13.4 三种重要的统计分布和分位数

正态分布是统计数据分析中最常见的分布，以标准正态分布为基础构造的 χ^2，t 和 F 分布通常被称为统计学中的“三大抽样分布”。

χ^2 分布：随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，都服从标准正态分布，则随机变量

$$Y = X_1^2 + \cdots + X_n^2$$

的分布称为自由度为 n 的 χ^2（卡方）分布，记为 $Y \sim \chi^2(n)$ 或 $Y \sim \chi^2_n$。

χ^2 分布随机变量具有可加性，即若 X_1, X_2 独立，$X_1 \sim \chi^2(m)$，$X_2 \sim \chi^2(n)$，则 $X_1 + X_2 \sim \chi^2(m+n)$。且若 $X \sim \chi^2(n)$，则 $E(X) = n$，$Var(X) = 2n$。

****

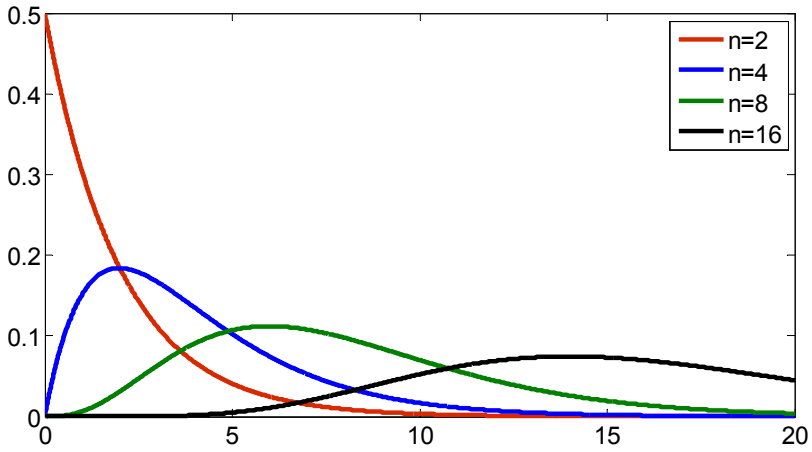

图 13.1 不同自由度 χ^2 分布的密度函数

****

χ^2 分布有重要意义的原因之一是下面的定理。

定理　设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，其样本均值和样本方差分别为 $\bar{X}$ 和 S^2，则有

（1） $\bar{X}$ 和 S^2 相互独立，

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

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

****

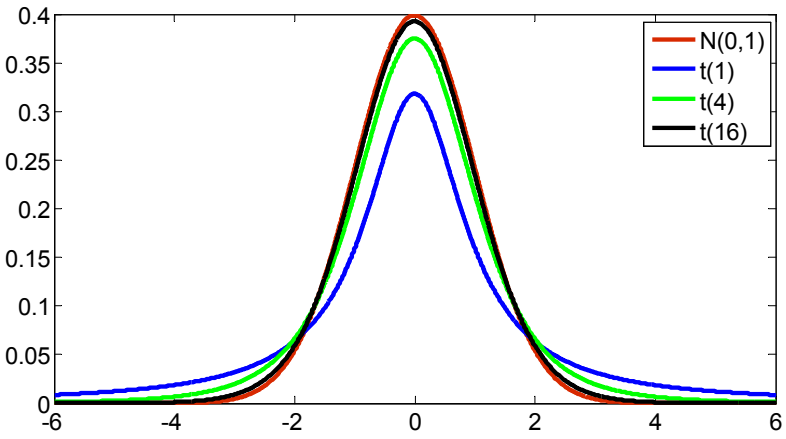

图 13.2 不同自由度 t 分布和标准正态分布的密度函数

****

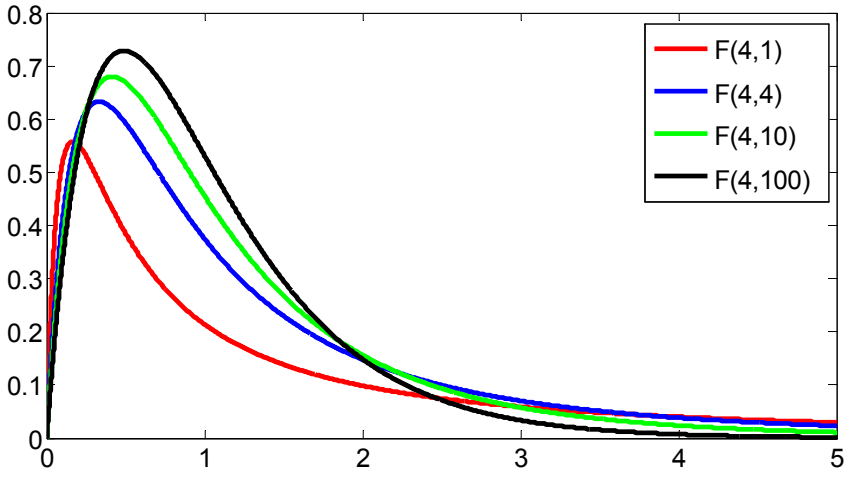

图 13.3 不同自由度 F 分布的密度函数

****

t 分布： 随机变量 $X_1 \sim N(0,1)$，$X_2 \sim \chi^2(n)$，X_1, X_2 相互独立，则 $Y = \dfrac{X_1}{\sqrt{X_2/n}}$

