概率论与数理统计

Lecture 7

统计学基本概念与抽样分布

曾靖

中国科学技术大学统计与金融系



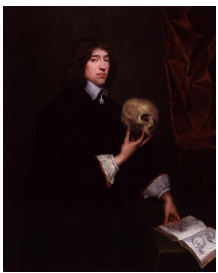
目录

5.1 统计学发展简史

5.2 数理统计的若干基本概念

5.3 统计量与抽样分布

统计学的起源



威廉 · 佩蒂
William Petty
1623–1687

- **从“国情学”出发：**早期记录人口、钱粮和土地等情况，为公共决策提供依据。
- **1662 年，格朗特：**整理伦敦死亡公报，从日常记录中分析人口与死亡的规律。
- **佩蒂的“政治算术”：**用统计数字分析政治、经济和社会问题，让事实参与论证。

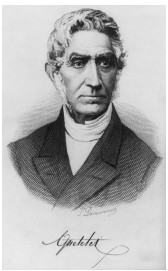
概率论与测量误差



高斯
Carl Friedrich Gauss
1777–1855

- **天文与地理测量：**同一对象的多次测量并不一致，怎样综合这些带误差的数据?
- **勒让德与高斯：**最小二乘法和正态误差理论，把测量问题与数学方法联系起来。
- **概率论提供语言：**伯努利、贝叶斯、拉普拉斯等人的工作，为描述随机性与不确定性奠定基础。

社会调查与随机抽样



凯特勒
Adolphe Quetelet
1796–1874

- **凯特勒：**将正态分布用于连续数据的分析，并提出“普通人”的概念，关注群体中的规律。
- **从普查到抽样：**拉普拉斯尝试用部分资料估计人口；凯尔发展“代表性抽样”思想。
- **鲍莱引入随机抽样：**抽取哪些个体，也成为统计方法需要研究的问题。

小样本与统计推断



戈塞特（Student）
William Sealy Gosset
1876–1937

- **高尔顿：**相关与回归思想的发展，使变量之间的关系成为研究对象。
- **K. 皮尔逊：**发展卡方检验，用数据检验模型与科学假设；1911 年创办大学统计系。
- **戈塞特：**以 Student 为笔名研究 t 分布，开创小样本推断的重要方法。

现代统计学的发展



费希尔
Ronald A. Fisher
1890–1962

- **费希尔：**从农业田间试验出发，发展试验设计、方差分析与似然方法。
- **奈曼与 E. 皮尔逊：**建立假设检验理论；随后瓦尔德发展序贯分析和统计决策理论。
- **图基：**推动探索性数据分析，强调借助图形发现数据中的结构与规律。

1970 年代以后：计算与数据驱动的发展

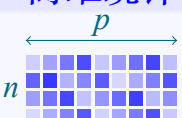
计算机的发展，使统计学走向更大规模、更复杂的数据。

大数据



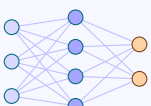
海量、多源数据

高维统计



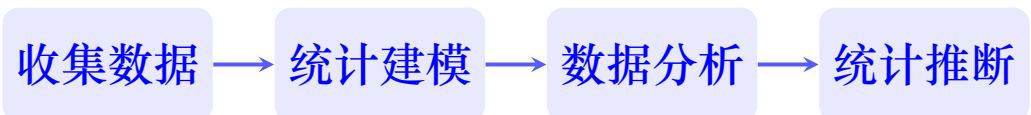
变量多，p 可超过 n

人工智能



学习与预测

数理统计学：从数据到结论



总体、个体与样本

例 1

为了解一批 10000 件产品的废品率，从中抽取 100 件进行检查。

- 研究对象的全体称为**总体**，每件产品是一个**个体**。
- 抽取的 100 件产品构成**样本**，样本量为 $n = 100$。

总体的概率描述

定义 5.2　总体

总体是所研究的某个指标的**取值全体**，以及**取这些值的概率分布**。通常用随机变量 X 来描述。

- 产品是否为废品：$X \sim B(1, p)$。
- 身高：可考虑 $X \sim N(\mu, \sigma^2)$。

样本与样本观察值

定义 5.3　样本

从总体中按一定的方式抽取的 n 个个体 $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$，称为是样本量（sample size）为 n 的一个样本。

- **抽样之前**：样本是随机向量 $(X_1, \ldots, X_n)$。
- **抽样之后**：得到具体的观察值 $(x_1, \ldots, x_n)$。
- 用**大写字母**表示随机变量，**小写字母**表示观察值。

