概率论与数理统计

第五章

杨青

yangq@ustc.edu.cn

第五章： 统计学基本概念

引言

- 陈希孺 (2009) 定义数理统计学为:

定义 5.1 数理统计学

数理统计学使用概率论和数学方法, 研究怎样**有效地收集** (试验或观测) 带有随机误差的数据, 并在设定的模型 (称为统计模型) 之下, 对这种数据进行**分析** (称为统计分析), 对所研究的问题作出**推断** (称为统计推断), 以便对所提问题作出**尽可能正确**的结论. ♣

1. 有效地收集数据

收集数据的方法有: **全面观察 (或普查)、抽样调查和安排试验**等方式.

—《抽样调查方法》 《试验的设计和分析》.

人口普查和抽样调查. 我国在 2000 年进行了第五次人口普查. 如果普查的数据是准确无误的, 无随机性可言, 不需用数理统计方法. 由于人口普查, 调查项目很多, 我国有 13 亿人口, 普查工作量极大, 而训练有素的工作人员缺乏. 因此虽是全面调查, 但数据并不可靠, 农村超计划生育瞒报、漏报人口的情况时有发生. 针对普查数据不可靠, 国家统计局在人口普查的同时还派出专业人员对全国人口进行抽样调查, 根据抽样调查的结果, 对人口普查的数字进行适当的修正. 抽样调查在普查不可靠时是一种补充办法.

2. 有效地使用数据（分析、推断）

- 获取数据后，需要用有效的方法，去集中和提取数据中的有关信息，以对所研究的问题作出一定的结论，在统计上称为“推断”.
- 为了有效的使用数据进行统计推断，需要对数据建立一个统计模型，并给定某些准则去评判不同统计推断方法的优劣.

某农村有 100 户农户，要调查此村农民是否脱贫. 脱贫的标准是每户年均收入超过 1 万元. 经调查此村 90 户农户年收入 5000 元，10 户农户年收入 10 万元，问此村农民是否脱贫?

(1) 用算术平均值计算该村农户年均收入如下：

$$\bar{x} = (90 \times 0.5 + 10 \times 10)/100 = 1.45(\text{万})$$

按此方法得出结论：该村农民已脱贫. 但 90% 的农户年均收入只有 5000 元，事实上并未脱贫.

(2) 用样本中位数计算该村农户年均收入：即将 100 户的年收入记为 $x_1, x_2, \cdots, x_{100}$，将其按大小排列为 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(100)}$. 样本中位数定义为排在最中间两户的平均值，即

$$(x_{(50)} + x_{(51)})/2 = 0.5(\text{万})$$

按此方法得出结论：该村农民尚未脱贫. 这与实际情况相符.

3. 尽可能正确的结论--数理统计方法的归纳性质

- 数理统计是数学的一个分支, 但是它的推理方法是不一样的. **统计方法的本质是归纳式**的, 而数学则是**演绎式**的.
- 统计方法的归纳性质, 源于它在作结论时, 是根据所观察到的大量的 “个别” 情况, “归纳” 起来所得. 而不是从一些假设、命题或已知事实出发按一定的逻辑推理得出来的 (这后者称为演绎推理) .
- 例如：统计学家通过大量的观察资料发现, 吸烟与某种呼吸系统的疾病有关. 他得出这一结论的根据是: 从观察到的大量例子, 看到吸烟者中患此种疾病的比例远高于不吸烟者. 他不可能用逻辑推理的方法证明这一点.

- 例如：在几何学中要证明“等腰三角形两底角相等”，只需从等腰这个前提出发，运用几何公理，一步步地推出这个结论 (这一方法属于演绎推理) . 而一个习惯于统计方法的人，就可能想出这样的方法：作很多大小形状不一的等腰三角形，实际测量它的底角查看区别如何，根据所得数据，看看可否作出底角相等的结论，这属于归纳推理的方法.
- **归纳推理是要冒风险的**. 人们不可能做出十分肯定的结论，因为归纳推理所依据的**数据一般带有随机性的误差**.即我们的样本来自于总体，至于哪些个体被抽到，在抽样方案实施之前是不知道的. 所以通过分析这些数据而作出结论，就难保不出差错.
- 统计学的作用之一就是提供归纳推理和计算不确定性程度的方法. 统计分析的要点就是使可能产生的错误越小越好.

