

第六章 数理统计的基本概念

本章导读

前五章通常从一个已经确定的分布出发, 计算概率或数字特征. 本章把方向反过来: 总体分布含有未知信息时, 怎样从有限个观测值推断总体? 这就是数理统计的起点.

本章的知识链是



前三个框说明“用什么数据、怎样加工数据”, 最后一个框说明“加工结果具有怎样的随机规律”. 第七至第九章的参数估计和假设检验, 都建立在这条链上.

6.1 总体与样本

6.1.1 总体不是一堆编号, 而是所研究指标的分布

假设我们关心一批灯泡的寿命. 每只灯泡是一个**个体**, 寿命是个体上的数量指标. 若从总体中随机抽取一只灯泡, 用随机变量 X 表示其寿命, 那么 X 的分布就刻画了总体.

定义 6.1 (总体、个体与总体分布)

研究对象的全体称为**总体**, 组成总体的每个基本单位称为**个体**. 在数理统计中, 通常用个体上某个数量指标的分布来代表总体, 也把服从该分布的随机变量 X 简称为总体.

总体分布可能完全已知, 也可能只知道分布族而不知道参数. 例如

$$X \sim N(\mu, \sigma^2)$$

表示已知总体属于正态分布族; 若 μ, σ^2 未知, 统计问题就是利用样本获得关于它们的信息.

注: [模型中的“已知”与“未知”] “总体是正态总体”不等于“总体分布完全已知”. 前者只说明分布类型; 只有当 μ 和 σ^2 也已知时, 分布才被完全确定.

6.1.2 简单随机样本

不可能观察整个总体时,就从中抽取一部分个体.为使样本能够代表总体,抽样方式必须受到条件约束.

定义 6.2 (简单随机样本)

设总体的分布函数为 $F(x)$. 若随机变量 $X_1, X_2, \dots, X_n$ 满足:

1. $X_1, \dots, X_n$ 相互独立;
2. 每个 X_i 都与总体 X 同分布, 分布函数均为 $F(x)$,

则称 $X_1, \dots, X_n$ 是来自总体 X 的容量为 n 的简单随机样本, 简称样本.

“同分布”保证每次抽取面对同一个总体规律;“相互独立”保证一次抽取的结果不改变其他抽取的概率规律. 本讲义后续所说的样本,若无特别说明,均指简单随机样本.

6.1.3 样本的二重性

抽样前, $X_1, \dots, X_n$ 的值尚未确定, 它们是随机变量; 抽样后得到

$$x_1, x_2, \dots, x_n,$$

这些是确定的观测值. 二者不能混用.

阶段	记号	含义
观测前	$X_1, \dots, X_n$	随机变量, 可讨论分布、期望和方差
观测后	$x_1, \dots, x_n$	已经得到的数, 可进行具体计算

例题 6.3 (区分样本与样本值)

随机抽取三只灯泡并记录寿命. 抽取前用 X_1, X_2, X_3 表示三次寿命, 它们构成随机样本; 若实际得到 1010, 980, 1050 小时, 这三个数是该样本的一组样本值.

为什么要区分: “样本均值服从什么分布”讨论的是随机变量的函数; “样本均值是多少”则把实际数字代入计算. 前者属于理论推断, 后者属于数据处理.

小结

总体用所研究指标的概率分布刻画; 简单随机样本必须同时满足独立和同分布. 使用样本前, 先辨明分布类型与参数哪些已知, 再辨明当前记号表示随机样本还是已经观测到的样本值.

6.2 经验分布函数

6.2.1 用样本中“不超过给定值的比例”模仿总体概率

总体分布函数是 $F(x) = P(X \leq x)$. 若 F 未知, 自然可以用样本中不超过 x 的观测所占比例来估计它.

定义 6.4 (经验分布函数)

设 $X_1, \dots, X_n$ 为样本, 定义

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}},$$

其中示性函数在事件 $\{X_i \leq x\}$ 发生时取 1, 否则取 0. 观测到样本值后, $F_n(x)$ 就是样本中不超过 x 的数据比例.

