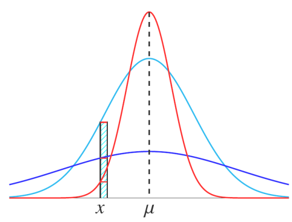
随机数学引论

第二卷 数理统计

作者：欧阳顺湘

时间：December 26, 2023

版本：3.14→ π



初稿，请多提建议；谨慎外传，以免错误流传

目录

前言

统计学是研究如何有效地收集和分析带有随机性的数据的科学与艺术. 统计学是一门与数据打交道的学科. 所以，统计发展史可以追溯到需要计数的原始社会. 但统计学成为一门系统的学科却是最近 300 年才开始的. 统计学的发展历程主要包括：古典记录统计学、近代描述统计学和现代推断统计学. 统计推断的主要内容是参数估计与假设检验. 我们将简单介绍描述统计，主要介绍推断统计. 此外，还将介绍方差分析、非参数统计和回归分析等内容. 它们与参数估计和假设检验也密切相关.

作者希望本书稿可以作为教师、学生的参考材料. 期待读者的批评指正.

第 1 章　多元正态分布

在第一卷中，我们介绍了二元正态分布. 本章我们介绍一般的多元正态分布. 在初等概率论与数理统计中，如果说正态分布如同平面几何中的直线一样重要，那么多元正态分布就如平面几何中的三角形一样重要.

1.1　记号与相关知识回顾

我们先介绍一些记号.

设 $n \geq 1$，将 $\mathbb{R}^n$ 中的向量 $\boldsymbol{x}, \boldsymbol{y}$ 视为列向量：

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

通常，为节省书写空间，利用向量或矩阵的转置符号 T，可以将 $\boldsymbol{x}, \boldsymbol{y}$ 表示为：

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}, \quad \boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}.$$

记 $\boldsymbol{x}, \boldsymbol{y}$ 的内积为

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y} = \sum_{j=1}^{n} x_j y_j.$$

使用内积符号有时可使公式的表现形式更美观.

设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$ 为概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$ 上的 n 维随机向量. 定义 $\boldsymbol{X}$ 的期望或均值为向量

$$\boldsymbol{\mu} := \mathbb{E}\boldsymbol{X} := \begin{pmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \vdots \\ \mathbb{E}X_n. \end{pmatrix}$$

用行向量转置的方式可以表示为：

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}} = \mathbb{E}\boldsymbol{X} = (\mathbb{E}X_1, \mathbb{E}X_2, \cdots, \mathbb{E}X_n)^{\mathrm{T}}.$$

利用期望的线性性，易得如下结论.

命题 1.1

设 $\boldsymbol{X}$ 为 n 维随机向量，$\boldsymbol{Y}$ 为 m 为随机向量，A 为 $m \times n$ 阶矩阵，$\nu \in \mathbb{R}^m$. 则

$$\mathbb{E}(A\boldsymbol{X} + \boldsymbol{Y} + \nu) = A\mathbb{E}\boldsymbol{X} + \mathbb{E}\boldsymbol{Y} + \nu.$$

一般地，若 $\boldsymbol{X} = (X_{ij})_{m\times n}$ 为一个 $m \times n$ 阶的随机矩阵，其第 i 行，第 j 列元素 X_{ij} 为随机变量. 定义 $\boldsymbol{X}$ 的期望（均值）为

$$\mathbb{E}\boldsymbol{X} = (\mathbb{E}X_{ij})_{m\times n} = \begin{pmatrix} \mathbb{E}X_{11} & \mathbb{E}X_{12} & \cdots & \mathbb{E}X_{1n} \\ \mathbb{E}X_{21} & \mathbb{E}X_{22} & \cdots & \mathbb{E}X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}X_{m1} & \mathbb{E}X_{n2} & \ldots & \mathbb{E}X_{mn} \end{pmatrix}.$$

随机矩阵的期望有如下性质.

命题 1.2. 随机矩阵的期望的性质

设 $\boldsymbol{X}_{n\times m}, \boldsymbol{Y}_{r\times p}$ 为随机矩阵，$A_{r\times n}, B_{m\times p}$ 为矩阵，则

$$\mathbb{E}(A\boldsymbol{X}B + \boldsymbol{Y}) = A(\mathbb{E}\boldsymbol{X})B + \mathbb{E}\boldsymbol{Y}. \tag{1.1}$$

设 $\boldsymbol{X}$ 为 n 维随机向量，$\boldsymbol{Y}$ 为 m 维随机向量，定义它们的协方差矩阵为

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) &= \mathbb{E}(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})^{\mathrm{T}} \\ &= (\mathbb{E}(X_i - \mathbb{E}X_i)(Y_j - \mathbb{E}Y_j))_{n\times m} \\ &= (\mathrm{Cov}(X_i, Y_j))_{n\times m}. \end{aligned}$$

特别，称

$$\mathrm{Var}(\boldsymbol{X}) := \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}}$$

为 $\boldsymbol{X}$ 的协方差矩阵. 若记 $\boldsymbol{\Sigma} = (\sigma_{ij})$，则它的第 i 行，第 j 列的元素 σ_{ij} 表示 X_i, X_j 的协方差，即

$$\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)], \quad i, j = 1, 2, \cdots, n.$$

协方差矩阵有如下性质.

命题 1.3. 协方差矩阵的性质

设 $\boldsymbol{X},\boldsymbol{Z}$ 为 n 维随机向量，$\boldsymbol{Y}$ 为 m 维随机向量，设 A 为 $r\times n$ 阶矩阵，B 为 $p\times m$ 阶矩阵，则

1. $\operatorname{Cov}(\boldsymbol{X}+\boldsymbol{Z},\boldsymbol{Y})=\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})+\operatorname{Cov}(\boldsymbol{Z},\boldsymbol{Y})$.
2. $\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})=\operatorname{Cov}(\boldsymbol{Y},\boldsymbol{X})^{\mathrm{T}}$.
3. $\operatorname{Cov}(A\boldsymbol{X},\boldsymbol{Y})=A\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})$.
4. $\operatorname{Cov}(\boldsymbol{X},B\boldsymbol{Y})=\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})B^{\mathrm{T}}$.
5. $\operatorname{Cov}(A\boldsymbol{X},B\boldsymbol{Y})=A\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})B^{\mathrm{T}}$.

证明

1. 显然.
2. $$\begin{aligned}\operatorname{Cov}(\boldsymbol{Y},\boldsymbol{X})^{\mathrm{T}}&=[\mathbb{E}(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})^{\mathrm{T}}]^{\mathrm{T}}\\&=\mathbb{E}([(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})^{\mathrm{T}}]^{\mathrm{T}})\\&=\mathbb{E}(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})^{\mathrm{T}}.\end{aligned}$$
3. 根据随机矩阵的期望的性质(1.1)，可得
$$\begin{aligned}\operatorname{Cov}(A\boldsymbol{X},\boldsymbol{Y})&=\mathbb{E}(A\boldsymbol{X}-\mathbb{E}(A\boldsymbol{X}))(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})^{\mathrm{T}}\\&=\mathbb{E}A(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})^{\mathrm{T}}\\&=A\mathbb{E}A(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{Y}-\mathbb{E}\boldsymbol{Y})^{\mathrm{T}}\qquad=A\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y}).\end{aligned}$$
4. $$\begin{aligned}\operatorname{Cov}(\boldsymbol{X},B\boldsymbol{Y})&=[\operatorname{Cov}(B\boldsymbol{Y},\boldsymbol{X})]^{\mathrm{T}}\\&=[B\operatorname{Cov}(\boldsymbol{Y},\boldsymbol{X})]^{\mathrm{T}}\\&=[\operatorname{Cov}(\boldsymbol{Y},\boldsymbol{X})]^{\mathrm{T}}B^{\mathrm{T}}\\&=\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})B^{\mathrm{T}}.\end{aligned}$$
5. 利用第 3，第 4 条结论即可.

□

命题 1.4

设 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 为 n 维正态随机向量，$A_{m\times n}$ 为矩阵，$\nu\in\mathbb{R}^m$，则
$$\operatorname{Var}(A\boldsymbol{X}+\nu)=A\boldsymbol{\Sigma}A^{\mathrm{T}}.$$

命题 1.5

设 $\boldsymbol{X}$ 为 n 维随机向量，则对任意 $\boldsymbol{x}\in\mathbb{R}^n$，
$$\langle\boldsymbol{x},\operatorname{Var}(X)\boldsymbol{x}\rangle=\operatorname{Var}(\langle\boldsymbol{x},\boldsymbol{X}\rangle).$$

证明

$$
\begin{aligned}
\langle \boldsymbol{x}, \mathrm{Var}(X)\boldsymbol{x}\rangle &= \langle \boldsymbol{x}, [\mathbb{E}(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})^{\mathrm{T}}]\boldsymbol{x}\rangle \\
&= \boldsymbol{x}^{\mathrm{T}}\mathbb{E}(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})^{\mathrm{T}}]\boldsymbol{x} \\
&= \mathbb{E}\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})^{\mathrm{T}}\boldsymbol{x} \\
&= \mathbb{E}(\langle \boldsymbol{x}, \boldsymbol{X}-\mathbb{E}\boldsymbol{X}\rangle)^2 \\
&= \mathbb{E}(\langle \boldsymbol{x}, \boldsymbol{X}\rangle - \mathbb{E}\langle \boldsymbol{x}, \boldsymbol{X}\rangle)^2 \\
&= \mathrm{Var}\langle \boldsymbol{x}, \boldsymbol{X}\rangle.
\end{aligned}
$$

□

上述结论说明 $\boldsymbol{X}$ 的协方差矩阵是对称非负定矩阵.

概述对称非负定矩阵的一些重要性质如下.

对 n 阶对称非负定方阵 $\boldsymbol{\Sigma}$，存在正交矩阵 P，对角阵

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_r, 0, \cdots, 0)$$

使得

$$P\Lambda P^{\mathrm{T}} = \boldsymbol{\Sigma},$$

其中 $\lambda_1, \lambda_2, \cdots, \lambda_r$ 为正的，是 $\boldsymbol{\Sigma}$ 的所有正特征值. 记

$$\boldsymbol{\Sigma}^{1/2} = P\Lambda^{1/2}P^{\mathrm{T}},$$

其中 $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_n})$. 另记 $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{1/2})^{-1}$，则 =

$$\boldsymbol{\Sigma}^{-1/2} = P\Lambda^{-1/2}P^{\mathrm{T}},$$

其中

$$\Lambda^{-1/2} = (\Lambda^{1/2})^{-1} = \mathrm{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \cdots, 1/\sqrt{\lambda_n}).$$

1.2 多元正态分布的定义

回忆一维正态分布的一个性质：设 $U \sim N(0,1)$,$\sigma \geq 0, \mu \in \mathbb{R}$，$X = \sigma U + \mu$. 若 $\sigma > 0$，则 $X \sim N(\mu, \sigma^2)$. 若 $\sigma = 0$，则 $X = \sigma U + \mu = \mu$ 也有意义，取值集中于一点，称之为退化的正态分布. 受此启发，我们通过标准的多元正态分布定义一般的多元正态分布.

定义 1.1. 标准多元正态分布

设 $U_1, U_2, \cdots, U_n$ 为独立同分布的标准正态随机变量，$\boldsymbol{U} = (U_1, U_2, \cdots, U_n)^{\mathrm{T}}$. 则称 $\boldsymbol{U}$ 服从均值为 0，协方差矩阵为 I_n 的 n 维标准正态分布，记为 $\boldsymbol{U} \sim N(0, I_n)$.

定义 1.2.（可能退化的）多元正态分布的一般定义

设 $\boldsymbol{\mu} \in \mathbb{R}^n$，$\boldsymbol{\Sigma}$ 为非负定矩阵，Γ 为 n 阶方阵，使得 $\boldsymbol{\Sigma} = \Gamma\Gamma^{\mathrm{T}}$. 设 $\boldsymbol{U} = (U_1, U_2, \cdots, U_n)^{\mathrm{T}} \sim N(0, I_n)$. 则称

$$\boldsymbol{X} = \Gamma\boldsymbol{U} + \boldsymbol{\mu}, \tag{1.2}$$

服从均值为 μ，协方差阵为 $\boldsymbol{\Sigma}$ 的 n 维正态分布，记为

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

如果 $\boldsymbol{\Sigma}$ 是奇异的，则称 $\boldsymbol{X}$ 的分布是退化的.

在定义1.2中，称 μ 为均值，$\boldsymbol{\Sigma}$ 为协方差阵下面的结论说明，它们的名字是名副其实的.

命题 1.6. 均值与协方差

设 n 为随机向量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 则

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}, \quad \boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X}).$$

即对任意的 $i, j = 1, 2, \cdot, n$，$\mu_i = \mathbb{E}X_i, \sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

证明 设 $\boldsymbol{X}$ 有如下表示

$$\boldsymbol{X} = \Gamma\boldsymbol{U} + \boldsymbol{\mu},$$

其中 $\boldsymbol{U} \sim N(0, I_n)$. 于是

$$\mathbb{E}\boldsymbol{X} = \mathbb{E}(\Gamma\boldsymbol{U} + \boldsymbol{\mu}) = \Gamma(\mathbb{E}\boldsymbol{U}) + \boldsymbol{\mu} = 0 + \boldsymbol{\mu} = \boldsymbol{\mu}.$$

又有

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(\Gamma\boldsymbol{U} + \boldsymbol{\mu}) \quad = \Gamma\,\mathrm{Cov}(\boldsymbol{U}, \boldsymbol{U})\Gamma^{\mathrm{T}} = \Gamma I_n \Gamma^{\mathrm{T}} = \boldsymbol{\Sigma}.$$

□

在定义1.2中，隐含着分布由 $\boldsymbol{\mu}$ 与 $\boldsymbol{\Sigma}$ 确定，与 Γ 的选取无关. 在 $\boldsymbol{\Sigma}$ 正定的情形，我们将证明，$\boldsymbol{X}$ 的有密度，且密度仅依赖于 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$. 在一般情形，我们将计算 $\boldsymbol{X}$ 的特征函数，特征函数仅依赖于 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$（参命题1.8）.

我们先看两个例子.

例 1.1 设 $X_1, X_2, \cdots, X_n$ 相互独立，且 $X_i \sim N(0, 1)$，记 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$. 则 $\boldsymbol{X} \sim N(0, I_n)$. $\boldsymbol{X}$ 的联合概率密度函数为

$$\begin{aligned} p(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \\ &= \frac{1}{\sqrt{\det(2\pi I_n)}} \exp\left(-\frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{x}\rangle\right). \end{aligned}$$

例 1.2 设 $X_1, X_2, \cdots, X_n$ 相互独立，且 $X_i \sim N(\mu_i, \sigma_i^2)$. $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$. 按

定义，$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Lambda)$，其中

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}}, \quad \Lambda = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2).$$

又 $\boldsymbol{X}$ 的联合概率密度函数为

$$\begin{aligned} p(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \\ &= \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \\ &= \frac{1}{\sqrt{\det(2\pi\Lambda)}} \exp\left(-\frac{1}{2}\left\langle \boldsymbol{x}-\boldsymbol{\mu}, \Lambda^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right), \end{aligned}$$

特别，若 $\sigma_i^2 = \sigma^2$，则 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \sigma^2 I_n)$，$\boldsymbol{X}$ 的联合概率密度为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{x}-\boldsymbol{\mu}\rangle\right).$$

当 Γ 非奇异时，可计算 $\Gamma\boldsymbol{U}+\boldsymbol{\mu}$ 的概率密度，但需要用到如下密度变换公式.

引理 1.1. 密度变换公式

设二元随机变量 $\boldsymbol{\Xi} = (\xi_1, \xi_2)$ 有联合概率密度函数 $f_{\boldsymbol{\Xi}}(\boldsymbol{\xi})$，$T$ 为可逆光滑映射，则 $\boldsymbol{\Phi} = T(\boldsymbol{\Xi})$ 有概率密度函数

$$f_{\boldsymbol{\Phi}}(\boldsymbol{\phi}) = f_{\boldsymbol{\Xi}}(T^{-1}\boldsymbol{\phi})|J_{T^{-1}}(\boldsymbol{\phi})|,$$

其中 $J_{T^{-1}}$ 为 T^{-1} 的 Jacobi 行列式.

命题 1.7. 正态分布的密度

设 $\boldsymbol{\Sigma}$ 为 n 阶正定矩阵，$\boldsymbol{\mu} \in \mathbb{R}^n$ 为常向量，$\boldsymbol{U} = (U_1, U_2, \cdots, U_n)^{\mathrm{T}}$，其中 $U_1, U_2, \cdots, U_n$ 为相互独立的标准正态随机变量. 设 $\boldsymbol{X} = \Gamma\boldsymbol{U}+\boldsymbol{\mu}$，其中 Γ 为满足 $\Gamma\Gamma^{\mathrm{T}} = \Sigma$ 的 n 阶方阵，则 $\boldsymbol{X}$ 的概率密度函数为

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right), \quad x \in \mathbb{R}^n. \tag{1.3}$$

证明 设 $\boldsymbol{u} = (u_1, u_2, \cdots, u_n)^{\mathrm{T}}$，则 $\boldsymbol{U}$ 的联合概率密度函数为

$$f_U(\boldsymbol{u}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\langle \boldsymbol{u}, \boldsymbol{u}\rangle\right).$$

考虑变换

$$\boldsymbol{x} = T\boldsymbol{u} := \Gamma\boldsymbol{u}+\boldsymbol{\mu}.$$

则

$$\boldsymbol{u} = T^{-1}\boldsymbol{x} = \Gamma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}).$$

因为 $\det\Sigma = \det(\Gamma\Gamma^{\mathrm{T}}) = (\det\Gamma)^2$，所以 T^{-1} 的 Jacobian 为

$$J_{T^{-1}} = \frac{1}{J_T} = \frac{1}{\det\Gamma} = \frac{1}{\sqrt{\det\Sigma}}.$$

根据引理1.1可知 $\boldsymbol{X}$ 的概率密度函数为

$$\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) =& f_{\boldsymbol{U}}(T^{-1}\boldsymbol{x})|J_{T^{-1}}(\boldsymbol{x})| \\
=& \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\left\langle\Gamma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),\Gamma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right)\cdot\frac{1}{\sqrt{\det\Sigma}} \\
=& \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}\left\langle\boldsymbol{x}-\boldsymbol{\mu},(\Gamma^{-1})^{\mathrm{T}}\Gamma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right) \\
=& \frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{1}{2}\left\langle\boldsymbol{x}-\boldsymbol{\mu},\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right).
\end{aligned}$$

□

回忆一般正态随机变量 $X \sim N(\mu,\sigma^2)$ 的概率密度函数可以写成

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

可见，它的形式与非退化的 n 维正态随机向量的概率密度函数(1.3)相同.

退化的 n 维正态分布的随机变量在 $\mathbb{R}^n$ 中没有概率密度函数，它的取值集中于 $\mathbb{R}^n$ 某子空间.

例 1.3 设 $\mu_1,\mu_2,\sigma_1>0,\sigma_2>0,-1\le\rho\le 1$ 为常数，如果二元随机向量 $\boldsymbol{X} = (X_1,X_2)^{\mathrm{T}}$ 的联合概率密度为

$$p(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}Q(x_1,x_2)\right), \tag{1.4}$$

其中

$$Q(x_1,x_2) = \frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right),$$

则称 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}. \tag{1.5}$$

命题 1.8. 正态分布的特征函数

设 $\boldsymbol{\Sigma}$ 为 n 阶非负定矩阵，$\boldsymbol{\mu}\in\mathbb{R}^n$ 为常向量，$\boldsymbol{U} = (U_1,U_2,\cdots,U_n)^{\mathrm{T}}$，其中 $U_1,U_2,\cdots,U_n$ 为相互独立的标准正态随机变量. 设 $\boldsymbol{X} = \Gamma\boldsymbol{U}+\boldsymbol{\mu}$，其中 Γ 为满足 $\Gamma\Gamma^{\mathrm{T}} = \Sigma$ 的 n 阶方阵，则 $\boldsymbol{X}$ 的概率密度函数为

设 n 维随机变量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则 $\boldsymbol{X}$ 的特征函数为

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle - \frac{1}{2}\langle \boldsymbol{t}, \Sigma \boldsymbol{t}\rangle\right), \quad \boldsymbol{t} \in \mathbb{R}^n. \tag{1.6}$$

证明 设 $\boldsymbol{t} \in \mathbb{R}^n$，则

$$\begin{aligned}
\phi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}\exp(i\langle \boldsymbol{t}, \boldsymbol{X}\rangle) = \mathbb{E}\exp(i\langle \boldsymbol{t}, \Gamma\boldsymbol{U} + \boldsymbol{\mu}\rangle) \\
&= \exp(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle)\mathbb{E}\exp(i\langle \boldsymbol{t}, \Gamma\boldsymbol{U}\rangle) \\
&= \exp(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle)\mathbb{E}\exp(i\langle \Gamma^{\mathrm{T}}\boldsymbol{t}, U\rangle) \\
&= \exp(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle)\mathbb{E}\exp(i\langle \Gamma^{\mathrm{T}}\boldsymbol{t}, U\rangle) \\
&= \exp(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle)\exp\left(-\frac{\Gamma}{1}2\langle \Gamma^{\mathrm{T}}\boldsymbol{t}, \Gamma^{\mathrm{T}}\boldsymbol{t}\rangle\right) \\
&= \exp(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle\exp\left(-\frac{1}{2}\langle \boldsymbol{t}, \Gamma\Gamma^{\mathrm{T}}\boldsymbol{t}\rangle\right) \\
&= \exp\left(i\langle \boldsymbol{t}, \boldsymbol{\mu}\rangle - \frac{1}{2}\langle \boldsymbol{t}, \boldsymbol{\Sigma}\boldsymbol{t}\rangle\right).
\end{aligned}$$

□

上述结论表明，定义1.2是无歧义的.

1.3 线性变换与多元正态分布随机变量的表示

推论 1.1

设 $\boldsymbol{\mu} \in \mathbb{R}^n$，$\boldsymbol{\Sigma}$ 为非定矩阵，$\boldsymbol{U} \sim N(0, I_n)$，则 $\Sigma^{1/2}\boldsymbol{U} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

证明 在(1.2)中取 $\Gamma = \Sigma^{1/2}$ 即可. □

设 L 为正定矩阵 $\boldsymbol{\Sigma}$ 的 Cholesky 分解，即 L 为 n 阶下三角矩阵使得 $LL^{\mathrm{T}} = \Sigma$，则由推论1.2可得多元正态随机变量的一个简单表示.

推论 1.2

设 $\boldsymbol{U} \sim N(0, I_n)$，$\boldsymbol{\mu} \in \mathbb{R}^n$，$L$ 为正定矩阵 $\boldsymbol{\Sigma}$ 的 Cholesky 分解. 则 $L\boldsymbol{U} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

例 1.4 设 $\boldsymbol{\mu} \in \mathbb{R}^2$，

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

其中 $\rho\in(-1,-1)$. 易知 $\mathbf{\Sigma}$ 的 Cholesky 分解为

$$L=\begin{pmatrix}\sigma_1 & 0\\ \rho\sigma_2 & \sqrt{1-\rho^2}\sigma_2\end{pmatrix}.$$

设 U_1,U_2 为相互独立的标准正态随机变量,

$$\begin{cases}X_1=\sigma_1U_1+\mu_1,\\ X_2=\sigma_2(\rho U_1+\sqrt{1-\rho^2}U_2)+\mu_2.\end{cases}\tag{1.7}$$

则 $\boldsymbol{X}=(X_1,X_2)^{\mathrm{T}}\sim N(\boldsymbol{\mu},\mathbf{\Sigma})$.

注 利用二元正态分布随机变量 $\boldsymbol{X}$ 的性质,不难理解(1.7)是自然的. 事实上,$\rho\frac{\sigma_2}{\sigma_1}X_1$ 可视作 X_2 在 X_1 上的投影,使得 $X_2-\rho\frac{\sigma_2}{\sigma_1}X_1$ 与 $\rho\frac{\sigma_2}{\sigma_1}X_1$ 不相关. 在我们讨论的情形,它们还是相互独立的. 从而可以将 X_2 分解为两个相互独立的随机变量之和:

$$X_2=\left(X_2-\rho\frac{\sigma_2}{\sigma_1}X_1\right)+\rho\frac{\sigma_2}{\sigma_1}X_1.$$

设 U_1,U_2 为相互独立的一维正态分布. 因 $X_2-\rho\frac{\sigma_2}{\sigma_1}X_1\sim N(\mu_2-\rho\frac{\sigma_2}{\sigma_1}\mu_1,(1-\rho^2)\sigma_2)$,与 $\sigma_2\sqrt{1-\rho^2}U_2+\mu_2-\rho\frac{\sigma_2}{\sigma_1}\mu_1$ 同分布. 而 $\rho\frac{\sigma_2}{\sigma_1}X_1\sim N(\rho\frac{\sigma_2}{\sigma_1}\mu_1,\rho^2\sigma_2^2)$, 可知 $\rho\frac{\sigma_2}{\sigma_1}X_1$ 与 $\rho\sigma_2U_1+\rho\frac{\sigma_2}{\sigma_1}\mu_1$ 同分布. 这样, 可知

$$\begin{aligned}X_2&=\rho\frac{\sigma_2}{\sigma_1}X_1+\left(X_2-\rho\frac{\sigma_2}{\sigma_1}X_1\right)\\&\sim\left(\rho\sigma_2U_1+\rho\frac{\sigma_2}{\sigma_1}\mu_1\right)+\left(\sigma_2\sqrt{1-\rho^2}U_2+\mu_2-\rho\frac{\sigma_2}{\sigma_1}\mu_1\right)\\&=\sigma_2(\rho U_1+\sqrt{1-\rho^2}U_2)+\mu_2.\end{aligned}$$

回顾 n 维多元正态分布的定义, 我们通过给出多元正态分布的概率密度函数来定义, 密度函数中涉及到 $\mathbf{\Sigma}$ 的逆, 因此要求 $\mathbf{\Sigma}$ 是可逆的.

利用命题1.7可以得到如下关于正态分布的线性变换的一般结论.

命题 1.9. 多元正态分布的线性变换

设 n 维随机变量 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\Sigma)$, A 为 n 阶方阵, 则 $A\boldsymbol{X}+\boldsymbol{\nu}\sim N(A\boldsymbol{\mu}+\boldsymbol{\nu},A\Sigma A^{\mathrm{T}})$.

证明 不妨设

$$\boldsymbol{X}=\Gamma\boldsymbol{U}+\boldsymbol{\mu},$$

其中 $\boldsymbol{U}$ 的两个分量为相互独立的标准正态随机变量, $\Gamma\Gamma^{\mathrm{T}}=\Sigma$. 则

$$A\boldsymbol{X}+\boldsymbol{\nu}=A\Gamma\boldsymbol{U}+A\boldsymbol{\mu}+\boldsymbol{\nu}.$$

因为

$$(A\Gamma)(A\Gamma)^{\mathrm{T}}=A\Gamma\Gamma^{\mathrm{T}}A^{\mathrm{T}}=A\Sigma A^{\mathrm{T}},$$

可见 $A\boldsymbol{X}+\boldsymbol{\nu}\sim N(A\mu+\boldsymbol{\nu},A\Sigma A^{\mathrm{T}})$. □

推论 1.3

设 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\Lambda)$, 其中 Λ 为 n 阶对角阵，设 P 为 n 阶正交矩阵，则 $P\boldsymbol{X}\sim N(P\boldsymbol{\mu},\Lambda)$.

证明 只要注意到 $P\boldsymbol{X}$ 的协方差阵为

$$P\Lambda P^{\mathrm{T}}=\Lambda PP^{\mathrm{T}}=\Lambda.$$

□

例 1.5 设 $X_1,X_2\sim N(\mu,\sigma^2)$，且相互独立，则二维随机向量 $\boldsymbol{X}=(X_1,X_2)^{\mathrm{T}}\sim N(\mu\mathbf{1},\sigma^2I_2)$. 其中 $\mathbf{1}$ 表示分量全为 1 的设

$$P=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta.\end{pmatrix}$$

则 $P\boldsymbol{X}\sim N(\mu P\mathbf{1},\sigma^2I_2)$. 特别，取

$$P=\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1.\end{pmatrix}$$

则可知 $\frac{1}{\sqrt{2}}(X_1+X_2)$，$\frac{1}{\sqrt{2}}(X_1-X_2)$ 都服从标准正态分布，且相互独立.

利用命题1.9可将多元正态分布随机变量标准化.

推论 1.4

设 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$ 正定，则 $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})\sim N(0,I_n)$.

进而可得如下表示.

推论 1.5

设 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$ 正定，则存在 $U\sim N(0,I_n)$ 使得 $\boldsymbol{X}=\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}+\boldsymbol{\mu}$.

证明 按推论1.4，$\boldsymbol{U}=\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})\sim N(0,I_n)$. 再利用推论1.2即可完成证明.
□

1.4 独立性

设 n 维随机向量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 对 $\boldsymbol{X}$ 进行分块分割：

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{pmatrix},$$

其中 $\boldsymbol{X}^{(1)}$ 是 m 维随机向量，$\boldsymbol{X}^{(2)}$ 是 $n-m$ 维随机向量. 类似地，设

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\boldsymbol{X}^{(1)} \\ \mathbb{E}\boldsymbol{X}^{(2)} \end{pmatrix},$$

其中 $\boldsymbol{\mu}^{(1)} \in \mathbb{R}^m, \boldsymbol{\mu}^{(2)} \in \mathbb{R}^{n-m}$,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(1)}) & \mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) \\ \mathrm{Cov}(\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(1)}) & \mathrm{Cov}(\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(2)}) \end{pmatrix},$$

其中 $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$ 分别为 $m, n-m$ 阶方阵，$\boldsymbol{\Sigma}_{12}$ 为 $m \times (n-m)$ 阶矩阵，$\boldsymbol{\Sigma}_{21}$ 为 $(n-m) \times m$ 阶矩阵.

命题 1.10. 独立性

设 n 为随机向量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 独立当且仅当 $\boldsymbol{\Sigma}_{12} = 0$.

证明

(1) 先证充分性. 设 $\boldsymbol{\Sigma}_{12} = 0$. 则

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

因 $\boldsymbol{\Sigma}$ 非负定，$\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$ 都非负定. 设 $U_1, U_2, \cdots, U_n$ 相互独立，且都服从标准正态分布，$\boldsymbol{U} = (U_1, U_2, \cdots, U_n)^{\mathrm{T}} \sim N(0, I_n)$,

$$\Gamma = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{1/2} & 0 \\ 0 & \boldsymbol{\Sigma}_{22}^{1/2} \end{pmatrix},$$

则

$$\tilde{\boldsymbol{X}} = \Gamma \boldsymbol{U} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

也就说 $\boldsymbol{X}$ 与 $\tilde{\boldsymbol{X}}$ 同分布，自然同时有 $\boldsymbol{X}^{(1)}, \tilde{\boldsymbol{X}}^{(1)}$ 同分布，$\boldsymbol{X}^{(2)}, \tilde{\boldsymbol{X}}^{(2)}$ 同分布. 因

$$\tilde{\boldsymbol{X}}^{(1)} = \Sigma_{11}\boldsymbol{U}^{(1)} + \boldsymbol{\mu}^{(1)}, \quad \tilde{\boldsymbol{X}}^{(2)} = \Sigma_{22}\boldsymbol{U}^{(2)} + \boldsymbol{\mu}^{(2)},$$

其中 $\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}$ 相互独立，可见 $\tilde{\boldsymbol{X}}^{(1)}, \tilde{\boldsymbol{X}}^{(2)}$ 相互独立. 从而 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 相互独立.

(2) 再证明必要性.

设 $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}$ 相互独立，则对任意 $i = 1, 2, \cdots, m$，$j = m+1, m+2, \cdots, n$，$\mathrm{Cov}(X_i, X_j) = 0$. 根据命题1.6，$\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. 所以 $\sigma_{ij} = 0$. 从而 $\boldsymbol{\Sigma}_{12} = 0$.

□

推论 1.6

设 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 若 $\boldsymbol{\Sigma}$ 为对角阵，则 $\boldsymbol{X}$ 的各个分量 $X_1, X_2, \cdots, X_n$ 相互独立.

推论 1.7

设 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则存在 n 维矩阵 A 使得 $\boldsymbol{Y} = A\boldsymbol{X}$ 的各个分量相互独立.

证明 对非负定矩阵 $\boldsymbol{\Sigma}$，总存在正交矩阵 A 使得 $A\Sigma A^{\mathrm{T}}$ 为对角阵. 此时 $\boldsymbol{Y}$ 的协方差矩阵为对角阵，因此 $\boldsymbol{Y}$ 的各分量相互独立. □

1.5 条件分布

引理 1.2

设 $\boldsymbol{U}, \boldsymbol{V}$ 分别为 $m, n-m$ 维随机向量，且相互独立，A 为 $m \times (n-m)$ 阶矩阵，

$$\begin{cases} \boldsymbol{X} = \boldsymbol{U} + A\boldsymbol{V}, \\ \boldsymbol{Y} = \boldsymbol{V} \end{cases}$$

则在给定 $\boldsymbol{Y} = \boldsymbol{y} \in \mathbb{R}^{n-m}$ 的条件下，$\boldsymbol{X}$ 的条件概率密度函数为 $f_{\boldsymbol{U}}(\boldsymbol{x} - A\boldsymbol{y})$，$\boldsymbol{x} \in \mathbb{R}^m$，其中 $f_{\boldsymbol{U}}$ 是 $\boldsymbol{U}$ 的概率密度函数. 即在 $\boldsymbol{Y} = \boldsymbol{y}$ 的条件下，$\boldsymbol{X}$ 的条件分布与 $\boldsymbol{U} + A\boldsymbol{y}$ 的分布相同. 且

$$f_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{x}, \boldsymbol{y}) = f_{\boldsymbol{U}}(\boldsymbol{x} - A\boldsymbol{y}) f_{\boldsymbol{V}}(\boldsymbol{y}).$$

证明 考虑变换

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

其逆变换为

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

逆变换的 Jacobian 矩阵为

$$\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y}. \end{pmatrix} = \begin{pmatrix} I_m & -A \\ 0 & I_{n-m} \end{pmatrix},$$

它的行列式为 1. 所以，由引理1.1可得，

$$\begin{aligned} f_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{x},\boldsymbol{y}) &= f_{(\boldsymbol{U},\boldsymbol{V})}(\boldsymbol{u}(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{v}(\boldsymbol{x},\boldsymbol{y})) \\ &= f_{(\boldsymbol{U},\boldsymbol{V})}(\boldsymbol{x}-A\boldsymbol{y},\boldsymbol{y}) = f_{\boldsymbol{U}}(\boldsymbol{x}-A\boldsymbol{y})f_{\boldsymbol{V}}(\boldsymbol{y}). \end{aligned}$$

由此可得在给定 $\boldsymbol{Y} = \boldsymbol{y}$ 的条件下，$\boldsymbol{X}$ 的条件分布密度为

$$\begin{aligned} f_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x}) &= \frac{f_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \\ &= \frac{f_{\boldsymbol{U}}(\boldsymbol{x}-A\boldsymbol{y})f_{\boldsymbol{V}}(\boldsymbol{y})}{f_{\boldsymbol{V}}(\boldsymbol{y})} \\ &= f_{\boldsymbol{U}}(\boldsymbol{x}-A\boldsymbol{y}). \end{aligned}$$

易知 $\boldsymbol{U} + A\boldsymbol{y}$ 的概率密度函数为 $f_{\boldsymbol{U}}(\boldsymbol{x}-A\boldsymbol{y})$，所以，给定 $\boldsymbol{Y} = \boldsymbol{y}$ 的条件下，$\boldsymbol{X}$ 的条件分布与 $\boldsymbol{U} + A\boldsymbol{y}$ 相同. □

定理 1.1. 正态分布的条件分布

设 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. 给定 $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ 的条件下，$\boldsymbol{X}^{(1)}$ 的条件分布为 $N(\mu_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})$，其中

$$\boldsymbol{\mu}_{1\cdot 2} = \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}^{(2)} - \boldsymbol{\mu}^{(2)}),$$

$$\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

2. 边缘分布 $\boldsymbol{x}^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11})$，$\boldsymbol{x}^{(2)} \sim N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$.
3. $\boldsymbol{X}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}^{(2)}$，$\boldsymbol{X}^{(2)}$ 相互独立，$\boldsymbol{X}^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}^{(1)}$，$\boldsymbol{X}^{(1)}$ 相互独立.
4. 有密度分解

$$p(\boldsymbol{x}) = p_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}=\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)})p_{\boldsymbol{X}^{(2)}}(\boldsymbol{x}^{(2)}), \quad \boldsymbol{x} \in \mathbb{R}^n, \tag{1.8}$$

其中 $p_{\boldsymbol{X}^{(2)}}(\boldsymbol{x}^{(2)})$ 为 $\boldsymbol{X}^{(2)}$ 的概率密度函数，且 $p_{\boldsymbol{X}^{(2)}}(\boldsymbol{x}^{(2)}) \sim N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$，$p_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}=\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)})$ 为给定 $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ 的条件下，$\boldsymbol{X}^{(1)}$ 的条件概率分布密度函数，且 $p_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}=\boldsymbol{x}^{(2)}}(\boldsymbol{x}^{(1)}) \sim N(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2})$.

证明 (1) 设

$$\begin{cases} \boldsymbol{Y}^{(1)} = \boldsymbol{X}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}^{(2)}, \\ \boldsymbol{Y}^{(2)} = \boldsymbol{X}^{(2)} \end{cases}$$

即设

$$\boldsymbol{Y} = B\boldsymbol{X},$$

其中

$$B = \begin{pmatrix} I_m & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ 0 & I_{n-m} \end{pmatrix}.$$

按命题1.9，$\boldsymbol{Y} \sim N(B\boldsymbol{\mu}, B\boldsymbol{\Sigma}B^{\mathrm{T}})$. 因为

$$B\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)} \\ \boldsymbol{\mu}^{(2)} \end{pmatrix}$$

$$B\boldsymbol{\Sigma}B^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{\Sigma}_{11\cdot 2} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

可知 $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ 相互独立，且

$$\boldsymbol{Y}^{(1)} \sim N(\boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{11\cdot 2}), \quad \boldsymbol{Y}^{(2)} \sim N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22}).$$

从而由

$$\begin{cases} \boldsymbol{X}^{(1)} = \boldsymbol{Y}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{Y}^{(2)}, \\ \boldsymbol{X}^{(2)} = \boldsymbol{Y}^{(2)} \end{cases}$$

并利用引理1.2可知

$$\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)} \sim L(\boldsymbol{Y}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}^{(2)}) \sim N(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}).$$

此处 $L(\xi)$ 表示 ξ 的分布.

(2) 由上述证明可见，$\boldsymbol{X}^{(2)} \sim \boldsymbol{Y}^{(2)} \sim N(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}_{22})$. 由对称性，类似可知 $\boldsymbol{X}^{(1)} \sim N(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11})$.

(3) 因为 $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ 独立，而 $\boldsymbol{X}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}^{(2)} = \boldsymbol{Y}^{(1)}$，$\boldsymbol{X}^{(2)} = \boldsymbol{Y}^{(2)}$. 由对称性，可知另一个结论也成立.

(4) 由联合分布是条件分布与边缘分布的乘积得到.

□

命题 1.11

设 n 维随机向量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$ 正定，$\boldsymbol{\Sigma}_{12} = 0$. 则 $\boldsymbol{X}$ 的联合密度有分解

$$p(\boldsymbol{x}) = p_{\boldsymbol{X}^{(1)}}(\boldsymbol{x}^{(1)})p_{\boldsymbol{X}^{(2)}}(\boldsymbol{x}^{(2)}), \quad \boldsymbol{x} \in \mathbb{R}^n.$$

推论 1.8

设 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ 为 n 维随机向量，B 为 $m \times n$ 阶矩阵，$m < n$，秩为 m，则 $B\boldsymbol{X} \sim N(B\boldsymbol{\mu}, B\Sigma B^{\mathrm{T}})$.

证明 可将矩阵 B 增添 $(n-m) \times n$ 阶矩阵 C 成为矩阵 A，即

$$A = \begin{pmatrix} B \\ C \end{pmatrix}$$

. 则

$$\boldsymbol{Y} = A\boldsymbol{X} = \begin{pmatrix} B\boldsymbol{X} \\ C\boldsymbol{X} \end{pmatrix} \sim N(A\boldsymbol{\mu}, A\Sigma A^{\mathrm{T}}).$$

因为

$$A\boldsymbol{\mu} = \begin{pmatrix} B\boldsymbol{\mu} \\ C\boldsymbol{\mu} \end{pmatrix}, \quad A\Sigma A^{\mathrm{T}} = \begin{pmatrix} B\Sigma B^{\mathrm{T}} & B\Sigma C^{\mathrm{T}} \\ C\Sigma B^{\mathrm{T}} & C\Sigma C^{\mathrm{T}} \end{pmatrix}.$$

因为 $B\boldsymbol{X}$ 成为 $A\boldsymbol{X}$ 的前 m 个分量，从而 $B\boldsymbol{X} \sim N(B\boldsymbol{\mu}, B\Sigma B^{\mathrm{T}})$. □

特别，若 B 为 $1 \times n$ 的矩阵，则有如下结论.

推论 1.9

设 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$，$\boldsymbol{b} \in \mathbb{R}^n$，则 $\langle \boldsymbol{b}, \boldsymbol{X} \rangle \sim N(\langle \boldsymbol{b}, \boldsymbol{\mu} \rangle, \langle \boldsymbol{b}, \Sigma \boldsymbol{b} \rangle)$.

1.6 另一种随机变量表示

定理 1.2

设 $\boldsymbol{\mu} \in \mathbb{R}^n$，$\boldsymbol{\Sigma}$ 为非负定 n 阶矩阵，$\boldsymbol{\Sigma}$ 的秩为 r. 则以下两种叙述等价.

1. $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，即 $\boldsymbol{X} = \Gamma \boldsymbol{U} + \boldsymbol{\mu}$，其中 $\boldsymbol{U} \sim I(0, I_n)$，$\Gamma$ 为 n 阶方阵，$\Gamma\Gamma^{\mathrm{T}} = \boldsymbol{\Sigma}$.
2. $\boldsymbol{X} = B\boldsymbol{V} + \mu$, 其中 $\boldsymbol{V} \sim N(0, I_r)$, B 为 $n \times r$ 阶矩阵, 秩为 r, $BB^{\mathrm{T}} = \boldsymbol{\Sigma}$.

证明 先证充分性.

设 $\boldsymbol{V} \sim N(0, I_r)$, 且有 $n \times r$ 阶矩阵 B 满足 $BB^{\mathrm{T}} = \boldsymbol{\Sigma}$, 同时有 $\boldsymbol{X} = B\boldsymbol{V} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

取 $\boldsymbol{W} \sim N(0, I_{n-r})$, 使得 $\boldsymbol{V}, \boldsymbol{W}$ 相互独立. 设

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{W} \end{pmatrix}, \quad \Gamma = \begin{pmatrix} B & O_{n\times(n-r)} \end{pmatrix}.$$

则 $\boldsymbol{U} \sim N(0, I_n)$, 且

$$\boldsymbol{X} = B\boldsymbol{V} + \boldsymbol{\mu} = \begin{pmatrix} B & O_{n\times(n-r)} \end{pmatrix} \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{W} \end{pmatrix} + \boldsymbol{\mu} = \Gamma\boldsymbol{U} + \boldsymbol{\mu},$$

所以, 由 $\boldsymbol{X} = B\boldsymbol{V} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 可知 $\boldsymbol{X} = \Gamma\boldsymbol{U} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 且满足

$$\Gamma\Gamma^{\mathrm{T}} = \begin{pmatrix} B & O_{n\times(n-r)} \end{pmatrix} \begin{pmatrix} B^{\mathrm{T}} \\ O_{n\times(n-r)}^{\mathrm{T}} \end{pmatrix} = BB^{\mathrm{T}} = \boldsymbol{\Sigma}.$$

再证必要性.

设 $\boldsymbol{U} \sim N(0, I_n)$, n 阶矩阵 Γ 满足 $\Gamma\Gamma^{\mathrm{T}} = \boldsymbol{\Sigma}$, 同时有 $\boldsymbol{X} = \Gamma\boldsymbol{U} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

因为 $\boldsymbol{\Sigma}$ 非负定, 秩为 r, 可知存在正交矩阵 P, 对角阵 $D = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_r)$, 其中对 $i = 1, 2, \cdots, r$, λ_i 为 $\boldsymbol{\Sigma}$ 的正特征值. 于是有

$$P\boldsymbol{\Sigma}P^{\mathrm{T}} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}.$$

设

$$\Lambda = \begin{pmatrix} D & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

$$\Lambda^{-1/2} = \begin{pmatrix} D^{-1/2} & 0 \\ 0 & I_{n-r} \end{pmatrix} := \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \cdots, \frac{1}{\sqrt{\lambda_r}}, 1, 1, \cdots, 1\right).$$

从而

$$\Lambda^{-1/2}P\boldsymbol{\Sigma}P\Lambda^{-1/2}=\Lambda^{-1/2}\begin{pmatrix}D & 0\\ 0 & I_{n-r}\end{pmatrix}\Lambda^{-1/2}=\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}.$$

它说明

$$\boldsymbol{Y}:=\Lambda^{-1/2}P(\boldsymbol{X}-\boldsymbol{\mu})\sim N\left(0,\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}\right).$$

由此可知，$\boldsymbol{Y}^{(1)}\sim N(0,I_r)$，而 $\boldsymbol{Y}^{(2)}$ 集中于 $\boldsymbol{0}$，是退化的. 设

$$P^{\mathrm{T}}\Lambda^{1/2}=\begin{pmatrix}B_{n\times r} & C_{n\times(n-r)}\end{pmatrix}.$$

其中 B,C 分别由 $P^{\mathrm{T}}\Lambda^{1/2}$ 的前 n，$n-r$ 列构成. 所以

$$\boldsymbol{X}-\boldsymbol{\mu}=P^{\mathrm{T}}\Lambda^{1/2}\boldsymbol{Y}=B\boldsymbol{Y}^{(1)}+C\boldsymbol{Y}^{(2)}=B\boldsymbol{Y}^{(1)}+C0=B\boldsymbol{Y}^{(1)}.$$

最终得到

$$\boldsymbol{X}=B\boldsymbol{V}+\boldsymbol{\mu}\sim N(\boldsymbol{\mu},BB^{\mathrm{T}}).$$

再注意到

$$\boldsymbol{\Sigma}=P^{\mathrm{T}}\Lambda^{1/2}\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}\Lambda^{1/2}P^{\mathrm{T}}=\begin{pmatrix}B & C\end{pmatrix}\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}B^{\mathrm{T}}\\ C^{\mathrm{T}}\end{pmatrix}=B^{\mathrm{T}}B$$

就可完成证明. □

1.7 特征函数

根据推论1.9，多元正态分布的分量的任意线性组合服从一元正态分布. 它可以作为多元正态分布的刻画. 事实上，根据 Cramér-Wold 机制 [1] 多元随机向量的分布由其在任意方向的一维投影的分布所决定.

命题 1.12. Cramér-Wold 机制

n 元随机向量 $\boldsymbol{X}$ 的分布由其线性组合 $Z=\sum_{i=1}^n a_iX_i$ 的分布唯一确定，其中 $a_i\in\mathbb{R}, i=1,2,\cdots,n$.

证明 Z 的特征函数为

$$\phi_Z(t)=\mathbb{E}\,\mathrm{e}^{itY}=\mathbb{E}\,\mathrm{e}^{it\sum_{i=1}^n a_iX_i}.$$

$\phi_Z(1)$ 恰好是 $\boldsymbol{X}$ 的特征函数在 $(a_1,a_2,\cdots,a_n)$ 处的值，因而确定了 $\boldsymbol{X}$ 的分布. □

命题 1.13

设 n 维随机变量 $\boldsymbol{X}$ 的均值为 $\boldsymbol{\mu}$，协方差矩阵为 Σ. 则它的特征函数形如

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\langle \boldsymbol{t},\boldsymbol{\mu}\rangle - \frac{1}{2}\langle \boldsymbol{t},\Sigma\boldsymbol{t}\rangle\right), \quad \boldsymbol{t}\in\mathbb{R}^n, \tag{1.9}$$

当且仅当对任意 $\boldsymbol{b}\in\mathbb{R}^n$，$Y=\langle \boldsymbol{b},\boldsymbol{X}\rangle$ 服从正态分布，即 Y 的特征函数为

$$\phi_Y(s) = \exp\left(is\langle \boldsymbol{a},\boldsymbol{\mu}\rangle - \frac{1}{2}s^2\langle \boldsymbol{a},\Sigma\boldsymbol{a}\rangle\right), \quad s\in\mathbb{R}.$$

证明 先证必要性. 设 $\boldsymbol{X}$ 的特征函数为(1.9)，则 $\boldsymbol{X}$ 的分量的任意线性组合 $Y=\langle \boldsymbol{a},\boldsymbol{X}\rangle$ 的特征函数为

$$\phi_Y(s) = \mathbb{E}\exp(isY) = \phi_{\boldsymbol{X}}(s\boldsymbol{a}) = \exp\left(is\langle \boldsymbol{a},\boldsymbol{\mu}\rangle - \frac{1}{2}s^2\langle \boldsymbol{a},\Sigma\boldsymbol{a}\rangle\right).$$

它表明 $\langle \boldsymbol{a},\boldsymbol{X}\rangle$ 服从均值为 $\langle \boldsymbol{a},\boldsymbol{\mu}\rangle$, 方差为 $\langle \boldsymbol{a},\Sigma\boldsymbol{a}\rangle$ 的一维正态分布.

充分性. 设对任意 $\boldsymbol{t}\in\mathbb{R}^n$，$Z=\langle \boldsymbol{t},\boldsymbol{X}\rangle$ 服从正态分布. 按引理**??**,

$$\mathbb{E}Z = \langle \boldsymbol{t},\boldsymbol{\mu}\rangle, \quad \operatorname{Var} Z = \langle \boldsymbol{t},\boldsymbol{\Sigma}\boldsymbol{t}\rangle.$$

所以 Z 的特征函数为

$$\phi_Z(s) = \mathbb{E}\exp(isZ) = \mathbb{E}\exp\left(is\langle \boldsymbol{t},\boldsymbol{\mu}\rangle - \frac{1}{2}s^2\langle \boldsymbol{t},\boldsymbol{\Sigma}\boldsymbol{t}\rangle\right).$$

注意到

$$\phi_Z(1) = \mathbb{E}\exp(iZ) = \mathbb{E}\exp(i\langle \boldsymbol{t},\boldsymbol{X}\rangle) = \phi_{\boldsymbol{X}}(\boldsymbol{t}).$$

所以(1.9)成立. □

至此，我们可以总结一般多元正态分布的等价定义.

设 $\boldsymbol{\mu}\in\mathbb{R}^n$，$\boldsymbol{\Sigma}$ 为 n 阶非负定矩阵. 则有如下五种等价描述 $\boldsymbol{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的方式（第五种方式要求 $\boldsymbol{\Sigma}$ 正定）:

1. $\boldsymbol{X}=\Gamma\boldsymbol{U}+\boldsymbol{\mu}$，其中 $\Gamma\Gamma^{\mathrm{T}}=\boldsymbol{\Sigma}$，$\boldsymbol{U}$ 为 n 维随机向量，n 分量独立同分布，都服从标准正态随机变量.
2. 设 $\boldsymbol{\Sigma}$ 的秩为 r，$\boldsymbol{X}=B\boldsymbol{V}+\mu$，其中 B 为 $n\times r$ 阶矩阵，秩为 r，$BB^{\mathrm{T}}=\boldsymbol{\Sigma}$，$\boldsymbol{V}$ 由 r 个独立同分布标准正态随机变量组成.
3. $\boldsymbol{X}$ 的特征函数为
$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\langle \boldsymbol{t},\boldsymbol{\mu}\rangle - \frac{1}{2}\langle \boldsymbol{t},\Sigma\boldsymbol{t}\rangle\right), \quad \boldsymbol{t}\in\mathbb{R}^n,$$
4. $\boldsymbol{X}$ 的分量的任意线性组合服从正态分布：对任意 $\boldsymbol{b}\in\mathbb{R}^n$，则 $\langle \boldsymbol{b},\boldsymbol{X}\rangle\sim N(\langle \boldsymbol{b},\boldsymbol{\mu}\rangle,\langle \boldsymbol{b},\boldsymbol{\Sigma}\boldsymbol{b}\rangle)$.