简单随机样本

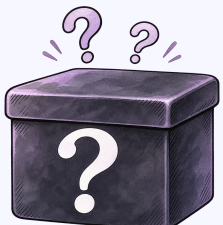
定义 5.4　简单随机样本

若 $X_1, \ldots, X_n$ 相互独立，且都与总体 X 同分布，则称其为来自该总体的**简单随机样本**。

- **同分布（identical distribution）**：每个 X_i 都服从总体分布 F。
- **独立性（independence）**：样本分量相互独立。
- 简记为 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$，其中 i.i.d. 是 independently and identically distributed 的缩写。

样本的二重性

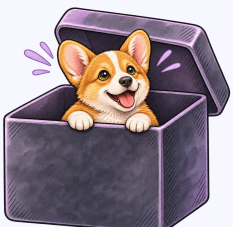
抽样之前：随机向量



$(X_1, \ldots, X_n)$

取值尚未确定，用于研究统计方法的性质。

抽样之后：样本值



$(x_1, \ldots, x_n)$

得到一组确定的数，可直接用于计算。

样本分布

- 样本 $(X_1, \ldots, X_n)$ 的联合分布称为**样本分布**。
- 对连续总体的简单随机样本，若总体密度为 f，则联合密度为

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

统计量

定义 5.5　统计量（statistic）

设 $X_1, \ldots, X_n$ 为样本。不含未知参数的样本函数

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

称为**统计量**。

- 统计量的表达式**不能含未知参数**。
- 统计量可以**直接从样本中计算得到**。

常见统计量：样本均值与样本方差

样本均值（sample mean）

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

反映总体均值的信息。

样本方差（sample variance）

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

反映总体方差的信息。$S = \sqrt{S^2}$ 称为样本标准差。

常见统计量：样本矩、偏度与峰度

样本矩（sample moment）

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

a_k 为样本 k 阶原点矩，m_k 为样本 k 阶中心矩；$a_1 = \overline{X}$。

样本偏度与峰度系数（$m_2 > 0$）

$$\widehat{\beta}_1 = \frac{m_3}{m_2^{3/2}}, \qquad \widehat{\beta}_2 = \frac{m_4}{m_2^2}.$$

分别反映总体偏度与峰度的信息。

常见统计量：样本相关系数

样本相关系数（sample coefficient of correlation）

设 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 为从二维总体中抽取的样本，则

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

- 也称为皮尔逊相关系数，反映总体相关系数的信息。

常见统计量：次序统计量

次序统计量（order statistics）：将样本按大小排列，

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

样本中位数（sample median）

$$m_n = \begin{cases} X_{((n+1)/2)}, & n \text{ 为奇数}, \\ \frac{1}{2}\left[X_{(n/2)} + X_{(n/2+1)}\right], & n \text{ 为偶数}. \end{cases}$$

反映总体中位数的信息。

极值：$X_{(1)}$ 和 $X_{(n)}$ 分别为样本极小值与极大值。

Pop Quiz 1：哪些是统计量？

设 $X_1, \ldots, X_n$ 来自 $N(\mu, \sigma^2)$，μ, σ^2 均未知。

问题 1（多选） 下列哪些是统计量？

(A) $\overline{X}$。

(B) $\overline{X} - \mu$。

(C) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$。

(D) S^2/σ^2。

问题 2（简答） 统计量的分布能否含有未知参数？举例说明。

Pop Quiz 1：答案

问题 1　**(A)、(C)**。

- 二者都能仅根据样本计算。
- (B)、(D) 的表达式含有未知参数，不是统计量。

问题 2　**可以**。例如 $\overline{X}$ 是统计量，但

$$\overline{X} \sim N(\mu, \sigma^2/n),$$

其分布依赖未知的 μ 和 σ^2。

抽样分布

定义 5.6　抽样分布

统计量 $T = T(X_1, \ldots, X_n)$ 的概率分布称为该统计量的**抽样分布**。

设 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$。

- **样本分布：** 随机向量 $(X_1, \ldots, X_n)$ 的联合分布，描述整个样本的取值规律。
- **抽样分布：** 统计量的分布。例如样本均值满足

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

这就是样本均值的抽样分布。

卡方分布

定义 5.7　χ^2 分布

设 $X_1, \ldots, X_n$ 相互独立且都服从 $N(0,1)$，则

$$X = \sum_{i=1}^{n} X_i^2$$

服从自由度为 n 的**卡方分布**，记为 $X \sim \chi_n^2$。

- 构造条件：**标准正态、相互独立**。
- 本讲的自由度 n 为正整数，$X \geq 0$。
- 常用于**正态总体方差的检验**，以及**拟合优度与列联表检验**。