例题 6.5 (从频数写出经验分布函数)

一组容量为 10 的样本值为

$$1, 2, 4, 3, 4, 2, 3, 4, 4, 2.$$

不同取值 1, 2, 3, 4 的频数依次为 1, 3, 2, 4, 故

$$F_{10}(x) = \begin{cases} 0, & x < 1, \\ 0.1, & 1 \leq x < 2, \\ 0.4, & 2 \leq x < 3, \\ 0.6, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

方法选择: 先排序并统计不同取值的频数, 再逐段累加相对频数. **结果检查:** 经验分布函数必须右连续、单调不减, 且两端极限分别为 0 和 1.

定理 6.6 (格里文科定理)

设 $F_n(x)$ 是来自总体分布函数 $F(x)$ 的经验分布函数, 则当 $n \rightarrow \infty$ 时, $F_n(x)$ 以概率 1 一致收敛于 $F(x)$, 即

$$P\left(\lim_{n \rightarrow \infty} \sup_x |F_n(x) - F(x)| = 0\right) = 1.$$

注: [为什么经验分布能代表总体] 对固定的 x , 每个 $1_{\{X_i \leq x\}}$ 都是参数为 $F(x)$ 的伯努利随机变量, $F_n(x)$ 正是它们的样本均值. 大数定律说明它会稳定到 $F(x)$. 格里文科定理进一步把“固定一个 x ”加强为“对所有 x 同时控制”. 严格证明涉及本阶段尚未学习的工具, **本阶段不要求复现**; 应会构造和检查经验分布函数.

小结

经验分布函数用样本比例估计总体分布函数, 是一条从数据到分布的直接通道. 构造时先排序、计频数、再累加; 检查时看它是否为右连续的阶梯函数, 并满足单调性和两端极限.

6.3 统计量与常用样本统计量

6.3.1 统计量: 只加工样本, 不偷用未知信息

定义 6.7 (统计量与统计量的观测值)

设 $X_1, \dots, X_n$ 为样本. 若函数

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

不含任何未知参数, 则称 T 为**统计量**. 把样本值代入后得到的数 $T(x_1, \dots, x_n)$, 称为统计量的**观测值**或**统计值**.

统计量仍是随机变量, 因此有自己的分布、期望和方差. 定义中禁止的是**未知参数**, 而不是一切常数: 已知参数可以出现在统计量中.

例题 6.8 (判断是否为统计量)

设 $X_1, \dots, X_n$ 来自 $N(\mu, \sigma^2)$, 且 μ, σ^2 均未知. 则 $\bar{X}$ 、 $\max_i X_i$ 和 $\sum X_i^2$ 都是统计量; $\bar{X} - \mu$ 与 S^2/σ^2 含有未知参数, 不是统计量.

判断方法: 先列出题目中的未知参数, 再逐个检查表达式中是否出现它们. 不要因为一个表达式具有重要分布就误称它为统计量.

6.3.2 五类常用统计量

定义 6.9 (样本均值、样本方差与样本矩)

设 $X_1, \dots, X_n$ 为样本, 定义

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{样本均值,}$$

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

$$S = \sqrt{S^2} \quad \text{样本标准差,}$$

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad k \text{ 阶样本原点矩,}$$

$$B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k \quad k \text{ 阶样本中心矩.}$$

特别地, $A_1 = \bar{X}$, 而

$$B_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2.$$

样本均值描述数据的中心位置; 样本方差和标准差描述离散程度; 样本矩则与总体矩对应, 是后面矩估计法的基础.

定理 6.10 (样本方差的计算恒等式)

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

因而

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

证明: 证明目标: 把“离均差平方和”改写为平方和与样本均值的组合.

出发点: 使用 $\sum_{i=1}^n X_i = n\bar{X}$.

关键转换: 展开平方, 再把与下标 i 无关的 $\bar{X}$ 提出求和.

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

证明小结: 真正起作用的是样本均值定义. 这个恒等式既便于手算, 也是在证明 $E(S^2) = \sigma^2$ 时消去交叉项的关键工具. ■

小结

统计量是只由样本和已知量构成的函数; 代入样本值后才得到统计值. 遇到常用统计量, 要分清 S^2 的分母是 $n-1$, 而二阶样本中心矩 B_2 的分母是 n .