若干基本概念–统计推断(主要为估计和假设检验)

1. 总体(population)：研究对象的全体，但只考虑我们感兴趣的指标

定义 5.2 统计总体

研究对象某个指标取值的全体以及取这些值的概率分布, 称为统计总体, 简称总体. ♣

- 由统计总体的定义及随机变量分布的定义, 总体就是一个分布.
- 总体也往往用随机变量 X 来表示. 例如我们称身高 $H \sim N(\mu, \sigma^2)$, 就是指身高 H 的取值是统计总体, 服从正态分布.

2. **样本**(sample)：抽取总体中的一部分个体进行观测或分析, 这些个体称为样本

定义 5.3 样本

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

- 从总体中按一定方式抽取样本的行为称为**抽样**. 最常用的一种抽样方法叫作"**简单随机抽样**".

- **简单随机抽样**要求满足下列两条:
  1. **代表性.** 总体中的每一个体都有同等机会被抽入样本，这意味着样本中每个个体与所考察的总体具有相同分布. 因此，任一样本中的个体都具有代表性.
  2. **独立性.** 样本中每一个体取什么值并不影响其它个体取什么值. 这意味着，样本中各个体 $X_1, X_2 \ldots, X_n$是相互独立的随机变量.

- 由简单随机抽样获得的样本 $(X_1, X_2 \ldots, X_n)$称为**简单随机样本**，也称为**简单样本**. 在没有歧义的时候也常常简称为**样本**.

性质 设 $X_1, \ldots, X_n$ 为从总体 F 中抽取的样本量为 n 的简单随机样本, 则

(i) $X_1, \ldots, X_n$ 相互独立,

(ii) $X_1, \ldots, X_n$ 相同分布, 即同有分布 F.

I. **有放回抽样**得到的样本**是简单随机样本**

II. **不放回抽样**，由于前面抽取的样本影响了后面样本取值的概率，所以**不是简单随机样本**

III. **n为样本量**，由于试验规定要抽取n个个体试验才结束，因此被抽到的n个个体放在一起称为一个样本，不是n个样本

IV. **样本的两重性**：样本既可看成具体的数，又可以看成随机向量.
一般，抽样方案实施之前，确定不了样本的具体值，视为随机向量 $(X_1, X_2 \dots, X_n)$；实施之后，确定了个体，也确定了指标的取值，这时样本是一组数，用小写的英文字母 $(x_1, x_2 \dots, x_n)$表示，常称为样本的一个实现（样本值）

- 例：假定一批产品有10000件, 其中有正品也有废品, 为估计废品率, 我们往往从中抽取一部分, 如100件进行检查. 此时这批10000件产品称为总体, 其中的每件产品称为个体, 而从中抽取的100件产品称为样本. 样本中个体的数目称为样本量. 而抽取样本的行为称为抽样.

3. 统计模型

- 样本既然可以视为随机变量，就有一定的概率分布，这个概率分布就叫作**样本分布**. 样本分布是样本所受随机性影响的最完整的描述.
- 一个问题的**统计模型**，就是指研究该问题时所抽样本的**样本分布**，也常称为**概率模型**或**数学模型**.
- 由于模型只取决于样本的分布，故常把**分布的名称作为模型的名称**. (如：样本分布为正态，可称其为正态模型).

例子

一大批产品共有 N 个, 其中废品 M 个, N 已知, 而 M 未知. 现在从中抽出 n 个加以检验, 用以估计 M 或废品率 $p = M/N$.

(1) 有放回抽样, 即每次抽样后记下结果, 然后将其放回去, 再抽第二个, 直到抽完 n 个为止. 求样本分布.

(2) 不放回抽样, 即一次抽一个, 依次抽取, 直到抽完 n 个为止. 求样本分布.

4. 统计量(statistic)：样本自身是一些杂乱无章的数字，要对这些数字进行加工整理，计算出一些有用的量.

定义 5.4 统计量

完全由样本 $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ 决定的量称为统计量 (statistic). ♣

由定义, 统计量是样本的函数.

(1) "完全"是指统计量中不能有其他未知参数, 例如当 μ 是一个给定的常数时 $\overline{X}-\mu$ 是统计量, 当 μ 作为未知参数时,$\overline{X}-\mu$ 就不是统计量.

(2) 由于样本有二重性, 所以统计量也有二重性, 既可以视为随机变量也可以视为具体数值.