5. 设 $\boldsymbol{\Sigma}$ 正定. $\boldsymbol{X}$ 的概率密度函数为

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\left\langle \boldsymbol{x}-\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\rangle\right), \quad x \in \mathbb{R}^n.$$

第 1 章 习题

1.

2.

3. 设三维随机向量 $\boldsymbol{X} = (X_1, X_2, X_3)^{\mathrm{T}} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，其中

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}.$$

如果 $Y_1 = X_1 + X_2 + X_3$ 与 $Y_2 = X_1 - X_2 - X_3$ 相互独立，则 ρ 取何值?

第 2 章 数理统计基础

本章将介绍总体、样本、统计量等基本概念，并介绍一些常用统计量，如样本均值、样本方差等样本矩，以及经验分布函数，次序统计量.

2.1 总体、样本与统计量

*总体*是研究的问题所涉及的对象的全体的集合. 总体中的每一个元素称为*个体*.

我们关心总体某个或某些数量指标 X，如国民身高、体重和身体质量指数. 在得到观察值之前，我们不知道 X 的取值，但它的取值有一定的分布，因而可以当做一个随机变量. 我们可以把这个随机变量当做总体. X 的分布也叫*总体分布*.

我们的目的是用抽样的方法去了解总体，了解总体的分布或其数字特征.

*抽样*是按一定规则从总体中取出一定数量的个体的过程. 有很多抽样的方法. 如分层抽样. 它是将总体按其属性、特征分成若干层，按一定比例从各层中随机抽取样本.

如果抽取个体的过程是随机的，独立的，就称这样的抽样为*简单随机抽样*. 抽出的个体就称为（简单随机）*样本*. 如果不特别指出，本书中的抽样均指简单随机抽样.

因为抽样是随机的，所以样本也是一个随机变量. 而且，它的分布与总体分布相同. 例如，我们研究一个班同学的成绩，如果这个班的同学的成绩优秀的比例为 20%，则我们随机抽取的这位同学的成绩优秀的可能性也为 20%. 又如，某委员会有 8 个人，年龄分别为 18,20,20,25,27,27,27，28. 年龄就是我们考察的总体 X，X 取值 27 的概率为 $\frac{3}{8}$. 随机抽样一人，其年龄 X_1 等可能地取 18,20,20,25,27,27,27，28 中的一个值，特别，取值为 27 的概率为 $\frac{3}{8}$. 从此可见，样本 X_1 与总体 X 同分布.

换言之，样本是总体的一个代表（样本），其分布及其概率特征与总体的分布及相应的概率特征相同.

假设抽取了 n 个样本，n 就称为*样本容量*. 我们也称 n 个样本 $X_1, X_2, \ldots, X_n$ 为一个样本容量为 n 的样本. 因为是简单随机抽样，可知 $X_1, X_2, \ldots, X_n$ 是相互独立，与总体同分布的.

样本的具体观测值 $x_1, x_2, \ldots, x_n$ 称为样本值.

容量为 n 的样本 $X_1, X_2, \ldots, X_n$ 也可以看做随机向量 $(X_1, X_2, \ldots, X_n)$. 它的分布称为样本分布或抽样分布. 特别，总体分布就是样本量为 1 时，样本的抽样分布. 样本值 $x_1, x_2, \ldots, x_n$ 可以看做是 $\mathbb{R}^n$ 中的一个点 $(x_1, x_2, \ldots, x_n)$，称为样本点.

数理统计的主要任务之一就是从样本信息推断总体的信息，具体而言，从抽样分布的信息得到总体分布的信息.

数学上我们如下抽象地理解总体与样本.

定义 2.1. 总体与样本

设 X 为随机变量，$X_1, X_2, \cdots, X_n$ 为相互独立，且与 X 同分布的随机变量，记为

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} X,$$

则称 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的容量为 n 的简单随机样本，简称样本. 它们的观测值称为样本值，或 X 的 n 个独立观测值.

我们一般利用一个或多个统计量来提取（概括）样本信息.

定义 2.2. 统计量

设

$$T = T(X_1, X_2, \ldots, X_n)$$

为样本 $X_1, X_2, \ldots, X_n$ 的函数，如果这个函数中不含未知参数，则称之为统计量.

统计量是随机变量的样本的函数，因此是随机变量. 设 $x_1, x_2, \cdots, x_n$ 为样本观察值，则 $T(x_1, x_2, \cdots, x_n)$ 为统计量的观察值.

描述总体 X 的重要数字特征有期望 $\mathbb{E}X$，方差 $\operatorname{Var} X = E(X - \mathbb{E}X)^2$ 和标准差 $\sqrt{\operatorname{Var} X}$.

更一般地，可以定义总体的（原点）矩和中心矩.

定义 2.3. 中心矩与原点矩

设 X 为随机变量，k 为整数，设 $\mathbb{E}X^k, \mathbb{E}(X - \mathbb{E}X)^k$ 存在且有限，则我们分别称

$$\mu_k = \mathbb{E}X^k, \quad \sigma_k = \mathbb{E}(X - \mathbb{E}X)^k$$

为 X 的 k 阶总体（原点）矩，k 阶总体中心矩.

人们有时还利用偏度、峰度来描述总体分布.

定义 2.4. 偏度

$$\nu_1 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\operatorname{Var} X}}\right)^3.$$

定义 2.5. 峰度

$$\nu_2 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\operatorname{Var} X}}\right)^4.$$

特别，若 $X \sim N(\mu, \sigma^2)$，则

$$\nu_1 = 0, \quad \nu_2 = 3.$$

如果一个随机变量的峰度小于 3, 则称之为低峰态的（platykurtic），如果等于 3, 则称之为常峰态的（mesokurtic），如果大于 3, 则称之为尖峰态的（leptokurtic）. 低峰态曲线比误差曲线（正态分布）有更短的尾巴，而尖峰态曲线有更长的尾巴. 戈赛特在其以笔名 “Student” 发表的文章 [2] 中，提到一个记忆方法：分别用鸭嘴兽和袋鼠表示低峰与高峰.

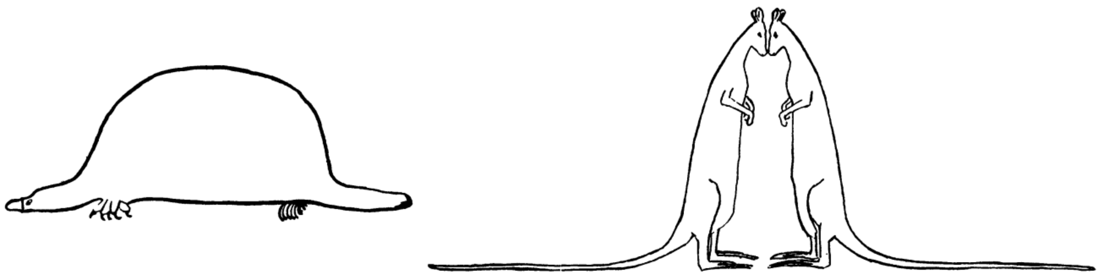


图 2.1　鸭嘴兽与袋鼠：低峰与尖峰（来自戈赛特的文章）

可以定义与总体数字特征相应的统计量.

分别于总体期望、方差对应，有样本均值

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

和样本方差

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

与标准差对应的样本统计量有样本标准差

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

定义 2.6. 样本中心矩与样本原点矩

设 $X_1, X_2, \ldots, X_n$ 为总体 X 的一个样本，k 为整数，则我们分别称

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X^k, \quad B_k = \frac{1}{n}\sum_{i=1}^{n}(X - \overline{X})^k$$

为 X 的k 阶样本（原点）矩，k 阶样本中心矩.

显然，样本一阶矩 A_1 即样本均值 $\overline{X}$. 另外，

$$A_2 - A_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2 \tag{2.1}$$

这对应于方差的计算公式 $\mathbb{E}X^2 - (\mathbb{E}X)^2 = \operatorname{Var} X$：

$$\mu_2 - \mu_1^2 = \sigma^2.$$

可以类似地定义样本偏度

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{S^3},$$

和样本峰度

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i \overline{X})^4}{S^4}.$$

从样本均值的信息可以对总体均值有所推断. 如设总体均值为 μ，则样本均值的期望为 μ. 由大数定律，在一定条件下，当 $n \to \infty$ 时，样本均值依概率收敛到 μ. 因此，我们通常将样本方差作为总体方差的近似值（估计量）. 样本方差的定义中用 $\frac{1}{n-1}$ 而非 $\frac{1}{n}$ 是因为它有如下性质 $\mathbb{E}S^2 = \sigma^2$(见(2.16))，我们称之为样本方差的无偏估计.

样本方差有一些有益的表示公式.

命题 2.1. 样本方差的表示

样本方差有如下表示

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{n}{n-1}\bar{X}^2, \tag{2.2}$$

$$S^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2. \tag{2.3}$$

证明 另外证明. □

2.2 描述统计

描述统计主要是对样本观察值的描述.

数据可以说话. 但需要正确使用，正确理解. 美国大文豪马克·吐温曾经在《我的自传》一书中写道：“数字经常欺骗我，特别是我自己整理它们时. 针对这一情况，本杰明·迪斯雷利的说法十分准确：‘世界上有三种谎言：谎言、该死的谎言、统计数字. ’” 理查德·冯·米塞斯（Richard von Misses）也曾写道：“我曾偶尔听到如下说法：‘世上有三种谎言，可验证的白谎言，无证据的一般谎言，以及统计学. ’” 这些观点说明，未经正确解释的数据，可能误导人们.

设 $x_1, x_2, \cdots, x_n$ 为样本观测值，则样本均值与样本方差，样本标准差的观测值为

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

$$s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

而样本 k 阶原点矩和 k 阶中心矩的观测值分别为

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad b_k := \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k.$$

显然

$$a_1 = \bar{x}, \quad b_2 = \frac{n-1}{n}s^2.$$

数据可视化可直观地显示数据别后的规律.

我国古代的河图洛书或许可以看做早期的数字可视化. 统计学中最早应用图形表示数据的是护理学先驱南丁格尔. 她发明了玫瑰图等工具.

表示数据的图形方法主要使用直方图、箱线图等.

2.3 经验分布函数

设 $X_1, X_2, \cdots, X_n$ 为来自 X 的样本.

定义 2.7. 经验分布函数

样本的经验分布函数定义为

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_k \le x\}}, \quad x \in \mathbb{R}.$$

经验分布函数 $F_n(x)$ 是样本值不大于 x 的频率，可以理解为随机变量取值不超过 x 这个事件发生的频率.

显然，经验分布函数是一个阶梯函数，因为

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ \dfrac{k}{n}, & X_{(k)} \le x < X_{(k+1)}, \\ 1, & x \ge X_{(n)}. \end{cases}$$

因为 $x \mapsto \mathbb{1}_{\{X_k \le x\}}$ 是增函数，所以经验分布函数 $F_n(x)$ 也是不降的：

$$F_n(x) \le F_n(y) \quad \text{a.s.}, \quad x \le y.$$

给定样本观测值 $(x_1, x_2, \cdots, x_n)$，作为 x 的函数，$F_n(x)$ 是左极限存在的右连续（即在任何 $x \in \mathbb{R}$，$F_n(x-) = \lim_{y \uparrow x F(y)}$ 存在，且 $F_n(x)$ 在 x 处右连续）增函数，是一个分布函数.

对给定的 $x \in \mathbb{R}$，$F_n(x)$ 是样本 $X_1, X_2, \cdots, X_n$ 的函数. 因为 $X_1, X_2, \cdots, X_n$ 独立同分布，可知

$$\mathbb{1}_{\{X_1 \le x\}}, \mathbb{1}_{\{X_2 \le x\}}, \cdots, \mathbb{1}_{\{X_n \le x\}}$$

相互独立同分布，与 $\mathbb{1}_{\{X \le x\}}$ 同分布，分布均为 $B(1, p)$，其中

$$p = \mathbb{P}(X \le x) = F_X(x).$$

由伯努利大数定律，当 $n \to \infty$ 时，

$$F_n(x) \xrightarrow{\mathbb{P}} \mathbb{E}\mathbb{1}_{\{X \le x\}} = p = F_X(x).$$

上述结论表明，经验分布函数依概率收敛于真实的分布函数.

利用 Kolmogorov 强大数定律，还可以证明经验分布函数几乎处处收敛于真实的分布函数.

实际上，还可以证明 $F_n(x)$ 均方收敛于 $F(x)$. 证明如下.

因为 $\mathbb{1}_{\{X_1 \le x\}}, \mathbb{1}_{\{X_2 \le x\}}, \cdots, \mathbb{1}_{\{X_n \le x\}}$ 独立同分布（即它们是 $\mathbb{1}_{\{X \le x\}}$ 的一个样本），所以，它们的平均 $F_n(x)$ 的期望 $F_X(x)$：

$$\mathbb{E}F_n(x) = \mathbb{E}\mathbb{1}_{X \le x} = \mathbb{P}(X \le x) = p = F_X(x).$$

又，$F_n(x)$ 的方差为

$$\operatorname{Var} F_n(x) = \frac{1}{n} \operatorname{Var} \mathbb{1}_{X \le x} = \frac{1}{n} p(1-p) = \frac{1}{n} F_X(x)(1 - F_X(x)).$$

所以，

$$\mathbb{E}\left(F_n(x) - F_X(x)\right)^2 = \operatorname{Var}\left(F_n(x) - F_X(x)\right) = \frac{1}{n} F_X(x)(1 - F_X(x)) \to 0, \quad n \to \infty.$$

利用棣莫弗-拉普拉斯中心极限定理，可知

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_X(x)(1 - F_X(x))}}$$

依分布收敛于服从标准正态分布的随机变量，也就是说，对充分大的 n，

$$F_n(x) - F(x) \sim N\left(F_X(x), \frac{1}{n}F_X(x)(1 - F_X(x))\right).$$

1933 年，格列坚科证明实际上，有更强的结论：

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|\right) = 1.$$

在这个基础上，柯尔莫哥洛夫-斯米尔诺夫得到了检验一个分布检验方法.（待补充于后面）

类似地，对任意 $a < b$，

$$\mathbb{1}_{\{b<X_1\le a\}}, \mathbb{1}_{\{b<X_2\le a\}}, \cdots, \mathbb{1}_{\{b<X_n\le a\}}$$

相互独立同分布，与 $\mathbb{1}_{a<\{X_n\le b\}}$ 同分布，分布均为 $B(1, p')$，其中

$$p' = \mathbb{P}(a < X \le b) = F_X(b) - F_X(a).$$

同样，由伯努利大数定律，可知当 $n \to \infty$ 时，

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{a<X_i\le b} \xrightarrow{\mathbb{P}} \mathbb{E}\mathbb{1}_{\{a<X\le b\}} = p' = F_X(b) - F_(a).$$

在描述统计中，人们常画出经验分布函数以近似分布函数；类似地，可以画出直方图以近似分布密度函数． 它们的理论依据就是上面使用大数定律做出的解释.

2.4 次序统计量

定义 2.8. 次序统计量

设 $X_1, X_2, \cdots, X_n$ 为 n 个随机变量. 对任意 $r = 1, 2, \cdots, n$，记 $X_1, X_2, \cdots, X_n$ 中第 r 个最小的为 $X_{(r)}$，即对任意样本点 ω，$X_{(r)}(\omega)$ 取值为 $X_1(\omega), X_2(\omega), \cdots, X_n(\omega)$ 按从小到大排序后，排在第 r 位的值. 即

$$X_{(1)}(\omega) \le X_{(2)}(\omega) \le \cdots \le X_{(r)}(\omega) \le \cdots \le X_{(n)}(\omega).$$

在 $X_1(\omega), X_2(\omega), \cdots, X_n(\omega)$ 中有 $r-1$ 个值不大于 $X_{(r)}(\omega)$ 的值，有 $n-r$ 个值不小于 $X_{(r)}(\omega)$ 的值，称 X_r 为 $X_1, X_2, \cdots, X_n$ 的第 r 个次序统计量.

显然

$$X_{(1)} \le X_{(2)} \le\le \cdots \le X_{(n)}.$$

因此

$$X_{(1)} = \min\{X_1, X_2, \cdots, X_n\},$$

$$X_{(n)} = \min\{X_1, X_2, \cdots, X_n\}.$$

以下设 $X_1, X_2, \cdots, X_n$ 独立同分布，为来自总体 X 的样本，分布函数为 F. 如果总体 X 是连续型随机变量，则记其概率密度函数为 f.

命题 2.2. 极值分布

极大值 $X_{(n)}$ 的分布函数为

$$F_{X_{(n)}}(x) = [F(x)]^n, \quad x \in \mathbb{R}. \tag{2.4}$$

极小值 $X_{(1)}$ 的分布函数为

$$F_{X_{(1)}} = 1 - [1 - F(x)]^n, \quad x \in \mathbb{R}. \tag{2.5}$$

证明 对极大值，有

$$\begin{aligned} F_{X_{(n)}}(x) &= \mathbb{P}(X_{(n)} \le x) \\ &= \mathbb{P}(X_1 \le x, X_2 \le x, \cdots, X_n \le x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \le x) \\ &= \prod_{i=1}^{n} F(x) \\ &= [F(x)]^n. \end{aligned}$$

对极小值，有

$$\begin{aligned} F_{X_{(1)}} &= \mathbb{P}(X_{(1)} \le x) \\ &= 1 - \mathbb{P}(X_{(1)} > x) \\ &= 1 - \mathbb{P}(X_1 > x, X_2 > x, \cdots, X_n > x) \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > x) \\ &= 1 - \prod_{i=1}^{n} [1 - F(x)] \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

□

从证明可见，$X_{(n)} \le x$ 表示 $X_1, X_2, \cdots, X_n$ 中有 n 个 X_i 的取值不超过 x，每个 X_i 的取值不超过 x 的概率为 $F(x)$.

一般地，对第 r 个次序统计量，有如下结论.

命题 2.3. 第 r 个次序统计量的分布函数

$X_{(r)}$ 的分布函数为

$$F_{X_{(r)}}(x)=\sum_{i=r}^{n}\binom{n}{i}F^{i}(x)(1-F(x))^{n-i},\quad x\in\mathbb{R}.\tag{2.6}$$

证明 设 $\omega\in\{X_{(r)}\le x\}$. 因为 $X_{(r)}(\omega)\le x$ 当且仅当

$$X_1(\omega),X_2(\omega),\cdots,X_n(\omega)$$

中至少有 r 个数不大于 x.

随机变量 $X_1,X_2,\cdots,X_n$ 中至少 r 个随机变量不大于 x 的事件是其中恰好有 $i\ge r$ 个随机变量不大于 x 的事件之并.

给定 $i\ge r$，从 $X_1,X_2,\cdots,X_n$ 中选定 i 个随机变量的组合数为 $\binom{n}{i}$.

$X_1,X_2,\cdots,X_n$ 中恰好有 $i\ge r$ 个随机变量不大于 x，意味着其中有 i 个随机变量的取值不大于 x，概率为 $F^{i}(x)$；同时另外 $n-i$ 个随机变量的取值小于 x，概率 $(1-F(x))^{n-i}$. 同时发生的概率为 $F^{i}(x)(1-F(x))^{n-i}$.

所以，

$$\begin{aligned}\mathbb{P}(X_{(r)}\le x)&=\mathbb{P}(X_1,X_2,\cdots,X_n\text{中至少有}r\text{个随机变量不大于}x)\\&=\sum_{i=r}^{n}\mathbb{P}(X_1,X_2,\cdots,X_n\text{恰好有}i\text{个随机变量不大于}x)\\&=\sum_{i=r}^{n}\binom{n}{i}F^{i}(x)(1-F(x))^{n-i}.\end{aligned}$$

□

注意到

$$\sum_{i=0}^{n}\binom{n}{i}F^{i}(x)(1-F(x))^{n-i}=[F(x)+(1-F(x))]^{n}=1,$$

可知(2.6)可以改写为

$$F_{X_{(r)}}(x)=1-\sum_{i=0}^{r}\binom{n}{i}F^{i}(x)(1-F(x))^{n-i},\quad x\in\mathbb{R}.\tag{2.7}$$

特别，由(2.7)可得(2.5).

如果总体分布有概率密度函数 f，则可以通过对次序统计量的分布函数求导得到其概率密度函数. 如从分别(2.5)，(2.4)可得极小值、极大值随机变量的概率密度函数.

命题 2.4. 极值随机变量的概率密度函数

设 $X\sim f$，则 $X_{(1)}$ 的概率密度函数为

$$f_{X_{(1)}}(x)=n[1-F(x)]^{n-1}f(x),$$

$X_{(n)}$ 的概率密度函数为

$$f_{X_{(n)}}(x)=n[F(x)]^{n-1}f(x).$$

但从(2.6)直接求导不易直接得到第 r 个次序统计量 $X_{(r)}$ 的概率密度函数. 为此寻求次序统计量的分布函数的另一表示.

记 F^{-1} 为 X 的分布函数 F 的广义逆

$$F^{-1}(y)=\sup\{x\colon F(x)\le y\}.$$

若设 U 服从 $(0,1)$ 上的均匀分布，则 $F^{(-1)}(U)$ 与 X 同分布. 所以，下面先考虑均匀分布随机变量的次序统计量.

设 $U_1,U_2,\cdots,U_n$ 为独立同分布随机变量，且 U_i 服从 $(0,1)$ 上均匀分布. 应用(2.4)于 $U_{(r)}$，可得

$$F_{U_{(r)}}(u)=\sum_{i=r}^{n}\binom{n}{i}u^i(1-u)^{n-i},\quad u\in(0,1). \tag{2.8}$$

命题 2.5. 均匀随机变量的次序统计量的概率密度函数

$U_{(r)}$ 的概率密度函数为

$$f_{U_{(r)}}(u)=\frac{n!}{(r-1)!(n-r)!}u^{r-1}(1-u)^{n-r}\mathbb{1}_{(0,1)}(u)=\frac{1}{B(r,n-r+1)}u^{r-1}(1-u)^{n-r}\mathbb{1}_{(0,1)}(u).$$

证明

$$\begin{aligned}
&f_{U_{(r)}}(u)\\
=&\frac{d}{du}F_{U_{(r)}}(u)\\
=&\sum_{i=r}^{n}\binom{n}{i}iu^{i-1}(1-u)^{n-i}-\sum_{i=r}^{n}\binom{n}{i}(n-i)u^{i-1}(1-u)^{n-i-1}\\
=&n\sum_{i=r}^{n}\binom{n-1}{i-1}u^{i-1}(1-u)^{(n-1)-(i-1)}-n\sum_{i=r}^{n}\binom{n-1}{i}u^{i-1}(1-u)^{(n-1)-i}\\
=&n\sum_{i=r-1}^{n}\binom{n-1}{i}u^{i}(1-u)^{(n-1)-i}-n\sum_{i=r}^{n}\binom{n-1}{i}u^{i-1}(1-u)^{(n-1)-i}\\
=&n\binom{n-1}{r-1}u^{r-1}(1-u)^{(n-1)-(r-1)}\\
=&\frac{n!}{(r-1)!(n-r)!}u^{r-1}(1-u)^{n-r}.
\end{aligned}$$

□

根据上述结论，$U_{(r)}$ 服从参数为 $r, n-r+1$ 的 Beta 分布．所以

$$F_{U_{(r)}}(u) = \int_0^u \frac{1}{B(r, n-r+1)} t^{r-1}(1-u)^{n-r} dt, \quad 0 < u < 1. \tag{2.9}$$

利用(2.6)，(2.8)可知

$$F_{U_{(r)}}(F(x)) = \sum_{i=r}^{n} \binom{n}{i} [F(x)]^i [(1-F(x)]^{n-i} = F_{X_{(r)}}(x),$$

从而在利用(2.9)可得 $X_{(r)}$ 的分布函数的另一表达.

命题 2.6. 次序统计量的分布函数的另一表示

$$F_{X_{(r)}}(x) = \int_0^{F(x)} \frac{1}{B(r, n-r+1)} t^{r-1}(1-u)^{n-r} dt, \quad x \in \mathbb{R}. \tag{2.10}$$

也可以以如下方式推导.

由于 $X_i \sim F^{-1}(U_i)$，且 $F^{-1}(U_1), F^{-1}(U_2), \cdots, F^{-1}(U_n)$ 相互独立，以及 F^{-1} 的单调性，可知

$$F^{-1}(U_{(r)}) \sim X_{(r)}.$$

由此可得

$$\begin{aligned}\mathbb{P}(X_{(r)} \le x) &= \mathbb{P}(F^{(-1)}(U_{(r)}) \le x) = \mathbb{P}(U_{(r)} \le F(x)) \\ &= \int_0^{F(x)} \frac{1}{B(r, n-r+1)} u^{r-1}(1-u)^{n-r} du.\end{aligned}$$

通过求导可从积分形式的次序统计量的分布函数(2.10)得到次序统计量的概率密度函数.

命题 2.7. 次序统计量的分布密度

设 F 有概率密度函数 f，

$$f_{X_{(r)}}(x) = \frac{1}{B(r, n-r+1)} f(x) F^{r-1}(x)(1-F(x))^{n-r}, \quad x \in \mathbb{R}.$$

如何直观地理解这个结果呢?

回忆对连续型随机变量 X，设它的概率密度函数为 f，则对充分小的 $\varepsilon > 0$，有

$$f(x)\varepsilon \approx \mathbb{P}(X \in (x, x+\varepsilon]).$$

我们先直观考虑 $X_{(n)}$ 的密度．因为对充分小的 $\varepsilon > 0$，

$$f_{X_{(n)}}(x) \approx \mathbb{P}(X_{(n)} \in (x, x+\varepsilon]).$$

故设 $X_{(n)}$ 落于 $(x, x+dx]$ 中. 这意味着 $X_1, X_2, \cdots, X_n$ 中有一个（一共有 n 中取

法）随机变量落在 $(x, x+\varepsilon]$ 中（相应的概率近似为 $f(x)\varepsilon$），其他 $n-1$ 个随机变量都落在 x 的左边，即取值小于或等于 x（相应的概率为 $F^{n-1}(x)$）. 因此，

$$\begin{aligned}
f_{X_{(n)}}(x)\varepsilon \approx & \mathbb{P}(X_{(n)} \in (x, x+\varepsilon]) \\
& = \mathbb{P}(X_1, X_2, \cdots, X_n \text{中某个随机变量} \in (x, x+\varepsilon], \text{其他} \le x) \\
& = \sum_{i=1}^{n} \mathbb{P}(X_i \in (x, x+\varepsilon], \text{其他} \le x) \\
& = n\mathbb{P}(X_1 \in (x, x+\varepsilon])\mathbb{P}(X_2 \le x, X_3 \le x, \cdots, X_n \le x) \\
& \approx nf(x)\varepsilon \cdot \prod_{i=2}^{n} \mathbb{P}(X_i \le x) \\
& = nf(x)\varepsilon \cdot \prod_{i=2}^{n} F(x) \\
& = n(F(x))^{n-1} f(x)\varepsilon.
\end{aligned}$$

所以，$X_{(n)}$ 的概率密度函数为

$$f_{X_{(n)}}(x) = nF^{n-1}(x)f(x).$$

类似地可以直观得到 $X_{(1)}$ 的概率密度函数. 设 $X_{(1)}$ 落于 $(x-\varepsilon, x]$ 中，这意味着 $X_1, X_2, \cdots, X_n$ 中有一个（一共有 n 种取法）随机变量落在 $(x-\varepsilon, x]$ 中（相应的概率近似为 $f(x)\varepsilon$），其他的 $n-1$ 个随机变量都落在 x 的右边（相应的概率为 $(\mathbb{P}(X > x))^{n-1} = (1-F(x))^{n-1}$）. 因此，$X_{(1)}$ 的概率密度函数为

$$f_{X_{(1)}}(x) = n(1-F(x))^{n-1} f(x).$$

$X_{(r)}$ 的密度如下考虑：设 $X_{(r)}$ 落在 $(x, x+\varepsilon]$ 中，则 $X_1, X_2, \cdots, X_n$ 中有一个（一共 n 种取法）落在 $(x, x+\varepsilon]$ 中，剩下的 $n-1$ 个随机变量中有 $k-1$ 个（一共有 $\binom{n-1}{k-1}$ 种取法）落在 x 的左边（相应的概率为 $F^{n-1}(x)$），剩下来的 $n-k$ 个随机变量都落在 $x+\varepsilon$ 的右边（相应的概率为 $(1-F(x))^{n-k}$）. 总的取法为

$$n\binom{n-1}{k-1} = n\frac{(n-1)!}{(k-1)!(n-k)!} = \frac{n!}{(k-1)!(n-k)!}.$$

因此，$X_{(r)}$ 的概率密度函数为

$$f_{X_{(r)}}(x) = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(x)(1-F(x))^{n-k} f(x).$$

利用上述“微元”法，次序统计量的联合分布也可以计算.

如极小值与极大值的联合分布概率密度函数计算如下.

设 $x - dx \le X_{(1)} \le x, y \le X_{(n)} \le y + dy$. 这意味着 $X_1, X_2, \cdots, X_n$ 中有一个随机变量（n 种取法）位于 x 的左边（相应的概率为 $nf(x)dx$），一个随机变量（$n-1$ 种取法）位于 y 的右边（相应的概率为 $(n-1)f(y)dy$），其余的随机变量在 x, y 之间（相应的概率为 $(F(y)-F(x))^{n-2}$）. 因此，我们可得下面的结论.

命题 2.8

$X_{(1)}, X_{(n)}$ 的联合概率密度函数为

$$f_{1,n}(x,y) = n(n-1)f(x)f(y)(F(y)-F(x))^{n-2}, \quad x \le y.$$

特别，若 F 为 $[0,1]$ 上的均匀分布，则

$$f_{1,n}(x,y) = n(n-1)(y-x)^{n-2}\mathbb{1}_{\{(u,v)\in\mathbb{R}^2:0\le u\le v\le 1\}}, \quad x,y \in \mathbb{R}.$$

由极值的联合概率密度函数可得极差 $R = X_{(n)} - X_{(1)}$ 的概率密度密度函数

$$f_R(r) = \int_{-\infty}^{+\infty} f_{1,n}(x+r,x)dx.$$

一般地，有下面的结论.

命题 2.9. 次序统计量的联合概率密度函数

设 $1 \le r_1 \le r_2 \le \cdots \le r_k \le n$，$1 \le k \le n$，则 $X_{(r_1)}, X_{(r_2)}, \cdots, X_{(r_k)}$ 的联合概率密度函数为

$$\begin{aligned}&f_{(X_{(r_1)},X_{(r_2)},\cdots,X_{(r_k)})}(x_1,x_2,\cdots,x_k)\\ =&n!\prod_{i=0}^{k}\frac{F(x_{i+1}-F(x_i))^{r_{i+1}-r_i-1}}{r_{i+1}-r_i-1}\prod_{i=0}^{k}f(x_i), \quad x_1\le x_2\le\cdots\le x_k,\end{aligned} \tag{2.11}$$

其中 $x_0 = -\infty$，$x_{k+1} = +\infty$，$r_0 = 0$，$r_{k+1} = n+1$.

由上述结论可特别得到如下推论.

推论 2.1. 两个次序统计量的联合分布

设 $1 \le r < s \le n$，则 $X_{(r)}$ 与 $X_{(s)}$ 的联合概率密度函数为

$$\begin{aligned}f_{(X_{(r)},X_{(s)})}(x,y) =& \frac{n!}{(r-1)!(s-r-1)!(n-s)!}\\ &\times (F(x))^{r-1}(F(y)-F(x))^{s-r-1}(1-F(y))^{n-s}f(x)f(y), \quad x\le y.\end{aligned} \tag{2.12}$$

推论 2.2. 前 r 个次序统计量的联合分布

设 $1 \le r \le n$，则 $X_{(1)}, X_{(2)}, \cdots, X_{(r)}$ 的联合概率密度函数为

$$\begin{aligned}&f_{(X_{(1)},X_{(2)},\cdots,X_{(r)})}(x_1,x_2,\cdots,x_r)\\ =&\frac{n!}{(n-r)!}(1-F(x_r))^{n-r}\prod_{i=1}^{r}f(x_i), \quad x_1\le x_2\le\cdots\le x_r.\end{aligned}$$

特别，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 的联合概率密度函数为

$$\begin{aligned}&f_{(X_{(1)},X_{(2)},\cdots,X_{(n)})}(x_1,x_2,\cdots,x_n)\\ =&n!f(x_1)f(x_2)\cdots f(x_n), \quad x_1\le x_2\le\cdots\le x_n.\end{aligned}$$

为简单起见，我们假设 X 是连续型随机变量. 否则中位数的定义稍微复杂些.

定义 2.9. 中位数

如果 η 满足

$$\mathbb{P}(X \geq \eta) = \mathbb{P}(X \leq \eta) = \frac{1}{2},$$

则称之为 X 的中位数（median）.

类似于期望使得均方误差 $\mathbb{E}(X-c)^2$ 最小的常数 c，中位数是使得绝对误差 $E|X-c|$ 最小的常数 c.

样本中位数的定义依赖于样本量.

定义 2.10. 样本中位数

样本中位数 M 定义为

$$M = \begin{cases} X_{\frac{1}{2}(n+1)}, & n\text{是奇数}, \\ \frac{1}{2}[X_{(\frac{1}{2}n)} + X_{(\frac{1}{2}n+1)}], & n\text{是偶数}. \end{cases}$$

当 $n = 2m+1$ 时，样本中位数的概率密度函数为

$$\frac{(2m+1)!}{m!m!}[F(x)]^m[1-F(x)]^m f(x), \quad x \in \mathbb{R}.$$

练 习 2.4

1. 设 $U_1, U_2, \cdots, U_n$ 独立同分布与区间 $(0,1)$ 上的均匀分布，$1 \leq r < s \leq n$ 求 (U_r, U_s) 的联合概率密度函数，并求 $U_{(s)} - U_{(r)}$ 的概率密度函数. 特别，求极差 $R = U_{(n)} - U_{(1)}$ 的概率密度函数. 并指出它们与 Beta 分布的联系.

2.5 χ^2 分布

定义 2.11. χ^2 分布

设 $Z_1, Z_2, \ldots, Z_n$ 相互独立同分布，$Z_i \sim N(0,1)$，则称

$$U = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

的分布为服从自由度为 n 的χ^2 分布，记为

$$U \sim \chi^2(n).$$

χ^2 分布一般读作 Chi 方分布或卡方分布（Chi-Squared distribution）. χ^2 分布的概率密度函数如下：

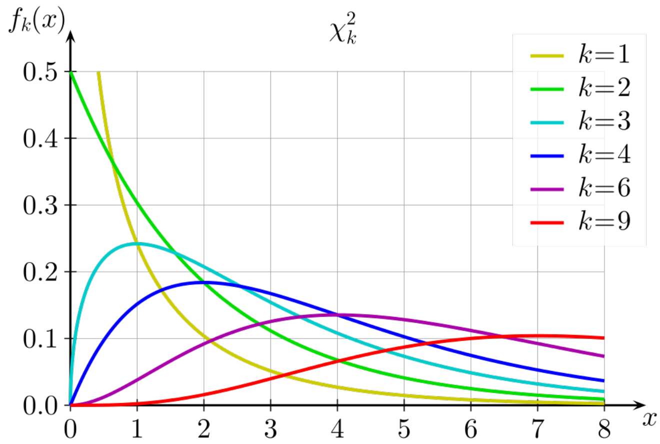


图 2.2 χ^2 分布的概率密度函数

命题 2.10. χ^2 分布的概率密度函数

设 X 服从自由度为 n 的 χ^2 分布，则 X 的概率密度为

$$f(x,n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}x^{(n-2)/2}\,\mathrm{e}^{-x/2}. & x\geq 0,\\ 0, & x<0,\end{cases} \tag{2.13}$$

它恰为参数为 $\frac{n}{2},\frac{1}{2}$ 的 Γ 分布.

证明 [方法一] 按 Γ 分布的定义以及卷一中对标准正态随机变量的平方的分布密度的计算结果可知，自由度为 1 的 χ^2 分布恰为参数为 $\frac{1}{2},\frac{1}{2}$ 的 Γ 分布，即

$$Z_1^2\sim\Gamma\left(\frac{1}{2},\frac{1}{2}\right).$$

因此，由 Γ 分布的可加性可知 U 服从参数为 $\frac{n}{2},\frac{1}{2}$ 的 Γ 分布，

$$U\sim\Gamma\left(\frac{n}{2},\frac{1}{2}\right).$$

□

证明 [方法二] 设 $u\in\mathbb{R}$，

$$B(u):=\{(z_1,z_2,\cdots,z_n):\sum_{i=1}^n z_i^2\leq u\},$$

则利用极坐标变化，可得

$$\begin{aligned}\mathbb{P}(U\leq u)&=(2\pi)^{-n/2}\int_{B(u)}\mathrm{e}^{-\frac{1}{2}\sum_{i=1}^n z_i^2}\,dz_1dz_2\cdots dz_n\\&=c_n\int_0^{\sqrt{u}}\mathrm{e}^{-\frac{1}{2}r^2}r^{n-1}dr,\end{aligned}$$

其中 c_n 为常数使得

$$1=\mathbb{P}(U<\infty)=c_n\int_0^{+\infty}\mathrm{e}^{-\frac{1}{2}r^2}r^{n-1}dr.$$

所以，

$$c_n = 2^{1-\frac{n}{2}}\Gamma\left(\frac{n}{2}\right).$$

U 的分布密度为

$$\begin{aligned} f_U(u) &= \frac{d}{du}\mathbb{P}(U \le u) \\ &= c_n\, \mathrm{e}^{-\frac{1}{2}u}\, u^{\frac{n}{2}-\frac{1}{2}} \cdot \frac{1}{\sqrt{u}} \\ &= \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}\, \mathrm{e}^{-\frac{u}{2}}\, u^{\frac{n}{2}-1}. \end{aligned}$$

□

推论 2.3

自由度为 2 的 χ^2 分布是参数为 $\frac{1}{2}$ 的指数分布.

利用 χ^2 分布与 Γ 分布的关系以及 Γ 分布的性质，可得 χ^2 分布有如下常用性质：

命题 2.11

设 $U \sim \chi^2(n)$，则

1. 矩母函数为 $M(t) = \mathbb{E}\,\mathrm{e}^{tX} = (1-2t)^{-n/2}$.
2. 设 $V \sim \chi^2(m)$ 且 U, V 相互独立，则 $U + V \sim \chi^2(n+m)$.

命题 2.12

设 $U \sim \chi^2(n)$，则

$$\mathbb{E}U = n, \quad \mathrm{Var}\, U = 2n.$$

证明 设 $Z_1, Z_2, \ldots, Z_n$ 相互独立同分布，$Z_i \sim N(0,1)$，则 $U \sim \sum_{i=1}^n Z_i^2$. 由例**??**可知

$$\mathbb{E}Z_i^2 = 1$$

以及

$$\mathrm{Var}\, Z_i^2 = \mathbb{E}Z_i^4 - (\mathbb{E}Z_i^2)^2 = 3 - 1^2 = 2.$$

所以，

$$\begin{aligned} \mathbb{E}U &= \mathbb{E}\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \mathbb{E}Z_i^2 = \sum_{i=1}^n 1 = n, \\ \mathrm{Var}\, U &= \mathrm{Var}\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \mathrm{Var}\, Z_i^2 = \sum_{i=1}^n 2 = 2n. \end{aligned}$$

□

2.6 样本均值与样本方差的分布

下面这个定理非常关键.

定理 2.1

设 $X_1, X_2, \ldots, X_n$ 为 $N(\mu, \sigma^2)$ 的一个样本，则

1.
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

2.
$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

3. S^2 与 $\bar{X}$ 独立.

在证明结论之前，我们一个有启发意义的例子和一个有启发意义的分解.

例 2.1 考虑 $n = 2$ 的情形. 设 X_1, X_2 为来自正态总体 $N(\mu, \sigma^2)$ 的样本. 此时样本均值为

$$\bar{X} = \frac{1}{2}(X_1 + X_2),$$

样本方差为（参公式(2.2)或直接用(2.3)）

$$S^2 = \frac{1}{2-1}[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2] = \frac{1}{4}(X_1 - X_2)^2,$$

根据例1.5，$X_1 + X_2$ 与 $X_1 - X_2$ 独立，所以 S^2 与 $\bar{X}$ 独立. 此外，

$$\frac{(n-1)S^2}{\sigma^2} = \frac{2(X_1 - X_2)^2}{4\sigma^2} \sim \chi^2(1).$$

例 2.2 设 $X_1, X_2, \cdots, X_n$ 为来自 $X \sim N(\mu, \sigma^2)$ 的样本. 由正态分布的线性组合仍是正态分布，可得

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\bar{X}\right),$$

因此

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

所以

$$\frac{n(\bar{X} - \mu)}{\sigma^2} \sim \chi^2(1). \tag{2.14}$$

因为

$$\frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

且 $\frac{X_1-\mu}{\sigma}, \frac{X_2-\mu}{\sigma}, \cdots, \frac{X_n-\mu}{\sigma}$ 相互独立，故

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n). \tag{2.15}$$

因为

$$\begin{aligned}
&\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \\
=&\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 \\
=&\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2+\frac{n(\bar{X}-\mu)}{\sigma^2} \\
=&\frac{(n-1)S^2}{\sigma^2}+\frac{n(\bar{X}-\mu)^2}{\sigma^2}.
\end{aligned}$$

也就是说有如下分解

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

注意到上式左边服从 $\chi^2(n)$ 分布，右边第二式服从 $\chi^2(1)$ 分布，而且通过计算右边两项的协方差为零证明它们相互独立，从而可以（如用矩母函数或概率密度）证明右边第一式服从自由度为 $n-1$ 的 χ^2 分布，即 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

证明 [定理2.1的矩阵证明] 取 n 阶正交矩阵 A 使得它的第一行为 $\frac{1}{\sqrt{n}}(1,1,\cdots,1)$. 考虑变换

$$Y := (Y_1,Y_2,\cdots,Y_n)^T := A(X_1,X_2,\cdots,X_n)^T.$$

于是，$Y_1,Y_2,\cdots,Y_n$ 相互独立，且

$$Y_i \sim N\left(\left(\sum_{j=1}^{n}a_{ij}\right)\mu,\sigma^2\right).$$

对任意的 $j=1,2,\cdots,n$，$a_{1j}=\frac{1}{\sqrt{n}}$，所以

$$\sum_{j=1}^{n}a_{1j}=n\cdot\frac{1}{\sqrt{n}}=\sqrt{n}.$$

而对 $i=2,3,\cdots,n$，

$$\sum_{j=1}^{n}a_{ij}=\sqrt{n}\cdot\sum_{j=1}^{n}a_{ij}\cdot\frac{1}{\sqrt{n}}=\sqrt{n}\cdot\sum_{j=1}^{n}a_{ij}\cdot a_{1j}=\sqrt{n}(AA^T)_{i1}=0.$$

上式用到如下结论：正交矩阵 A 的第 i 行与第 1 行正交. 事实上，$AA^T=I$ 的第 i（$i\neq 1$）行，第 1 列的元素为 0，表明 A 的第 i 与 A^T 的第 1 列正交. 而 A^T 的第 1 列即 A 的第 1 行. 所以，

$$Y_1 \sim N(\sqrt{n}\mu,\sigma^2), \quad Y_i \sim N(0,\sigma^2), \quad i=2,3,\cdots,n.$$

注意到

$$\begin{aligned}\bar{X} &= \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{\sqrt{n}}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{n}}X_j \\ &= \frac{\sqrt{n}}{n}\sum_{j=1}^{n} a_{1j}X_j = \frac{\sqrt{n}}{n}(AX)_1 = \frac{\sqrt{n}}{n}Y_1.\end{aligned}$$

所以，利用 $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$ 可得

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right).$$

这就证明了第一条结论.

正交变换不改变距离：

$$\sum_{i=1}^{n} Y_i^2 = Y^T Y = (AX)^T(AX) = X^T(A^T A)X = X^T X = \sum_{i=1}^{n} X_i^2.$$

因此，

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.\end{aligned}$$

注意到

$$\frac{Y_i}{\sigma} \sim N(0,1), \quad i = 2, 3, \cdots, n$$

可知

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i}{\sigma}\right)^2$$

是 $n-1$ 个独立的标准正态随机变量之和，因此服从自由度为 $n-1$ 的 χ^2 分布. 这证明了第二个结论.

注意到 $\bar{X}$ 仅依赖于 Y_1，S^2 仅依赖于 $Y_2, Y_3, \cdots, Y_n$，而 Y_1 与 $Y_2, Y_3, \cdots, Y_n$ 独立，所以 $\bar{X}$ 与 $\frac{(n-1)S^2}{\sigma^2}$ 独立. 这证明了第三个结论. □

注 上述证明中的正交矩阵可以具体构造如下：

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & \cdots & 0 & 0 & 0 \\ \frac{1}{\sqrt{(n-1)(n-2)}} & \frac{1}{\sqrt{(n-1)(n-2)}} & \frac{1}{\sqrt{(n-1)(n-2)}} & \frac{1}{\sqrt{(n-1)(n-2)}} & \cdots & \frac{1}{\sqrt{(n-1)(n-2)}} & \frac{-(n-2)}{\sqrt{(n-1)(n-2)}} & 0 \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}$$

该矩阵第 i 行第 j 列元素 a_{ij} 为

$$a_{ij} = \begin{cases} \dfrac{1}{\sqrt{i(i-1)}}, & j \le i, \\ \dfrac{-(i-1)}{\sqrt{i(i-1)}}, & j = i, \\ 0, & j > i. \end{cases}$$

2.6.1 样本方差的期望与方差

推论 2.4. 样本方差的期望与方差

样本方差的期望为

$$\mathbb{E}S^2 = \sigma^2, \tag{2.16}$$

证明 [(2.16)的直接证明] 利用(2.3)，可得

$$\begin{aligned} \mathbb{E}S^2 &= \mathbb{E}\left(\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2\right) \\ &= \frac{1}{2n(n-1)}\sum_{i,j=1}^{n}\mathbb{E}(X_i - X_j)^2 \\ &= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Var}(X_i - X_j) \\ &= \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}2\sigma^2 \\ &= \frac{1}{2n(n-1)}\cdot n(n-1)\cdot 2\sigma^2 = n\sigma^2. \end{aligned}$$

□

推论 2.5. 样本方差的方差

设 $X_1, X_2, \cdots, X_n$ 独立同分布于 $N(\mu, \sigma^2)$，则样本方差的方差为

$$\operatorname{Var} S^2 = \frac{2\sigma^4}{n-1}, \tag{2.17}$$

证明 因为

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

根据 χ^2 分布的方差计算公式可得

$$\mathbb{E}\frac{(n-1)S^2}{\sigma^2} = n-1, \quad \operatorname{Var}\frac{(n-1)S^2}{\sigma^2} = 2(n-1).$$

所以

$$\mathbb{E}S^2 = \sigma^2, \quad \operatorname{Var} S^2 = \frac{2\theta^4}{n-1}.$$

□

2.7 t 分布

t 分布由戈赛特提出．他当时在从事啤酒酿造行业，他想知道发酵时加多少酵母最合适．他经过研究后以笔名“student”在 1908 年发表文章 The Probable error of the mean，提出了 t 分布，贡献了小样本检验的思想．

图 2.3 戈赛特，笔名“学生”

定义 2.12. t 分布

设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，X, Y 相互独立，则称

$$t := \frac{X}{\sqrt{\frac{Y}{n}}}$$

服从自由度为 n 的t 分布（或学生分布），记为

$$t \sim t(n).$$

命题 2.13. t **分布的概率密度函数**

自由度为 n 的 t 分布随机变量的概率密度函数如下

$$f(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

证明 它是两个分布易得的独立随机变量的商的分布. □

可以注意到，T 的分布 t 分布可以看做为正态分布 $N(0, W)$ 与随机变量 W 的混合分布，其中 W 服从逆 Gamma 分布. 具体而言，设

$$T|W \sim N(0, W)$$

而 W 服从逆 Gamma 分布：

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T|W}(t|w) f_W(w) dw \\ &\propto \frac{1}{\sqrt{2\pi w}} \exp\left(-\frac{x^2}{2w}\right) \cdot w_{-\frac{n}{2}-1} \exp\left(-\frac{n}{2w}\right) dw \\ &\propto \int_0^\infty w^{-\frac{n+1}{2}-1} \exp\left(\frac{x^2+n}{2}\right)^{-\frac{n+1}{2}} dw. \end{aligned}$$

易从 t 分布的定义知道它的概率密度函数是关于纵轴左右对称的. 通过作图，还可以见到 t 分布的分布密度曲线是钟形曲线.

从分布密度表达式可知，自由度为 1 的 t 分布就是标准柯西分布. 按定义，设 X, Y 为相互独立的标准正态随机变量，则 $\frac{X}{|Y|}$ 服从自由度为 1 的 t 分布. 我们曾证明 $\frac{X}{Y}$ 服从柯西分布. 这两个结论不矛盾，因为容易见到 $\frac{X}{Y}$ 与 $\frac{X}{|Y|}$ 同分布.

当 $n \to \infty$ 时，t 分布的概率密度函数逐点收敛于标准正态分布密度函数：对任意 $t \in \mathbb{R}$，

$$\lim_{n\to\infty} f(t) = \varphi(t).$$

证明参习题解答. 一个直观解释如下：因为 Y 服从自由度为 n 的 χ^2 分布，不妨视之为 $Z_1^2 + Z_2^+ \cdots Z_n^2$，其中 $Z_1, Z_2, \cdots, Z_n$ 独立同分布，且 $Z_i \sim N(0, 1)$. 由大数定律，$\frac{1}{n}V$ 依概率收敛到 $\mathbb{E}Z_1^2 = 1$. 从而可知，当 n 充分大时，T 近似于 U.

根据 Scheffe 定理，服从自由度为 n 的 t 分布的随机变量依分布收敛于标准正态随机变量.

一般，当 $n \geq 20$ 时，$f(t)$ 与 $\varphi(t)$ 非常接近.

当 $|t| \to \infty$，$f(t)$ 趋于 0 的速度比 $\varphi(t)$ 慢.

当 n 比较小时，t 分布的作用凸显. 因此，戈塞特被认为小样本理论之父. t 分布的尾概率比标准正态分布的尾概率大，即对充分大的 $x > 0$，

$$\mathbb{P}(|t| \geq x) \geq \mathbb{P}(|Z| \geq x) = 2\Phi(-x),$$

其中 $Z \sim N(0, 1)$.

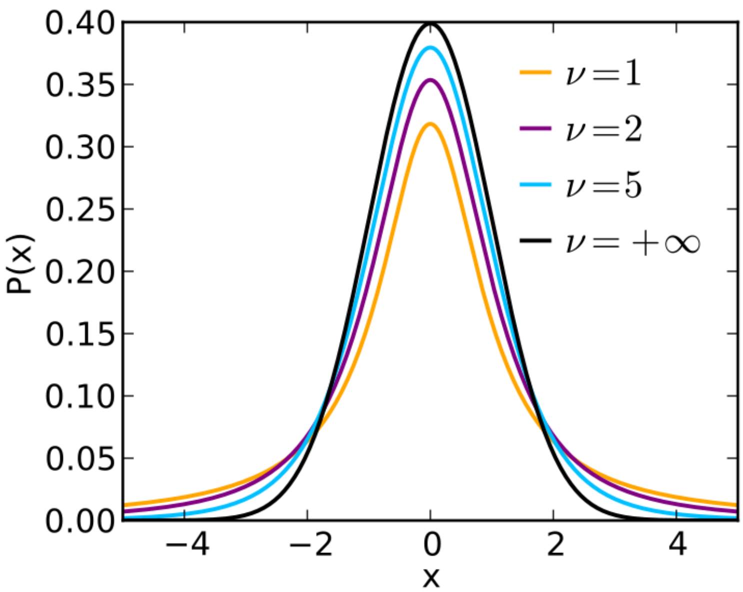


图 2.4 t 分布的概率密度函数

2.8 F 分布

F 分布由著名统计学家 Fisher 提出.

