《概率论与数理统计》第十二周要点

第六章样本及抽样分布——复习要点

1. 总体与样本

(1) 总体的概念

- 对实际问题中的数量指标进行观察，其全部可能取值构成总体。
- 总体中每一个可能观察值称为个体。
- 容量有限的为有限总体；容量极大或被视为无限的称为无限总体。
- 总体的分布即随机变量 X 的分布，如“灯泡寿命为指数分布总体”等。

(2) 简单随机样本　设 $X_1, X_2, \ldots, X_n$ 是在相同条件下获得的独立观测值，且与总体 X 具有相同分布，则称

$$X_1, X_2, \ldots, X_n$$

为来自总体 X 的容量为 n 的简单随机样本。

- 样本值：$x_1, x_2, \ldots, x_n$；
- 样本可视为随机向量 $(X_1, \ldots, X_n)$。

(3) 获取样本的方式

- 有限总体：放回抽样 $\rightarrow$ 严格独立；不放回抽样在 $N \gg n$ 时可近似为独立；
- 无限总体：抽取一个不影响总体分布，视为独立抽样。

—

2. 描述性统计图形：直方图与箱线图

(1) 频率直方图

- 依次步骤：确定区间 $\rightarrow$ 分组（组距与组限）$\rightarrow$ 统计每组频数/频率；
- 每个矩形面积 = 该组的频率；
- 当样本量大时，直方图外廓接近于总体概率密度曲线；
- 可用于判断总体形态（如本章图示中呈现单峰、近似对称似正态分布）。

(2) 样本分位数　对样本按从小到大排列 $x_{(1)} \le \cdots \le x_{(n)}$，$p$ 分位数满足：

$$\text{至少有} np \text{ 个样本值} \le x_p, \quad \text{且至少有} n(1-p) \text{ 个样本值} \ge x_p.$$

- 中位数：$p = 0.5$；
- 四分位数：$Q_1 = x_{0.25}$，$Q_3 = x_{0.75}$；
- 四分位距：$\mathrm{IQR} = Q_3 - Q_1$。

(3) 箱线图　由 Min, Q_1, M, Q_3, Max 五个数构成，反映：

- 中心位置（由中位数给出）；
- 离散程度（箱体长度、全距）；
- 对称性（中位数在箱体位置）；
- 尾部长短（触须延伸长度）。

(4) 异常值检测（修正箱线图）　若数据满足：

$$x < Q_1 - 1.5\,\mathrm{IQR} \quad \text{或} \quad x > Q_3 + 1.5\,\mathrm{IQR},$$

则视为疑似异常值，并在图中以符号标出。

—

3. 统计量与经验分布函数

(1) 统计量　若 $T = g(X_1, \ldots, X_n)$ 含有样本值但不含未知参数，则称 T 为统计量。

(2) 常用统计量

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

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k.$$

(3) 大数定律的应用　若总体存在 k 阶矩，则：

$$A_k \xrightarrow{P} E(X^k), \qquad B_k \xrightarrow{P} E[(X - E(X))^k].$$

(4) 经验分布函数（EDF）　定义：

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}.$$

格里汶科定理（Glivenko–Cantelli）给出一致收敛：

$$\sup_x |F_n(x) - F(x)| \xrightarrow{P} 0.$$

经验分布函数可作为总体分布 $F(x)$ 的估计。

—

4. 抽样分布：χ^2 分布、t 分布、F 分布（不考）

(1) χ^2 分布　若 $X_i \sim N(0,1)$ 独立，则：

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n).$$

性质：

$$E(\chi_n^2) = n, \qquad D(\chi_n^2) = 2n.$$

用于样本方差的分布推断。

—

(2) t 分布（Student 分布）　设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$ 且独立，则：

$$T = \frac{X}{\sqrt{Y/n}} \sim t(n).$$

性质：

- 图形对称、尾部较重；
- 当 $n \to \infty$ 时，$t(n) \to N(0,1)$。

—

(3) F 分布　若 $U \sim \chi^2(n_1)$、$V \sim \chi^2(n_2)$ 独立，则：

$$F = \frac{U/n_1}{V/n_2} \sim F(n_1, n_2).$$

F 分布用于比较方差、方差比检验等。

—

5. 正态总体下样本均值与样本方差的抽样分布（不考）

设总体 $X \sim N(\mu, \sigma^2)$，样本 $X_1, \ldots, X_n$。

(1) 样本均值的分布

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

(2) 样本方差的分布

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

(3) 样本均值与样本方差相互独立　这是正态分布的重要性质（非正态时一般不成立）。

(4) t 分布的由来

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

(5) 两个正态总体的比较（双样本情形）　若两总体 $N(\mu_1, \sigma_1^2)$ 与 $N(\mu_2, \sigma_2^2)$ 独立：

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(\text{适当自由度}),$$

若假设方差相等，则样本方差可合并，得到标准两样本 t 检验形式。

—

小结

- 样本是统计推断的基础；统计量是样本的函数。
- 直方图与箱线图为描述性统计的重要工具。
- 经验分布函数用于估计总体分布函数，并满足一致收敛。
- χ^2、t、F 分布是三类重要抽样分布，均来源于正态总体。
- 正态总体具有特殊性质：$\bar{X}$ 与 S^2 独立，并产生 t 与 F 分布。