卡方分布的性质

- 若 $X \sim \chi^2_n$，则

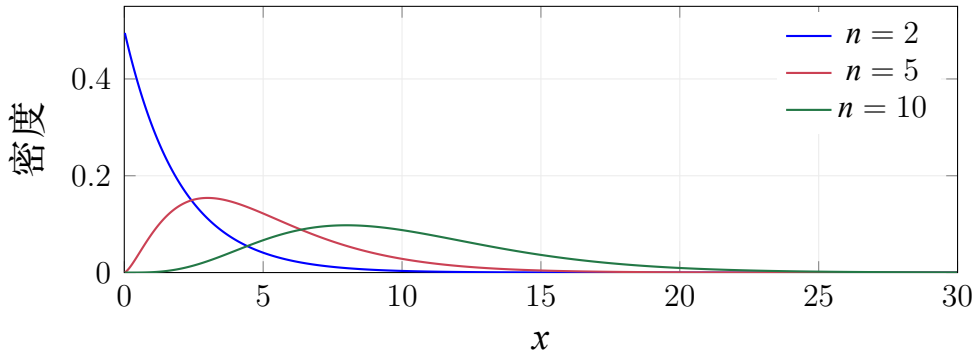
$$EX = n, \qquad \mathrm{Var}(X) = 2n.$$

- 若 $X \sim \chi^2_m$、$Y \sim \chi^2_n$，且相互独立，则

$$X + Y \sim \chi^2_{m+n}.$$

- 一般呈右偏；自由度增大时，相对偏斜程度减小。

卡方分布：自由度与形状



t 分布

定义 5.8　t 分布

设 $X \sim N(0,1)$、$Y \sim \chi^2_n$，且二者**相互独立**，则

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n.$$

- t 分布关于 0 对称，取值范围为 $\mathbb{R}$。
- 常用于总体方差未知时，对正态总体均值进行 t 检验。

t 分布的性质

- t_1 是**标准柯西分布**，期望不存在。
- 若 $n > 1$，则 $ET = 0$；若 $n > 2$，则

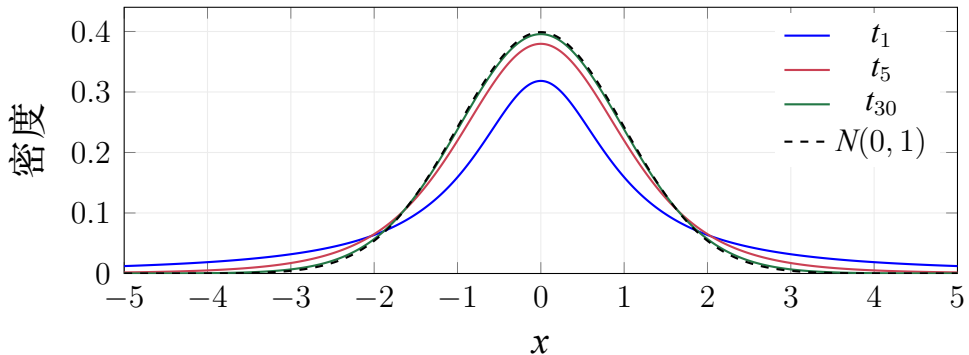
$$\mathrm{Var}(T) = \frac{n}{n-2}.$$

- 有限自由度的 t 分布比标准正态分布的尾部更重。
- 当 $n \to \infty$ 时，t_n 分布趋于 $N(0,1)$。

注

若 $T \sim t_n$，则当 $0 < k < n$ 时，$E|T|^k < \infty$，即 k 阶绝对矩存在。

t 分布与标准正态分布



F 分布

定义 5.9　F 分布

设 $X \sim \chi^2_m$、$Y \sim \chi^2_n$，且二者**相互独立**，则

$$F = \frac{X/m}{Y/n} \sim F_{m,n}.$$

- m 为分子自由度，n 为分母自由度，**顺序不能随意交换**。
- $F > 0$，常用于研究两个独立正态总体的方差比。

F 分布的性质

- 若 $F \sim F_{m,n}$，则 $1/F \sim F_{n,m}$。
- 若 $T \sim t_n$，则 $T^2 \sim F_{1,n}$。
- **上分位数**：$P\{F > F_{m,n}(\alpha)\} = \alpha$，则