定义 2.13. F 分布

设随机变量 X, Y 相互独立，且 $X \sim \chi^2(n), Y \sim \chi^2(m)$，则称

$$F = \frac{X/n}{Y/m}$$

服从自由度为 n, m 的F 分布，记为

$$F \sim F(n.m).$$

命题 2.14. F 分布的概率密度函数

自由度为 n, m 的 F 分布的概率密度函数为

$$f(x; n, m) = \frac{1}{\mathrm{B}\left(\frac{n}{2}, \frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m} x\right)^{-\frac{n+m}{2}},$$

其中 $B(n, m)$ 为 Beta 函数.

证明 它是两个独立分 χ^2 分布的简单函数的商. □

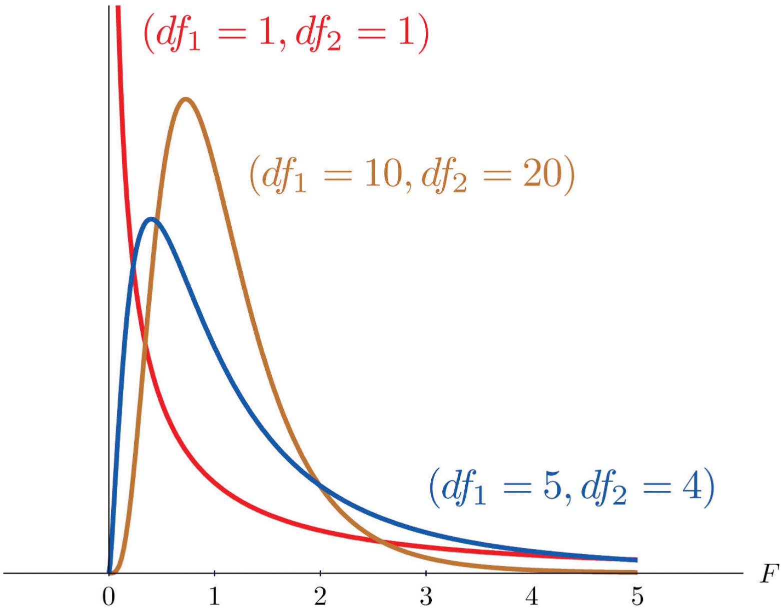


图 2.5 F 分布的概率密度函数

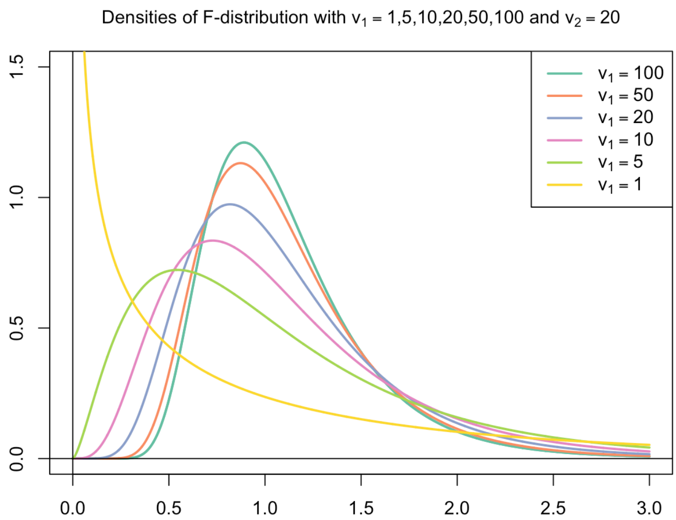


图 2.6 F 分布的概率密度函数

对任意 $\alpha \in (0,1)$，记 $F_\alpha(n,m)$ 为自由度为 n,m 的 F 分布的上 α 分位点，即若 $F \sim F(n,m)$，则

$$\mathbb{P}(F > F_\alpha(n,m)) = \alpha.$$

命题 2.15

1. 如果 $F \sim F(n,m)$，则 $\frac{1}{F} \sim F(m,n)$.
2. 对任意 $\alpha \in (0,1)$，有

$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}.$$

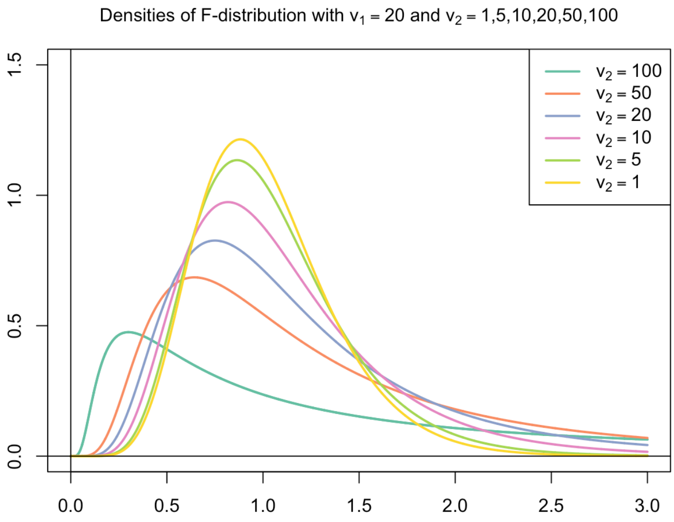


图 2.7 F 分布的概率密度函数

证明 （1）设随机变量 X,Y 相互独立，且 $X\sim\chi^2(n), Y\sim\chi^2(m)$，则

$$\frac{X/n}{Y/m}\sim F(n,m)\sim F.$$

对任意 $x>0$，

$$\begin{aligned}\mathbb{P}(\frac{1}{F}\leq x)&=\mathbb{P}(F\geq\frac{1}{x})\\&=\mathbb{P}\left(\frac{X/n}{Y/m}\geq\frac{1}{x}\right)\\&=\mathbb{P}\left(\frac{Y/m}{X/n}\leq x\right).\end{aligned}$$

因为

$$\frac{Y/m}{X/n}\sim F(m,n),$$

可知 $\frac{1}{F}$ 的分布与 $F(m,n)$ 相同.

（2）设 $F\sim F(n,m)$，则

$$\begin{aligned}1-\alpha&=1-\mathbb{P}(F>F_\alpha(n,m))\\&=1-\mathbb{P}\left(\frac{1}{F}<\frac{1}{F_\alpha(n,m)}\right)\\&=\mathbb{P}\left(\frac{1}{F}>\frac{1}{F_\alpha(n,m)}\right).\end{aligned}$$

由（1）中已证的结论，$1/F\sim F(m,n)$，故

$$1-\alpha=\mathbb{P}\left(\frac{1}{F}>F_{1-\alpha}(m,n)\right)$$

比较上面两个结果可得

$$F_{1-\alpha}(m,n)=\frac{1}{F_\alpha(n,m)}.$$

□

根据定理2.1，

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$$

服从标准正态分布. 下面的定理告诉我们，如果用样本标准差 S 替换上式中的 σ，则所得的随机变量服从自由度为 $n-1$ 的 t 分布.

定理 2.2

设 $X_1, X_2, \ldots, X_n$ 为 $N(\mu, \sigma^2)$ 的一个样本，则

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t(n-1).$$

证明 根据定理2.1，

$$Z := \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1),$$

$$U := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

且因 $\bar{X}$ 与 S^2 独立可知 Z 与 U 独立，所以

$$\frac{Z}{\sqrt{\frac{U}{n-1}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1).$$

□

设总体服从正态分布. 根据推论?? 和推论??，样本方差 S^2 的期望为 σ^2，且当样本量足够大时，S^2 的方差接近于 0. 所以，当样本量足够大时，样本方差近似于总体方差，因此，可以理解，当 n 充分大时，

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$$

与

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S}$$

很接近，从而它们各自的分布——标准正态分布与自由度为 $n-1$ 的 t 分布——很接近. 这就可以解释之前提到的，当自由度 $n \to +\infty$ 时, 自由度为 n 的 t 的概率密度函数的极限为标准正态分布的密度函数.

下面的结论在进行两总体的均值差的区间估计、假设检验以及两总体方差的比区间估计、假设检验时要用到.

定理 2.3

设 $X_1, X_2, \ldots, X_{n_1}$ 为来自总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本，$Y_1, Y_2, \ldots, Y_{n_2}$ 为来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本，且 $X_1, X_2, \ldots, X_{n_1}$ 和 $Y_1, Y_2, \ldots, Y_{n_2}$ 相互独

立，样本均值分别为

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_j,$$

样本方差分别为

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2, \quad S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2.$$

则

1.
$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

2.
$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

3. 如果 $\sigma_1 = \sigma_2 = \sigma$，则
$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1),$$
$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

   其中
$$S_W^2 := \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

   被称为联合样本方差（pooled sample variance）.

证明

(1)
$$\overline{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \quad \overline{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2}).$$

因为 $\overline{X}, \overline{Y}$ 相互独立，所以它们的线性组合 $\overline{X}-\overline{Y}$ 也服从正态分布，注意到

$$\mathbb{E}(\overline{X}-\overline{Y}) = \mu_1-\mu_2,$$
$$\operatorname{Var}(\overline{X}-\overline{Y}) = \operatorname{Var}\overline{X} + \operatorname{Var}\overline{Y} = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2},$$

有

$$\overline{X}-\overline{Y} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right).$$

因而可得 (1).

(2) 注意到

$$U := \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad V := \frac{(n_2-1)S_1^2}{\sigma_2^2} \sim \chi^2(n_2-1),$$

且因为 X_i, Y_j 独立，可知 U, V 独立，可得

$$\frac{\frac{U}{n_1-1}}{\frac{V}{n_2-1}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

(3) 因为利用 (1) 中的结论可知

$$Z := \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1).$$

由 χ^2 分布的可加性，

$$W := \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2).$$

根据定理2.1，$\bar{X}$ 与 S_1^2 独立，又由 X_i, Y_j 独立，所以 $\bar{X}$ 与 S_2^2 也独立，从而 $\bar{X}$ 与 W 独立，同理可证 $\bar{Y}$ 与 W 独立，从而 Z 与 W 独立. 根据 t 分布的定义，可知

$$\frac{Z}{\sqrt{\frac{W}{n_1+n_2-2}}} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2).$$

□

2.9 随机变量列的收敛

定义 2.14. 几乎处处收敛

设 $(X_n)n \geq 1$，X 为概率空间 $(\Omega, \mathscr{A}, \mathbb{P})$ 上的随机变量序列，如果

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = \mathbb{1},$$

则称 X_n 依概率 1 收敛到 X(convergence with probability 1) 或几乎处处收敛到 X，或几乎必然收敛到 X(convergence almost surely，简记为 a.s.)，记为

$$X_n \xrightarrow{\text{a.s.}} X \quad (n \longrightarrow \infty).$$

定义 2.15. 依概率收敛

如果对任意 $\varepsilon > 0$，

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

即 X_n 依概率测度 $\mathbb{P}$ 收敛到 X，则称 X_n 依概率收敛到 X(convergence in probability).

定义 2.16. p 次平均收敛

设 $p \geq 1$，如果对任意 $n \geq 1$，$\mathbb{E}|X_n|^p < \infty$ 且

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0,$$

则称 X_n 依p 次平均收敛到 X(Convergence in p-th mean)，记为

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

特别，若 $X_n \xrightarrow{1} X$，则称 X_n 平均收敛到 X(Convergence in mean),
若 $X_n \xrightarrow{2} X$，则称 X_n 依均方收敛到 X(Convergence in mean square), 也记为

$$X_n \xrightarrow{\text{m.s.}} X.$$

在概率论、统计学中，研究随机变量列的收敛非常普遍，称为极限理论. 很多与 n 有关的随机变量（统计量），例如，样本 $Y_1, ...Y_n$ 的均值 $\bar{Y}_n$，样本方差

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n),$$

都与 n 有关，自然要考虑它们的极限分布. 考虑 X_n 的极限分布可以应用于近似计算可能异常复杂的 X_n 的分布等问题.

命题 2.16

设 $P > 0, X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X.$

证明 由 Markov 不等式，对任意 $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}|X_n - X|^P}{\varepsilon^p}.$$

所以，若 $X_n \xrightarrow{L^p} X$, 则 $X_n \xrightarrow{\mathbb{P}} X$. □

定义 2.17. 依分布收敛

设 $\{X_n\}_{n=1}^{\infty}$，X 为随机变量，X_n 的分布函数为 F_n，X 的分布函数为 F. 如果对 F 的任一连续点 x，都有

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{2.18}$$

则称 X_n 依分布收敛到X(convergence in law 或 convergence in distribution)，记为

$$X_n \xrightarrow{L} X \quad (n \longrightarrow \infty),$$

或

$$X_n \xrightarrow{D} X \ (n \longrightarrow \infty).$$

我们称 F 为 X_n 的极限分布 (limiting distribution), 有时也称之为弱收敛（后文详述）.

注意定义中不要求对任何 x，$F_n(x)$ 收敛到 $F(x)$，事实上也做不到.

例 2.3 设 $X_n = \frac{1}{n}$，$X = 0$，则 X_n 依分布收敛到 X.

证明 X_n 的分布函数为

$$F_{X_n}(x) = \begin{cases} 0, & x \le \dfrac{1}{n}, \\ 1, & x > \dfrac{1}{n}, \end{cases}$$

而

$$F_X(x) = \begin{cases} 0, & x \le 0, \\ 1, & x > 0, \end{cases}$$

注意到 F_X 的唯一不连续点为 $x = 0$. 显然，对任何 $x < 0$，$F_{X_n}(x) = 0 = F_X(x)$，而对任何 $x > 0$，只要 n 充分大，总有 $x > \frac{1}{n}$，从而 $F_{X_n}(x) = 1$ 收敛到 $F_X(x) =$.

□

例 2.4 设 $X_1, X_2, ...X_n$ 是来自 $[0, \theta]$ 上均匀分布的一个随机样本，即 $\{X_i\}_{i=1}^n$ 独立同分布于服从 $[0, \theta](\theta > 0)$ 上均匀分布的随机变量. 记 $X_{(n)}$ 为 $X_1, X_2, ...X_n$ 的 n 阶次序统计量，即

$$X_{(n)} = \max(X_1, X_2, ...X_n),$$

则 $X_{(n)}$ 的分布函数 $F_{X(n)}$ 可如下计算：

$$\begin{aligned} F_{X_{(n)}}(X) &= \mathbb{P}(F_{X_{(n)}} \le x) = \mathbb{P}(X_1 \le x, X_2 \le x, ..., X_n \le x) \\ &= \prod_{i=1}^n \mathbb{P}(X_i \le x) = \begin{cases} 0, & y < 0, \\ \left(\dfrac{x}{\theta}\right)^n, & 0 \le y < \theta, \\ 1, & y \ge \theta. \end{cases} \end{aligned}$$

显然

$$\lim_{n\to\infty} F_{X_{(n)}}(x) = \begin{cases} 0, & x < \theta, \\ 1, & x \ge \theta. \end{cases}$$

因为随机变量 $X = \theta$ 的分布函数为

$$F_X(x) = \begin{cases} 0, & x < \theta, \\ 1, & x \ge \theta. \end{cases}$$

可知 $X_{(n)}$ 依分布收敛到 $X = \theta$. 这也合乎常理，$X_1, X_2, ..., X_n$ 中的最大值应该以某种方式“收敛”到最大可能值 θ.

例 2.5 设 $X_1, X_2, ...X_n$ 为来自标准正态分布的随机样本，$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 我们知道

$$\bar{X}_n \sim N\left(0, \frac{1}{n}\right).$$

即以 0 为均值，方差为 $\frac{1}{n}$ 的正态分布. 当 $n \longrightarrow \infty$ 时，$\bar{X}_n$ 的方差越来越小，趋于 0，也就是说 $\bar{X}_n$ 越来越集中于其均值 0 附近. 直观上，$\bar{X}_n$ 应该以某种方式收敛于 0. 事实上，可以证明

$$\bar{X}_n \xrightarrow{L} 0.$$

证明如下：

因为 $\bar{X}_n \sim N(0, \frac{1}{n})$, 它的分布函数为

$$F_{\bar{X}_n}(x) = \int_{-\infty}^{x} \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{nt^2}{2}\right) dt.$$

作变量代换 $s = \sqrt{n}t$ 可得

$$F_{\bar{X}_n}(x) = \int_{-\infty}^{\sqrt{n}x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

易知 $F_{\bar{X}_n}(x) = \Phi(\sqrt{n}x)$, 其中 Φ 为标准正态分布函数，所以

$$\lim_{n\to\infty} F_{\bar{X}_n}(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x = 0, \\ 1, & x > 0. \end{cases}$$

注意到 $X = 0$ 的分布函数

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

显然，对任何 $x \neq 0$, 即对任意 F_X 的连续点 x，都有

$$\lim_{n\to\infty} F_{\bar{X}_n}(x) = F_X(x).$$

注意当 $n \longrightarrow \infty$ 时，$F_{X_n}(0)$ 的极限为 $\frac{1}{2}$，不等于 $F(0) = 1$，$x = 0$ 不是 $F(x)$ 的连续点. 所以当 $n \longrightarrow \infty$ 时，F_{X_n} 的极限不是右连续的，而分布函数必须是右连续的，所以 F_{X_n} 的逐点极限不是一个分布函数.

由此可见，在依分布收敛定义中，我们只要求对 F 的连续点 x 有(2.18)成立.

命题 2.17. 依概率收敛强于依分布收敛

设 X_n 依概率收敛到 X，则 X_n 依分布收敛到 X，即

$$X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{L} X.$$

证明 分别记 X_n, X 的分布函数为 F_n, F, 即

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

对任意 $\varepsilon > 0$, 由于 $X_n \xrightarrow{\mathbb{P}} X$ 可知

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

因为

$$\begin{aligned} F_n(x) = \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \\ &\leq \mathbb{P}(X \leq x+\varepsilon) + \mathbb{P}(X - X_n > \varepsilon) \\ &= F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

所以

$$F_n(x) \leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

在上式中，令 $n \longrightarrow \infty$ 可得

$$\varlimsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon). \tag{2.19}$$

同理可知

$$\begin{aligned} F(x-\varepsilon) = \mathbb{P}(X \leq x-\varepsilon) &= \mathbb{P}(X \leq x-\varepsilon, X_n \leq x) + \mathbb{P}(X \leq x-\varepsilon, X_n > x) \\ &\leq \mathbb{P}(X \leq x) + \mathbb{P}(X_n - X > \varepsilon) \\ &\leq F_n(x) + \mathbb{P}(|X - X_n| > \varepsilon). \end{aligned}$$

所以

$$F(x-\varepsilon) \leq F_n(x) + \mathbb{P}(|X - X_n| > \varepsilon).$$

在上式中，令 $n \longrightarrow \infty$ 可得

$$F(x-\varepsilon) \leq \varliminf_{n\to\infty} F_n(x). \tag{2.20}$$

综合(2.19)和(2.20)可得

$$F(x-\varepsilon) \leq \varliminf_{n\to\infty} F_n(x) \leq \varlimsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

注意到 F 是增函数，令 $\varepsilon \downarrow 0$ 有

$$F(x-0) = \lim_{\varepsilon\downarrow 0} F(x-\varepsilon) \leq \varliminf_{n\to\infty} F_n(x) \leq \varlimsup_{n\to\infty} F_n(x) \leq \lim_{\varepsilon\downarrow 0} F(x+\varepsilon) = F(x+0).$$

如果 x 是 F 的连续点，则

$$F(x-0) = F(x+0) = F(x).$$

因而由上式可得

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

这就证明了 X_n 依分布收敛于 X. □

依分布收敛是一种很弱的收敛方式. 例如，考虑随机变量 $X \sim B(1, \frac{1}{2})$, 即

$$\mathbb{P}(X=0) = \frac{1}{2},\ \ \mathbb{P}(X=1) = \frac{1}{2}.$$

设 $X_n = X$, 则 $X_n \xrightarrow{L} X$. 同时还有 $X_n \xrightarrow{L} 1-X$. 然而，X_n 不是几乎处处收敛于 $1-X$，不依 L^p 收敛到 $1-X$，也不依概率收敛到 $1-X$.

一般地，我们有如下关系

$$\begin{array}{l} \boxed{X_n \xrightarrow{\text{a.s.}} X} \searrow \\ \qquad\qquad\qquad\quad \boxed{X_n \xrightarrow{\mathbb{P}} X} \longrightarrow \boxed{X_n \xrightarrow{L} X}. \\ \boxed{X_n \xrightarrow{L^p} X} \nearrow \end{array}$$

当 X 为常数 c 时，最后的关系可逆，即有以下命题

命题 2.18

设 c 为常数，则 $X_n \xrightarrow{L} c \Longleftrightarrow X_n \xrightarrow{\mathbb{P}} c$.

证明 由命题**??**，我们只要证明必要性. 设 X_n 的分布函数为 F_n, 随机变量 $X=c$ 的分布函数为

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

对任意 $\varepsilon > 0, c+\varepsilon, c-\varepsilon$ 都是 F 的连续点，所以，当 $n \longrightarrow \infty$ 时，

$$\begin{aligned} \mathbb{P}(|X_n - c| \geq \varepsilon) &= \mathbb{P}(X_n \geq c+\varepsilon) + \mathbb{P}(X_n \leq c-\varepsilon) \\ &= 1 - F_n(c+\varepsilon) + F_n(c-\varepsilon) \\ &\longrightarrow 1 - F(c+\varepsilon) + F(c-\varepsilon) \\ &= 1 - 1 + 0 = 0. \end{aligned}$$

这就证明了 $X_n \xrightarrow{\mathbb{P}} c$. □

我们现在知道，依分布收敛是几种收敛方式中最弱的一种收敛方式. 而且，当我们说

$$X_n \xrightarrow{L} X$$

时，我们不要求 X_n, X 定义于同一概率空间. 但借助于几乎处处收敛我们有如下有用的 Skorokhod 表示定理.

定理 2.4. Skorokhod 表示定理

设 $\{X_n\}, X$ 的分布函数分别为 $\{F_n\}, F$，且

$$X_n \xrightarrow{L} X, \ (n \longrightarrow \infty).$$

则存在概率空间 $(\Omega, \mathscr{A}, \mathbb{P})$ 及其上随机变量 $\{Y_n\}, Y$, 使得

(a) $\{Y_n\}, Y$ 的分布函数分别为 $\{F_n\}, F$.

(b) $Y_n \xrightarrow{\text{a.s.}} Y, \ (n \longrightarrow \infty)$.

Skorokhod 表示定理有如下两个初等应用.

定理 2.5. Skorokhod 表示定理的应用

设 $X_n \xrightarrow{L} X, g : \mathbb{R} \longrightarrow \mathbb{R}$ 为连续函数，则 $g(X_n) \xrightarrow{L} g(X)$.

证明 设 $\{Y_n\}, Y$ 如 Skorokhod 定理中所给. 则由 g 的连续性可知

$$\{Y_n \longrightarrow Y\} \subset \{g(Y_n) \longrightarrow g(Y)\}.$$

因此，

$$g(Y_n) \xrightarrow{\text{a.s.}} g(Y), \ (n \longrightarrow \infty).$$

所以，

$$g(Y_n) \xrightarrow{L} g(Y).$$

然而，$g(Y_n), g(Y)$ 分别与 $g(X_n), g(X)$ 有相同的分布. 这就证明了 $g(X_n) \xrightarrow{L} g(X)$.

□

定义 2.18

如果对任意有界连续函数 g 都有 $\mathbb{E}g(X_n) \longrightarrow \mathbb{E}g(X)\ (n \longrightarrow \infty)$，则称 X_n 弱收敛于 X(weak convergence), 记为 $X_n \xrightarrow{\omega} X$.

定理 2.6

$X_n \xrightarrow{L} X$ 当且仅当 $X_n \xrightarrow{\omega} X$.

内容提要

- 总体
- 样本
- 统计量
- 抽样分布
- χ^2 分布
- t 分布
- F 分布

第 2 章 习题

1. 设总体 $X \sim U(-\frac{1}{2}, \frac{1}{2})$, $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本, $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为次序统计量. 分别求 $X_{(1)}, X_{(n)}$, 以及 $(X_{(1)}, X_{(n)})$ 的分布.
2. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本，$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为次序统计量.
   (a). 如果 $X \sim U(a, b)$，求证
   $$\mathbb{E}X_{(r)} = a + \frac{r}{n+1}(b-a).$$
   (b). 如果 X 服从参数为 $\lambda > 0$ 的指数分布，求 $\mathbb{E}X_{(r)}$.
3. 证明 $n \to \infty$ 时，服从自由度为 n 的 t 分布的随机变量 T_n 依分布收敛于标准正态分布随机变量，且 T_n 的概率密度函数逐点收敛于标准正态随机变量的概率密度函数.
4. 已知随机变量 $X \sim t(n)$，求证：$X^2 \sim F(1, n)$.
5. 求总体 $N(20, 3)$ 的容量分别为 $10, 15$ 的两个独立样本均值的差的绝对值大于 0.3 的概率.
6. 设 $X_1, \cdots, X_n, X_{n+1}, \cdots, X_{n+m}$ 是分布 $N(0, \sigma^2)$ 的容量为 $n+m$ 的样本，试求下列统计量的概率分布
   (a).
   $$Y_1 = \frac{\sqrt{m}\sum_{i=1}^{n} X_i}{\sqrt{n}\sqrt{\sum_{i=n+1}^{n+m} X_i^2}}.$$
   (b).
   $$Y_2 = \frac{m\sum_{i=1}^{n} X_i^2}{n\sum_{i=n+1}^{n+m} X_i^2}$$
7. 设 $X_1, \cdots, X_n, X_{n+1}$ 是来自总体 $N(\mu, \sigma^2)$ 的样本，
   $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$
   试求统计量
   $$T = \frac{X_{n+1} - \bar{X}}{S^*} \cdot \sqrt{\frac{n-1}{n+1}}$$
   的分布.
8. 设总体 $X \sim N(\mu, \sigma^2)$，$X_1, X_2, \cdots, X_n$ 为来自 X 的样本.
   (a). 写出 $X_1, X_2, \cdots, X_n$ 的联合概率密度.
   (b). 写出 $\bar{X}$ 的概率密度.
9. 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的简单随机样本，已知 $\mathbb{E}X^k = a_k, k = 1, 2, 3, 4$，证明：当 n 充分大时，$Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2$ 近似服从正态分布，并指出其分布参数.
10. 设 $\overline{X}$ 为来自容量为 25 的标准正态分布的样本的样本均值. 求 $|\overline{X}|$ 的上 90%,95%,99% 分位数，即分别求 $c_{0.9}, c_{0.95}, c_{0.99}$ 使得
    $$\mathbb{P}(|\overline{X}| > c_{0.9}) = 0.1, \quad \mathbb{P}(|\overline{X}| > c_{0.95}) = 0.05, \quad \mathbb{P}(|\overline{X}| > c_{0.99}) = 0.01..$$

11. 从正态总体 $N(3.4, 6^2)$ 中抽取容量为 n 的样本，如果需要样本均值位于区间 $(1.4, 5.4)$ 的概率不小于 0.95，问样本容量 n 至少应多大？
12. 设 $X_1, X_2, \cdots, X_n$ 为独立同分布随机变量序列，期望为 μ，方差为 σ^2. 记

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

(a). 设 a 为常数，证明对任意形如 aS^2 的 σ^2 的估计，有

$$\mathrm{MSE}(aS^2) = \mathbb{E}(aS^2 - \sigma^2)^2 = a^2 \operatorname{Var} S^2 + (a-1)^2\sigma^4.$$

(b). 证明

$$\operatorname{Var}(S^2) = \frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\right)\sigma^4,$$

其中 $\kappa = \frac{\mathbb{E}(X-\mu)^4}{\sigma^4}$ 为 X 的峰度 (kurtosis).

(c). 设 X_i 服从正态分布. 证明

(i) $\mathrm{MSE}(S^{*2}) < \mathrm{MSE}(S^2)$.

(ii) $\kappa = 3$.

(iii) 形如 aS^2 的估计中，MSE 最小的是 $\frac{n-1}{n+1}S^2$.

(d). 不做正态性假设，证明当

$$a = \frac{n-1}{(n+1) + \frac{(\kappa-3)(n-1)}{n}}$$

时 $\mathrm{MSE}(S^2)$ 取得最小值.

13. 设 n 维随机向量 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，求证 $\left\langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right\rangle$ 服从自由度为 n 的 χ^2 分布.
14. 设随机变量 U, V 相互独立，且 $U \sim \chi^2(n), V \sim \chi^2(m)$，其中 $n, m \geq 1$，求证 $U + V \sim \chi^2(n+m)$.
15. 设 $X_1, X_2, \cdots, X_n$ 相互独立且都服从参数为 $\lambda > 0$ 的指数分布（分布函数为 $F(x) = 1 - \mathrm{e}^{-\lambda x}, x > 0$），求证 $2\lambda(X_1 + X_2 + \cdots + X_n) \sim \chi^2(2n)$.
16. 设总体 $X \sim N(\mu, \sigma^2)$，$X_1, X_2, \cdots, X_n$ 为来自 X 的样本.

(a). 写出 $X_1, X_2, \cdots, X_n$ 的联合概率密度函数.

(b). 写出 $\bar{X}$ 的概率密度函数.

(c). 请写出 $\bar{X}$ 与 X_1 的协方差.

第 3 章 点估计

在很多统计问题中，总体的分布类型往往知道，具体分布往往由一个或多个参数决定. 如设总体的概率函数 $f(x;\theta)$ 已知，仅其中参数 $\theta = (\theta_1, \theta_2, \cdots, \theta_m)$ 中的一个或多个 θ_i 未知. 因此我们只要确定分布的参数. 参数估计有两种形式，一是点估计，二是区间估计，它们分别确定参数的大小，参数所在的范围. 本章先介绍点估计的方法.

定义 3.1. 点估计

设总体分布依赖于未知参数 $\theta = (\theta_1, \theta_2, \cdots, \theta_m)$，$\theta$ 取值于参数空间 Θ. 设 $X_1, X_2, \ldots, X_n$ 是来自总体的一个样本，用于估计未知参数 θ（或更一般地，未知参数 θ 的函数 $g(\theta)$）的统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$（$\hat{g} = \hat{g}(X_1, X_2, \cdots, X_n)$）称为 θ（$g(\theta)$）的点估计，简称估计. 设 $x_1, x_2, \cdots, x_n$ 为样本观测值，则 $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$（$\hat{g} = \hat{g}(x_1, x_2, \cdots, x_n)$）称为 θ（$g(\theta)$）的估计值.

我们将介绍三种进行点估计的方法：

1. 矩估计法；
2. 极大似然估计法；
3. 贝叶斯估计.

3.1 矩估计

设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本.

回忆我们曾介绍过总体 X 的 $k \geq 1$ 阶（原点）矩

$$\mu_k = \mathbb{E}X^k$$

样本 k 阶（原点）矩为

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

一般来说，总体的各阶矩是未知参数 θ 的函数. 如

$$\begin{cases} \mu_1 = \mu_1(\theta_1, \theta_2, \cdots, \theta_m), \\ \mu_2 = \mu_2(\theta_1, \theta_2, \cdots, \theta_m), \\ \cdots\cdots \\ \mu_m = \mu_m(\theta_1, \theta_2, \cdots, \theta_m). \end{cases}$$

从而，由上述方程（组）可以知各参数可以用各阶矩来表示：

$$\begin{cases} \theta_1 = \theta_1(\mu_1, \mu_2, \cdots, \mu_m), \\ \theta_2 = \theta_2(\mu_1, \mu_2, \cdots, \mu_m), \\ \cdots\cdots \\ \theta_m = \theta_m(\mu_1, \mu_2, \cdots, \mu_m). \end{cases}$$

如果 μ_k 有限，根据大数定律，A_k 依概率收敛于 μ_k：

$$A_k \xrightarrow{\mathbb{P}} \mu_k, \quad n \to \infty.$$

用各阶样本矩 $A_1, A_2, \cdots, A_m$ 分别替换 $\theta_k(\mu_1, \mu_2, \cdots, \mu_m)$ 中各总体矩 $\mu_1, \mu_2, \cdots, \mu_m$ 得到 θ_k 的估计：

$$\hat{\theta}_n := \theta(A_1, A_2, \cdots, A_m).$$

这样的估计是合理的.

如果参数 $\theta(\mu_1, \mu_2, \cdots, \mu_m)$ 是各阶总体矩的连续函数，则当 $n \to \infty$ 时，因 A_k 依概率收敛于 μ_k，可得 $\hat{\theta}_n = \theta(A_1, A_2, \cdots, A_m)$ 依概率收敛于 $\theta = \theta(\mu_1, \mu_2, \cdots, \mu_m)$.

因此，当 n 充分大时，$\hat{\theta}_k := \theta(A_1, A_2, \cdots, A_m)$ 近似于 $\theta_k = \theta_k(\mu_1, \mu_2, \cdots, \mu_m)$.

所以，用样本矩代替总体矩，可以给出参数的估计. 这样得到的估计就称为*矩估计*. 这个方法源自 1900 年统计学家卡尔・皮尔逊（Karl Pearson）提出的*替换原理*.

例 3.1 设 $X_1, X_2, \ldots, X_n$ 为总体 X 的一个样本，总体 X 的概率密度函数为

$$f(x;\theta) = \begin{cases} \dfrac{1}{\theta}, & 0 \leq x \leq \theta, \\ 0, & \text{其他}, \end{cases}$$

其中 $\theta > 0$ 为未知参数. 求 θ 的矩估计量.

解 总体 X 的一阶原点矩为

$$\mu_1 = \mathbb{E}X = \frac{\theta}{2}.$$

所以，参数 θ 可以表示为一阶矩 μ_1 的函数：

$$\theta = 2\mu_1.$$

因此，用样本一阶矩（即样本均值）$\overline{X}$ 代替总体一阶矩 μ_1 可得 θ 的矩估计量为

$$\hat{\theta} = 2\overline{X}.$$

□

例 3.2 设总体 $X \sim B(1,p)$，$0 < p < 1$，$X_1, X_2, \dots, X_n$ 是一个样本，求 θ 的矩估计量.

解 X 的一阶矩为

$$\mu_1 = \mathbb{E}X = p.$$

因此，$p = \mu_1$. 用样本均值替换一阶矩可得 p 的矩估计量为

$$\hat{p} = \overline{X}.$$

□

上面的例题相当于用频率估计概率.

例 3.3 设总体 X 的期望为 μ，方差为 σ^2，$X_1, X_2, \dots, X_n$ 是一个样本，求 μ, σ^2 的矩估计量.

解 注意到

$$\mu_1 = \mathbb{E}X = \mu, \quad \mu_2 = \mathbb{E}X^2 = \mu_1^2 + \sigma^2,$$

可得

$$\mu = \mu_1, \quad \sigma^2 = \mu_2 - \mu_1^2.$$

所以，分别用样本一阶矩 $\overline{X}$ 和二阶矩 A_2 替换总体一阶矩 μ_1 和二阶矩 μ_2 可得 μ, σ^2 的矩估计量为

$$\begin{aligned}
\hat{\mu} &= A_1 = \overline{X}, \\
\hat{\sigma}^2 &= A_2 - A_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2.
\end{aligned}$$

实际上，注意到 σ^2 是总体的二阶中心矩，我们也可以直接用二阶样本中心矩 B_2 作为 σ^2 的矩估计量. 上面仅用原点矩的计算也表明可以用二阶样本中心距估计二阶总体中心矩.

□

例 3.4 设 X 的概率质量函数如下

X	0	1	2	3
$\mathbb{P}$	θ^2	$2\theta(1-\theta)$	θ^2	$1-2\theta$

其中 $0 < \theta < \frac{1}{2}$ 为未知参数，样本值为 3,1,3,0,3,1,2,3. 求 θ 的矩估计值.

解

$$\mu_1 = \mathbb{E}X = 0 \cdot \theta^2 + 1 \cdot 2\theta(1-\theta) + 2 \cdot \theta^2 + 3 \cdot (1-2\theta) = 3 - 4\theta.$$

所以，参数 θ 可以用一阶总体矩表示为

$$\theta = \frac{3-\mu_1}{4}.$$

用样本一阶矩 $\overline{x}$ 替换总体一阶矩 μ_1 可得参数 θ 的矩估计值为

$$\hat{\theta} = \frac{3-\overline{x}}{4}.$$

代入样本均值

$$\overline{x} = \frac{3+1+3+0+3+1+2+3}{8} = 2,$$

可得 θ 的矩估计值为

$$\hat{\theta} = \frac{3-2}{4} = \frac{1}{4}.$$

□

上面这个例子中，样本值中有 4 个 3，2 个 1，1 个 0 和 1 个 2. 从观测值可以看出 3 出现的频率大些. 从估计结果，也可以看出来. 根据估计值，我们可以得到 X 的概率质量函数如下：

X	0	1	2	3
$\mathbb{P}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$	$\frac{1}{4}$

例 3.5 设总体 X 服从 $[a,b]$ 区间上的均匀分布，$X_1, X_2, \cdots, X_n$ 是来自 X 的样本，求 a,b 的矩估计量.

解 易知

$$\mu_1 = \mathbb{E}X = \frac{a+b}{2},$$

$$\mu_2 = \mathbb{E}X^2 = \operatorname{Var}X + (\mathbb{E}X)^2 = \frac{(b-a)^2}{12} + \mu_1^2.$$

由此可得

$$\begin{aligned} a+b &= 2\mu_1, \\ b-a &= \sqrt{12(\mu_2-\mu_1^2)}. \end{aligned}$$

解方程可得

$$\begin{aligned} a &= \mu_1 - \sqrt{3(\mu_2-\mu_1^2)}, \\ b &= \mu_1 + \sqrt{3(\mu_2-\mu_1^2)}. \end{aligned}$$

分别用 $A_1 = \overline{X}, A_2$ 代替 μ_1, μ_2，注意到(2.1)

$$A_2 - A_1^2 = \frac{n-1}{n} S^2,$$

可得 a, b 的矩估计量如下

$$\hat{a} = \overline{X} - \sqrt{\frac{3(n-1)}{n}} S,$$
$$\hat{b} = \overline{X} + \sqrt{\frac{3(n-1)}{n}} S.$$

□

例 3.6 设 $X_1, X_2, \cdots, X_n$ 为来自参数为 θ 的 Beta 分布，即 X 有概率密度函数为

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

则

$$\mu_1 = \mathbb{E}X = \int_0^1 x \cdot f(x, \theta) dx = \frac{\theta}{\theta+1}.$$

由此可得

$$\theta = \frac{\mu_1}{\mu_1 - 1}.$$

因此 θ 的矩估计量为

$$\hat{\theta} = \frac{\overline{X}}{\overline{X} - 1}.$$

3.2 极大似然估计的直观思想

极大似然估计（或最大似然估计，Maximal Likelihood Estimation, 简写为 MLE）的思想最早是由高斯提出来的. 高斯在研究误差分布的时候，认为误差分布的极大似然估计导出的应该是算术平均. 高斯由此反过去确定误差分布应该是正态分布.

1922 年，费希尔（Fisher）再次提出极大似然估计的思想，并证明极大似然估计有一些好的性质从而使极大似然估计得到广泛应用.

图 3.1　年轻的费希尔

极大似然估计的思想极其简单. 如新手、熟手完成某件事情成功的概率一低一高. 现在已知任务没成功，则我们可以估计是新手做的，因为新手失败的概率高. 我们下面先以例子用数学语言叙述，进行定量计算.

简化的例子

问题 3.1 设粉笔盒中有粉笔 10 支，其中有彩色粉笔 1 支或 2 支. 从中随机抽取 1 支粉笔. 请根据抽取到的粉笔的颜色，估计盒中彩色粉笔的支数.

直观上，如果抽到的是白色粉笔，则可以认为盒子中彩色粉笔数为 1 支，因为这是抽到白色粉笔的概率较大；如果抽到是彩色粉笔，则可以认为盒子中彩色粉笔数为 2 支，因为这时抽到彩色粉笔的可能性更大.

这就是极大似然估计的朴素思想. 盒子中粉笔的分布最可能的情形是使发生可能性最大的事件发生.

下面将上述讨论改用数学的语言表达.

估计粉笔盒中的粉笔的支数，相当于估计粉笔盒中彩色粉笔的比例 θ，它的取值范围为 $\Theta = \left\{\frac{1}{10}, \frac{2}{10}\right\}$.

用随机变量 X 来表示从粉笔盒中随机抽取一支粉笔所得的结果：用 $X = 1$ 表示取到彩色粉笔，用 $X = 0$ 表示取到白色粉笔，则 X 的分布为

$$\mathbb{P}(X = 1) = \theta$$

$$\mathbb{P}(X = 0) = 1 - \theta.$$

其中 θ 就是需要估计的总体 X 的参数，参数空间为 Θ.

设从盒中随机抽取一支粉笔，所得结果为 X_1，它可以视为 X 的一个样本.

如果盒子中只有一只彩色粉笔 $\theta = \frac{1}{10}$，则抽到彩色粉笔和白色粉笔的概率分别为 $\frac{1}{10}$ 和 $\frac{2}{10}$，即

$$\mathbb{P}\left(X_1 = 1|\theta = \frac{1}{10}\right) = \frac{1}{10}, \quad \mathbb{P}\left(X_1 = 0|\theta = \frac{1}{10}\right) = \frac{9}{10}.$$

类似地，若盒子中有 2 只彩色粉笔，则抽到彩色粉笔和白色粉笔的概率分别为 $\frac{2}{10}$ 和 $\frac{8}{10}$，即

$$\mathbb{P}\left(X_1 = 1|\theta = \frac{2}{10}\right) = \frac{2}{10}, \quad \mathbb{P}\left(X_1 = 0|\theta = \frac{2}{10}\right) = \frac{8}{10}.$$

概率	$X_1 = 1$	$X_1 = 0$
$\theta = \frac{1}{10}$	$\frac{1}{10}$	$\frac{9}{10}$
$\theta = \frac{2}{10}$	$\frac{2}{10}$	$\frac{8}{10}$

如果抽到是彩色粉笔，即 $X_1 = 1$. 参数 θ (真实情形) 应取使得该事件更容易发生的值. 因为当 $\theta = \frac{2}{10}$ 时，$X_1 = 1$ 的概率较大. 所以，θ 的估计为 $\hat{\theta} = \frac{2}{10}$. 类似地，若抽到的是白色粉笔，即 $X_1 = 0$，则取 θ 的估计为 $\hat{\theta} = \frac{1}{10}$.

X_1	$X_1 = 1$	$X_1 = 0$
$\hat{\theta}$	$\frac{2}{10}$	$\frac{1}{10}$

问题 3.2 设粉笔盒中有粉笔 10 支，其中有彩色粉笔 1 支或 2 支. 从中有放回地随机抽取一支粉笔两次. 请根据抽取到的粉笔的颜色，估计其中彩色粉笔的支数.

续前面一个问题的讨论. 用 X_1, X_2 分别表示两次抽取的结果. 用 $T = X_1 + X_2$ 两次抽取中所记录的彩色粉笔数之和. 则可以计算在不同参数情形下，T 取各值的概率如下表：

$\mathbb{P}(T = t)$	$t = 0$	$t = 1$	$t = 2$
$\theta = \frac{1}{10}$	$\frac{81}{100}$	$\frac{18}{100}$	$\frac{1}{100}$
$\theta = \frac{2}{10}$	$\frac{64}{100}$	$\frac{32}{100}$	$\frac{4}{100}$

由上表可见，两次抽取所得结果是一次白色一次彩色的概率在 $\theta = \frac{2}{10}$ 时比在 $\theta = \frac{1}{10}$ 时对应的概率大. 按极大似然估计的思想，若两次抽得的结果不同，可以取 θ 的估计为 $\hat{\theta} = \frac{2}{10}$. 其他情况可类似考虑. 按 T 的取值可得 θ 的估计如下表.

T	0	1	2
$\hat{\theta}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{2}{10}$

下面的问题将上述问题稍作修改.

问题 3.3 设某盒粉笔中彩色粉笔的比例为 θ. 有放回地抽取两次，设先后抽到彩色和白色粉笔，求 θ 的估计.

分别以随机变量 X_1, X_2 记录所得的记录. 分别取到彩色和白色粉笔，即样本值为 $x_1 = 1, x_2 = 0$，概率为

$$\mathbb{P}(X_1 = 1, X_2 = 0) = \theta \cdot (1 - \theta).$$

极大似然估计的思想要求 θ 应使 $\theta(1-\theta)$ 最大. 由于

$$\theta(1-\theta) \leqslant \left(\frac{\theta + (1-\theta)}{2}\right)^2 = \left(\frac{1}{2}\right)^2$$

且 $\theta = 1 - \theta$ 时取到等号，可知当 $\hat{\theta} = \frac{1}{2}$ 时，$\mathbb{P}(X_1 = 1, X_2 = 0)$ 最大.

因此，此时可认为 $\hat{\theta} = \frac{1}{2}$.

抽样不同，得到的估计也不同. 两次抽样所得 p 的极大似然 $\hat{p}$ 如下表：

(X_1, X_2)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$\hat{p}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0

易见

$$\hat{p}(X_1, X_2) = \frac{X_1 + X_2}{2} = \overline{X}$$

下面一般地考虑上述问题，它可以当做极大似然估计方法的一般介绍.

问题 3.4 设总体 X 服从两点分布：$X \sim B(1, p), 0 < p < 1$. 求 p 的极大似然估计.

解 注意到 X 的概率质量函数

X	1	0
$\mathbb{P}$	p	$1-p$

可统一表示为

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

设 $(X_1,\cdots,X_n)$ 为总体的一个随机样本，样本值为 $(x_1,x_2,\cdots,x_n)$，则由独立性与同分布性，得到这样一个样本值的概率为

$$\begin{aligned}
&\mathbb{P}\left((X_1,\cdots,X_n)=(x_1,\ldots x_n)\right)\\
=&\mathbb{P}\left((X_1=x_1,X_2=x_2,\cdots,X_n=x_n)\right)\\
=&\mathbb{P}\left((X_1=x_1)\,\mathbb{P}\left(X_2=x_2\right)\cdots\mathbb{P}\left(X_n=x_n\right)\right)\\
=&\prod_{i=1}^n\mathbb{P}(X_i=x_i)=\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}\\
=&p^{\sum_{i=1}^n x_i}(1-p)^{n-\Sigma x_i}=p^{n\overline{x}}(1-p)^{n(1-\overline{x})}.
\end{aligned}$$

对给定的样本值，这个概率与参数 p 有关，记之为

$$L(p)=p^{n\overline{x}}(1-p)^{n(1-\overline{x})}.$$

我们称它为 p 的似然函数. 我们想求 $L(p)$ 的极大值点. 但我们考虑它的对数

$$\log L(p)=n\overline{x}\log p+n(1-\overline{x})\log(1-p).$$

因为对数函数是单调增函数，$L(p)$ 与 $\log L(p)$ 有相同的极大值点. 对 $\log L(p)$ 关于 p 求导数，可得

$$\frac{d\log p}{dp}=\frac{n\overline{x}}{p}-\frac{n(1-\overline{x})}{1-p}=0.$$

我们称之为似然方程. 从中可解得 p 的似然估计值为

$$\hat{p}=\overline{x}.$$

p 的似然估计量为

$$\hat{p}=\overline{X}.$$

□

连续随机变量情形

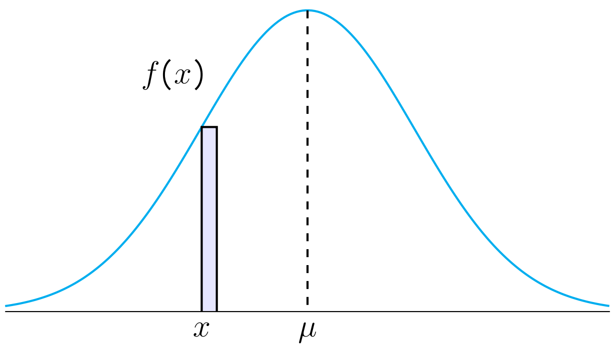
设 X 服从正态分布 $N\left(\mu,\sigma^2\right)$，分布密度函数为：

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

则 X 取值在 x 附近的概率为

$$\mathbb{P}(X \in (x, x+dx]) \approx f(x)dx.$$

直观上，它是概率密度函数下方以 $f(x)$ 为高，dx 为底的小矩形的面积.



$f(x)$ 越大，小矩形越高，X 取值于 x 附近的概率就越大.

方差已知情形

现设方差 σ_0^2 已知，均值 μ 未知：

$$X \sim N\left(\mu, \sigma_0^2\right).$$

进行一次抽样，样本值为 x，μ 的值应该“使” X 取值于 x 附近的概率最大，这个概率由 $f(x)$ 来决定.

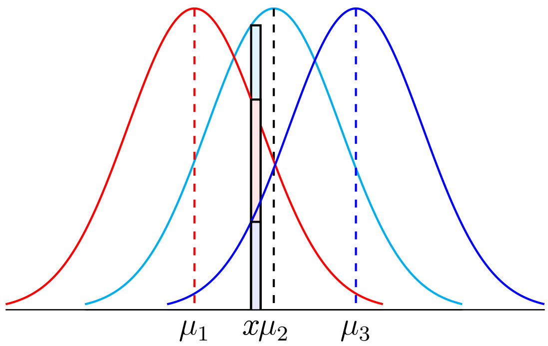


图 3.2　不同的期望条件下，在 x 附近取值的概率

易从图中可见，当 $\mu = x$ 时，可使 $f(x)$ 最大. 从数学上也可见：

$$\begin{aligned} f\left(x, \mu, \sigma_0^2\right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \quad (x = \mu). \end{aligned}$$

例如，设 X 为一个班学生的成绩，随机访问班上一名学生，得知其成绩为 80，有理由估计全班学生的平均成绩为 80.