(3) 统计量不是人为随意造出来的, 它们是为了解决种种统计推断问题而产生的.

- 样本均值 (sample mean)：

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

它反映了总体均值的信息.

- 样本方差 (sample variance)：

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

它反映总体方差的信息. 而 S 称为样本标准差, 它反映了总体标准差的信息.

- 样本矩 (sample moments):

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$$

称为样本 k 阶原点矩, 特别 $k=1$ 时, $a_1 = \overline{X}$ 即样本均值. 称

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, \quad k = 2, 3, \cdots$$

称为样本 k 阶中心矩.

- 样本偏度系数 (sample skewness coefficient)

$$\hat{\beta}_1 = \frac{m_3}{m_2^{3/2}} = \sqrt{n}\sum_{i=1}^{n}(X_i - \overline{X})^3 \Bigg/ \left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{3/2}$$

它反映了总体偏度 (见(4.17)) 的信息.

- 样本峰度系数 (sample kurtosis coefficient)

$$\hat{\beta}_2 = \frac{m_4}{m_2^2} - 3 = n\sum_{i=1}^{n}(X_i - \overline{X})^4 \Bigg/ \left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{2} - 3$$

它反映了总体峰度 (见(4.18)) 的信息.

- 样本相关系数 (sample coefficient of correlation)

设 $(X_1, Y_1), \cdots, (X_n, Y_n)$ 为从二维总体 $F(x, y)$ 中抽取的样本, 则称

$$\rho_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{5.6}$$

为样本相关系数, 也称为 Pearson 相关系数. 它反映总体相关系数 (见(4.22)) 的信息.

- 次序统计量 (order statistics)及其有关统计量：把样本按大小排列为

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)},$$

则称 $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ 为次序统计量, $(X_{(1)}, \ldots, X_{(n)})$ 的任一部分也称为次序统计量.

利用次序统计量可以定义下列统计量:

(1) 样本中位数 (sample median):

$$m_n = \begin{cases} X_{(\frac{n+1}{2})}, & \text{当 n 为奇数} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}], & \text{当 n 为偶数} \end{cases} \tag{5.7}$$

样本中位数反映总体中位数的信息. 当总体分布关于某点对称时, 对称中心既是总体中位数又是总体均值, 故此时 m_n 也反映总体均值的信息.

(2) 极值: $X_{(1)}$ 和 $X_{(n)}$ 称为样本的极小值和极大值. 极值统计量在关于灾害问题和材料试验的统计分析中是常用的统计量. $\max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i$ 称为极差 (range).

(3) 样本 p 分位数: $X_{([(n+1)p])}$, 其中 $[(n+1)p]$ 表示 $(n+1)p$ 的整数部分. 常见样本分位数包括样本四分位数 $X_{([(n+1)/4])}, X_{([(n+1)/2])}, X_{([3(n+1)/4])}$ 以及样本内四分位距 $X_{([3(n+1)/4])} - X_{([(n+1)/4])}$.

- 经验分布函数 (empirical distribution function):

$$\begin{aligned} F_n(x) &= \{X_1, X_2, \ldots, X_n \text{中} \le x \text{的个数}\}/n \\ &= \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_{(i)}) \\ &= \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \end{aligned}$$

称为样本 $X_1, \ldots, X_n$ 的经验分布函数. 它的图形是一个阶梯型的分布函数 (图5.1), 在 $X_{(i)}$ 处有跳 $1/n$. 由大数定律, 对几乎每个 x, $F_n(x)$ 依概率收敛于总体分布函数 $F(x)$.

经验分布函数图例:

