$ 变为 1，而是后验概率 $P(A|A) = 1$.

1.7 事件的独立性

定义 1.9. 若 $P(AB) = P(A)P(B)$，则称事件 A 和 B 相互独立.

从条件概率的角度也可以理解独立性. 如果 $P(B) > 0$，则事件 A 和 B 相互独立等价于条件概率 $P(A|B) = P(A)$. 换句话说，在已知事件 B 发生的情况下，事件 B 的发生与否对事件 A 的发生概率没有影响.

下面例子中事件之间的独立性不容易直接观察到，这时可以利用独立性的定义来加以验证.

例 1.6. 假设 A_i 表示掷两个骰子的点数之和为 i 的倍数，其中 $i = 2, 3, 5$. 我们需要判断事件 A_2 与 A_3 以及事件 A_2 与 A_5 是否独立，并给出理由.
通过计算概率，可以得出以下结果：

$$P(A_2) = \frac{1}{2}, \quad P(A_3) = \frac{1}{3}, \quad P(A_5) = \frac{7}{36},$$

$$P(A_2A_3) = \frac{1}{6}, \quad P(A_2A_5) = \frac{1}{12}.$$

由独立性定义可知，A_2 与 A_3 独立，而 A_2 与 A_5 不独立.

命题 1.3. 若事件 A 和 B 相互独立，则事件 A^c 和 B 也相互独立，其中 A^c 为事件 A 的余事件.

当涉及三个或更多事件时，独立性的定义需要更加严格的条件.

定义 1.10. 若 $P(ABC) = P(A)P(B)P(C)$，且事件 A、B 和 C 两两独立，则称事件 A、B 和 C 独立.

需要注意的是，仅仅满足两两独立并不意味着三者相互独立，独立性还要求三个事件的联合概率等于它们各自概率的乘积. 下面的例子说明了这一点.

例 1.7. 一个均匀四面体的其中三个面分别标记数字 $1, 2, 3$，另外一个面同时标记数字 $1, 2, 3$. 令 A_i 表示掷四面体时向下的面含有数字 i（$i = 1, 2, 3$）. 直接验证可知事件 A_1, A_2, A_3 不独立但两两独立.

类似地，多个事件同样有独立性的刻画.

定义 1.11. 若对于一列（至多可数多个）事件 $\{A_i\}$，任意取其中有限个不同事件 $A_{i_1}, A_{i_2}, \ldots, A_{i_r}$，都有 $P(A_{i_1}A_{i_2}\cdots A_{i_r}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_r})$，则称事件列 $\{A_i\}$ 相互独立.

例 1.8. 假设每周开奖的彩票中奖率为 10^{-5}，奖金不累积且各次独立开奖. 求连续十年（520 周）不中奖的概率.
令事件 A_i 表示第 i 周不中奖，则每周不中奖的概率为 $P(A_i) = 1 - 10^{-5}$. 根据事件的独立性，连续 520 周不中奖的概率为

$$P(A_1 \cdots A_{520}) = (1 - 10^{-5})^{520} \approx 0.9948.$$

定义 1.12. 若事件 A, B, E 满足 $P(AB|E) = P(A|E)P(B|E)$，则称事件 A 和 B 在条件 E 下是条件独立的.

值得注意的是，条件独立性和事件间的独立性没有蕴涵关系. 条件独立性仅在给定条件 E 的前提下成立，而普通的独立性则不依赖于任何条件事件.

1.8 Bayes 公式

定理 1.1 (全概率公式)**.** 设 $\{B_i\}$ 是样本空间 Ω 的一个分割，即满足以下条件：

1. $\sum_i B_i = \Omega$，即所有事件 B_i 的和事件为样本空间 Ω.
2. $B_iB_j = \varnothing, \forall i \neq j$，即各事件 B_i 两两互斥.
3. $P(B_i) > 0, \forall i$，即每个事件 B_i 的概率为正.

则

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

证明. 根据分割的含义，利用加法公理和乘法法则，可得

$$P(A) = P(\sum_i (AB_i)) = \sum_i P(AB_i) = \sum_i P(A|B_i)P(B_i).$$

□

全概率公式为复杂事件的概率计算提供了一种分解策略，通过将事件概率分解为条件概率的加权和，有效简化了计算过程，使其在实际应用中更易处理.

例 1.9 (敏感问题调查). 在调查问卷中，受访者可能对敏感问题（如“你是否有某病史”）有所顾虑，从而给出虚假回答。为保护隐私并取得信任，可以引入一个不具有敏感性的“保护性问题”（如“你的电话号码尾号是否为偶数”），并让受访者通过抛硬币择一问题回答（如约定掷得正面则回答敏感问题，掷得反面回答保护性问题），从而无需担心隐私暴露.
假设人群中敏感问题答案为“是”的比例为 p（未知），保护性问题答案为“是”的比例为 q（已知）. 若从 n 个被调查者中收集到 k 个“是”的答案，则根据全概率公式，有

$$\text{回答“是”的理论比例} = \frac{1}{2}p + \frac{1}{2}q \approx \frac{k}{n}.$$

因此，

$$p \approx 2\frac{k}{n} - q.$$

得到敏感问题答案为“是”的比例估计值.

接下来讨论概率论中一个非常重要的定理——Bayes 公式.

定理 1.2 (Bayes 公式). 设 $\{B_i\}$ 是样本空间 Ω 的一个分割，则对任意事件 A（其中 $P(A) > 0$），有

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}, \quad i = 1, 2, \ldots, n.$$

Bayes 公式提供了一种动态更新概率的机制，在医学诊断、机器学习等领域有广泛应用，为基于数据的科学决策提供了理论支持.

例 1.10 (假阳性悖论). 考虑某种罕见病的检测. 设事件 A 表示检测呈阳性，B 表示实际患病. 已知

$$P(B) = 10^{-4}, \quad P(A|B) = 0.99, \quad P(A|B^c) = 10^{-3}.$$

由 Bayes 公式得到检测呈阳性时实际患病的概率为

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \approx 9\%.$$

该结果表明，在给定的假设下，即使检测结果为阳性，实际患病概率的绝对数值仍然较低，不足以断言其患病. 但该人患病概率的增幅显著，值得引起关注.

如果该人接受第二次检测，且结果仍为阳性，假设两次检测相互独立且检测方法相同，则再次检测仍为阳性后，其患病的概率变为

$$\begin{aligned} P(B|A_1A_2) &= \frac{P(A_1A_2|B)P(B)}{P(A_1A_2|B)P(B) + P(A_1A_2|B^c)P(B^c)} \\ &= \frac{P(A_1|B)P(A_2|B)P(B)}{P(A_1|B)P(A_2|B)P(B) + P(A_1|B^c)P(A_2|B^c)P(B^c)} \\ &\approx 99\%. \end{aligned}$$

实践中，这个结果已可作为确诊依据.

第二章　随机变量

2.1　一维随机变量

定义 2.1.（一维）随机变量是样本空间上的实值函数.

上述定义是初等概率论教科书中常见的形式，强调了随机变量是实值函数的本质. 事实上，并非样本空间上所有的实值函数都可以称为随机变量.

对于实值函数 X 和 $\mathbb{R}$ 的可测子集 I（例如 $I=(a,b]$），定义其原像集为

$$X^{-1}(I):=\{\omega\in\Omega\mid X(\omega)\in I\}\subset\Omega.$$

基于此，随机变量更严谨的定义如下.

定义 2.2. [随机变量] 设 $(\Omega,\mathscr{F})$ 是可测空间，$X:\Omega\to\mathbb{R}$ 是一个函数. 若对任意可测集 $I\subset\mathbb{R}$，都有 $X^{-1}(I)\in\mathscr{F}$，则称 X 是 $(\Omega,\mathscr{F})$ 上的随机变量.

这里“可测空间”是指 $\mathscr{F}$ 是样本空间 Ω 上的 σ-代数. 注意到，随机变量的定义不依赖概率测度 P 的存在，因此此处并不要求“概率空间”.

事实上，由测度论知识可知，满足随机变量的严谨定义等价于要求 $\forall x\in\mathbb{R},\{\omega|X(\omega)\leq x\}\in\mathscr{F}$.

例 2.1. 以下是两个随机变量的例子.

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

在实际应用中经常用如“$X_1=20$”，“$X_2>100000$”等简化的记号来表示与随机变量相关的事件. 例如，“$X_1=20$”实际上表示事件 $\{\omega\in\Omega_1|X_1(\omega)=20\}$，即“恰好有20人支持”的所有可能结果构成的集合.

随机变量是试验结果的数值摘要，起到概括的作用. 随机变量的“随机”要素来自于样本点 $\omega \in \Omega$ 的随机选择. 在实际应用中，随机变量常常比样本空间更直观.

随机变量可以分类为：

1. 离散型：取值至多可数多个.
2. 连续型：取值为区间型（非严格定义）.
3. 其他：例如混合型随机变量，其同时具有离散和连续的特性.

定义 2.3. 记号 $P(X \in I)$ 表示“X 的取值在 I 中的概率”，其值为 $P(X^{-1}(I))$，即

$$P(X \in I) := P(\{\omega \in \Omega \mid X(\omega) \in I\}).$$

例如，随机变量 X 取值在区间 $(a, b]$ 内的概率可以表示为

$$P(a < X \leq b) = P(\{\omega \mid X(\omega) \in (a, b]\}).$$

定义 2.4 (分布函数). 对于随机变量 X，其累积分布函数（Cumulative Distribution Function，简记为 CDF）定义为

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

当没有歧义时，可以省略下标 X，简写为 $F(x)$.

可以通过 CDF 计算某个区间内随机变量的概率，例如：

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

这表明，CDF 的差值可以直接给出一个半开闭区间内随机变量取值的概率.

例 2.2. 设 X 表示掷两个均匀六面骰子所得的点数之和，X 的分布如下表所示：

X	2	3	4	5	6	7	8	9	10	11	12
P	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

对应的 CDF 图像见图 2.1，可以直观地看到该分布函数是一个阶梯函数.

注：由于绘图软件的限制，各个阶跃点的绘制方式可能存在不规范的情况，例如 $F(3) = \frac{3}{36}$ 而非 $\frac{1}{36}$. 另外，对于 $x < 2$，$F(x) = 0$；对于 $x \geq 12$，$F(x) = 1$.

命题 2.1. 累积分布函数（CDF）具有以下基本性质：

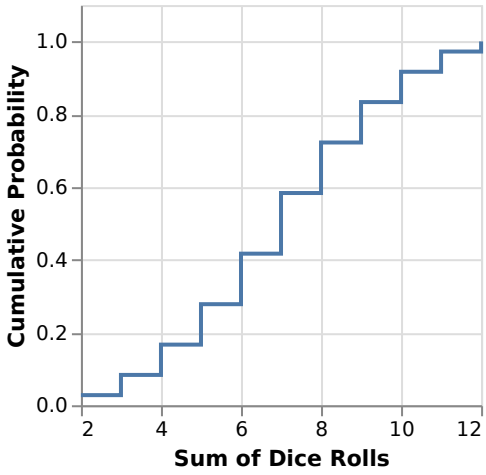

图 2.1: 例 2.2 中 X 的 CDF 图像

1. $F_X(x)$ 是单调递增的（但不一定严格单调递增）.
2. $\lim\limits_{x\to+\infty} F_X(x) = 1, \quad \lim\limits_{x\to-\infty} F_X(x) = 0.$
3. $F_X(x)$ 是右连续的.

事实上，可以证明这些性质是一个函数 $F : \mathbb{R} \to \mathbb{R}$ 成为累积分布函数的充要条件.

命题 2.2. 若 X 和 Y 是随机变量，则 $aX + bY$, XY, $\dfrac{X}{Y}$（当 $Y \neq 0$ 时）仍然是随机变量，这里 a, b 为常数. 一般来说，如果 g 是一个可测函数，则 $g(X, Y)$ 也是随机变量.

定义 2.5. 若对任意可测集 $I \subset \mathbb{R}$，都有

$$P(X_1 \in I) = P(X_2 \in I).$$

则称随机变量 X_1 和 X_2 具有相同分布.

根据测度论知识，可以得到更为方便使用的随机变量同分布的判定方法.

命题 2.3. 设随机变量 X_1 和 X_2 的 CDF 分别为 F_1 和 F_2，则 X_1 和 X_2 具有相同分布的充要条件是：对所有的 $x \in \mathbb{R}$，都有 $F_1(x) = F_2(x)$.

需要注意的是，同分布并不等同于“同变量”. 即使两个随机变量同分布，作为函数它们也不一定相同.

例 2.3. 在掷硬币的例子中，设 X 表示正面向上次数，Y 表示反面向上次数. 显然，X 和 Y 是同分布的，但它们的取值并不相等，因为 X 只能取 0 或 1，而 Y 也只能取 0 或 1，并且 $X + Y = 1$.

2.2 离散随机变量

定义 2.6. 离散随机变量 X 的*概率质量函数*（Probability Mass Function, 简记为 PMF）f 定义为

$$f(x) := P(X = x), \quad \forall x \in \mathbb{R}.$$

还可以通过（形如例 2.2 中的）*分布表*展示一个离散随机变量的所有可能取值及其对应的概率（通常称为离散随机变量的*分布律*）.

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

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

定义 2.7. 离散随机变量 X 的（数学）期望定义为

$$\mathrm{E}(X) := \sum_i x_i f(x_i).$$

称 X 的期望*存在*，当且仅当 $\sum_i |x_i| f(x_i) < +\infty$.

当期望存在时，X 的*方差*定义为

$$\mathrm{Var}(X) := \sum_i (x_i - \mathrm{E}(X))^2 f(x_i).$$

当方差有限时，称其算术平方根为 X 的*标准差*，记作 $\mathrm{SD}(X)$.

通常也将一个随机变量的数学期望 $\mathrm{E}(X)$ 称为该随机变量的*均值*（mean）. 利用均值和标准差可以将随机变量进行*标准化*，即对随机变量 X 进行线性变换：

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

其中，$\mu = \mathrm{E}(X)$，$\sigma = \mathrm{SD}(X)$. 标准化后的随机变量 Z 无量纲，且具有均值 0 和标准差 1.

对于可测函数 g，$g(X)$ 也是随机变量，可以验证其期望为

$$\mathrm{E}(g(X)) = \sum_i g(x_i) f(x_i).$$

特别地，有

$$\mathrm{Var}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right) = \mathrm{E}(X^2) - \mathrm{E}^2(X).$$

期望反映了随机变量取值的集中趋势，而方差或者标准差则反映了随机变量取值的分散程度. 标准差与随机变量本身具有相同的量纲，使其在实际应用中更容易直观理解.

2.3 常见离散分布

定义 2.8. 若随机变量 X 的取值集合为 $\{0,1\}$，且存在 $p \in (0,1)$，使得 $P(X=1)=p$，$P(X=0)=1-p$，则称 X 服从 *Bernoulli* 分布，记作 $X \sim B(p)$.

在此定义中，p 被称为 Bernoulli 分布的参数. 通常将 X 的两个取值分别对应试验“成功”和“失败”，例如 $X=1$ 表示成功，$X=0$ 表示失败，则试验“成功”的概率为 p，而试验“失败”的概率为 $1-p$. 这样的试验通常被称为 *Bernoulli* 试验.

直接计算可得， Bernoulli 分布的期望和方差分别为

$$\mathrm{E}(X)=p, \quad \mathrm{Var}(X)=p(1-p).$$

定义 2.9. 若随机变量 X 的取值集合为 $\{0,1,\ldots,n\}$ （$n \in \mathbb{N}^*$），且存在 $p \in (0,1)$ 使得

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

则称 X 服从二项分布，记作 $X \sim B(n,p)$.

这里可以将 X 理解为“n 次独立 Bernoulli 试验成功的次数”. 其中 p 是每次试验的成功概率. 而 $B(1,p)$ 即为 Bernoulli 分布 $B(p)$. 若 $X \sim B(n,p)$，则该随机变量的期望和方差分别为

$$\mathrm{E}(X)=np, \quad \mathrm{Var}(X)=np(1-p).$$

例 2.4. 考虑观察时间区间 $(0,1]$ 内某路口的交通事故数 X，将该区间等分为 n 个小区间，即有每段小区间

$$l_i=(\frac{i-1}{n},\frac{i}{n}], \ i=1,2,\ldots,n.$$

随着 n 趋向无穷大，每个小区间的长度趋于零. 假设以下条件成立：

1. 每个小区间内，至多发生一次事故；
2. 每个小区间内发生一次事故的概率与区间长度（即 $\frac{1}{n}$）成正比，记为 $p=\frac{\lambda}{n}$；
3. 各个区间内是否发生事故是相互独立的.

可以看到，当 n 充分大时，这些假设是合理的. 在这些条件下，有 $X \sim B(n,p)$，且

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \to \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{当 } n \to \infty.$$

定义 2.10. 若随机变量 X 的取值集合为 $\mathbb{N}$，且存在 $\lambda>0$ 使得 $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$, $k \in \mathbb{N}$，则称 X 服从 *Poisson* 分布，记作 $X \sim P(\lambda)$.

若 $X \sim P(\lambda)$，则该随机变量的期望和方差为

$$\mathrm{E}(X) = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

Poisson 分布通常用于描述一定时间或空间内小概率事件的发生次数，其中 λ 是此时间（或空间）内该事件的平均发生次数.

例 2.5. 设某医院平均每天出生婴儿数为 λ，则接下来 t 天内出生婴儿数服从参数为 λt 的 Poisson 分布.

例 2.6. 假设一只昆虫的产卵次数服从参数为 λ 的 Poisson 分布，而虫卵能发育成虫的概率为 p（$0 < p < 1$）. 假设每个虫卵是否能够发育成虫是独立的. 可以证明，有 k 个后代的概率由参数为 λp 的 Poisson 分布确定.

记 X 为产卵个数，Y 为发育成虫的个数. 则有

$$\begin{aligned} P(Y=k) &= \sum_{n=k}^{\infty} P(Y=k|X=n)P(X=n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \frac{(\lambda p)^k}{k!} e^{-\lambda p}. \end{aligned}$$

因此， $Y \sim P(\lambda p)$.

通过例 2.4 的讨论可见，对于二项分布 $X \sim B(n,p)$，X 可以近似服从参数为 $\lambda = np$ 的 Poisson 分布，记作 $X \overset{\text{近似}}{\sim} P(np)$. 当 p 很小，n 很大，且 np 不太大时近似效果比较好. 近似误差的大小不超过 $\min\{p, np^2\}$.（证明复杂，此处略）

当 n 次 Bernoulli 试验之间不完全独立，但满足弱相依条件时，Poisson 分布仍然可以作为一种较好的近似.

例 2.7 (配对问题续). 设 A_i 表示第 i 个人拿到自己的帽子，则 $P(A_i) = \dfrac{1}{n}$，而 $P(A_i|A_j) = \dfrac{1}{n-1}$（当 $j \neq i$ 时）. 当 n 很大时，$\dfrac{1}{n}$ 和 $\dfrac{1}{n-1}$ 非常接近，可以认为满足所谓的弱相依条件. 记 X 为拿到自己帽子的人数，则 X 近似服从 Poisson 分布，其参数为 $\lambda = np = n \cdot \dfrac{1}{n} = 1$，即

$$P(X=k) \approx \frac{e^{-1}}{k!}.$$

为了验证这种近似的合理性，采用常规方法重新讨论. 考虑指定的某 k 个人，令事件 E 表示这 k 个人都拿到自己的帽子，事件 F 表示其余 $(n-k)$ 个人都没有拿到自己的帽子. 则有

$$P(EF) = P(E)P(F|E) = \frac{(n-k)!}{n!} \cdot P_{n-k}\,,$$

其中，P_{n-k} 表示当总人数为 $(n-k)$ 时无人拿到自己帽子的概率. 由此，

$$P(X=k)=\binom{n}{k}P(EF)=\frac{1}{k!}P_{n-k}\to\frac{e^{-1}}{k!},\quad \text{当 } n\to\infty.$$

这说明前述近似是合理的.

2.4 连续随机变量

定义 2.11. 对随机变量 X，若存在函数 $f:\mathbb{R}\to[0,+\infty)$，使得对于任意可测集 $I\subset\mathbb{R}$，都有

$$P(X\in I)=\int_I f(x)\,\mathrm{d}x,$$

则称 X 为 *连续型随机变量*，$f(x)$ 为 X 的*概率密度函数*（Probability Density Function，简记为 PDF）.

命题 2.5. 连续随机变量 X 的概率密度函数 f 满足以下性质：

1. $\displaystyle\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x=1.$
2. 对于任意实数 a 和 b，有

   $$P(a<X\le b)=\int_a^b f(x)\,\mathrm{d}x=P(a\le X\le b)=P(a\le X<b)=P(a<X<b).$$

   这表明连续型随机变量在区间中各类端点的定义对概率无影响.
3. 对于任意的实数 a，有 $P(X=a)=0$. 表明连续型随机变量在某一特定取值的概率总是零.
4. 如果 $f(x)$ 在某点 x_0 连续，，则对于一个小区间 $(x_0-\delta,x_0+\delta)$，有

   $$P(x_0-\delta<X<x_0+\delta)=\int_{x_0-\delta}^{x_0+\delta}f(t)\,\mathrm{d}t\approx f(x_0)\cdot 2\delta.$$

   这意味着随机变量 X 在小区间内的概率与区间长度近似成正比，比例系数即为区间中心处的密度值 $f(x_0)$.
5. 累积分布函数（CDF）$F(x)$ 可以表示为

   $$F(x)=\int_{-\infty}^{x}f(t)\,\mathrm{d}t,$$

   从而 $F(x)$ 是关于 x 的连续函数；如果 f 在 x 处连续，则有 $F'(x)=f(x)$.
6. 概率密度函数不唯一. 若随机变量的概率密度函数存在，则在一个零测集上对其值进行任意修改，所得到的新函数仍是一个合法的概率密度函数.

与离散型随机变量类似，可以在连续型情形定义随机变量的数学期望（均值）和方差.

定义 2.12. 连续随机变量 X 的期望定义为

$$\mathrm{E}(X) := \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x.$$

称 X 的期望存在，当且仅当

$$\int_{-\infty}^{+\infty} |x| f(x)\,\mathrm{d}x < +\infty.$$

当期望存在时，X 的方差定义为

$$\mathrm{Var}(X) := \int_{-\infty}^{+\infty} (x-\mathrm{E}(X))^2 f(x)\,\mathrm{d}x = \mathrm{E}((X-\mathrm{E}(X))^2) = \mathrm{E}(X^2) - \mathrm{E}^2(X).$$

当方差有限时，称其算术平方根为 X 的标准差，记作 $\mathrm{SD}(X)$.

此外，对于任意可测函数 g，$g(X)$ 也是随机变量，可以验证其期望为

$$\mathrm{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x)\,\mathrm{d}x.$$

同样地，有

$$\mathrm{Var}(X) = \mathrm{E}\left((X-\mathrm{E}(X))^2\right) = \mathrm{E}(X^2) - \mathrm{E}^2(X).$$

随机变量的期望可以直观理解为随机变量取值的中心位置，而方差则衡量了随机变量的取值如何围绕期望值分布. 在物理模型中，期望与分布的重心位置一致，方差则对应系统的惯性特性.

2.5 常见连续分布

定义 2.13. 若一个连续型随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a,b), \\ 0, & \text{其他}, \end{cases}$$

则称 X 服从区间 (a,b) 上的均匀分布，记作 $X \sim U(a,b)$.

在应用中，$X \sim U(0,1)$ 被称为随机数，在计算机模拟中经常使用.

简单计算可得，若 $X \sim U(a,b)$，则该随机变量的期望和方差分别为

$$\mathrm{E}(X) = \frac{a+b}{2}, \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

定义 2.14. 若一个连续型随机变量的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty,$$

则称 X 服从正态分布，记作 $X \sim N(\mu, \sigma^2)$，其中 μ 和$\sigma^2 > 0$ 为参数.

正态分布是最重要的一类连续型分布，广泛应用于自然和社会科学等众多领域. 直接计算可得，若 $X \sim N(\mu, \sigma^2)$，则其期望和方差分别为

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

图 2.2 展示了正态分布的“经验法则”：约 68% 的数据落在均值 μ 的正负一个标准差 σ 范围内，95% 的数据落在正负两个标准差范围内，99.7% 的数据落在正负三个标准差范围内.