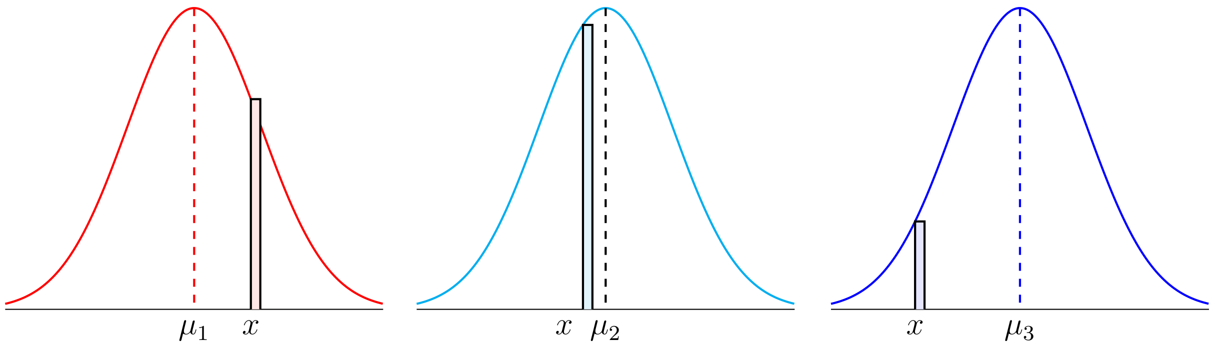


图 3.3　不同的期望条件下，在 x 附近取值的概率（图像分开考察）

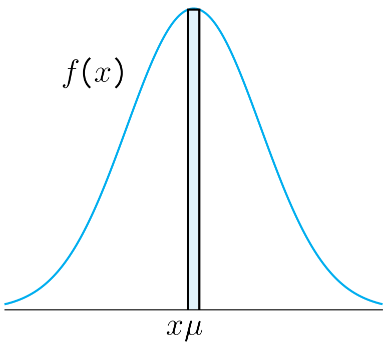


图 3.4　不同的期望条件下，在 x 附近取值的概率（图像分开考察）

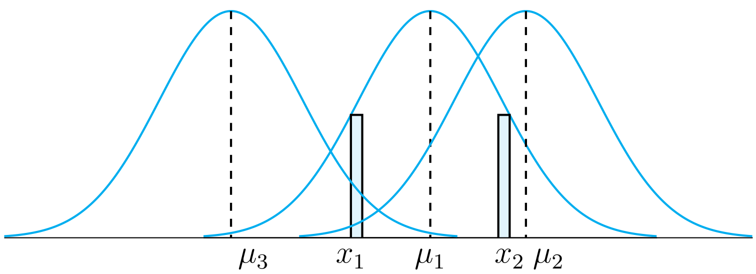
设抽样两次，样本值为 x_1, x_2. 则

$$
\begin{aligned}
&\mathbb{P}\left(X_1 \text{ 取值于 } x_1 \text{ 附近，} X_2 \text{ 取值于 } x_2\right) \\
=&\ \mathbb{P}\left(X_1 \text{ 取值于 } x_1 \text{ 附近}\right) \cdot \mathbb{P}\left(X_2 \text{ 取值于 } x_2 \text{ 附近}\right) \\
=&\ f\left(x_1\right) dx_1 \cdot f\left(x_2\right) dx_2 \\
=&\ f\left(x_1\right) f\left(x_2\right) dx_1 dx_2.
\end{aligned}
$$

μ 的值应该使 x_1, x_2 被取到的可能性比较大，即 $f(x_1)f(x_2)$ 最大.

可以证明，当 μ 取 x_1, x_2 的平均值时，$f(x_1)f(x_2)$ 最大.

比如，从班级中随机取两名学生的成绩，分别为 70，90 则可估计平均成绩越 80.



$$
\hat{\mu} = \frac{x_1 + x_2}{2}
$$

问题 3.5 设 μ_0 已知，方差 σ^2 未知. 随机抽样，样本量为 1.

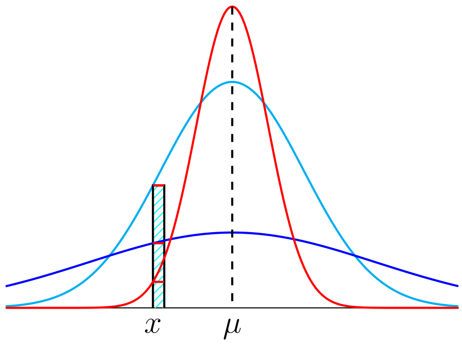


图 3.5　不同方差条件下，取值于 x 附近的概率

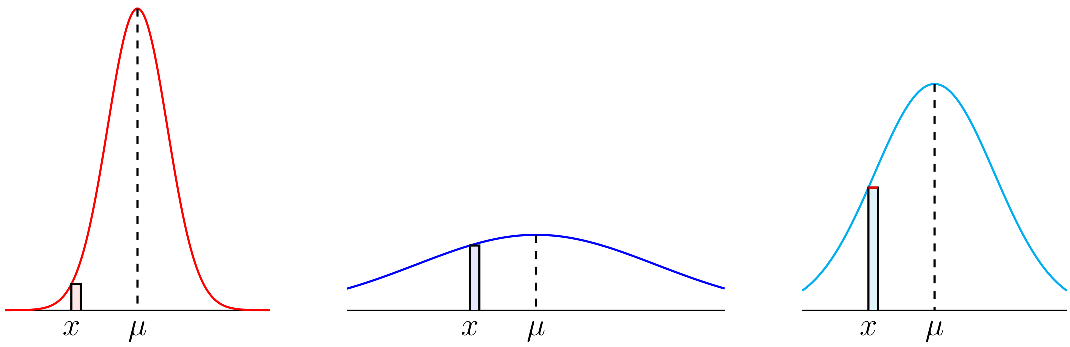


图 3.6　不同方差条件下，取值于 x 附近的概率（分别作图）

$$\left.\begin{array}{l}\sigma^2\text{ 小，质量集中在中心，}f(x)\text{ 小}\\ \sigma^2\text{ 大，质量分散到尾部，}f(x)\text{ 小}\end{array}\right. \Longrightarrow \sigma^2\text{ 有使 }f(x)\text{ 极大的值}$$

3.3 极大似然估计的一般做法

设总体 X 的概率函数为 $f(x;\theta)$，其中 $\theta = (\theta_1, \theta_2, \cdots, \theta_m)$ 为未知参数. $(X_1, \cdots, X_n)$ 是来自总体 X 的样本.

若 X 是离散型随机变量，则 $f(x;\theta)$ 是 X 的概率质量函数. 样本的联合概率质量函数为

$$\begin{aligned}
&\mathbb{P}((X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n))\\
=&\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)\\
=&\mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)\cdots\mathbb{P}(X_n = x_n)\\
=&\prod_{i=1}^{n}\mathbb{P}(X_i = x_i)\\
=&\prod_{i=1}^{n} f(x_i;\theta).
\end{aligned}$$

若 X 是连续型随机变量，样本 $(X_1, X_2, \cdots, X_n)$ 在样本值为 $(x_1, x_2, \cdots, x_n)$ 附近取值的概率为

$$\begin{aligned}
&\mathbb{P}(X_1 \in (x_1, x_1 + h_1], X_2 \in (x_2, x_2 + h_2], \cdots, X_n \in (x_n, x_n + h_n]) \\
&= \prod_{i=1}^{n} \mathbb{P}(X_i \in (x_i, x_i + h_i]) \\
&\approx \prod_{i=1}^{n} f(x_i; \theta) h_i.
\end{aligned}$$

可见，该概率取决于

$$\prod_{i=1}^{n} f(x_i; \theta).$$

因此，在离散型或连续型随机变量的情形，样本取值 $(x_1, \cdots, x_n)$ 或取值 $(x_1, \cdots, x_n)$ 附近的概率决定于样本的联合概率函数

$$\prod_{i=1}^{n} f(x_i; \theta).$$

它依赖于参数 θ 的取值，因而将它视为参数的函数

$$L(\theta) := \prod_{i=1}^{n} f(x_i; \theta),$$

称之为参数 θ 的似然函数.

样本值 $(x_1, \cdots, x_n)$ 出现了，则应该认为参数 θ 当使 $(x_1, \cdots, x_n)$ 最有可能出现，即 θ 应使

$$\prod_{i=1}^{n} f(x_i, \theta) \tag{3.1}$$

极（最）大.

使 $L(\theta)$ 达到最大的 $\hat{\theta}$ 称之为 θ 的极大似然估计，即

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

考虑到 $L(\theta)$ 往往是乘积的形式，总体的概率函数 $f(x; \theta)$ 多是指数函数，因而为计算方便，长考虑极大似然函数的对数

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta),$$

并称之为对数似然函数.

因为 $\log x$ 是 x 的增函数，所以 $L(\theta)$ 与 $l(\theta)$ 的极大值点相同.

实际中，我们可以用微积分方法求极值. 若 $L(\theta)$ 或 $l(\theta)$ 是可微函数，则可考察极大似然方程

$$\frac{\partial \log L}{\partial \theta} = 0$$

或

$$\frac{\partial \log l}{\partial \theta} = 0.$$

通过求驻点来找极大点.

例 3.7 设总体 X 服从参数为 $\lambda > 0$ 的指数分布，其概率密度函数为

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

求未知参数 λ 的极大似然估计.

解 似然函数为

$$L(\lambda) = \prod_{i=1}^{n} \left(x \, \mathrm{e}^{-\lambda x_i} \right) = \lambda^n \, \mathrm{e}^{-\lambda \sum_{i=1}^{n} x_i}.$$

取对数，可得对数似然函数为

$$l(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

对数似然方程为

$$\frac{dl(\lambda)}{d\lambda} = \frac{d \log L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

似然估计为

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}}.$$

由此可得 λ 的极大似然估计量为

$$\hat{\lambda} = \frac{1}{\overline{X}}.$$

□

例 3.8 求正态分布总体 $N(\mu, \sigma^2)$ 的参数 μ, σ^2 的极大似然估计.

解 似然函数为

$$\begin{aligned} L\left(\mu, \sigma^2\right) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right] \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right). \end{aligned}$$

考虑 L 的对数，

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

对数似然方程组为

$$\begin{cases}\dfrac{\partial l}{\partial \mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0,\\ \dfrac{\partial l}{\partial \sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0.\end{cases}$$

可得

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x},$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

所以，μ,σ^2 的极大似然估计样本量分别为 $\overline{X}$ 和 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})$. □

例 3.9 设总体 X 服从区间 $[0,\theta]$ $(\theta>0)$ 上的均匀分布，参数为 θ. 求 θ 的极大似然估计.

解 设 $x_{(n)}:=\max\{x_1,x_2,\dots,x_n\}$. 随机变量 X 的概率密度函数为

$$f(x;\theta)=\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)=\frac{1}{\theta}\mathbb{1}_{[x,+\infty)}(\theta).$$

因此，似然函数为

$$\begin{aligned}L(\theta)&=L(\theta;x_1,x_2,\dots,x_n)\\&=\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_1)\cdot\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_2)\cdots\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_n)\\&=\frac{1}{\theta^n}\mathbb{1}_{[x_1,+\infty)}(\theta)\cdot\mathbb{1}_{[x_2,+\infty)}(\theta)\cdots\mathbb{1}_{[x_n,+\infty)}(\theta)\\&=\frac{1}{\theta^n}\mathbb{1}_{[x_1,+\infty)\cap[x_2,+\infty)\cdots\cap[x_n,+\infty)}(\theta)\\&=\frac{1}{\theta^n}\mathbb{1}_{[x_{(n)},+\infty)}(\theta).\end{aligned}$$

用分段函数表示就是

$$L(\theta)=\begin{cases}\frac{1}{\theta^n}, & 0\le x_{(n)}\le\theta,\\ 0, & \text{其他}.\end{cases}$$

样本值是给定的，$L(\theta)$ 的图形如下.

显然，为使 $L(\theta)$ 极大，θ 应满足 $\theta\ge x_{(n)}$. 当 $\theta\ge x_{(n)}$ 时，$L(\theta)$ 是单调减函数，从而

$$L(\theta)=\frac{1}{\theta^n}\le\frac{1}{x_{(n)}^n}.$$

故 $L(\theta)$ 在 $\hat{\theta}=x_{(n)}$ 时达到最大值. 因此，θ 的极大似然估计值为 $x_{(n)}$，极大似然估计量为

$$\hat{\theta}=X_{(n)}:=\max\{X_1,X_2,\cdots,X_n\}.$$

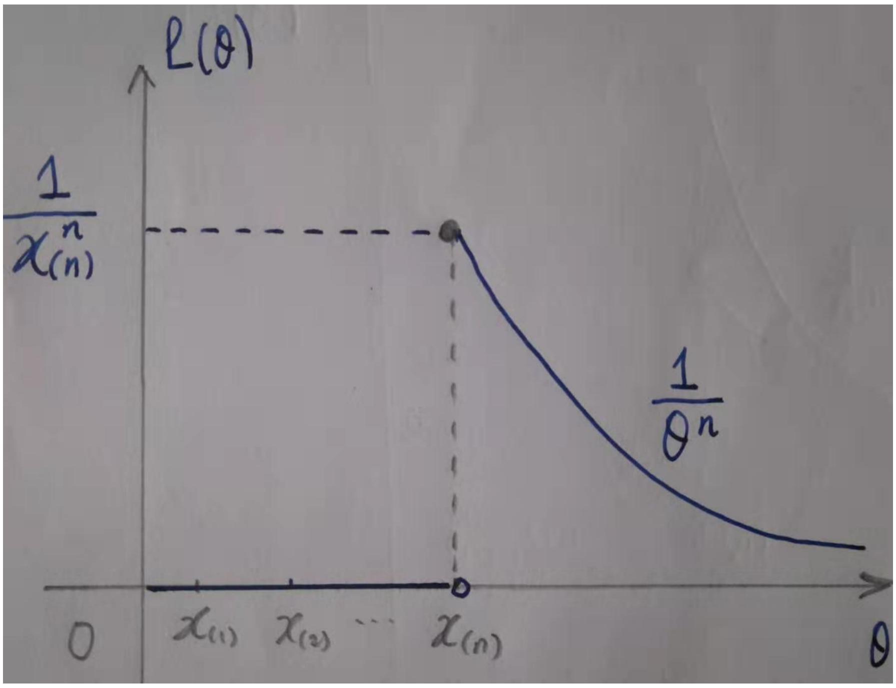


图 3.7 极大似然函数

□

例 3.10 设总体 X 服从区间 $[\theta_1, \theta_2]$（$\theta_1 < \theta_2$）上的均匀分布，求参数 θ_1, θ_2 的极大似然估计.

解 设 $X_1, X_2, \ldots, X_n$ 为来自总体 X 的样本，$x_1, x_2, \ldots, x_n$ 为样本值. 总体 X 的概率密度函数可以表示为

$$f(x; \theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} \mathbb{1}_{[\theta_1, \theta_2]}(x).$$

似然函数可以表示为

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2) \\ &= \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} \mathbb{1}_{[\theta_1, \theta_2]}(x_i) \\ &= \frac{1}{(\theta_2 - \theta_1)^n} \prod_{i=1}^{n} \mathbb{1}_{[\theta_1, \theta_2]}(x_i) \\ &= \frac{1}{(\theta_2 - \theta_1)^n} \mathbb{1}_{[\theta_1, \theta_2]}(x_{(1)}) \cdot \mathbb{1}_{[\theta_1, \theta_2]}(x_{(n)}). \end{aligned}$$

显然，我们要在条件 $\theta_1 \geq x_{(1)}$，$\theta_2 \leq x_{(n)}$ 下考虑 $L(\theta_1, \theta_2)$ 的极大值，因为否则 $L(\theta_1, \theta_2)$ 的值为 0. 又因为，$L(\theta_1, \theta_2)$ 对 θ_1 是单调减的，对 θ_2 是单调增的，所以当

$$\theta_1 = x_{(1)}, \quad \theta_2 = x_{(n)}$$

时，$L(\theta_1, \theta_2)$ 取到极大值.

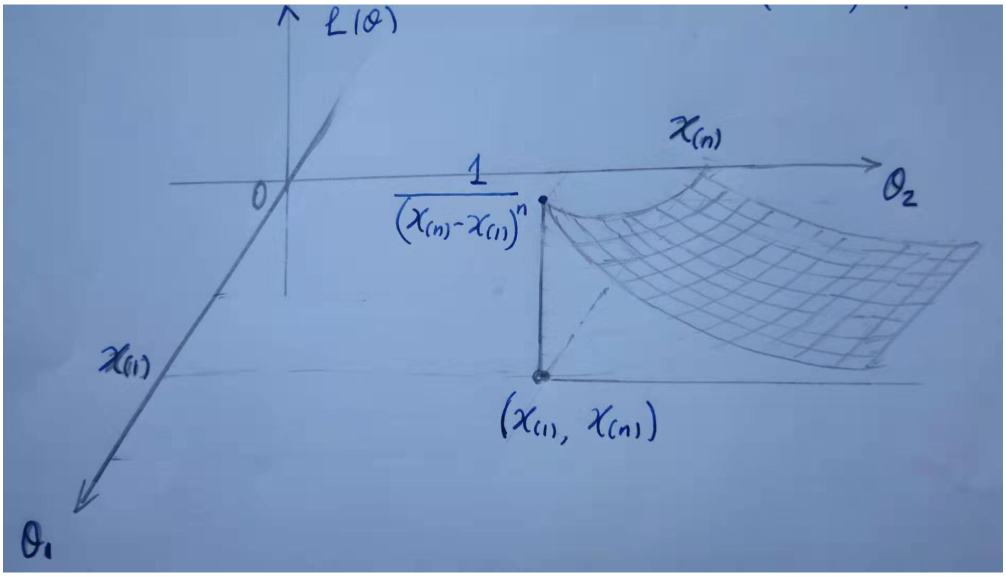


图 3.8 极大似然函数

□

例 3.11（Poisson 分布） 参数为 λ 的 Poisson 分布的参数 λ 的极大似然估计.

解 设 $X_1, X_2, \cdots, X_n$ 为一个样本，则

$$\mathbb{P}\left(X=x_i\right)=\frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \quad i=1,2,\cdots n.$$

似然函数为

$$\begin{aligned} R(\lambda) &= \prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}\right)e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_i!\cdots x_n!}\cdot e^{-n\lambda} \end{aligned}$$

对数似然函数为

$$\log L(\lambda)=\left(\sum_{i=1}^{n}x_i\right)\log\lambda-\log\left(\pi x_i!\right)-n\lambda.$$

似然方程为

$$\frac{d\log L(\lambda)}{d\lambda}=\Sigma x_i\cdot\frac{1}{\lambda}-n=0.$$

由此可得似然估计值为

$$\hat{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}.$$

因此，λ 的极大似然估计量为 $\overline{X}$. □

例 3.12（多项分布） 设 $X_1, X_2, \cdots, X_n$ 服从参数为 $p_1, p_2, \cdots, p_n$ 的多项分布：

$$\mathbb{P}(X_1=x_1, X_2=x_2, \cdots, X_m=x_m)=\frac{n!}{\prod_{i=1}^{n}x_i!}\prod_{i=1}^{n}p_i^{x_i}.$$

对数似然函数为

$$l(p_1,p_2,\cdots,p_m)=\log n!-\sum_{i=1}^n \log x_i!+\sum_{i=1}^n x_i\log p_i.$$

因为 $\sum_{i=1}^m p_i=1$，故用拉格朗日乘子法，考虑函数

$$l(p_1,p_2,\cdots,p_m,\lambda):=\log n!-\sum_{i=1}^n \log x_i!+\sum_{i=1}^n x_i\log p_i+\lambda\left(\sum_{i=1}^n p_i-1\right).$$

由

$$\frac{\partial l}{\partial p_i}=0$$

可得

$$\frac{x_i}{p_i}+\lambda=0.$$

因此

$$\hat{p}_i=-\frac{x_i}{\lambda}.$$

又由 $\sum_{i=1}^n \hat{p}_i=1$ 可得 $\lambda=-n$，于是有

$$\hat{p}_i=\frac{x_i}{n}.$$

例 3.13（Hardy-Weinberg 平衡） Hardy–Weinberg 平衡定律说对无限大的群体，假设随机婚配，没有突变，没有选择，没有迁移，没有遗传漂变，则群体内一个位点上的基因型频率和基因频率将代代保持不变，处于遗传平衡状态. 设某基因类型 AA,Aa,aa 在群体中出现的频率为 $(1-\theta)^2,2\theta(1-\theta),\theta^2$.

1937 年对香港居民的一次血型检查中，得到如下结果

M	MN	N	全
342	500	187	1029

表 3.1 血型频率

求 θ 的极大似然估计.

解 设 X_1,X_2,X_3 分别表示 n 个个体中基因类型为 AA,Aa,aa 的次数，因此基因类型 AA,Aa,aa 各出现 x_1,x_2,x_3 次的概率为

$$\mathbb{P}(X_1=x_1,X_2=x_2,X_3=x_3)=\frac{n!}{x_1!x_2!x_3!}(1-\theta)^{2x_1}[2\theta(1-\theta)]^{x_2}\theta^{2x_3}.$$

因此，对数似然函数为

$$
\begin{aligned}
l(\theta) &= \log n! - \sum_{i=1}^{3} \log x_i! + x_1 \log(1-\theta)^2 + x_2 \log 2\theta(1-\theta) + x_3 \log \theta^2 \\
&= \log n! - \sum_{i=1}^{3} \log x_i! + x_2 \log 2 + (x_2 + 2x_3) \log \theta + (2x_1 + x_2) \log(1-\theta).
\end{aligned}
$$

令 $l'(\theta) = 0$ 可得

$$
\frac{x_2 + 2x_3}{\theta} - \frac{2x_1 + x_2}{1-\theta} = 0.
$$

因此 θ 的极大似然估计值为

$$
\hat{\theta} = \frac{2x_2 + x_3}{2x_1 + 2x_2 + 2x_3} = \frac{2x_2 + x_3}{2n\bar{x}}.
$$

带入具体数值可得

$$
\hat{\theta} = 0.4247.
$$

□

例 3.14（对数正态分布） 设 $Y \sim N(\mu, \sigma^2)$，$X = \exp(Y)$. 称 X 服从参数为 μ, σ 的对数正态分布. 随机变量 X 的概率密度函数如下

$$
f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi} x\sigma} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} & x > 0 \\ 0 & \text{其他}. \end{cases}
$$

设 Y 的对数似然函数为 $\log L_\phi$，则对数正态分布的似然函数为

$$
\begin{aligned}
&\log L_L(\mu, \sigma | x_1, x_2, \cdots, x_n) \\
&= -\sum_k \log x_k + \log L_\varphi(\mu, \sigma | \log x_1, \log x_2, \cdots, \log x_n) \\
&= \text{常数} + \log L_\varphi(\mu, \sigma | \log x_1, \log x_2, \cdots, \log x_n).
\end{aligned}
$$

根据正态分布的最大似然估计，可得对数正态分布参数最大似然估计值为：

$$
\hat{\mu} = \frac{\sum_{k=1}^{n} \log x_k}{n}, \quad \hat{\sigma}^2 = \frac{\sum_{k=1}^{n} (\log x_k - \hat{\mu})^2}{n}.
$$

所以，对数正态分布中参数的极大似然估计为

$$
\hat{\mu} = \frac{}{n} \sum_{k=1}^{n} \log X_k, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (\log X_k - \hat{\mu})^2.
$$

例 3.15 设总体 X 服从参数为 $\theta \in \mathbb{R}$ 的 Laplace 分布，即它有概率密度函数为

$$
f(x, \theta) = \frac{1}{2} \mathrm{e}^{-|x-\theta|}, \quad x \in \mathbb{R}.
$$

设 $X_1, X_2, \cdots, X_n$ 为来自 X 的样本. 求 θ 的极大似然估计.

解 样本 $X_1, X_2, \cdots, X_n$ 的联合密度为

$$f(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{2} \mathrm{e}^{-|x_i - \theta|} = \left(\frac{1}{2}\right)^n \mathrm{e}^{-\sum_{i=1}^{n} |x_i - \theta|}.$$

对数似然函数为

$$l(\theta) = -\log 2 - \sum_{i=1}^{n} |x_i - \theta|.$$

利用符号函数

$$\operatorname{sign}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0, \end{cases}$$

可以将 $l'(\theta)$ 表示为

$$l'(\theta) = \sum_{i=1}^{n} \operatorname{sign}(x_i - \theta).$$

若 n 为偶数，则 $(x_{(n/2)}, x_{(n/2+1)})$ 中的任何数都是 θ 的极大似然估计；若 n 为奇数，则中位数 $x_{((n+1)/2)}$ 是 θ 的唯一极大似然估计.

□

3.4 高斯如何发现正态分布

设随机误差的概率密度函数形如 $f(x - \theta)$，θ 为未知参数，f 的具体形式未知. 设样本观察值为 $x_1, x_2, \cdots, x_n$，则似然函数为

$$L(\theta) = \prod_{i=1}^{n} f(x_i - \theta).$$

对数似然函数为

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i - \theta).$$

求导可得

$$\frac{dl}{d\theta} = -\sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)}.$$

记

$$g(x) = \frac{f'(x)}{f(x)}.$$

高斯认为 θ 的极大似然估计应该为均值，所以

$$\sum_{i=1}^{n} g(x_i - \bar{x}) = 0. \tag{3.2}$$

设有观察值 $x_1 = x, x_2 = -x, \bar{x} = 0$，则函数方程(3.2)意味着

$$g(x) + g(-x) = 0.$$

这说明 $g(x)$ 为偶函数.

一般地，设有如下 $n+1$ 个观察值，

$$x_1 = x_2 = \cdots = x_n = -x, x_{n+1} = nx$$

则 $\bar{x} = 0$，由(3.2)可得

$$ng(-x) = -g(nx).$$

因此，利用 $g(-x) = g(x)$ 可得

$$ng(x) = g(nx).$$

假设 g 是连续函数，根据熟知的程序可知 g 是线性函数

$$g(x) = kx + b,$$

其中 k, b 为常数. 所以

$$f'(x) = (kx + b)f(x).$$

由此可知 f 的形式如下：

$$f(x) = c\exp\left(-\frac{1}{2}rx^2\right).$$

注意到 f 是密度函数，可知 $r > 0$.

3.5 贝叶斯估计

在经典估计方法中，总体分布的参数 θ 被视为未知的，但是确定的. 贝叶斯估计的一个关键想法是将未知参数当做随机变量. 我们不说参数的值应该为多少，而是对参数的取值赋予相应的概率. 参数的初始分布被称为先验分布，它依赖于估计者的经验，是主管的. 这其实是非常自然的. 按贝叶斯的观点，在得到总体的观测值之前，假设参数有先验分布，抽样得到样本，在得到观察值之后，再利用观察值所提供的信息，修正相应的先验分布而得到参数的后验分布. 由后验分布的均值或中位数等可以得到参数的点估计.

设总体 X 的概率函数为 $f(x;\theta)$. 它依赖于未知参数 θ. 设 Θ 为随机变量，先验分布为 $f_\Theta(\theta)$. 我们将参数 θ 看作随机变量 Θ 的取值，因而概率函数 $f(x;\theta)$ 可以被视为在参数 $\Theta = \theta$ 时条件分布：

$$f_{X|\Theta}(x|\theta).$$

因此，(X,Θ) 的联合概率函数为

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_{\Theta}(\theta).$$

如果已知 $X=x$，则 Θ 的条件概率函数（分布密度），即后验分布为

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)},$$

其中关于 X 的边缘概率函数 $f_X(x)$ 可以如下计算：

$$f_X(x) = \int f_{X,\Theta}(x,\theta)d\theta = \int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta.$$

于是 Θ 的后验分布可以表示为

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta}.$$

它表示在观察到 X 的取值后，Θ 更有可能服从的分布.

在具体计算中，我们可以先不计算 $\int f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta$，而是先利用

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)f_{\Theta}(\theta) \tag{3.3}$$

得到后验分布的形式再计算分布的具体形式.

(3.3)说参数的后验密度正比于似然函数与参数的先验分布的乘积.

例 3.16 设总体服从参数为 λ 的 Poisson 分布，$X_1,X_2,\cdots,X_n$ 为来自 X 的样本，参数 λ 的先验分布为 $f_{\Lambda}(\lambda)$. 求 λ 的贝叶斯估计.

解 设参数 λ 为随机变量 Λ 的取值，则似然函数为

$$f_{X|\Lambda}(x|\lambda) = \prod_{i=1}^{n}\frac{\lambda^{x_i}\mathrm{e}^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\lambda}}{\prod_{i=1}^n x_i}.$$

于是，λ 的后验分布为

$$f(\lambda|x_1,x_2,\cdots,x_n) \propto \lambda^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\lambda}\cdot f_{\Lambda}(\lambda).$$

设 Λ 服从参数为 α,ν 的 Γ 分布

$$f_{\Lambda}(\lambda) = \frac{\nu^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\nu\lambda},$$

则 λ 的后验分布为

$$f(\lambda|x_1,x_2,\cdots,x_n) \propto \lambda^{\alpha-1+\sum_{i=1}^n x_i}\mathrm{e}^{-(n+\nu)\lambda}.$$

即 λ 的后验分布为参数为 $\alpha+\sum_{i=1}^n x_i, n+\nu$ 的 Γ 分布，后验分布的均值为

$$\mathbb{E}(\Lambda|x_1,x_2,\cdots,x_n) = \frac{\alpha+\sum_{i=1}^n x_i}{n+\nu}.$$

□

例 3.17 设 $X_1,X_2,\cdots,X_n$ 服从参数为 θ 的两点分布 $B(1,\theta)$. 设参数 θ 为随机变量 Θ 的取值，服从参数为 α,β 的 Beta 分布，求 θ 的贝叶斯估计.

解 X_i 的概率函数为

$$f(x_i,\theta)=\theta^{x_i}(1-\theta)^{1-x_i},\quad x_i=0,1.$$

样本 $X_1,X_2,\cdots,X_n$ 与 Θ 的联合概率函数为

$$\begin{aligned}f(x_1,x_2,\cdots,x_n,\theta)&=\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\&=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\theta^{\alpha+\sum_{i=1}^n x_i-1}(1-\theta)^{n+\beta-1-\sum_{i=1}^n x_i},\end{aligned}$$

其中 $x_i=0,1,i=1,2,\cdots,n,\theta\in(0,1)$. 由此可见，$\theta$ 的后验分布

$$f(\theta|x_1,x_2,\cdots,x_n)\propto\theta^{\alpha+\sum_{i=1}^n x_i-1}(1-\theta)^{n+\beta-\sum_{i=1}^n x_i-1}.$$

即 θ 的后验分布服从 $\mathrm{Beta}(\alpha+\sum_{i=1}^n x_i,n+\beta-\sum_{i=1}^n x_i)$. θ 的后验均值为

$$\mathbb{E}(\theta|x_1,x_2,\cdots,x_n)=\frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\beta+n}.$$

□

内容提要

- 点估计
- 矩估计
- 极大（最大）似然估计
- 贝叶斯估计
- 高斯与正态分布

第 3 章 习题

1. 第二次世界大战期间，德国给自己制造的坦克进行编号. 盟军通过缴获、情报，记录下了部分坦克生产编号. 如何通过已知的编号估计德军的总坦克数?
2. 设总体 X 服从参数为 N 和 p 的二项分布，$x_1,\cdots,x_n$ 为取自 X 的样本，试求参数 N,p 的矩估计.
3. 设总体概率密度为
$$f(x;a)=\begin{cases}(a+1)x^a, & 0<x<1,\\0, & \text{其他},\end{cases}$$
其中 $a>-1$. 试用样本 $x_1,x_2,\cdots,x_n$ 求参数 a 的矩估计和极大似然估计.
4. 设总体的概率密度为
$$f(x;\theta)=\begin{cases}\exp(-(x-\theta)), & x\geq\theta,\\0, & \text{其他},\end{cases}$$
试用样本 $x_1,x_2,\cdots,x_n$ 求参数 θ 的极大似然估计.
5. 设总体 X 服从几何分布
$$\mathbb{P}(X=k)=p(1-p)^{k-1},\quad k=1,2,\cdots,$$

其中 $0 < p < 1$，试利用样本 $x_1, x_2, \cdots, x_n$ 求 p 的极大似然估计.

第 4 章 估计的性质

用不同方法得到的总体分布中未知参数的估计不一定相同. 哪一个好呢? 有什么标准来评估? 本章介绍常用的标准：无偏性、有效性与一致性. 还将介绍统计量的充分性和完备性以及点估计的大样本理论.

4.1 无偏性

定义 4.1. 无偏性

设 $X_1, X_2, \cdots, X_n$ 为具有参数 θ 的总体的样本. 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$ 为 θ 的估计量，如果

$$\mathbb{E}\hat{\theta} = \theta, \tag{4.1}$$

则称 $\hat{\theta}$ 为 θ 的无偏估计量（unbiased estimator），$\hat{\theta}(x_1, x_2, \cdots, x_n)$ 为无偏估计值，否则，称 $\hat{\theta}$ 为 θ 的有偏估计，偏差为

$$b(\theta) = \hat{\theta} - \theta.$$

在不引起混淆的场合，粗略地称无偏估计值和无偏估计量为无偏估计.

例 4.1 样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ 是总体均值 μ 的无偏估计；由(2.16)可知，样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ 是总体方差的无偏估计. 而 $\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ 是 σ^2 的有偏估计.

一般地，设总体 X 的 k 阶原点矩 μ_k 存在，则样本 k 阶原点矩为总体 k 矩的无偏估计. 事实上，

$$\begin{aligned} \mathbb{E}A_k &= \mathbb{E}\frac{1}{n}\sum_{i=1}^n X_i^k = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^k \\ &= \frac{1}{n}\cdot n\mu_k = \mu_k. \end{aligned}$$

但样本中心距不一定是总体中心距的无偏估计. 此外，若 $\hat{\theta}$ 为 θ 的无偏估计，则 $\hat{\theta}$ 的函数 $f(\hat{\theta})$ 不一定是 $f(\theta)$ 的无偏估计.

例 4.2 设 X 服从参数为 θ 的指数分布，即对任意的 $x > 0$，$\mathbb{P}(X > x) = \mathrm{e}^{-\theta x}$. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本，则 $\overline{X}$，$nX_{(1)}$ 是 $\frac{1}{\theta}$ 的无偏估计，这里

$$X_{(1)} := \min\{X_1, X_2, \cdots, X_n\}.$$

证明 因为

$$\mathbb{E}\overline{X} = \mathbb{E}X = \frac{1}{\theta},$$

所以显然 $\overline{X}$ 是 $\frac{1}{\theta}$ 的无偏估计.

另对任意 $x > 0$，

$$\begin{aligned}\mathbb{P}(X_{(1)} > x) =& \mathbb{P}(X_1 > x, X_2 > x, \cdots, X_n > x)\\ =& \prod_{i=1}^{n}\mathbb{P}(X_i > x) = \prod_{i=1}^{n}\mathrm{e}^{-\theta x} = \mathrm{e}^{-n\theta x},\end{aligned}$$

所以 $X_{(1)}$ 服从参数为 $n\theta$ 的指数分布，因此，

$$\mathbb{E}nX_{(1)} = n\cdot\frac{1}{n\theta} = \frac{1}{\theta}.$$

所以 $nX_{(1)}$ 也是 $\frac{1}{\theta}$ 的无偏估计. □

例 4.3 设 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$，则

1. $\frac{n+1}{n}X_{(n)}$ 为 θ 的无偏估计；
2. $2\overline{X}$ 为 θ 的无偏估计.

解 （1）$X_{(n)}$ 的概率密度为

$$f_n(x) = \frac{n}{\theta^n}x^{n-1}, \quad x \in (0, \theta).$$

因此，易计算得到

$$\begin{aligned}\mathbb{E}\left[\frac{n+1}{n}X_{(n)}\right] &= \frac{n+1}{n}\int_0^\theta x\cdot\frac{n}{\theta^n}x^{n-1}dx\\ &= (n+1)\int_0^\theta \frac{1}{\theta^n}x^n dx\\ &= \theta.\end{aligned}$$

因此，$\frac{n+1}{n}X_{(n)}$ 为 θ 的无偏估计.

（2）显然

$$\mathbb{E}2\overline{X} = 2\mathbb{E}\overline{X} = 2\mathbb{E}X_1 = 2\times\frac{\theta}{2} = \theta.$$

□

定义 4.2. 均方误差

设 $\hat{\theta}$ 为 θ 的估计，称

$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}(\hat{\theta} - \theta)^2$$

为 $\hat{\theta}$ 的均方误差.

命题 4.1

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}\,\hat{\theta} + (\mathbb{E}\hat{\theta} - \theta)^2.$$

特别，如果 $\hat{\theta}$ 是 θ 的无偏估计，则

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}\,\hat{\theta}.$$

证明

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \theta)^2 \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))) + (\mathbb{E}\hat{\theta} - \theta)^2 \\
&= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + ((\mathbb{E}(\hat{\theta}) - \theta))^2 \\
&= \mathrm{Var}\,\hat{\theta} + (\mathbb{E}\hat{\theta} - \theta)^2,
\end{aligned}$$

其中利用了

$$\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) = (\mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}(\hat{\theta}) - \theta) = 0.$$

□

4.2 有效性

在例4.2中，$\overline{X}$ 与 $nX_{(1)}$ 都是 $\frac{1}{\theta}$ 的无偏估计，在例4.3中，$2\overline{X}$，$\frac{n+x}{n}X_{(n)}$ 都是 θ 的无偏估计. 选择哪一个更好呢？估计量是一个随机变量，我们希望它的方差足够小. 这就引出有效性的概念.

定义 4.3. 无偏估计的有效性

设 $\hat{\theta}_1$,$\hat{\theta}_2$ 均为 θ 的无偏估计. 如果

$$\mathrm{Var}_\theta\,\hat{\theta}_1 \le \mathrm{Var}_\theta\,\hat{\theta}_2,$$

且存在某 θ_0 使得

$$\mathrm{Var}_{\theta_0}\,\hat{\theta}_1 < \mathrm{Var}_{\theta_0}\,\hat{\theta}_2,$$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效. 如果 $\hat{\theta}$ 是 θ 的所有无偏估计中，方差最小的，则称之为有效估计量（efficient estimator）.

例 4.4 续例4.2.

$$\mathrm{Var}\,\overline{X} = \frac{1}{n\theta^2}.$$

$X_{(1)}$ 服从参数为 $n\theta$ 的指数分布，所以

$$\mathrm{Var}\,nX_{(1)} = n^2 \cdot \frac{1}{n^2\theta^2} = \frac{1}{\theta^2}.$$

可见，当 $n > 1$ 时，$\overline{X}$ 比 $nX_{(1)}$ 更有效.

例 4.5 设 $X_1, X_2, \cdots, X_n$ 为样本，$\overline{X}_p = \sum_{i=1}^n p_i X_i$，其中 $p_1, p_2, \cdots, p_n$ 为非负数且 $\sum_{i=1}^n p_i = 1$. 则 $\overline{X}_p$ 是无偏估计. 因为

$$\operatorname{Var} \overline{X}_p = \sum_{i=1}^n p_i^2 \operatorname{Var} X$$

由柯西—施瓦茨不等式，

$$1 = \left(\sum_{i=1}^n 1 \cdot p_i\right)^2 \leq \sum_{i=1}^n 1^2 \cdot \sum_{i=1}^n p_i^2 = n \sum_{i=1}^n p_i^2$$

可得

$$\operatorname{Var} \overline{X}_p \geq \frac{1}{n} \operatorname{Var} X = \operatorname{Var} \overline{X}.$$

因此，$\overline{X}$ 比 $\overline{X}_p$ 有效.

例 4.6 续例4.3. 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $X \sim U(0, \theta)$ 的样本，则 $\frac{n+1}{n} X_{(n)}$，$2\overline{X}$ 均为 θ 的无偏估计. 但

$$\operatorname{Var}(2\overline{X}) = 4 \cdot \frac{1}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

而

$$\mathbb{E} X_{(n)}^2 = \int_0^\theta x^2 \cdot \frac{n}{\theta^n} x^{n-1} = \int_0^\theta \frac{n}{\theta^n} x^{n+1} = \frac{n}{n+2} \theta^2.$$

所以，

$$\begin{aligned}
\operatorname{Var}\left(\frac{n+1}{n} X_{(n)}\right) &= \mathbb{E}\left(\frac{n+1}{n} X_{(n)}\right)^2 - \theta^2 \\
&= \left(\frac{n+1}{n}\right)^2 \cdot \mathbb{E} X_{(n)}^2 - \theta^2 \\
&= \left(\frac{n+1}{n}\right)^2 \cdot \left(\frac{n}{n+2} \theta^2\right) - \theta^2 \\
&= \left(\frac{(n+1)^2}{n(n+2)} - 1\right) \theta^2 \\
&= \frac{\theta^2}{n(n+2)}.
\end{aligned}$$

因此，当 $n > 1$ 时，$\frac{n+1}{n} X_{(n)}$ 比 $2\overline{X}$ 更有效.

4.3 Crámer-Rao 不等式

苏格拉底曾要求柏拉图到一片麦地里摘一株最饱满的麦穗回来. 但只准往前走，不准后退. 结果柏拉图两手空空的走出了麦地，原因就是柏拉图不知道什么是最好的.

在参数的无偏估计中，有没有最有效的估计？最有效的估计是什么？这就涉及到估计的方差是否有下界. Cramér-Rao 不等式给出了肯定的回答.

Cramér-Rao 不等式由 Rao 于 1945 年证明. 同一时期的 Harald Cramér 也得到了该结果，并收入他于 1946 年出版的书《统计学中的数学方法》中. Dembo, Thomas Cover (1991) 在"Information Theoretic Inequalities"中揭示了 Weyl-Heisenberg 不确定性准则可以由 Cramér-Rao 不等式得到.

Cramér-Rao 不等式借助 Fisher 信息给出了无偏估计的方差的下界.

设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本，X 的概率函数为 $f(x|\theta)$.

定义 4.4. Fisher 信息

设 f 满足一定的光滑条件，定义

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2,$$

称之为 θ 的 Fisher 信息.

定理 4.1. Cramér-Rao 不等式

设 $T = T(X_1, X_2, \cdots, X_n)$ 为 θ 的无偏估计，f 满足一定的光滑条件，则

$$\operatorname{Var} T \geq \frac{1}{nI(\theta)}.$$

先证明一些要用到的结论.

引理 4.1

设 f 满足一定的光滑条件，则

1. $$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = 0.$$

2. $$\operatorname{Var}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = I(\theta).$$

3. $$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right].$$

证明 (1) 注意到

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= \int\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]f(x|\theta)dx \\
&= \int\frac{\partial}{\partial\theta}f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \frac{\partial}{\partial\theta}1 = 0.
\end{aligned}$$

(2) 上面的结论说明（证明中也用到）

$$\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)dx = 0.$$

对上式两边再次关于 θ 求偏导数，可得

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)dx \\
&= \int \frac{\partial}{\partial\theta}\left(\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\right) dx \\
&= \int \frac{\partial^2}{\partial\theta^2}\left[\log f(x|\theta) f(x|\theta)\right] d\theta + \frac{\partial}{\partial\theta}\log f(x|\theta)\frac{\partial}{\partial\theta}f(x|\theta)dx \\
&= \int \frac{\partial^2}{\partial\theta^2}\left[\log f(x|\theta)\right] f(x|\theta)dx + \int \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 f(x|\theta)dx \\
&= \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2,
\end{aligned}$$

其中用到如下等式

$$\frac{\partial}{\partial\theta}f(x|\theta) = \frac{\partial}{\partial\theta}\left(\log f(x|\theta)\right) f(x|\theta).$$

由此可得 (2).

(3) 因为 $\mathbb{E}\frac{\partial}{\partial\theta}\log f(X|\theta) = 0$ 可知

$$\operatorname{Var}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 = I(\theta).$$

□

引理 4.2

设 $X_1, X_2, \cdots, X_n$ 为来自总体 $X \sim f(x|\theta)$ 的样本，

$$Z := \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta),$$

则

$$\mathbb{E}Z = 0, \quad \operatorname{Var} Z = nI(\theta).$$

证明 因 $X_1, X_2, \cdots, X_n$ 独立同分布，所以

$$\frac{\partial}{\partial\theta}\log f(X_1|\theta), \frac{\partial}{\partial\theta}\log f(X_2|\theta), \cdots, \frac{\partial}{\partial\theta}\log f(X_n|\theta)$$

独立同分布，因此

$$\mathbb{E}Z = n\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = 0.$$

类似地，

$$\operatorname{Var} Z = n\operatorname{Var}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = nI(\theta).$$

□

现在可以证明 Cramér-Rao 不等式.

证明 [定理4.1的证明] 由协方差不等式，有

$$\mathrm{Cov}^2(Z,T) \leq \mathrm{Var}\, Z\, \mathrm{Var}\, T.$$

由前面的引理，已知

$$\mathrm{Var}\, Z = nI(\theta).$$

我们只要证明

$$\mathrm{Cov}(Z,T) = 1$$

即可得证本结论.

因为 $\mathbb{E}Z = 0$，所以

$$\mathrm{Cov}(Z,T) = \mathbb{E}(ZT) - \mathbb{E}Z \cdot \mathbb{E}T = \mathbb{E}(ZT).$$

而

$$\begin{aligned}
\mathbb{E}(ZT) &= \int_{\mathbb{R}^n} T(x_1,x_2,\cdots,x_n)\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i|\theta)\prod_{j=1}^n f(x_j|\theta)dx_1dx_2\cdots dx_n \\
&= \int_{\mathbb{R}^n} T(x_1,x_2,\cdots,x_n)\sum_{i=1}^n \frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\prod_{j=1}^n f(x_j|\theta)dx_1dx_2\cdots dx_n \\
&= \int_{\mathbb{R}^n} T(x_1,x_2,\cdots,x_n)\frac{\partial}{\partial\theta}\prod_{j=1}^n f(x_j|\theta)dx_1dx_2\cdots dx_n \quad \text{（统计量}T\text{与}\theta\text{无关）} \\
&= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} T(x_1,x_2,\cdots,x_n)\prod_{j=1}^n f(x_j|\theta)dx_1dx_2\cdots dx_n \\
&= \frac{\partial}{\partial\theta}\mathbb{E}T(X_1,X_2,\cdots,X_n) \quad \text{（}T\text{是无偏统计）} \\
&= \frac{\partial}{\partial\theta}\theta = 1.
\end{aligned}$$

□

例 4.7（Poisson 分布） 设 $X_1,X_2,\cdots,X_n$ 为来自 $X \sim \mathrm{Poisson}(\lambda)$ 的样本. 由例3.11，λ 的极大似然估计为 $\hat\lambda = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. 因为 $\mathrm{Var}\, X = \lambda$，所以

$$\mathrm{Var}\,\hat\lambda = \mathrm{Var}\,\overline{X} = \frac{1}{n^2}\cdot n\lambda = \frac{\lambda}{n}.$$

Poisson 分布的概率函数为

$$f(x|\lambda) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},$$

所以

$$\log f(x|\lambda) = x\log\lambda - \lambda - \log x!.$$

$$\frac{\partial}{\partial\lambda}\log f(x|\lambda) = \frac{x}{\lambda} - 1,$$

$$\frac{\partial^2}{\partial\lambda^2}\log f(x|\lambda) = -\frac{x}{\lambda^2}.$$

所以，

$$I(\lambda) = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right] = -\mathbb{E}\left(-\frac{X}{\lambda^2}\right)$$
$$= \frac{\mathbb{E}X}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

也可以用如下较为复杂的方法计算：

$$I(\lambda) = \mathbb{E}\left[\frac{\partial}{\partial\lambda}\log f(X|\lambda)\right]^2 = \mathbb{E}\left(\frac{X}{\lambda} - 1\right)^2$$
$$= \frac{1}{\lambda^2}\mathbb{E}(X-\lambda)^2 = \frac{1}{\lambda^2}\mathbb{E}(X-\mathbb{E}X)^2 = \frac{1}{\lambda^2}\operatorname{Var} X$$
$$= \frac{1}{\lambda^2}\cdot\lambda = \frac{1}{\lambda}.$$

因此，λ 的 Cramér-Rao 下界为

$$\frac{1}{nI(\lambda)} = \frac{\lambda}{n} = \operatorname{Var}\overline{X}.$$

即 λ 的极大似然估计达到 Cramér-Rao 下界.

定义 4.5. UMVUE

如果估计 $\hat{\theta}$ 为 θ 的无偏估计，且在所有无偏估计中，$\hat{\theta}$ 有最小的方差，则称之为一致最小方差无偏估计（记为 UMVUE，Uniform Minimum Variace Unbiased Estimate 的缩写）. 达到 Cramér-Rao 下界的无偏估计称为有效的.

显然，有效的估计是 UMVUE.

例 4.8 设 $X_1,\cdots,X_n$ 为来自 $B(1,\theta)$ 的样本，则 $\overline{X}$ 为 θ 的 UMVUE.

证明 总体 $X \sim B(1,\theta)$ 的概率函数为

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

因此

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}.$$

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 = \mathbb{E}\left(\frac{X-\theta}{\theta(1-\theta)}\right)^2$$
$$= \frac{\operatorname{Var} X}{[\theta(1-\theta)]^2} = \frac{\theta(1-\theta)}{[\theta(1-\theta)]^2} = \frac{1}{\theta(1-\theta)}.$$

又，

$$\operatorname{Var}\overline{X} = \frac{1}{n}\operatorname{Var} X = \frac{\theta(1-\theta)}{n} = \frac{1}{nI(\theta)}.$$

所以，$\overline{X}$ 为 θ 的 UMVUE. □

例 4.9 设 $X_1,X_2,\cdots,X_n$ 为来自总体 $X \sim N(\mu,\sigma_0^2)$ 的样本，其中 σ_0^2 已知，μ 未知，则 $\overline{X}$ 为 μ 的 UMVUE.

证明 易知 $\overline{X}$ 为 μ 无偏估计. 因为 X 的概率密度函数为

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right),$$

所以

$$\log f(x|\mu) = -\log(\sqrt{2\pi}\sigma_0) - \frac{(x-\mu)^2}{2\sigma_0^2},$$

而

$$\frac{d}{d\mu}\log f(x|\mu) = -\frac{x-\mu}{\sigma_0^2}.$$

所以

$$\begin{aligned}I(\mu) &= \mathbb{E}\left[\frac{d}{d\mu}\log f(X|\mu)\right]^2 = \mathbb{E}\left(\frac{X-\mu}{\sigma_0^2}\right)^2\\ &= \frac{\operatorname{Var} X}{\sigma_0^4} = \frac{\sigma_0^2}{\sigma_0^4} = \frac{1}{\sigma_0^2}.\end{aligned}$$

所以 μ 的 Cramér-Rao 下界为

$$\frac{1}{nI(\mu)} = \frac{\sigma_0^2}{n} = \operatorname{Var}\overline{X}.$$

故 $\overline{X}$ 为 μ 的 UMVUE. □

注 根据证明，Cramér-Rao 不等式中的等号成立当且仅当证明中用到的协方差不等式取到等式，或等价地说，无偏估计量 T 与 $Z := \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i|\theta)$ 线性相关. 在上述例子中，这个事实是显然的. 具体而言，在例4.7, 4.8, 4.9中，概率函数的对数关于参数的偏导数分为为

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log f(x|\lambda) &= \frac{x}{\lambda} - 1,\\ \frac{\partial}{\partial\theta}\log f(x|\theta) &= \frac{x-\theta}{\theta(1-\theta)},\\ \frac{d}{d\mu}\log f(x|\mu) &= -\frac{x-\mu}{\sigma_0^2},\end{aligned}$$

因此，Cramér-Rao 不等式中构造的随机变量 Z 在例4.7, 4.8, 4.9中对应地分别为

$$\begin{aligned}Z(\lambda) &= \frac{n\overline{X}}{\lambda} - n,\\ Z(\theta) &= \frac{n\overline{X} - n\theta}{\theta(1-\theta)},\\ Z(\mu) &= -\frac{n\overline{X} - n\mu}{\sigma_0^2}.\end{aligned}$$

这表明，这些例子中构造的 Z 均与样本均值线性相关. 在这三个例子中，样本均值恰好为对应参数的无偏估计. 根据注4.3，可知它们也是 UMVUE.

4.4 估计的效率

有的无偏估计的方差不能达到 Cramér-Rao 下界，我们考虑它在多大程度上接近下界.

定义 4.6. 无偏估计的效率

参数 θ 的无偏估计 $\hat{\theta}$ 的效率（efficiency）定义为

$$\mathrm{eff}(\hat{\theta}) = \frac{\frac{1}{nI(\theta)}}{\mathrm{Var}(\hat{\theta})}.$$

显然，$\mathrm{eff}(\hat{\theta})$ 是参数 θ 的函数，它是无偏估计最小可能的方差与 $\hat{\theta}$ 的方差之比. 根据 Cramér-Rao 不等式，$\mathrm{eff}(\hat{\theta}) \leq 1$.

例 4.10 设 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} N(\mu, \theta)$，其中 μ 已知，θ 为未知参数. 求样本方差 S^2 的效率.