6.4 统计量的数字特征

设总体 X 的期望和方差存在, 并记

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

定理 6.11 (样本均值的期望与方差)

对简单随机样本 $X_1, \dots, X_n$,

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

证明: 证明目标: 求样本均值的期望和方差.

出发点: 样本同分布给出 $E(X_i) = \mu$ 、 $\text{Var}(X_i) = \sigma^2$; 样本独立使和的方差可以直接相加.

关键转换: 把常数 $1/n$ 提到期望或方差之外.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu, \\ \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}. \end{aligned}$$

证明小结: 期望结论只需同分布, 方差相加还用到了独立性. 样本量扩大时, 样本均值的中心不变, 而波动按 $1/n$ 缩小. ■

定理 6.12 (样本方差的无偏性)

若总体方差为 σ^2 , 则

$$E(S^2) = \sigma^2.$$

因此用分母 $n - 1$ 定义的样本方差是总体方差的无偏估计.

证明: 证明目标: 证明离均差平方和除以 $n - 1$ 后, 期望恰为 σ^2 .

出发点: 使用恒等式

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

关键转换: 分别求右侧两项的期望; 第一项含 n 个总体方差, 第二项含样本均值的方差.

$$\begin{aligned} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] &= \sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \\ &= n\sigma^2 - n\text{Var}(\bar{X}) \\ &= n\sigma^2 - n \cdot \frac{\sigma^2}{n} \\ &= (n - 1)\sigma^2. \end{aligned}$$

两边除以 $n - 1$, 得到 $E(S^2) = \sigma^2$.

证明小结: 估计均值 μ 消耗了一个自由度, 所以离均差平方和的期望只有 $(n - 1)\sigma^2$. 若仍除以 n , 所得 B_2 的期望为 $\frac{n-1}{n}\sigma^2$, 会系统性偏小. ■

例题 6.13 (先识别结构, 再求期望)

若总体 $X \sim B(m, p)$, 求

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right].$$

二项分布的方差为 $mp(1 - p)$. 由上一证明中的中间结论, 直接得到

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = (n - 1)mp(1 - p).$$

方法选择: 看到离均差平方和, 应先联系 $(n - 1)S^2$, 不必逐项展开. **结果检查:** 当 $n = 1$ 时离均差恒为零, 公式也给出零.

小结

样本均值满足 $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \sigma^2/n$; 样本方差满足 $E(S^2) = \sigma^2$. 使用这些结论前, 先检查是否为简单随机样本, 以及所需的总体矩是否存在.

6.5 三大抽样分布与上分位点

6.5.1 抽样分布与上分位点

定义 6.14 (抽样分布)

统计量的概率分布称为该统计量的**抽样分布**. 它描述重复抽样时统计量怎样波动, 是进行区间估计和假设检验的概率基础.

定义 6.15 (上 α 分位点)

设连续型随机变量 Y 的密度为 $f(y)$. 若数 y_α 满足

$$P(Y > y_\alpha) = \int_{y_\alpha}^{\infty} f(y) dy = \alpha,$$

则称 y_α 为 Y 的上 α 分位点.

“上”表示右尾. 看到上 α 分位点, 应立刻画出右侧面积为 α 的示意图, 避免把 $P(Y \leq y_\alpha) = \alpha$ 写反.

6.5.2 卡方分布: 独立标准正态的平方和

定义 6.16 (χ^2 分布)

若 $Z_1, \dots, Z_k$ 相互独立且均服从 $N(0, 1)$, 则

$$V = \sum_{i=1}^k Z_i^2$$

服从自由度为 k 的卡方分布, 记为 $V \sim \chi^2(k)$.

常用性质为

$$E(V) = k, \quad \text{Var}(V) = 2k.$$

若 $V_1 \sim \chi^2(k_1)$ 、 $V_2 \sim \chi^2(k_2)$ 且相互独立, 则

$$V_1 + V_2 \sim \chi^2(k_1 + k_2).$$

自由度为 k 的卡方分布上 α 分位点记为 $\chi_\alpha^2(k)$.