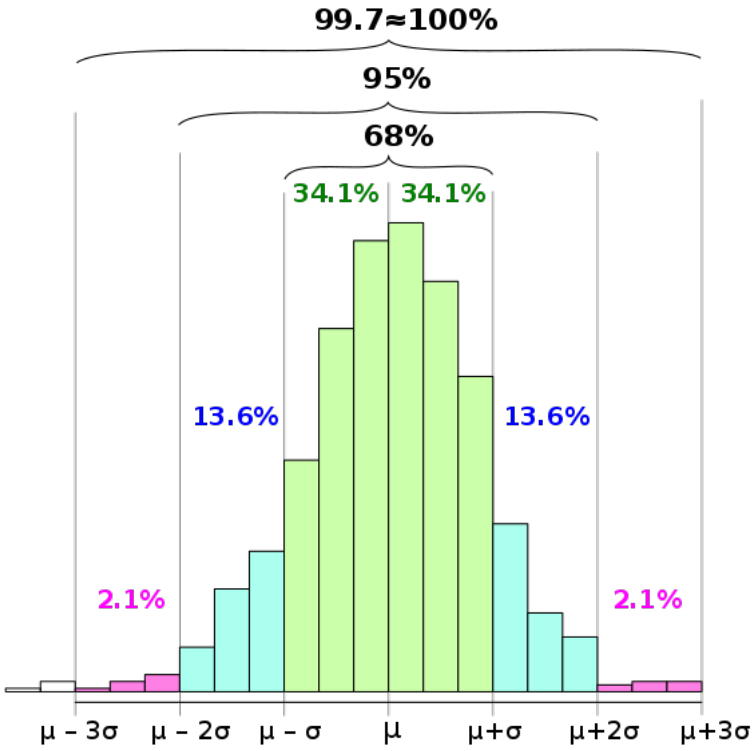

图 2.2: 经验法则

易见，$X \sim N(\mu, \sigma^2)$ 的充要条件是 $Y = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$. 其中 $N(0, 1)$ 称为标准正态分布.

定义 2.15. 若一个连续型随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

则称 X 服从指数分布，记作 $X \sim \mathrm{Exp}(\lambda)$.

指数分布常用于描述等待时间、寿命等随机过程，特别是在可靠性分析和排队理论中有重要应用.

若 $X \sim \mathrm{Exp}(\lambda)$，则其期望和方差分别为

$$\mathrm{E}(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

需要注意的是，指数分布有另一种常见符号约定，以 $\beta = \dfrac{1}{\lambda}$ 为参数，某些数学软件可能采用这种方式.

指数分布的累积分布函数为

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

其尾概率为

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

例 2.8. 设某医院每天出生婴儿的平均数量为 λ，现在观察到一名婴儿出生. 则接下来 t 天内再有婴儿出生的概率为 $P(X \le t)$，其中 X 表示到下一个婴儿出生所需的时间.
记 $N(t)$ 为 t 天内出生婴儿的数量，由例 2.5 知道 $N(t) \sim P(t\lambda)$，因此，$P(X > t) = P(N(t) = 0) = e^{-\lambda t}$. 故 $P(X \le t) = 1 - e^{-\lambda t}$，这表明 X 服从参数为 λ 的指数分布.

指数分布还可以通过失效率（或危险率）的角度来理解. 假设 X 表示某种零件的寿命，其累积分布函数为 $F(x)$，且 $F(0) = 0$. 考虑条件概率 $P(x < X < x + \mathrm{d}x \mid X > x)$，即零件寿命大于 x 时，它在 $(x, x + \mathrm{d}x)$ 时间内失效的概率. 根据条件概率公式有

$$P(x < X < x + \mathrm{d}x \mid X > x) = \frac{F(x + \mathrm{d}x) - F(x)}{1 - F(x)} \approx \frac{F'(x)}{1 - F(x)}\,\mathrm{d}x.$$

这里 $\dfrac{F'(x)}{1 - F(x)}$ 是一个条件概率密度，称为瞬时失效率，记其为 $\lambda(x)$，它描述了零件“年龄”为 x 时无法继续工作的瞬时失效率.

若假设零件的瞬时失效率是常数，即 $\lambda(x) \equiv \lambda$，不随时间变化（无老化），则其累积分布函数为

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

此即表明零件寿命 X 服从参数为 λ 的指数分布.

另一方面，指数分布具有一个重要的性质——无记忆性. 具体来说，若 X 服从指数分布，则对于任意的 $t, s > 0$，有

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

这个条件概率与条件中的 s 无关. 可以证明，指数分布是唯一满足无记忆性的连续型随机变量. 显然，很多实际应用中“无老化”假设并不合理. 为了改进这一点，可以尝试考虑更一般

的瞬时失效率. 例如，若

$$\lambda(x)=\frac{\alpha x^{\alpha-1}}{\beta^{\alpha}},\quad x>0,$$

则计算可得此时对应的分布函数为

$$F(x)=1-e^{-(\frac{x}{\beta})^{\alpha}},\quad x>0.$$

这个分布称为 Weibull 分布. 当 $\alpha=1$ 时，Weibull 分布退化为指数分布.

可以看到，已介绍的分布都含有某些参数，这些分布参数可以按照其在概率分布中的作用主要分为以下几类：

1. 位置参数：描述分布在数轴上的位置，决定分布的中心位置，如正态分布中的均值 μ；
2. 尺度参数：衡量分布的分散程度或变化幅度，用于描述分布的“散布”或“宽度”，如正态分布中的标准差 σ 和 Weibull 分布中的参数 β；
3. 形状参数：描述分布的特性，决定分布曲线的形状特征，如 Weibull 分布中的参数 α.

2.6 随机变量的函数

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

例 2.9. 设 $X\sim\mathrm{Exp}(\lambda)$，$Y=\begin{cases}0, & X\le t_0,\\ 1, & X>t_0,\end{cases}$ 其中 $t_0>0$ 为常数，则 $Y\sim B(e^{-\lambda t_0})$.

例 2.10. 设 X 为连续型随机变量，其概率密度函数为 $f(x)$，考虑 $Y=X^2$.
直接计算可得，$\forall y>0$，

$$P(Y\le y)=P(X^2\le y)=P(-\sqrt{y}\le X\le\sqrt{y})=\int_{-\sqrt{y}}^{\sqrt{y}}f(x)\mathrm{d}x.$$

因此 Y 的概率密度函数为

$$l(y)=\frac{\mathrm{d}}{\mathrm{d}y}P(Y\le y)=\frac{1}{2\sqrt{y}}(f(\sqrt{y})+f(-\sqrt{y})),\quad y>0.$$

特别地，若 $X\sim N(0,1)$，称 Y 服从自由度为 1 的 χ^2-分布，记作 $Y\sim\chi^2(1)$.

若 $Y=g(X)$ 为随机变量，可以计算 Y 的分布如下：

(1) 对于离散型随机变量 Y，$P(Y=y)=P(g(X)=y)=P(X\in g^{-1}(y))$.

(2) 对于连续型随机变量 Y，$P(Y\le y)=P(g(X)\le y)=P(X\in g^{-1}((-\infty,y]))$.

例 2.11 (对数正态分布). 设随机变量 $Y>0$，$\log Y$ 服从正态分布 $N(\mu,\sigma^2)$, 则 Y 的概率密度函数为

$$f_Y(y)=f_X(\log y)\left|\frac{\mathrm{d}}{\mathrm{d}y}\log y\right|=\frac{1}{\sqrt{2\pi}\sigma y}\exp\left(-\frac{(\log y-\mu)^2}{2\sigma^2}\right),\ y>0.$$

例 2.12. 设随机变量 X 的累积分布函数 $F(x)$ 连续且严格单调.

(1) （概率积分变换）求 $Y=F(X)$ 的分布.

(2) 证明：若 $Y\sim U(0,1)$，则 $F^{-1}(Y)$ 的累积分布函数为 $F(x)$，这里 F^{-1} 为 F 的反函数.

根据假设直接可得

(1) $F(y)=P(Y\le y)=P(F(X)\le y)=P(X\le F^{-1}(y))=F(F^{-1}(y))=y$，即 $Y\sim U(0,1)$.

(2) 令 $X=F^{-1}(Y)$，则 $F_X(x)=P(F^{-1}(Y)\le x)=P(Y\le F(x))=F(x)$.

第三章　联合分布

3.1　随机向量

定义 3.1. 设 $(\Omega, \mathscr{F}, P)$ 为概率空间，若 $\{X_i\}_{i=1}^n$ 均为该空间上的随机变量，则称 $(X_1, X_2, \ldots, X_n) : \Omega \to \mathbb{R}^n$ 为（n 维）随机向量.

随机向量多元统计分析的核心工具，广泛应用于多变量数据的建模与分析，是例如多维正态分布、多元线性回归以及贝叶斯分析模型等.

定义 3.2. n 维随机向量的（联合）累积分布函数（CDF）定义为

$$F(x_1, \ldots, x_n) := P(X_1 \le x_1, \cdots, X_n \le x_n), \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

对于 $n = 2$ 的情况，即二元分布，通常用随机向量 (X, Y) 表示，其对应的累积分布函数为

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

例 3.1.

$$P(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1).$$

联合分布函数具有如下重要性质：

1. 单调性：对每个分量 x_i 均单调非递减.
2. 右连续性：对每个分量 x_i 均右连续.
3. 边界条件：

$$\lim_{x_1, \ldots, x_n \to -\infty} F(x_1, \ldots, x_n) = 0,$$

$$\lim_{x_1, \ldots, x_n \to +\infty} F(x_1, \ldots, x_n) = 1.$$

这些性质对于刻画随机向量的分布特性及其分量之间的依赖关系具有重要意义.

3.2 离散分布

定义 3.3. 若 n 维随机向量 $(X_1,\dots,X_n)$ 的分量 $\{X_i\}_{i=1}^n$ 均为离散随机变量，则称 $(X_1,\dots,X_n)$ 是*离散随机向量*. 离散随机向量 $(X_1,\dots,X_n)$ 的（联合）*概率质量函数*（PMF）定义为

$$f(x_1,\dots,x_n) := P(X_1=x_1,\dots,X_n=x_n),\quad \forall (x_1,\dots,x_n)\in\mathbb{R}^n.$$

换言之，离散随机向量是由多个离散随机变量组成的向量，概率质量函数 $f(x_1,\dots,x_n)$ 刻画了随机向量取值为 $(x_1,x_2,\dots,x_n)$ 的概率. 离散随机向量在实际应用中常用于描述多维离散事件的组合. 例如，在问卷调查中，多位受访者的选择可以构成一个离散随机向量，用以描述所有受访者选择的联合概率分布.

命题 3.1. 离散随机向量 $(X_1,\dots,X_n)$ 的概率质量函数具有如下性质：

1. 非负性：$f(x_1,\dots,x_n)\geq 0,\quad \forall (x_1,\dots,x_n)\in\mathbb{R}^n$.
2. 归一性：$\sum f(x_1,\dots,x_n)\equiv 1$.

归一性中的求和范围是至多可数的，这是因为多个离散随机变量的笛卡尔积仍然是至多可数集.

例 3.2. 设 $\{B_i\}_{i=1}^n$ 为样本空间 Ω 的一个分割（分割的定义见 1.8 节），$P(B_i)=p_i>0,\forall i\in\{1,\dots,n\}$. 进行 N 次独立试验，随机变量 X_i 表示结果 B_i 的发生次数（$i\in\{1,\dots,n\}$）. 则随机向量 $(X_1,\dots,X_n)$ 的联合概率质量函数为

$$P(X_1=k_1,\dots,X_n=k_n)=\binom{N}{k_1,\dots,k_n}p_1^{k_1}\cdots p_n^{k_n},$$

其中 $k_1+k_2+\cdots+k_n=N$，$k_i\geq 0$ 为非负整数，$\binom{N}{k_1,\dots,k_n}=\dfrac{N!}{k_1!\cdots k_n!}$ 为多项式 $(a_1+\cdots+a_n)^N$ 展开式中项 $a_1^{k_1}\cdots a_n^{k_n}$ 的系数，称为多项系数. 称随机向量 $(X_1,\dots,X_n)$ 服从*多项分布*，记作

$$(X_1,\dots,X_n)\sim M(N;p_1,\dots,p_n).$$

多项分布是离散随机向量的重要实例，用于描述 N 次独立试验中各事件（对应于分割 $B_1,B_2,\dots,B_n$）的发生次数. 例如，在 N 次投掷骰子试验中，记录点数（1 到 6）出现的次数，随机向量 $(X_1,\dots,X_6)$ 服从多项分布 $M(N;\frac{1}{6},\dots,\frac{1}{6})$. 多项系数 $\binom{N}{k_1,\dots,k_n}$ 刻画了试验中各可能结果的组合数，反映了离散事件的联合特性.

3.3 连续分布

定义 3.4. 对于 n 维随机向量 $(X_1, \ldots, X_n)$，若存在一个函数 $f : \mathbb{R}^n \to [0, +\infty)$，使得对任意可测集 $Q \subset \mathbb{R}^n$，均有

$$P((X_1, \ldots, X_n) \in Q) = \int_Q f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n,$$

则称 $(X_1, \ldots, X_n)$ 为连续型随机向量，f 为其（联合）概率密度函数（PDF）.

在上述定义中，概率密度函数 f 描述了随机向量 $(X_1, \ldots, X_n)$ 在 $\mathbb{R}^n$ 中各个区域的概率分布，其积分值表示随机向量取值落在特定区域的概率.

命题 3.2. 连续型随机向量 $(X_1, \ldots, X_n)$ 的概率密度函数具有以下性质：

1. 非负性：$f(x_1, \ldots, x_n) \geq 0, \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$
2. 归一性：$\int_{\mathbb{R}^n} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \equiv 1.$
3. 与 CDF 的关系：当 $n = 2$ 时，联合分布函数 $F(x, y)$ 与概率密度函数f(x, y)的关系为：

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s)\, \mathrm{d}s\, \mathrm{d}t,$$

且对于几乎所有的 $(a, b) \in \mathbb{R}^2$，有

$$f(a, b) = \frac{\partial^2 F}{\partial x \partial y}(a, b),$$

其中“几乎所有”（almost everywhere, 简记为 a.e.）指除去测度为零的点集外均成立.

上述性质表明，联合分布函数为概率密度函数的累积，而概率密度函数则可以通过对联合分布函数（形式上）求二阶混合偏导数还原.

例 3.3. 矩形域上均匀分布的概率密度函数可以表示为：

$$f(x, y) = \begin{cases} \dfrac{1}{(b-a)(d-c)}, & (x, y) \in (a, b) \times (c, d), \\ 0, & \text{其他情况.} \end{cases}$$

该例子描述的是在矩形区域 $(a, b) \times (c, d)$ 上的均匀分布，PDF 为常数，表示在该矩形区域内每个点的概率密度相等. 均匀分布是概率论中最简单的一类分布，可用于描述无偏向的随机现象，例如随机生成点或模拟等概率事件.

例 3.4. 二元正态分布（Bivariate Normal Distribution）是描述两个连续型随机变量关系的核心分布模型. 设随机向量 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，其联合概率密度函数为

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\left(\frac{x-\mu_1}{\sigma_1} \right)^2 + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 - 2\rho \frac{x-\mu_1}{\sigma_1} \frac{y-\mu_2}{\sigma_2} \right) \right),$$

其中 $(x,y)\in\mathbb{R}^2$，σ_1， $\sigma_2>0$， $|\rho|<1$. 为了简化表达，引入向量形式：

$$\boldsymbol{x}=\begin{bmatrix}\frac{x-\mu_1}{\sigma_1}\\ \frac{y-\mu_2}{\sigma_2}\end{bmatrix},\quad W=\frac{1}{1-\rho^2}\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix},$$

则概率密度函数的指数项可简写为：

$$-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}W\boldsymbol{x},$$

其中 W 是正定矩阵，其 Cholesky 分解为 $W=A^{\mathrm{T}}A$，这里

$$A=\frac{1}{\sqrt{1-\rho^2}}\begin{bmatrix}1 & -\rho\\ 0 & \sqrt{1-\rho^2}\end{bmatrix}.$$

二元正态分布在统计学、金融学、信号处理等诸多领域有着广泛应用. 其中，相关系数 ρ 刻画了两个随机变量 X 和 Y 之间的线性相关程度，μ_1,μ_2 表示边缘分布的均值，而 σ_1^2,σ_2^2 则反映了各自的离散程度.

3.4 边际分布

定义 3.5. 对于 n 维随机向量 $(X_1,\ldots,X_n)$，其第 i（$i=1,\ldots,n$）个分量 X_i 的*边际分布函数*（Marginal Distribution Function）为

$$F_{X_i}(x)=P(X_i\le x)=P(X_i\le x,-\infty<X_j<+\infty,\forall j\ne i).$$

边际分布是描述高维随机向量中单个或部分随机变量分布的一种方法. 通过将联合分布中其他分量的取值范围扩展至整个空间 $\mathbb{R}$，可以获得个别变量的边际分布. 换句话说，边际分布可视为高维联合测度在低维子空间上的投影. 下面以低维情形具体说明.

当 $n=2$ 时，设随机向量 (X,Y) 的联合分布函数为 $F(x,y)$，则 X 和 Y 的边际分布函数分别为

$$F_X(x)=\lim_{y\to+\infty}F(x,y),\quad F_Y(y)=\lim_{x\to+\infty}F(x,y).$$

当 $n=3$ 时，设随机向量 (X,Y,Z) 的联合分布函数为 $F(x,y,z)$，则 X 的边际分布函数为

$$F_X(x)=\lim_{y,z\to+\infty}F(x,y,z),$$

(X,Y) 的联合边际分布函数为

$$F_{X,Y}(x,y)=\lim_{z\to+\infty}F(x,y,z).$$

例 3.5. 对于二维随机向量 (X,Y)，

$$P(X>a,Y>b)=1-F_X(a)-F_Y(b)+F(a,b).$$

对于不同类型的随机变量，边际分布的计算方法也有所不同. 设二元离散随机变量 (X,Y) 的联合概率质量函数为 $P(X=x,Y=y)$，则的边际概率质量函数

$$P(X=x)=\sum_y P(X=x,Y=y).$$

即通过对其他分量求和，获得目标随机变量的边际分布. 连续型随机向量情形类似. 设 (X,Y) 的联合概率密度函数为 $f(x,y)$，则 X 的边际分布函数为

$$F_X(x)=P(X\le x,-\infty<Y<+\infty)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}f(t,y)\,\mathrm{d}y\,\mathrm{d}t,$$

进而 X 的边际概率密度函数为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\,\mathrm{d}y.$$

例 3.6. 考虑二元正态分布 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$. 直接计算可得 X 的边际概率密度函数为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\,\mathrm{d}y=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right),$$

即 $X\sim N(\mu_1,\sigma_1^2)$. 同理，$Y$ 的边际概率密度函数为

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right),$$

即 $Y\sim N(\mu_2,\sigma_2^2)$.

边际分布可以由联合分布唯一确定，但反之，联合分布不能仅通过边际分布唯一恢复. 二元正态分布就是一个典型的例子，其参数 ρ 的信息在 X 和 Y 的边际分布中是无法捕捉到的. 甚至可以构造这样的例子，两个分量的边际分布皆为正态分布，但联合分布不是二元正态分布，具体请参见习题.

3.5 条件分布

下面以二维情形为例进行条件分布的概念解析. 考虑随机向量 (X,Y).

若 (X,Y) 为离散型，设其联合概率质量函数为

$$P(X=a_i,Y=b_j)=p_{ij}\ge 0,\quad \sum_{i,j}p_{ij}=1.$$

则在事件 $Y = b_j$ 下，X 的条件概率质量函数定义为

$$P(X = a_i \mid Y = b_j) = \frac{P(X = a_i, Y = b_j)}{P(Y = b_j)} = \frac{p_{ij}}{\sum_k p_{kj}}.$$

该函数满足

$$\sum_i P(X = a_i \mid Y = b_j) = 1, \quad \forall j.$$

例 3.7. 考虑两次掷硬币试验. X 表示第一次掷得正面次数（0 或 1），Y 表示两次总正面次数（0, 1, 2），其联合分布如下表所示：

$X\backslash Y$	0	1	2
0	0.25	0.25	0
1	0	0.25	0.25

当观测到$Y = 1$时，条件分布计算为：

$$P(X = 0 \mid Y = 1) = \frac{0.25}{0.5} = 0.5, \quad P(X = 1 \mid Y = 1) = \frac{0.25}{0.5} = 0.5$$

若 (X, Y) 为连续型，设其联合概率密度函数为 $f(x, y)$. 先考虑条件概率

$$P\left(X \le x \mid y \le Y \le y + \mathrm{d}y\right) = \frac{P(X \le x,\, y \le Y \le y + \mathrm{d}y)}{P(y \le Y \le y + \mathrm{d}y)} = \frac{\displaystyle\int_{-\infty}^{x} \int_{y}^{y+\mathrm{d}y} f(t, s)\, \mathrm{d}s\, \mathrm{d}t}{\displaystyle\int_{y}^{y+\mathrm{d}y} f_Y(s)\, \mathrm{d}s}.$$

当 $\mathrm{d}y \to 0$ 时，对 x 求导即可得到在 $Y = y$ 条件下 X 的条件概率密度函数形式为 $\dfrac{f(x, y)}{f_Y(y)}$.

定义 3.6. 对于连续型随机向量 (X, Y)，若联概率密度函数为 $f(x, y)$ 且 $f_Y(y) > 0$，则定义 X 在 $Y = y$ 条件下的*条件概率密度函数*（PDF）为

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

相应地，其*条件累积分布函数*（CDF）定义为

$$F_{X|Y}(a \mid y) = P(X \le a \mid Y = y) = \int_{-\infty}^{a} f_{X|Y}(x \mid y)\, \mathrm{d}x.$$

可以直接验证，$f_{X|Y}(x \mid y)$ 符合概率密度函数的所有性质，并且熟知的基本定理在连续型随机向量情形下也同样成立：

1. 乘法法则：

$$f(x, y) = f_{X|Y}(x \mid y) f_Y(y) = f_{Y|X}(y \mid x) f_X(x).$$

2. 全概率公式：

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y = \int_{-\infty}^{+\infty} f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y.$$

3. Bayes 公式：

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\displaystyle\int_{-\infty}^{+\infty} f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y}.$$

例 3.8. 设 $(X,Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则有

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\,\sigma_2} \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\left(y - \left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right)^2}{2(1-\rho^2)\sigma_2^2} \right).$$

由此确认，$X = x$ 条件下 Y 服从正态分布 $N\left(\mu_2 + \rho\dfrac{\sigma_2}{\sigma_1}(x-\mu_1),\, (1-\rho^2)\sigma_2^2\right)$.

例 3.9 (Farlie-Morgenstein 族)**.** 设 $F(x)$ 和 $G(y)$ 是一维随机变量的累积分布函数，对任意的参数 $\theta \in [-1,1]$，定义二元函数

$$H(x,y) := F(x)G(y)\left[1 + \theta\left(1 - F(x)\right)\left(1 - G(y)\right)\right].$$

可以验证， $H(x,y)$ 是一个二元随机向量 (X,Y) 的联合累积分布函数. 同时，X 和 Y 的边际分布分别为 $F(x)$ 和 $G(y)$：

$$\lim_{y\to+\infty} H(x,y) = F(x), \quad \lim_{x\to+\infty} H(x,y) = G(y).$$

通过选取不同的 θ 值（例如 $\theta = -1$ 或 $\theta = 1$），可以构造具有特定边际分布的不同联合分布.

3.6 随机变量的独立性

定义 3.7. 设 (X,Y) 为二维随机向量，其联合分布函数为 $F(x,y)$. 若对任意 $x, y \in \mathbb{R}$ 有

$$F(x,y) = F_X(x)F_Y(y),$$

其中 $F_X(x)$ 和 $F_Y(y)$ 分别为 X 和 Y 的边际分布函数，则称 X 与 Y 相互独立.

对于离散型或连续型随机向量 (X,Y)，独立性等价于其联合概率质量函数或概率密度函数满足：

$$f(x,y) = f_X(x)f_Y(y), \quad \forall x, y \in \mathbb{R},$$

其中 $f_X(x)$ 和 $f_Y(y)$ 分别为 X 和 Y 的边际概率质量函数或概率密度函数.

例 3.10. 设 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则以下等价：

1. X 与 Y 独立.
2. 相关系数 $\rho=0$.
3. 联合密度可分解 $f(x,y)=\dfrac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\cdot\dfrac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$.

定义 3.8. 推广到 n 维情形，设 $(X_1,\dots,X_n)$ 为 n 维随机向量，其联合分布函数为 $F(x_1,\dots,x_n)$. 若对任意 $x_1,\dots,x_n\in\mathbb{R}$ 都有

$$F(x_1,\dots,x_n)=F_1(x_1)\dots F_n(x_n),$$

其中 $F_i(x_i)$ 为 X_i 的边际分布函数，则称 $X_1,\dots,X_n$ 相互独立.

类似地，对于离散型或连续型随机向量，独立性等价于联合概率质量函数或概率密度函数满足：

$$f(x_1,\dots,x_n)=f_1(x_1)\dots f_n(x_n),\quad \forall x_1,\dots,x_n\in\mathbb{R},$$

其中 $f_i(x_i)$ 为 X_i 的边际概率质量函数或概率密度函数.

定理 3.1. 关于独立性有以下两个重要结果：

1. 若 $X_1,\dots,X_n$ 相互独立，则对任意 $m\in\{1,\dots,n-1\}$ 和可测函数 g_1,g_2，随机变量

   $$Y_1=g_1(X_1,\dots,X_m)\quad 与\quad Y_2=g_2(X_{m+1},\dots,X_n)$$

   也相互独立. 这表明独立性在函数变换下是保持的.
2. （**因子分解定理**）设 $(X_1,\dots,X_n)$ 为连续型随机向量，若其联合概率密度函数可表示为各分量函数乘积：

   $$f(x_1,\dots,x_n)=g_1(x_1)\dots g_n(x_n),\quad \forall x_1,\dots,x_n\in\mathbb{R},$$

   其中 $g_i\geq 0$，$i=1,\dots,n$，则 $X_1,\dots,X_n$ 相互独立，且每个 X_i 的边际概率密度函数 f_i 与 g_i 仅相差一个常数因子.