解 总体的概率函数为

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{(x-\mu)^2}{2\theta}\right),$$

因此

$$\log f(x|\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\theta - \frac{(x-\mu)^2}{2\theta}.$$

由此可得

$$\begin{aligned}\frac{d^2 \log f(x|\theta)}{d\theta^2} &= \frac{d}{d\theta}\left[-\frac{1}{2\theta} + \frac{(x-\mu)^2}{2\theta^2}\right] \\ &= \frac{1}{2\theta^2} - \frac{(x-\mu)^2}{\theta^3}.\end{aligned}$$

所以

$$\begin{aligned}I(\theta) &= \mathbb{E}\left[-\frac{d^2 \log f(X|\theta)}{d\theta^2}\right] \\ &= \mathbb{E}\left[-\frac{1}{2\theta^2} + \frac{(X-\mu)^2}{\theta^3}\right] \\ &= -\frac{1}{2\theta^2} + \frac{\mathbb{E}(X-\mu)^2}{\theta^3} \\ &= -\frac{1}{2\theta^2} + \frac{\theta}{\theta^3} \\ &= \frac{1}{2\theta^2}.\end{aligned}$$

由此可见

$$\frac{1}{nI(\theta)} = \frac{2\theta^2}{n}.$$

又注意到（参(2.17)）

$$\operatorname{Var} S^2 = \frac{2\theta^2}{n-1}.$$

所以，S^2 的效率为

$$\operatorname{eff}(S^2) = \frac{\frac{2\theta^2}{n}}{\frac{2\theta^2}{n-1}} = \frac{n-1}{n}.$$

□

在上面的例子中，当 $n \to \infty$ 时，S^2 的效率趋于 1，此时我们称估计是渐进有效的.

定义 4.7. 估计的相对效率与渐进相对效率

计 $\hat{\theta}$ 相对于两一个估计 $\tilde{\theta}$ 的相对效率（relative efficiency）定义为

$$\operatorname{eff}(\hat{\theta}, \tilde{\theta} | \theta) = \frac{\operatorname{Var} \operatorname{Var} \hat{\theta}}{\operatorname{Var} \tilde{\theta}}.$$

当样本量 $n \to \infty$ 时，相对效率的极限称为渐进相对效率（ARE,aysmptotic relative efficiency）.

随着分布的改变，效率可能变化很大. 因此有人研究稳健统计. 如对正态总体，样本均值是有效的，但对混合分布，如 $N(\mu, \sigma^2)$ 与 $N(\mu, 100\sigma^2)$ 的混合，来自后者的极端值会影响均值的效率. 相反，截断平均可能更加有效.

例 4.11 对正态分布，中位数相对于均值的渐进相对效率为 $\frac{2}{\pi} \approx 64\%$.

自由度为 5 的 t 分布，中位数相对于均值的渐进相对效率约为 96%.

设有来自总体 $N(\mu, \sigma^2)$ 的样本 $X_1, X_2, \cdots, X_n$，我们想估计 σ^2. 考虑

$$\hat{\sigma}^2, \quad \tilde{\sigma}^2 = \frac{d^2\pi}{2},$$

其中

$$d = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}|.$$

这里的常数 $\sqrt{\frac{\pi}{2}}$ 是因为根据大数定律，$d \to \mathbb{E}|X_1 - \bar{X}| = \sqrt{\frac{2}{\pi}}\sigma$.

$\tilde{\sigma}^2, \hat{\sigma}^2$ 的相对效率为 0.876.

若 X_i 由混合分布组成：$1-\varepsilon$ 的 $N(\mu, \sigma^2)$，ε 的 $N(\mu, 9\sigma^2)$，则相对效率与 ε 有关.

有更好的估计：

$$X_{(3n/4)} - X_{(1n/4)}$$

等.

4.5 估计的渐进理论

定义 4.8. 相合性

设 $\hat{\theta}_n$ 为参数 θ 基于样本量为 n 的样本的估计.

1. 如果当 $n \to \infty$ 时，$\hat{\theta}_n$ 依概率收敛于 θ，即
$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = 0,$$
则称 $\hat{\theta}_n$ 是依概率相容的估计，或称其为弱相合估计.

2. 如果当 $n \to \infty$ 时，$\hat{\theta}_n$ 几乎处处收敛于 θ，即
$$\mathbb{P}(\lim_{n\to\infty} \hat{\theta}_n = \theta) = 1,$$
则称 $\hat{\theta}_n$ 是强相合估计.

3. 如果当 $n \to \infty$ 时，$\hat{\theta}_n$ 以 L^p 收敛于 θ，即
$$\lim_{n\to\infty} \mathbb{E}|\hat{\theta}_n - \theta|^p = 0,$$
则称 $\hat{\theta}_n$ 是 p 阶矩相合估计.

根据弱大数定律，各阶样本矩（设存在）依概率收敛于相应阶的总体矩，如果估计量是有限个样本矩的连续函数，则估计是弱相合的. 这是矩估计的合理性.

下面的定理说明，在一定的条件下，极大似然估计也是弱相合的.

定理 4.2. 极大似然估计的弱相合性

设总体的概率函数 f 满足一定的光滑条件，则样本的极大似然估计是弱相合的.

证明 设
$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$
由弱大数定律，
$$\begin{aligned} \frac{1}{n} l(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta) \\ &\to \mathbb{E} f(X|\theta) = \int \log f(x|\theta) f(x|\theta_0) dx, \end{aligned}$$
其中 θ_0 为真值. 因为样本反映的是参数为真值时的信息，其平均自然应该收敛到真实参数所对应的期望.

作为极大似然估计，$\hat{\theta}_n$ 是 $\frac{1}{n} l(\theta)$ 的极大值点. 因此，在一定条件下，$\hat{\theta}_n$ 应依概率收敛于 $\mathbb{E} f(X|\theta)$ 的极大值点. 下面证明，$\mathbb{E} f(X|\theta)$ 的极大值点为 θ_0.

为此，只要证明 $\mathbb{E}f(X|\theta)$ 关于 θ 的导函数在 $\theta=\theta_0$ 时为 0 即可. 因为

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbb{E}f(X|\theta) &= \frac{\partial}{\partial\theta}\int \log f(x|\theta)f(x|\theta_0)dx \\ &= \int \frac{\partial}{\partial\theta}\log f(x|\theta)f(x|\theta_0)dx \\ &= \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}f(x|\theta_0)dx.\end{aligned}$$

因此，

$$\begin{aligned}\left[\frac{\partial}{\partial\theta}\mathbb{E}f(X|\theta)\right]\Bigg|_{\theta=\theta_0} &= \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)|_{\theta=\theta_0}}{f(x|\theta_0)}f(x|\theta_0)dx \\ &= \int \left[\frac{\partial}{\partial\theta}f(x|\theta)\right]\Bigg|_{\theta=\theta_0}dx \\ &= \left[\frac{\partial}{\partial\theta}\int f(x|\theta)dx\right]\Bigg|_{\theta=\theta_0} \\ &= \left[\frac{\partial}{\partial\theta}1\right]\Bigg|_{\theta=\theta_0} \\ &= 0.\end{aligned}$$

□

定理 4.3. 极大似然估计的近似服从正态分布

设总体的概率函数 f 满足一定的光滑条件，则 $\sqrt{nI(\theta_0)}(\hat{\theta}_n-\theta_0)$ 依分布收敛于标准正态分布.

证明 设总体为 X. 回忆

$$l(\theta)=\sum_{i=1}^n \log f(X_i|\theta).$$

对 $l'(\hat{\theta}_n)$ 在 θ_0 附近作泰勒展开，可得

$$l'(\hat{\theta}_n)\approx l'(\theta_0)+(\hat{\theta}_n-\theta_0)l''(\theta_0).$$

因为 $\hat{\theta}_n$ 是极大似然估计，所以 $l'(\hat{\theta}_n)=0$. 因此，

$$\hat{\theta}_n-\theta_0\approx-\frac{l'(\theta_0)}{l''(\theta_0)}.$$

简单计算可知

$$l'(\theta_0)=\sum_{i=1}^n\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]\Bigg|_{\theta=\theta_0}=\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i|\theta_0),$$

进而有

$$l''(\theta_0)=\sum_{i=1}^n\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0).$$

因为 $X_1, X_2, \cdots, X_n$ 是独立同分布随机变量，根据弱大数定律，当 $n \to \infty$ 时，

$$\frac{1}{n} l''(\theta_0) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(X|\theta_0)\right] = -I(\theta_0).$$

所以，

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \frac{-\sqrt{n} l'(\theta_0)}{l''(\theta_0)} = \frac{-\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\theta_0)} \approx \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{I(\theta_0)}.$$

根据引理4.2（注意到 $l'(\theta_0)$ 即 Cramér-Rao 不等式证明中构造的随机变量 Z(取 $\theta = \theta_0$)），有

$$\mathbb{E} l'(\theta_0) = 0, \quad \operatorname{Var} l'(\theta_0) = nI(\theta_0).$$

因此，

$$\mathbb{E} \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{I(\theta_0)} = 0, \quad \operatorname{Var} \frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{I(\theta_0)} = \frac{1}{I(\theta_0)}.$$

注意到 $\frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{I(\theta_0)}$ 是独立随机变量之和，由中心极限定理，可知它依分布收敛于均值为 0，方差为 $\frac{1}{I(\theta_0)}$ 的正态分布. 由此可知 $\sqrt{n}(\hat{\theta}_n - \theta_0)$ 依分布收敛到 $N\left(0, \frac{1}{I(\theta_0)}\right)$，换言之 $\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0)$ 依分布收敛于标准正态随机变量. □

由上述结论，我们可以说最大似然估计是渐进无偏估计，且称极限正态分布的方差为最大似然估计的渐进方差. 注意到渐进方差恰好为无偏估计的方差的 Cramér-Rao 的下界，所以，从近似的角度，最大似然估计已经足够好.

例 4.12 设 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$，其中

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{其他}. \end{cases}$$

求 θ 的极大似然估计的极限性质.

解 似然函数为

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}.$$

$$l(\theta) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i.$$

由 $l'(\theta) = 0$ 可得 θ 的极大似然估计

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log x_i}.$$

注意到

$$\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{\partial}{\partial \theta} (\log \theta + (\theta - 1) \log x) = \frac{1}{\theta} + \log x.$$

所以

$$\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = -\frac{1}{\theta^2},$$

$$I(\theta) = -\mathbb{E}\left[-\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right] = \frac{1}{\theta^2}.$$

因此，

$$\frac{\sqrt{n}}{\theta}\left(-\frac{n}{\sum_{i=1}^n \log x_i} - \theta\right) \overset{\text{近似}}{\sim} N(0,1).$$

□

4.6 充分统计量

为简洁起见，我们将用如下记号

$$\mathbf{X} = (X_1, X_2, \cdots, X_n), \quad \mathbf{x} = (x_1, x_2, \cdots, x_n).$$

来分别表示样本和样本点（样本观察值）.

我们用统计量 T，如样本平均，样本方差，最大值、最小值等来规约、提炼数据. 对两个样本 X, Y，如果 $T(X) = T(Y)$，则对统计量 T 而言，这两个样本被视为相同的.

我们希望用于估计参数 θ 的统计量能包含样本足够多的信息以能较好地给出 θ 的估计.

参数 θ 的充分统计量是指在某种意义上提取了样本中有关 θ 的所有信息的统计量. 任何关于样本的其他信息，只要不是充分统计量的值相关的信息，不含有任何参数的更多信息. 具体而言，设 $T(\mathbf{X})$ 是参数 θ 的充分统计量，则关于 θ 的任何推断，仅通过 $T(X)$ 依赖于 $\mathbf{X}$，即若 $\mathbf{x}, \mathbf{y}$ 是两个样本点，如果 $T(\mathbf{x}) = T(\mathbf{y})$，则不论 $\mathbf{x}, \mathbf{y}$ 哪个被观察到，关于 θ 的推断都相同.

定义 4.9. 充分统计量

统计量 $T(\mathbf{X})$ 称为 θ 的充分统计量，如果给定 $T(\mathbf{X})$ 的值的条件下，$\mathbf{X}$ 的分布不依赖于 θ.

这个定义的直观解释是：给定了 T 的值，不能从样本的分布获取 θ 的信息；或说知道了 T，可以不需要样本的信息而不丧失任何信息.

我们先直接计算一个例子.

例 4.13 设 $X_1, X_2, \cdots, X_n \sim \text{Poisson}(\theta)$，证明 $T(\mathbf{X}) := \sum_{i=1}^n X_i$ 是 θ 的充分统计量.

证明 由 Poisson 分布的可加性，可得 $T(\mathbf{X}) = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$. 因此

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n | T(\mathbf{x}) = t) \\
=&\frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, \cdots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i)}{\mathbb{P}(\sum_{i=1}^{n} X_i = t)} \\
=&\frac{\left(\prod_{i=1}^{n-1} \frac{\theta^{x_i} \mathrm{e}^{-\theta}}{x_i!}\right) \frac{\theta^{t-\sum_{i=1}^{n-1} x_i} \mathrm{e}^{-\theta}}{\left(t-\sum_{i=1}^{n-1} x_i\right)!}}{\frac{(n\theta)^t \mathrm{e}^{-n\theta}}{t!}} \\
=&\frac{t!}{\prod_{i=1}^{n} x_i!} \cdot \frac{1}{n^t}.
\end{aligned}$$

所得结果与 θ 无关，因此 $T(\mathbf{X})$ 是充分统计量. □

定理 4.4. 充分统计量的判别

设 $p(\mathbf{x}|\theta)$ 是样本 $\mathbf{X}$ 的联合概率函数，$q(t|\theta)$ 是统计量 $T(X)$ 的概率函数，如果对样本空间中的任何 $\mathbf{x}$，比例

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$

是 θ 的常值函数，则 $T(\mathbf{X})$ 是 θ 的充分统计量.

证明 仅就离散型随机变量的情形证明.

根据条件概率定义，

$$\begin{aligned}
\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) &= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T(\mathbf{X}) = t)} \\
&= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{x}) = t)}{\mathbb{P}_\theta(T(\mathbf{X}) = t)}.
\end{aligned} \tag{4.2}$$

设 t 和 $\mathbf{x}$ 给定.

如果 $T(\mathbf{x}) \neq t$，则

$$\{\mathbf{X} = \mathbf{x}, T(\mathbf{x}) = t\} = \varnothing,$$

从而(4.2)的分子为 0.

如果 $T(\mathbf{x}) = t$，则由于

$$\{\boldsymbol{X} = \boldsymbol{x}\} \subset \{T(\boldsymbol{X}) = T(\boldsymbol{x})\},$$

所以

$$\begin{aligned}
\{\mathbf{X} = \mathbf{x}, T(\mathbf{x}) = t\} &= \{X = x, T(\mathbf{X}) = T(\mathbf{x})\} \\
&= \{X = x\}.
\end{aligned} \tag{4.3}$$

从而(4.2)的分子为

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x}|\theta).$$

在 $T(\mathbf{x})=t$ 时，(4.2)的分母为

$$\mathbb{P}(T(\boldsymbol{X})=t)=\mathbb{P}(T(\boldsymbol{X})=T(\boldsymbol{x}))=q(T(\boldsymbol{x})|\theta).$$

综合可得，对给定的 t 和 $\mathbf{x}$,

$$\mathbb{P}(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t)=\begin{cases}0, & T(\mathbf{x})\neq t,\\ \dfrac{p(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)}, & T(\mathbf{x})=t.\end{cases}$$

所以，如果 $\frac{p(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)}$ 不依赖于 θ，则 $\mathbb{P}(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t)$ 与 θ 无关，从而 $T(\mathbf{X})$ 是充分统计量. □

例 4.14 设样本 $X_1,X_2,\cdots,X_n$ 为来自总体 $X\sim B(1,\theta)$ 的样本，则 $T(\mathbf{X})=\sum_{i=1}^n X_i$ 是 θ 的充分统计量.

证明 设 $T(\mathbf{x})=t$，即 $\sum_{i=1}^n x_i=t$. 显然 $\mathbf{X}$ 的联合概率质量函数为

$$\begin{aligned}p(\mathbf{x}|\theta)&=\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}\\&=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\\&=\theta^t(1-\theta)^{n-t}.\end{aligned}$$

由二项分布的可加性，可得 $T(\mathbf{X})=\sum_{i=1}^n X_i=n\overline{X}\sim B(n,\theta)$. 因此 $T(\mathbf{X})$ 的概率质量函数为

$$q(t|\theta)=\binom{n}{t}\theta^t(1-\theta)^{n-t}.$$

从而

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}=\frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}=\frac{1}{\binom{n}{t}}.$$

由此可见，所得条件概率与 θ 无关，因此 $T(\mathbf{X})$ 是 θ 的充分统计量. □

例 4.15（正态均值的充分统计量） 设总体 $X\sim N(\mu,\sigma^2)$，其中 σ^2 已知，μ 为参数，$X_1,X_2,\cdots,X_n$ 为样本，则 $T(\mathbf{X})=\bar{X}$ 是 μ 的充分统计量.

证明 样本 $\mathbf{X}$ 的联合概率密度函数为

$$\begin{aligned}f(\mathbf{x}|\mu)&=\prod_{i=1}^n(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right)\\&=(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right).\end{aligned}$$

因为 $T(\mathbf{X})=\bar{X}$ 的概率密度函数为

$$q(\bar{x}|\theta)=\left(2\pi\frac{\sigma^2}{n}\right)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\cdot n(\bar{x}-\mu)^2\right).$$

因此

$$\frac{f(\mathbf{x}|\mu)}{q(T\mathbf{x})|\mu} = n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right).$$

它不依赖于参数 μ，因此是 μ 的充分统计量. □

例 4.16 设 X 的概率函数为 f，$X_1, X_2, \cdots, X_n$ 为样本，则样本的联合概率函数为

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)})$$

由此可知次序统计量是充分统计量.

定理 4.5. 因子分解定理

设样本的联合概率函数为 $f(\mathbf{x}|\theta)$，则统计量 $T(\mathbf{X})$ 为 θ 的充分统计量当且仅当存在函数 $g(t|\theta)$，$h(\mathbf{x})$ 使得对任何样本点 $\mathbf{x}$ 以及参数 θ，有如下分解

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)\cdot h(\mathbf{x}), \tag{4.4}$$

其中 $h(x)$ 不依赖于 θ.

证明 仅对离散情形给出证明.

先证必要性.

设 $T(\mathbf{X})$ 是充分统计量，取

$$g(t|\theta) = \mathbb{P}_\theta(T(\mathbf{x}) = t),$$
$$h(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(x)).$$

因为 $T(\mathbf{X})$ 是充分统计量，故 $h(\mathbf{x})$ 不依赖于 θ. 注意到(4.3)，可得

$$\begin{aligned}
f(\mathbf{x}|\theta) &= \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) \\
&= \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) \\
&= \mathbb{P}_\theta(T(\mathbf{X}) = T(\mathbf{x}))\cdot\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \\
&= g(T(\mathbf{x})|\theta)h(\mathbf{x}).
\end{aligned}$$

证充分性.

设分解(4.4)成立，$q(t|\theta)$ 为 $T(\mathbf{X})$ 的概率质量函数. 为证明 $T(\mathbf{X})$ 的充分统计量，要证明

$$\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$$

与 θ 无关. 记 A_t 为样本空间中使得 $T(\mathbf{y})$ 取值为 t 的样本点 $\mathbf{y}$ 的集合，即

$$A_t := \{y \in \mathscr{X} : T(\mathbf{y}) = t\}.$$

因此

$$A_{T(\mathbf{x})} := \{y \in \mathscr{X} : T(\mathbf{y}) = T(\mathbf{x})\}.$$

则

$$\begin{aligned}
q(T(\mathbf{x})|\theta) &= \mathbb{P}_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\
&= \mathbb{P}_\theta\left(\bigcup_{\mathbf{y}\in\mathscr{X}} \{\mathbf{X} = \mathbf{y}, T(\mathbf{X}) = T(\mathbf{x})\}\right) \\
&= \mathbb{P}_\theta\left(\bigcup_{\mathbf{y}\in A_{T(\mathbf{x})}} \{\mathbf{X} = \mathbf{y}, T(\mathbf{X}) = T(\mathbf{y})\}\right) \\
&= \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} \mathbb{P}_\theta\left(\mathbf{X} = \mathbf{y}\right) \\
&= \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} f(\mathbf{y}|\theta) \\
&= \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} g(T(\mathbf{y}|\theta))h(\mathbf{y}) \\
&= \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} g(T(\mathbf{x}|\theta))h(\mathbf{y}) \\
&= g(T(\mathbf{x}|\theta)) \sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y}).
\end{aligned}$$

因此

$$\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} = \frac{g(T(\mathbf{x})|\theta)\cdot h(\mathbf{x})}{g(T(\mathbf{x}|\theta))\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{T(\mathbf{x})}} h(\mathbf{y})}$$

与参数 θ 无关，从而 $T(\mathbf{X})$ 是充分统计量. □

例 4.17 续例4.14. 设 $\sum_{i=1}^n x_i = t$，则样本的联合概率函数可以分解为

$$\begin{aligned}
f(x|\theta) &= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\
&= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}(1-\theta)^n \\
&= \left(\frac{\theta}{1-\theta}\right)^t(1-\theta)^n.
\end{aligned}$$

取 $g(t,\theta) = \left(\frac{\theta}{1-\theta}\right)^t(1-\theta)^n, h(\mathbf{x}) = 1$，由分解定理，可见 $\sum_{i=1}^n X_i$ 是充分统计量.

为应用因子分解定理，我们需要将样本的联合概率函数分解为两部分，一部分为 $g(T(\mathbf{x})|\theta)$，一部分为 $h(\mathbf{x})$，前者包含未知参数，且仅通过统计量依赖于样本，后者与参数无关，仅与样本有关.

例 4.18 设样本来自正态总体，未知均值，方差已知. 则 $\mathbf{X}$ 的联合概率密度函数

为

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]\right) \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right). \end{aligned}$$

取

$$h(\mathbf{x}) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right),$$

它仅依赖于样本 $\mathbf{x}$，取

$$g(T(\mathbf{x})|\theta) = \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right),$$

仅通过统计量依赖于样本，其中

$$g(t|\theta) = \exp\left(-\frac{n}{2\sigma^2}(t-\mu)^2\right).$$

由此可见统计量 $\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2$ 是充分统计量.

例 4.19（离散均匀分布） 设 $X_1, X_2, \cdots, X_n$ 为来自 $\{1, 2, \cdots, \theta\}$ 上的离散均匀分布，θ 为未知参数. 总体的概率质量函数为

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & x = 1, 2, \cdots, \theta, \\ 0, & \text{其他}. \end{cases}$$

利用示性函数，可以将 $f(x|\theta)$ 表示如下：

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta}\mathbb{1}_{\{1,2,\cdots,\theta\}}(x)\cdots\mathbb{1}_{\{1,2,\cdots\}}(x) \\ &= \frac{1}{\theta}\mathbb{1}_{\{x,x+1,\cdots\}}(\theta)\cdots\mathbb{1}_{\{1,2,\cdots\}}(x). \end{aligned}$$

因此，样本的联合概率质量函数为

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{1}_{\{x_i,x_i+1,\cdots\}}(\theta)\cdots\prod_{i=1}^{n}\mathbb{1}_{\{1,2,\cdots\}}(x_i) \\ &= \frac{1}{\theta^n}\mathbb{1}_{\{x_{(n)},x_{(n)}+1,\cdots\}}(\theta)\cdots\prod_{i=1}^{n}\mathbb{1}_{\{1,2,\cdots\}}(x_i) \\ &= g(x_{(n)}|\theta)h(\mathbf{x}), \end{aligned}$$

其中

$$g(t|\theta) = \frac{1}{\theta^n}\mathbb{1}_{t,t+1,\cdots\}}(\theta) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]}(t),$$

$$h(\mathbf{x}) = \prod_{i=1}^{n}\mathbb{1}_{\{1,2,\cdots\}}(x_i),$$

由此可见，

$$T(\mathbf{X}) = X_{(n)}$$

是充分统计量.

充分统计量与极大似然估计：

设 $T(\mathbf{X})$ 为充分统计量，则似然函数为

$$L(\theta) = f(\theta|\mathbf{x}) = g(T(\mathbf{X})|\theta))h(\mathbf{x}).$$

作为 θ 的函数，极大似然函数仅通过 $g(T(\mathbf{X})|\theta))$ 依赖于 θ，因此，关于 θ 的极大似然估计就是使得

$$g(T(\mathbf{X})|\theta)$$

达到最大值的 θ 的取值 $\hat{\theta}$，显然，它是 $T(\mathbf{x})$ 的函数.

充分统计量与贝叶斯估计：

后验分布正比于似然函数与先验分布的乘积，因此，对任何观测值 $\mathbf{x}$，只要 $T(\mathbf{x})$ 相同，则后验分布就相同. 即 $T(\mathbf{x})$ 包含了所需要的用来做估计的信息. 这也正好说明充分统计量的作用.

4.7 指数分布族

定义 4.10. 单参数指数分布族

称分布族 $f(x|\theta)$ 为单参数指数分布族，如果

$$f(x|\theta) = \begin{cases} \exp\left[c(\theta)T(x) + d(\theta) + s(x)\right], & x \in A, \\ 0, & x \notin A. \end{cases}$$

其中 A 与 θ 无关.

设 $X_1, X_2, \cdots, X_n$ 为样本，则

$$f(\mathbf{x}|\theta)] = \prod_{i=1}^{n}\exp\left[c(\theta)T(x_i) + d(\theta) + s(x_i)\right]$$

$$= \exp\left[c(\theta)\sum_{i=1}^{n}T(x_i) + d(\theta)\right]\exp\left[\sum_{i=1}^{n}s(x_i)\right].$$

由此可见，$\sum_{i=1}^n T(x_i)$ 是充分统计量.

例 4.20（Bernoulli 分布） 设总体 $X \sim B(1,\theta)$，则其概率函数为

$$\begin{aligned}\mathbb{P}(X=x) &= \theta^x(1-\theta)^{1-x} \\ &= \exp\left[x\log\frac{\theta}{1-\theta}+\log(1-\theta)\right]\end{aligned}$$

为指数分布族，$T(x)=x$，因此 $\sum_{i=1}^n X_i$ 为充分统计量.

定义 4.11. 多参数指数分布族

称 $f(\boldsymbol{x}|\theta)$ 为 k 参数指数分布族，如果

$$f(\boldsymbol{x}|\theta)=\begin{cases}\exp\left[\sum_{j=1}^k c_j(\theta)T_j(\boldsymbol{x})+d(\theta)+s(\boldsymbol{x})\right], & \boldsymbol{x}\in A,\\ 0, & \boldsymbol{x}\notin A,\end{cases}$$

其中 A 与参数无关.

4.8 Rao-Blackwell 定理

定理 4.6. Rao-Blackwell 定理

设 $\hat\theta$ 为 θ 的估计，$\mathbb{E}\hat\theta^2<\infty$. 设 T 为 θ 的充分统计量，$\tilde\theta=\mathbb{E}(\hat\theta|T)$，则

$$\mathbb{E}_\theta(\tilde\theta-\theta)^2\le\mathbb{E}_\theta(\hat\theta-\theta)^2,$$

且等号成立当且仅当 $\hat\theta=\tilde\theta$. 特别，如果 $\hat\theta$ 是 θ 的无偏估计，则 $\tilde\theta$ 也是 θ 的无偏估计，且

$$\operatorname{Var}_\theta\tilde\theta\le\operatorname{Var}_\theta\hat\theta.$$

证明 由命题4.1，

$$\begin{aligned}\mathbb{E}(\tilde\theta-\theta)^2&=\operatorname{Var}\tilde\theta+(\mathbb{E}\tilde\theta-\theta)^2,\\ \mathbb{E}(\hat\theta-\theta)^2&=\operatorname{Var}\hat\theta+(\mathbb{E}\hat\theta-\theta)^2,\end{aligned}$$

以及

$$\mathbb{E}\tilde\theta=\mathbb{E}[\mathbb{E}(\hat\theta|T)]=\mathbb{E}\hat\theta,$$

可知我们只要证明

$$\operatorname{Var}\tilde\theta\le\operatorname{Var}\hat\theta.$$

由条件方差等式

$$\begin{aligned}\operatorname{Var}\hat\theta&=\mathbb{E}\operatorname{Var}(\hat\theta|T)+\operatorname{Var}(\mathbb{E}(\hat\theta|T))\\ &=\mathbb{E}\operatorname{Var}(\hat\theta|T)+\operatorname{Var}\tilde\theta..\end{aligned}$$

注意到

$$\operatorname{Var}(\hat{\theta}|T) = \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}|T))^2|T) \geq 0,$$

故 $\mathbb{E}\operatorname{Var}(\hat{\theta}|T) \geq 0$，从而有

$$\operatorname{Var}\tilde{\theta} \leq \operatorname{Var}\hat{\theta}.$$

且等号成立当且仅当 $\mathbb{E}\operatorname{Var}(\hat{\theta}|T) = 0$. 它成立当且仅当

$$0 = \operatorname{Var}(\hat{\theta}|T) = \mathbb{E}((\hat{\theta} - \mathbb{E}(\hat{\theta}|T))^2|T),$$

因条件方差是非负的，它等于零也等价于它的期望为零：

$$E\operatorname{Var}(\hat{\theta}|T) = \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}|T))^2 = 0,$$

此即

$$\hat{\theta} = \mathbb{E}(\hat{\theta}|T) = \tilde{\theta}.$$

由于 $\tilde{\theta}$ 是一个估计，不依赖于 θ，而 T 是充分统计量，也不依赖于 θ，从而 $\tilde{\theta}$ 是估计. □

下面的例子表明，对不是充分统计量的随机变量取条件期望，得不到估计.

例 4.21 对不充分统计量取条件期望

设 X_1, X_2 为来自 $N(\mu, 1)$ 的样本，$\bar{X} = \frac{1}{2}(X_1 + X_2)$，则 $\bar{X}$ 是充分统计量，且

$$\mathbb{E}_\mu \bar{X} = \mu, \quad \operatorname{Var}_\mu \bar{X} = \frac{1}{2}.$$

即 $\bar{X}$ 是参数 μ 的无偏估计.

我们考虑

$$\phi(X_1) = \mathbb{E}_\mu(\bar{X}|X_1).$$

易知 X_1 不是充分统计量，我们考虑直接用条件期望的性质得：

$$\mathbb{E}_\mu\phi(X_1) = \mathbb{E}_\mu[\mathbb{E}(\bar{X}|X_1)] = \mathbb{E}\bar{X} = \mu,$$

根据 Rao-Blackwell 定理中的证明过程，可得

$$\operatorname{Var}_\mu \phi(X_1) \leq \operatorname{Var}_\mu \bar{X},$$

由此看起来，$\phi(X_1)$ 是比 $\bar{X}$ 方差更小的无偏估计？事实上，$\phi(X_1)$ 依赖于参数 μ，不是估计：

$$\begin{aligned}\phi(X_1) &= \mathbb{E}_\mu(\bar{X}|X_1) = \mathbb{E}_\mu(\frac{1}{2}(X_1 + X_2)|X_1)\\ &= \mathbb{E}_\mu(\frac{1}{2}X_1 + \frac{1}{2}X_2|X_1) = \frac{1}{2}X_1 + \frac{1}{2}\mathbb{E}_\mu X_2 = \frac{1}{2}X_1 + \frac{1}{2}\mu.\end{aligned}$$

注 根据 $\phi(X_1)$，可以直接计算它的期望与方差如下：

$$\mathbb{E}_\mu\phi(X_1) = \mathbb{E}_\mu(\frac{1}{2}X_1 + \frac{1}{2}\mu) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu,$$

$$\operatorname{Var}\phi(X_1)=\operatorname{Var}(\frac{1}{2}X_1+\frac{1}{2}\mu)=\operatorname{Var}\frac{1}{2}X_1=\frac{1}{4}.$$

4.9 完备统计量

定义 4.12. 完备统计量

设 T 为一个统计量，若对任何满足

$$\mathbb{E}(l(T(X))=0,\quad \theta\in\Theta$$

的 $l(T)$，都有

$$\mathbb{P}(l(T(X))=0)=1,\quad \theta\in\Theta,$$

则称 T 是一个完备统计量.

例 4.22 设 $X_i\overset{\text{iid}}{\sim}U(0,\theta)$，$\Theta=(0,+\infty)$. 证明 $X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$ 是完备的.

证明 $X_{(n)}$ 的概率密度函数为

$$f(t|\theta)=\begin{cases} n\dfrac{t^{n-1}}{\theta^n}, & 0<t<\theta,\\ 0, & \text{其他}. \end{cases}$$

或将 $X_{(n)}$ 的概率密度函数写成 $f(x|\theta)=\theta^{-n}nx^{n-1}\mathbb{1}_{(0,\theta)}$. 设 l 满足

$$\mathbb{E}_\theta l(X_{(n)})=0,\quad \theta>0,$$

可知

$$\mathbb{E}l(X_{(n)})=\int_0^\theta l(x)\theta^{-n}nx^{n-1}dx=0.$$

所以

$$\int_0^\theta l(x)x^{n-1}dx=0.$$

求导可知对任意 $\theta>0$，$l(\theta)\theta^{n-1}=0$. 由此可知对任意 $\theta>0$，$l(\theta)=0$. □

定理 4.7. Lehmann-Scheffe 定理

设 T 为参数 θ 的完备充分统计量，$\hat{\theta}$ 为 θ 的无偏估计，则 $\tilde{\theta}=\mathbb{E}(\hat{\theta}|T)$ 为 θ 的唯一 UMVUE.

证明 设 $\hat{\theta}_1$ 是 θ 的另一个无偏估计，则

$$\tilde{\theta}_1=\mathbb{E}(\hat{\theta}_1|T)$$

也是 θ 的无偏估计，且有 Rap-Blackwell 定理，

$$\operatorname{Var}\tilde{\theta}_1 \le \operatorname{Var}\hat{\theta}_1. \tag{4.5}$$

由条件期望的性质，$\tilde{\theta}, \tilde{\theta}_1$ 都是 T 的函数，且都是 θ 的无偏估计，从而 $\tilde{\theta} - \tilde{\theta}_1$ 也是 T 的函数，且有

$$\mathbb{E}(\tilde{\theta} - \tilde{\theta}_1) = 0.$$

因为 T 是完备的，所以

$$\mathbb{P}(\tilde{\theta} - \tilde{\theta}_1) = 1.$$

由此可知，无偏估计 $\tilde{\theta}$ 与任取的无偏估计 $\hat{\theta}_1$ 相比，

$$\operatorname{Var}\tilde{\theta} = \operatorname{Var}\tilde{\theta}_1 \le \operatorname{Var}\hat{\theta}_1.$$

因此，$\tilde{\theta}$ 是一致方差最小的无偏估计.

另一方面，对任取的无偏估计 $\hat{\theta}$，因不等式(4.5)，只要将 $\tilde{\theta}$ 与 $\tilde{\theta}_1$ 比较. 但此时 $\tilde{\theta}, \tilde{\theta}_1$ 几乎处处相等，所以 $\tilde{\theta}_1$ 是唯一的.

□

例 4.23 设 $X_1, X_2, \cdots, X_n$ 为来自 $B(1,\theta)$ 的样本，$\theta \in (0,1)$，证明 $T = \sum_{i=1}^n X_i$ 为 θ 的完备充分统计量，并求 θ 的 UMVUE.

解 设 $t = \sum_{i=1}^n x_i$，则样本联合分布为

$$f(x_1, x_2, \cdots, x_n | \theta) = \theta^t (1-\theta)^{n-t}.$$

由分解定理易知 T 是充分的. 往证完备性. 设 $\mathbb{E}l(T) = 0$，注意到 $T \sim B(n,\theta)$，可知

$$\mathbb{E}l(T) = \sum_{t=0}^n l(t)\binom{n}{t}\theta^t(1-\theta)^{n-t}.$$

设 $u = \frac{\theta}{1-\theta}$，可得

$$\sum_{t=0}^n l(t)\binom{n}{t}y^t = 0.$$

上式对任意 $y > 0$ 都成立，因此

$$l(t)\binom{n}{t} = 0, \quad t = 0, 1, 2, \cdots, n.$$

这表明 $l(T) \equiv 0$，因此 T 是完备的.

显然 $\overline{X} = \frac{1}{n}T$ 是 θ 的无偏估计，又因为 $\overline{X}$ 是 T 的函数，所以

$$\mathbb{E}(\overline{X}|T) = \overline{X}.$$

根据 Lehman-Scheffe 定理，$\overline{X}$ 是 θ 的唯一的 UMVUE. □

内容提要

- 点估计
- 极大似然估计
- 无偏估计
- 有效性
- Cramér-Rao 不等式
- 一致最小方差无偏估计
- 充分统计量
- 完备统计量

第 4 章 习题

1. 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu, \sigma^2)$ 的一个样本.
   (a). 试求适当选择的常数 c 使得 $c\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2$ 是 σ^2 的无偏估计;
   (b). 求 k 使得 $k\sum_{i=1}^{n}|X_i - \bar{X}|$ 为 σ 的无偏估计.
2. 设 $\hat{\theta}$ 是参数 θ 的无偏估计，且 $\mathrm{Var}\,\hat{\theta} > 0$，试证明 $\hat{\theta}^2 = (\hat{\theta})^2$ 不是 θ 的无偏估计.
3. 设总体 X 的数学期望 $\mu = \mathbb{E}X$ 已知，试证统计量 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ 是总体方差 $\sigma^2 = \mathrm{Var}\,X$ 的无偏估计.
4. 设 X 为离散型随机变量，概率质量函数为
$$\mathbb{P}(X = 1) = \theta, \quad \mathbb{P}(X = 2) = 1 - \theta.$$
   三个独立样本观测值为 $x_1 = 1, x_2 = 2, x_3 = 1$. 如果 Θ 的先验分布为 $[0, 1]$ 上的均匀分布，求其后验密度.
5. 设总体的概率密度函数如下
$$f(x|\theta) = \frac{1}{2}(1 + \theta x), \quad -1 < x < 1, -1 < \theta < 1,$$
   $X_1, X_2, \cdots, X_n$ 为样本，求 θ 的矩估计，并证明它是弱相容的.
6. 设样本 $X_1, X_2, \cdots, X_n$ 来自正态总体 $N(\theta, \theta)$，其中 $\theta > 0$ 为参数.
   (a). 求 θ 的极大似然估计 $\hat{\theta}$.
   (b). 求 $\hat{\theta}$ 的渐进方差.
7. 设 $X_1, X_2, \cdots, X_n$ 为来自几何分布的样本，用因子分解定理证明 $T = \sum_{i=1}^{n} X_i$ 是充分统计量.
8. 设 X 为来自正态总体 $N(0, \sigma^2)$ 的样本. $|X|$ 是否为充分统计量?
9. 设 $X_1, X_2, \cdots, X_n$ 为来自总体概率密度函数如下
$$f(x|\mu, \sigma) = \frac{1}{\sigma}\exp\left(-\frac{x - \mu}{\sigma}\right), \quad \mu < x < \infty, 0 < \sigma < \infty$$
   的样本. 求 (μ, σ) 的二维充分统计量.
10. 证明二项分布属于指数分布族.
11. 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma^2)$ 的样本，其中 μ 未知，σ^2 已知. 求 μ^3 的 UMVUE.

12. 设 $X_1, X_2, \cdots, X_n$ 是来自参数为 $\lambda > 0$ 的 Poisson 分布. 设 $\overline{X}, S^2$ 分别为样本均值与样本方差.
   (a). 求证 $\overline{X}$ 是 λ 的 UMVUE.
   (b). 证明 $\mathbb{E}(S^2|\overline{X}) = \overline{X}$.
   (c). 证明 $\operatorname{Var} S^2 > \operatorname{Var} \overline{X}$.

第 5 章 区间估计

设总体 X 有某参数 θ 未知，$X_1, X_2, \ldots, X_n$ 为样本. 参数的点估计是用一个数值估计参数，区间估计用一个范围（区间）以及一个概率来共同叙述这个估计.

定义 5.1. 区间估计

设 $\hat{\theta}_1, \hat{\theta}_2$ 为 θ 的估计量，且

$$\mathbb{P}(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) = 1 - \alpha,$$

其中 $0 < \alpha < 1$，则称 $(\hat{\theta}_1, \hat{\theta}_2)$ 为 θ 的置信水平（置信度）为 $1 - \alpha$ 的置信区间（Confidence Interval，缩写为 CI）.

有时，只要

$$\mathbb{P}(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) \ge 1 - \alpha,$$

就称 $(\hat{\theta}_1, \hat{\theta}_2)$ 为 θ 的置信水平（置信度）为 $1 - \alpha$ 的置信区间. 按此定义，如果 $(\hat{\theta}_1, \hat{\theta}_2)$ 是置信水平为 $1 - \alpha$ 的置信区间，则对任何 $\alpha' > \alpha$，$(\hat{\theta}_1, \hat{\theta}_2)$ 也是置信水平为 $1 - \alpha'$ 的置信区间. 此时，称最大的置信水平为置信系数.

显然，置信区间是一个随机区间，区间的长短反映估计的精度，置信度的高低反映估计的可靠性. 精度和可靠性一般不可兼得. 人们往往在同一置信度下寻求精度尽量高（长度最短）的置信区间. 未知参数 θ 不一定恰估计的区间中. 置信水平的理解如下：在 100 次抽样得到 100 个区间中，平均有 $100(1 - \alpha)$ 个区间包含 θ.

得到参数的区间估计的一般步骤如下：

1. 求枢轴量 $W(\mathbf{X}, \theta)$，即找到一个包含未知参数 θ，但其分布与未知参数无关的统计量.
2. 对给定的置信水平，求出常数 a, b，使得
$$\mathbb{P}(a < W(X_1, X_2, \cdots, X_n; \theta) < b) = 1 - \alpha.$$
3. 计算置信区间：通过等价变换得到
$$\mathbb{P}(\theta_1 < \theta < \theta_2) = 1 - \alpha.$$

5.1 正态总体的均值的区间估计

设 $X \sim N(\mu, \sigma^2)$，$X_1, X_2, \ldots, X_n$ 为来自该总体的样本.

5.1.1 方差已知

假设方差 σ^2 已知. 此时，

$$Z := \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

因此有 $u_{\alpha/2} > 0$ 使得

$$\mathbb{P}(|Z| \leq z_{\alpha/2}) = 1 - \alpha.$$

显然，$z_{\alpha/2}$ 是 Z 的 $\alpha/2$ 上侧分位数. 如果 $\alpha = 0.05$，则 $z_{0.025} = 1.96$. 由此可知

$$\mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

因此，

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

这表明

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$

是均值 μ 的置信度为 $1 - \alpha$ 的置信区间.

有时，我们也用

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

表示如上对称置信区间.

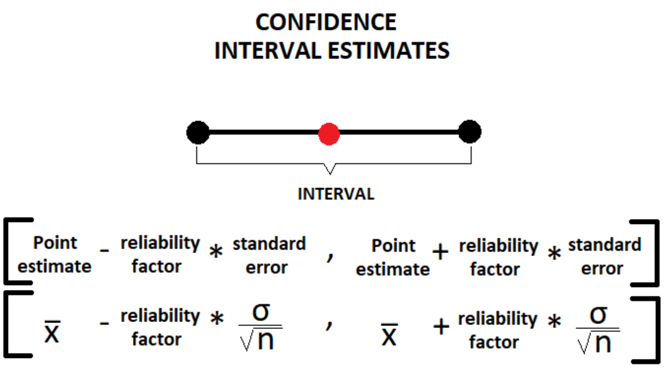


图 5.1 点估计与区间估计的关系

这个区间是随机区间. μ 不一定恰在这个区间中. 但在 100 次抽样得到 100 个区间中，平均有 $100(1-\alpha)$ 个区间包含 μ.

区间长度与样本量有关，样本量越大，估计越精确.

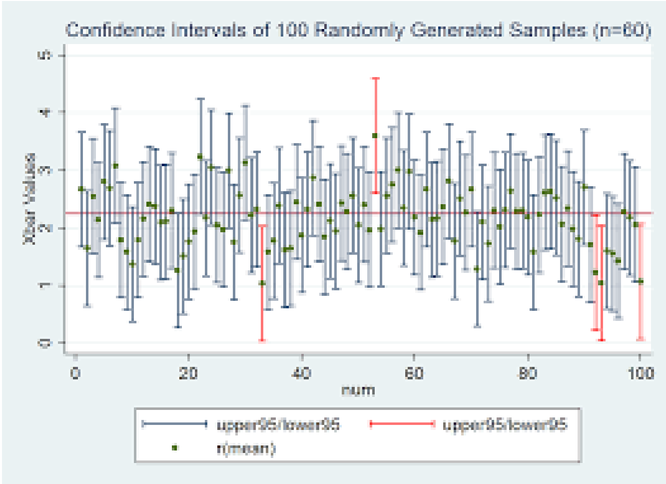


图 5.2　置信度的意义

5.1.2 方差未知

假设方差 σ^2 未知. 此时，我们用样本标准差 S 代替 Z 中的 σ，可得

$$T := \frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t(n-1).$$

因此有 $t_{\alpha/2}(n-1) > 0$（如无歧义，简记为 $t_{\alpha/2}$）使得

$$\mathbb{P}(|T| \leq t_{\alpha/2}) = 1-\alpha.$$

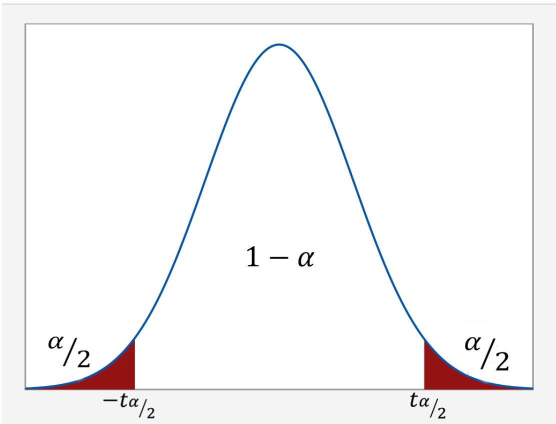


图 5.3　t 分布的分位数

显然，$t_{\alpha/2}$ 是 T 的 $\alpha/2$ 上侧分位数. 如果 $\alpha = 0.05$，$n = 10$，则 $t_{0.025} = ***$. 由此可知

$$\mathbb{P}(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1-\alpha.$$

因此，

$$\mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}}t_{\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{\alpha/2}\right) = 1-\alpha.$$

这表明

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}t_{\alpha/2},\overline{X}+\frac{\sigma}{\sqrt{n}}t_{\alpha/2}\right)$$

是均值 μ 的置信度为 $1-\alpha$ 的置信区间.

一般，当 α 较小时，$t_{\alpha/2}>z_{\alpha/2}$，所以，在相同可靠度条件下，方差未知的时候，均值的区间估计比已知时的区间估计范围更广（不精确）.

但当 $n\to\infty$ 时，差别逐渐缩小.

5.2 方差的区间估计

5.2.1 均值未知

根据

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

可知

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1)\leq\frac{(n-1)S^2}{\sigma^2}\leq\chi^2_{\alpha/2}(n-1)\right)=1-\alpha.$$

因此，

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\leq\sigma^2\leq\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)=1-\alpha.$$

所以，可得 σ^2 的置信水平为 $1-\alpha$ 的区间估计为

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

5.2.2 均值已知

因为对任意 $i=1,2,\cdots,n$，

$$\frac{X_i-\mu}{\sigma}\sim N(0,1)$$

所以，

$$\frac{(X_i-\mu)^2}{\sigma^2}\sim\chi^2(1)$$

再有 χ^2 分布的可加性，可得

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

因此,

$$\mathbb{P}\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\alpha/2}(n)} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right) = 1-\alpha,$$

所以,

$$\left[\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right]$$

是 σ^2 的置信水平为 $1-\alpha$ 的置信区间.

比较下面的做法?

根据

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$$

可得

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(1) \le \frac{n(\overline{X}-\mu)^2}{\sigma^2} \le \chi^2_{\alpha/2}(1)\right) = 1-\alpha.$$

因此,

$$\mathbb{P}\left(\frac{n(\overline{X}-\mu)^2}{\chi^2_{\alpha/2}(1)} \le \sigma^2 \le \frac{n(\overline{X}-\mu)^2}{\chi^2_{1-\alpha/2}(1)}\right) = 1-\alpha.$$

所以,可得 σ^2 的置信水平为 $1-\alpha$ 的区间估计为

$$\left[\frac{n(\overline{X}-\mu)^2}{\chi^2_{\alpha/2}(1)}, \frac{n(\overline{X}-\mu)^2}{\chi^2_{1-\alpha/2}(1)}\right].$$

例 5.1 假设某人测量了 16 个面包,它们的重量分别为

994 977 975 988 1000 982 971 983 999 985 991 951 1010 973 982 1020.

设每个面包的重量服从正态分布 $N(\mu,\sigma^2)$.

1. 求均值 μ 的置信水平为 0.95 的置信区间;
2. 求方差 σ^2 的置信水平为 0.95 的置信区间.

Excel 计算文件为 CI-bread.xls

解 1. 样本量为 $n=16$,样本均值为 $\bar{x}=986.313$,样本方差为 $s^2=270.096$,标准样本方差为 $s=16.436$,$t_{0.025}(15)=2.490$. 于是,μ 的置信水平为 95% 的置信区间为

$$986.313 \pm 2.132 \times \frac{16.435}{\sqrt{4}}$$

区间估计为 $[976.082, 996.543]$.

因为 $(n-1)s^2 = 4051.438$,分位点为 $\chi^2_{0.025}(15) = 27.488$,$\chi^2_{0.975}(15) = 6.262$.

所以 σ^2 的区间估计为

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right] = [147.387, 646.974].$$

□

5.3 两总体情形

设有两个相互独立的总体 $X \sim N(\mu_1, \sigma_1^2)$,$Y \sim N(\mu_2, \sigma_2^2)$,各有样本 $X_1, X_2, \cdots, X_{n_1} \sim X$，$Y_1, Y_2, \cdots, Y_{n_2} \sim Y$，样本均值分别为 $\overline{X}, \overline{Y}$，样本方差分别为 S_1^2, S_2^2.

5.3.1 均值差 $\mu_1 - \mu_2$ 的区间估计

方差 σ_1^2, σ_2^2 已知的情形

根据定理2.3，注意到

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

由此可得

$$\left((\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

为 $\mu_1 - \mu_2$ 的置信水平为 $1-\alpha$ 的置信区间.

方差相等 $\sigma_1^2 = \sigma_2^2$ 但未知的情形

根据定理2.3，注意到

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中

$$S_W^2 := \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

我们可得 $\mu_1 - \mu_2$ 的一个置信水平为 $1-\alpha$ 的置信区间

$$\left[(\overline{X}-\overline{Y}) - t_{\alpha/2}S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\overline{X}-\overline{Y}) + t_{\alpha/2}S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

5.3.2 方差之比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的估计

根据定理2.3，

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1,n_2-1),$$

所以

$$\mathbb{P}\left(F_{1-\alpha/2}(n_1-1,n_2-1) < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{\alpha/2}(n_1-1,n_2-1)\right) = 1-\alpha.$$

因此

$$\mathbb{P}\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right) = 1-\alpha.$$

由此可得

$$\left[\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right]$$

为 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信水平为 $1-\alpha$ 的置信区间.

5.4 指数分布的参数的区间估计

例 5.2 设 $X_1, X_2, \cdots, X_n$ 为来自参数为 θ 的指数分布的样本，求 θ 的置信水平为 $100(1-\alpha)\%$ 的置信区间.