例题 6.17 (线性组合先标准化, 再平方求和)

设 $X_1, \dots, X_5$ 相互独立且均服从 $N(0, 1)$, 令

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 - \sqrt{2}X_5)^2.$$

求常数 c , 使 $cY \sim \chi^2(2)$.

两个括号均服从 $N(0, 3)$, 且因使用了互不重叠的独立变量而相互独立. 因此

$$\frac{X_1 + X_2 + X_3}{\sqrt{3}}, \quad \frac{X_4 - \sqrt{2}X_5}{\sqrt{3}}$$

是两个独立标准正态变量, 故 $Y/3 \sim \chi^2(2)$, 即 $c = 1/3$.

结果检查: 自由度等于独立标准正态平方项的个数, 而不是原始变量的总数.

6.5.3 学生分布: 标准正态除以独立卡方均方的平方根

定义 6.18 (t 分布)

若 $Z \sim N(0, 1)$ 、 $V \sim \chi^2(k)$, 且 Z 与 V 相互独立, 则

$$T = \frac{Z}{\sqrt{V/k}}$$

服从自由度为 k 的 t 分布, 记为 $T \sim t(k)$.

t 分布关于原点对称, 因此其上分位点满足

$$t_{1-\alpha}(k) = -t_{\alpha}(k).$$

当自由度增大时, $t(k)$ 逐渐接近 $N(0, 1)$; $t(1)$ 是标准柯西分布, 期望不存在. t 分布密度的严格推导属于本阶段不要求复现的深层理论证明.

6.5.4 F 分布: 两个独立卡方均方之比

定义 6.19 (F 分布)

若 $U \sim \chi^2(k_1)$ 、 $V \sim \chi^2(k_2)$, 且 U, V 相互独立, 则

$$F = \frac{U/k_1}{V/k_2}$$

服从自由度为 (k_1, k_2) 的 F 分布, 记为 $F \sim F(k_1, k_2)$.

定理 6.20 (t 与 F 的平方关系)

若 $T \sim t(k)$, 则

$$T^2 \sim F(1, k).$$

证明: 证明目标: 把 T^2 改写成两个独立卡方变量分别除以自由度后的比.

出发点: 由定义, $T = Z/\sqrt{V/k}$, 其中 $Z \sim N(0, 1)$ 、 $V \sim \chi^2(k)$ 且独立.

关键转换: 利用标准正态的平方服从 $\chi^2(1)$.

$$T^2 = \frac{Z^2/1}{V/k} \sim F(1, k).$$

证明小结: 平方把 t 分布分子中的标准正态变成自由度为 1 的卡方; 独立性由原定义保留. ■

定理 6.21 (F 分布的倒数与分位点)

若 $F \sim F(k_1, k_2)$, 则

$$\frac{1}{F} \sim F(k_2, k_1),$$

并且

$$F_{1-\alpha}(k_1, k_2) = \frac{1}{F_{\alpha}(k_2, k_1)}.$$

证明: 证明目标: 说明取倒数会交换两个自由度, 并推出相应的上分位点关系.

出发点: 写成 $F = (U/k_1)/(V/k_2)$, 其中两个卡方变量独立.

关键转换: 取倒数交换分子与分母.

$$\frac{1}{F} = \frac{V/k_2}{U/k_1} \sim F(k_2, k_1).$$

再由连续性,

$$P(F \leq F_{1-\alpha}(k_1, k_2)) = \alpha.$$

取倒数后不等号方向改变, 故

$$P\left(\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(k_1, k_2)}\right) = \alpha.$$

右端临界值应为 $F_{\alpha}(k_2, k_1)$, 整理即得结论.

证明小结: 倒数不仅交换自由度, 还会交换左右尾, 所以分位点下标由 $1 - \alpha$ 变为 α . ■

目标分布	典型结构	必查条件	自由度来源
χ^2	标准正态平方和	各项独立且已标准化	独立平方项数
t	标准正态除以卡方均方平方根	分子、分母独立	分母卡方自由度
F	两个卡方均方之比	两个卡方独立	分子、分母分别计数

注: [识别结构不等于只看外形] “平方和”不自动服从卡方, “正态除以根号平方和”也不自动服从 t . 每次都要完成三步: 先把正态变量标准化, 再核对自由度, 最后验证组成部分的独立性.