的分布称为自由度为 n 的 t 分布，记为 $Y \sim t(n)$。

F 分布： 随机变量 X_1, X_2 相互独立，$X_1 \sim \chi^2(m)$，$X_2 \sim \chi^2(n)$，则 $Y = \dfrac{X_1/m}{X_2/n}$

的分布称为自由度为 m 与 n 的 F 分布，记为 $Y \sim F(m,n)$。

****

分位数： X 为一连续分布随机变量，其分布函数为 $F(x)$，如果

$F(a) = P(X \le a) = \alpha$，则 a 称为该分布的（下侧）α 分位点，也称为 α 分位数。

标准正态分布的 α 分位点记为 u_α，

自由度为 n 的 χ^2 分布的 α 分位点记为 $\chi^2_\alpha(n)$，

自由度为 n 的 t 分布的 α 分位点记为 $t_\alpha(n)$，

自由度为 (m,n) 的 F 分布的 α 分位点记为 $F_\alpha(m,n)$。

****

例 13.4.1 $X \sim N(0,1)$，求 $P(|X| < u_{0.975})$。

解 利用标准正态分布的对称性，

$$P(|X| < u_{0.975}) = P(-u_{0.975} < X < u_{0.975}) = 2 \cdot P(0 < X < u_{0.975})$$

$$= 2 \cdot \left[\frac{1}{2} - P(X \ge u_{0.975})\right] = 1 - 2P(X \ge u_{0.975})$$

$$= 1 - 2 \cdot (1 - 0.975) = 0.95.$$

****

例 14.4.2 $X_1, X_2, \cdots, X_n$ 是来自均匀总体 $U(0,\theta)$ 的样本，求参数 θ 的矩估计量和最大似然估计量，并判断是否为无偏估计，若不是无偏估计尝试给出无偏校正，并比较估计的有效性。

验证 $\hat{\theta}_1 = 2\bar{X}$ 和 $\hat{\theta}_2 = \frac{n+1}{n}\max_{1\le k\le n} X_k$ 都是参数 θ 的无偏估计，并比较它们的有效性。

例 14.4.2 设元件的寿命服从指数分布 $f(x) = \lambda \cdot e^{-\lambda x}\ (x>0)$。为了了解元件寿命的期望值，即参数 $\frac{1}{\lambda}$，人们进行 n 次抽样，得到样本值 $x_1, x_2, \cdots, x_n$。具体的抽样方式为指定 $T>0$，取随机样本 $x_k = \begin{cases} \text{元件 } k \text{ 的寿命，当此寿命小于 } T \text{ 时} \\ T, \text{当时刻 } T \text{ 元件 } k \text{ 还没有失效} \end{cases}$，即试验进行到元件失效或时刻 T 停止。试基于以下思路对元件寿命参数 $\frac{1}{\lambda}$ 进行估计：

（1）利用取值小于 T 的样本数占总样本数的比例得到估计量；

（2）利用所有取值小于 T 的样本的均值得到估计量；

（3）利用样本的平均值得到估计量。

设总体 $X \sim E(\lambda)\ f(x) = \lambda \cdot e^{-\lambda x}$

某工厂正在对一种零件的生产过程进行技术革新，现在通过抽样试验，比较现有方法和新的方法，以判断新方法是否是零件质量有所改进。如果现有方法中 1500 件中有 75 件发现是不合格品，新方法中 2000 件有 80 件不合格品，

****

例 15.3.3 在《环境污染杂志》（Journal of Environmental Pollution）上发

表的一篇题为“大型无脊椎动物的聚集是酸矿污染的指示表”的论文中，报告了一项在美国亚拉巴马州某地进行的一项调查结果，调查的目的是研究生理化学参数与大型无脊椎动物聚集情况的不同测量结果之间的关系。调查中一项有效因素是，在酸矿的排水区域内，种类分歧度指数代表了水栖动物的退化程度，从概念上说，大型无脊椎动物种类的分歧度的高指数应该代表一个不重要的水栖动物系统，而种类的低指数则相反。

****

例 15.3.3（续）在本项研究中，选取两个独立水电站进行抽样，每月收集一个样品，在酸矿的排水区域的上游和下游各有一个采样点。下游水电站收集了 12 月的 12 个观测值，得到种类分歧度指数的均值为 $\bar{x}_1 = 3.11$，标准差为 $s_1 = 0.771$，而在上游收集了 10 个月中的 10 个观测值，相应的种类分歧度指数的均值为 $\bar{x}_2 = 2.04$，标准差为 $s_2 = 0.448$。当方差相等，总体近似服从正态分布时，求两个水电站总体均值差的 90%置信区间。

****

例 15.3.3（续）μ_1 和 μ_2 分别表示两个正态总体的期望，计算 $\mu_1 - \mu_2$ 的 90%置信区间。

如果方差相等，对公共方差进行估计 $s_w^2 = \dfrac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$

$$s_w^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} = \frac{11 \cdot 0.771^2 + 9 \cdot 0.448^2}{12+10-2} = 0.417 = 0.646^2$$

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

$$\sqrt{\frac{m+n}{mn}} s_w t_{0.95}(m+n-2) = \sqrt{\frac{1}{12} + \frac{1}{10}} \cdot 0.646 \cdot 1.725 = 0.477$$

$$0.593 \le \mu_1 - \mu_2 \le 1.547$$

字幕：如果两个总体方差未知或不相等，则无法进行进一步的估计。只有方差相等时，才能对期望差做最近一步的估计。实际得到的两个样本标准差的观测值分别是0.771和0.448的差别并不是特别悬殊，姑且可以认为方差相等。（稍停顿）