解

因为 X_i 服从参数为 θ 的指数分布，所以它的尾概率为

$$\mathbb{P}(X_i > x) = \mathrm{e}^{-\theta x}, \quad x > 0.$$

从而对任意 $x > 0$，

$$\begin{aligned}\mathbb{P}(2\theta X_i > x) &= \mathbb{P}(X_i > \frac{x}{2\theta}) \\ &= \mathrm{e}^{-\theta\cdot\frac{x}{2\theta}} = \mathrm{e}^{-\frac{x}{2}}.\end{aligned}$$

它表明 $2\theta X_i$ 服从参数为 $\frac{1}{2}$ 的指数分布. 而参数为 $\frac{1}{2}$ 的指数分布恰为自由度为 2 的卡方分布. 所以

$$2\theta X_i \sim \chi^2(2).$$

由 χ^2 分布的可加性可知

$$2\theta n\overline{X} = 2\theta\sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

于是，

$$\mathbb{P}(\chi^2_{1-\alpha/2}(2n) \le 2\theta n\overline{X} \le \chi^2_{\alpha/2}(2n)) = 1-\alpha.$$

它等价于

$$\mathbb{P}\left(\frac{\chi^2_{1-\alpha/2}(2n)}{2n\overline{X}} \le \theta \le \frac{\chi^2_{\alpha/2}(2n)}{2n\overline{X}}\right) = 1-\alpha.$$

这表明，

$$\left[\frac{\chi^2_{1-\alpha/2}(2n)}{2n\overline{X}}, \frac{\chi^2_{\alpha/2}(2n)}{2n\overline{X}}\right]$$

是 θ 的置信水平为 $1-\alpha$ 的置信区间. □

注关于 $2\theta n\overline{X}$ 的分布，也可如下推理. 参数为 θ 的指数分布即参数为 $1,\theta$ 的 Γ 分布. 由 Γ 分布的可加性，

$$n\overline{X} = \sum_{i=1}^{n} X_i \sim \mathrm{Gamma}(n,\theta).$$

再由 Γ 分布的尺度变换性质，

$$2\theta n\overline{X} \sim \mathrm{Gamma}\left(n, \frac{1}{2}\right).$$

由此可见

$$2\theta n\overline{X} \sim \chi^2(2n).$$

5.5 用近似分布做区间估计

例 5.3（0-1 分布） 设样本 $X_i \sim B(1,\theta)$，则 $\sum_{i=1}^n X_i \sim B(n,\theta)$，由 De Moivre-Laplace 中心极限定理，

$$\frac{n\overline{X}-n\theta}{\sqrt{n\theta(1-\theta)}} \overset{\text{近似}}{\sim} N(0,1).$$

由此可得

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le \overline{X} + z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}\right) = 1-\alpha.$$

用 $\overline{X}$ 近似代替 θ，可得 θ 的置信水平为 $1-\alpha$ 的近似置信区间为

$$\left[\overline{X} - z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}, \overline{X} + z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right].$$

一般地，设总体的均值为 μ，则由中心极限定理，

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \overset{\text{近似}}{\sim} N(0,1).$$

由此可得

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \le z_{\alpha/2}\right) = 1-\alpha.$$

如果方差 σ^2 为均值 μ 的函数 $\sigma^2 = \sigma^2(\mu)$，则我们可能能从上式获得 μ 的置信区

间. 但其中的方程可能较难求解. 形式上，μ 的置信水平为 $1-\alpha$ 的近似“置信区间”

$$\left[\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

未知参数 σ 可用其参数估计 $\hat{\sigma}$ 代替，从而可得 μ 的置信水平为 $1-\alpha$ 的近似“置信区间”

$$\left[\overline{X}-z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}},\overline{X}+z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right].$$

设 $X_1,X_2,\cdots,X_n$ 为总体 $X\sim f(x|\theta)$ 样本，$\hat{\theta}_n$ 为参数 θ 的极大似然估计，则由定理4.3，

$$\sqrt{nI(\theta)}(\hat{\theta}_n-\theta)\xrightarrow{L}N(0,1).$$

因此，

$$\mathbb{P}(|\sqrt{nI(\theta)}(\hat{\theta}_n-\theta)|<z_{\alpha/2})\approx 1-\alpha.$$

由此可得 θ 的一个置信水平为 $1-\alpha$ 的近似置信区间

$$\left[\hat{\theta}_n-z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}},\hat{\theta}_n+z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}}\right]$$

例 5.4 设总体服从参数为 θ 的指数分布，则概率函数为

$$f(x|\theta)=\theta\,\mathrm{e}^{-\theta x},\quad x\in(0,+\infty).$$

因此，似然函数为

$$L(\theta)=\prod_{i=1}^{n}\theta\,\mathrm{e}^{-\theta x_i}=\theta^n\,\mathrm{e}^{-\sum_{i=1}^{n}x_i}.$$

对数似然函数为

$$l(\theta)=\log L(\theta)=n\log\theta-\sum_{i=1}^{n}x_i.$$

由 $l'(\theta)=\frac{n}{\theta}-n\bar{x}$ 可得参数 θ 的极大似然估计

$$\hat{\theta}_n=\frac{1}{\overline{X}}.$$

参数 θ 的 Fisher 信息为

$$I(\theta)=-\mathbb{E}\left(\frac{\partial}{\partial\theta^2}\log f(X|\theta)\right)=-\mathbb{E}\left(\frac{\partial}{\partial\theta^2}\log\theta\,\mathrm{e}^{-\theta X}\right)=\frac{1}{\theta^2}.$$

因此，

$$\frac{\sqrt{n}}{\theta}\left(\frac{1}{\overline{X}}-\theta\right)\overset{\text{近似}}{\sim}N(0,1).$$

于是我们有

$$\mathbb{P}(-z_{\alpha/2}<\frac{\sqrt{n}}{\theta}\left(\frac{1}{\overline{X}}-\theta\right)<z_{\alpha/2})\approx 1-\alpha.$$

由此可得

$$\mathbb{P}\left(\frac{1}{\overline{X}}(1-\frac{1}{\sqrt{n}}z_{\alpha/2}) \leq \theta \leq \frac{1}{\overline{X}}(1+\frac{1}{\sqrt{n}}z_{\alpha/2})\right) \approx 1-\alpha.$$

例 5.5 设样本 $X_1, X_2, \cdots, X_n \sim \text{Poisson}(\lambda)$. 根据例3.11，参数 λ 的极大似然估计为 $\hat{\lambda} = \overline{X}$. 由 Poisson 分布的可加性，

$$n\hat{\lambda} = n\overline{X} \sim \text{Poisson}(n\lambda).$$

由此可见，$n\hat{\lambda}$ 的分布依赖于 λ，不是枢轴量. 精确的置信区间需要费一些功夫才能得到（参 Pearson and Hartley 1966）.

根据例4.7，参数 λ 的 Fiser 信息为

$$I(\lambda) = \frac{1}{\lambda}.$$

所以，λ 的一个置信水平为 $1-\alpha$ 的近似置信区间为

$$\mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{\frac{n}{\lambda}}(\overline{X}-\lambda) \leq z_{\alpha/2}\right) = 1-\alpha.$$

用 λ 的极大似然估计 $\overline{X}$ 代替 $\sqrt{\frac{n}{\lambda}}$ 中的 λ，可得 λ 的一个置信水平为 $1-\alpha$ 的近似置信区间

$$\left[\overline{X} - z_{\alpha/2}\cdot\sqrt{\frac{\overline{X}}{n}}, \overline{X} + z_{\alpha/2}\cdot\sqrt{\frac{\overline{X}}{n}}\right]$$

5.6 中位数的区间估计

本小节考虑中位数及其区间估计. 一般地可以考虑一般分位数的区间估计.

设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本. 我满假设总体的分布未知，否则，例如，对正态分布，中位数与均值相等，中位数的区间估计等价于均值的区间估计.

为简单起见，我们假设 X 是连续型随机变量. 否则中位数的定义稍微复杂些.

定义 5.2. 中位数

如果 η 满足

$$\mathbb{P}(X \geq \eta) = \mathbb{P}(X \leq \eta) = \frac{1}{2},$$

则称之为 X 的中位数（median）.

类似于期望使得均方误差 $\mathbb{E}(X-c)^2$ 最小的常数 c，中位数是使得绝对误差 $E|X-c|$ 最小的常数 c.

样本中位数的定义依赖于样本量.

定义 5.3. 样本中位数

样本中位数 M 定义为

$$M = \begin{cases} X_{\frac{1}{2}(n+1)}, & n\text{是奇数}, \\ \frac{1}{2}[X_{(\frac{1}{2}n)} + X_{(\frac{1}{2}n+1)}], & n\text{是偶数}. \end{cases}$$

样本中位数可以自然视为总体中位数的估计. 很自然地可以考虑用次序统计量来构造总体中位数的区间估计. 中位数的区间估计形如

$$(X_{(s)}, X_{(t)}).$$

实际上，我们常考虑对称区间：

$$(X_{(k)}, X_{(n-k+1)}).$$

这个区间的观察值为

$$(x_{(k)}, x_{(n-k+1)}),$$

在 n 个样本观察值 $x_1, x_2, \cdots, x_n$ 中，有 $k-1$ 个观察值 $x_{(1)}, x_{(2)}, \cdots, x_{(k-1)}$ 位于区间的左边，$n-(n-k+1) = k-1$ 个观察值 $x_{(n-k+2)}, \cdots, x_n$ 在区间的右边. 即这个区间的构造有对称性：去掉的极端（极大、极小）观察值数目相同.

注意到

$$\mathbb{P}(X_i < \eta) = \mathbb{P}(X_i > \eta) = \frac{1}{2},$$

故 $X_1, X_2, \cdots, X_n$ 中小于中位数 η 的个数 Y 服从二项分布：$Y \sim B(n, \frac{1}{2})$. 因此

$$\mathbb{P}(Y = k) = 2^{-n}\binom{n}{k}.$$

从而累积分布和生存函数分别为：

$$\mathbb{P}(Y \le r) = \mathbb{P}(Y < r+1) = 2^{-n}\sum_{i=0}^{r}\binom{n}{i},$$

$$\mathbb{P}(Y \ge r) = \mathbb{P}(Y > r-1) = 2^{-n}\sum_{i=r}^{n}\binom{n}{i}.$$

再注意到 $\eta < X_{(s)}$ 表示样本 $X_1, X_2, \cdots, X_n$ 中至多有 $s-1$ 个观察值小于 η，因此

$$\mathbb{P}(\eta < X_{(s)}) = \mathbb{P}(Y \le s-1) = 2^{-n}\sum_{i=0}^{s-1}\binom{n}{i}.$$

类似地，$\eta > X_{(t)}$ 表示有样本 $X_1, X_2, \cdots, X_n$ 中至少有 t 个观察值小于 η，因此

$$\mathbb{P}(\eta > X_{(t)}) = \mathbb{P}(Y \ge t) = 2^{-n}\sum_{i=t}^{n}\binom{n}{i} = \sum_{j=0}^{n-t} 2^{-n}\binom{n}{n-j} = \sum_{j=0}^{n-t} 2^{-n}\binom{n}{j}.$$

从而

$$\mathbb{P}(X_{(s)} \le \eta \le X_{(t)}) = 1 - \mathbb{P}(\eta < X_{(s)}) - \mathbb{P}(\eta > X_{(t)}) = 1 - \sum_{i=s}^{t-1} 2^{-n}\binom{n}{i}.$$

特别，考虑 $s = k, t = n - k + 1$ 的对称情形，则注意到

$$\mathbb{P}(\eta < X_k) = \mathbb{P}(Y \le k-1) = 2^{-n}\sum_{i=0}^{k-1}\binom{n}{i},$$

$$\mathbb{P}(\eta > X_{n-k+1}) = \mathbb{P}(Y \ge n-k+1) = 2^{-n}\sum_{i=0}^{k-1}\binom{n}{i},$$

由此可见

$$\mathbb{P}(\eta < X_k) = \mathbb{P}(\eta > X_{n-k+1}) = 2^{-n}\sum_{i=0}^{k-1}\binom{n}{i},$$

从而

$$\mathbb{P}(X_{(k)} \le \eta \le X_{(n-k+1)}) = 1 - \mathbb{P}(\eta < X_{(k)}) - \mathbb{P}(\eta > X_{(n-k+1)}) = 1 - 2\sum_{i=0}^{k-1} 2^{-n}\binom{n}{i}.$$

例 5.6 参考文件 Hampson-Walker1961-Platinum.xlsx：有铂热升华温度的 26 个试验数据，$Y \sim B(26, \frac{1}{2})$，由于 $\mathbb{P}(Y \le 7) = 0.0145$，可知 $[X_{(8)}, X_{(19)}]$ 是中位数的置信水平为 $97\% \approx 1 - 2 \times 0.0145$ 的置信区间.

内容提要

- 区间估计
- 置信水平

附录

第 5 章 习题

1. 下面的 16 个数来自计算机的正态随机数生成器[1]

5.3885	3.2387	0.2195	0.3231	−3.6310	−0.3892	1.7771	1.4031
−1.5357	−1.6683	4.9765	2.0393	−0.5556	−0.5957	−0.9547	1.0123

(a). 你认为这个正态分布的均值 μ 和方差 σ^2 是多少?

(b). 给出 μ 的 90%, 95%, 99% 的置信区间.

(c). 给出 σ^2 的 90%, 95%, 99% 的置信区间.

[1]本例使用 WPS 的电子表格，命令为 = norminv(rand(), μ, σ)).

(d). 给出 σ 的 90%, 95%, 99% 的置信区间.

(e). 如果将 μ 的置信区间长度减半，样本需要增加多少?

2. 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，μ, σ 未知. 常数 c 应取多少时才能保证区间 $(-\infty, \bar{X}+c]$ 是 μ 的 95% 的置信区间，即求 c 使得

$$\mathbb{P}(-\infty < \mu \leq \bar{X}+c) = 0.95.$$

3. 设 $L(x), U(x)$ 满足

$$\mathbb{P}(L(\mathbf{X}) \leq \theta) = 1-\alpha_1, \quad \mathbb{P}(U(\mathbf{X}) \geq \theta) = 1-\alpha_2,$$

且对任意 $\mathbf{x}$，

$$L(\mathbf{x}) \leq U(\mathbf{x}).$$

证明

$$\mathbb{P}(L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})) = 1-\alpha_1-\alpha_2.$$

第 6 章 假设检验

假设检验是推断统计的另一重要内容. 检验问题是对总体分布的检验. 可以是对已知总体未知参数的检验，也可以是非参数检验.

6.1 从女士品茶谈起

1935 年，费歇尔在他的《试验设计》(The Design of Experiments) 第 2 章中，描述名为“女士品茶”的试验.

一位女士宣称她能品尝出奶茶是先加奶后加茶还是先加茶后加奶. 费歇尔为检验女士是否有超能力而设计了一个试验. 随机地给这位女士 8 杯茶品尝，其中 4 杯奶茶是先加奶，另外 4 杯奶茶是后加奶. 这里就有一个基本的假设检验问题. 假设这位女士没有超能力，则她对任意一杯奶茶能作出正确判断的概率为 $\frac{1}{2}$. 若用 X 表示 b8 杯奶茶中，她能正确判断的奶茶杯数，则 X 取值 $x \in \{0, 1, 2, \cdots, 8\}$ 的概率

$$\mathbb{P}(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^8,$$

以及取值不小于 x 的累积概率

$$\mathbb{P}(X \geq x) = \sum_{k=x+1}^{n}\binom{n}{x}\left(\frac{1}{2}\right)^8$$

如下表所示.

$X = x$	0	1	2	3	4	5	6	7	8
$\mathbb{P}(X = x)$	0.0039	0.0312	0.1093	0.2187	0.2734	0.2187	0.1093	0.0312	0.0039
$\mathbb{P}(X \geq x)$	1	0.9960	0.9648	0.8554	0.6367	0.3632	0.1445	0.0351	0.0039

表 6.1 正确判断的概率

可见，正确判断 7 杯以及 7 杯以上的概率为 0.0351. 这是一个较小的概率. 也就是说，如果这位女士能正确判断出 7 杯或 8 杯奶茶的制作方法，则可以认为她是随机判断的假设不成立. 用这种方法做推断，可能犯错. 这位女士没有超能力却被误认为有超能力的概率为 0.0351.

也可以考虑相关的问题：这位女士没有超能力，即正确判断一杯奶茶的概率为 $\frac{1}{2}$，或这位女士能正确判断一杯奶茶的概率为某个确定的数值. 为方便起见，我们用硬币投掷来替代女士品茶问你题.

假设投掷一枚硬盘正面朝上的概率为 θ，$\theta = 0.5$ 或 $\theta = 0.8$. 投掷 8 次，记录正面朝上的次数 X. 希望通过正面朝上的次数来推断 θ 的取值.

将 $\theta = 0.5$ 或 $\theta = 0.7$ 分别称为假设 H_0 和 H_1. 在这两种假设下，正面朝上的次数 $X = x$ 的概率分别为

$$\mathbb{P}(X = x|H_0) = \binom{8}{x} 0.5^8, \quad \mathbb{P}(X = x|H_1) = \binom{8}{x} 0.7^x 0.3^{1-x}.$$

概率 $\mathbb{P}(X = x|H_0)\mathbb{P}(X = x|H_1)$ 也可以理解为似然函数 $\mathbb{P}(H_0|x), \mathbb{P}(H_1|x)|$. 这两个似然函数分别表示观察到 $X = x$ 的条件下，假设 H_0 或 H_1 为真的可能性. 于是在不同假设下，这两个概率之比为

$$\frac{\mathbb{P}(X = x|H_0)}{\mathbb{P}(X = x|H_1)} = \frac{0.5^8}{0.7^x 0.3^{1-x}}.$$

称之可理解为似然函数之比 $\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}$，简称似然比.

将似然函数 $\mathbb{P}(X = x|H_0), \mathbb{P}(X = x|H_1)$ 以及似然比 $\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}$ 列表如下.

X=x	0	1	2	3	4	5	6	7	8
$\mathbb{P}(H_0\|x)$	0.00390	0.03125	0.10937	0.21875	0.27343	0.21875	0.10937	0.03125	0.00390
$\mathbb{P}(H_1\|x)$	0.00006	0.00122	0.01000	0.04667	0.13613	0.25412	0.29647	0.19765	0.05764
$\frac{\mathbb{P}(H_0\|x)}{\mathbb{P}(H_1\|x)}$	59.53	25.51	10.93	4.68	2.00	0.86	0.36	0.15	0.06

表 6.2 投掷硬币的似然函数与似然比

由表可见，似然比随着 x 的增大而减小. 似然比越大，越倾向于认为 H_0 为真. 可以设置一个临界值 c，若似然比小于这个值，则认为 H_1 为真；若似然比大于这个值，则认为 H_0 为真. 具体取什么值，需要综合考虑犯错的概率是否可以接受.

若设 $c = 0.5$，则当 $X \geq 6$ 时，拒绝 H_0，认为 H_1 为真；否则，若 $X < 6$，则接受 H_0，拒绝 H_1. 此时可能犯两类错误，第一类错误是 H_0 为真却拒绝了它，相应的概率为

$$\mathbb{P}(X \geq 6|H_0) = 0.14453.$$

第二类错误是 H_1 为真却接受了 H_0，相应的概率为

$$\mathbb{P}(X < 6|H_1) = 0.44822.$$

若设 $c = 0.3$，则当 $X \geq 7$ 时，拒绝 H_0，认为 H_1 为真；否则，若 $X < 7$，则接受 H_0，拒绝 H_1. 此时犯犯第一类错误的概率为 H_0 为真却拒绝了它的概率

$$\mathbb{P}(X \geq 7|H_0) = 0.03515.$$

犯第二类错误的概率为 H_1 为真却接受了 H_0 的概率

$$\mathbb{P}(X < 7|H_1) = 0.74470167.$$

可见，一般来说，若第一类错误较小，则第二类错误较大.

6.2 基本概念

定义 6.1. 统计假设

一个统计假设是关于总体分布的一个论断. 如果某统计假设能够完全确定总体的分布，则称这个假设为简单假设；否则称该假设为复合假设.

一个假设检验问题是指判断一对假设哪一个为真的问题. 我们称其中一个假设为零假设，又称原假设)；另一个假设为备择假设，或称对立假设，分别记为 H_0, H_1.

设总体的分布为 P，属于分布族 $\mathscr{P}$，则可将 H_0, H_1 表示为

$$H_0\colon P \in \mathscr{P}_0 \quad \leftrightarrow \quad H_1\colon P \in \mathscr{P}_1,$$

其中 $\mathscr{P}_0, \mathscr{P}_1 \subset \mathscr{P}$，且 $\mathscr{P}_0, \mathscr{P}_1$ 不相交.

如果总体的分布由其参数 θ 决定，则统计假设可归结为关于总体参数的假设，原假设和备择假设可分别表示为

$$H_0\colon \theta \in \Theta_0 \quad \leftrightarrow \quad H_1\colon \theta \in \Theta_1,$$

其中 $\Theta_0, \Theta_1 \subset \Theta$，且 Θ_0, Θ_1 不相交，Θ 为参数空间.

一个检验是指基于总体的样本 $\boldsymbol{X}$ 做出一个决定，给出一个拒绝域或临界域. 拒绝域 C 是指样本所有可能取值 $\mathscr{X}$ 的一个子集. C 的余集称为接受域. 如果样本 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n) \in C$，则接受原假设，拒绝原假设；否则拒绝原假设，接受备择假设.

有时人们将"接受 H_0"表述为"不拒绝 H_0"，这里有细微的差别. 前者表示检验人愿意相信 H_0 是对的，后者表示检验人内心不相信 H_0，只是没有足够的证据来否定这个假设.

我们也可以用统计决策理论来表述上面介绍的假设检验.

统计决策是指根据样本采取的行动. 用 A 表示可采取的行动的集合. 判决准则（判决函数）是由样本空间 $\mathscr{X}$ 到样本空间的一个函数.

具体到假设检验，只有两个行动：拒绝 H_0 或接受 H_0，分别用 1 和 0 来表示这两个行动. 行动空间为 $A = \{0,1\}$，而判决准则 $d(\cdot)$ 满足：对任意样本 $\boldsymbol{X}$，若 $d(\boldsymbol{x}) = 1$，则拒绝 H_0；若 $d(\boldsymbol{x}) = 1$，则接受 H_0. 显然

$$d(\boldsymbol{x}) = \mathbb{1}_C(\boldsymbol{X}).$$

样本观察值 $\boldsymbol{X}$ 是样本 $\boldsymbol{X}$ 的实现，从而有

$$d(\boldsymbol{x}) = \mathbb{1}_C(\boldsymbol{X}).$$

因为样本 $\boldsymbol{X}$ 是随机的，一个检验可能会出现两类错误：弃真和采伪（纳伪）. 我们称弃真为第一类错误，即 H_0 为真，但被拒绝了. 它发生的概率为

$$\begin{aligned}\alpha(P) &= \mathbb{P}(\text{发生第一类错误})\\ &= \mathbb{P}(\text{拒绝}H_0|H_0\text{为真})\\ &= \mathbb{P}(\boldsymbol{X} \in C|H_0\text{为真})\\ &= \mathbb{E}(d(\boldsymbol{x})|H_0\text{为真})..\end{aligned}$$

往后将它简记为

$$\alpha(P) = \mathbb{P}(\boldsymbol{X} \in C|H_0) = \mathbb{E}(d(\boldsymbol{x})|H_0)..$$

其他情形类似.

我们称采伪，即接受错误的假设为第二类错误，它发生的概率为

$$\begin{aligned}\beta(P) &= \mathbb{P}(\text{发生第二类错误})\\ &= \mathbb{P}(\text{接受}H_0|H_1)\\ &= \mathbb{P}(\boldsymbol{X} \in C^{\mathsf{C}}|H_1)\\ &= \mathbb{E}(\mathbb{1}_{C^{\mathsf{c}}}(\boldsymbol{X})|H_1).\end{aligned}$$

称 $1 - \beta(P)$ 为检验的势.

检验的功效函数（也称势函数）定义为

$$p(P) = \mathbb{P}(\text{拒绝}H_0|P) = \mathbb{P}(\boldsymbol{X} \in C|P) = \mathbb{E}(\mathbb{1}_C|P).$$

显然，若 $P \in \mathscr{P}_0$，则 $p(P) = \alpha(P)$，为犯第一类错误的概率；而若 $P \in \mathscr{P}_1$，则 $p(P) = 1 - \beta(P)$，为犯第二类错误的概率.

若总体的分布 P 由其参数 $\theta \in \Theta$ 确定，假设检验问题为 $H_0\colon \theta \in \Theta_0 \leftrightarrow H_1\colon \theta \in$

Θ_1，则分别记 $\alpha(\theta), \beta(\theta)$ 为犯第一类、第二类错误的概率，即

$$\alpha(\theta) = \mathbb{P}(\boldsymbol{X} \in C | \theta \in \Theta_0),$$
$$\beta(\theta) = \mathbb{P}(\boldsymbol{X} \in C^{\mathsf{c}} | \theta \in \Theta_1).$$

此时检验的势函数为

$$p(\theta) = \mathbb{P}(\boldsymbol{X} \in C|\theta) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0, \\ 1 - \beta(\theta), & \theta \in \Theta_1. \end{cases}$$

	接受 H_0	拒绝 H_0
H_0 为真	正确决策	第一类错误
H_1 为真	第一类错误	正确决策

表 6.3　决策及结果

理想的决策（检验）自然是没有错误的检验. 但因为样本的随机性，两类错误是客观存在不可避免的. 犯两类错误的概率同时最小化的检验也不存在. 有时可以通过加大样本量来将两类错误都减小. 现实中，发生第一类错误的后果更严重（这也是选择哪个假设为零假设的一个依据），在样本量固定的情形，一般会要求发生第一类错误的概率不超过 α，再最小化第二类错误，即使得

$$\sup_{P \in \mathscr{P}_0} \alpha(P) \le \alpha, \tag{6.1}$$

同时最小化 $\beta(P), P \in \mathscr{P}_0$.

称(6.1)中的 α 为显著性水平，称(6.1)的左边，即

$$\alpha^* := \sup_{P \in \mathscr{P}_0} \alpha(P)$$

为检验水准.

另一个常用的概念是 p 值. p 值是基于样本观察数据对原假设的真实性的一个判断，是拒绝原假设的最小的显著性水平，或说，在原假设为真时，等于或更极端于实际观察结果的概率.

显著性水平的选取是主观的，但习惯上常取 $5\%, 1\%$ 等值.

例 6.1 设总体服从正态分布 $N(\mu, \sigma^2)$，且 σ^2 已知. 考虑假设检验

$$H_0: \mu = \mu_0 \quad \leftrightarrow \quad H_1: \mu = \mu_1.$$

设样本均值为 $\bar{x}$，临界值为 c，拒绝域为

$$W = \{\boldsymbol{x} \in \mathscr{X} : \bar{x} > c\}.$$

假设 $\mu_0 < \mu_1$. 下图形象显示了 α, β 以及 p 值.

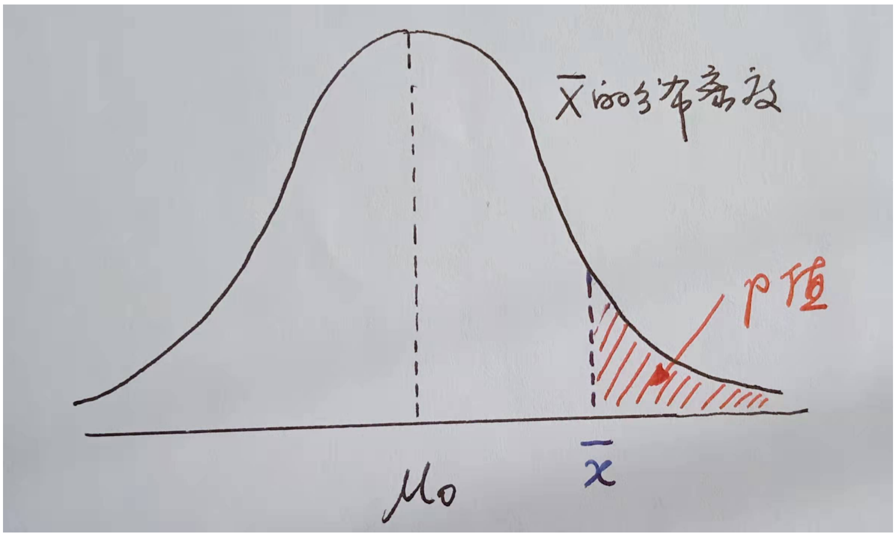


图 6.1　p 值

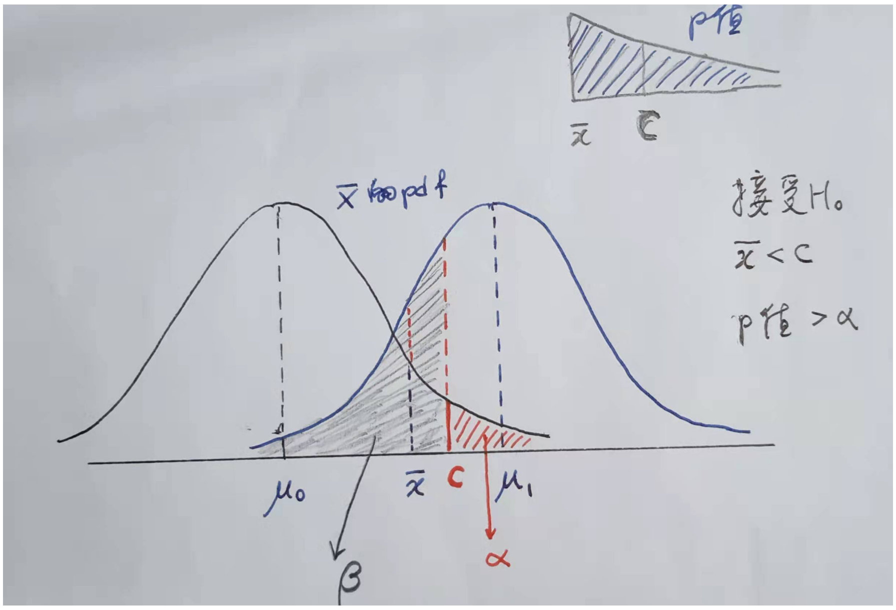


图 6.2　p 值与 α, β

例 6.2 投掷一枚硬币，设正面朝上的概率为 θ. 用 X 表示投掷一次，正面朝上的次数，则 $X \sim B(1, \theta)$. 为检验该硬币正面朝上的概率是 θ_0 还是 θ_1（设 $\theta_0 < \theta_1$），提出假设检验问题：

$$H_0 : \theta = 0.5 \leftrightarrow H_1 : \theta = 0.6.$$

为此，投掷该硬币 n 次，样本为 $X_1, X_2, \cdots, X_n$，样本观察值为 $x_1, x_2, \cdots, x_n$. 设拒绝域为

$$C = \left\{ \boldsymbol{X} \in \mathscr{X} : \sum_{i=1}^{n} x_i > N \right\}.$$

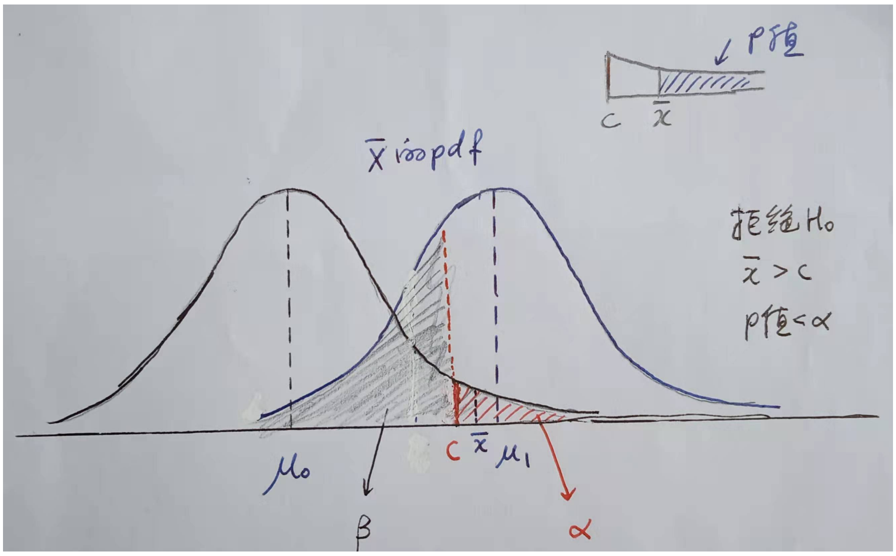


图 6.3 p 值与 α,β

注意到 $\sum_{i=1}^{n} \sim B(n,\theta)$，则发生第一类错误的概率为

$$\alpha(\theta_0)=\mathbb{P}\left(\sum_{i=1}^{n}X_i>N|\theta_0\right)=\sum_{x=N+1}^{n}\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x},$$

发生第二类错误的概率为

$$\beta(\theta_1)=\mathbb{P}\left(\sum_{i=1}^{n}X_i\leq N|\theta_1\right)=\sum_{x=0}^{N}\binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}.$$

例 6.3 设总体 $X\sim N(\mu,\sigma_0^2)$，σ_0^2 已知. 检验问题：$H_0\colon\mu=\mu_0\leftrightarrow H_1\colon\mu=\mu_1$.

6.3 似然比检验之初见

设有一枚硬币，正面朝上的概率 θ 可能为 θ_0 或 θ_1，假设 $\theta_0>\theta_1$.（它可能是如下情况：有两枚硬币，正面朝上的概率分别为 θ_0 或 θ_1，从这两枚硬币中随机选择一枚，希望通过试验来判断这枚硬币是那一枚.）为检验该硬币正面朝上的概率为 θ_0 还是 θ_1，我们随机投掷该硬币 n 次.

我们用统计学语言描述该问题. 现有总体 $X\sim B(1,\theta)$. 考虑检验问题

$$H_0\colon\theta=\theta_0\leftrightarrow H_1\colon\theta=\theta_1.$$

设 $X_1,X_2,\cdots,X_n$ 为总体 X 的简单随机样本. 求显著性水平为 α 的检验.

我们先用直观方法考虑下该问题. 正面朝上的次数为 $T\sim B(n,\theta)$，因此

$$\mathbb{P}(T=t|\theta)=\binom{n}{t}\theta^t(1-\theta)^{n-t},\quad t=0,1,2,\cdots,n.$$

设试验观察到正面朝上的总次数为 $T=t$，则直观的想法是，如果在 $\theta=\theta_0$

时 $T=t$ 的概率远小于 $\theta=\theta_1$ 时 $T=t$ 的概率，也就是如果硬币正面朝上的可能性很小，则我们应该认为硬币正面朝上的概率偏小，即拒绝原假设. 考虑

$$\frac{\mathbb{P}(T=t|\theta_0)}{\mathbb{P}(T=t|\theta_1)} \propto \left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)^t.$$

注意到由 $\theta_0>\theta_1$ 有

$$\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} = \frac{\theta_0-\theta_0\theta_1}{\theta_1-\theta_0\theta_1} > 1$$

由此可知，存在常数 c_1 使不等式

$$\left(\frac{\mathbb{P}(T=t|\theta_0)}{\mathbb{P}(T=t|\theta_1)}\right)^t \le c_1$$

成立等价于存在常数 c_2 使得

$$t \le c_2.$$

设检验的显著性水平为 α，则拒绝域为

$$C=\{t:t\le c\}$$

其中 c 使得

$$\mathbb{P}(T\le c)=\sum_{i=0}^{c}\binom{n}{i}\theta_0^i(1-\theta_0)^{n-i}=\alpha.$$

下面用似然函数来处理我们的问题. 我们将看到它与上述直观是相同的.

设样本值为 $x_1,x_2,\cdots,x_n$. 用 $f(x|\theta)$ 表示总体 X 的概率函数：

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

则似然函数为

$$\begin{aligned} L(\theta) &= L(\theta|x_1,x_2,\cdots,x_n)\\ &= \prod_{i=1}^{n} f(x_i|\theta)=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

回忆，似然函数表示在获得观察值的条件下，参数 θ 的后验“分布”.

对我们的检验问题 $H_0{:}\,\theta=\theta_0 \leftrightarrow H_1{:}\,\theta=\theta_1$，直观上，拒绝域中的样本值 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 应使似然比

$$\Lambda(\boldsymbol{x}) := \frac{L(\theta_0|x_1,x_2,\cdots,x_n)}{L(\theta_1|x_1,x_2,\cdots,x_n)} < \lambda,$$

其中 λ 满足

$$\mathbb{P}(\Lambda(\boldsymbol{X})<\lambda)=\alpha,$$

因为（利用此前的计算经验）

$$\Lambda(\boldsymbol{x})=\left(\frac{\theta_0-\theta_0\theta_1}{\theta_1-\theta_0\theta_1}\right)^{\sum_{i=1}^n x_i},$$

所以，拒绝域的形式为

$$C := \{\boldsymbol{x} \in \mathscr{X} : \sum_{i=1}^{n} x_i \le c\},$$

其中

$$c = \frac{\log \lambda - n \log \frac{1-\theta_1}{1-\theta_0}}{\log \frac{\theta_0 - \theta_0\theta_1}{\theta_1 - \theta_0\theta_1}}$$

满足

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \le c\right) = \alpha.$$

一般地，设总体的概率函数（概率密度或概率质量函数）为 $f(x|\theta)$，其中 θ 为参数，可能为向量. 似然函数定义为

$$L(\theta|x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

我们也常将其简记为

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta).$$

定义 6.2. 似然比检验

设有检验问题 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta = \theta_1$，

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})}.$$

似然比检验（Likelihood Ratio Test，简记为 LRT）是拒绝域形如

$$\{\boldsymbol{x} \in \mathscr{X} : \lambda(\boldsymbol{x}) \le c\}$$

的检验，其中 $0 \le c \le 1$.

常数 c 的选择与显著性水平有关：

$$\mathbb{P}(\lambda(X) \le c) = \alpha.$$

注 如果用贝叶斯的观点，将参数 θ 当做随机变量，则我们要考虑的似然比应为如下量

$$f(\theta_0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta_0)f(\theta_0)}{f(\boldsymbol{x})}$$

和

$$f(\theta_1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta_1)f(\theta_1)}{f(\boldsymbol{x})}$$

的比：

$$\frac{f(\theta_0|\boldsymbol{x})}{f(\theta_1|\boldsymbol{x})} = \frac{f(\theta_0)}{f(\theta_1)} \cdot \frac{L(\theta_0)}{L(\theta_1)}.$$

由此可见，拒绝域除与似然比有关，与参数的先验概率之比也有关.

6.4 Neyman-Pearson 引理

定理 6.1. Neyman-Pearson 引理

设 H_0, H_1 为简单假设. 有如下似然比检验：当似然比小于或等于 c 时，拒绝 H_0，显著性水平为 α，则任何其他显著性水平不大于 α 的检验的势不大于似然比检验的势，即似然比检验是最优势检验.

证明 设 f 为总体的概率函数，检验问题为 $H_0\colon f(x) = f_0(x) \leftrightarrow H_1\colon f(x) = f_1(x)$. 似然比检验的拒绝域为

$$C = \{\boldsymbol{x} \in \mathscr{X} : f_0(\boldsymbol{x}) \le c f_1(\boldsymbol{x}).\}$$

其中

$$f_0(\boldsymbol{x}) = \prod_{i=1}^{n} f_0(x_i), \quad f_1(\boldsymbol{x}) = \prod_{i=1}^{n} f_1(x_i).$$

似然比检验的决策函数为 $d(\boldsymbol{x}) = \mathbb{1}_C(\boldsymbol{x})$. 于是，似然比检验的显著性水平为

$$\alpha = \mathbb{E}_0(d(\boldsymbol{X})) = \mathbb{E}(d(\boldsymbol{X})|H_0),$$

检验的势为

$$\mathbb{E}_1 d(\boldsymbol{X}) := \mathbb{E}(d(\boldsymbol{X})|H_1).$$

设 d^* 为其他检验对应的判决函数. 它的显著性水平为

$$\mathbb{E}_0(d^*(\boldsymbol{X})) := \mathbb{E}(d^*(\boldsymbol{X})|H_0)$$

由假设，该检验的显著性水平不超过 α，即不超过似然比检验的显著性水平，所以

$$\mathbb{E}_0(d^*(\boldsymbol{X})) \le \mathbb{E}_0(d(\boldsymbol{X})). \tag{6.2}$$

该检验的势为 $\mathbb{E}_1(d^*(\boldsymbol{X}))$. 要证明该检验的势不大于似然比检验的势，即证明

$$\mathbb{E}_1(d^*(\boldsymbol{X})) \le \mathbb{E}_1(d(\boldsymbol{X})). \tag{6.3}$$

为此，要用到关键不等式

$$d^*(\boldsymbol{X})[c f_1(\boldsymbol{x}) - f_0(\boldsymbol{x})] \le d(\boldsymbol{x})[c f_1(\boldsymbol{x}) - f_0(\boldsymbol{x})]. \tag{6.4}$$

在关键不等式(6.4)的两边同时关于 $\boldsymbol{x}$ 积分（或求和，根据 f 是概率密度或概率质量函数而定），可得

$$c\mathbb{E}_1 d^*(\boldsymbol{X}) - \mathbb{E}_0 d^*(\boldsymbol{X}) \le c\mathbb{E}_1(d(\boldsymbol{X})) - \mathbb{E}_0 d(\boldsymbol{X}).$$

从而有

$$\mathbb{E}_0 d(\boldsymbol{X}) - \mathbb{E}_0 d^*(\boldsymbol{X}) \le c[\mathbb{E}_1 d(\boldsymbol{X}) - \mathbb{E}_1 d^*(\boldsymbol{X})].$$

由(6.2)，上式左边非负，故右边也非负，从而可得(6.3)，这就证明了所求的结论.

下面证明(6.4).

如果 $d(\boldsymbol{x})=1$，则 $\boldsymbol{x}$ 在拒绝域 C 中，从而 $f_0(\boldsymbol{x})\le cf_1(\boldsymbol{x})$. 从而(6.4)等价于 $d^*(\boldsymbol{x})\le d(\boldsymbol{x})$. 注意到 $d(\boldsymbol{x})=1$，$d^*(\boldsymbol{x})=1$ 或 0，它自然成立.

如果 $d(\boldsymbol{x})=0$，则 $cf_1(\boldsymbol{x})-f_0(\boldsymbol{x})<0$. 因此，注意到 $d^*(\boldsymbol{x})$ 取值 0 或 1，所以不等式(6.4)的左边非正，而其右边为零，从而不等式成立.

□

例 6.4 设样本 $X_1,X_2,\cdots,X_n$ 来自总体 $N(\mu,\sigma_0^2)$，σ_0^2 已知. 有假设检验问题 $H_0\colon\mu=\mu_0\leftrightarrow H_1\colon\mu=\mu_1$，其中 μ_0,μ_1 为给定常数. 求显著性水平为 α 的检验.

考虑似然比

$$\begin{aligned}\Lambda&=\frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})}\\&=\frac{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\mu_0)^2\right)}{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(x_i-\mu_1)^2\right)}\\&=\exp\left(\frac{1}{2\sigma_0^2}(2n\bar{x}(\mu_0-\mu_1)+n(\mu_1^2-\mu_0^2).)\right)\end{aligned}$$

较小的似然比拒绝原假设是最大势检验. 当 $\mu_0<\mu_1$ 时，较小的似然比等价于 $\bar{x}$ 较小；当 $\mu_0>\mu_1$ 时，较小的似然比等价于 $\bar{x}$ 较大.

考虑 $\mu_0<\mu_1$ 的情形. 拒绝域的形式为

$$\{\boldsymbol{x}\in\mathscr{X}:\bar{x}>x_0\},$$

其中 x_0 满足

$$\mathbb{P}(\overline{X}>x_0|\mu=\mu_0)=\alpha.$$

因为

$$\alpha=\mathbb{P}\left(\overline{X}>x_0\middle|\,\mu=\mu_0\right)=\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0}>\frac{\sqrt{n}(x_0-\mu_0)}{\sigma_0}\middle|\,\mu=\mu_0\right)$$

且注意到在 H_0 成立的条件下（即 $\mu=\mu_0$），

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0}\sim N(0,1)$$

可知

$$\frac{\sqrt{n}(x_0-\mu_0)}{\sigma_0}=z_\alpha.$$

由此可得

$$x_0=\mu_0+z_\alpha\frac{\sigma_0}{\sqrt{n}}.$$

例 6.5 设样本来自正态总体 $N(\mu_0,\sigma^2)$，其中 μ_0 已知. 考虑检验问题

$$H_0\colon \sigma^2=\sigma_0^2 \leftrightarrow H_1\colon \sigma^2=\sigma_1^2,$$

其中 $\sigma_1>\sigma_0$. 求显著性水平为 α 的检验.

解 注意到 $\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}<0$，易知对某常数 c_1 有似然比

$$\frac{L(\sigma_0^2|\boldsymbol{x})}{L(\sigma_1^2|\boldsymbol{x})}=\frac{\sigma_1^2}{\sigma_0^2}\exp\left[\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)\sum_{i=1}^n(x_i-\mu_0)^2\right]\le c_1$$

等价于有某常数 c_2 使得

$$\sum_{i=1}^n(x_i-\mu_0)^2\ge c_2$$

因为当 $\sigma^2=\sigma_0^2$ 时，

$$\sum_{i=1}^n\frac{(X_i-\mu_0)^2}{\sigma_0^2}\sim\chi^2(n)$$

可知

$$\mathbb{P}\left(\sum_{i=1}^n\frac{(X_i-\mu_0)^2}{\sigma_0^2}>\chi_\alpha^2(n)\right)=\alpha.$$

因此，由 N-P 引理，

$$C=\left\{\boldsymbol{x}\in\mathscr{X}\colon\sum_{i=1}^n(x_i-\mu_0)^2>\sigma_0^2\chi_\alpha^2(n)\right\}$$

是检验水平为 α 的最优势检验.

□

6.5 广义似然比检验

对检验问题

$$H_0\colon\theta=\theta_0\leftrightarrow H_1\colon\theta=\theta_1,$$

似然比为

$$\lambda(\boldsymbol{x})=\frac{L(\theta_0|\boldsymbol{x})}{L(\theta_1|\boldsymbol{x})}=\frac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\theta_1)}.$$

定义广义似然比为

$$\begin{aligned}\Lambda(\boldsymbol{x})&=\frac{L(\theta_0|\boldsymbol{x})}{\max_{\theta=\theta_0,\theta_1}L(\theta|\boldsymbol{x})}\\&=\begin{cases}1, & L(\theta_0|\boldsymbol{x})\ge L(\theta_1|\boldsymbol{x}),\\ \lambda(\boldsymbol{x}), & L(\theta_0|\boldsymbol{x})<L(\theta_1|\boldsymbol{x}).\end{cases}\end{aligned}$$

显然，广义似然比的值不大于 1. 一般而言，检验的拒绝域 $\{\boldsymbol{x}\in\mathscr{X}\colon\lambda(\boldsymbol{x})\le$

$c\}$ 中的 c 比较小，如 $c < 1$. 因为若广义似然比的值等于 1，则说明备择假设中没有参数使样本出现的概率更大，因而不能拒绝原假设. 又，对任何 $c < 1$，$\lambda(\boldsymbol{x}) \le c$ 等价于 $\Lambda(\boldsymbol{x}) \le c$.

一般地，设有假设检验问题

$$H_0\colon \theta \in \Theta_0 \quad \leftrightarrow \quad H_1\colon \theta \in \Theta_1.$$

其中 $\Theta_0 \cap \Theta_1 = \varnothing$，$\Theta_0 \cup \Theta_1 = \Theta$.

记似然函数为 $L(\theta)$. 则似然比为

$$\Lambda^*(\boldsymbol{x}) = \frac{\max_{\Theta \in \Theta_0} L(\theta|\boldsymbol{x})}{\max_{\Theta \in \Theta_1} L(\theta|\boldsymbol{x})}.$$

较小的 $\Lambda^*(\boldsymbol{x})$ 倾向于拒绝 H_0. 拒绝域形如

$$\{\boldsymbol{x} \in \mathscr{X} : \Lambda^*(\boldsymbol{x}) < \lambda^*\},$$

其中 λ^* 为待定常数.

考虑广义似然比

$$\Lambda(\boldsymbol{x}) = \frac{\max_{\Theta \in \Theta_0} L(\theta|\boldsymbol{x})}{\max_{\Theta \in \Theta} L(\theta|\boldsymbol{x})} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})},$$

其中 $\hat{\theta}_0$ 是使 $L(\theta|\boldsymbol{x})$ 达到最大值的参数 $\theta \in \Theta_0$，即参数 $\theta \in \Theta_0$ 的极大似然估计；$\hat{\theta}$ 是使 $L(\theta|\boldsymbol{x})$ 达到最大值的参数 $\theta \in \Theta$，即参数 $\theta \in \Theta$ 的极大似然估计.

因为

$$\Lambda(\boldsymbol{x}) = \begin{cases} \dfrac{\max_{\Theta \in \Theta_0} L(\theta|\boldsymbol{x})}{\max_{\Theta \in \Theta_1} L(\theta|\boldsymbol{x})} = \Lambda^*(\boldsymbol{x}), & \max\limits_{\Theta \in \Theta_0} L(\theta|\boldsymbol{x}) < \max\limits_{\Theta \in \Theta_1} L(\theta|\boldsymbol{x}), \\ 1, & \text{其他}. \end{cases}$$

因此，

$$\Lambda(\boldsymbol{x}) = \min\{\Lambda^*(\boldsymbol{x}), 1\},$$

从而可知，若 $\lambda < 1$，则

$$\{\boldsymbol{x} \in \mathscr{X} : \Lambda^*(\boldsymbol{x}) < \lambda\} = \{\boldsymbol{x} \in \mathscr{X} : \Lambda(\boldsymbol{x}) < \lambda\}.$$

因此，根据似然比检验的思想，也利用广义似然比做出检验. 设显著性水平为 α，则拒绝域为

$$\{\boldsymbol{x} \in \mathscr{X} : \Lambda(\boldsymbol{x}) < \lambda\},$$

其中 λ 使得

$$\mathbb{P}(\Lambda(\boldsymbol{X}) < \lambda | H_0) = \alpha.$$

例 6.6 (正态均值检验) 设 $X_1, X_2, \cdots, X_n$ 为来自正态总体 $N(\mu, \sigma_0^2)$ 的样本，σ_0 已

知，检验问题为

$$H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0.$$

按上面的一般做法，取

$$\Theta_0 = \{\mu_0\}, \quad \Theta_1 = \{\mu: \mu \neq \mu_0\}, \quad \Theta = \mathbb{R}.$$

似然函数为

$$\begin{aligned} L(\mu) &= \frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2\right]\right). \end{aligned}$$

注意到 μ 的极大似然估计值为 $\overline{x}$，故广义似然比为

$$\begin{aligned} \Lambda :&= \frac{\max_{\mu\in\Theta_0} L(\mu)}{\max_{\mu\in\Theta} L(\mu)} \\ &= \frac{\frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \mu_0)^2\right)}{\frac{1}{(\sigma_0\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \overline{x})^2\right)} \\ &= \exp\left(-\frac{n(\overline{x} - \mu_0)^2}{2\sigma_0^2}\right). \end{aligned}$$

因此

$$-2\log\Lambda(\boldsymbol{X}) = \frac{n(\overline{X} - \mu_0)^2}{\sigma^2}.$$

似然比检验告诉我们，当 $\frac{n(\overline{x}-\mu_0)^2}{\sigma^2}$ 比较大时，拒绝 H_0. 因为

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1),$$

所以

$$\frac{n(\overline{X} - \mu)^2}{\sigma^2} \sim \chi^2(1).$$

因此，置信水平为 α 的拒绝域为

$$\left\{\boldsymbol{x} \in \mathscr{X}: \frac{n(\overline{x} - \mu_0)^2}{\sigma^2} > \chi^2_\alpha(1).\right\}.$$

它即

$$\left\{\boldsymbol{x} \in \mathscr{X}: \frac{|\sqrt{n}(\overline{x} - \mu_0)|}{\sigma} > z_{\alpha/2}.\right\}.$$

换言之，拒绝域为

$$\left\{\boldsymbol{x} \in \mathscr{X}: \overline{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad \text{或} \quad \overline{x} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.\right\}.$$

将正态均值检验中的思想推广，可得如下一般结论.

定理 6.2. χ^2 检验

设概率函数满足一定的光滑条件，则当 $n \to \infty$ 时，$-2\log\Lambda$ 依分布收敛自由度为 $\dim\Theta - \dim\Theta_0$ 的 χ^2 分布.