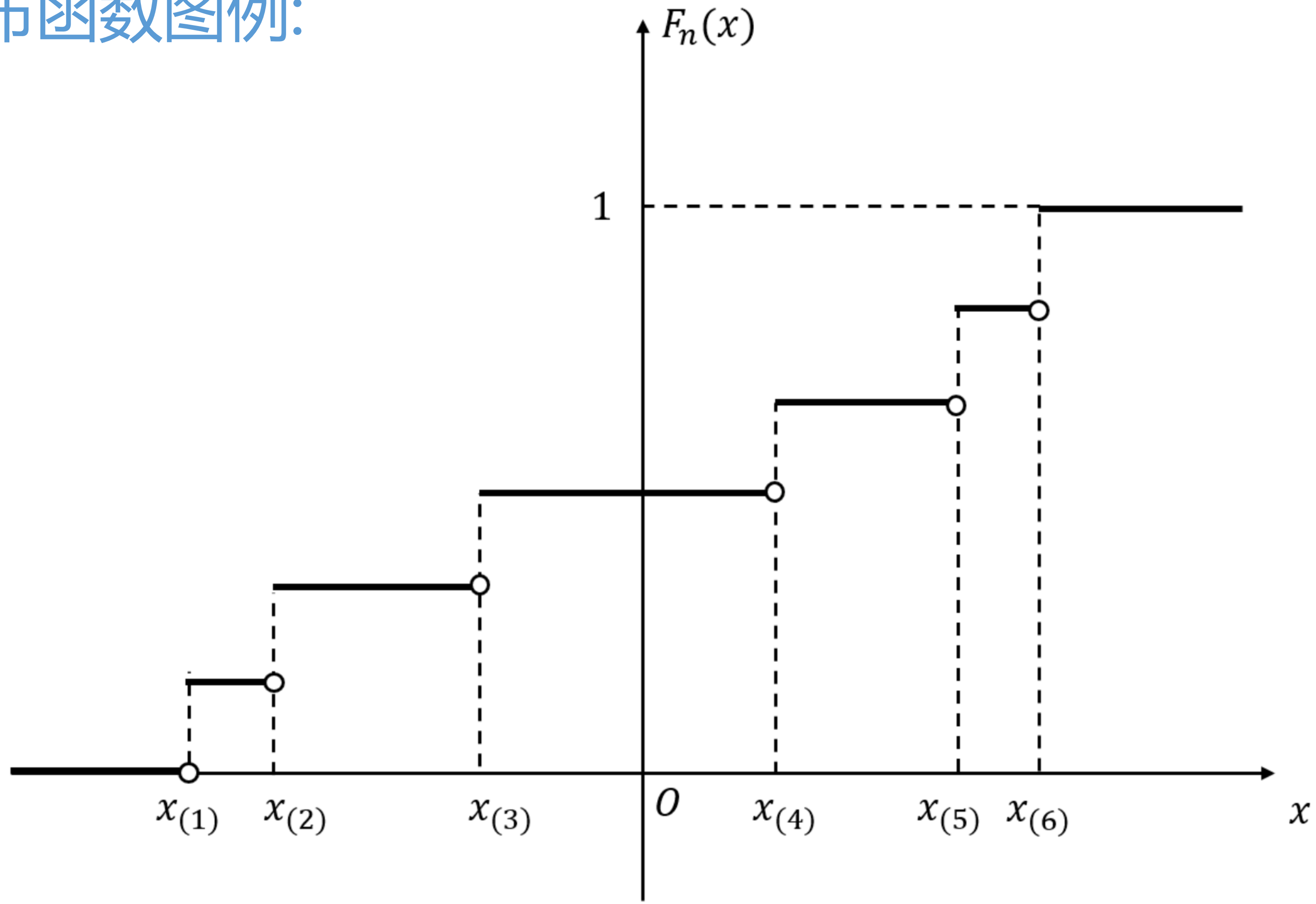

图 5.1: 经验分布函数

抽样分布

定义 5.5 抽样分布

设 $(X_1, X_2, \ldots, X_n)$ 为一个样本, 统计量 $T = T(X_1, X_2, \ldots, X_n)$ 的分布称为抽样分布(sampling distribution). ♣

注意： 样本分布与抽样分布不同

例 5.5 设 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 为来自总体 $X \sim B(1, p)$ 的简单样本, 求 $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ 的抽样分布.

例 5.6 设 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 为来自总体 $X \sim U(0, \theta)$ 的简单样本, 其中 $\theta > 0$ 为参数. 求统计量 $T(\mathbf{X}) = \max_{i \le n} X_i = X_{(n)}$ 的抽样分布.