小结

三大抽样分布都由独立标准正态变量构造. 识别时按“标准化—独立性—自由度”检查; 查表时先确认使用的是上分位点, 并特别留意 F 分布取倒数会同时交换自由度和尾部.

6.6 一个正态总体的抽样分布

设 $X_1, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的简单随机样本, $\bar{X}$ 和 S^2 分别为样本均值和样本方差.

定理 6.22 (正态总体的四个核心抽样结论)

1. 样本均值仍服从正态分布:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

2. $\bar{X}$ 与 S^2 相互独立.
3. 两个常用平方和满足

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

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

4. 用 S 代替未知的 σ 后,

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

注: [结论怎样串起来] 第一条来自独立正态变量的线性组合. 第三条中, 减去已知 μ 得到 n 个独立标准正态平方, 所以自由度为 n ; 减去样本均值后, 偏差满足

$$\sum_{i=1}^n (X_i - \bar{X}) = 0,$$

存在一个线性约束, 故只剩 $n - 1$ 个自由方向. 第二条“正态样本均值与样本方差独立”的严格证明依赖正交变换, **本阶段不要求复现**; 但必须记住它依赖正态总体. 结合第二、三条就能按定义构造第四条.

证明: 证明目标: 由标准正态与卡方的组合, 推出学生化统计量服从 $t(n - 1)$.

出发点: 令

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

正态总体保证 $\bar{X}$ 与 S^2 独立, 所以 Z 与 V 独立.

关键转换: 按 t 分布定义组成 $Z/\sqrt{V/(n-1)}$.

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

证明小结: 核心条件是总体正态和均值、方差的独立性. 若用 $\sum (X_i - \mu)^2$ 构造分母, 它一般不与 $\bar{X}$ 独立, 不能据此得到 $t(n)$. ■

例题 6.23 (样本均值的概率)

从 $N(5, 16)$ 中抽取容量为 36 的简单随机样本, 求

$$P(4 \leq \bar{X} \leq 6).$$

因为

$$\bar{X} \sim N\left(5, \frac{16}{36}\right),$$

其标准差为 $2/3$. 标准化得

$$P(4 \leq \bar{X} \leq 6) = P(-1.5 \leq Z \leq 1.5) = 2\Phi(1.5) - 1.$$

方法选择: 问题只涉及 $\bar{X}$ 且总体方差已知, 直接使用样本均值的正态分布. **结果检查:** 区间以总体均值 5 为中心, 标准化后应得到对称区间.

例题 6.24 (用卡方分布求统计量的方差)

设总体为 $N(0, 1)$, 令

$$T = \bar{X}^2 - \frac{S^2}{n}.$$

求 $E(T)$ 与 $\text{Var}(T)$.

由 $E(\bar{X}^2) = \text{Var}(\bar{X}) + E(\bar{X})^2 = 1/n + 1$ 和 $E(S^2) = 1$, 得 $E(T) = 0$. 又因 $\bar{X}$ 与 S^2 独立, 平方后仍独立, 故

$$\text{Var}(T) = \text{Var}(\bar{X}^2) + \frac{1}{n^2} \text{Var}(S^2).$$

由 $n\bar{X}^2 \sim \chi^2(1)$ 和 $(n-1)S^2 \sim \chi^2(n-1)$,

$$\text{Var}(\bar{X}^2) = \frac{2}{n^2}, \quad \text{Var}(S^2) = \frac{2}{n-1}.$$

因此

$$\text{Var}(T) = \frac{2}{n^2} + \frac{2}{n^2(n-1)} = \frac{2}{n(n-1)}.$$

关键转折: 遇到“正态变量的平方”的方差, 先把整体改造成卡方变量, 比直接计算四阶矩更简洁.

小结

一个正态总体有四个必须熟练调用的结论: 样本均值的正态分布、均值与方差独立、两个平方和的卡方分布、学生化均值的 t 分布. 使用时先判断分母中减的是 μ 还是 $\bar{X}$, 再确定自由度是 n 还是 $n-1$.