例 3.11. 令三角形区域 $D=\{(x,y)\in\mathbb{R}^2\mid x\geq 0,\,y\geq 0,\,x+y\leq 1\}$. 设随机向量 (X,Y) 在区域 D 上服从均匀分布，则其概率密度为

$$f(x,y)=\begin{cases}2, & (x,y)\in D,\\ 0, & 其他.\end{cases}$$

直接计算可得，当 $(x,y)\in D$ 时，

$$f_X(x)f_Y(y)=4(1-x)(1-y)\neq 2=f(x,y).$$

故 X 与 Y 不独立.

3.7 随机向量的函数

本节讨论对于一个给定的随机向量 $(X_1,\dots,X_n)$ 和可测函数 g，如何确定随机变量 $Y=g(X_1,\dots,X_n)$ 的分布.

首先介绍**直接法**，指通过概率定义（离散型的概率质量函数或连续型的概率密度函数）直接求出目标随机变量的分布.

例 3.12. 设 $X_1\sim B(n_1,p)$, $X_2\sim B(n_2,p)$ 且相互独立，考虑随机变量 $Y=X_1+X_2$. 对于任意整数 $k\in\{0,1,\dots,n_1+n_2\}$，有

$$\begin{aligned}
P(Y=k)&=P(X_1+X_2=k)\\
&=\sum_{k_1=0}^{k}P(X_1=k_1,X_2=k-k_1)\\
&=\sum_{k_1=0}^{k}P(X_1=k_1)P(X_2=k-k_1)\quad\text{（由独立性）}\\
&=\sum_{k_1=0}^{k}\binom{n_1}{k_1}p^{k_1}(1-p)^{n_1-k_1}\binom{n_2}{k-k_1}p^{k-k_1}(1-p)^{n_2-(k-k_1)}\\
&=\binom{n_1+n_2}{k}p^k(1-p)^{n_1+n_2-k}.
\end{aligned}$$

因此，$Y\sim B(n_1+n_2,p)$.

例 3.13. 设随机向量 (X_1,X_2) 的联合概率密度函数为 $f(x_1,x_2)$，且 $X_1>0$. 考虑随机变量

$$Y=\frac{X_2}{X_1},$$

其分布函数为

$$\begin{aligned}
P(Y\le y)&=P\left(\frac{X_2}{X_1}\le y\right)=P(X_2\le X_1y)\\
&=\int_D f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2,
\end{aligned}$$

其中积分区域

$$D=\{(x_1,x_2)\mid x_1>0,x_2\le x_1y\}$$

对应图 3.1 所示的范围.

进一步作变量替换 $x_2=x_1t$，则

$$P(Y\le y)=\int_0^{+\infty}\int_{-\infty}^{y}f(x_1,x_1t)x_1\,\mathrm{d}t\,\mathrm{d}x_1.$$

因此，Y 的概率密度函数为

$$l(y)=\int_0^{+\infty} x_1 f(x_1, yx_1)\,\mathrm{d}x_1.$$

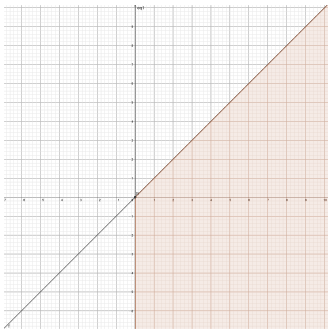

图 3.1: 区域 D：边界线斜率即为 y

接下来介绍**密度函数变换法**.

对于连续型随机向量，还可以根据可逆变换的观点，利用积分坐标变换求解目标变量或向量的分布. 设随机向量 (X_1, X_2) 的联合概率密度函数为 $f(x_1, x_2)$，并存在可逆变换

$$\begin{cases} Y_1 = g_1(X_1, X_2), \\ Y_2 = g_2(X_1, X_2), \end{cases}$$

且其逆变换为

$$\begin{cases} X_1 = h_1(Y_1, Y_2), \\ X_2 = h_2(Y_1, Y_2). \end{cases}$$

进一步假设变换在映射域上可微，Jacobi 行列式为

$$J=\det\begin{bmatrix} \dfrac{\partial h_1}{\partial y_1} & \dfrac{\partial h_1}{\partial y_2} \\ \dfrac{\partial h_2}{\partial y_1} & \dfrac{\partial h_2}{\partial y_2} \end{bmatrix}.$$

则根据积分坐标变换，随机向量 (Y_1, Y_2) 的联合概率密度函数为

$$l(y_1, y_2) = f\left(h_1(y_1, y_2), h_2(y_1, y_2)\right)|J|.$$

例 3.14. 设随机向量 (X_1, X_2) 的联合概率密度函数为 $f(x_1, x_2)$，考虑 $Y = X_1 + X_2$. 引入辅助变量 $Z = X_1$，有可逆变换

$$\begin{cases} X_1 = Z, \\ X_2 = Y - Z. \end{cases}$$

Jacobi 行列式为

$$J = \det \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} = -1.$$

因此，随机向量 (Y, Z) 的联合概率密度为

$$l(y, z) = f(z, y - z)\,|J| = f(z, y - z).$$

对 z 进行积分得到 Y 的概率密度为

$$l_Y(y) = \int_{-\infty}^{+\infty} f(z, y - z)\,\mathrm{d}z.$$

若进一步假设 X_1 和 X_2 相互独立，f_1 和 f_2 分别为 X_1 和 X_2 的概率密度，则 $f(x_1, x_2) = f_1(x_1)f_2(x_2)$，从而 Y 的概率密度为

$$l_Y(y) = \int_{-\infty}^{+\infty} f_1(z)f_2(y - z)\,\mathrm{d}z = (f_1 * f_2)\,(y),$$

其中 $f_1 * f_2$ 表示函数 f_1 和 f_2 的**卷积**.

一般地，通过密度变换法可得到正态分布的可加性：若 $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则

$$X_1 + X_2 \sim N\left(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2\right).$$

统计中的三大重要抽样分布（χ^2 ，t 和 F 分布）均为正态随机变量的函数，详细推导请参考多元统计分析相关资料.

第四章　随机变量的数字特征

4.1　期望

（数学）期望用于描述随机变量分布的中心趋势，是概率论中最重要最基本的数字特征之一. 离散型和连续型随机变量的期望分别参见定义 2.7 和定义 2.12 . 对于 n 维随机向量 $(X_1, \ldots, X_n)$ 的期望定义为各分量期望的有序组：

$$\mathrm{E}((X_1, \ldots, X_n)) := (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_n)).$$

命题 4.1. 期望具有如下重要性质：

1. **随机向量函数的期望：** 设 $g : \mathbb{R}^n \to \mathbb{R}$ 为可测函数，$(X_1, \ldots, X_n)$ 为离散型或连续型随机向量，其联合概率密度或概率质量函数为 f，则

$$\mathrm{E}(g(X_1, \ldots, X_n)) = \begin{cases} \sum g(x_1, \ldots, x_n) f(x_1, \ldots, x_n), & \text{离散型,} \\ \displaystyle\int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, & \text{连续型.} \end{cases}$$

2. **线性性质：** 对任意随机变量 X, Y 和常数 $a, b \in \mathbb{R}$，有

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y).$$

3. **独立性与乘积期望：** 若随机变量 $X_1, \ldots, X_n$ 相互独立，则

$$\mathrm{E}(X_1 \cdots X_n) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_n).$$

关于上述性质，有以下补充说明：第一条性质中，$g(X_1, \ldots, X_n)$ 的期望需要存在（即相应的求和或积分要绝对收敛）；线性性质可推广至任意有限项线性组合；独立性条件对乘积期望性质的成立是充分非必要的. 事实上，只要随机变量之间的相关系数为零，该性质就成立.

例 4.1. 设 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则

$$\mathrm{E}(XY) = \mu_1\mu_2 + \rho\sigma_1\sigma_2$$

特别地，当且仅当 X, Y 相互独立（此时等价于 $\rho = 0$）时，有 $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

例 4.2. 设 $X_1, \ldots, X_n$ 独立同分布且都服从 $N(0,1)$，则

$$\mathrm{E}(X_1^2 + \cdots + X_n^2) = n.$$

事实上，

$$X_1^2 + \cdots + X_n^2 \sim \chi^2(n),$$

即服从自由度为 n 的 χ^2 分布.

4.2 分位数

定义 4.1. 设连续型随机变量 X 的累积分布函数为 $F(x)$. 若实数 m 满足

$$F(m) = P(X \le m) = 0.5,$$

则称 m 为 X 的中位数（Median），记为 $\mathrm{Med}(X)$. m 将概率质量等分，满足 $P(X \le m) = P(X \ge m) = 0.5$

中位数也是一种反映随机变量集中趋势的重要统计量，它将连续型随机变量的取值分为等概率的两部分. 与均值类似，中位数也能够描述数据分布的中心位置. 然而，与均值相比，中位数受极端值（outliers）的影响较小，具有更高的稳健性. 容易看到中位数取值不一定唯一.

命题 4.2. 设连续型随机变量 X 有中位数 m，则

$$\mathrm{E}(|X - c|) \ge \mathrm{E}(|X - m|), \quad c \in \mathbb{R}\text{fi}$$

且等号成立当且仅当 $c = m$.

命题证明留给读者. 对于离散型随机变量，前述中位数定义不再适用. 下面给出中位数的一般形式刻画，当随机变量是连续型时，与定义 4.1 一致.

定义 4.2. 对任意随机变量 X，若实数 m 满足

$$P(X \le m) \ge 0.5 \quad \text{且} \quad P(X \ge m) \ge 0.5,$$

则称 m 为 X 的中位数.

中位数 m 的定义等价于满足

$$P(X < m) \le 0.5 \quad \text{且} \quad P(X > m) \le 0.5.$$

X	1	2	3	4
P	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{12}$	$\frac{1}{12}$

例 4.3. 设离散型随机变量 X 的分布表为： 由此可知，

$$P(X \le 2) = \frac{5}{6} \ge 0.5 \quad 且 \quad P(X \ge 2) = \frac{4}{6} \ge 0.5,$$

因此 $m = 2$ 是 X 的（唯一）中位数.

定义 4.3. 对随机变量 X 和给定 $\alpha \in (0,1)$，若实数 a_α 满足

$$P(X \le a_\alpha) \ge \alpha \quad 且 \quad P(X \ge a_\alpha) \ge 1-\alpha,$$

则称 a_α 为 X 的（下） α-分位数. 当 $\alpha = 0.5$ 时即为中位数.

例 4.4. 给定 $\alpha \in (0,1)$，定义分布函数的广义逆（分位数函数）为

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \ge \alpha\}.$$

可验证此函数左连续，且

$$P(X < F^{-1}(\alpha)) \le \alpha \le P(X \le F^{-1}(\alpha)).$$

显然，$F^{-1}(\alpha)$ 是 X 的一个下 α-分位数.

除了均值和中位数，众数（Mode）也是一种常用的随机变量取值集中趋势的描述. 对于离散型随机变量，众数为使概率质量函数 $f(x)$ 达到最大值的点. 对于连续型随机变量，众数是其概率密度函数 $f(x)$ 的局部极大值点. 需要指出的是，这一定义尽管直观但并不严谨. 例如，由于概率密度函数在零测集上的值可以任意调整，所以连续型随机变量的众数往往缺乏测度意义上的一致性，. 对于离散型随机变量，众数总是存在的；而对于连续型随机变量，其众数可能不存在，即使存在也不一定具有解析或唯一性.

4.3 方差

离散型和连续型随机变量的方差以及标准差分别详见定义 2.7 和定义 2.12. 对于一般随机变量 X，其方差定义为

$$\mathrm{Var}(X) := \mathrm{E}((X - \mathrm{E}(X))^2),$$

标准差记为

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

方差和标准差主要用于刻画随机变量取值的分散程度或波动情况.

特别地，当 X 表示金融市场中的收益率时， $\mathrm{SD}(X)$ 通常称为波动率（Volatility），用于衡量收益的不确定性或风险水平. 在实际分析中，为消除尺度影响，通常引入**变异系数**（Coefficient of Variation, 简记为 CV）：

$$\mathrm{CV} := \frac{\mathrm{SD}(X)}{\mu},$$

其中 $\mu = \mathrm{E}(X) \neq 0$ 为 X 的均值. 变异系数反映了标准差相对于均值的相对大小，常用于比较不同量纲数据的波动程度.

命题 4.3. 方差具有以下重要性质：

(1) 对任意常数 c，有 $\mathrm{Var}(c) \equiv 0$.

(2) 对任意常数 c 和随机变量 X，有 $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

(3) 对任意随机变量 X 和 Y，有

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{E}\left((X-\mathrm{E}(X))\,(Y-\mathrm{E}(Y))\right).$$

特别地，若 X 和 Y 相互独立，则 $\mathrm{E}((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))) = 0$，从而

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

例 4.5. 设随机变量 $X_1, X_2, \ldots, X_n$ 独立同分布，其公共期望为 μ，公共方差为 σ^2. 定义：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

其中 $\bar{X}$ 称为**样本均值**，S^2 称为**样本方差**. 易见 $\mathrm{E}(\bar{X}) = \mu$，下面分别计算 $\mathrm{Var}(\bar{X})$ 和 $\mathrm{E}(S^2)$.

由于 $X_1, X_2, \ldots, X_n$ 独立同分布，根据方差性质有

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

对于样本方差 S^2，将其展开可得

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right).$$

利用独立性和线性性质，对其求期望：

$$\mathrm{E}(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\cdot\frac{\sigma^2}{n}\right) = \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2.$$

这表明样本方差 S^2 是总体方差 σ^2 的无偏估计量，相关性质将在统计推断部分详细讨论.

在更广泛的统计分析中，方差的概念可进一步推广，例如协方差矩阵、条件方差等.

4.4 协方差与相关系数

定义 4.4. 对于随机变量 X 和 Y，其*协方差*定义为

$$\mathrm{Cov}(X,Y) := \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$$

其中 $\mu_X = \mathrm{E}(X)$，$\mu_Y = \mathrm{E}(Y)$ 分别为 X 和 Y 的期望.

命题 4.4. 协方差运算满足下列基本性质：

(1) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(2) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$（对称性）.

(3) $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.

(4) 对任意常数 $a,b,c \in \mathbb{R}$，

$$\mathrm{Cov}(aX_1 + bX_2 + c, Y) = a\,\mathrm{Cov}(X_1,Y) + b\,\mathrm{Cov}(X_2,Y).$$

协方差符号体现变量间协同变化方向. 在讨论诸如 $\mathrm{Cov}\left(\sum X_i, \sum Y_j\right)$ 等问题时，可利用性质 (4) 进行逐项展开计算.

定义 4.5. 随机变量 X 与 Y 的*Pearson 相关系数*为标准化协方差：

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \mathrm{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right],$$

其中 $\sigma_X = \sqrt{\mathrm{Var}(X)}$，$\sigma_Y = \sqrt{\mathrm{Var}(Y)}$. 相关系数满足 $\rho(X,Y) \in [-1,1]$，其绝对值度量相关性强度，符号指示相关方向.

注：X,Y 的 Pearson 相关系数有时也记为 $\mathrm{Corr}(X,Y)$.

定理 4.1. 相关系数具有以下特征：

1. 若 X,Y 独立，则 $\rho(X,Y) = 0$（称 X 与 Y *不相关*）；但反之未必成立.
2. $|\rho(X,Y)| \le 1$，且等号成立当且仅当存在常数 a,b，使得 $P(Y = aX+b) = 1$，即称 $Y = aX + b$ *几乎必然成立*（almost surely，简记为 a.s.）.

定理 4.1 证明. 性质 1 由定义直接可得.

对于性质 2，引入标准化变量 $U = \dfrac{X-\mu_X}{\sigma_X}$ 与 $V = \dfrac{Y-\mu_Y}{\sigma_Y}$，则

$$\rho(X,Y) = \mathrm{Cov}(U,V) = \mathrm{E}(UV)$$

应用 Cauchy-Schwarz 不等式可得

$$|\rho(X,Y)| \le \sqrt{\mathrm{E}(U^2)\mathrm{E}(V^2)} = 1,$$

等号成立当且仅当存在常数 t_0 使 $V = t_0 U$ a.s.，整理可得所要结论，这里 $a = \dfrac{\sigma_Y}{\sigma_X}$. □

例 4.6. 设 $X \sim N(0,1)$，$Y = X^2$，则

$$\mathrm{Cov}(X,Y) = \mathrm{E}(X^3) - \mathrm{E}(X)\mathrm{E}(X^2) = 0,$$

从而 $\rho(X,Y) = 0$，即 X 与 Y 不相关. 然而 X,Y 显然存在确定性二次关系，说明"不相关"与"独立"并不等价.

例 4.7. 对于二元正态分布 $(X,Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$，有

$$\rho(X,Y) = \rho.\text{（参数直接对应相关系数）}$$

证明要点：通过正交变换将其化为独立标准正态变量的线性组合.

例 4.8. （配对问题）考虑 n 个人随机取帽子的配对问题，令 X 表示拿到自己帽子的人数，I_i 表示第 i 个人匹配成功的示性变量，则

$$X = \sum_{i=1}^{n} I_i, \quad \mathrm{E}(X) = \sum_{i=1}^{n} \frac{1}{n} = 1.$$

根据方差性质和协方差定义，计算可得

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(I_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(I_i, I_j), \\
\mathrm{Var}(I_i) &= \frac{1}{n}\left(1 - \frac{1}{n}\right) = \frac{n-1}{n^2}, \\
\mathrm{Cov}(I_i, I_j) &= \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}.
\end{aligned}$$

由此可得

$$\mathrm{Var}(X) = n \cdot \frac{n-1}{n^2} + \binom{n}{2} \cdot \frac{1}{n^2(n-1)} = 1.$$

4.5 矩

定义 4.6. 对于随机变量 X 和常数 c，称 $\mathrm{E}((X-c)^k)$ 为 X 关于 c 点的 k 阶矩，其中 $k = 1, 2, \ldots$. 特别地，当 $c = 0$ 时，称为 k 阶原点矩，记为 μ_k；当 $c = \mathrm{E}(X)$ 时，称为 k 阶中心矩，记为 m_k.

根据定义，1 阶原点矩 $\mu_1 \equiv \mathrm{E}(X)$，1 阶中心矩 $m_1 \equiv 0$，而 2 阶中心矩即为随机变量的方差

$$m_2 = \mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X).$$

若随机变量 X 的分布关于均值对称，则其所有奇数阶中心矩均为零.

定义 4.7. 设 $\mu = \mathrm{E}(X)$, $\sigma = \sqrt{\mathrm{Var}(X)}$，则称

$$\tilde{\mu}_k := \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^k\right] = \frac{m_k}{\sigma^k}$$

为 X 的 k 阶标准化矩.

特别地，1 阶标准化矩 $\tilde{\mu}_1 \equiv 0$；2 阶标准化矩 $\tilde{\mu}_2 \equiv 1$；而 3 阶标准矩称为 X 的*偏度系数*（Skewness），记为

$$\mathrm{Skew}(X) := \frac{\mathrm{E}((X-\mu)^3)}{\sigma^3},$$

用于衡量分布的非对称性.

例 4.9. 若 $X \sim N(0,1)$，则其偏度系数为

$$\mathrm{Skew}(X) = \int_{-\infty}^{+\infty} x^3 f(x)\,\mathrm{d}x = 0,$$

其中 $f(x)$ 为 X 的概率密度函数. 事实上，所有对称分布的偏度均为零.

偏度系数的符号表明分布的偏态方向：

- 若 $\mathrm{Skew}(X) < 0$，则称分布为“负偏”或“左偏”，分布通常呈现为左侧长尾（见图 4.1）.
- 若 $\mathrm{Skew}(X) > 0$，则称分布为“正偏”或“右偏”.
- 若 $\mathrm{Skew}(X) = 0$，则提示无显著方向性偏态.

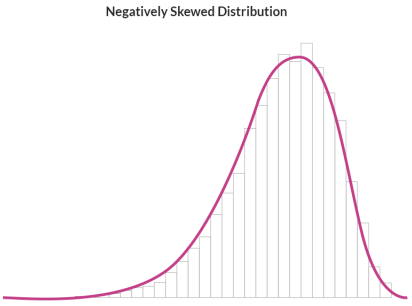

图 4.1: 负偏（左偏）分布示意图

4 阶标准矩称为 X 的*峰度系数*（Kurtosis），记为

$$\mathrm{Kurt}(X) := \frac{\mathrm{E}((X-\mu)^4)}{\sigma^4}.$$

例 4.10. 若 $X \sim N(\mu, \sigma^2)$，根据偶次幂的对称性，其 4 阶中心矩为 $3\sigma^4$，因此，

$$\mathrm{Kurt}(X) = \frac{\mathrm{E}((X-\mu)^4)}{\sigma^4} = 3.$$

由于正态分布的峰度系数为 3，通常将 $\mathrm{Kurt}(X)-3$ 定义为超额峰度系数.表征分布的尖峰厚尾特性：（超额）峰度衡量分布的尖峰程度和尾部特征衡量尾部的厚度和集中程度，常用于检测极端值的可能性：

- $\mathrm{Kurt}(X)>3$ 表示分布通常呈现“尖峰厚尾”，称为高峰度分布（Leptokurtic）；
- $\mathrm{Kurt}(X)<3$ 表示分布通常呈现“平峰薄尾”（platykurtic）；
- $\mathrm{Kurt}(X)=3$ 表明分布与正态分布相似（如图 4.2）.

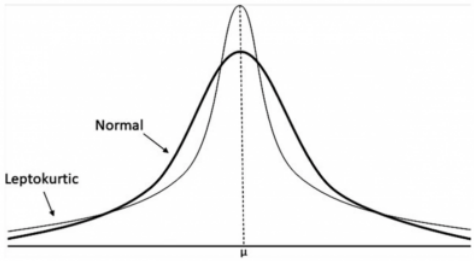

图 4.2: 高峰度分布与正态分布对比

例 4.11. 考虑自由度为 ν 的 t 分布，其概率密度函数为

$$f(t;\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},\quad t\in\mathbb{R},$$

其中 $\Gamma(x)$ 是 Gamma 函数. t 分布的峰度系数为

$$\mathrm{Kurt}(X)=3+\frac{6}{\nu-4}\quad(\nu>4).$$

当 $\nu\to 4^+$ 时，峰度趋向无穷大，呈现显著”尖峰厚尾”特性；而当 $\nu\to\infty$ 时，t 分布收敛到正态分布.

对于 $k\geq 5$ 的高阶标准矩，虽然理论上也可以刻画分布的偏态和尾部行为，但由于更高阶的幂次计算容易受噪声和离群值影响，在实际中较少使用.

4.6 矩母函数

定义 4.8. 对于随机变量 X，若函数

$$M_X(t)=\mathrm{E}(e^{tX})$$

在 $t=0$ 的某个邻域内存在，则称其为 X 的矩母函数（Moment Generating Function，简记为 MGF）.

矩母函数得名于其生成随机变量各阶矩的特性. 其存在性要求随机变量的指数函数期望收敛，但这一条件并非对所有随机变量都成立（例如柯西分布、对数正态分布等厚尾分布的的矩母函数不存在）.

命题 4.5. 矩母函数具有以下基本性质：

(1) $M_X(0) = 1$.

(2) 对于线性变换 $Y = aX + b$，有

$$M_Y(t) = e^{bt} M_X(at).$$

(3) 若 X 和 Y 独立，则 $Z = X + Y$ 的矩母函数为

$$M_Z(t) = M_X(t) M_Y(t).$$

例 4.12. 若 $X \sim \mathrm{Exp}(\lambda)$ 为指数分布，由定义得

$$M_X(t) = \mathrm{E}(e^{tX}) = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} \mathrm{d}x = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

这表明指数分布的矩母函数在 $t < \lambda$ 的区域内存在.

例 4.13. 若 $X \sim N(0, 1)$，则其矩母函数为

$$M_X(t) = \mathrm{E}(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = e^{\frac{t^2}{2}}, \quad t \in \mathbb{R}.$$

对于一般正态随机变量 $Y \sim N(\mu, \sigma^2)$，令 $Y = \sigma X + \mu$，则

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

矩母函数的主要应用如下：

- **生成各阶矩**：通过求导数，$M_X^{(n)}(0) = \mathrm{E}(X^n)$.
- **分布的唯一性**：（在某个邻域内等价的）矩母函数唯一确定对应分布.
- **独立随机变量的和**：借助矩母函数可以简化随机变量加和的分析.

定理 4.2. 若随机变量 X 的矩母函数 $M_X(t)$ 存在，则 X 的 n 阶原点矩为

$$\mathrm{E}(X^n) = M_X^{(n)}(0),$$

其中 $M_X^{(n)}(t)$ 表示 $M_X(t)$ 的 n 阶导数.