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

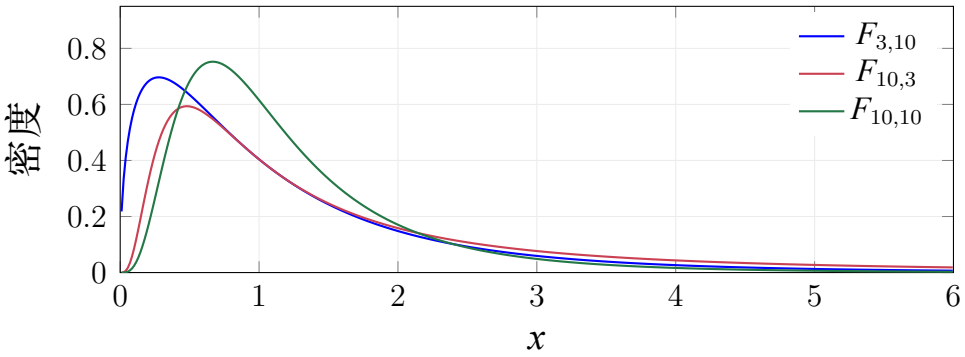
思考

若下分位数 $q_{m,n}(\alpha)$ 满足

$$P\{F \le q_{m,n}(\alpha)\} = \alpha, \qquad 0 < \alpha < 1,$$

它与 $q_{n,m}(1-\alpha)$ 有什么关系？

F 分布：自由度的顺序



Pop Quiz 2：分布的构造与性质

多选题　下列说法中，正确的是：

(A) 若 Z_1, Z_2 各服从 $N(0,1)$，则 $Z_1^2 + Z_2^2 \sim \chi_2^2$。

(B) t_1 关于 0 对称，因此其期望等于 0。

(C) 若 $T \sim t_5$，则 $T^2 \sim F_{1,5}$。

(D) 若 $W \sim F_{3,8}$，则 $1/W \sim F_{8,3}$。

Pop Quiz 2：答案

答案　**(C)、(D)**。

正态总体下的样本均值与样本方差

定理 5.1

设 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$n \geq 2$、$\sigma^2 > 0$，则

1. $\overline{X} \sim N(\mu, \sigma^2/n)$。
2. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$。
3. $\overline{X}$ 与 S^2 相互独立。

- 对一般非正态总体，上述分布与独立性结论不一定成立。

由正态总体得到 t 分布

推论 5.1

在上述正态总体条件下，

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

推导： 令

$$Z=\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1), \qquad Q=\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

由 Z, Q 独立，$\dfrac{Z}{\sqrt{Q/(n-1)}}=\dfrac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}$。

Pop Quiz 3：精确结论还是渐近结论?

问题 1（单选） 正态总体方差未知，用 S 替代 σ 后，$\sqrt{n}(\overline{X}-\mu)/S$ 的精确分布是:

(A) $N(0,1)$。 (B) t_n。 (C) t_{n-1}。

问题 2（简答） 若简单随机样本的总体非正态，但具有有限方差，$E\overline{X}=\mu$、$\mathrm{Var}(\overline{X})=\sigma^2/n$ 是否仍成立? $\overline{X}$ 与 S^2 是否一定独立?

Pop Quiz 3：答案

问题 1 **(C)**。自由度来自样本方差对应的 χ^2_{n-1}。

问题 2

- 对独立同分布、有限方差的样本，均值与方差公式仍然成立。
- $\overline{X}$ 与 S^2 不一定独立。

两个独立正态样本：均值差

推论 5.2　共同方差

设 $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma^2)$，$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma^2)$，两组样本相互独立，$m, n \geq 2$、$\sigma^2 > 0$。记两组样本均值为 $\overline{X}, \overline{Y}$，样本方差为

$$S_X^2 = \frac{\sum_{i=1}^m (X_i - \overline{X})^2}{m-1}, \quad S_Y^2 = \frac{\sum_{j=1}^n (Y_j - \overline{Y})^2}{n-1}.$$

令 $S_w^2 = \dfrac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$，则

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{1/m + 1/n}} \sim t_{m+n-2}.$$

两个独立正态样本：方差比

推论 5.3

设 $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$，$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$，两组样本相互独立，$m, n \geq 2$、$\sigma_1^2, \sigma_2^2 > 0$。记两组样本均值为 $\overline{X}, \overline{Y}$，样本方差为

$$S_X^2 = \frac{\sum_{i=1}^m (X_i - \overline{X})^2}{m-1}, \quad S_Y^2 = \frac{\sum_{j=1}^n (Y_j - \overline{Y})^2}{n-1}.$$

则

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F_{m-1,n-1}.$$

指数总体的一个抽样分布结论

推论 5.4

设 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$，其中 $\lambda > 0$ 为率参数，则

$$2\lambda \sum_{i=1}^{n} X_i = 2\lambda n \overline{X} \sim \chi^2_{2n}.$$

采用形状–率参数记法，由独立 Gamma 变量的可加性，

$$X_i \sim \text{Ga}(1, \lambda) \quad \Longrightarrow \quad \sum_{i=1}^{n} X_i \sim \text{Ga}(n, \lambda).$$

因此 $2\lambda \sum_{i=1}^{n} X_i \sim \text{Ga}(n, 1/2) = \chi^2_{2n}$。

Pop Quiz 4：使用公式前先检查条件

多选题　合并方差的两样本 t 公式需要：

(A) 两组样本相互独立，且各自为简单随机样本。
(B) 两总体为正态分布。
(C) 两总体方差相等。
(D) 两组样本量相等。

Pop Quiz 4：答案

答案　**(A)、(B)、(C)**。两组样本量可以不同。