$\dim\Theta, \dim\Theta_0$ 分别为 Θ, Θ_0 中自由参数的个数.

在前面的例子中，σ_0 已知，零假设完全决定了 μ，$\dim\Theta_0 = 0$，在备择假设中，μ 可为 $\mathbb{R}$ 中任意值，故 $\dim\Theta = 1$. 因此，自由度为 $\dim\Theta - \dim\Theta_0 = 1 - 0 = 1$.

例 6.7 (指数分布的广义似然比检验) 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 服从自位置参数为 θ, 尺度参数为 1 的指数分布的样本，即 X 的概率密度函数为

$$f(x|\theta) = \begin{cases} \mathrm{e}^{-(x-\theta)}, & x \geq \theta, \\ 0 & x < \theta, \end{cases}$$

其中 $-\infty < \theta < \infty$. 对假设检验问题

$$H_0\colon \theta \leq \theta_0 \quad \leftrightarrow \quad H_1\colon \theta > \theta_0.$$

求检验水平为 α 的检验.

解 概率密度函数用示性函数可以表示为

$$f(x|\theta) = \mathrm{e}^{-(x-\theta)}\,\mathbb{1}_{[\theta,+\infty)}(x) = \mathrm{e}^{-(x-\theta)}\,\mathbb{1}_{(-\infty,x]}(\theta),$$

因此，似然函数为

$$\begin{aligned} L(\theta|\boldsymbol{x}) &= \exp\left(-\sum_{i=1}^n x_i + n\theta\right)\prod_{i=1}^n \mathbb{1}_{(-\infty,x_i]}(\theta) \\ &= \exp\left(-\sum_{i=1}^n x_i + n\theta\right)\mathbb{1}_{(-\infty,x_{(1)}]}(\theta), \end{aligned}$$

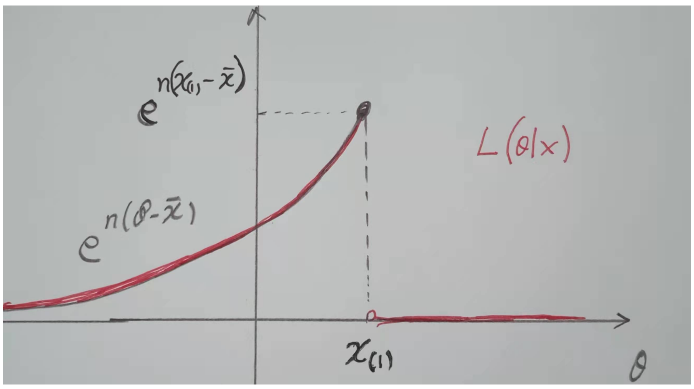


图 6.4　指数分布参数的似然函数

注意到 $L(\theta|\boldsymbol{x})$ 是 $\theta \in (-\infty, x_{(1)}]$ 的增函数，而当 $\theta > x_{(1)}$ 时，$L(\theta|\boldsymbol{x}) = 0$，故

$$\max_{\theta\in\mathbb{R}} L(\theta|\boldsymbol{x}) = L(x_{(1)}|\boldsymbol{x}) = \exp\left(-\sum_{i=1}^n x_i + nx_{(1)}\right),$$

$$\max_{\theta \le \theta_0} L(\theta|\boldsymbol{x}) = \begin{cases} = L(x_{(1)}|\boldsymbol{x}) = \exp\left(-\sum_{i=1}^{n} x_i + n x_{(1)}\right), & x_{(1)} \le \theta_0, \\ = L(\theta_0|\boldsymbol{x}) = \exp\left(-\sum_{i=1}^{n} x_i + n\theta_0\right), & x_{(1)} > \theta_0, \end{cases}$$

因此，广义似然比为

$$\lambda(\boldsymbol{x}) = \begin{cases} 1, & x_{(1)} \le \theta_0, \\ \exp(-n x_{(1)} + n\theta_0), & x_{(1)} > \theta_0, \end{cases}$$

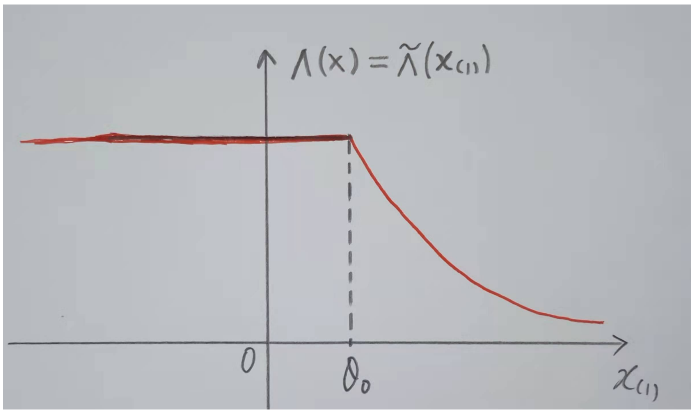


图 6.5　指数分布参数的广义似然比

因此，拒绝域为 $\{\boldsymbol{x} \in \mathscr{X} : \Lambda(\boldsymbol{x}) \le \lambda\}$，它等价于 $\{\boldsymbol{x} \in \mathscr{X} : x_{(1)} \ge \theta_0 - \frac{1}{n}\log\lambda\}$. 即拒绝域为

$$\{\boldsymbol{x} \in \mathscr{X} : x_{(1)} > c\},$$

其中常数 c 使得对预先给定的显著性水平 $0 < \alpha < 1$ 有

$$\mathbb{P}(X_{(1)} > c) = \alpha.$$

□

上面的例子显示，广义似然比检验的统计量仅依赖于充分统计量. 这是如下一般结论的反映.

定理 6.3. 广义似然比与充分统计量

如果 $T(\boldsymbol{X})$ 是 θ 的充分统计量，$\tilde{\Lambda}(T(\boldsymbol{X})), \Lambda(\boldsymbol{X})$ 分别为基于 $T(\boldsymbol{X}), \boldsymbol{X}$ 的广义似然比检验统计量，则对任意 $\boldsymbol{x} \in \mathscr{X}$，$\tilde{\Lambda}(T(\boldsymbol{x})) = \Lambda(\boldsymbol{x})$.

证明 由因子分解定理，有

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x}),$$

其中 $g(t|\theta)$ 是 $T(\boldsymbol{X})$ 的概率函数，于是

$$\begin{aligned}\Lambda(\boldsymbol{x}) &= \frac{\sup_{\theta\in\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\theta\in\Theta} L(\theta|\boldsymbol{x})} \\ &= \frac{\sup_{\theta\in\Theta_0} f(\boldsymbol{x}|\theta)}{\sup_{\theta\in\Theta} f(\boldsymbol{x}|\theta)} \\ &= \frac{\sup_{\theta\in\Theta_0} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{\sup_{\theta\in\Theta} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})} \\ &= \frac{\sup_{\theta\in\Theta_0} g(T(\boldsymbol{x})|\theta)}{\sup_{\theta\in\Theta} g(T(\boldsymbol{x})|\theta)} \\ &= \tilde{\Lambda}(T(\boldsymbol{x})).\end{aligned}$$

□

6.6 注记

图 6.6 年老时的费歇尔

在女士品茶试验中，费歇尔讨论了很多问题，例如：

1. 应该为奉上多少杯茶？
2. 是否要告诉她 8 杯奶茶中有多少杯是先加奶？
3. 怎样剔除茶水温度、奶与茶的混合度等随机因素？假设女士有超能力，8 杯奶茶中也可能有一杯没猜对，这样的概率有多大？

费歇尔曾在农业实验站工作. 在费歇尔之前，这个实验站已经进行了约 90 年的肥料构成（称之为人工肥料）实验.

统计学人物简介

Neyman（奈曼）：出生于俄罗斯，父母为波兰人. 大学先学物理，后学数学. 老师有 Bernstein，曾到巴黎与 Lebesgue,Borel 等工作. 前往伦敦（当时的世界统

计中心）与 Karl Pearson 工作，但与 K. Pearson 不和，但幸运的是与 Karl Pearson 的儿子 Egon Pearson 关系良好，有很多合作. Neyman-Pearson 引理即他们的合作成果. 1938 年，奈曼到美国，在伯克利建立了世界统计中心. 许宝騄为其中国学生，前面提到 Blackwell, Lehman, Scheffe 也是他的学生.

卡尔・皮尔逊（Karl Pearson，1857-1936）提出过矩估计法，分布密度函数族，χ^2 拟合优度检验. 他在科学哲学方面也非常有名，著有《科学的语法》[1]《自由思想的伦理》. 他原名 Carl Pearson，随 Karl Max 改名 Karl Pearson. 《科学的语法》对奈曼有很大的影响.

费歇（Ronald Aylmer Fisher，1890-1962），1909 年入剑桥，1912 年以 wrangler 毕业（考试成绩最好），1914 年因视力差而被军队拒绝，到高中任教直至 1919 年. 1919 年，入 Rothamsted 农业试验站（Rothamsted Experiment Station）做统计员. 那里有积累了 90 年的数据. 费歇尔指出采用哪种肥力指数并没有什么差异，所有修正都不足以调整不同地块上的肥力差异. 他还通过检查历史雨量和收成数据，指出年度间不同气候的影响远远大于不同肥力的影响.

1933 年，费歇到伦敦 University College 应用统计系作系主任. 著作有《研究人员用统计方法》《实验设计》. 费歇与 K. 皮尔逊关系不佳，不能得到许可在皮尔逊创立的《计量生物学》上发表文章. 费歇提倡优生学. 2020 年，由于 BLM 运动 (Black Lives Matter) 引起反种族歧视，反优生学，因而剑桥大学被迫将橱窗中将对费歇的纪念撤掉. 费歇尔是假设检验的主要奠基人，被称为推断统计之父.

埃贡（Egon Pearson，1895-1980）是卡尔・皮尔逊的唯一的儿子，与 Neyman 多有合作.

内容提要

- 假设检验
- 第一类错误
- 第二类错误
- Neyman-Pearson 引理
- 一致最优势检验

第 6 章 习题

1. 独立投掷硬币 10 次，记硬币正面朝上的概率为 p，假设检验问题为 $H_0: p = \frac{1}{2} \leftrightarrow p \neq \frac{1}{2}$. 如果正面出现 0 次或 10 次时拒绝检验原假设.
   (a). 检验的显著性水平是多少？
   (b). 如果 $p = 0.1$，检验的势是多少？
2. 设 $X_1, \cdots, X_n$ 为来自参数为 λ 的 Poisson 分布的样本.
   (a). 考虑检验问题 $H_0: \lambda = \lambda_0 \leftrightarrow H_1: \lambda = \lambda_1$（$\lambda_0 < \lambda_1$）. 求检验的似然比.

[1]The Grammar of Science, 李醒民有中译《科学的规范》

(b). 利用 Poisson 分布的可加性解释如何确定上述假设检验问题的显著性水平为 α 的拒绝域.

(c). 证明对于假设 $H_0\colon \lambda = \lambda_0 \leftrightarrow H_1\colon \lambda > \lambda_0$，上述检验是一致最优势的.

3. 设样本 $X_1, X_2, \cdots, X_n$ 来自总体 $X \sim B(1, \theta)$. 现在有假设检验

$$H_0\colon \theta = 0.48 \leftrightarrow H_1\colon \theta = 0.52.$$

检验方法如下：若 $\sum_{i=1}^n x_i$ 较大则拒绝 H_0. 利用中心极限定理进行近似计算，求样本量应该至少为多大才能使两类错误发生的概率都约为 0.01.

4. 设 $X_1, X_2, \cdots, X_n$ 为来自 Beta$(\mu, 1)$ 的样本，$Y_1, Y_2, \cdots, Y_m$ 为来自 Beta$(\theta, 1)$ 的样本，且假设这两个样本之间也相互独立.

(a). 求 $H_0\colon \theta = \mu \leftrightarrow H_1\colon \theta \neq \mu$ 的广义似然比检验.

(b). 证明上述检验可以基于如下统计量

$$T = \frac{\sum_{i=1}^n \log X_i +}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m Y_i}.$$

5. 设 X_1, X_2 独立同分布，为来自均匀分布 $(\theta, 1+\theta)$ 的样本. 为检验 $H_0\colon \theta = 0 \leftrightarrow H_1\colon \theta > 0$，考虑如下两种检验方法：

$$\phi_1\text{：} X_1 > 0.95\text{时，拒绝}H_0,$$

$$\phi_2\text{：} X_1 + X_2 > C\text{时，拒绝}H_0.$$

(a). 求 C 的值使得检验 ϕ_1, ϕ_2 的显著性水平相同.

(b). 求两个检验各自的势函数.

(c). 证明否否定：ϕ_2 的势比 ϕ_1 的势更大.

第 7 章 拟合优度检验

7.1 多项分布的检验

设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_m)$ 服从参数为 $p_1, p_2, \cdots, p_m$ 的多项分布. 设观察值为 $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$，似然函数为

$$L(p_1, p_2, \cdots, p_m|\boldsymbol{x}) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}.$$

按例3.12，p_j 的极大似然估计为 $\hat{p}_j = \frac{x_j}{n}$.

假设检验问题为

$$H_0\colon p_i = p_i(\theta), \quad \theta \in \Theta_0,$$

$$H_1\colon p_1 + p_2 + \cdots + p_m = 1.$$

广义似然比为

$$\Lambda = \frac{\max_{\theta\in\Theta_0} L(p_1(\theta), p_2(\theta), \cdots, p_m(\theta)|\boldsymbol{x})}{\max_{p_1+p_2+\cdots+p_m=1} L(p_1, p_2, \cdots, p_m|\boldsymbol{x})} = \prod_{i=1}^n \left(\frac{p_i(\hat{\theta})}{\hat{p}_i}\right)^{x_i}.$$

其中 $\hat{\theta}$ 是使得 $L(p_1(\theta), p_2(\theta), \cdots, p_m(\theta))$ 达到最大值的 $\theta \in \Theta_0$，而 $\hat{p}_i$ 满足 $x_i = n\hat{p}_i$.

因此，

$$-2\log\Lambda(\boldsymbol{x}) = -2n\sum_{i=1}^n \hat{p}_i \log\frac{p_i(\hat{\theta})}{\hat{p}_i} = 2\sum_{i=1}^n O_i \log\frac{O_i}{E_i},$$

其中

$$O_i = n\hat{p}_i, \quad E_i = np_i(\hat{\theta}),$$

分别表示观察频数和期望频数.

注意到 $p_1, p_2, \cdots, p_m$ 只有一个限制条件 $p_1+p_2+\cdots+p_m = 1$，故 $\dim\Theta = m-1$. 设 $\dim\Theta_0 = k$，则在 H_0 成立时，

$$-2\log\Lambda(\boldsymbol{X}) \sim \chi^2(m-k-1).$$

当 H_0 成立时，对充分大的 n，$\hat{p}_i \approx p_i(\hat{\theta})$. 考察函数

$$h(x) = x\log\left(\frac{x}{x_0}\right),$$

在 x_0 处做泰勒展开，可得

$$f(x) \approx (x - x_0) + \frac{1}{2}(x - x_0)^2\frac{1}{x_0}.$$

因此，

$$-2\log\Lambda(\boldsymbol{X}) \approx X^2 := \sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i} \sim \chi^2(m-k-1).$$

X^2 被称为皮尔逊 χ^2 统计量. 这里要将 O_i, E_i 视为 $O_i(\boldsymbol{X}), E_i(\boldsymbol{X})$.

$-2\log\Lambda(\boldsymbol{x})$ 的近似可以通过泰勒展开得到.

特别，对 $\dim\Theta_0 = 0$ 的情形，有如下定理

定理 7.1. 多项分布

设 $(X_1, X_2, \cdots, X_m) \sim \mathrm{Multinomial}(n, p_1, p_2, \cdots, p_m)$，$n = \sum_{i=1}^m X_i$，$\sum_{i=1}^m p_i = 1$，则当 $n \to \infty$ 时，

$$\sum_{i=1}^{m}\frac{(X_i-np_i)^2}{np_i} = \sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i}$$

依分布收敛于自由度为 $m-1$ 的 χ^2 分布：

$$Q_n \xrightarrow{L} \chi^2(m-1).$$

上述定理可以如下粗略理解. 首先注意到多项分布的边缘分布服从二项分布，即 $X_i \sim B(n, p_i)$. 设 n 充分大，p_i 较小，可知

$$\mathbb{E}X_i = np_i, \quad \operatorname{Var}X_i = np_i(1-p_i) \approx np_i.$$

由中心极限定理

$$\frac{X_i-np_i}{\sqrt{np_i}} \overset{\text{近似}}{\sim} N(0,1).$$

所以

$$\frac{(X_i-np_i)^2}{np_i} \sim \chi^2(1).$$

再由“可加性”（实际缺独立性），可知 $\sum_{i=1}^m \frac{(X_i-np_i)^2}{np_i}$ 近似服从自由度为 m 的 χ^2 分布. 注意到 $\sum_{i=1}^n (X_i - np_i) = 0$，故将自由度修正为 $m-1$.

当 $m=2$ 时，定理7.1的证明很显然. 此时 $p_1+p_2=1$，$X_1+X_2=n$，从而可得

$$X_2 - np_2 = (n-X_1) - n(1-p_1) = np_1 - X_1.$$

因此

$$\begin{aligned}\frac{(X_1-np_1)^2}{np_1} + \frac{(X_2-np_2)^2}{np_2} &= \frac{1}{n}(X_1-np_1)^2\left(\frac{1}{p_1}+\frac{1}{p_2}\right)\\ &= \frac{(X_1-np_1)^2}{np_1(1-p_1)}.\end{aligned}$$

注意到 $X_1 \sim B(n,p_1)$，由棣莫佛-拉普拉斯中心极限定理，

$$\frac{X_1-np_1}{\sqrt{np_1(1-p_1)}} \overset{\text{近似}}{\sim} N(0,1).$$

所以

$$\left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \overset{\text{近似}}{\sim} \chi^2(1).$$

例 7.1 某校高中生投掷硬币 17950 次，正面朝上 9207 次，反面 8743 次. 这是否足以说明该硬币不是均匀的?

解 使用 χ^2 检验 可见，在显著性水平 $\alpha = 0.01$ 的条件下，不能接受硬币是均匀

coin toss test						
		observed	expected	diff	diff^2	
正面	n1	9207	8975	232	5.997103064	
反面	n2	8743	8975	-232	5.997103064	
零假设概率	p1	0.5				
总投掷次数	n	17950			11.99420613	Chi^2 统计量
自由度	df	1				
显著性水平	alpha	0.05	0.01		0.000533662	p值
临界值	chi^2_alpha(1)	3.841458821	6.634896601			

图 7.1 硬币均匀性检验

的.

用电子表格 “=BINOM.DIST(9207,17950,0.5,1)” 可得正面次数不超过 9207 的概率为 0.000259352.

利用电子表格函数“=BINOM.INV(17950,0.5,0.005)”和“=BINOM.INV(17950,0.5,0.995)”可分别计算得到 8802,9148，即正面朝上的次数位于 $[8802, 9148]$ 之间的概率为 0.99，可见，正面朝上超过 9207 次的概率小于 0.01.

如果考虑正态分布近似，则因

$$np = 17950 \times \frac{1}{2} = 8975, \quad \sqrt{n \times \frac{1}{2} \times \frac{1}{2}} = 66.9888.$$

可知

$$\frac{X - 8975}{66.9888}$$

近似服从标准生态分布. 标准正态分布随机变量取值于 $[-2.575829304, 2.575829304]$ 之外的概率为 0.01. 因

$$\frac{9207 - 8975}{66.9888} = 3.463265501,$$

落在 $[-2.575829304, 2.575829304]$ 之外，可见，显著性水平为 0.01 时，应该拒绝原假设.

□

例 7.2 哈代-温伯格平衡

参 Excel 文件 Hardy-Weinberg-Equilibrium.xlsx.

例 7.3 细菌凝块 将 0.01 毫升的牛奶散放在 1 平方厘米的载玻片上，用显微镜观

Hardy-Weinberg平衡				
blood type	M	MN	N	sum
	(1-theta)^2	2theta(1-theta)	theta^2	1.0000
observed counts=O_i	342.0000	500.0000	187.0000	1029.0000
MLE of theta	0.4247			
p_i hat	0.3324	0.4859	0.1817	
p_i(theta^)	0.3310	0.4887	0.1804	
expected counts=E_i	340.5682	502.8310	185.6008	1029.0000
(O-E)^2/E	0.0060	0.0159	0.0105	0.0325
O*log(O/E)	1.4348	-2.8230	1.4044	0.0325
chi^2(1,0.05)	3.8415			
p-value	0.8569			

图 7.2 孟德尔豌豆试验

察细菌凝块数. 计算了 400 个方格上的凝块数.

number per square			observed		pi	expected counts	(Oi-Ei)^2/Ei
0	56	0	0	56	0.087160851	34.86434058	12.81298001
1	104	104	1	104	0.212672478	85.06899103	4.212852373
2	80	160	2	80	0.259460423	103.7841691	5.4506068
3	62	186	3	62	0.21102781	84.41112416	5.950145686
4	42	168	4	42	0.128726964	51.49078574	1.749342385
5	27	135	5	27	0.062818759	25.12750344	0.139538071
6	9	54	6	9	0.025546295	10.21851807	0.145303484
7	9	63	>=7	20	0.01258642	5.034567926	44.48527866
8	5	40				Pearson Chi^2	74.94604747
9	3	27	mle of	2.44		chisq(0.005,6)	18.54758418
10	2	20	n	400		p 值小于0.005	
19	1	19					

图 7.3 细菌凝块数

例 7.4 孟德尔试验 孟德尔在一次试验中测量得到各种豌豆的数据如下 请检验这四种豌豆出现的比例是否为 9:3:3:1.

$$-2\log\sum_{i=1}^{4} O_i \log\frac{O_i}{E_i} = 0.618.$$

利用自由度为 3 的 χ^2 分布，统计量的 p 值为 0.8922.

Pearson 的 χ^2 统计量为

$$\sum_{i=1}^{4}\frac{(O_i-E_i)^2}{E_i} = 0.604.$$

对应的 p 值为 0.8954.

类型	频数
光滑黄色	315
光滑绿色	108
皱黄色	102
皱绿色	31

type	Observed counts	Expected counts	Probablity	Oi/Ei	2*Oi*log(Oi/Ei)	(Oi-Ei)^2/Ei
Smooth yellow	315	312.75	9/16	1.007	4.5161	0.0162
Smooth green	108	104.25	3/16	1.036	7.6333	0.1349
Wrinkled Yellow	102	104.25	3/16	0.978	-4.4511	0.0486
Wrinkled green	31	34.75	1/16	0.892	-7.0799	0.4047
sum	556				0.6184	0.6043
degree of freedom			3	p value	0.8922	0.8954

图 7.4　孟德尔豌豆试验

费歇曾将孟德尔多次试验的结果放一起考虑，得出的 p 值非常大，因而令人对孟德尔是否操控数据而生疑.

参 Excel 文件 Fisher-Mendel-Reexam.xlsx.

7.2 分布的检验

设样本 $X_1, X_2, \cdots, X_n$ 来自分布函数为 F 的总体 X. 检验问题：

$$H_0 : F = F_0 \quad \leftrightarrow \quad H_1 : F \neq F_0.$$

将 X 的值域分割为 m 个互不相交的子集 $A_1, A_2, \cdots, A_m$. 设 N_i 为落入 A_i 中的样本个数

$$N_i = \#\{X_j : X_j \in A_i, j = 1, 2, \cdots, n\}, \quad i = 1, 2, \cdots, m.$$

则 $(N_1, N_2, \cdots, N_m)$ 服从参数为 $n, p_1, p_2, \cdots, p_m$ 的多项分布，其中样本量 n 满足

$$n = \sum_{i=1}^{m} N_i,$$

而 p_i 为随机变量 X 取值于 A_i 的概率：

$$p_i = \mathbb{P}(X \in A_i) = \int_{A_i} dF(x).$$

设 $x_1, x_2, \cdots, x_n$ 为样本 $X_1, X_2, \cdots, X_n$ 的观察值，则 $N_1, N_2, \cdots, N_m$ 的观察值

$n_1, n_2, \cdots, n_m$ 为

$$n_i = \#\{x_j : x_j \in A_i, j = 1, 2, \cdots, n\}, \quad i = 1, 2, \cdots, m.$$

样本取值 A_i 中的频率为 $\frac{n_i}{n}$，这个频率是随机变量 $\frac{N_i}{n}$ 的观测值. 按大数定律，$\frac{N_i}{n}$ 依概率收敛收敛于 p_i.

原检验问题可以通过检验如下问题（依赖于区间的划分）来处理：

$$H_0' : p_i = p_{i_0} := \int_{A_i} dF_0(x), \quad i = 1, 2, \cdots, m.$$

$$H_1' : \text{存在} i = 1, 2, \cdots, m, p_i \neq p_{i_0}.$$

注意到 $\sum_{i=1}^m p_i = 1$，所以 m 个等式

$$p_i = p_{i_0}, i = 1, 2, \cdots, m$$

实际上只有 $m - 1$ 个等式.

即我们使用检验多项分布的方法来检验总体分布. 此时，落入“盒子” A_i 中的观察频数 O_i 和期望频数 E_i 分别为：

$$O_i = N_i = n\hat{p}_i, \quad E_i = np_{i0}, \quad i = 1, 2, \cdots, m.$$

按此前的检验方法，若 H_0 为真，则统计量在 $n \to \infty$ 时，

$$Q_n := \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^m \frac{(N_i - np_{i_0})^2}{np_{i_0}}$$

依分布收敛于自由度为 $m - 1$ 的 $\chi^2(m - 1)$：

$$\text{当} n \to \infty \text{时}, \quad Q_n = \sum_{i=1}^m \frac{(N_i - np_{i_0})^2}{np_{i_0}} \xrightarrow{L} \chi^2(m - 1).$$

因为 Q_n 近似服从自由度为 $m - 1$ 的卡方分布，从而

$$\alpha \approx \mathbb{P}(Q_n > \chi^2_\alpha(m - 1)).$$

统计量 Q_n 的观察值为

$$q_n = \frac{(n_i - np_{i_0})^2}{np_{i_0}}.$$

显著性水平为 $\alpha \in (0, 1)$ 的拒绝域为

$$\{(n_1, n_2, \cdots, n_m) : q_n \geq \chi^2_\alpha(m - 1)\}.$$

例 7.5 用 Excel 生成 $U_i \sim U(0, 1)$，检验 $\sum_{i=1}^{12} U_i - 6$ 是否服从标准正态分布.

参考 excel 文件 Uniform-Sum2Norm.xlsx.

例 7.6 Clark 曾对伦敦受袭的数据做 χ^2 检验：自由度为 4，χ^2 值为 1.17，得到不小于这个值的概率为.88.

袭击数（k）	0	1	2	3	4	≥ 5
遭 k 个袭击的区域	229	211	93	35	7	1
Poisson 预测	226.74	211.39	98.54	30.62	7.14	1.57

表 7.1　导弹袭击分布

7.3 概率图与正态性检验

设 $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} U(0,1)$，则由前面的知识，

$$X_{(r)} \sim \mathrm{Beta}(k, n+1-k),$$

所以（参考命题10.4）

$$\mathbb{E}X_{(r)} = \frac{r}{n+1}.$$

将散点 $(x_{(r)}, \frac{r}{n+1})$，$r = 1, 2, \cdots, n$，在直角坐标系上作出，则它们近似在一条斜率为 1 直线上．这一点可以如下理解：总体服从均匀分布，则可以认为 $X_1, X_2, \cdots, X_n$ 均匀地分布在 $(0,1)$ 上. 将 $(0,1)$ 分成 n 份，理想情况，可以认为排序后的样本点从小到大分布在这 n 份里，这就可以合理的理解 $X_{(r)}$ 近似于 $\frac{r}{n+1}$.

反过来利用这个线性性可以检验数据是否服从均匀分布.

现设 $X_i \overset{\text{iid}}{\sim} F \sim X$，则 X 的概率积分变换

$$F(X) \sim U(0,1).$$

事实上，

$$\mathbb{P}(F(X) \leq y) = \mathbb{P}(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y.$$

所以

$$\mathbb{E}F(X_{(r)}) = \frac{r}{n+1}.$$

因此，

$$F(X_{(r)}) \approx \frac{r}{n+1}.$$

由此可得

$$X_{(r)} \approx F^{-1}(\frac{r}{n+1}).$$

也就是说，如果 $X \sim F$，则散点图

$$\left(X_{(r)}, F^{-1}\left(\frac{r}{n+1}\right)\right)$$

近似在一条直线.

现设 F 有如下形式

$$F(x)=G\left(\frac{x-\mu}{\sigma}\right).$$

则由 $F(X_{(r)})\approx\frac{r}{n+1}$ 可得

$$G\left(\frac{X_{(r)}-\mu}{\sigma}\right)\approx\frac{r}{n+1}.$$

等价地有

$$\frac{X_{(r)}-\mu}{\sigma}\approx G^{-1}\left(\frac{r}{n+1}\right),$$

或

$$X_{(r)}\approx\sigma G^{-1}\left(\frac{r}{n+1}\right)+\mu.$$

注意到正态分布的分布函数有如下形式

$$F(x|\mu,\sigma^2)=\Phi\left(\frac{x-\mu}{\sigma}\right),$$

其中 Φ 为标准正态分布的分布函数.

可见，如果散点图

$$\left(x_{(r)},\Phi^{-1}\left(\frac{r}{n+1}\right)+\mu\right)$$

近似在一条直线，则可以判断，分布近似为正态分布.

参考 excel 文件 probability-plot.xlsx

内容提要

- 多项分布检验
- χ^2 检验

第 7 章 习题

1. 假设基因频率是均衡的，基因型 AA,Aa,aa 出现的频率分比为 $(1-\theta)^2, 2\theta(1-\theta)$，$\theta^2$. Plato 等发表了如下 190 个人类样本中触珠蛋白型的数据：

Hp1-1	Hp1-2	Hp2-2
10	68	112

表 7.2 触珠蛋白型

请验证遗传模型的拟合优度.

2. Bhattacharjee 等为研究在盈月期间动物是否更易咬人而收集了某医疗结构处理被动物咬伤的病例. 月亮周期被划分为 10 个阶段，每个阶段被咬伤的人数如下表所示. 29 日是盈月. 咬伤事件是否有时间趋势?

日期	2,3,4	5,6,7	8,9,10	11,12,13	14,15
人数	155	142	146	148	110

日期	16,17,18	19,20,21	22,23,24	25,26,27	28,29,1
人数	137	150	163	201	269

3. 某人投掷一枚正四面体骰子 80 次，以检验它是否为均匀的. 得到的各点数分别为

点数	1	2	3	4
频数	17	18	22	23

4. 文件 Distribution-Test-Beta-Data.xlsx 中有 100 个数据，请检验这些数据是否为来自 Beta 分布 $\mathrm{Beta}(2,1)$ 的观察值.
5. 生成容量为 100 的指数分布随机数（参数可以自定或取为 1），作出概率图.

第 8 章　两样本比较

为研究某种条件对研究对象的影响，进行除该种条件下，其他条件均相同的实验. 称此类实验是对照实验. 在对照实验中，实验组接受处理；控制组（又称对照组）不接受处理. 从这两类实验可得到两个样本. 希望对这两个样本进行检验，以考察接受处理的效果.

两样本检验，配对或不配对，都有对应的非参数方法，不需要假设原数据服从正态分布.

8.1　独立样本检验：基于正态分布的方法

设有两个独立样本 $X_1, X_2, \cdots, X_n$ 和 $Y_1, Y_2, \cdots, Y_m$，分别来自正态总体 $X \sim N(\mu_X, \sigma^2)$ 和 $Y \sim N(\mu_Y, \sigma^2)$. 分别记两样本的样本均值为 $\overline{X}, \overline{Y}$：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

两样本的样本方差为 S_X^2, S_Y^2：

$$S_X^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad S_Y^2 := \frac{1}{m-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

不妨设 X, Y 分别来自实验组和控制组. 实验效果以 μ_X 与 μ_Y 的差来体现. 所以，假设检验问题原假设为

$$H_0 \colon \mu_X - \mu_Y = 0,$$

而备择假设可以为双边备择假设

$$H_1 \colon \mu_X - \mu_Y \neq 0,$$

或单边备择假设

$$H_2 \colon \mu_X - \mu_Y > 0,$$

或

$$H_3 \colon \mu_X - \mu_Y < 0.$$

易知 $\overline{X} - \overline{Y}$ 是 $\mu_X - \mu_Y$ 的极大似然估计. 拒绝域依赖于 $\overline{X} - \overline{Y}$ 的取值.

样本来自正态总体，所以

$$\overline{X} \sim N\left(\mu_X, \frac{1}{n}\sigma^2\right), \quad \overline{Y} \sim N\left(\mu_Y, \frac{1}{m}\sigma^2\right).$$

因为两样本独立，所以 $\overline{X},\overline{Y}$ 独立，从而有

$$\overline{X}-\overline{Y}\sim N\left[\mu_X-\mu_Y,\left(\frac{1}{n}+\frac{1}{m}\right)\sigma^2\right].$$

若 σ^2 已知，则

$$U:=\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim N(0,1).$$

因为

$$\frac{(n-1)S_X^2}{\sigma^2}\sim\chi^2(n-1),\quad \frac{(m-1)S_Y^2}{\sigma^2}\sim\chi^2(m-1)]$$

且两者独立，从而

$$V=\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}\sim\chi^2(m+n-2).$$

由抽样分布定理，$\overline{X}$ 与 S_X^2 独立，而 $\overline{Y}$ 与 S_X^2 分别是两个独立样本的函数，从而独立，所以 U 与 V 独立. 从而

$$T:=\frac{U}{\sqrt{\frac{V}{m+n-2}}}=\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(m+n-2).$$

其中

$$S_p^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{m+n-2},$$

为联合样本方差. 在方差 σ^2 未知时，用 S_p 替换 U 中的 σ 就得到 T.

因此，$\mu_X-\mu_Y$ 的置信水平为 $100(1-\alpha)\%$ 的置信区间为

$$\left[(\overline{X}-\overline{Y})-t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}},(\overline{X}-\overline{Y})+t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right].$$

由检验与置信区间的对偶性，对检验问题 $H_0\boldsymbol{s}H_1$，在显著性水平为 α 时，若

$$\bar{x}-\bar{y}<t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\quad 或\quad \bar{x}-\bar{y}>t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}$$

则拒绝原假设.

8.2 独立样本检验：非参数方法（秩和检验）

秩和检验指的是 Wilcoxon 秩和检验（威尔科克森秩和检验），也被称为 Mann-Whitney 检验（曼恩-威特尼检验），或 Mann-Whitney-Wilcoxon 检验（缩写为 MWW 检验）.

检验量的选取在不同文献中有不同方法，对应的临界值表也有些差异. 但其中有着联系. 我们简单介绍两种有联系的方法.

8.2.1 秩的计算

定义 8.1. 秩

设 $x_1, x_2, \cdots, x_n$ 为样本观察值. 设

$$x_{(1)} < x_{(2)} < \cdots < x_{(r)} < \cdots < x_{(n)}$$

则称 r 为 $x_{(r)}$ 的秩.

如果有重复数据，用秩的平均值代替. 在电子表格中（如 Excel），可以用函数

rank.avg(数值，引用范围，升序 1）

得到秩.

非参数方法是不需要假设样本的分布，一般将数据用其秩来代替，通过分析秩来进行推断.

设从两个总体中分别抽取 n_1, n_2 个样本，将这 $n_1 + n_2$ 个观察值放一起，从小到大排序，算出各个观察值的秩. 将属于第一个总体的样本的观察值的秩相加，称为第一个总体的秩和，记为 R_1，其余观察值的秩的总和称为第 2 个总体的的秩和，记为 R_2.

显然，R_1, R_2 是离散型随机变量，因为

$$R_1 + R_2 = \frac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1),$$

所以只要研究其中一个，R_1 或 R_2 即可，如仅考虑秩和较小的那个，或样本量较小的那个样本的秩和.

R_1 的最小值是当样本 1 的观察值在全部观察值中排在前 n_1 个，秩取值于 $1, 2, \cdots, n_1$，秩和为

$$1 + 2 + \cdots + n_1 = \frac{1}{2}n_1(n_1 + 1).$$

R_1 的最大值是当样本 1 的观察值排在后 n_1 个，秩取值于 $n_2+1, n_2+2, \cdots, n_2+n_1$，秩和为

$$(n_2 + 1) + (n_2 + 2) + \cdots + (n_2 + n_1) = \frac{1}{2}n_1(n_1 + 2n_2 + 1).$$

所以，R_1 的取值范围为

$$\frac{1}{2}n_1(n_1 + 1) \le R_1 \le \frac{1}{2}n_1(n_1 + 2n_2 + 1).$$

8.2.2 U 统计量及其他

设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本，$Y_1, Y_2, \cdots, Y_m$ 为来自 Y 的样本. 零假设 H_0 为两个总体的分布相同，备择假设为两个总体的分布不同：

$$H_0: F_X = F_Y \leftrightarrow H_1: F_X \neq F_Y.$$

如果分布相同，则

$$\mathbb{P}(X_i > Y_j) = \mathbb{P}(X_i < Y_j) = \frac{1}{2}.$$

秩和检验，符号秩就是利用这一点来构造统计量.

严格说，我们其实是检验两个总体的分布的中位数是否有差异，因此实际需要假设两总体的概率函数 f_1, f_2 有平移关系：$f_1(x-\mu_1) = f_2(x-\mu_2)$. 检验问题就相当于检验 $H_0: \mu_1 = \mu_2$ 是否成立. 此时备择假设可以为 $H_1: \mu_1 \neq \mu_2$，或其他备择假设，如 $H_1: \mu_1 > \mu_2$ 或 $H_1: \mu_1 \neq \mu_2$.

Mann-Whitney 的统计量定义为

$$U(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} S(X_i, Y_j),$$

其中

$$S(X,Y) = \begin{cases} 1, & X > Y, \\ \frac{1}{2}, & X = Y, \\ 0, & X < Y. \end{cases}$$

例 8.1 考察如下两组数据：

X	Y
7	6
8	7
	9

表 8.1 两个样本

易计算得到两个 U 统计量为

$$U_1 = U(\mathbf{X}, \mathbf{Y}) = (1 + 0.5 + 0) + (1 + 1 + 0) = 3.5,$$
$$U_2 = U(\mathbf{Y}, \mathbf{X}) = (0 + 0) + (0.5 + 0) + (1 + 1) = 2.5.$$

一般取 $U=\min(U_1,U_2)=2.5$ 作为检验值. U_1,U_2 有如下联系：

$$U_1+U_2=nm.$$

所以，实际上只要计算出 U_1，就有

$$U_2=nm-U_1=2\times3-3.5=2.5.$$

实际中，可使用统计量

$$U=\min\{U_1,U_2\}=\min\{3.5,2.5\}=2.5.$$

需要查对应的临界值表.

将两个样本合并排序后可得它们的秩如下（样本 **X** 的数据用红色标出）

原始数据	6	7	7	8	9
序数	1	2	3	4	5
秩	1	2.5	2.5	4	5

表 8.2 秩

两个样本 **X**, **Y** 的秩和分别为

$$R_1=2.5+4=6.5,\quad R_2=1+2.5+5=8.5.$$

两个秩和有如下关系：

$$R_1+R_2=\frac{1}{2}(n+m)(n+m+1),$$

因此实际只要计算一个秩和，如计算出 R_1 就可以得到

$$R_2=\frac{1}{2}(n+m)(n+m+1)-R_1=\frac{1}{2}(2+3)(2+3+1)-6.5=15-8.5=8.5.$$

我们一般如下计算秩和的临界值：使用如下步骤：

1. 考虑样本量小的样本的秩和 R.
2. 取 $n'=\min\{n,m\}$，$R'=n'(n+m+1)-R$.
3. 取 $R^*=\min\{R,R'\}$.
4. 将 R^* 与临界值作比较.

对我们上面的例子，可得

$$n'=\min\{2,3\}=2,\quad R'=2(2+3+1)-6.5=5.5,\quad R^*=\min\{6.5,5.5\}=5.5.$$

U_1, U_2, U 与 R_1, R_2, R^* 有如下联系（待证明）：

$$U_1 = R_1 - \frac{1}{2}n(n+1) = 6.5 - \frac{1}{2} \times 2(2+1) = 6.5 - 3 = 3.5,$$
$$U_2 = R_2 - \frac{1}{2}m(m+1) = 8.5 - \frac{1}{2} \times 3(3+1) = 8.5 - 6 = 2.5,$$
$$U = R^* - \frac{1}{2}n'(n'+1) = 5.5 - 3 = 2.5$$

R 也可以如下计算.

将原始数据合并，倒序排列计算相应的秩和：

原始数据	9	8	7	7	6
序数	1	2	3	4	5
秩	1	2	3.5	3.5	5

表 8.3 倒序：秩

则

$$R' = 2 + 3.5 = 5.5.$$

8.2.3 计算例子

秩和检验的基本思想是：如果两个总体没差别，则两个样本的数据应该相互“渗透”，秩和应该不极端（和样本量有关.）

考虑冰融化的潜热的例子.

两个样本的样本量分别为 $n = 8, m = 13$.

使用如下步骤：

1. 考虑样本量小的样本的秩和 $R = 51$.
2. 取 $n' = \min\{n, m\}$，
$$R' = n'(n+m+1) - R.$$
在我们的例子中，$R' = 8(8+13+1) - 51 = 176 - 51 = 125$.
3. 取 $R^* = \min\{R, R'\} = \min\{51, 125\}$.
4. 查表，显著性水平为 0.01 的临界值为 53（8,13 对应的数据），大于 51，拒绝原假设.

Excel 文件“冰潜热-两样本比较的参数法和非参数法.xlsx”

			rank			
method A	method B	合并数据	从大到小	从小到大		
79.98	80.02	80.02	10.5	11.5	n	13
80.04	79.94	79.94	21	1	m	8
80.02	79.98	79.98	14.5	7.5	较小样本量n'	8
80.04	79.97	79.97	17.5	4.5	R	51
80.03	79.97	79.97	17.5	4.5	R'	125
80.03	80.03	80.03	6.5	15.5	根据倒序算的秩	125
80.04	79.95	79.95	20	2	R*	51
79.97	79.97	79.97	17.5	4.5	临界值	53
80.05		79.98	14.5	7.5		
80.03		80.04	3	19	E T_Y	88
80.02		80.02	10.5	11.5	var T_Y	190.6667
80.00		80.04	3	19	Z	-2.6796
80.02		80.03	6.5	15.5	p value	0.0074
		80.03	6.5	15.5	alpha	0.05
		80.04	3	19	临界值	-1.959964
		79.97	17.5	4.5		
		80.05	1	21		
		80.03	6.5	15.5		
		80.02	10.5	11.5		
		80.00	13	9		
		80.02	10.5	11.5		

图 8.1

TABLE 8 Critical Values of Smaller Rank Sum for the Wilcoxon Mann-Whitney Test (Continued)

	α for Two-Sided Test	α for One-Sided Test	n_1 (Smaller Sample)																			
n_2			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10	.20	.10	1	6	12	20	28	38	49	60	73	87										
	.10	.05		4	10	17	26	35	45	56	69	82										
	.05	.025		3	9	15	23	32	42	53	65	78										
	.01	.005			6	12	19	27	37	47	58	71										
11	.20	.10	1	6	13	21	30	40	51	63	76	91	106									
	.10	.05		4	11	18	27	37	47	59	72	86	100									
	.05	.025		3	9	16	24	34	44	55	68	81	96									
	.01	.005			6	12	20	28	38	49	61	73	87									
12	.20	.10	1	7	14	22	32	42	54	66	80	94	110	127								
	.10	.05		5	11	19	28	38	49	62	75	89	104	120								
	.05	.025		4	10	17	26	35	46	58	71	84	99	115								
	.01	.005			7	13	21	30	40	51	63	76	90	105								
13	.20	.10	1	7	15	23	33	44	56	69	83	98	114	131	149							
	.10	.05		5	12	20	30	40	52	64	78	92	108	125	142							
	.05	.025		4	10	18	27	37	48	60	73	88	103	119	136							
	.01	.005			7	*13	22	31	41	53	65	79	93	109	125							
14	.20	.10	1	*8	16	25	35	46	59	72	86	102	118	136	154	174						
	.10	.05		*6	13	21	31	42	54	67	81	96	112	129	147	166						
	.05	.025		4	11	19	28	38	50	62	76	91	106	123	141	160						
	.01	.005			7	14	22	32	43	54	67	81	96	112	129	147						

图 8.2

如果样本量比较大，可以用正态分布近似.

定理 8.1

用 T_Y 表示 $Y_1, Y_2, \cdots, Y_m$ 的秩和. 如果 $F_X = F_Y$，则

$$\mathbb{E}T_Y = \frac{1}{2}m(m+n+1),$$

$$\operatorname{Var} T_Y = \frac{1}{12}nm(m+n+1).$$

因 T_Y 近似正态分布，则

$$\frac{T_Y - \mu_{T_Y}}{\sigma_{T_Y}}$$

近似服从正态分布.

对冰融化潜热的例子，可以计算如下：

$$\begin{aligned} T_Y &= 51, \\ \mathbb{E}T_Y &= \frac{1}{2} \times 8(8+13+1) = 88, \\ \operatorname{Var} T_Y &= \sqrt{\operatorname{Var} T_Y} = \sqrt{\frac{1}{12} \times 8 \times 13(8+13+1)} \approx 13.8, \end{aligned}$$

统计量为

$$\frac{T_Y - \mathbb{E}T_Y}{\sigma_{T_Y}} = -2.68.$$

具体计算参 Excel 文件或截图.

8.2.4 如何算概率

样本 1 的观察值的秩的取法有 $\binom{n_1+n_2}{n_1}$ 种.

下面是一个 toy 模型. 考虑 $n_1 = n_2 = 2$ 的情形，第一个样本 $\binom{4}{2} = 6$ 种可能的秩的取法及秩和的取值、相应的概率、累积概率如下：

秩	秩和 R_1
1 2	3
1 3	4
1 4	5
2 3	5
2 4	6
3 4	7

秩和的可能取值为 3,4,5,6,7，相应概率、累积概率如下表

秩和 R_1	概率	累积概率
3	$\frac{1}{6}$	$\frac{1}{6}$
4	$\frac{1}{6}$	$\frac{2}{6}$
5	$\frac{2}{6}$	$\frac{4}{6}$
6	$\frac{1}{6}$	$\frac{5}{6}$
7	$\frac{1}{6}$	1

在 H_0 成立的条件下，两个样本独立同分布，因而可以放在一起研究. 第一个样本的观察值应该随机地分散在 $n_1 + n_2$ 个观察值中. R_1 的值不应该很大，也不应该很小. 如果过大或过小（根据备择假设来定）就应该拒绝原假设. 例如，对备择假设 H_1: $\mu_1 \neq \mu_2$，拒绝域为

$$R_1 \leq C_U(\alpha/2) \quad 或 \quad R_1 \geq C_L(\alpha/2),$$

其中 C_U, C_L 应满足

$$\mathbb{P}(R_1 \leq C_U(\alpha/2)) \leq \frac{\alpha}{2}, \quad 且 \quad \mathbb{P}(R_1 \geq C_L(\alpha/2)) \leq \frac{\alpha}{2}.$$

实际上，C_U, C_L 取满足上述条件的最小整数. 此时有

$$\mathbb{P}(R_1 \leq C_U(\alpha/2)\,或\,R_1 \geq C_L(\alpha/2)) \leq \alpha.$$

下面考虑 $n_1 = 3, n_2 = 4$ 的情况. 第一个样本的各个观察值的秩的分布情况有 $\binom{3+4}{3} = 35$ 种情况. 列表如下：

秩	R_1	秩	R_1	秩	R_1	秩	R_1	秩	R_1
1 2 3	6	1 3 6	10	1 6 7	14	2 4 7	13	3 5 6	14
1 2 4	7	1 3 7	11	2 3 4	9	2 5 6	13	3 5 7	15
1 2 5	8	1 4 5	10	2 3 5	10	2 5 7	14	3 6 7	16
1 2 6	9	1 4 6	11	2 3 6	11	2 6 7	15	4 5 7	15
1 2 7	10	1 4 7	12	2 3 7	12	3 4 5	12	4 5 7	16
1 3 4	8	1 5 6	12	2 4 5	11	3 4 6	13	4 6 7	17
1 3 5	9	1 5 7	13	2 4 6	12	3 4 7	14	5 6 7	18

秩和的可能取值为 $6, 7, \cdots, 17, 18$，相应的概率、累积概率如下表

秩和 R_1	6	7	8	9	10	11	12	13	14	15	16	17	18
概 率	$\frac{1}{35}$	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{3}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{5}{35}$	$\frac{4}{35}$	$\frac{4}{35}$	$\frac{3}{35}$	$\frac{2}{35}$	$\frac{1}{35}$	$\frac{1}{35}$
累积概率	$\frac{1}{35}$	$\frac{2}{35}$	$\frac{4}{35}$	$\frac{7}{35}$	$\frac{11}{35}$	$\frac{15}{35}$	$\frac{20}{35}$	$\frac{24}{35}$	$\frac{28}{35}$	$\frac{31}{35}$	$\frac{33}{35}$	$\frac{34}{35}$	1

考虑检验问题 $H_0 \colon \mu_1 = \mu_2$，$H_1 \colon \mu_1 > \mu_2$，取显著性水平 $\alpha = 0.05$，则因

$$\mathbb{P}(R_1 \geq 18) = \frac{1}{35} < 0.05, \quad \mathbb{P}(R_1 \geq 17) = \frac{2}{35} = 0.057 > 0.05,$$

可知拒绝域为 $r_1 = 18$.

例 8.2 考虑两组作比较的数据

组 1	1.9	0.5	0.9	2.1	
组 2	3.1	5.3	1.4	4.6	2.8

将全部数据排序，

数据	0.5	0.9	1.4	1.9	2.1	2.8	3.1	4.6	5.3
秩	1	2	3	4	5	6	7	8	9

零假设为 $H_0 \colon \mu_1 = \mu_2$，备择假设为 $H_1 \colon \mu_1 < \mu_2$.

在这个例子中，$n_1 = 4, n_2 = 5$. 在样本 1 有一个 4 个样本，样本 2 有 5 个样本的情形，样本 1 的观察值的秩的取法有 $\binom{4+5}{4} = 126$ 种等可能的取法. R_1 的最小可能值为 $\frac{1}{2} \times 4 \times 5 = 10$，最大可能值为 $\frac{1}{2} \times 4 \times (4 + 2 \times 5 + 1) = 30$.

在我们的例子中，R_1 的观察值为

$$r_1 = 1 + 2 + 4 = 5 = 12.$$

在显著性水平 $\alpha = 0.05$ 时，拒绝域形如 $\{R_1 \leq c\}$，其中 c 使得

$$\mathbb{P}(R_1 \leq c | H_0) \leq 0.05.$$

下表列出了样本 1 的秩不大于 14 时，各观察值秩的可能取值为

其中，秩和为 10,11,12,13,14 的情况的频数及相应概率、累积概率如下表

由此可见，在原假设成立的条件下，

$$\mathbb{P}(R_1 \leq 12 | H_0) \approx 0.032, \quad \mathbb{P}(R_1 \leq 13 | H_0) \approx 0.056, \quad \mathbb{P}(R_1 \leq 14 | H_0) \approx 0.1.$$

由此可见，我们可取 R_1 的临界值为 12. 现在观察值为 12，可见要拒绝原假设.