证明. 仅给出形式验证. 利用 Taylor 展开，矩母函数可以表示为

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right) = \sum_{n=0}^{\infty} \mathrm{E}(X^n)\frac{t^n}{n!}.$$

将其与 $M_X(t)$ 在 $t=0$ 处的 Taylor 展开式比较系数即可得 $M_X^{(n)}(0) = \mathrm{E}(X^n)$. □

例 4.14 (正态分布的矩计算). 若 $X \sim N(0,1)$，其矩母函数为 $M_X(t) = e^{\frac{t^2}{2}}$. 利用指数函数幂级数展开有

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\left(\frac{t^2}{2}\right)^n}{n!}.$$

比较系数可得

$$\mathrm{E}(X^{2n}) = \frac{(2n)!}{2^n n!}, \quad \mathrm{E}(X^{2n+1}) = 0, \quad n = 0, 1, 2, \ldots.$$

特别地，

$$\mathrm{Var}(X) = \mathrm{E}(X^2) = 1, \quad \mathrm{Kurt}(X) = \mathrm{E}(X^4) = 3.$$

定理 4.3. 若存在常数 $a > 0$，使得对所有 $t \in (-a, a)$，矩母函数 $M_X(t) = M_Y(t)$ 成立，则随机变量 X 和 Y 具有相同的分布.

定理的证明涉及到解析延拓原理和逆拉普拉斯变换，此处从略.

例 4.15. 若已知随机变量 X 的矩母函数为

$$M_X(t) = \frac{1}{2}e^{-t} + \frac{1}{4} + \frac{1}{8}e^{4t} + \frac{1}{8}e^{5t},$$

由于对任意离散随机变量 Y，其矩母函数可表示为 $M_Y(t) = \sum_y e^{ty}P(Y=y)$，将 $M_X(t)$ 与此形式比较，可得 X 为离散随机变量，其分布表如下：

X	-1	0	4	5
P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

例 4.16. 设 $X_1 \sim N(\mu_1, \sigma_1^2)$，$X_2 \sim N(\mu_2, \sigma_2^2)$ 且相互独立，则

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = e^{(\mu_1+\mu_2)t + \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}.$$

这是正态分布的矩母函数，根据定理 4.3可得，$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

除矩母函数之外，还有与之密切相关的生成函数工具：

- **概率母函数**（Probability Generating Function, 简记为 PGF）：对非负离散随机变量 X，定义为

$$G_X(t) = \mathrm{E}(t^X) = \sum_{k=0}^{\infty} P(X=k)t^k, \quad |t| \le 1.$$

概率母函数适用于分析计数过程，特别是在泊松过程等应用中. 它与矩母函数的关系为 $M_X(t) = G_X(e^t)$.

- **特征函数**（Characteristic Function）：定义为

$$\phi_X(t) = \mathrm{E}(e^{itX}), \quad i^2 = -1.$$

特征函数没有存在性问题，适用于所有类型的随机变量，若随机变量 X 的矩母函数存在，则 $\phi_X(t) = M_X(it)$. 特征函数通过傅里叶逆变换可唯一确定分布.

4.7 条件期望

条件期望是概率论中分析随机变量关系的核心工具，它不仅度量了随机变量在给定信息下的集中趋势，还为预测和统计推断提供了理论基础.

定义 4.9 (条件期望). 给定随机变量 X, Y，对任意可测集 $I \subset \mathbb{R}$，称

$$\mathrm{E}(Y|X \in I) := \begin{cases} \sum\limits_i y_i P(Y=y_i|X \in I), & \text{若 } (X,Y) \text{ 为离散型,} \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|X \in I)\mathrm{d}y, & \text{若 } (X,Y) \text{ 为连续型,} \end{cases}$$

为**条件期望**. 当条件具体化为 $X = x$ 时，记为 $\mathrm{E}(Y|x)$. 定义函数

$$h(x) := \mathrm{E}(Y|x) = \begin{cases} \sum\limits_i y_i P(Y=y_i|X=x), & \text{离散型,} \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)\mathrm{d}y, & \text{连续型.} \end{cases}$$

将 h 作用于随机变量 X 得到随机变量 $h(X) = \mathrm{E}(Y|X)$，称为 Y 对 X 的**条件期望**，也称为 Y 对 X 的**回归函数**.

例 4.17. 若 $(X,Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则

$$\mathrm{E}(Y|x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

这表明在二维正态分布中，条件期望相对于 x 呈线性关系，对于线性回归理论具有重要意义.

例 4.18. 设某产品有两种型号，平均寿命分别为 10 年和 15 年，市场占有率分别为 60% 和 40%. 令 X 表示产品型号（对应取值 1 和 2），Y 表示寿命（单位：年），则

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}(\mathrm{E}(Y|X)) \\ &= \mathrm{E}(Y|X=1)P(X=1) + \mathrm{E}(Y|X=2)P(X=2) \\ &= 10 \times 0.6 + 15 \times 0.4 = 12.\end{aligned}$$

此计算体现了条件期望的一个基本性质——全期望公式.

定理 4.4 (全期望公式). 对随机向量 (X,Y)，若相关期望存在，则

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X)).$$

证明. 以连续型为例. 设 (X,Y) 的联合密度为 $f(x,y)$，X 的边际密度为 $f_X(x)$，Y 在给定 $X=x$ 条件下的条件密度为 $f_{Y|X}(y|x)$，则

$$\begin{aligned}\mathrm{E}(\mathrm{E}(Y|X)) &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} y f_{Y|X}(y|x) \mathrm{d}y \right) f_X(x) \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x,y) \mathrm{d}y \mathrm{d}x \\ &= \mathrm{E}(Y)\end{aligned}$$

□

例 4.19. 一个矿工在井下迷路，所处位置有三个门：

- 从第一个门出去，2 小时可达安全之处；
- 从第二个门出去，3 小时后会回到原地；
- 从第三个门出去，1 小时后会回到原地.

该矿工随机选择门，而且始终无法区分三个门. 设他到达安全处的期望时间为 $\mathrm{E}[Y]$. 令 X 为其首次选择的门号，则

$$\mathrm{E}[Y|X=1] = 2, \quad \mathrm{E}[Y|X=2] = 3 + \mathrm{E}[Y], \quad \mathrm{E}[Y|X=3] = 1 + \mathrm{E}[Y].$$

由全期望公式，

$$\mathrm{E}[Y] = \frac{1}{3}(2 + (3 + \mathrm{E}[Y]) + (1 + \mathrm{E}[Y])),$$

解得 $\mathrm{E}[Y] = 6$.

定理 4.5. 对任意可测函数 g，有

$$\mathrm{E}((Y - g(X))^2) \geq \mathrm{E}((Y - \mathrm{E}(Y|X))^2).$$

即条件期望是均方误差意义下的最优预测.

证明. 这里给出一个不够严格但富有启发性的证明. 根据期望的性质

$$\mathrm{E}((Y-c)^2) \geq \mathrm{E}((Y-\mathrm{E}(Y))^2)\,(\forall c \in \mathbb{R}),$$

对于任意可测函数 $g(X)$，则有

$$\mathrm{E}((Y-g(X))^2|X) \geq \mathrm{E}((Y-\mathrm{E}(Y|X))^2|X).$$

两边对 X 求期望即得

$$\mathrm{E}((Y-g(X))^2) \geq \mathrm{E}((Y-\mathrm{E}(Y|X))^2).$$

□

由于条件期望依赖 X 和 Y 的联合分布，实际中往往难以直接获取，即使是数值近似计算，因此常使用最优线性预测，即最小化均方误差 $\min\limits_{a,b}\mathrm{E}((Y-(aX+b))^2)$，对应的就是经典的最小二乘法.

命题 4.6. 设 $\hat{Y}=\mathrm{E}(Y|X)$，$\tilde{Y}=Y-\hat{Y}$，则

(1) $\mathrm{E}(\tilde{Y})=0$.

(2) $\mathrm{E}(\tilde{Y}\hat{Y})=0$.

(3) $\mathrm{Var}(Y)=\mathrm{Var}(\hat{Y})+\mathrm{Var}(\tilde{Y})$.

上述命题在统计学中具有重要应用，证明可通过直接验证得到，留给读者. 其中，性质（2）表明预测值与误差正交，性质（3）将方差分拆为可解释与不可解释两部分.

定义 4.10 (条件方差). 随机变量 Y 在给定 $X=x$ 条件下的*条件方差*定义为

$$\mathrm{Var}(Y|x) := \mathrm{E}\left((Y-\mathrm{E}(Y|x))^2\,|x\right) = \mathrm{E}(Y^2|x)-(\mathrm{E}(Y|x))^2.$$

将其视为 x 的函数，作用于随机变量 X 得到随机变量 $\mathrm{Var}(Y|X)$，称为在给定 X 条件下 Y 的*条件方差*.

定理 4.6 (全方差定律). 对任意随机向量 (X,Y)，若相关期望和方差存在，则

$$\mathrm{Var}(Y)=\mathrm{E}(\mathrm{Var}(Y|X))+\mathrm{Var}(\mathrm{E}(Y|X)).$$

全方差定律也称为方差分解公式，易见其与命题 4.6 中性质（3）等价.

第五章　不等式与极限定理

5.1　概率不等式

在概率论中，许多不等式为随机变量的尾部概率分布提供了重要的界限. 以下将详细介绍三种经典的概率不等式：马尔可夫不等式、切比雪夫不等式和切尔诺夫不等式.

定理 5.1 (Markov 不等式)**.** 若随机变量 $Y \geq 0$，则对任意 $a > 0$，有

$$P(Y \geq a) \leq \frac{\mathrm{E}(Y)}{a}.$$

证明. 定义一个示性变量 I 表示事件 $Y \geq a$ 是否发生：

$$I = \begin{cases} 1, & Y \geq a, \\ 0, & Y < a. \end{cases}$$

则显然有 $I \leq \dfrac{Y}{a}$. 由期望的线性性质，

$$P(Y \geq a) = \mathrm{E}(I) \leq \mathrm{E}\left(\frac{Y}{a}\right) = \frac{\mathrm{E}(Y)}{a}.$$

□

Markov 不等式利用期望值来界定非负随机变量取极端值的概率，是一种非常通用的界限，但这种界限通常较松，尤其在尾部区域（远离期望值）效果较差.

定理 5.2 (Chebyshev 不等式)**.** 若随机变量 Y 的方差 $\mathrm{Var}(Y)$ 存在，则对任意 $a > 0$，有

$$P(|Y - \mathrm{E}(Y)| \geq a) \leq \frac{\mathrm{Var}(Y)}{a^2}.$$

证明. 应用 Markov 不等式，有

$$P(|Y - \mathrm{E}(Y)| \geq a) = P((Y - \mathrm{E}(Y))^2 \geq a^2) \leq \frac{\mathrm{E}((Y - \mathrm{E}(Y))^2)}{a^2} = \frac{\mathrm{Var}(Y)}{a^2}.$$

□

Chebyshev 不等式进一步利用了随机变量的方差（二阶矩）信息，对其围绕期望值集中的趋势进行了分析. 这表明随机变量的方差越小，其偏离期望值的概率越小. 特别地， Chebyshev 不等式的一个直接推论是：当 $\mathrm{Var}(Y) \equiv 0$ 时，则有$Y = \mathrm{E}(Y)$ 几乎必然成立.

定理 5.3 (Chernoff 不等式). 对任意随机变量 Y 和任意常数 a 和任意 $t > 0$，有

$$P(Y \geq a) \leq \frac{\mathrm{E}(e^{tY})}{e^{ta}}.$$

证明. 通过 Markov 不等式，

$$P(Y \geq a) = P(e^{tY} \geq e^{ta}) \leq \frac{\mathrm{E}(e^{tY})}{e^{ta}}.$$

□

Chernoff 不等式通过引入指数变换，将尾部概率与随机变量矩母函数联系起来，利用参数 t 调整（优化选取），可获得更精确的概率界限，在机器学习和随机算法分析中被广泛应用.

例 5.1. 设随机变量 $X \sim N(0,1)$（标准正态分布），计算 $P(|X| \geq 3)$，并利用上述三个不等式进行估计.

1. 根据 Markov 不等式：

$$P(|X| \geq 3) \leq \frac{\mathrm{E}(|X|)}{3} = \frac{1}{3}\sqrt{\frac{2}{\pi}} \approx 0.27.$$

2. 根据 Chebyshev 不等式：

$$P(|X| \geq 3) \leq \frac{\mathrm{Var}(X)}{3^2} = \frac{1}{9} \approx 0.11.$$

3. 根据 Chernoff 不等式，使用一个优化参数 $t > 0$：

$$P(|X| \geq 3) = 2P(X \geq 3) \leq 2\frac{\mathrm{E}(e^{tX})}{e^{3t}} = 2e^{\frac{t^2}{2}-3t}.$$

令 $t = 3$ 最小化上界，有

$$P(|X| \geq 3) \leq 2e^{-\frac{9}{2}} \approx 0.022.$$

4. 根据经验法则（正态分布的性质）， $P(|X| \geq 3) \approx 0.003$.

这个例子清楚地展示了三个不等式的估计精度：切尔诺夫不等式提供了最紧的界限，而马尔可夫不等式的估计最为松散. 这也说明了高阶矩信息在概率估计中的重要性.

5.2 大数定律

大数定律（Law of Large Numbers，简记为 LLN）是概率论中最基本也是最重要的定理之一，从数学上说明了随机现象在大量重复试验中所呈现的统计规律性.

设随机变量 $X_1, \ldots, X_n$ 独立同分布，期望为 $\mathrm{E}(X_i) = \mu$，方差为 $\mathrm{Var}(X_i) = \sigma^2 > 0$，则对于样本均值 $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$，有 $\mathrm{E}(\bar{X}) = \mu$，且 $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n} \to 0\,(n \to +\infty)$. 这表明随着样本量增大，样本均值的波动会越来越小.

定理 5.4 (Khinchin 弱大数定律). 设随机变量序列 $\{X_i\}_{i=1}^{n}$ 独立同分布，且 $\mathrm{E}(X_i) = \mu$，$\mathrm{Var}(X_i) = \sigma^2 > 0$，则对任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0,$$

或等价地，

$$\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) = 1.$$

证明. 由 Chebyshev 不等式可得

$$\begin{aligned} P(|\bar{X} - \mu| \geq \epsilon) &\leq \frac{\mathrm{Var}(\bar{X})}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0 \quad (n \to \infty). \end{aligned}$$

□

弱大数定律的一个重要应用是参数估计. 对任给精度 $\epsilon > 0$ 和置信度 $1 - \alpha$ $(\alpha > 0)$，存在样本量阈值 $N \in \mathbb{N}^+$，使得当 $n \geq N$ 时，

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \alpha.$$

即样本容量充分大时样本均值以高概率落入真值邻域.

通过更进一步的讨论可以证明（内容超出本讲义范围），上述定理中关于方差有限的条件可以去掉，结论仍正确. 当 $X_i \sim B(p)$ 时，即为 *Bernoulli 大数定律*. 此外，还有其他形式的弱大数定律，如

1. 要求 X_i 两两不相关，$\mathrm{Var}(X_i)$ 一致有界，得到 Chebyshev 弱大数定律；
2. 要求 $\mathrm{Var}(\bar{X}) \to 0\,(n \to +\infty)$，得到 Markov 弱大数定律.

定义 5.1 (依概率收敛). 称随机变量序列 $\{Y_n\}$ 依概率收敛于 Y（记作 $Y_n \xrightarrow{P} Y$），若对任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0.$$

利用依概率收敛的概念，弱大数定律结论可以重新表述为 $\bar{X} \xrightarrow{P} \mu$.

例 5.2 (样本方差的收敛性). 设 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ 为样本方差，则

$$S^2 \xrightarrow{P} \sigma^2.$$

证明. 关键步骤：

1. 展开差平方和：

$$S^2 - \sigma^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum(X_i-\mu)^2 - \sigma^2\right] - \frac{n}{n-1}(\bar{X}-\mu)^2 - \frac{\sigma^2}{n-1}.$$

2. 应用弱大数定律及依概率收敛的运算性质即得证.

□

定理 5.5 (Kolmogorov 强大数定律). 设 $\{X_i\}$ 独立同分布，$\mathrm{E}(X_i) = \mu$，则

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

定理的证明这里略去，当 X_i 的四阶矩有限时留给读者作为练习.

对 $X_i \sim B(p)$，样本均值（频率）满足

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

这为概率的频率解释提供了理论基础.

定义 5.2 (几乎必然收敛). 称随机变量序列 $\{Y_n\}$ 几乎必然收敛于 Y（记作 $Y_n \xrightarrow{a.s.} Y$），若

$$P\left(\lim_{n\to\infty} Y_n = Y\right) = 1$$

使用上述定义，强大数定律结论可以重新表述为 $\bar{X} \xrightarrow{a.s.} \mu$.

例 5.3 (Monte Carlo 积分). 计算 $\displaystyle\int_a^b g(x)\,dx$ $(g > 0)$：

1. 取 $c > \sup\limits_{x\in[a,b]} g(x)$.
2. 生成 $(X_i, Y_i) \sim U([a,b]\times[0,c])$.
3. 定义 $I_i = \mathbf{1}_{\{Y_i \le g(X_i)\}}$.
4. 由大数定律，

$$\int_a^b g(x)\,dx \approx c(b-a)\cdot\frac{1}{n}\sum_{i=1}^{n} I_i.$$

例 5.4 (收敛性质对比). 设 $\Omega=[0,1]$，在 Ω 上取均匀分布. 定义随机变量序列如下：$\forall\omega\in\Omega$，

$$Y_1(\omega)=\omega+I_{[0,1]}(\omega),\quad Y_2(\omega)=\omega+I_{[0,1/2]}(\omega),\quad Y_3(\omega)=\omega+I_{[1/2,1]}(\omega),$$

$$Y_4(\omega)=\omega+I_{[0,1/3]}(\omega),\quad Y_5(\omega)=\omega+I_{[1/3,2/3]}(\omega),\quad Y_6(\omega)=\omega+I_{[2/3,1]}(\omega),\quad \cdots.$$

则

$$Y_n\xrightarrow{P}\omega \text{ 成立，但 } Y_n\xrightarrow{a.s.}\omega \text{ 不成立.}$$

事实上，可以证明几乎必然收敛蕴含依概率收敛.

5.3 中心极限定理

中心极限定理（Central Limit Theorem, 简记为 CLT）是概率论中最重要的定理之一，最初由法国数学家 De Moivre 在 1733 年研究二项分布的极限性质时发现，后经 Laplace、Lindeberg 和 Lyapunov 等人进一步发展将这一理论逐渐完善.

定理 5.6 (Lindeberg-Lévy 中心极限定理). 设随机变量 $X_1,X_2,\ldots,X_n$ 独立同分布，均值 $\mathrm{E}(X_i)=\mu$ 且方差 $\mathrm{Var}(X_i)=\sigma^2>0$. 则对于任意 $x\in\mathbb{R}$，有

$$\lim_{n\to+\infty}P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\le x\right)=\Phi(x),$$

其中 $\bar{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$ 为样本均值，$\Phi(x)$ 为标准正态分布的累积分布函数. 此外，等价地有

$$\lim_{n\to+\infty}P\left(\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}\le x\right)=\Phi(x).$$

证明. 这里仅在矩母函数存在的情形下给出证明. 不失一般性，假设随机变量 X_i 的均值 $\mu=0$ 且方差 $\sigma^2=1$. 考察标准化和的矩母函数

$$\mathrm{E}\left(e^{t\frac{X_1+\cdots+X_n}{\sqrt{n}}}\right)=M^n\left(\frac{t}{\sqrt{n}}\right),$$

其中 $M(t)=\mathrm{E}(e^{tX_i})$ 为单个随机变量 X_i 的矩母函数，满足 $M(0)=1,M'(0)=0,M''(0)=1$.

通过 Taylor 展开，对矩母函数 $M(t)$ 进行展开得：

$$M\left(\frac{t}{\sqrt{n}}\right)=1+0+\frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2+o\left(\frac{t^2}{n}\right),$$

将其代入并考虑乘积的渐近行为，有

$$M^n\left(\frac{t}{\sqrt{n}}\right)=\left(1+\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n\to e^{t^2/2}\quad(n\to+\infty).$$

这正是标准正态分布 $N(0,1)$ 的矩母函数，表明 $\dfrac{X_1+\cdots+X_n}{\sqrt{n}}$ 的分布趋近于 $N(0,1)$. □

该定理揭示了一个深刻的统计现象：大量独立同分布随机变量的均值或加和会近似服从正态分布. 这一发现为统计推断提供了重要理论基础. 具体地，根据中心极限定理，当 n 足够大时，样本均值 $\bar{X}$ 的分布可以近似为正态分布：

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

类似地，样本和 $X_1 + X_2 + \cdots + X_n$ 的分布可以近似为为正态分布：

$$X_1 + X_2 + \cdots + X_n \overset{\text{近似}}{\sim} N(n\mu, n\sigma^2).$$

例 5.5 (De Moivre-Laplace 中心极限定理). 设 $X_i \sim B(p)$，则 $\sum_{i=1}^{n} X_i \sim B(n,p)$. 当 n 足够大时，二项分布可以近似为正态分布：

$$\sum_{i=1}^{n} X_i \overset{\text{近似}}{\sim} N(np, np(1-p)).$$

因此，关于 $\sum_{i=1}^{n} X_i$ 的取值概率可以用正态分布近似计算，例如

$$P(t_1 \le \sum_{i=1}^{n} X_i \le t_2) \approx \Phi(y_2) - \Phi(y_1),$$

其中

$$y_1 = \frac{t_1 - np - \frac{1}{2}}{\sqrt{np(1-p)}}, \quad y_2 = \frac{t_2 - np + \frac{1}{2}}{\sqrt{np(1-p)}},$$

这里的 $\frac{1}{2}$ 是连续性修正项，用以提高近似精度.

例 5.6 (选举问题). 设 p 表示选民的真实支持度（未知），通过随机抽样调查 n 人（假设 $n \ll N$，可以近似为有放回抽样），得到样本支持比例

$$P_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$

其中 X_i 为第 i 个被调查者的态度，1 表示支持，0 表示不支持，则 $X_i \sim B(p)$（$i = 1, 2, \ldots, n$）且近似相互独立. 设定精度 $\epsilon = 0.03$，置信度 $1 - \alpha = 95\%$，要求确定最小样本量 n，使得

$$P(|P_n - p| < \epsilon) \ge 1 - \alpha.$$

根据中心极限定理，有

$$P(|P_n - p| \ge \epsilon) \approx 2\left(1 - \Phi\left(\frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}}\right)\right) \le \alpha,$$

从而得到

$$n \geq \frac{z_{\alpha/2}^2 p(1-p)}{\epsilon^2},$$

其中 $z_{\alpha/2}$ 为标准正态分布的上 $\frac{\alpha}{2}$ 分位数. 保守估计对 $p(1-p)$ 取其最大值，此时对应 $p=0.5$，有

$$n \geq \frac{z_{\alpha/2}^2}{4\epsilon^2}.$$

代入 $\epsilon = 0.03, \alpha = 0.05$，得到 $n \geq 1068$. 值得注意的是，这一结果与总人数 N 无关!

定义 5.3 (依分布收敛). 称随机变量序列 $\{Y_n\}$ **依分布收敛**于随机变量 Y （记作 $Y_n \xrightarrow{d} Y$），如果对于任意实数 x，当 $n \to \infty$ 时，随机变量 Y_n 的分布函数 $F_{Y_n}(x)$ 满足

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

结合该定义，中心极限定理可以重新表述为：

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, \quad \text{其中 } Z \sim N(0,1),$$

简记为

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

可以证明，依概率收敛蕴含依分布收敛.

第二部分

统计推断

统计引言

统计学是一门从数据中获得信息的学问. 根据 Claude Shannon 的信息论，所谓的信息是不确定性的分解和减少. 通过获取数据与分析数据，逐步降低对未知现象的不确定性，从而获得有价值的知识. 这种不确定性的减少过程，本质上就是统计学.

数理统计通常包括以下三部分：

- 数据收集：涉及实验设计、抽样方案制定等，是获得可靠数据的关键.
- 数据分析：借助统计模型，通过描述性统计和探索性分析揭示数据特征.
- 统计推断：从样本信息推断总体特征，是统计学的重要目标，如参数估计与假设检验.

例. 假设需要检测某厂的一大批电子元件产品的寿命，关注的问题是“判断这批产品是否合格”. 此问题的“总体”即为这批元件的寿命，更具体地说，是元件寿命这一随机变量 X 的分布. 这个案例体现了从数据收集（抽检元件）到数据分析（计算统计量）再到统计推断（判断是否合格）的完整过程.

统计学中的总体（Population）指的是一个概率分布. 研究对象的全体（随机变量 X）所服从的分布即为总体分布. 统计分析的核心目标是通过研究总体变量 X 的某些数字特征，来了解总体分布的性质.

总体可以进一步细分为有限总体与无限总体. 在个体数量足够多时，一个有限总体可以近似看作无限总体. 而所谓虚拟总体是一种假想的无限总体，其实际上并不存在具体的个体集合. 例如，同一生产工艺下可能生产的所有产品，某种实验条件下可能出现的所有实验结果等，都可以视为虚拟总体.

统计模型则指代一族概率分布，例如正态分布族

$$\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\},$$

这是一个典型的参数模型（Parametric Model），因为它可以通过少量参数（如 μ 与 σ^2）来完全刻画相应的概率分布. 与此对应的是非参数模型，这类模型无法用有限数量的参数完整表征分布族，例如，仅假定随机变量 X 是连续的且其期望 $\mathrm{E}(X)$ 存在，但对其具体分布形态不做进一步限制，此时所讨论的分布族就是一个非参数模型.