- 正态总体在统计推断中具有举足轻重的地位. 对正态总体中的参数进行推断时候，常常需要用到样本均值和样本方差及其函数的分布

例：考虑正态分布总体

- 用样本均值$\bar{X}$估计总体均值μ, 需要估计$\bar{X}$与μ的误差，则需要知道$\bar{X}$的分布

- 当用样本方差来构造区间来估计方差范围时，要知道样本方差的分布

- 方差未知的时候，若要检验均值是否等于某个给定数μ_0，要用到统计量$\sqrt{n}(\bar{X}-\mu_0)/S$的分布

- 样本均值和样本方差的分布与“统计三大分布”有着密切关系

定义 5.6 (卡方分布)

设样本 $(X_1, X_2, \ldots, X_n)$ 为标准正态总体的一个简单随机样本, 称

$$X = X_1^2 + X_2^2 + \cdots + X_n^2$$

是自由度为 n 的 χ^2 分布, 记为 $X \sim \chi_n^2$. ♣

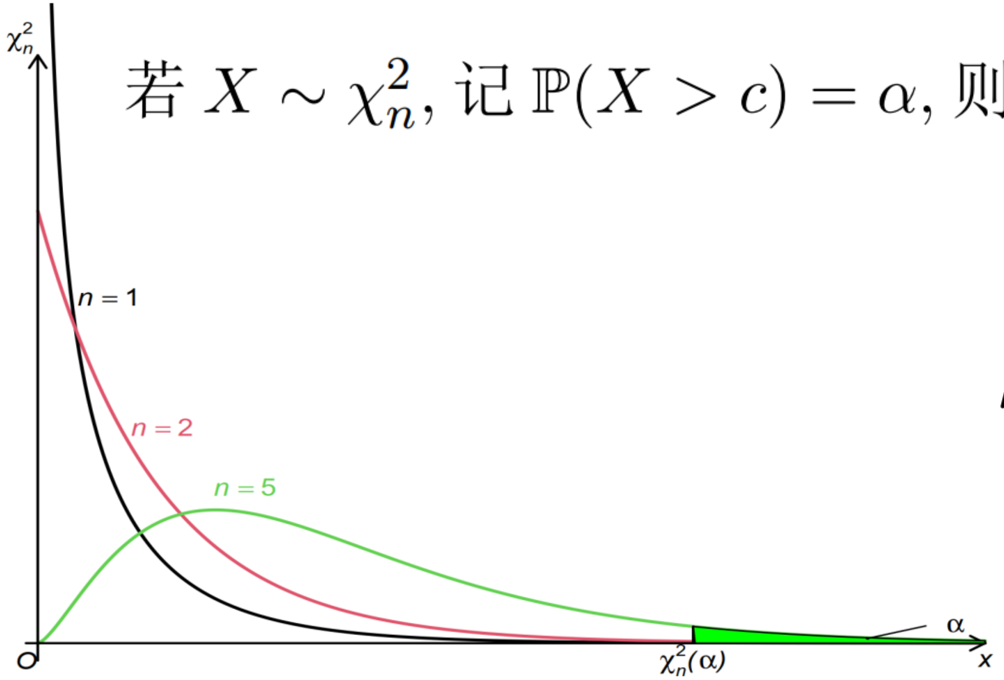

图 5.2: χ^2 分布密度图示

若 $X \sim \chi_n^2$, 记 $\mathbb{P}(X > c) = \alpha$, 则 $c = \chi_n^2(\alpha)$ 称为 χ_n^2 分布的上侧 α 分位数.

χ_n^2 分布的密度函数为

$$k_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} e^{-x/2} x^{(n-2)/2} I_{(0,\infty)}(x).$$

回忆: Gamma函数 $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$.

$\Gamma(x+1) = x\Gamma(x)$, $\Gamma(n) = (n-1)!$,

$\Gamma(1/2) = \sqrt{\pi}$

补：卡方分布密度函数推导

$Y_1 = X_1^2$ 的概率密度函数为 $f_1(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} I(y>0) = \boxed{\frac{1}{2^{1/2}\Gamma(\frac{1}{2})} y^{1/2-1} e^{-y/2} I(y>0)}$

由卷积公式知 $Y_2 = X_1^2 + X_2^2$ 的概率密度函数为

$$
\begin{aligned}
f_2(z) &= \int_R f_1(y) f_2(z-y)dy \\
&= \frac{1}{2\Gamma^2(\frac{1}{2})} e^{-z/2} I(z>0) \int_0^1 t^{-1/2}(1-t)^{-1/2}dt \\
&= \boxed{\frac{1}{2^{2/2}\Gamma(\frac{2}{2})} z^{2/2-1} e^{-z/2} I(z>0)}
\end{aligned}
$$