秩	秩和 R_1
1 2 3 4	10
1 2 3 5	11
1 2 3 6	12
1 2 3 7	13
1 2 3 8	14
1 2 4 5	12
1 2 4 6	13
1 2 4 7	14
1 2 5 6	14
1 3 4 5	13
1 3 4 6	14
2 3 4 5	14

可以证明，当 H_0 为真时，

$$\mu_{R_1} = \mathbb{E}R_1 = \frac{n_1(n_1+n_2+1)}{2},$$

$$\sigma^2_{R_1} = \operatorname{Var} R_1 = \frac{n_1 n_2(n_1+n_2+1)}{12}.$$

如果 $n_1, n_2 \geq 10$，则

$$R_1 \overset{\text{近似}}{\sim} N(\mu_{R_1}, \sigma^2_{R_1}),$$

秩和 R_1	概率	累积概率
10	$\frac{1}{126}$	$\frac{1}{126}$
11	$\frac{1}{126}$	$\frac{2}{126}$
12	$\frac{2}{126}$	$\frac{4}{126}$
13	$\frac{3}{126}$	$\frac{7}{126}$
14	$\frac{5}{126}$	$\frac{12}{126}$

因而

$$\frac{R_1 - \mu_{R_1}}{\sigma_{R_1}} \sim N(0,1).$$

8.3 配对样本检验：基于正态性假设

很多对照实验是配对试验，控制组与试验组是相关的，这样可以有效考察实验效果. 数学上，利用配对试验可以降低方差.

设 $(X_i, Y_i), i = 1, 2, \cdots, n$ 为配对样本，$\{X_i\}$ 来自正态总体 $X \sim N(\mu_X, \sigma_X^2)$，$\{Y_i\}$ 来自正态总体 $Y \sim N(\mu_Y, \sigma_Y^2)$. 假设不同的配对是独立的，但配对样本是正相关的，即 $\mathrm{Cov}(X_i, Y_i) = \sigma_{XY} > 0$.

设

$$D_i = X_i - Y_i, \quad i = 1, 2, \cdots, n.$$

则 $\{D_i\}$ 是相互独立的，且

$$\mathbb{E}D_i = \mu_X - \mu_Y,$$

$$\mathrm{Var}\, D_i = \mathrm{Var}(X_i - Y_i) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_{XY},$$

其中 $\rho > 0$ 为 X, Y 的相关系数.

记

$$\overline{D} = \frac{1}{n}\sum_{i=1}^n D_i = \overline{X} - \overline{Y},$$

则

$$\mathbb{E}\overline{D} = \mu_X - \mu_Y,$$

$$\mathrm{Var}\, \overline{D} = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_{XY}).$$

显然，

$$\mathrm{Var}\, \overline{D} < \frac{1}{n}(\sigma_X^2 + \sigma_Y^2).$$

即比独立样本时的方差更小.

即 $S_{\overline{D}}$ 为 $\overline{D}$ 的样本标准差，

$$S_{\overline{D}}^2 = \frac{1}{n}\mathscr{S}_D^2 = \frac{1}{n}\left(\frac{1}{n-1}\sum_{i=1}^n (D_i - \overline{D})^2\right).$$

则

$$t = \frac{\overline{D} - \mu_D}{S_{\overline{D}}} = \frac{\sqrt{n}(\overline{D} - \mu_D)}{S_D} \sim t(n-1).$$

μ_D 的置信水平为 $100(1-\alpha)\%$ 的置信区间为

$$\left[\overline{D} - t_{\alpha/2}(n-1)s_{\overline{D}}, \overline{D} + t_{\alpha/2}(n-1)s_{\overline{D}}\right].$$

对检验问题

$$H_0: \mu_D = 0 \boldsymbol{s} H_1: \mu_D \neq 0,$$

显著性水平为 α 时的拒绝域为

$$\overline{D} > t_{\alpha/2}(n-1)s_{\overline{D}}.$$

8.4 配对样本检验：非参数方法（符号秩检验）

Wilcoxon 符号秩检验用于配对检验. 如前，设 $(X_i, Y_i), i = 1, 2, \cdots, n$，为来自 (X, Y) 的样本. 设

$$D_i := X_i - Y_i, \quad i = 1, 2, \cdots, n.$$

检验的原假设为 H_0: $F_X = F_Y$，备择假设按习惯，分为双边假设和单边假设检验.

（一般假设两样本只是中位数有差别，因而 H_0: 中位数之差为零，H_1: 中位数之差不是零）.

符号秩检验的步骤如下：

1. 将原始数据分两列列出（行也可以）;
2. 一列计算数据之差 $D_i = X_i - Y_i$;（忽略差为零的配对，即相等的数据删去不用）
3. 一列求差的绝对值 D_i;
4. 一列对 $\{D_i\}$ 求秩;（秩相同的用平均秩）
5. 一列将 D_i 的符号赋给相应的秩得到符号秩;
6. 计算正符号秩之和 W_+、负符号秩之和的相反数 W_-，分别简称为正秩和、负秩和.（负秩和：符号为负的样本的秩和）
7. 将 $W = \min\{W_+, W_-\}$ 与临界值作比较.

下面是一个简单的例子.

符号秩和为

$$W_+ = 2 + 4 = 6.$$

为进行检验，需要计算 W_+ 的概率分布.

在原假设 H_0 成立的条件下，D_i 的分布关于 0 对称. 因此 D_i 的符号秩可以等

配对数据 1	配对数据 2	差	|差|	秩	符号秩
25	27	2	2	2	2
29	25	-4	4	3	-3
60	59	-1	1	1	-1
27	37	10	10	4	4

可能地取 2^n 个值 $\pm 1, \pm 2, \cdots, \pm n$ 之中的任何一个. 由此可以计算正秩和 W_+、负秩和 W_- 的精确分布，进而可以构造拒绝域.

例 8.3 设 $n = 3$，排除数据相等的情况，则 $|D_1|, |D_2|, |D_3|$ 的可能的秩 R_1, R_2, R_3 为 $\{1, 2, 3\}$ 的任意排列. 因为这些排列都是等可能的，因而我们可以只考虑秩为 $1, 2, 3$ 的情况. 此时，R_i 可以等可能地取正或负，因此，R_1, R_2, R_3 带上符号后有 $2^3 = 8$ 种情况，对每一种情况都可以计算出相应的 W_+，从而得到 W_+ 的概率分布.

符	1	1	1	1	-1	-1	-1	-1
号	2	2	-2	-2	2	2	-2	-2
秩	3	-3	3	-3	3	-3	3	-3
W_+	6	3	4	1	5	2	3	0

由此可得 W_+ 的分布列如下：

W_+	0	1	2	3	4	5	6
$\mathbb{P}(W_+)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

易知 W_+ 的最小值为 0，最大值为 $\sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$.（中间值 $\frac{1}{4}n(n+1)$ 恰好为 $\mathbb{E}W_+$.）

我们可以注意到 W_+ 的精确分布是对称的. 这样就可以使得我们取

$$W = \min\{W_+, W_-\}$$

为检验统计量.

在 H_0 成立的条件下，W 不应偏小或偏大.

注意到

$$W_+ + W_- = \frac{1}{2}n(n+1),$$

所以，只要得到了 W_+，就可以由

$$W_- = \frac{1}{2}n(n+1) - W_+$$

得到负秩和 W_-.

给定显著性水平 α. 拒绝域的形式为 $\{W \le W^*\}$，W^* 的选择依赖于是双边检验还是单边检验. 对双边检验，W^* 满足

$$\mathbb{P}(W \le W^*) \le \frac{\alpha}{2},$$

对单边检验，W^* 满足

$$\mathbb{P}(W \le W^*) \le \alpha.$$

例 8.4

血液配对检验，用符号秩检验.

试验前	试验后	差	差的 绝对值	秩	符号秩		
X_i	Y_i	D_i	|D_I|	R_i	sgn(D_i) R_i		
25	27	2	2	2	2	样本量n	11
25	29	4	4	3.5	3.5	W-	-1
27	37	10	10	6	6	W_+	
44	56	12	12	7	7	用n(n+1)/2-|W-|	65
30	46	16	16	10	10	用sumif计算	65
67	82	15	15	8.5	8.5	alpha	0.05
53	57	4	4	3.5	3.5	临界值	
53	80	27	27	11	11	双边检验临界值	11
52	61	9	9	5	5	单边检验临界值	7
60	59	-1	1	1	-1		
28	43	15	15	8.5	8.5		

图 8.3 符号秩检验

参考 Excel 文件“blood-before-after.xlxs”.

8.4.1 用近似正态分布

容易理解，在 H_0 成立的条件下，W_+ 与如下随机变量

$$\sum_{i=1}^{n} kB_k,$$

同分布，其中 $B_k \sim B(1, \frac{1}{2})$. 实际上，$W_+$ 有如下表示

$$W_+ = \sum_{i=1}^{n} kI_k,$$

其中

$$I_k = \begin{cases} 1, & \{|D_i|\}\text{中秩为}k\text{的}|D_i|\text{所对应的}D_i\text{的符号为正}, \\ 0, & \text{其他}, \end{cases}$$

且 $I_k \sim B(1, \frac{1}{2})$ 且相互独立.

定理 8.2

设 D_i 独立，以 0 为对称中心，

$$\mathbb{E}W_+ = \frac{n(n+1)}{4}, \quad \operatorname{Var} W_+ = \frac{n(n+1)(2n+1)}{24}.$$

证明 因为

$$\mathbb{E}I_k = \frac{1}{2}, \quad \operatorname{Var} I_k = \frac{1}{4},$$

所以

$$\mathbb{E}W_+ = \sum_{k=1}^{n} k\mathbb{E}I_k = \frac{1}{2}\sum_{i=1}^{n} k = \frac{1}{4}n(n+1),$$

$$\operatorname{Var} W_+ = \sum_{k=1}^{n} k^2 \operatorname{Var} I_k = \frac{1}{2}\sum_{i=1}^{n} k^2 = \frac{1}{24}n(n+1)(2n+1).$$

□

因为 W_+ 可以表示为独立随机变量之和，根据中心极限定理，在样本量足够大时，W_+ 近似服从正态分布，因而

$$Z := \frac{W_+ - \mathbb{E}W_+}{\sqrt{\operatorname{Var} W_+}}$$

近似服从标准正态分布.

注意到

$$-W_- - \mathbb{E}W_+ = \frac{n(n+1)}{2} - W_+ - \frac{n(n+1)}{2} = \frac{n(n+1)}{4} - W_+,$$

也就是说，

$$\frac{W_- - \mathbb{E}W_+}{\sqrt{\operatorname{Var} W_+}} = -\frac{W_+ - \mathbb{E}W_+}{\sqrt{\operatorname{Var} W_+}}.$$

例 8.5 鱼的汞含量的两种测量方法比较.

t 检验：

符号秩检验：用符号秩检验的临界值以及用正态分布近似两种方法.

参考 Excel 文件“Fish-Mercury-检验.xlsx”

数据序	选择性还原X	高锰酸钾Y	差D		
1	0.32	0.39	0.07	样本量	25
2	0.40	0.47	0.07	X的平均值	0.4172
3	0.11	0.11	0.00	Y的平均值	0.4576
4	0.47	0.43	-0.04	D的平均值	0.0404
5	0.32	0.42	0.10	D的标准差	0.1158
6	0.35	0.30	-0.05		
7	0.32	0.43	0.11	t统计量	1.7448
8	0.63	0.98	0.35	自由度	24
9	0.50	0.86	0.36	双边p值	0.0938
10	0.60	0.79	0.19	alpha	0.05
11	0.38	0.33	-0.05	临界值	2.0639
12	0.46	0.45	-0.01		
13	0.20	0.22	0.02	tinv算的是双边概率	
14	0.31	0.30	-0.01		
15	0.62	0.60	-0.02	结论	接受原假设
16	0.52	0.53	0.01		
17	0.77	0.85	0.08		
18	0.23	0.21	-0.02		
19	0.30	0.33	0.03		
20	0.70	0.57	-0.13		
21	0.41	0.43	0.02		
22	0.53	0.49	-0.04		
23	0.19	0.20	0.01		
24	0.31	0.35	0.04		
25	0.48	0.40	-0.08		

图 8.4

数据序	选择性还原X	高锰酸钾Y	差D	绝对值|D|	序数	绝对值的秩	符号秩		
12	0.46	0.45	-0.01	0.01	1	2.5	-2.5	去掉相同的值	
14	0.31	0.30	-0.01	0.01	2	2.5	-2.5		
16	0.52	0.53	0.01	0.01	3	2.5	2.5	样本量n	24
23	0.19	0.20	0.01	0.01	4	2.5	2.5	正秩和W_+	+194.5
13	0.20	0.22	0.02	0.02	5	6.5	6.5	负秩和W_-	-105.5
15	0.62	0.60	-0.02	0.02	6	6.5	-6.5	alpha	0.05
18	0.23	0.21	-0.02	0.02	7	6.5	-6.5	双边临界值	81
21	0.41	0.43	0.02	0.02	8	6.5	6.5	接受原假设	81<105.5
19	0.30	0.33	0.03	0.03	9	9	9		
4	0.47	0.43	-0.04	0.04	10	11	-11	E W_+	+150
22	0.53	0.49	-0.04	0.04	11	11	-11	Var W_+	+1225.0
24	0.31	0.35	0.04	0.04	12	11	11		
6	0.35	0.30	-0.05	0.05	13	13.5	-13.5	Z统计量	1.2714
11	0.38	0.33	-0.05	0.05	14	13.5	-13.5	双边检验p值	0.2036
1	0.32	0.39	0.07	0.07	15	15.5	15.5		
2	0.40	0.47	0.07	0.07	16	15.5	15.5		
17	0.77	0.85	0.08	0.08	17	17.5	17.5		
25	0.48	0.40	-0.08	0.08	18	17.5	-17.5		
5	0.32	0.42	0.10	0.10	19	19	19		
7	0.32	0.43	0.11	0.11	20	20	20		
20	0.70	0.57	-0.13	0.13	21	21	-21		
10	0.60	0.79	0.19	0.19	22	22	22		
8	0.63	0.98	0.35	0.35	23	23	23		
9	0.50	0.86	0.36	0.36	24	24	24		

图 8.5

附录：人物小传

Wilcoxon 是一位化学家. 1945 年，他发现了秩和检验，不知道是否是新方法，于是投稿到统计期刊. 1947 年，Mann 及其学生 Whitney 也发现了类似的检验方法.

例 8.6

内容提要

- 配对检验
- 符号秩检验
- 秩和检验

第 8 章 习题

1. 见上一章相关习题.

第 9 章 方差分析

在试验中，将要考察的指标称为试验指标，影响指标的条件称为因素. 因素所处的状态称为因素的水平. 只有一个因素在改变，称为单因素试验，如试验温度对面包烘烤的试验；否则称为多因素试验，如试验盐、糖对面包风味的影响.

我们通过分析方差的方法检验各水平之间是否有差异，所以实际上是分析均值.

常用 ANOVA，即方差分析（Analysis of Variance）的英文缩写来表示方差分析.

9.1 单因素方差分析

设某因素有 I 个水平，每个水平有 J 个样本，以 Y_{ij} 记在水平 i 下第 j 个观察值，于是

$$Y_{i1}, Y_{i1}, \cdots, Y_{iJ}$$

为水平 i 下的 J 个观察值.

记全体 IJ 个观察值的总的平均为

$$\overline{Y}_{\cdot\cdot} := \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij},$$

则这 IJ 个观察的样本方差为

$$MSS_{TOT} := \frac{1}{IJ-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{\cdot\cdot})^2.$$

称

$$SS_{TOT} := (IJ-1)MSS_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{\cdot\cdot})^2$$

为总的离差平方和.

记水平 i 下的 J 个观察值的平均值为

$$\overline{Y}_{i\cdot} := \frac{1}{J}\sum_{j=1}^{J} Y_{ij}.$$

从而水平 i 的 J 个观察值的样本方差为

$$S_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\cdot})^2.$$

称

$$SS_W := \sum_{i=1}^{I}(J-1)S_i^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\cdot})^2.$$

为组内离差平方和. 它反映的数据的差异来自于各水平下观察值的随机差异.

各水平下观察值的平均

$$\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, \cdots, \overline{Y}_{I\cdot}$$

的样本方差为

$$S_B^2 := \frac{1}{I-1}\sum_{i=1}^{I}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2.$$

定义各水平之间的离差平方和

$$SS_B = J(I-1)S_B^2 = J\sum_{i=1}^{I}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2.$$

这里用到如下事实：

$$\frac{1}{I}\sum_{i=1}^{I}\overline{Y}_{i\cdot} = \overline{Y}_{\cdot\cdot}.$$

SS 表示平方和（英文 Sum of Squares 的缩写），SS_{TOT} 表示总的离差平方和，SS_W 表示总的组内离差平方和，SS_B 表示总的组间平方和，其中 W 来自英文单词 Within 的首字母，B 来自英文单词 Between 的首字母.

	Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7			
	4.13	3.86	4	3.88	4.02	4.02	4		I	7.00000
	4.07	3.85	4.02	3.88	3.95	3.86	4.02		J	10.00000
	4.04	4.08	4.01	3.91	4.02	3.96	4.03			
	4.07	4.11	4.01	3.95	3.89	3.97	4.04			
	4.05	4.08	4.04	3.92	3.91	4	4.1			
	4.04	4.01	3.99	3.97	4.01	3.82	3.81		SS_B+SS_W	0.35614
	4.02	4.02	4.03	3.92	3.89	3.98	3.91			
	4.06	4.04	3.97	3.9	3.89	3.99	3.96		total mea	3.98457
	4.1	3.97	3.98	3.97	3.99	4.02	4.05		total var	0.00509
	4.04	3.95	3.98	3.9	4	3.93	4.06		SS_TOT	0.35614
	Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7		var	0.00178
mean	4.062	3.997	4.003	3.92	3.957	3.955	3.998		SS	0.01247
var	0.000956	0.007241	0.000481	0.001	0.002941	0.004045	0.006476		SS_B	0.12474
SS	0.00956	0.07241	0.00481	0.01	0.02941	0.04045	0.06476		SS_W	0.23140

图 9.1 离差平方和的计算

参考文件“Lab-Chap12AoV.xlsx”

下面的离差平方和分解很关键.

定理 9.1

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\cdot})^2 + J\sum_{i=1}^{I}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2.$$

即

$$SS_{TOT} = SS_W + SS_B.$$

证明 给定的 $i=1,2,\cdots,I$，对 $Y_{i1},Y_{i2},\cdots,Y_{iJ}$ 应用等式(**??**)（取 $a=\overline{Y}_{..}$），可得

$$\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2=\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2+J(\overline{Y}_{i.}-\overline{Y}_{..})^2,$$

这里利用了的平均为 $\overline{Y}_{i.}$. 对上述等式中的 i 从 1 到 I 进行求和即可完成证明.

□

注对上述离差平方和的分解的证明，常见的证明方法是在 $Y_{ij}-\overline{Y}_{..}$ 中减去 $\overline{Y}_{i.}$ 再加上 $\overline{Y}_{i.}$. 具体如下：

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2\\
=&\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(Y_{ij}-\overline{Y}_{i.})+(\overline{Y}_{i.}-\overline{Y}_{..})\right]^2\\
=&\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2+J\sum_{i=1}^{I}(\overline{Y}_{i.}-\overline{Y}_{..})^2+\sum_{i=1}^{I}\sum_{j=1}^{J}2(Y_{ij}-\overline{Y}_{i.})(\overline{Y}_{i.}-\overline{Y}_{..})\\
=&\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2+J\sum_{i=1}^{I}(\overline{Y}_{i.}-\overline{Y}_{..})^2,
\end{aligned}$$

其中用到交叉项之和为零，因根据(**??**)，有

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{j=1}^{J}2(Y_{ij}-\overline{Y}_{i.})(\overline{Y}_{i.}-\overline{Y}_{..})\\
=&2\sum_{i=1}^{I}\left[(\overline{Y}_{i.}-\overline{Y}_{..})\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})\right]\\
=&2\sum_{i=1}^{I}\left[(\overline{Y}_{i.}-\overline{Y}_{..})\left(\sum_{j=1}^{J}Y_{ij}-J\overline{Y}_{i.}\right)\right]\\
=&0.
\end{aligned}$$

假设数据满足如下模型

$$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij},\quad i=1,2,\cdots,I,j=1,2,\cdots,J,$$

其中 μ,α_i 为常数，μ 表示总的平均，α_i 表示水平 i 带来的影响，且

$$\sum_{i=1}^{I}\alpha_i=0.$$

如果 $\sum_{i=1}^{I}\alpha_i\neq 0$，我们可以调整 μ 的值以使得 $\sum_{i=1}^{I}\alpha_i=0$. 设 ε_{ij} 表示随机误差，且

$$\varepsilon_{ij}\sim N(0,1).$$

$$\mathbb{E}Y_{ij}=\mu+\alpha_i.$$

因此

$$\alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

表示各水平之间无差别. 因为假设 $\sum_{i=1}^{I} \alpha_i = 0$，可知所有的 α_i 如果相同，则只能为 0. 因此，我们考虑零假设

$$H_0\colon \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0.$$

备择假设为至少有 $\alpha_i \neq \alpha_j, i \neq j$：

$$H_0\colon \alpha_i \neq \alpha_j,\ i \neq j.$$

我们将要研究在零假设为真的条件下，SS_W, SS_B, SS_{TOT} 的分布.

为此，我们先介绍一个将要两次使用的结论.

引理 9.1

设 $X_i, i = 1, 2, \cdots, n$ 相互独立，$\mathbb{E}X_i = \mu_i, \operatorname{Var} X_i = \sigma^2$，则

$$\mathbb{E}(X_i - \overline{X})^2 = (\mu_i - \bar{\mu})^2 + \frac{n-1}{n}\sigma^2,$$

其中 $\bar{\mu} = \mathbb{E}\bar{X}$，即 $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$.

证明 注意到

$$\mathbb{E}(X_i - \overline{X})^2 = \operatorname{Var}(X_i - \overline{X}) + (\mathbb{E}(X_i - \overline{X}))^2,$$

而

$$\mathbb{E}(X_i - \overline{X}) = \mathbb{E}X_i - \mathbb{E}\overline{X} = \mu_i - \bar{u},$$

我们只要证明

$$\operatorname{Var}(X_i - \overline{X}) = \frac{n-1}{n}\sigma^2.$$

事实上，

$$\begin{aligned}
\operatorname{Var}(X_i - \overline{X}) &= \operatorname{Var} X_i + \operatorname{Var} \overline{X} - 2\operatorname{Cov}(X_i, \overline{X}) \\
&= \sigma^2 + \frac{1}{n}\sigma^2 - \frac{2}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2,
\end{aligned}$$

其中用到

$$\begin{aligned}
2\operatorname{Cov}(X_i, \overline{X}) &= 2\operatorname{Cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) \\
&= \frac{2}{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i, X_j) \\
&= \frac{2}{n}\operatorname{Cov}(X_i, X_i) \\
&= \frac{2}{n}\sigma^2.
\end{aligned}$$

□

定理 9.2. SS_W **的期望**

$$\mathbb{E}SS_W = I(J-1)\sigma^2.$$

证明 注意到

$$Y_{i1}, Y_{i1}, \cdots, Y_{iJ}$$

独立同分布，且

$$\mathbb{E}Y_{ij} = \mu + \alpha_i, \quad \mathbb{E}\bar{Y}_{i\cdot} = \mu + \alpha_i,$$

同时还有

$$\operatorname{Var} Y_{ij} = \sigma^2.$$

根据引理9.1，

$$\mathbb{E}(Y_{ij} - \bar{Y}_{i\cdot})^2 = \frac{J-1}{J}\sigma^2,$$

从而有

$$\begin{aligned}
\mathbb{E}SS_W &= \mathbb{E}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{E}(Y_{ij} - \overline{Y}_{i\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{J-1}{J}\sigma^2 \\
&= IJ \times \frac{J-1}{J}\sigma^2 \\
&= I(J-1)\sigma^2.
\end{aligned}$$

□

也可以如下快速计算.

定义联合样本方差为

$$S_p^2 = \frac{1}{I}\sum_{i=1}^{I} S_i^2.$$

易知

$$\begin{aligned}
S_p^2 &= \frac{1}{I}\sum_{i=1}^{I} S_i^2 \\
&= \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J-1}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\cdot})^2 \\
&= \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i\cdot})^2 \\
&= \frac{1}{I(J-1)}SS_W.
\end{aligned}$$

因为

$$\mathbb{E}S_i^2 = \sigma^2,$$

所以

$$\mathbb{E}\frac{SS_W}{I(J-1)} = \mathbb{E}S_p^2 = \frac{1}{I}\sum_{i=1}^{I}\mathbb{E}S_i^2 = \frac{1}{I}\times I\sigma^2 = \sigma^2.$$

由此可知

$$S_p^2 = \frac{SS_W}{I(J-1)}$$

是 σ^2 的无偏估计.

定理 9.3. SS_B 的期望

$$\mathbb{E}SS_B = J\sum_{i=1}^{n}\alpha_i^2 + (I-1)\sigma^2.$$

证明 由假设,

$$\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \cdots, \bar{Y}_{I\cdot}$$

相互独立,方差相同,为

$$\operatorname{Var}\bar{Y}_{i\cdot} = \frac{1}{J}\sigma^2,$$

但期望各不相同,

$$\mathbb{E}\bar{Y}_{i\cdot} = \mu + \alpha_i.$$

于是

$$\mathbb{E}\bar{Y}_{\cdot\cdot} = \frac{1}{I}\sum_{i=1}^{I}(\mu+\alpha_i) = \mu + \frac{1}{I}\sum_{i=1}^{I}\alpha_i = \mu.$$

所以,根据引理9.1,

$$\mathbb{E}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2 = (\mu+\alpha_i-\mu)^2\frac{I-1}{I}\cdot\frac{1}{J}\sigma^2 = \alpha_i^2 + \frac{I-1}{I}\cdot\frac{1}{J}\sigma^2.$$

从而有

$$\begin{aligned}
\mathbb{E}SS_B &= \mathbb{E}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{E}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\alpha_i^2+\frac{I-1}{I}\cdot\frac{1}{J}\sigma^2\right] \\
&= J\sum_{i=1}^{I}\alpha_i^2+IJ\times\frac{I-1}{I}\cdot\frac{1}{J}\sigma^2 \\
&= J\sum_{i=1}^{I}\alpha_i^2+(I-1)\sigma^2.
\end{aligned}$$

□

因此，当 H_0 成立时，

$$S_p^2=\frac{SS_W}{I(J-1)},\frac{SS_B}{I-1}$$

都是 σ^2 的无偏估计，从而两者大致相等. 而若有某 $\alpha_i \neq 0$，则

$$F := \frac{\frac{SS_W}{I(J-1)}}{\frac{SS_B}{I-1}}$$

偏大. 从而可以根据 F 的取值来检验 H_0.

定理 9.4

1. 假设误差服从均值为 0，方差为 σ^2 的正态分布，则 $\frac{SS_W}{\sigma^2}$ 服从自由度为 $I(J-1)$ 的 χ^2 分布，
2. 假设误差服从均值为 0，方差为 σ^2 的正态分布，且 H_0 成立，则 $\frac{SS_B}{\sigma^2}$ 服从自由度为 $I-1$ 的 χ^2 分布.
3. 假设误差服从均值为 0，方差为 σ^2 的正态分布，且 H_0 成立，则 SS_B 与 SS_W 独立，且

$$F=\frac{SS_B/(I-1)}{SS_W/I(J-1)}\sim F(I-1,I(J-1)).$$

证明

1. 注意到对任意给定的 $i=1,2,\cdots,I$，$Y_{i1},Y_{i2},\cdots,Y_{iJ}$ 独立同分布与 $N(\mu+\alpha_i,\sigma^2)$，因此，按基本定理，

$$\xi_i := \frac{1}{\sigma^2}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i\cdot})^2\sim\chi^2(J-1).$$

因为 $\xi_1, \xi_2, \cdots, x_I$ 是相互独立的，由 χ^2 分布的可加性可得

$$SS_W = \sum_{i=1}^{I} \xi_i \sim \chi^2(I(J-1)).$$

2. 由于对任意的 $i = 1, 2, \cdots, I$，$\alpha_i = 0$，可得知 $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \cdots, \bar{Y}_{I\cdot}$ 相互独立，且分布均为 $N(\mu, \frac{1}{J}\sigma^2)$. 因此，

$$\frac{1}{\sigma^2} SS_B = \frac{J}{\sigma^2} \sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \frac{\sum_{i=1}^{I} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\frac{1}{J}\sigma^2} \sim \chi^2(I-1).$$

3. 因为对来自正态分布的样本，样本均值与样本方差独立，所以 $\bar{Y}_{i\cdot}$ 与 $\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2$ 独立. 又，若 $k \neq i$，则因 $\bar{Y}_{i\cdot}$ 是样本 $Y_{i1}, Y_{i2}, \cdots, Y_{iJ}$ 的函数，$\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\cdot})^2$ 是样本 $Y_{k1}, Y_{k2}, \cdots, Y_{kJ}$ 的函数，而这两个样本是相互独立的，从而可知 $\bar{Y}_{i\cdot}$ 与 $\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\cdot})^2$ 独立. 由此可知 $\bar{Y}_{i\cdot}$ 的函数 $SS_B = J\sum_{i=1}^{I}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$ 与 $SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\cdot})^2$ 独立.

由于 SS_B 与 SS_W 相互独立，且由本定理前两个结论可知，在 H_0 成立的条件下，

$$F = \frac{SS_B/(I-1)}{SS_W/I(J-1)} = \frac{SS_B/[\sigma^2(I-1)]}{SS_W/[\sigma^2 I(J-1)]} \sim F(I-1, I(J-1)).$$

□

例 9.1

Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
4. 13	3. 86	4	3. 88	4. 02	4. 02	4
4. 07	3. 85	4. 02	3. 88	3. 95	3. 86	4. 02
4. 04	4. 08	4. 01	3. 91	4. 02	3. 96	4. 03
4. 07	4. 11	4. 01	3. 95	3. 89	3. 97	4. 04
4. 05	4. 08	4. 04	3. 92	3. 91	4	4. 1
4. 04	4. 01	3. 99	3. 97	4. 01	3. 82	3. 81
4. 02	4. 02	4. 03	3. 92	3. 89	3. 98	3. 91
4. 06	4. 04	3. 97	3. 9	3. 89	3. 99	3. 96
4. 1	3. 97	3. 98	3. 97	3. 99	4. 02	4. 05
4. 04	3. 95	3. 98	3. 9	4	3. 93	4. 06

图 9.2　7 个实验室

方差分析表

来源	离差平方和	自由度	平均离差平方和	F 值
组间	0.1247	6	0.02079	5.6601
组内	0.2314	63	0.00367	
总	0.3561	69		

p 值为 0.0001，对显著性水平 $\alpha = 0.01$，有显著性差异. 或：临界值 $F_0.05(6, 63) =$

$3.10277 < 5.6601$，所以有显著性差异.

9.2 多重比较

在单因素方差分析中，零假设是各水平的均值相同，备择假设为至少有两个水平的均值有差异. 因此，方差分析可以指出各水平之间是否有显著差异，但当各水平的均值之间有显著差异时，并没有告诉我们具体是哪两个或多个水平之间的均值有差异.

为知道具体是哪两个或多个水平下均值有差异，一个办法是进行多个检验，如对 n 个水平下的任何两个水平进行 t 检验. 但这样就要进行 $\binom{n}{2} = \frac{1}{2}n(n-1)$ 个检验. 同时进行多个检验，不但费时，且容易出现所谓的“多重比较谬误”. 如进行 m 个独立检验，原假设为 $H_{0k}, k = 1, 2, \cdots, m$，每个检验犯第一类错误的概率按习惯控制为 α，即

$$\mathbb{P}(\text{拒绝}H_{0k}|H_{0k}\text{为真}) = 0.05, \quad k = 1, 2, \cdots, m.$$

如果将这 m 个检验作为一个整体考虑，联合显著水平为

$$\begin{aligned}&\mathbb{P}(\text{拒绝某}H_{0k}, k = 1, 2, \cdots, m|H_{01}, H_{02}, \cdots H_{0m}\text{为真})\\=&1 - \mathbb{P}(\text{接受所有}H_{0k}, k = 1, 2, \cdots, m|H_{01}, H_{02}, \cdots H_{0m}\text{为真}) = \quad 1 - (1-\alpha)^m.\end{aligned}$$

由此可见，当 $\alpha > 0$ 且 m 较大时，$1-(1-\alpha)^m$ 接近 1，如 $1-(1-0.05)^{100} \approx 0.994$. 这说明，在原假设全为真的条件下，发生一次拒绝某 H_{0k} 的错误的概率是较高的.

这里的关键原因在于 n 个水平下需要进行 $\frac{1}{2}n(n-1)$ 个两两配对的检验，$\frac{1}{2}n(n-1)$ 这个数目往往比较大.

所谓的生日悖论与此类似：在 23 人中，至少两人生日相同的概率大于 $\frac{1}{2}$，就是因为这 23 人两两配对比较生日，需要进行 $\binom{23}{2} = 253$ 个配对. 在 253 个配对中，每个配对生日相同的概率确实比较小，为 $\frac{1}{365}$. 但这 253 个配对中至少有一个配对有相同生日的概率却为

$$1 - \left(1 - \frac{1}{365}\right)^{253}.$$

9.2.1 图基（Tukey）法

图基法是约翰·图基（John Tukey，1915-2000）发现的，图基是快速傅里叶变换、刀切法、箱形图、投影寻踪法等方法的提出者.

我们知道置信区间可以用于假设检验. 图基法即基于置信区间. 图基法用到所谓的学生化极差分布. 所谓学生化是指用样本标准差代替总体标准差的过程, 如有服从正态分布的随机变量 $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ 得到服从 t 分布的随机变量: $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$.

定义 9.1. 学生化极差分布

设 $X_1, X_2, \cdots, X_n$ 为来自标准正态分布的样本, U 与之独立且服从自由度为 k 的 χ^2 分布, 则称

$$q_{n,k} = \frac{X_{(n)} - X_{(1)}}{\sqrt{U/k}} = \max_{i,j=1,2,\cdots n} \frac{X_i - X_j}{\sqrt{U/k}}$$

服从组数为 n, 自由度为 k 的学生化极差分布, 或称之为参数为 n, k 的学生化极差分布, 记为

$$q_{n,k} \sim t(n, k).$$

现考虑相互独立服从正态分布的随机变量

$$Y_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \cdots, I, j = 1, 2, \cdots, J,$$

则有独立同分布随机变量列

$$\bar{Y}_{i\cdot} - \mu_i \sim N\left(0, \frac{\sigma^2}{J}\right), \quad i = 1, 2, \cdots, I.$$

或说

$$\frac{\sqrt{J}(\bar{Y}_{i\cdot} - \mu_i)}{\sigma} \sim N(0, 1), \quad i = 1, 2, \cdots, I.$$

显然, 联合样本方差

$$S_p^2 = \frac{SS_W^2}{I(J-1)}$$

是 σ^2 的估计. 且由定理 ** 可知

$$\frac{I(J-1)S_p^2}{\sigma^2} \sim \chi^2(I(J-1)).$$

于是, 按学生化极差分布的定义, 可知

定理 9.5. 学生化极差分布

$$q := \max_{i_1, i_2} \frac{\sqrt{J}\left|(\bar{Y}_{i_1\cdot} - \mu_{i_1}) - (\bar{Y}_{i_2\cdot} - \mu_{i_2})\right|}{S_p} \sim t(I, I(J-1)).$$

记上 100α 分位点为 $q_{I,I(J-1)}(\alpha)$，即

$$\begin{aligned}&\mathbb{P}\left(\text{对任意} i_1,i_2,\left|(\bar{Y}_{i_1\cdot}-\mu_{i_1})-(\bar{Y}_{i_2\cdot}-\mu_{i_2})\right|\le q_{I,I(J-1)}(\alpha)\frac{S_p}{\sqrt{J}}\right)\\ =&\mathbb{P}\left(\max_{i_1,i_2}\left|(\bar{Y}_{i_1\cdot}-\mu_{i_1})-(\bar{Y}_{i_2\cdot}-\mu_{i_2})\right|\le q_{I,I(J-1)}(\alpha)\frac{S_p}{\sqrt{J}}\right)\\ =&1-\alpha.\end{aligned}$$

因此，$\mu_{i_1}-\mu_{i_2}$ 的置信水平为 $100(1-\alpha)\%$ 的置信区间为

$$(\bar{Y}_{i_1\cdot}-\bar{Y}_{i_2\cdot})\pm q_{I,I(J-1)}(\alpha)\frac{S_p}{\sqrt{J}}.$$

由置信区间与假设检验的对偶性，在显著性水平 α 下，若 0 不属于上述置信区间，即区间的两端点的符号相同，或说观察值满足[1]

$$|\bar{y}_{i_1\cdot}-\bar{y}_{i_2\cdot}|>q_{I,I(J-1)}(\alpha)\frac{S_p}{\sqrt{J}}$$

则拒绝原假设.

例 9.2 在单因素方差分析的 7 个实验室例子中，

$$s_p=0.06,\quad q_{7,63}(0.05)=4.31,\quad q_{7,63}(0.05)\frac{s_p}{\sqrt{J}}=0.082$$

各实验室的平均值为

实验室	Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
平均值	4.062	3.997	4.003	3.92	3.957	3.955	3.998

可见 Lab1 与 Lab4 的平均值的差为 $0.142>0.082$，可见这两个实验室之间有差别，而 Lab6 和 Lab7 之间的平均值差为 3.998 − 3.995 = 0.003 < 0.082，可见这两个实验室没差别.

两样本比较的 t 检验可以与上面的方法做比较.

注意到 $Y_{ij}\sim N(\mu_i,\sigma^2)$，因此

$$\bar{Y}_{i\cdot}=\frac{1}{J}\sum_{i=1}^{J}Y_{ij}\sim N\left(\mu_i,\frac{1}{J}\sigma^2\right)$$

类似地，

$$\bar{Y}_{k\cdot}\sim N\left(\mu_k,\frac{1}{J}\sigma^2\right),$$

因为 $\bar{Y}_{i\cdot},\bar{Y}_{k\cdot}$ 相互独立，可知

$$\bar{Y}_{i\cdot}-\bar{Y}_{k\cdot}\sim N\left(\mu_i-\mu_k,\frac{2}{J}\sigma^2\right)$$

[1] $a-b,a+b$ 符号相同当且仅当 $(a+b)(a-b)=(a^2-b^2)>0$，又等价于 $|a|>|b|$.

因此

$$\frac{(\overline{Y}_{i\cdot}-\overline{Y}_{k\cdot})-(\mu_i-\mu_k)}{\sqrt{\frac{2}{J}}\sigma}\sim N(0,1).$$

用联合样本标准差 S_p 代替 σ，可得

$$\frac{(\overline{Y}_{i\cdot}-\overline{Y}_{k\cdot})-(\mu_i-\mu_k)}{\sqrt{\frac{2}{J}}S_p}\sim t(I(J-1)).$$

在我们的问题中，$I=7, J=10$，于是自由度为 $I(J-1)=63$，$\sqrt{\frac{2}{J}}=\sqrt{\frac{2}{10}}=\frac{1}{\sqrt{5}}$.

如果仅两两作 t 检验，只要差大于

$$t_{\alpha/2}(63)\sqrt{\frac{2}{J}}s_p=t_{0.025}(63)\frac{1}{\sqrt{5}}0.06=0.053$$

即认为有差别.

前者比后者大，是因为前者是同时检验. 但两者差别不大，说明前者的方法比较好.

9.2.2 Bonferroni 方法

这种方法是将显著性水平调低以得到共同的检验.

邦弗罗尼法有美国女统计学家 Olive Jean Dunn（1915-2000）提出，以意大利数学家 Bonferroni 的名字命名，因该方法用到简单的 Bonferroni 不等式.

参考第**??**节中有关 Bonferroni 不等式的介绍.

由

$$\mathbb{P}\left(\cup_{i=1}^{n}A_i\right)\leq\sum_{i=1}^{n}\mathbb{P}(A_i). \tag{9.1}$$

设有 k 个零假设 $H_{0i}, i=1,2,\cdots,k$ 待检验，检验每个零假设的显著性水平为 α/k，设 $R_i, i=1,2,\cdots,k$ 表示第 i 个零假设被拒绝，则

$$\mathbb{P}(R_i|H_{0i})\leq\frac{\alpha}{k},\quad i=1,2,\cdots,k.$$

则检验 k 个零假设同时成立的显著性水平为这 k 个零假设同时成立的条件下，k 个零假设之一被拒绝的概率，则由(9.1)可得

$$\mathbb{P}(\cup_{i=1}^{k}R_i|H_{0i},i=1,2,\cdots,k)\leq\sum_{i=1}^{k}\mathbb{P}(R_i|H_{0i},i=1,2,\cdots,k)\leq k\cdot\frac{\alpha}{k}=\alpha.$$

它等价于说，若有 k 个置信水平为 $100(1-\alpha/k)\%$ 的置信区间，则它们也是置信水平为 $100(1-\alpha)\%$ 的同时置信区间.

例 9.3

仍然考虑前面的问题. 因为 $\binom{7}{2} = 21$，置信水平为 95% 的同时置信区间为

$$\mu_{i1} - \mu_{i2} \in (\bar{y}_{i_1\cdot} - \bar{y}_{i_2\cdot}) \pm t_{0.025/21}(63)\frac{s_p}{\sqrt{5}}$$

其中，$\frac{1}{\sqrt{5}}$ 来自 $\sqrt{\frac{2}{J}} = \sqrt{\frac{2}{10}}$，$t_{0.025/21}(63)$ 的值可以用 excel 计算或查表 $t_{0.025/20}(60) = 3.16$ 得到近似值. 于是置信区间为

$$(\bar{y}_{i_1\cdot} - \bar{y}_{i_2\cdot}) \pm 0.085.$$

9.2.3 克鲁斯卡尔-沃利斯（Kruskal-Wallis）检验

克鲁斯卡尔-沃利斯（Kruskal-Wallis）检验是曼恩-惠特尼检验的推广. 该方法假设样本是独立的，但不需要假设样本服从正态分布.

设有 I 个水平，第 i 个水平下有 J_i 个观测值，即 Y_{ij} 为水平 i 下的第 j 个观测值. 记 R_{ij} 为组合样本中 Y_{ij} 的秩，第 i 组样本的平均秩为

$$\overline{R}_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} R_{ij},$$

总的平均秩为

$$\overline{R}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2},$$

其中 $N = \sum_{i=1}^{I} J_i$ 为总观测数.

$$SS_B = \sum_{i=1}^{I} J_i(\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2.$$

检验问题：各水平是否相同. 检验方法：SS_B 相对较大，则拒绝零假设.

查表 Lehman（1975）

$$I = 3, \quad J_i > 5$$

或

$$I > 3, \quad J_i \geq 4$$

标准化的 SS_B 近似服从 χ^2 分布：

$$K := \frac{12}{N(N+1)} SS_B \overset{\text{近似}}{\sim} \chi^2(I-1).$$

K 也可以表示为

$$K := \frac{12}{N(N+1)}\left(\sum_{i=1}^{I} J_i \overline{R}_{i\cdot}^2\right) - 3(N+1).$$

例 9.4 $K = 29.5$, 自由度为 6，p 值比 0.005 小.

例 9.5 红楼梦作者归属

将全书分为三部分（1–40 回，41–80 回，81–120 回）作对比. 用各部分之间的词频作为检验对象. 前两部分的检验如果没有差异，说明这个方法是可靠的，如果后两部分有差异，则说明前 80 回与后 40 回有显著差异.

9.3 两因素方差分析

某些总体可能不止受到一个因素的影响，如农作物的产量，可能与气温、降水量有关，也可能与土壤营养有关. 下面研究两个因素影响.

设某试验设计两个因素 A, B，因素 A 有 I 个水平，因素 B 有 J 个水平，共有 $I \times J$ 个因素组合，每个组合因素下有 K 个观测.

令 Y_{ijk} 表示单元 ij 的第 k 个观测，设

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk},$$

其中 $\alpha_i, \beta_j, \delta_{ij}$ 为常数，满足

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{i=1}^{I} \delta_{ij} = \sum_{j=1}^{J} \delta_{ij} = 0,$$

且

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

为随机变量，相互独立同分布.

样本 $\{Y_{ijk}: i = 1, 2, \cdots, I, j = 1, 2, \cdots, J, k = 1, 2, \cdots, K\}$ 的对数自然函数为

$$l = -\frac{1}{2} IJK \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \mu - \alpha_i - \beta_j - \delta_{ij})^2.$$

由此易由最大似然估计得到各参数的估计如下：

$$\hat{\mu} = \bar{Y}_{\dots},$$

$$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{\dots},$$

$$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{\dots},$$

$$\hat{\delta}_{ijk} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{\dots}.$$

记

$$SS_A = JK\sum_{i=1}^{I}\hat{\alpha}_i^2,$$

$$SS_B = IK\sum_{j=1}^{J}\hat{\beta}_i^2,$$

$$SS_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\delta}_{ij}^2,$$

$$SS_E = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij.})^2,$$

$$SS_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{...})^2,$$

则类似于单因素方差分析中离差平方和分解，对双因素，也有如下离差平方和分解：

$$SS_{TOT} = SS_A + SS_B + SS_{AB} + SS_E.$$

这些量的函数在合适的条件下具有明确的分布，因此可以进行假设检验. 方差分析表如下：

方差分析表

来源	离差平方和	自由度	平均离差平方和	F 值
因素 A	SS_A	$I-1$	$MS_A = \frac{SS_A}{I-1}$	$\frac{MS_A}{MS_E}$
因素 B	SS_B	$J-1$	$MS_B = \frac{SS_B}{J-1}$	$\frac{MS_B}{MS_E}$
A,B 交互	SS_{AB}	$(I-1)(J-1)$	$MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}$	$\frac{MS_{AB}}{MS_E}$
误差	SS_E	$IJ(K-1)$	$MS_E = \frac{SS_E}{[IJ(K-1)]}$	
总	SS_{TOT}	$IJK-1$		

附录

内容提要

- ❑ 方差分析
- ❑ 水平
- ❑ 单因素方差分析
- ❑ 多因素方差分析
- ❑ 组内离差平方和
- ❑ 组间离差平方和
- ❑ 多重比较
- ❑ 图基法
- ❑ Bonferroni 法
- ❑ Kruskal-Wallis 检验

第 9 章 习题

1.

第 10 章 附录：统计分布

10.1 Gamma 分布

Gamma 分布和很多分布密切相关.

定义 10.1. Gamma 分布

设 $\alpha > 0, \beta \in \mathbb{R}$，如果随机变量 X 的概率密度函数如下

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

则称 X 服从参数为 α, β 的 Gamma 分布，记为

$$X \sim \mathrm{Gamma}(\alpha, \beta).$$

命题 10.1. Gamma 分布的期望和方差

设 $X \sim \mathrm{Gamma}(\alpha, \beta)$，则

$$\mathbb{E}X = \frac{\alpha}{\beta}, \quad \mathrm{Var}\, X = \frac{\alpha}{\beta^2}$$

证明

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

类似地，可以计算得到

$$\begin{aligned}
\mathbb{E}X^2 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\beta^{(}\alpha+2)} = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\beta^2} = \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\Gamma(\alpha)\beta^2} \\
&= \frac{\alpha^2 + \alpha}{\beta^2}.
\end{aligned}$$

所以，

$$\mathrm{Var}\, X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{\alpha}{\beta^2}.$$

□

命题 10.2. Gamma 分布的尺度变换

设 $X \sim \mathrm{Gamma}(\alpha, \beta)$，$c > 0$，则 $cX \sim \mathrm{Gamma}(\alpha, \beta/c)$.

命题 10.3. Gamma 分布的可加性

设 $X_i \sim \mathrm{Gamma}(\alpha_i, \beta)$，$i = 1, 2, \cdots, n$，且 $X_1, X_2, \cdots, X_n$ 相互独立，则

$$\sum_{i=1}^{n} X_i \sim \mathrm{Gamma}\left(\sum_{i=1}^{n} \alpha_i, \beta\right).$$

一些常见分布是 Gamma 分布的的特殊情形. 例如，参数为 $1,\beta$ 的 Gamma 分布是参数为 β 的指数分布，参数为 $\frac{n}{2}, \frac{1}{2}$ 的 Gamma 分布为自由度为 n 的 χ^2 分布.

10.2 Beta 分布

定义 10.2. Beta 分布

设 $a, b > 0$，如果随机变量 X 的概率密度函数为

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

则称 X 服从参数为 a, b 的 Beta 分布，记为 $X \sim \mathrm{Beta}(a, b)$.

定义中我们用到 Beta 函数：

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

Beta 函数与 Gamma 函数有如下关系.

引理 10.1. Beta 函数与 Gamma 函数的关系

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a, b)}. \tag{10.1}$$

证明

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty t^{a-1} \mathrm{e}^{-t}\, dt \int_0^\infty s^{a-1} \mathrm{e}^{-s}\, ds \\ &= \int_0^\infty \int_0^\infty t^{a-1} s^{b-1} \mathrm{e}^{-(t+s)}\, dtds.\end{aligned}$$

考虑变量代换：$y = t + s$，以及 $x = \frac{t}{t+s}$. 此时，$t = x(t+s), s = (1-x)(t+s)$，且

$t = xy, s = (1-x)y$. 所以，变换的 Jacobian 为

$$\begin{vmatrix} \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \\ \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \end{vmatrix} = \begin{vmatrix} y & x \\ -y & 1-x \end{vmatrix} = y(1-x) + xy = y.$$

因此，

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^1 \int_0^\infty x^{a-1}(1-x)^{b-1} y^{a+b-2} \mathrm{e}^{-y} y dx dy \\ &= \int_0^1 x^{a-1}(1-x)^{b-1} dx \int_0^\infty y^{a+b-1} \mathrm{e}^{-y} dy \\ &= B(a,b)\Gamma(a+b). \end{aligned}$$

□

命题 10.4. Beta 分布的期望和方差

设 $X \sim \mathrm{Beta}(a,b)$，则

$$\mathbb{E}X = \frac{a}{a+b}, \quad \mathrm{Var}\, X = \frac{ab}{(a+b)^2(a+b+1)}.$$

证明

$$\begin{aligned} \mathbb{E}X &= \frac{1}{B(a,b)} \int_0^1 x \cdot x^{a-1}(1-x)^{b-1} dx \\ &= \frac{1}{B(a,b)} \int_0^1 x^{a+1-1}(1-x)^{b-1} dx \\ &= \frac{B(a+1,b)}{B(a,b)} \\ &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\ &= \frac{a}{a+b}. \end{aligned}$$

为计算方程，先计算二阶矩：

$$\begin{aligned}
\mathbb{E}X^2 &= \frac{1}{B(a,b)}\int_0^1 x^2\cdot x^{a-1}(1-x)^{b-1}dx\\
&= \frac{1}{B(a,b)}\int_0^1 x^{a+2-1}(1-x)^{b-1}dx\\
&= \frac{B(a+2,b)}{B(a,b)}\\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\\
&= \frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\\
&= \frac{a(a+1)}{(a+b)(a+b+1)}.
\end{aligned}$$

所以，

$$\begin{aligned}
\operatorname{Var} X &= \mathbb{E}X^2-(\mathbb{E}X)^2\\
&= \frac{a(a+1)}{(a+b)(a+b+1)}-\frac{a^2}{(a+b)^2}\\
&= \frac{a(a+1)(a+b)-a^2(a+b+1)}{(a+b)^2(a+b+1)}\\
&= \frac{ab}{(a+b)^2(a+b+1)}.
\end{aligned}$$

□

命题 10.5. Gamma 分布于 Beta 分布的关系

设 $X\sim\text{Gamma}(a,1)$，$Y\sim\text{Gamma}(b,1)$，且 X,Y 独立，则

$$\frac{X}{X+Y}\sim\text{Beta}(a,b).$$

注意到 $a=b=1$ 时，$\text{Beta}(1,1)=U(0,1)$.

我们称参数为 $a,1$ 的 Beta 分布为参数为 a 的 Beta 分布.

参考文献

[1] CRAMéR W H, H. Some theorems on distribution functions[J]. J. London Math. Soc.

[2] Errors of routine analysis[J]. Biometrika, 19.

索　　引

人名索引

符号索引