样本是指从总体中抽取的一组观测值，记为 $X_1, \ldots, X_n$，其中每个 X_i 源自总体 X，而 n 称为样本容量. 样本的获取通常通过随机试验或观测进行. 在实际应用中，完整的样本数据并不总能获得，部分信息缺失或未能完整记录，这种情形被称为不完全观测.

简单随机抽样是指从一个总体（总体大小为 N）中，无放回地随机抽取 n 个个体组成样本，使得所有可能的、容量为 n 的样本出现的概率相等. 对于每一个样本，抽中的概率为

$$P = \frac{1}{\binom{N}{n}}.$$

若 $X_1, \ldots, X_n$ 独立同分布（Independent and Identically Distributed，简记为 IID），且 $X_i \sim X$，则称 $X_1, \ldots, X_n$ 为来自总体 X 的一个随机样本.

统计量是样本的函数，记作 $T(X_1, \ldots, X_n)$，其完全由样本决定. 例如，样本均值 $\bar{X}$ 与样本方差 S^2 就是最常用的统计量. 统计量是对数据进行简化的方式，本质上是一个随机变量，可以用来推断总体的特性. 一个好的统计量应该尽可能包含样本中关于总体的信息.

统计推断的核心目标是通过样本信息推断总体特性，基本内容包括参数估计和假设检验两个方面.

例. 假设元件寿命服从指数分布，即 $X \sim \mathrm{Exp}(\lambda)$.

1. 如何通过样本数据估计 λ？这属于参数估计问题；
2. 假设这批元件的合格标准是 $\mathrm{E}(X) \geq L$，而实际 $\mathrm{E}(X)$ 未知，因此需制定一个可操作的检验标准. 可以考虑设定一个临界值 l，当样本均值 $\bar{X} \geq l$ 时，即认为这批元件合格，该临界值 l 的确定可通过假设检验进行讨论.

例. 设模型为 $Y = aX + \epsilon$，其中 X 为自变量，Y 为因变量，而 ϵ 表示误差项. 考虑样本 $(X_1, Y_1), \ldots, (X_n, Y_n)$.

- 若参数 a 未知，则可通过样本对 a 进行估计，此为参数估计在模型推断方面的应用；
- 若参数 a 已知（或已被估计），则可利用观测到的 Y_i 对相应的 X_i 进行推断，这是关于变量推断的模型应用.

例 5.7 (简单随机抽样)**.** 设总体大小为 N, 总体均值和方差分别为 μ, σ^2. 令 X_i $(i = 1, \ldots, n)$ 为从总体中无放回抽取的简单随机样本.

- 由于简单随机抽样的每个个体被选中的概率均等为 $\dfrac{1}{N}$，有

$$E(X_i) = \sum_{j=1}^{N} x_j \cdot \frac{1}{N} = \mu.$$

同理，

$$\mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{1}{N}\sum_{j=1}^{N} x_j^2 - \mu^2 = \sigma^2.$$

- 利用线性性质可得

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

考虑样本间的协方差关系，

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j).$$

当 $i = j$ 时，$\mathrm{Cov}(X_i, X_i) = \sigma^2$；当 $i \neq j$ 时，由于无放回抽样的性质，

$$\mathrm{Cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}.$$

代入整理得

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\left(-\frac{\sigma^2}{N-1}\right)\right] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

上述方差公式中的因子 $\dfrac{N-n}{N-1}$ 通常称为有限总体修正因子，当 $N \to \infty$ 或 $n \ll N$ 时，该因子趋近于 1，这时的抽样便可近似视为有放回抽样.

第六章　参数估计

6.1　矩估计

设 $X_1, \ldots, X_n$ 为独立同分布的随机样本，定义样本矩如下：

1. k 阶样本原点矩 $\hat{\mu}_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k$
2. k 阶样本中心矩 $\hat{m}_k := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$

根据大数定律，

$$\hat{\mu}_k \xrightarrow{P} \mu_k = \mathrm{E}(X^k), \quad n \to \infty.$$

类似地，样本中心矩收敛于相应的总体中心矩. 矩估计是利用样本矩来估计分布参数的方法，其基本思想是将总体矩表示为参数的函数，然后用相应的样本矩替代，求解参数的估计值.

例 6.1. 设 $X_1, \ldots, X_n$ 是来自正态分布 $N(\mu, \sigma^2)$ 的独立同分布样本，则

$$\mu = \mathbb{E}(X) \approx \hat{\mu}_1 = \bar{X}, \quad \sigma^2 = \mathrm{Var}(X) \approx \hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

上述矩估计中，$\bar{X}$ 是 μ 的估计，$\hat{m}_2$ 是 σ^2 的估计，为常用的估计总体（不一定为正态分布）均值和方差的方法.

例 6.2. 设 $X_1, \ldots, X_n$ 是来自指数分布 $Exp(\lambda)$ 的独立同分布样本，则根据指数分布的性质，

$$\lambda = \mathbb{E}(X)^{-1} \approx \hat{\mu}_1^{-1} = \frac{1}{\bar{X}},$$

或

$$\lambda = (\mathrm{Var}(X))^{-1/2} \approx \hat{m}_2^{-1/2}.$$

在此实例中，参数 λ 可以通过两种不同的矩估计方法得出，其中基于 $\hat{\mu}_1$ 的估计（低阶矩）更为简单且常用.

由指数分布的例子可以看出，对于同一个参数可能存在多种不同的矩估计. 在实际应用中，通常优先使用低阶矩的估计，因为低阶矩通常具有更小的抽样误差和更高的估计效率. 矩估计虽然在计算上简便，但并不总是最优的. 在某些情况下，极大似然估计等其他方法可能提供更有效的估计量.

6.2 极大似然估计

极大似然估计（Maximum Likelihood Estimation，简记为 MLE）是参数估计的基本方法之一，其思想是在给定观测数据的条件下，选择使得观测数据出现概率或概率密度最大的参数作为未知参数的估计值. 以下详细介绍极大似然估计的定义、性质及其应用.

设随机向量 $(X_1,\dots,X_n)$ 的联合分布（概率质量函数或概率密度函数）为 $f(x_1,\dots,x_n;\theta)$，其中 θ 是未知参数. 观测到的具体样本数据记为 $(x_1,\dots,x_n)$.

定义 6.1 (似然函数). *似然函数*（likelihood function）定义为

$$L(\theta)=f(X_1,\dots,X_n;\theta).$$

在离散情形下，$L(\theta)$ 即当参数为 θ 时出现观测 $(X_1,\dots,X_n)$ 的概率.

如果 $X_1,\dots,X_n$ 是来自同一总体的随机样本，则 $X_1,\dots,X_n$ 独立同分布. 假设总体分布为 $f_1(x;\theta)$，则联合分布为

$$f(x_1,\dots,x_n;\theta)=\prod_{i=1}^{n}f_1(x_i;\theta),$$

从而似然函数为

$$L(\theta)=\prod_{i=1}^{n}f_1(X_i;\theta).$$

例 6.3. 设 $X_1,\dots,X_n$ 为来自正态总体 $N(\mu,\sigma^2)$ 的随机样本，其中 μ 和 σ^2 为未知参数. 此时 $(X_1,\dots,X_n)$ 的联合概率密度函数为

$$f(x_1,\dots,x_n;\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

对应的似然函数为

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}.$$

定义 6.2. *极大似然估计*（MLE）定义为

$$\theta^*=\underset{\theta}{\operatorname{argmax}}L(\theta),$$

即选取使似然函数达到最大值的参数作为估计值.

由定义可知 $\theta^* = \theta^*(X_1, \ldots, X_n)$ 是样本 $X_1, \ldots, X_n$ 的函数，因此它本身也是一个随机变量，通常称为参数 θ 的极大似然估计量.

例 6.4. 以上述正态分布的例子为例，计算 μ 和 σ^2 的极大似然估计. 为方便计算，考虑相应的对数似然函数，即

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

分别对 μ 和 σ^2 求导并令其等于零，得到似然方程组：

$$\begin{aligned}
\frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0, \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.
\end{aligned}$$

解得

$$\mu^* = \bar{X}, \quad (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

易验证， $\mu^*, (\sigma^2)^*$ 分别为 μ 和 σ^2 的极大似然估计.

命题 6.1 (极大似然估计不变性)**.** 若 θ^* 是 θ 的极大似然估计，则 $g(\theta^*)$ 是 $g(\theta)$ 的极大似然估计，其中 g 为任意可测函数.

例 6.5. 设 $X_1, \ldots, X_n$ 相互独立，且均来自区间 $(0, \theta]$ 上的均匀分布总体，其中 $\theta > 0$ 为未知参数. 由均匀分布的概率密度函数可知，

$$f(x; \theta) = \frac{1}{\theta} \cdot \mathbf{1}_{\{0 < x \le \theta\}},$$

故似然函数为

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & 0 < \max\{X_1, \ldots, X_n\} \le \theta, \\ 0, & \text{其他}. \end{cases}$$

最大化 $L(\theta)$ 可得 θ 极大似然估计为

$$\theta^* = \max\{X_1, \ldots, X_n\}.$$

例 6.6. 考虑随机样本 $X_1, \ldots, X_n$，其总体为 Cauchy 分布，概率密度函数为

$$f(x; \theta) = \frac{1}{\pi\left[1 + (x - \theta)^2\right]}, \quad x \in \mathbb{R}.$$

其中 θ 为未知参数.

- 由于 Cauchy 分布的各阶矩均不存在，所以矩估计方法无法使用.
- 考虑极大似然估计方法. 似然函数为

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\pi\left[1 + (X_i - \theta)^2\right]}.$$

对数似然函数为

$$\log L(\theta) = -n \log \pi - \sum_{i=1}^{n} \log\left[1 + (X_i - \theta)^2\right].$$

计算对 θ 的偏导，得到似然方程

$$\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0.$$

该方程通常不存在显式解，而且当样本量较大时可能出现多个解，导致计算复杂，即使是采用数值迭代法求解.

- 使用样本中位数作为参数 θ 的估计是一种合理方案，因为此例中 Cauchy 分布的中位数即为 θ.

这一例子说明，参数估计方法的选择应根据具体问题和分布特性而定，方法不是唯一的，也没有绝对的优劣. 需要指出，极大似然估计不一定是唯一的；极大似然估计需要分布的具体函数形式，而矩估计不需要. 此外，如果似然函数在最大值点附近变化过于平缓，则可能不利于通过迭代等方法有效计算.

6.3 优良性准则

无论是矩估计还是极大似然估计，都是利用样本的函数来对总体参数进行估计. 根据具体的样本数据，这些方法能够为每个待估参数提供一个确定的数值，而非一个区间或分布，这种直接给出参数单一估计值的方法称为*点估计*.

用于估计参数的函数 $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ 称为*估计量*. 其分布（依赖于 θ）称为*抽样分布*，而其标准差 $\sqrt{\mathrm{Var}(\hat{\theta})}$ 称为*标准误（差）*（Standard Error），记作 $\mathrm{Se}(\hat{\theta})$. 在实际研究或应用中，通常根据不同准则来选择合适的估计量. 下面首先介绍最基本的*无偏性*准则.

定义 6.3. 设 $\hat{\theta}$ 是 θ 的估计量，称 $\mathrm{E}(\hat{\theta} - \theta) = \mathrm{E}(\hat{\theta}) - \theta$ 为 $\hat{\theta}$ 的*偏差*（Bias）. 若对于参数 θ 的任意可能取值，总有 $\mathrm{E}(\hat{\theta} - \theta) = 0$，则称 $\hat{\theta}$ 为 θ 的一个*无偏估计（量）*.

由上述定义可知，无偏性指的是估计量的期望值等于被估计参数的真值，即不存在系统性偏差. 一般地，若 $\hat{g}(X_1, \ldots, X_n)$ 是对 θ 的函数 $g(\theta)$ 的估计，且对于参数 θ 的任意可能取

值，总有 $\mathrm{E}(\hat{g}(X_1,\ldots,X_n)) = g(\theta)$，则称 $\hat{g}(X_1,\ldots,X_n)$ 是 $g(\theta)$ 的一个无偏估计. 对于无偏估计 $\hat{g}(X_1,\ldots,X_n)$，若进行 N 组独立抽样，第 m 组样本记作 $X_1^{(m)},\ldots,X_n^{(m)}$，则由大数定律可知，

$$\frac{1}{N}\sum_{m=1}^{N}\hat{g}(X_1^{(m)},\ldots,X_n^{(m)}) \xrightarrow{a.s.} \mathrm{E}(\hat{g}(\theta)) = g(\theta).$$

在应用中，估计量是否无偏的重要性要视具体情形而定.

例 6.7. 若总体的均值 μ 和方差 σ^2 均未知，则由 $\mathrm{E}(\bar{X}) = \mu$ 知 $\bar{X}$ 是 μ 的无偏估计. 而二阶矩

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2,$$

其期望满足 $\mathrm{E}(m_2) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2$，因此 m_2 不是 σ^2 的无偏估计（估计值系统性偏小）. 样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}m_2,$$

此时 $\mathrm{E}(S^2) = \sigma^2$，故 S^2 是 σ^2 的无偏估计，其中 $\dfrac{n}{n-1}$ 被称为*无偏修正因子*.

例 6.8. 若总体为均匀分布 $X \sim U(0,\theta)$，参数 $\theta > 0$ 未知，则矩估计 $\hat{\theta} = 2\bar{X}$ 为 θ 的无偏估计. 而极大似然估计为

$$\theta^* = \max\{X_1,\ldots,X_n\},$$

有 $\mathrm{E}(\theta^*) = \dfrac{n}{n+1}\theta$，因此 θ^* 不是 θ 的无偏估计.

这些例子表明，极大似然估计并不一定是无偏的.

下面介绍另一个常用的均方误差准则. 定义*均方误差*（MSE）为

$$\mathrm{MSE}(\hat{\theta}) := \mathrm{E}\big[(\hat{\theta}-\theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \big[\mathrm{E}(\hat{\theta}-\theta)\big]^2.$$

其中，方差项 $\mathrm{Var}(\hat{\theta})$ 体现估计方法的*精确度*（Precision），而偏差平方项 $\big[\mathrm{E}(\hat{\theta}-\theta)\big]^2$ 体现估计方法的*准确度*（Accuracy）.

定义 6.4 (均方误差准则)**.** 设 $\hat{\theta}_1, \hat{\theta}_2$ 均为 θ 的无偏估计. 若对于 θ 的任意可能取值都有

$$\mathrm{Var}(\hat{\theta}_1) \leq \mathrm{Var}(\hat{\theta}_2),$$

且至少在某个 θ 值处严格不等号成立，则称在均方误差意义下，$\hat{\theta}_1$ 优于 $\hat{\theta}_2$.

例 6.9. 若总体的均值 μ 未知，方差为 σ^2，则易见 $\bar{X}$，$\dfrac{1}{2}(X_1+X_2)$ 以及 X_1 都是 μ 的无偏估计，其方差分别为 $\dfrac{\sigma^2}{n}$，$\dfrac{\sigma^2}{2}$ 与 σ^2. 若 $n > 2$，显然

$$\mathrm{Var}(\bar{X}) < \mathrm{Var}\left(\frac{1}{2}(X_1+X_2)\right) < \mathrm{Var}(X_1),$$

因此在均方误差意义下，$\bar{X}$ 优于 $\frac{1}{2}(X_1+X_2)$，而后者又优于 X_1.

定义 6.5. 若 $\hat{\theta}_0$ 是 θ 的无偏估计，并且对所有无偏估计 $\hat{\theta}$，对于 θ 的任意可能取值，都有

$$\mathrm{Var}(\hat{\theta}_0) \le \mathrm{Var}(\hat{\theta}),$$

则称 $\hat{\theta}_0$ 为 θ 的*最小方差无偏估计*（MVUE）.

例 6.10. 若 $X \sim N(\mu, \sigma^2)$，则 $\mathrm{E}(m_2) = \frac{n-1}{n}\sigma^2$，$\mathrm{E}(S^2) = \sigma^2$，而

$$\mathrm{E}((m_2-\sigma^2)^2) < \mathrm{E}((S^2-\sigma^2)^2).$$

此例表明，尽管 m_2 带有偏差，但其均方误差反而更小，一定程度上体现了偏差与方差的某种权衡.

接下来介绍估计量的大样本性质，即指样本容量 n 趋于无穷时估计量的渐近行为和特性. 在难以获得估计量精确分布的情况下，大样本性质通常成为理论分析和实践应用的核心工具.

（1）**渐近无偏性** 若对于参数 θ 的任意可能取值都有

$$\lim_{n\to+\infty} \mathrm{E}(\hat{\theta}-\theta) = 0,$$

则称估计量 $\hat{\theta}$ 具有*渐近无偏性*.

（2）**相合性** 若对任意 $\epsilon > 0$，对于参数 θ 的任意可能取值都有

$$\lim_{n\to+\infty} P(|\hat{\theta}-\theta| \ge \epsilon) = 0,$$

则称 $\hat{\theta}$ 是 θ 的*相合估计*.

根据定义，$\hat{\theta}$ 是 θ 的相合估计当且仅当 $\hat{\theta} \xrightarrow{P} \theta$，即 $\hat{\theta}$ 依概率收敛于 θ. 相合性衡量点估计是否“随样本量增加而趋于参数真实值”，是良好点估计的一项自然要求. 例如，根据弱大数定律，样本均值 $\bar{X}$ 是总体均值 μ 的相合估计. 相合性也称为一致性.

例 6.11. 若总体的均值为 μ，方差为 σ^2，则对于

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\bar{X}-\mu)^2,$$

由大数定律可得

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \xrightarrow{P} \sigma^2 \quad \text{且} \quad (\bar{X}-\mu)^2 \xrightarrow{P} 0,$$

故

$$m_2 \xrightarrow{P} \sigma^2,$$

即 m_2 是 σ^2 的相合估计. 同理，

$$S^2 = \frac{n}{n-1} m_2 \xrightarrow{P} \sigma^2,$$

因此 S^2 也是 σ^2 的相合估计.

（3）**渐近正态性** 若对于 θ 的任意可能取值都有

$$\frac{\hat{\theta} - \theta}{\mathrm{Se}(\hat{\theta})} \xrightarrow{d} Z \sim N(0,1),$$

则称 $\hat{\theta}$ 是 θ 的*渐近正态估计*.

例如，由中心极限定理可知，样本均值 $\bar{X}$ 是总体均值 μ 的渐近正态估计，且

$$\mathrm{Se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

渐近正态性为处理复杂统计模型提供了简便而强大的工具，尤其在参数估计量的精确分布难以求解的情形下，是构建参数的置信区间和假设检验的重要基础. 当样本容量 n 充分大时，可以认为

$$\hat{\theta} \overset{\text{近似}}{\sim} N\left(\theta, \mathrm{Se}^2(\hat{\theta})\right).$$

这一性质使得在大样本情况下能够有效地量化估计量的误差范围并进行可靠的统计推断.

6.4 置信区间

定义 6.6. 给定 $\alpha \in (0,1)$，设 $\hat{\theta}_i = \hat{\theta}_i(X_1, \ldots, X_n)$（$i = 1, 2$）为统计量，若

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1 - \alpha,$$

则称区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 为参数 θ 的一个 $(1-\alpha)$-置信的*（双侧）区间估计*. $(1-\alpha)$ 称为*置信水平*. $\hat{\theta}_1, \hat{\theta}_2$ 分别称为*置信上、下限*.

通常用*置信系数*或*置信度*表示所有可能置信水平中的最大者. 需特别注意，这三个术语都是针对估计方法而言的，是对方法可靠性的刻画. 在实际应用中，α 通常取 0.05，0.01 或 0.1 等较小的正值，分别对应 95%，99% 或 90% 的置信水平.

区间估计的精度通常可用区间长度的期望 $\mathrm{E}(\hat{\theta}_2 - \hat{\theta}_1)$ 来衡量. 统计推断中遵循可靠度优先原则，即先保证置信水平达到要求，然后再尽量提升精度（缩小区间期望长度）.

例 6.12. 设 $X_1, \ldots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的随机样本，其中 μ 未知，但 σ^2 已知.

为给出 μ 的区间估计，目标是找到适当的常数 $c_1, c_2 > 0$，使得

$$P(\bar{X} - c_1 < \mu < \bar{X} + c_2) \geq 1 - \alpha,$$

这等价于

$$P(-c_2 < \bar{X} - \mu < c_1) \geq 1 - \alpha.$$

容易看出，$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$. 记 $\alpha_1 = P(\bar{X} - \mu \leq -c_2)$ 和 $\alpha_2 = P(\bar{X} - \mu \geq c_1)$，考虑到正态分布的对称性，一个自然且最优的选择是令 $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$（此时可证明区间长度的期望最小）. 记 $z_{\alpha/2}$ 为标准正态分布的上 $\dfrac{\alpha}{2}$-分位数，即 $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. 则有

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right) = 1 - \alpha.$$

由此可得

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

因此，$\left(\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$ 是 μ 的一个 $(1-\alpha)$-置信的区间估计.

当 $\alpha = 0.05$ 时，$z_{\alpha/2} \approx 1.96 \approx 2$，这就是统计实践中常用的“$2\sigma$ 法则”的来源. 上述区间估计也可以这样解释：如果用样本均值 $\bar{X}$ 来估计 μ，则在 $(1-\alpha)$ 的置信水平下，绝对误差 $|\bar{X} - \mu|$ 不会超过 $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$. 注意到区间的半长度为 $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$. 若对精度有特定要求，比如给定 $\epsilon > 0$，要求

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \epsilon,$$

则必须满足

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{\epsilon}\right)^2.$$

这表明，当样本容量至少达到 $\left(\dfrac{z_{\alpha/2}\sigma}{\epsilon}\right)^2$ 时，可以在 $(1-\alpha)$-置信水平下保证绝对误差不超过 ϵ. 这揭示了置信水平 $1-\alpha$ 与精度要求 ϵ 以及样本容量 n 三者之间的重要关系，对实际抽样设计具有指导意义.

例 6.13. 设 $X_1, \ldots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的随机样本，其中 μ 和 σ^2 均未知.

首先需要估计 σ^2. 可以证明，

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 \sim \chi^2(n-1).$$

尽管卡方分布是非对称的，为适当简化，在牺牲一些区间精度要求的情况下，可同样令 $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$，得到

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

是 σ^2 的一个 $(1-\alpha)$-置信的区间估计，其中 $\chi^2_{\alpha/2}(n-1)$ 和 $\chi^2_{1-\alpha/2}(n-1)$ 分别为自由度为 $(n-1)$ 的卡方分布的上 $\dfrac{\alpha}{2}$-分位数和下 $\dfrac{\alpha}{2}$-分位数.

接下来估计 μ. 注意到 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$，且可证明其与 $\dfrac{(n-1)S^2}{\sigma^2}$ 相互独立，因此

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

故

$$\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

是 μ 的一个 $(1-\alpha)$-置信的区间估计，其中 $t_{\alpha/2}(n-1)$ 为自由度为 $(n-1)$ 的 t 分布的上 $\dfrac{\alpha}{2}$-分位数.

例 6.14 (两样本比较). 若 $X \sim N(\mu_1, \sigma^2)$, $Y \sim N(\mu_2, \sigma^2)$，且 X 与 Y 相互独立，μ_1, μ_2 以及 σ^2 均未知，下面估计均值差 $\mu_1 - \mu_2$.

设两个总体的随机样本分别为 $X_1, \ldots, X_n$ 和 $Y_1, \ldots, Y_m$，则

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right),$$

因此

$$\frac{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1).$$

同时，由

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1)$$

和

$$\frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{\sigma^2} = \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1),$$

以及 $\dfrac{(n-1)S_1^2}{\sigma^2}$ 与 $\dfrac{(m-1)S_2^2}{\sigma^2}$ 相互独立，可知

$$\frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2).$$

进一步可得

$$\frac{\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{\frac{(n-1)S_1^2}{\sigma^2}+\frac{(m-1)S_2^2}{\sigma^2}}{n+m-2}}} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2).$$

其中 $S^2 = \dfrac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ 是将两个样本方差合并后对两总体公共方差的估计，这里用到了 $\bar{X}-\bar{Y}$ 与 S^2 相互独立的事实. 因此

$$\left(\bar{X}-\bar{Y} - t_{\alpha/2}(n+m-2)S\sqrt{\frac{1}{n}+\frac{1}{m}},\ \bar{X}-\bar{Y} + t_{\alpha/2}(n+m-2)S\sqrt{\frac{1}{n}+\frac{1}{m}}\right)$$

是 $\mu_1 - \mu_2$ 的一个 $(1-\alpha)$-置信的区间估计.

当待估计参数的精确分布难以确定时，区间估计有相应的大样本方法，即所谓的**渐近置信区间**（Asymptotic Confidence Interval）. 这种方法基于统计量的渐近分布理论，特别是中心极限定理和各类极限定理，在样本量足够大时为参数提供了良好的区间估计近似.

例 6.15 (选举问题). 设 p 为未知的真实支持率，样本容量 $n = 1200$，调查中有 684 人表示支持，即支持率的观测比例为 $\dfrac{684}{1200} \approx 0.57$. 下面给出 p 的一个 95%-置信的区间估计.