从而由归纳法，假设 $Y_{n-1} = X_1^2 + \cdots + X_{n-1}^2$ 的概率密度函数为

$$
f_{n-1}(z) = \boxed{\frac{1}{2^{(n-1)/2}\Gamma(\frac{n-1}{2})} z^{(n-1)/2-1} e^{-z/2} I(z>0)}
$$

则 $Y_n = Y_{n-1} + X_n^2$ 的概率密度函数可由卷积公式得

$$
\begin{aligned}
f_n(z) &= \int_R f_{n-1}(y) f_1(z-y)dy \\
&= \frac{1}{2^{n/2}\Gamma(\frac{n-1}{2})\Gamma(\frac{1}{2})} z^{n/2-1} e^{-z/2} I(z>0) \int_0^1 t^{(n-1)/2-1}(1-t)^{1/2-1}dt \\
&= \boxed{\frac{1}{2^{n/2}\Gamma(\frac{n}{2})} z^{n/2-1} e^{-z/2} I(z>0)}
\end{aligned}
$$

性质 χ^2 分布具有以下性质:

(1) 若 $X \sim \chi_n^2$, 容易算得 $\mathbb{E}(X) = n, \mathrm{Var}(X) = 2n$.

(2) 若 $X \sim \chi_m^2, Y \sim \chi_n^2$, 且 X, Y 独立, 则 $Z = X + Y \sim \chi_{m+n}^2$.

再生性

定义 5.7

设 $X \sim N(0,1), Y \sim \chi_n^2$, 且 X, Y 相互独立, 称

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

的分布是自由度为 n 的 t 分布, 记为 $T \sim t_n$ ♣

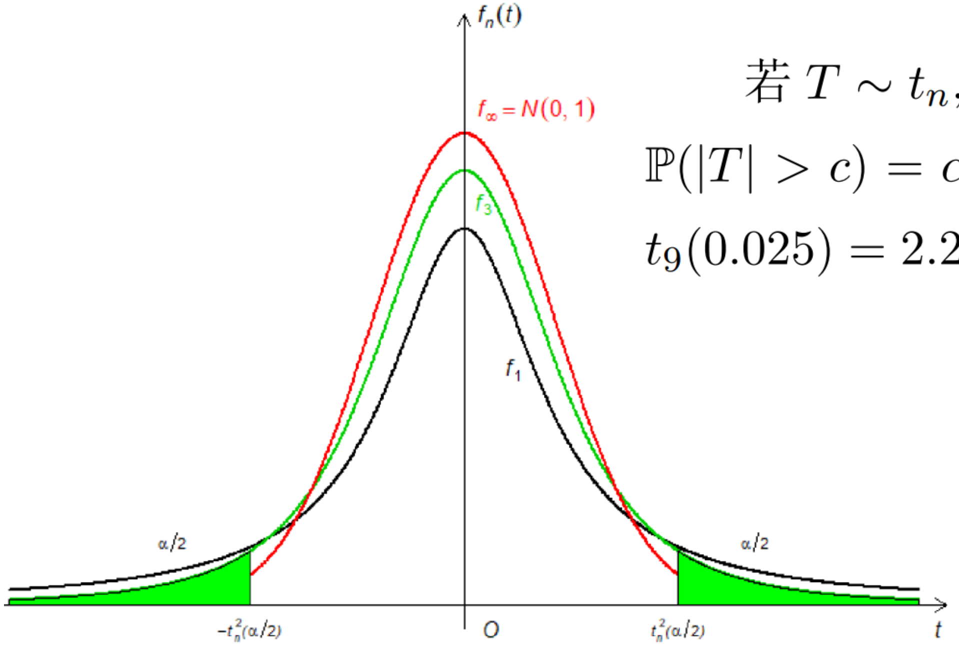

图 5.3: t 分布密度图

若 $T \sim t_n$, 记 $c = t_n(\alpha/2)$ 为自由度为 n 的 t 分布的上 $\alpha/2$ 分位数 (如图5.3所示), 则 $\mathbb{P}(|T| > c) = \alpha$. 当给定 α 时, $t_n(\alpha), t_n(\alpha/2)$ 等可通过查表求出. 例如 $t_{12}(0.05) = 1.782$, $t_9(0.025) = 2.262$ 等.

t_n 分布的密度函数为

$$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \in \mathbb{R}.$$

性质 (1) 当 $n=1$ 时 t_1 就是柯西分布, 此时密度函数为

$$f_1(t)=\frac{1}{\pi(1+t^2)}, t\in\mathbb{R}.$$

(2) 设 $T\sim t_n$, 则当 $n\geq 2$ 时, 由对称性可得 $\mathbb{E}(T)=0$; 当 $n\geq 3$ 时, $\mathrm{Var}(T)=n/(n-2)$.