6.7 两个正态总体的抽样分布

设

$$X_1, \dots, X_{n_1} \sim N(\mu_1, \sigma_1^2), \quad Y_1, \dots, Y_{n_2} \sim N(\mu_2, \sigma_2^2),$$

两组均为简单随机样本, 并且两组样本相互独立. 分别记样本均值和样本方差为 $\bar{X}, S_1^2$ 与 $\bar{Y}, S_2^2$.

定理 6.25 (两个样本均值之差)

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

从而

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

证明: 证明目标: 求两个独立正态样本均值之差的分布.

出发点:

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right),$$

且两者独立.

关键转换: 独立正态变量的线性组合仍为正态变量; 作差时期望相减, 方差相加.

$$\begin{aligned} E(\bar{X} - \bar{Y}) &= \mu_1 - \mu_2, \\ \text{Var}(\bar{X} - \bar{Y}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

由此得到第一式, 再按其均值和标准差标准化即得第二式.

证明小结: 作差不会让方差相减; 只有期望带负号, 方差仍相加. 两组样本相互独立是去掉协方差项的关键条件. ■

定理 6.26 (两个样本方差之比)

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

证明: **证明目标:** 把样本方差之比改造成两个独立卡方均方之比.

出发点:

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

且两组样本独立, 所以 U, V 独立.

关键转换: 分别除以各自自由度后作比.

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

证明小结: 分子自由度来自第一组, 分母自由度来自第二组; 交换方差之比的次序时也要交换两个自由度. ■

定理 6.27 (等方差时的合并方差与两样本 t 统计量)

若 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 定义合并样本方差

$$S_w^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

则

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2).$$

注: [合并方差公式的构造骨架] 两个独立卡方变量具有可加性, 因此

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

另一方面, 均值差标准化后服从 $N(0, 1)$, 且均值部分与两个样本方差构成的 S_w^2 独立. 用前者除以后者除以自由度的平方根, 约去 σ , 便得到上述 t 统计量. 应掌握这条构造链; 完整的多元独立性证明本阶段不要求复现.

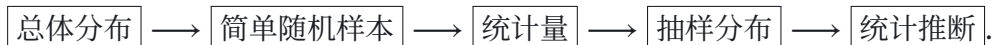
注: [三个公式怎样选择] 比较两个总体均值时, 若两个总体方差已知, 使用均值差的标准正态统计量; 若方差未知但已知相等, 使用合并方差构造的两样本 t 统计量. 比较两个总体方差时, 则从样本方差之比的 F 分布出发. 不要在未给出等方差条件时擅自使用 S_w^2 .

小结

两个正态总体的公式都可由一个总体的结论现推: 均值差用独立正态线性组合, 方差比用两个独立卡方均方之比, 等方差下的均值差再用合并方差学生化. 使用前必须检查两组样本相互独立, 以及合并方差所需的等方差条件.

本章总结

一、知识结构



二、核心公式及适用条件

结论	公式	关键条件
样本均值的数字特征	$E(\bar{X}) = \mu, \text{Var}(\bar{X}) = \sigma^2/n$	简单随机样本, 二阶矩存在
样本方差无偏性	$E(S^2) = \sigma^2$	简单随机样本, 二阶矩存在
一个正态总体	$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$	正态总体
一个正态总体	$(\bar{X} - \mu)/(S/\sqrt{n}) \sim t(n-1)$	正态总体
两个均值之差	标准化后服从 $N(0, 1)$	两组独立正态样本, 方差已知
两个方差之比	$(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2) \sim F(n_1-1, n_2-1)$	两组独立正态样本
合并方差统计量	学生化后服从 $t(n_1 + n_2 - 2)$	两组独立正态样本且方差相等

三、常见题型与方法选择

1. **构造经验分布函数:** 排序、计频数、累计相对频数, 再检查右连续与两端极限.
2. **判断统计量:** 先列未知参数, 再检查表达式中是否含有它们.
3. **求期望或方差:** 优先识别 $\bar{X}$ 、 S^2 、离均差平方和等整体, 不要急于逐项展开.
4. **判断抽样分布:** 先标准化, 再验证独立性, 最后按平方项个数确定自由度.
5. **正态总体平方问题:** 看到标准化正态的平方, 联想到 $\chi^2(1)$; 看到两个卡方均方之比, 联想到 F .
6. **两个总体问题:** 先问目标是均值差还是方差比, 再问方差已知、未知相等还是未知且不说明相等.