根据例 5.6 中讨论，样本支持比例记为

$$P_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$

X_i 为第 i 个被调查者的态度， $X_i \sim \mathrm{B}(p)$（$i = 1, 2, \ldots, n$）且（近似）相互独立，且

$$\mathrm{E}(P_n) = p, \quad \mathrm{Var}(P_n) = \frac{p(1-p)}{n}.$$

根据中心极限定理，当 n 足够大时，有

$$\frac{P_n - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\text{近似}}{\sim} N(0,1).$$

由于 p 未知，所以分母中的标准误也未知，无法直接利用这一分布给出置信区间. 记 $\sigma^2 = p(1-p)$，下面采用几种不同方法给出其估计. 注意到，已知调查结果给出了 P_n 的一个观测值 $\dfrac{684}{1200}$，记为 p_n.

1. 用样本方差 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 估计 σ^2. 计算所给调查结果的样本方差为

$$s^2 = \frac{1}{n-1}\left[np_n(1-p_n)^2 + (n - np_n)p_n^2\right] \approx 0.2475,$$

于是有

$$\frac{P_n - p}{\sqrt{\frac{s^2}{n}}} \overset{\text{近似}}{\sim} N(0,1),$$

对应的置信区间为

$$\left(P_n - z_{\alpha/2}\sqrt{\frac{s^2}{n}}, P_n + z_{\alpha/2}\sqrt{\frac{s^2}{n}}\right) \approx (0.542, 0.598).$$

2. 用二阶中心矩 $m_2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = P_n(1-P_n)$ 估计 σ^2，于是有

$$\frac{P_n - p}{\sqrt{\frac{p_n(1-p_n)}{n}}} \overset{\text{近似}}{\sim} N(0,1),$$

对应的置信区间为

$$\left(P_n - z_{\alpha/2}\sqrt{\frac{p_n(1-p_n)}{n}}, P_n + z_{\alpha/2}\sqrt{\frac{p_n(1-p_n)}{n}}\right) \approx (0.542, 0.598).$$

3. 使用 $p(1-p)$ 的最大值 $\frac{1}{4}$ 来估计 σ^2（这种方法更为保守），于是有

$$\frac{P_n - p}{\sqrt{\frac{1}{4n}}} \overset{\text{近似}}{\sim} N(0,1),$$

对应的置信区间为

$$\left(P_n - z_{\alpha/2}\frac{1}{2\sqrt{n}}, P_n + z_{\alpha/2}\frac{1}{2\sqrt{n}}\right) \approx (0.542, 0.598).$$

需要注意的是，以上方法采用了近似分布，因此只能说置信水平近似是 $(1-\alpha)$，且近似的精确程度取决于总体分布的特征及样本容量 n 的大小.

极大似然估计具有良好的渐近性质，因此可以结合这些性质来构建参数的置信区间. 下面通过理论分析和具体示例，进一步说明如何利用极大似然估计构建置信区间.

设总体变量 X 的概率质量函数或概率密度函数为 $f(x;\theta)$，$X_1,\ldots,X_n$ 为来自该总体的随机样本，则其对应的似然函数和对数似然函数分别为

$$L(\theta) = \prod_{i=1}^{n} f(X_i;\theta), \quad \ell(\theta) := \log L(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta).$$

定理 6.1. 若 f 满足一定的光滑性条件，θ^* 为参数 θ 的极大似然估计，则存在 $\sigma_n > 0$，使得

$$\frac{\theta^* - \theta}{\sigma_n} \xrightarrow{d} Z \sim N(0,1).$$

定理的证明超出本书范围，此处从略. 根据定理，在样本量足够大时有 $\frac{\theta^* - \theta}{\sigma_n} \overset{\text{近似}}{\sim} N(0,1)$.

定义 6.7. 随机样本 $X_1,\ldots,X_n$ 的*Fisher* 信息量定义为

$$I_n(\theta) := \mathrm{E}\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right].$$

当 $n=1$ 时，$I_1(\theta)$ 为单个观测的 Fisher 信息量，记为 $I(\theta)$. 注意到

$$\frac{\partial \log f(X;\theta)}{\partial \theta} = \frac{f_\theta(X;\theta)}{f(X;\theta)},$$

其中 f_θ 表示 f 对 θ 的（偏）导数. 易见

$$\mathrm{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right) = 0,$$

直接验证可得

$$I_n(\theta) = \sum_{i=1}^{n} \mathrm{E}\left[\left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right)^2\right] + \sum_{i\neq j} \mathrm{E}\left[\frac{\partial \log f(X_i;\theta)}{\partial \theta} \cdot \frac{\partial \log f(X_j;\theta)}{\partial \theta}\right] = nI(\theta).$$

根据对数似然函数的 Taylor 展开，对于 θ^* 附近的 θ，有

$$0 = \ell'(\theta^*) = \ell'(\theta) + \ell''(\theta)(\theta^* - \theta) + o(\theta^* - \theta),$$

近似可得

$$\theta^* - \theta \approx -\frac{\ell'(\theta)}{\ell''(\theta)}.$$

进一步考察 $\ell'(\theta)$ 和 $\ell''(\theta)$ 的渐近性质：

- 记

$$Y_i = \frac{f_\theta(X_i;\theta)}{f(X_i;\theta)}, \quad i = 1, \dots, n.$$

则 $Y_1, \dots, Y_n$ 独立同分布，且 $\mathrm{E}(Y_i) = 0$，$\mathrm{Var}(Y_i) = I(\theta)$. 由

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{f_\theta(X_i;\theta)}{f(X_i;\theta)},$$

根据中心极限定理有

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \to N(0, I(\theta)).$$

- 注意到

$$\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left(\frac{f_\theta(X;\theta)}{f(X;\theta)}\right) = \frac{f_{\theta\theta}(X;\theta)}{f(X;\theta)} - \left(\frac{f_\theta(X;\theta)}{f(X;\theta)}\right)^2,$$

直接验证可得

$$\mathrm{E}\Big(\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}\Big) = -\mathrm{E}\Big[\Big(\frac{f_\theta(X_i;\theta)}{f(X_i;\theta)}\Big)^2\Big] = -I(\theta).$$

由

$$\ell''(\theta) = \sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2},$$

根据弱大数定律可得

$$-\tfrac{1}{n}\ell''(\theta) \xrightarrow{P} I(\theta).$$

综上所述，可得

$$\sqrt{n}(\theta^* - \theta) \xrightarrow{d} N\Big(0, \tfrac{1}{I(\theta)}\Big),$$

即

$$\frac{\theta^* - \theta}{\sqrt{\frac{1}{nI(\theta)}}} \xrightarrow{d} N(0,1).$$

利用上述极大似然估计量的渐近正态性结论可以构造参数 θ 的置信区间. 虽然实际中 $I(\theta)$ 一般不可知，但可用一致估计量 $I(\theta^*)$ 进行替代，因此有

$$\frac{\theta^* - \theta}{\sqrt{\frac{1}{nI(\theta^*)}}} \xrightarrow{d} N(0,1).$$

由此可得 θ 的 $(1-\alpha)$-置信区间为

$$\left(\theta^* - z_{\alpha/2}\sqrt{\frac{1}{nI(\theta^*)}},\ \theta^* + z_{\alpha/2}\sqrt{\frac{1}{nI(\theta^*)}}\right).$$

例 6.16 (选举问题). 假设条件同前例所述不变. 此时总体分布可表示为 $f(x;p) = p^x(1-p)^{1-x}$，相应的 Fisher 信息量为

$$I(p) = \mathrm{E}\left[\left(\frac{X-p}{p(1-p)}\right)^2\right] = \frac{1}{p(1-p)}.$$

参数 p 的极大似然估计为 $p^* = p_n = \dfrac{684}{1200}$，根据前面的讨论有

$$\frac{P_n - p}{\sqrt{\frac{1}{nI(p^*)}}} = \frac{P_n - p}{\sqrt{\frac{p_n(1-p_n)}{n}}} \overset{\text{近似}}{\sim} N(0,1),$$

而这与前例中用二阶中心矩 m_2 估计 σ^2 的结果完全一致.

例 6.17 (两总体比较). 设两总体 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$ 相互独立. 随机样本分别为 $X_1, \ldots, X_n$ 和 $Y_1, \ldots, Y_m$，样本均值分别为 $\bar{X}, \bar{Y}$，样本方差分别为 S_1^2, S_2^2. 根据联合正态分布的性质，有

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

将未知的 σ_1^2 和 σ_2^2 分别用其相合估计量 S_1^2 和 S_2^2 替代，得到近似分布

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \overset{\text{近似}}{\sim} N(0,1).$$

由此可得 $\mu_1 - \mu_2$ 的 $(1-\alpha)$-置信的区间估计为

$$\left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}},\ \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}\right).$$

6.5 Bayes 估计

Bayes 学派对于世界的认知视角与频率学派不同. 简单来说，在 Bayes 方法中，对未知参数 θ 的认识可由一个概率分布来刻画，设对应的随机变量为 Θ，则 θ 为随机变量 Θ 的实现值. 在搜集数据之前对 Θ 的分布认知 $f_\Theta(\theta)$ 称为*先验分布*. 将试验观测抽象为随机变量 X，当参数为 θ 时，观测数据的分布 $f_{X|\Theta}(x|\theta)$ 称为*样本分布*. 当观测到数据 x 后，可利用 Bayes 公式更新对 Θ 的认识，得到所谓的*后验分布*

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)}{f_X(x)}.$$

这样，就可以通过后验分布对 Θ 进行推断.

例 6.18. 某枚硬币正面向上的概率记为 θ，未知. 假设对硬币事先一无所知，可设先验分布为 $f_\Theta(\theta) = 1$ $(\theta \in [0,1]$（无信息先验，体现了所谓的“同等无知”原则）. 做抛硬币试验 n 次，记 X 为 n 次试验中正面向上的次数，则给定 θ 时，$X \sim B(n,\theta)$，即样本分布为

$$f_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0,1,\ldots,n.$$

于是 X 与 Θ 的联合分布为

$$f(x,\theta) = f_{X|\Theta}(x|\theta)\, f_\Theta(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

而 X 的边际分布为

$$f_X(x) = \int_0^1 f(x,\theta)\,\mathrm{d}\theta = \binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}\,\mathrm{d}\theta = \binom{n}{x}\frac{\Gamma(x+1)\,\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{1}{n+1}.$$

因此后验分布为

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)}{f_X(x)} = \frac{\Gamma(n+2)}{\Gamma(x+1)\,\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x},$$

由此可知，当 $X = x$ 时， $\Theta \sim \mathrm{Beta}(x+1, n-x+1)$.

其中，$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$ 为 Gamma 函数，对正整数 n，有 $\Gamma(n+1) = n!$. 而 $\mathrm{Beta}(a,b)$ 表示 *Beta* 分布，其概率密度函数为

$$f(t) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,t^{a-1}(1-t)^{b-1}, \quad t \in [0,1].$$

若 $T \sim \mathrm{Beta}(a,b)$，则

$$\mathrm{E}(T) = \frac{a}{a+b}, \quad \mathrm{Var}(T) = \frac{ab}{(a+b)^2(a+b+1)}.$$

均匀分布 $U(0,1)$ 即为 $\mathrm{Beta}(1,1)$.

在上述示例中，若 $n = 20$ 且正面向上的次数 $x = 13$，则后验分布为 $\mathrm{Beta}(14,8)$，其概率密度函数如图 6.1 所示. 计算可知，$P(\Theta > \frac{1}{2}) \approx 0.91$，而 $\Theta < \frac{1}{4}$ 的可能性则极小.

得到后验分布后，如何给出参数 θ 的合理估计呢？以下是几种常用的点估计方法：

1. 后验众数 $\hat{\theta}_1$，即后验分布 $\mathrm{Beta}(x+1, n-x+1)$ 的概率密度函数的最大值点，计算可得 $\hat{\theta}_1 = \dfrac{x}{n}$.（结果与极大似然估计相一致，原因是选取了无信息先验，此时后验分布正比于似然函数）.
2. 后验均值 $\hat{\theta}_2$，即后验分布的期望， $\hat{\theta}_2 = \mathrm{E}(\Theta \mid X = x) = \dfrac{x+1}{n+2}$.
3. 后验中位数 $\hat{\theta}_3$. （无法直接求闭式解，可通过数值方法或查表计算）

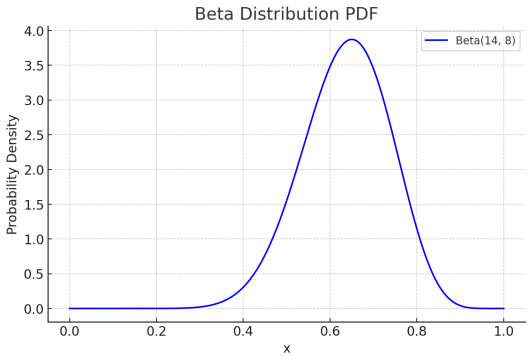

图 6.1: Beta$(14, 8)$ 的 PDF 图像

在该示例中还可进一步证明，若先验分布为 Beta(a, b)，则后验分布为 Beta$(x + a, n - x + b)$. 此时后验均值为

$$\frac{x+a}{n+a+b} = \frac{a+b}{n+a+b}\,\frac{a}{a+b} + \frac{n}{n+a+b}\,\frac{x}{n},$$

即后验均值是先验均值 $\dfrac{a}{a+b}$ 与样本均值 $\dfrac{x}{n}$ 的加权平均，权重分别为 $\dfrac{a+b}{n+a+b}$ 和 $\dfrac{n}{n+a+b}$.

Bayes 方法根据后验分布给出的区间估计称为**可信区间**. 具体而言，对于给定的 $\alpha \in (0, 1)$，确定实数 a, b 使得

$$P\big(a < \Theta < b \mid X = x\big) \geq 1 - \alpha.$$

这里 $1 - \alpha$ 称为**可信水平**. 常见构造方式有：

1. 最大后验区间，即区间内任意点的后验密度不低于区间外任意点的后验密度. （通常适用于单峰后验分布）
2. 等尾区间，即令 $P\big(\Theta < a \mid X = x\big) = P\big(\Theta > b \mid X = x\big) = \dfrac{\alpha}{2}$.

例 6.19. 设总体分布为 $X \sim N(\mu, \sigma^2)$，其中 σ^2 已知，μ 未知，$x_1, \ldots, x_n$ 为随机样本 $X_1, \ldots, X_n$ 的一个具体观测. 取 μ 的先验分布 $f(\mu) \propto 1$（无信息先验，实际上并不是真正的概率分布，可理解为广义概率密度函数），则样本分布为

$$f(x_1, \ldots, x_n \mid \mu) \propto \prod_{i=1}^{n} \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big).$$

根据 Bayes 公式，后验分布正比于先验分布与似然函数的乘积，即有

$$f(\mu \mid x_1, \ldots, x_n) \propto f(x_1, \ldots, x_n \mid \mu)\, f(\mu),$$

因此

$$f(\mu \mid x_1, \ldots, x_n) \propto \exp\left(-\frac{1}{2\sigma^2}\left[n\mu^2 - 2\mu \sum_{i=1}^{n} x_i\right]\right) \propto \exp\Big(-\frac{n(\mu - \bar{x})^2}{2\sigma^2}\Big).$$

这表明 μ 的后验分布为 $N\big(\bar{x}, \dfrac{\sigma^2}{n}\big)$. 由此可得 μ 的一个 $(1-\alpha)$-可信区间为

$$(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

由于正态分布是对称的单峰分布，所以该区间既是最大后验区间也是等尾区间. 值得注意的是，上述结果与经典方法结果相同，反映了在没有实质性先验信息的情况下，推断完全依赖于样本数据.

第七章　假设检验

7.1　基本概念

假设检验的现代框架形成于 20 世纪初. 1900 年 Karl Pearson 提出卡方检验，开创了基于统计量的假设验证方法；1925年 Ronald Fisher 系统化了显著性检验理论，引入 P 值作为衡量数据与假设不相容程度的指标；1930 年代 Jerzy Neyman 与 Egon Pearson（Karl Pearson之子）进一步构建了假设检验的决策理论框架，明确区分了原假设和备择假设，并引入了两类错误的概念，强调控制错误概率的重要性，同时提出了功效（power）的概念以优化检验设计. 20世纪 50 年代以后，随着计算技术的发展和应用统计学的普及，假设检验成为几乎所有实证研究领域的标准工具.

例 7.1. 在统计学史上著名的"女士品茶"实验中，Ronald Fisher 设计了一个精巧的随机对照研究. 某女士声称自己可以辨别奶茶的制作方法是先加奶还是先加茶. 为检验她的话是否为真，Ronald Fisher 设计了如下实验：分别用两种方法（先加奶和先加茶）各制作 4 杯奶茶，然后将 8 杯奶茶以随机顺序排列，让该女士进行盲测鉴别，并告知她两种制作方法的奶茶各有 4 杯. 实验结果是她全部判断正确. 据此能否能否合理地认为该女士确实具备这种鉴别能力呢？为回答这个问题，Fisher 引入了一个假设：

$$H_0:\ \text{该女士不具备鉴别能力},$$

即假设该女士完全是随机猜测. 那么，在 H_0 成立的前提下，该女士全部猜对的概率为

$$\frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} \approx 0.014.$$

根据小概率事件原理，即小概率事件在单次实验中不易发生，我们有充分理由怀疑假设 H_0 不成立，从而认为该女士确实具备鉴别能力.

这个经典案例不仅清晰展示了假设检验的基本流程，还体现了 Fisher 对实验随机化设计的重视. 实验的随机化处理（随机排序奶茶）确保了结果的可靠性，排除了其他干扰因素的影

响. 一个自然的问题是：概率要小到什么程度才算“小”？研究者通常需要结合实际情况设定一个阈值，称为显著性水平（Significance Level），通常用字母 α 表示. 常见的显著性水平选择为 $\alpha = 0.05$、 0.01 或 0.1 等. 若观测结果在假设 H_0 为真情况下的发生概率不超过 α，则可以认为该观测结果作为证据指向拒绝 H_0.

例如，上述品茶实验中，若该女士只猜对了 3 杯，则在 H_0 成立的前提下，猜对至少 3 杯的概率为

$$\frac{\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{16+1}{70} = \frac{17}{70} \approx 0.243.$$

这一概率是在假设 H_0 成立的情况下“出现观测结果以及更极端观测的概率”，称为P 值. 因为 $0.243 > \alpha$（若取 $\alpha = 0.05$ 或 $\alpha = 0.01$），所以不能拒绝假设 H_0，即这一观测结果作为证据不足以支持该女士具有鉴别能力.

Fisher 引入 P 值衡量观测结果与要检验的假设 H_0 所预期结果之间存在差异的程度. P 值越小表明差异越大，则观测结果与假设 H_0 越不匹配，反对假设 H_0 的理由就越充分. 当这种差异足够大时，认为观测结果是统计显著的（Statistically Significant），意味着样本中观察到的效应或差异不太可能仅仅是由随机变异引起的. 这种假设检验方法现在通常称为 *Fisher* 显著性检验. 值得注意的是，若认可某组样本观测，则用其来证实和证伪某个假设具有天然的不对等. 当 P 值不够小时，不能断言该假设成立，只能说该假设在这组观测下没有被拒绝.

什么情况下“不拒绝”可以升级为“接受”？另外，对于同一个假设可以有不同的检验方法，如何择优也是一个问题. Jerzy Neyman 与 Egon Pearson 将假设检验视为决策过程，提出了形式化的假设检验框架.

统计假设是关于一个或多个总体某些性质的断言或猜测. 在现代假设检验框架中，通常考虑两个对立的统计假设：原假设（或零假设，Null Hypothesis）和备择假设（Alternative Hypothesis），分别用 H_0 和 H_1 表示. 原假设是被检验的假设，通常表示“无效应”、“无差异”或“满足某种已知状态”. 备择假设则是在拒绝原假设后可供选择的假设. 很多情况下统计假设可以表示为参数形式，即

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1,$$

其中 θ 是总体的未知参数， Θ_0 和 Θ_1 是互不相交的参数集合（$\Theta_0 \cap \Theta_1 = \varnothing$），且 $\Theta_0 \cup \Theta_1$ 包含 θ 的所有可能取值.

例 7.2. 设总体分布为 $X \sim N(\mu, \sigma^2)$，其中 σ^2 已知，μ_0 是给定常数，以下是一些常见的关于正态总体均值检验的假设：

1. $H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$（双侧检验）
2. $H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$（单侧检验）

3. $H_0 : \mu \le \mu_0, \quad H_1 : \mu > \mu_0$（单侧检验）

检验是双侧的还是单侧的取决于备择假设的形式.

对于两个总体的比较，假设总体 $X \sim N(\mu_1, \sigma^2)$ 和 $Y \sim N(\mu_2, \sigma^2)$，$X$ 与 Y 相互独立，且 σ^2 已知，可检验两总体均值是否相等：

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

这是一个典型的双侧检验，用于比较两个总体的差异.

按照假设涉及的总体分布数量，可将假设分为简单假设和复合假设. 对应单一总体分布的假设称为简单假设，包含多个总体分布的假设称为复合假设. 例如，上例中的 $H_0 : \mu = \mu_0$ 是简单假设，而 $H_0 : \mu \le \mu_0$ 是复合假设. 注意，若 σ^2 未知，则此时 $H_0 : \mu = \mu_0$ 对应的总体有无穷多个（$\{N(\mu_0, \sigma^2) | \sigma \ge 0\}$），是一个复合假设.

根据样本观测结果对原假设进行决策（即拒绝 H_0 或不拒绝 H_0）的过程称为假设检验. 一个具体的检验（准则）是在何种样本观测结果下应拒绝 H_0 的明确规则. 设所有可能的样本观测结果组成集合 $\{(X_1(\omega), \ldots, X_n(\omega)) | \omega \in \Omega\}$，其中样本容量 n 固定，则可按检验准则将该集合划分为 R 和 R^c 互补的两部分，当观测值落在 R 内时拒绝 H_0，否则（观测值落在 R^c 内）则不拒绝 H_0，其中，R 称为拒绝域或临界域，R^c 称为接受域. 由于是根据样本观测做出决策，所以拒绝域一般可抽象地表示为

$$R = \{(X_1, \ldots, X_n) \,|\, T(X_1, \ldots, X_n) \ge c\},$$

其中 $T(X_1, \ldots, X_n)$ 是检验统计量，c 是检验的临界值，此时检验准则可以表述为：

$$\text{若 } T(X_1, \ldots, X_n) \ge c, \text{ 则拒绝 } H_0.$$

若对于给定的显著性水平 $\alpha \in (0, 1)$，有

$$P(T(X_1, \ldots, X_n) \ge c \,|\, H_0) \le \alpha,$$

则称该检验是显著性水平为 α 的检验. 综上所述，在显著性检验中，如果观测结果落入拒绝域 R，则表明观测到了一件（在原假设 H_0 为真前提下的）小概率事件发生，因此可以根据小概率事件原理拒绝 H_0；反之，如果观测结果落入接受域 R^c，则表明样本结果与 H_0 相容，没有足够证据拒绝 H_0.

例 7.3. 设总体分布为 $X \sim N(\mu, \sigma^2)$，其中 σ^2 已知. 分别考虑以下两个假设检验：

1. （双侧检验） $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$. 对于正态总体而言，样本均值 $\bar{X}$ 是总体均值 μ 的一个良好点估计，于是 $\bar{X}$ 与 μ_0 的偏离越大，越有理由怀疑 H_0 的正确性. 因此，可构造检验统计量 $|\bar{X} - \mu_0|$，拒绝域形式为

$$R = \{(X_1, \ldots, X_n) \,|\, |\bar{X} - \mu_0| \geq c\}.$$

对于给定的显著性水平 $\alpha \in (0, 1)$，确定临界值 c 使得

$$P(|\bar{X} - \mu_0| \geq c \,|\, H_0) \leq \alpha.$$

由于当 H_0 为真时， $\bar{X} - \mu_0 \sim N(0, \dfrac{\sigma^2}{n})$，标准化后有

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1),$$

可取 $c = z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$，其中 $z_{\alpha/2}$ 为标准正态分布的上 $\dfrac{\alpha}{2}$-分位数. 由此得检验准则为：

$$\text{当 } |\bar{X} - \mu_0| \geq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ 时拒绝 } H_0.$$

2. （单侧检验）$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$. 在这种情况下，只有当 $\bar{X}$ 显著小于 μ_0 时，才有理由拒绝 H_0. 因此拒绝域可设为

$$R = \{(X_1, \ldots, X_n) \,|\, \bar{X} - \mu_0 \leq c\}.$$

对于给定的显著性水平 $\alpha \in (0, 1)$，确定临界值 c 使得

$$P(\bar{X} - \mu_0 \leq c \,|\, H_0) \leq \alpha.$$

由于当 H_0 为真时， $\bar{X} - \mu \sim N(0, \dfrac{\sigma^2}{n})$，标准化后有

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

所以

$$P(\bar{X} - \mu_0 \leq c \,|\, H_0) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{c + \mu_0 - \mu}{\sigma/\sqrt{n}} \,|\, H_0) = \Phi(\frac{c + \mu_0 - \mu}{\sigma/\sqrt{n}}),\ \mu \geq \mu_0.$$

可取 $c = -z_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}$. 由此得检验准则为：

$$\text{当 } \bar{X} \leq \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \text{ 时拒绝 } H_0.$$

本例中，标准化后的检验统计量在 H_0 为真时的分布（称为零分布）为标准正态分布，因此称为 Z-检验.

上例中，若 σ^2 未知，则检验统计量可调整为 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$，其中 S^2 为样本方差. 当 H_0 为真时，

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1),$$

这里 $t(n-1)$ 是自由度为 $n-1$ 的 t 分布. 相应地，检验临界值也要改用 t 分布的分位数. 以前面的单侧检验为例，对于给定的显著性水平 $\alpha \in (0,1)$，类似讨论可得检验准则为：

$$\text{当 } \bar{X} \leq \mu_0 - t_\alpha(n-1)\cdot\frac{S}{\sqrt{n}} \text{ 时拒绝 } H_0.$$

其中 $t_\alpha(n-1)$ 是 $t(n-1)$ 上 α-分位数. 这种检验被称为 t-检验，是实际应用中最常用的假设检验方法之一，特别适用于小样本情况.

7.2 Neyman-Pearson 假设检验

由于样本观测结果具有随机性，根据样本作决策，错误是不可能根本避免的. 在检验一个假设 H_0 时，有可能犯以下两类错误：

- H_0 实际为真但被拒绝了，称为第 *I 类错误*（Type I error），又称*弃真错误*.
- H_0 实际为假但未被拒绝，称为第 *II 类错误*（Type II error），又称*取伪错误*.

两类错误发生的概率分别记作 $P(\mathrm{I})$ 和 $P(\mathrm{II})$. 将 $1-P(\mathrm{II})$ 称为检验的*功效*（Power）（或检验力）. 一次决策不会同时犯两种错误. 当检验准则对应的拒绝域为 R 时，有

$$P(\mathrm{I}) = P((X_1,\cdots,X_n)\in R \,|\, H_0), \quad P(\mathrm{II}) = P((X_1,\cdots,X_n)\in R^c \,|\, H_1).$$

利用上述概念，显著性水平为 α 的检验实际上就是犯第 I 类错误的概率不超过 α 的检验.

实际上，当样本容量 n 固定时，两类错误的发生概率呈现此消彼长的关系. 下边的例子有助于直观感受到这一点.

例 7.4. 考虑检验元件是否合格的问题，H_0 和 H_1 分别表示“元件合格”与“元件不合格”.

1. 若检验策略是从不拒绝 H_0（即总认为元件合格），则 $P(\mathrm{I})=0$，但 $P(\mathrm{II})=1$，意味着所有不合格产品都将被错误地接受.
2. 一般来说，当使 $P(\mathrm{I})$ 变小，即更谨慎地判断元件为不合格时，不合格元件就更难以被检测出来，可见 $P(\mathrm{II})$ 会随之增大，反之亦然.

例 7.5. 设总体分布为 $X\sim N(\mu,=\sigma^2)$. 考虑假设检验： $H_0:\mu\geq\mu_0$, $H_1:\mu<\mu_0$. 根据例 7.3 讨论可知，当取检验统计量为 $\bar{X}$ 时，拒绝域形状为

$$R=\{(X_1,\ldots,X_n)\,|\,\bar{X}-\mu_0\leq c\}.$$

第 I 类错误概率为

$$P(\bar{X}-\mu_0 \le c \,|\, H_0) = \Phi(\frac{c+\mu_0-\mu}{\sigma/\sqrt{n}}),\ \mu \ge \mu_0,$$

第 II 类错误概率为

$$P(\bar{X}-\mu_0 > c \,|\, H_1) = 1-\Phi(\frac{c+\mu_0-\mu}{\sigma/\sqrt{n}}),\ \mu < \mu_0.$$

当样本容量 n 固定时，找不到一个 c 使得两类错误概率均尽可能小. 保持 α 不变，当 n 增加，则检验统计量 $\bar{X}$ 的临界值 $\mu_0 - z_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}$ 变大，第 II 类错误概率变小.

Neyman-Pearson 范式：固定样本容量 n，对于预先给定的 $\alpha \in (0,1)$，优先保证第 I 类错误概率不超过 α，再在此限制下尽可能减小第 II 类错误概率. α 称为检验水平，满足要求的检验称为水平为 α 的检验.

由前面的讨论可知，检验水平与显著性检验的显著性水平等价. 若存在 $\alpha, \beta > 0$，使得检验满足 $P(\mathrm{I}) \le \alpha$，$P(\mathrm{II}) \le \beta$，则 α, β 为预先设定的可接受的长期错误率，是检验程序的属性.

在 Neyman-Pearson 范式下，H_0 与 H_1 的地位通常不对等. 原假设 H_0 往往是受保护的，若证据不充分则不能予以拒绝；备择假设 H_1 常常是研究者真正感兴趣的内容，也称为研究假设.

在 Fisher 显著性检验中，“不拒绝 H_0”并不等同于“接受 H_0”，但在 Neyman-Perason 检验框架下，如果第 II 类错误概率得到有效控制，即$P(\mathrm{II})$ 足够小（等价于说功效 $1-P(\mathrm{II})$ 足够大）时，则不拒绝 H_0 的决策可以升级为接受 H_0（相当于拒绝了 H_1）.

7.3 假设检验与置信区间

在统计推断中，区间估计和假设检验是两种主要的方法，二者虽然形式不同，但存在密切的关系，以下通过具体示例进行说明.

例 7.6. 设总体分布为 $X \sim N(\mu, \sigma^2)$，其中 σ^2 已知. 对于任意显著性水平 $\alpha \in (0,1)$，$(1-\alpha)$-置信的（双侧）区间估计为

$$\mu \in \left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right).$$

若进行 μ 的双侧假设检验：

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \ne \mu_0,$$

以 $\bar{X}$ 为检验统计量，检验准则为

$$\text{若 } |\bar{X}-\mu_0| \ge z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ 则拒绝 } H_0,$$

即所谓的 接受域

$$R^c = \left\{ (X_1, \cdots, X_n) \mid |\bar{X} - \mu_0| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\},$$

或接受条件为

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

由此可见，当 μ_0 落在置信区间 $\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$ 内时，等价于以 $\bar{X}$ 为检验统计量，对上述 H_0 和 H_1 进行假设检验时的结论为“不拒绝 H_0”，这一对应关系清晰地体现了区间估计与假设检验之间存在的对偶关系. 换言之，置信区间实际上可以被视为所有未被以给定显著性水平拒绝的参数值的集合.

例 7.7. 假设总体服从正态分布 $N(\mu, \sigma^2)$，但 μ 和 σ^2 均未知. 若进行如下单侧假设检验：

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0,$$

则以 $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ 作为检验统计量（其中S^2为样本方差），在显著性水平 α 下，接受域为

$$R^c = \{ (X_1, \cdots, X_n) \mid T \geq -t_\alpha(n-1) \}.$$

接受条件等价于

$$\mu_0 \leq \bar{X} + \frac{S}{\sqrt{n}} t_\alpha(n-1),$$

而$\bar{X} + \dfrac{S}{\sqrt{n}} t_\alpha(n-1)$ 正是 μ 的 $(1-\alpha)$-置信上限. 这进一步说明了单侧区间估计与单侧假设检验之间的对偶关系.

总之，区间估计与假设检验在理论上具有对偶性，本质上可以相互转化，实现等价的统计决策. 然而，置信区间相比假设检验，能够提供关于参数的更丰富信息，并且具有更直观、易于理解的解释，避免了“拒绝/不拒绝”这种简单二元的结论，因而更不容易被误解或误用.

7.4 检验的 P 值

定义 7.1. 当原假设 H_0 为真时，出现观测结果以及更极端观测的概率称为该检验的 P 值.

其中，所谓“更极端”的具体含义由备择假设 H_1 决定. P 值越小，表明原假设下观测到当前或更极端结果的可能性越低，观测结果与原假设 H_0 越不相容，从而对原假设提出越强的质疑.

例 7.8 (选举问题). 设 p 为未知的真实支持率，样本容量 $n = 1200$，调查中有 684 人表示支持，即支持率的观测比例为 $p_n = \dfrac{684}{1200} \approx 0.57$.

给定 $p_0 \in (0,1)$，考虑检验

$$H_0 : p = p_0, \quad H_1 : p > p_0.$$

以样本支持比例 $P_n = \bar{X}$ 作为检验统计量，其中 $X_i \sim \mathrm{B}(p)$（$i = 1, 2, \ldots, n$）且（近似）相互独立. 由中心极限定理，

$$\frac{P_n - p}{\mathrm{Se}(P_n)} \overset{\text{近似}}{\sim} N(0,1).$$

其中标准误 $\mathrm{Se}(P_n) = \sqrt{\dfrac{p(1-p)}{n}}$. 当 H_0 为真时，标准误 $\mathrm{Se}(P_n) = \sqrt{\frac{p_0(1-p_0)}{n}}$，故 P 值计算为

$$\begin{aligned} P(P_n \geq p_n | H_0) &= P\left(\frac{P_n - p_0}{\mathrm{Se}(P_n)} \geq \frac{p_n - p_0}{\mathrm{Se}(P_n)} | H_0 \right) \\ &\approx P\left(Z \geq \frac{p_n - p_0}{\mathrm{Se}(P_n)} \right), \end{aligned}$$

其中 $Z \sim N(0,1)$.

代入具体数值：

- 若 $p_0 = 0.5$，得 $\mathrm{Se}(P_n) \approx 0.014$，计算得此时 P 值 $\ll 0.001$；
- 若 $p_0 = 0.55$，得 $\mathrm{Se}(P_n) \approx 0.014$，计算得此时 P 值 ≈ 0.082.

根据 P 值的定义，在实际检验中，当 P 值 $\leq \alpha$（显著性水平）时拒绝 H_0，当 P 值 $> \alpha$ 时不拒绝 H_0. 非形式地，P 值通常被用作反对原假设证据强度的度量（尽管理论上不是一个好指标）. 然而，需要特别强调的是，P 值与 $P(H_0|\text{观测值})$（即给定观测数据下原假设成立的后验概率）并不同. 若 P 值不小，则不拒绝 H_0，其可能的原因为：一是 H_0 本身为真，二是 H_0 不真但检验的功效不够大. 同时，P 值也不等于“错误拒绝 H_0 的概率”.

值得注意的是，当原假设是复合假设时，按照定义 7.1 计算的 P 值不再是一个值，而是一个关于被检验参数的函数，相应的定义需要调整为原假设为真时出现观测结果以及更极端观测的最大概率. 例如，上例中若 H_0 改为 $p \leq p_0$，则 P 值应修正为

$$\sup_{p \leq p_0} P(P_n \geq p_n)$$

此时，当 H_0 为真时，标准误 $\mathrm{Se}(P_n)$ 依赖未知参数，所以需给出合理的估计 $\widehat{\mathrm{Se}}(P_n)$，进而有

$$\frac{P_n - p}{\widehat{\mathrm{Se}}(P_n)} \overset{\text{近似}}{\sim} N(0,1).$$

P 值的计算公式也相应修正为

$$P(P_n \geq p_n|H_0) = P\left(\frac{P_n - p}{\widehat{\mathrm{Se}}(P_n)} \geq \frac{p_n - p}{\widehat{\mathrm{Se}}(P_n)}|H_0\right) \approx P\left(Z \geq \frac{p_n - p}{\widehat{\mathrm{Se}}(P_n)}\right),$$

其中 $Z \sim N(0,1)$.

下面给出 P 值的一般定义.

定义 7.2. 若检验准则为“当 $T(X_1, \cdots, X_n) \geq C$ 时拒绝 $H_0 : \theta \in \Theta_0$”，则检验的 P 值定义为

$$\sup_{\theta \in \Theta_0} P(T(X_1, \cdots, X_n) \geq T(x_1, \cdots, x_n))$$

其中 $T(x_1, \cdots, x_n)$ 为检验统计量在所观测到样本 $(x_1, \cdots, x_n)$ 上的取值.

现在被普遍应用的假设检验是零假设显著性检验（Null Hypothesis Significance Test，简记为 NHST），通常先设定一个显著性水平（如 $\alpha = 0.05$），计算 P 值再与其比较进行决策. 零假设显著性检验其实是 Fisher 显著性检验与 Neyman-Pearson 假设检验二者的混合，虽然简化了实际应用流程，但过程中往往略功效分析，并且容易让人误解或滥用 P 值.

例 7.9 (P 值的常见误解)**.** 考虑比较两总体均值差异的假设检验：

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

取检验水平 $\alpha = 0.05$，已知采集样本后通过计算得到 P 值为 0.001. 请判断以下说法的对错并简要说明理由.

(1) 已经完全排除了原假设（两总体均值之间没有差异）.

(2) 已经得到了原假设为真的概率.

(3) 已经完全证明了备择假设（两总体均值之间存在差异）.

(4) 能够推断出备择假设为真的概率.

(5) 当你拒绝原假设时，你知道自己出错的概率.

(6) 你得到了一个可靠的实验结果（拒绝假设 H_0），即假设大量重复这个实验，那么你将在 99.9% 的情况下得到显著的结果（拒绝假设 H_0）.

- **(1) 和 (3) 错误：**假设检验框架中不存在绝对的证伪或证真. P 值小仅表明若 H_0 成立，观测到当前或更极端数据的概率很小，但不代表完全不可能.
- **(2) 和 (4) 错误：**P 值是在 H_0 成立的条件下计算的条件概率，不等于 H_0 或 H_1 成立的概率.

- **(5) 错误：**对单次具体决策而言，要么正确要么错误，没有概率可言. 显著性水平 α 描述的是长期重复实验中可接受的错误率，而非单次决策的错误概率.
- **(6) 错误：**这个描述混淆了 P 值和检验功效的概念，两者是不同的概念.

7.5 拟合优度检验

拟合优度检验是一种用于判断观测数据是否符合某一特定理论分布的统计方法，常见方法之一为 *Pearson* 卡方检验（Pearson's Chi-squared Test），也称为卡方拟合优度检验，是一种常用的非参数统计方法. Pearson 卡方检验通过比较观测频数与理论分布下的期望频数之间的差异，来判断样本数据是否符合某个特定的概率分布. 这种检验方法广泛应用于分类数据分析和模型验证等领域.

例 7.10. 想要检验一枚六面骰子是否为均匀骰子. 现连续掷该骰子 60 次，得到各点数出现频数如下表所示：

点数	1	2	3	4	5	6	总计
观测频数	4	6	17	16	8	9	60
期望频数	10	10	10	10	10	10	60

提出假设：

$$H_0: \text{骰子是均匀的（即各面出现的概率相等）}, \quad H_1: \text{骰子不均匀}$$

Pearson 卡方检验的实现主要依据以下结果.

定理 7.1. 考虑 $H_0: p_i = p_i^0, i = 1, \ldots, k$. 这里 k 是单元数量，p_i 是第 i 个单元的发生概率. 定义*Pearson 卡方检验统计量*为

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

其中

- O_i 是第 i 个单元的观测频数；
- E_i 是当 H_0 为真时第 i 个单元的期望频数，计算公式为 $E_i = np_i^0$，其中 n 为总观测次数.

若原假设 H_0 为真，则当 $n \to \infty$ 时，统计量 χ^2 的分布趋向于自由度为 $k-1$ 的 χ^2 分布，即 $\chi^2(k-1)$.

根据表中数据，代入公式，计算检验统计量 χ^2 的观测值为

$$\begin{aligned}\chi_0^2 &= \sum_{i=1}^{6}\frac{(O_i-E_i)^2}{E_i}\\ &= \frac{(4-10)^2}{10}+\frac{(6-10)^2}{10}+\frac{(17-10)^2}{10}+\frac{(16-10)^2}{10}+\frac{(8-10)^2}{10}+\frac{(9-10)^2}{10}\\ &= 14.2\end{aligned}$$

在 H_0 为真条件下，χ^2 统计量近似服从 $\chi^2(5)$ 分布（自由度为 $k-1=5$）. 根据 χ^2 分布表或数值计算，可得

$$P\text{值} = P(\chi^2 \geq \chi_0^2) = P(\chi^2 \geq 14.2) \approx 0.014.$$

在显著性水平 $\alpha = 0.05$ 下，由于 P 值 $< \alpha$，因此拒绝原假设 H_0，认为该骰子不均匀.

为了保证上述检验的适用性，期望频数通常需满足 $E_i = np_i^0 \geq 5$，即每个单元的期望频数不应小于5，否则需要合并相邻单元格以满足此条件. 如果无法满足此条件，比如说单元数过少无法合并，可考虑使用其他检验方法，如精确概率检验.

例 7.11. 对 $k=2$ 的简单情形，即检验 $H_0: p_1 = p_1^0, p_2 = 1-p_1^0$. 此时检验统计量为

$$\chi^2 = \frac{(O_1-np_1^0)^2}{np_1^0}+\frac{(O_2-n(1-p_1^0))^2}{n(1-p_1^0)} = \frac{(O_1-np_1^0)^2}{np_1^0(1-p_1^0)}.$$

根据定理 7.1， $\chi^2 \overset{近似}{\sim} \chi^2(1)$. 另一方面，$\mathrm{E}(O_1) = np_1^0, \mathrm{Var}(O_1) = np_1^0(1-p_1^0)$，而 $\chi^2(1)$ 分布正是标准正态分布的平方. 实际上，对于上述二项分布检验，也可以使用正态近似构造 Z 检验，即根据中心极限定理有

$$Z = \frac{O_1-np_1^0}{\sqrt{np_1^0(1-p_1^0)}} \overset{近似}{\sim} N(0,1),$$

此时有 $Z^2 = \chi^2$，两种检验方法等价.

需要特别强调，统计显著不一定意味着实际显著，尤其是在样本量极大的情况下. 以下例子说明了这一点.

例 7.12. 设连续掷一个六面骰子 6×10^{10} 次，得到结果如下：

点数	1	2	3	4	5	6
观测频数 -10^{10}	-10^6	1.5×10^6	-2×10^6	4×10^6	-3×10^6	0.5×10^6

考虑假设检验： H_0 : 骰子是均匀的，H_1 : 骰子不均匀.

计算 χ^2 统计量的观测值，得到

$$\begin{aligned}\chi_0^2 &= \frac{(-10^6)^2}{10^{10}} + \frac{(1.5\times 10^6)^2}{10^{10}} + \frac{(-2\times 10^6)^2}{10^{10}} + \frac{(4\times 10^6)^2}{10^{10}} + \frac{(-3\times 10^6)^2}{10^{10}} + \frac{(0.5\times 10^6)^2}{10^{10}} \\ &= 3250\end{aligned}$$

根据定理 7.1，检验统计量 $\chi^2 \overset{\text{近似}}{\sim} \chi^2(5)$，可以得到 $P(\chi^2 \geq 3250) \ll 0.00001$，即有检验的 P 值非常小，故拒绝原假设 H_0.

然而，观测值给出各点数出现概率的极大似然估计为：

$$\begin{aligned}p_1^* &= \frac{10^{10}-10^6}{6\times 10^{10}} = \frac{1}{6} - \frac{1}{6}\times 10^{-4} \\ p_2^* &= \frac{10^{10}+1.5\times 10^6}{6\times 10^{10}} = \frac{1}{6} + \frac{1.5}{6}\times 10^{-4} \\ &\vdots\end{aligned}$$

可以看出，各点数出现概率的估计值与理论值 $\frac{1}{6}$ 的差异极小，从通常的实际角度看应该可以认为骰子足够均匀了，这表明虽然从统计意义上拒绝了均匀性假设（统计显著），但实际上这种差异可能并不重要. 当样本量极大时，很小的差异多次累积也会导致统计显著. 总之，统计显著并不等于实际显著，统计显著的差异是否实际显著还需结合实际问题综合多方面的考虑，才能得到合理的解释.

Pearson 卡方检验也可以用于检验分类变量之间是否存在显著的相关性，这种应用通常被称为列联表检验（Contingency Table Test）.

例 7.13 (独立性检验). 希望研究对某项议题的态度与年龄段是否相互独立，调查结果如下：

	青	中	老	总计
支持	20	40	20	80
反对	30	30	10	70
总计	50	70	30	150

考虑假设检验 H_0：年龄段与态度相互独立，H_1：年龄段与态度不相互独立.

取检验统计量

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

其中 a, b 分别为行数和列数，O_{ij} 为观测频数，E_{ij} 为期望频数.

设 p_{ij} 表示样本属于 (i,j) 单元的概率，p_{i+} 和 p_{+j} 分别表示第 i 行和第 j 列的边际概率. 在 H_0 成立（独立）的情况下，有 $p_{ij} = p_{i+}p_{+j}$.

在 H_0 为真时，可给出 p_{ij} 的极大似然估计 $p_{ij}^* = (p_{i+}p_{+j})^* = p_{i+}^* p_{+j}^*$，其中 $p_{i+}^* = \frac{O_{i+}}{n}$，$p_{+j}^* = \frac{O_{+j}}{n}$，为边际概率的极大似然估计，$O_{i+}$ 为第 i 行的观测频数之和，O_{+j} 为第 j 列的观测频数之和. 此时期望频数 $E_{ij} = np_{ij} \approx np_{ij}^* = \frac{O_{i+}O_{+j}}{n}$，$n$ 为总样本量. 统计量的卡方分布自由度为 $ab-1-(a+b-2) = (a-1)(b-1)$，这里自由度由 $k-1 = ab-1$ 下降为 $ab-1-(a+b-2)$，所减少的自由度正好等于要估计的参数个数 $a-1+b-1 = a+b-2$，这种自由度的修正是由 Fisher 在 1924 年给出的，相关讨论比较复杂，超出了本书范围. 本例中 $a = 2$，$b = 3$，卡方自由度为 $(2-1)(3-1) = 2$. 代入数据计算得到检验统计量的观测值为

$$\chi_0^2 \approx 6.12.$$

从而

$$P\text{值} = P(\chi^2 \geq \chi_0^2) \approx P(\chi^2 \geq 6.12) \approx 0.0469.$$

在显著性水平 $\alpha = 0.05$ 下，由于 P 值 $\leq \alpha$，所以拒绝原假设，认为年龄段与对议题的态度不相互独立，即二者之间存在统计学上的关联.

7.6 似然比检验

似然比检验（Likelihood Ratio Test, 简记为 LRT）是经典统计推断中重要的一种假设检验方法，通过比较在不同假设下观测数据出现的最大可能性，来判断哪一个假设更为合理.

例 7.14. 设有两种硬币，正面朝上的概率分别为 $p = 0.5$ 和 $p = 0.7$. 现从中取一枚，未知其类型. 将其抛掷 $n = 10$ 次，观测到 $X = x$ 次正面向上. 考虑假设检验： $H_0 : p = 0.5$, $H_1 : p = 0.7$.

似然比为

$$\lambda(x) = \frac{P(X = x|H_0)}{P(X = x|H_1)} = \frac{\binom{n}{x}0.5^x(1-0.5)^{n-x}}{\binom{n}{x}0.7^x(1-0.7)^{n-x}} = \frac{0.5^x \cdot 0.5^{n-x}}{0.7^x \cdot 0.3^{n-x}}.$$

检验准则为：若似然比 $\lambda(x) \leq c$ 则拒绝 H_0，其中临界值 c 需满足 $P(\lambda(X) \leq c \,|\, H_0) \leq \alpha$，可据此确定 c 的具体数值.

x	0	1	$\cdots$	7	8	9	10
$p = 0.5$	0.0010	0.0098	$\cdots$	0.1172	0.0439	0.0098	0.0010
$p = 0.7$	0.0000	0.0001	$\cdots$	0.2668	0.2335	0.1211	0.0282
似然比	165.4	70.88	$\cdots$	0.4392	0.1882	0.0807	0.0346

若取显著性水平 $\alpha = 0.05$，则从表中数据可以得出

$$P(\text{拒绝}H_0 \,|\, H_0) = P(X \geq 9 \,|\, H_0) \approx 0.0108,$$

似然比临界值此时可以取 $c = 0.1$.

当 H_0 和 H_1 均为简单假设时，Neyman 与 Pearson 在其基础性工作中证明了似然比检验具有最优性（即在给定显著性水平的条件下，检验的功效最大）. 这一结论被称为 *Neyman-Pearson 引理*，是假设检验理论中的一个重要里程碑. 值得注意的是，当 H_0 或 H_1 不是简单假设时，似然比检验一般来说不能保证最优性，但在实际应用中通常表现良好.

一般地，若要检验假设 $H_0 : \theta \in \Theta_0$fi $H_1 : \theta \in \Theta_1$，给定随机样本 $X_1, \ldots, X_n$，则广义似然比定义为

$$\Lambda^* := \frac{\sup\limits_{\theta\in\Theta_0} L(\theta)}{\sup\limits_{\theta\in\Theta_1} L(\theta)},$$

其中 $L(\theta)$ 表示参数为 θ 时的似然函数. 在实际应用中，通常采用其修正形式（也称为广义似然比）

$$\Lambda := \frac{\sup\limits_{\theta\in\Theta_0} L(\theta)}{\sup\limits_{\theta\in\Theta_0\cup\Theta_1} L(\theta)} = \min\{\Lambda^*, 1\}.$$

分母 $\sup\limits_{\theta\in\Theta_0\cup\Theta_1} L(\theta)$ 等于 $L(\theta^*)$，其中 θ^* 为 θ 的极大似然估计. 显然，$\Lambda = \Lambda(X_1, \ldots, X_n)$ 的值越小，越倾向于反对原假设 H_0，因此可根据 $P(\Lambda \le \lambda_0 \,|\, H_0) \le \alpha$ 确定临界值 λ_0 以及拒绝域. 一般情况下，可以应用下述定理进行计算.