四、重点辨析

概念	正确理解	常见误判
样本与样本值	前者是随机变量组, 后者是观测后的数	用小写数据讨论抽样分布
统计量	只要求不含未知参数	误以为不能含任何参数或常数
S^2 与 B_2	分母分别为 $n-1$ 与 n	把二者视为同一个量
n 与 $n-1$	减已知 μ 得 n ; 减估计量 $\bar{X}$ 得 $n-1$	只按平方项表面个数计数
独立与不相关	构造 t, F 必须验证独立性	只看表达式外形便套分布
已知方差与等方差	已知方差用正态; 未知但相等才用合并方差	未给等方差条件仍强行合并

五、易错点清单

- 把总体当作有限数据表, 而忘记统计模型研究的是指标分布.
- 只检查样本“同分布”, 漏掉简单随机样本还要求相互独立.
- 把含未知参数的枢轴量误称为统计量.
- 把 $\text{Var}(\bar{X})$ 写成 $\sigma^2/\sqrt{n}$; 方差是 σ^2/n , 标准差才是 $\sigma/\sqrt{n}$.
- 忘记 F 分布的两个自由度有先后顺序.
- 把 $\bar{X}$ 与 S^2 的独立性推广到任意总体; 该核心结论依赖正态总体.
- 作差时误把方差也相减; 独立变量之差的方差仍相加.

六、证明掌握清单

- **A 级, 应能独立复现:** 样本均值的期望与方差; 样本方差恒等式与无偏性; 由独立标准正态变量构造 $\chi^2, t, F; T^2 \sim F(1, n)$ 、 F 倒数和分位点关系; 两个独立正态样本均值之差与方差之比.
- **B 级, 应掌握证明骨架:** 样本方差卡方分布自由度为何为 $n-1$; 等方差时合并方差和两样本 t 统计量的构造; 正态抽样公式之间怎样通过标准化连接.
- **C 级, 准确记条件和用途:** 正态样本均值与样本方差独立的严格证明; 格里文科定理严格证明; 三大抽样分布的密度推导; 正态抽样定理的完整多元几何证明. 这些内容本阶段不要求复现.

七、自测题

练习 6.28

设 $X_1, \dots, X_8$ 来自总体 $N(\mu, 4)$, μ 未知. 判断下列各式哪些是统计量:

$$\bar{X}, \quad \bar{X} - \mu, \quad \sum_{i=1}^8 (X_i - \bar{X})^2, \quad \frac{\sqrt{8}(\bar{X} - \mu)}{2}.$$

练习 6.29

一组样本值为 0, 1, 1, 2, 3. 写出其经验分布函数, 并检查右连续性、单调性和两端极限.

练习 6.30

设总体方差为 9, 样本容量为 25. 求 $E(\bar{X})$ 、 $\text{Var}(\bar{X})$ 与

$$E\left[\sum_{i=1}^{25} (X_i - \bar{X})^2\right].$$

练习 6.31

设 $X_1, \dots, X_{10}$ 来自 $N(0, 1)$. 分别判断

$$\sum_{i=1}^{10} X_i^2, \quad \sum_{i=1}^{10} (X_i - \bar{X})^2, \quad \frac{\sqrt{10}\bar{X}}{S}$$

服从什么分布, 并说明自由度来源.

练习 6.32

若 $T \sim t(12)$, 说明 T^2 和 $1/T^2$ 的分布; 若 $F \sim F(5, 8)$, 说明 $1/F$ 的分布, 并写出相应上分位点关系.

练习 6.33

两组相互独立的正态样本容量分别为 $n_1 = 10, n_2 = 15$. 说明在以下三种任务中应构造哪类分布: 总体方差均已知时比较均值; 总体方差未知但相等时比较均值; 比较两个总体方差.

完成自测后, 应能用一句话概括本章: **从总体抽取独立同分布样本, 用不含未知参数的统计量压缩数据, 再借助其抽样分布进行统计推断.**