定理 7.2. 在一定的正则条件下，在原假设 H_0 为真的前提下，当 $n \to \infty$ 时，有

$$-2\log\Lambda \xrightarrow{d} \chi^2(d),$$

其中自由度 $d = \dim(\Theta_0 \cup \Theta_1) - \dim(\Theta_0)$，dim 表示参数空间的自由参数个数.

例 7.15 (多项分布检验). 设 $X = (X_1, \ldots, X_k)$ 服从多项分布，参数为 $n, p_1, \ldots, p_k$，检验 $H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0$. 观测频数为 $n_1, \ldots, n_k$，满足 $n_1 + \cdots + n_k = n$. 此时似然函数为

$$L(p_1, \ldots, p_k) = \tbinom{n}{n_1,\ldots,n_k} p_1^{n_1} \cdots p_k^{n_k},$$

故（广义）似然比为

$$\Lambda = \frac{L(p_1^0, \ldots, p_k^0)}{L(p_1^*, \ldots, p_k^*)} = \frac{{p_1^0}^{n_1} \cdots {p_k^0}^{n_k}}{{p_1^*}^{n_1} \cdots {p_k^*}^{n_k}},$$

其中 $p_i^* = \dfrac{n_i}{n}$ 为参数 p_i 的极大似然估计. 记 $O_i = np_i^* = n_i, E_i = np_i^0$，则

$$-2\log\Lambda = -2\sum_{i=1}^{k} n_i \log\frac{p_i^0}{p_i^*} = 2\sum_{i=1}^{k} O_i \log\frac{O_i}{E_i}.$$

利用 Taylor 展开 $x\log\frac{x}{x_0}=0+(x-x_0)+\frac{(x-x_0)^2}{2x_0}+\cdots$，有

$$-2\log\Lambda=2\sum_{i=1}^{k}(O_i-E_i)+\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}+\cdots=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}+\cdots$$

由定理 7.2，在 H_0 为真的前提下，当 $n\to\infty$ 时，有 $-2\log\Lambda\xrightarrow{d}\chi^2(d)$，其中 $d=(k-1)-0=k-1$. 由此可见，多项分布检验中似然比检验与 Pearson 卡方检验两者渐近等价.

7.7 两总体比较

实际应用中经常会进行两个总体的比较，针对两个总体的参数（常见有均值、方差、比例等）进行差异性检验. 首先考虑两个**独立**总体，如下表所示.

总体	均值	方差	随机样本
X	μ_1	σ_1^2	$X_1,\cdots,X_n$
Y	μ_2	σ_2^2	$Y_1,\cdots,Y_m$

例 7.16 (正态总体比较). 若 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，给定显著性水平 $\alpha>0$，比较均值：

$$H_0:\mu_1=\mu_2,\quad H_1:\mu_1\neq\mu_2.$$

若 σ_1^2,σ_2^2 已知，则检验统计量为

$$Z=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}},$$

其中 $\bar{X},\bar{Y}$ 分别为两个样本均值. 根据之前章节的讨论，在 H_0 为真前提下，$Z\sim N(0,1)$，检验准则为：

$$\text{当 }|Z|\geq z_{\alpha/2}\text{ 时拒绝 }H_0.$$

若 $\sigma_1^2=\sigma_2^2=\sigma^2$ 但未知，则检验统计量为

$$T=\frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

这里 S^2 为合并样本方差，即

$$S^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2},$$

其中，S_X^2,S_Y^2 分别为两样本方差. 在 H_0 为真前提下，$T\sim t(n+m-2)$，检验准则为：

$$\text{当 }|T|\geq t_{\alpha/2}(n+m-2)\text{ 时拒绝 }H_0,$$

其中 $t_{\alpha/2}(n+m-2)$ 是自由度为 $n+m-2$ 的 t 分布的上 $\frac{\alpha}{2}$-分位数.

若 σ_1^2, σ_2^2 未知且不一定相等，则可用大样本方法，此时取检验统计量为

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}.$$

根据中心极限定理，Z 近似服从标准正态分布 $N(0,1)$.

检验两总体方差是否相等：

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

检验统计量为

$$F = \frac{S_X^2}{S_Y^2}$$

在 H_0 为真前提下，$F \sim F(n-1, m-1)$，检验准则为：

$$\text{当 } F \leq F_{\alpha/2}(n-1, m-1) \quad \text{或 } F \geq F_{1-\frac{\alpha}{2}}(n-1, m-1), \text{时拒绝 } H_0,$$

其中 $F_{\alpha/2}(n-1, m-1), F_{1-\frac{\alpha}{2}}(n-1, m-1)$ 分别是自由度为 $F(n-1, m-1)$ 的上 $\frac{\alpha}{2}$-分位数和上 $(1-\frac{\alpha}{2})$-分位.

例 7.17 (比例比较). 研究阿司匹林对降低心脏病发病率的有效性. 样本数据如下：

	心脏病发作	未发作	总计	发作率
阿司匹林	$k_1 = 139$	10898	$n_1 = 11037$	1.26%
安慰剂	$k_2 = 239$	10795	$n_2 = 11034$	2.17%

考虑假设检验： $H_0 : p_1 = p_2$， $H_1 : p_1 < p_2$，其中 p_1, p_2 分别为服用阿司匹林和安慰剂两总体心脏病理论发作率. 记试验组和对照组的样本发作率分别为 P_1, P_2. 检验统计量为 $P_1 - P_2$（样本发作率之差）. 根据中心极限定理，

$$Z = \frac{(P_1 - P_2) - (p_1 - p_2)}{\mathrm{Se}} \overset{\text{近似}}{\sim} N(0,1),$$

其中

$$\mathrm{Se}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

在 H_0 为真前提下，记 $p_1 = p_2 = p$，可用其极大似然估计 $p^* = \dfrac{k_1 + k_2}{n_1 + n_2}$ 替代 p 来估计标准误，得到

$$\widehat{\mathrm{Se}}^2 = p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

代入数据得 $p^* \approx 0.0171$， $\widehat{\mathrm{Se}}^2 \approx 0.00176^2$. 计算得检验统计量的观测值为

$$Z_0 \approx \frac{1.26\% - 2.17\%}{0.176\%} \approx -5.2$$

对应的 P 值极小（约 10^{-7}），因此拒绝 H_0，认为阿司匹林对降低心脏病发病率有效.

上例的实验设计要点包括：（1）随机分组；（2）双盲试验；（3）样本容量足够大.

例 7.18. 某大型出租车公司比较两种汽油 A 和 B 对行驶里程的影响. 随机将 100 部车辆分两组加等量的油进行实验，得到的数据总结如下：

	样本容量	平均里程	标准差
油 A	50	25	5.00
油 B	50	26	4.00

考查两均值差异，考虑假设检验： $H_0 : \mu_A = \mu_B$， $H_1 : \mu_A \neq \mu_B$. 检验统计量为

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A^2}{n_1} + \frac{S_B^2}{n_2}}},$$

其观测值为

$$Z_0 = \frac{25 - 26}{\sqrt{\frac{5^2}{50} + \frac{4^2}{50}}} \approx -1.10.$$

所以检验的 P 值为 $P(|Z| \geq Z_0) \approx 0.27$，因此不拒绝 H_0，即差异统计不显著.

下面讨论两**相关**总体的比较. 注意到上例中样本数据的标准差比较大，这很可能是车况和出租车司机（与出租车绑定）驾驶习惯差异造成的，为尽可能地减少这方面的影响，可采用配对试验设计.

例 7.19 (配对比较)**.** 延续上例，将每辆车在不同日子分别加油 A 与 B，可以通过掷硬币的方式决定先加油 A 还是油 B. 事实上，实验的车辆数目甚至可以减少到 10 就可以检验到差异. 具体数据如下：

车号	油 A	油 B	差异 d_i
1	27.01	26.95	0.06
2	20.00	20.44	−0.44
⋮	⋮	⋮	⋮
10	25.22	26.01	−0.79
均值	25.00	25.60	−0.60
标准差	4.27	4.10	0.61

表格中油 A 与油 B 两列不再相互独立，此时可以进行配对检验. 考虑这两列的差异列，可以看出差异列的标准差明显较小，这表明配对设计减小了车辆间的变异影响. 考虑假设检验： $H_0 : \mu_d = 0, H_1 : \mu_d \neq 0$，根据生活实际，不妨假设近似有 $d_i \sim N(\mu_d, \sigma_d^2)$，则

$$\frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} \sim t(n-1).$$

$n = 10$，检验的 P 值为

$$P\left(|t_9| \geq \left|\frac{-0.60}{0.61/\sqrt{10}}\right|\right) \approx P(|t_9| \geq 3.10) \approx 0.012$$

其中 $t_9 \sim t(9)$. 由于 P 值 ≤ 0.05，因此拒绝 H_0，认为两种油的平均里程存在显著差异.

7.8 Bayes 假设检验

贝叶斯假设检验是一种基于贝叶斯理论的方法，与经典的频率学派假设检验不同，通过引入先验分布，直接计算假设为真的后验概率，从而进行决策.

在贝叶斯假设检验中，首先要设定假设的先验概率，然后结合观测数据 $X = x$，通过贝叶斯公式计算后验概率. 检验的核心在于后验概率比，即

$$\frac{P(H_0|X=x)}{P(H_1|X=x)} = \frac{P(H_0)\,P(X=x|H_0)}{P(H_1)\,P(X=x|H_1)},$$

其中 $P(H_0), P(H_1)$ 是先验概率，$P(X = x|H_0)$ 与 $P(X = x|H_1)$ 是似然函数. 决策准则为：若后验概率比小于 k，则拒绝 H_0. 其中 k 为决策阈值，通常取 $k = 1$（即选择后验概率更大的假设）.

与经典方法中的 P 值检验不同，贝叶斯检验能直接反映假设为真的概率，比如在观测数据 $X = x$ 的条件下 H_0 为真的概率，并量化决策的可信度. 注意到，后验概率比可以分解为先验概率比与似然比的乘积，即

$$\frac{P(H_0|X=x)}{P(H_1|X=x)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(X=x|H_0)}{P(X=x|H_1)}.$$

这一分解突显了先验信息和数据证据在贝叶斯决策中的各自贡献. 似然比

$$BF_{01} := \frac{P(X=x|H_0)}{P(X=x|H_1)}$$

也称为贝叶斯因子，是贝叶斯统计中衡量证据强度的一种方式. 通常贝叶斯因子可以作如下解释：

- $BF_{01} > 10$：数据强支持 H_0

- $1 < BF_{01} \leq 10$：数据支持 H_0
- $BF_{01} = 1$：数据不偏向任何假设
- $\frac{1}{10} \leq BF_{01} < 1$：数据支持 H_1
- $BF_{01} < \frac{1}{10}$：数据强支持 H_1

例 7.20. 假设袋中有两种硬币，正面朝上的概率分别为 $p = 0.5$ 和 $p = 0.7$. 选择一枚抛掷 10 次，设 X 为正面次数. 考虑假设检验：

$$H_0 : p = 0.5, \quad H_1 : p = 0.7.$$

(1) 假设袋中两类硬币数量相同，则 $P(H_0) = P(H_1)$. 此时拒绝域为 $\{X \geq 7\}$.

(2) 假设袋中 $p = 0.5$ 的硬币数量是 $p = 0.7$ 的硬币数量的10倍，则 $\frac{P(H_0)}{P(H_1)} = 10$. 此时拒绝域为 $\{X \geq 9\}$.

前面的讨论显示，贝叶斯方法是将假设检验视为模型比较问题，则可自然地扩展到多个假设的比较，同时评估多个竞争假设或模型，通过后验概率直接量化各模型的相对可信度，而不仅限于经典方法中的二元假设（原假设 vs 备择假设）框架.

第八章　线性回归

8.1　简单线性回归

典型的回归问题可表述为

$$Y = f(X_1, \ldots, X_n) + \epsilon,$$

其中 Y 称为因变量或响应变量；$X_1, \ldots, X_n$ 称为自变量、回归变量或预测变量；ϵ 为随机误差项，代表一些无法观测或尚不清楚的影响因素.

若假定 $\mathrm{E}(\epsilon \,|\, X_1, \ldots, X_n) = 0$，则

$$\mathrm{E}(Y \,|\, X_1, \ldots, X_n) = f(X_1, \ldots, X_n).$$

f 称为 Y 关于 $X_1, \ldots, X_n$ 的（均值）回归函数. 通过对 $(X_1, \ldots, X_n, Y)$ 的样本观测，可以利用有监督学习方法估计函数 f 的结构或参数.

根据问题背景，自变量 $X_1, \ldots, X_n$ 在建模时可以是随机的，也可以是事先确定的控制变量. 为便于理论推导和理解，本讲义约定 $X_1, \ldots, X_n$ 为非随机的.

对于随机误差项，通常假设 $\mathrm{E}(\epsilon) = 0$，$\mathrm{Var}(\epsilon) = \sigma^2$，其中 σ^2 是一个未知常数. 影响 σ^2 的典型因素包括：

1. 对 Y 影响重要的自变量是否完全
2. 函数 f 的形式是否准确

模型若遗漏了关键变量或函数形式选择不合理，则会导致 σ^2 较大，从而影响拟合和解释能力.

简单线性回归的理论模型为

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

其中 β_0 为截距或常数项；β_1 为斜率或回归系数，二者合称为回归参数，均为未知数. 定义中“简单”是指自变量只有一个；“线性”是指模型关于参数 β_0, β_1 为线性关系.

设对 (X,Y) 进行 n 次独立观测，得到样本 $(x_1,y_1),\dots,(x_n,y_n)$，则模型具体形式为

$$\begin{cases} y_i=\beta_0+\beta_1x_i+\epsilon_i, \quad i=1,2,\dots,n, \\ \epsilon_i \text{ 独立同分布，} \mathrm{E}(\epsilon_i)=0,\ \mathrm{Var}(\epsilon_i)=\sigma^2. \end{cases}$$

此为简单线性回归或者一元线性回归模型. 由此可得，

$$\mathrm{E}(y_i)=\beta_0+\beta_1x_i, \quad \mathrm{Var}(y_i)=\sigma^2, \quad i=1,2,\dots,n.$$

8.2 回归参数推断

为估计参数 β_0,β_1，定义损失函数

$$S(\beta_0,\beta_1)=\sum_{i=1}^{n}\left[y_i-(\beta_0+\beta_1x_i)\right]^2. \tag{8.1}$$

最小化该损失函数可得回归系数的最小二乘估计

$$\begin{aligned} \hat\beta_1&=\frac{\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y)}{S_{xx}}=\frac{S_{xy}}{S_{xx}}, \\ \hat\beta_0&=\bar y-\hat\beta_1\bar x. \end{aligned}$$

其中

$$\bar x=\frac1n\sum_{i=1}^n x_i, \quad \bar y=\frac1n\sum_{i=1}^n y_i, \quad S_{xx}=\sum_{i=1}^n(x_i-\bar x)^2, \quad S_{xy}=\sum_{i=1}^n(x_i-\bar x)(y_i-\bar y).$$

注意到 $\hat\beta_1$ 可以写成

$$\hat\beta_1=\frac{\sum_{i=1}^n(x_i-\bar x)y_i}{S_{xx}},$$

而 $\hat\beta_0$ 可以表示为

$$\hat\beta_0=\sum_{i=1}^n\left(\frac1n-\frac{(x_i-\bar x)\bar x}{S_{xx}}\right)y_i.$$

拟合直线的方程为

$$y=\hat\beta_0+\hat\beta_1x.$$

易见该直线经过点 $(\bar x,\bar y)$.

命题 8.1. $\hat\beta_1$, $\hat\beta_0$ 分别为 β_1, β_0 的无偏估计.

证明. 由 $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$ 以及样本 x_i 非随机可得，

$$\begin{aligned}\mathrm{E}(\hat{\beta}_1) &= \frac{\sum_{i=1}^n (x_i - \bar{x})\mathrm{E}(y_i)}{S_{xx}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} \\ &= \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i}{S_{xx}} \\ &= \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2}{S_{xx}} = \beta_1.\end{aligned}$$

类似地，

$$\begin{aligned}\mathrm{E}(\hat{\beta}_0) &= \mathrm{E}(\bar{y} - \hat{\beta}_1 \bar{x}) \\ &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}(y_i) - \bar{x}\mathrm{E}(\hat{\beta}_1) \\ &= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 = \beta_0.\end{aligned}$$

□

可对 x_i 进行中心化处理，得到

$$y_i = \beta_0 + \beta_1 \bar{x} + \beta_1(x_i - \bar{x}) + \epsilon_i = \tilde{\beta}_0 + \beta_1(x_i - \bar{x}) + \epsilon_i,$$

其中 $\tilde{\beta}_0 = \beta_0 + \beta_1 \bar{x}$，其估计 $\hat{\tilde{\beta}}_0 = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$.

在 $x = x_i$ 处，回归模型的拟合值为

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

其残差为

$$e_i := y_i - \hat{y}_i.$$

残差平方和（Residual Sum of Squares, 简记为 RSS）定义为

$$\mathrm{RSS} := \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

命题 8.2. $\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n-2}$ 是 σ^2 的无偏估计.

证明是直接的，过程留给读者. 进一步计算可以得出，

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}}, \\ \mathrm{Var}(\hat{\beta}_0) &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) = \frac{\sigma^2}{S_{xx}}\frac{\sum_{i=1}^n x_i^2}{n}.\end{aligned}$$

据此可得标准误的估计为

$$\widehat{\mathrm{Se}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}},$$
$$\widehat{\mathrm{Se}}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

假设 $\epsilon_i \sim N(0, \sigma^2)$，$i = 1, 2, \ldots, n$，则 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ 且相互独立. 在此假设下，似然函数为

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right).$$

可以证明，β_1, β_0 的极大似然估计正是其最小二乘估计 $\hat{\beta}_1, \hat{\beta}_0$.

下面检验 X 与 Y 之间是否存在线性关系. 考虑假设检验：

$$H_0\colon\ \beta_1 = 0, \quad H_1\colon\ \beta_1 \neq 0.$$

选取检验统计量

$$T = \frac{\hat{\beta}_1}{\widehat{\mathrm{Se}}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}.$$

基于正态性与独立性，可证明

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1), \quad \frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2(n-2),$$

且二者相互独立，进而

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\mathrm{RSS}/\left[\sigma^2(n-2)\right]}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2).$$

在 H_0 为真的条件下，$T \sim t(n-2)$，则给定显著性水平 $\alpha > 0$，检验准则为：

$$\text{若 } |T| \geq t_{\alpha/2}(n-2), \text{ 则拒绝 } H_0,$$

即认为 β_1 显著不为零.

类似地，可以对于截距进行检验，考虑假设检验：

$$H_0\colon\ \beta_0 = 0, \quad H_1\colon\ \beta_0 \neq 0.$$

在 H_0 为真的条件下，检验统计量

$$T_0 = \frac{\hat{\beta}_0}{\widehat{\mathrm{Se}}(\hat{\beta}_0)} \sim t(n-2).$$

具体讨论这里略去.

基于 t 分布，可分别构造参数 β_1 和 β_0 的 $(1-\alpha)$-置信区间：

$$\left(\hat{\beta}_1 - t_{\alpha/2}(n-2) \cdot \widehat{\mathrm{Se}}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2}(n-2) \cdot \widehat{\mathrm{Se}}(\hat{\beta}_1)\right),$$
$$\left(\hat{\beta}_0 - t_{\alpha/2}(n-2) \cdot \widehat{\mathrm{Se}}(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\alpha/2}(n-2) \cdot \widehat{\mathrm{Se}}(\hat{\beta}_0)\right).$$

8.3 预测

本节探讨简单线性回归中的预测问题. 当自变量取新值 x_0 时，相应的响应变量为

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0,$$

其中 $\epsilon_0 \sim N(0, \sigma^2)$ 与样本误差 $\{\epsilon_i\}_{i=1}^n$ 相互独立. 记 $\mu_0 = \mathrm{E}(y_0) = \beta_0 + \beta_1 x_0$，则 μ_0 的点预测为拟合直线在 x_0 处的值，即

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) y_i.$$

直接验证可得下述结果.

命题 8.3. $\hat{y}_0$ 是 μ_0 的无偏估计，即 $\mathrm{E}(\hat{y}_0) = \mu_0$.

由于 $\hat{y}_0$ 的方差为

$$\mathrm{Var}(\hat{y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right),$$

使用残差估计 $\hat{\sigma}$ 后，有

$$\widehat{\mathrm{Se}}(\hat{y}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

且同理可得

$$\frac{\hat{y}_0 - \mu_0}{\widehat{\mathrm{Se}}(\hat{y}_0)} \sim t(n-2).$$

由此得到 μ_0 的 $(1-\alpha)$-置信区间为

$$\left(\hat{y}_0 - t_{\alpha/2}(n-2)\widehat{\mathrm{Se}}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2}(n-2)\widehat{\mathrm{Se}}(\hat{y}_0) \right)$$

实际应用中往往更关心对具体观测值 y_0 的预测. 由于 $y_0 \sim N(\mu_0, \sigma^2)$，$\mu_0$（若已知）是 y_0 在均方意义下的最优估计. 由于 μ_0 未知，用 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 代替，也是 y_0 的一个自然估计. 可以证明 y_0 与 $\hat{y}_0$ 相互独立，且 $y_0 - \hat{y}_0$ 服从正态分布. 可验证， $\mathrm{E}(y_0 - \hat{y}_0) = 0$，且

$$\mathrm{Var}(y_0 - \hat{y}_0) = \mathrm{Var}(\hat{y}_0) + \mathrm{Var}(y_0) = \sigma^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}).$$

因此

$$\frac{y_0 - \hat{y}_0}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1).$$

又 $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ 且与 $\dfrac{y_0 - \hat{y}_0}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$ 相互独立，进而有

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2).$$

可以据此给出 y_0 的 $(1-\alpha)$-置信的（双侧）区间估计为

$$\left(\hat{y}_0 - t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}, \hat{y}_0 + t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right),$$

称为预测区间.

当样本量 n 增大时，预测精度提高，但预测区间的宽度不会趋于零（因为有“1”项）. 预测精度还取决于 $(x_0-\bar{x})^2$，x_0 越接近样本中心 $\bar{x}$，预测越精确，反之则预测误差越大. 另外，模型外推时需特别谨慎，线性关系可能在数据范围外不成立，模型假设可能失效.

需要特别注意，回归模型中 X 与 Y 的地位不对等，因此不能将回归方程逆转使用.

例 8.1. 设 $(X,Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则

$$\mathrm{E}(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$
$$\mathrm{E}(X|Y=y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)$$

从第一个方程形式上解出 x，得到

$$x = \mu_1 + \frac{\sigma_1}{\rho\sigma_2}(y-\mu_2),$$

这与第二个方程不一致（除非 $\rho = \pm 1$），说明回归关系具有方向性.

若回归模型固定截距为零，即 $Y = \beta_1 X + \epsilon$，此时

- 参数估计：$\hat{\beta}_1 = \dfrac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$
- 自由度：$(n-2)$ 改为 $(n-1)$
- 置信区间和预测区间的公式需作相应调整

讨论细节留给读者补充.

参考文献

[1] Dimitri P. Bertsekas，John N. Tsitsiklis. 概率导论 [M]. 郑忠国，童行伟译. 北京: 人民邮电出版社, 2016.（原书第2版，2008）

[2] Joseph K. Blitzstein，Jessica Hwang. 概率论导论 [M]. 张景肖译. 北京: 机械工业出版社, 2019.（原书第1版，2015）

[3] George Casella, Roger Berger. *Statistical inference* [M]. 北京: 机械工业出版社, 2002.（原书第2版，2002）

[4] Samprit Chatterjee，Ali S. Hadi. 例解回归分析 [M]. 郑忠国，许静译. 北京: 机械工业出版社, 2013.（原书第5版，2012）

[5] 陈希孺. 概率论与数理统计 [M]. 合肥: 中国科技大学出版社, 2009.

[6] Kai Lai Chung，Farid AitSahlia. *Elementary Probability Theory* [M]. 4th ed. New York：Springer, 2003.

[7] Zoltán Dienes. 如何理解心理学 [M]. 孙里宁译. 上海: 华东师范大学出版社, 2018.

[8] John A. Rice. 数理统计与数据分析 [M]. 田金方译. 北京: 机械工业出版社, 2016.（原书第3版，2007）

[9] Sheldon M. Ross. 概率论基础教程 [M]. 童行伟，梁宝生译. 北京: 机械工业出版社, 2014.（原书第9版，2014）

[10] 王兆军，邹长亮. 数理统计教程 [M]. 北京: 高等教育出版社, 2014.

[11] Larry Wasserman. *All of Statistics* [M]. New York：Springer, 2004.

[12] 张尧庭，陈汉峰. 贝叶斯统计推断 [M]. 北京: 科学出版社, 1991.