(3) 当 $n\to\infty$ 时, t_n 分布的密度趋于标准正态密度函数, 即 $\lim_{n\to\infty} f_n(t)=\phi(t)$.

定义 5.8 F 分布

设 $X \sim \chi_m^2, Y \sim \chi_n^2$, 且 X, Y 相互独立, 称

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

的分布是自由度为 m, n 的 F 分布, 记为 $F \sim F_{m,n}$. ♣

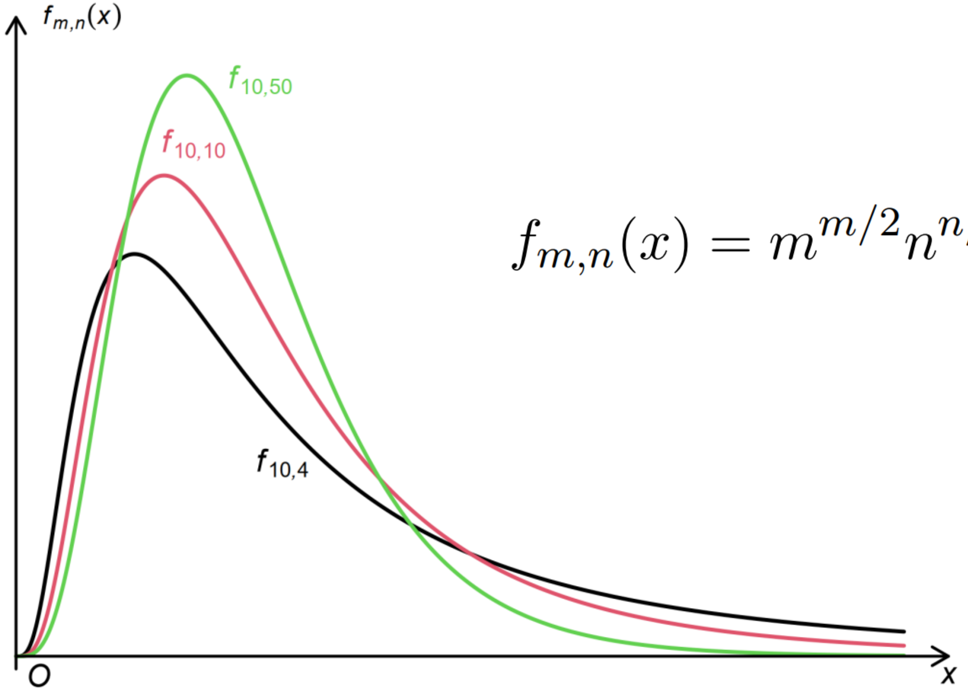

图 5.4: F 分布密度函数图

$F_{m,n}$ 的密度函数为

$$f_{m,n}(x) = m^{m/2} n^{n/2} \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} x^{m/2-1} (mx+n)^{-(m+n)/2} I_{(0,\infty)}(x).$$

性质 F 分布变量具有下列的性质:

(1) 若 $Z \sim F_{m,n}$, 则 $1/Z \sim F_{n,m}$;

(2) 若 $T \sim t_n$, 则 $T^2 \sim F_{1,n}$;

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

若 $F \sim F_{m,n}$, 记 $\mathbb{P}(F > c) = \alpha$, 则 $c = F_{m,n}(\alpha)$ 称为 F 分布的上侧 α 分位数

通常人们把χ^2分布,t分布,F分布这三个分布合称为“统计三大分布”，在正态总体的估计和假设检验中有着重要作用.

定理 5.1

设随机变量 $X_1, \ldots, X_n$ $i.i.d. \sim N(a, \sigma^2)$, $c_1, c_2, \ldots, c_n$ 是不全为零的常数, 则有

(1) 独立的正态随机变量线性组合服从正态分布, 即

$$T = \sum_{k=1}^{n} c_k X_k \sim N\left(a \sum_{k=1}^{n} c_k, \sigma^2 \sum_{k=1}^{n} c_k^2\right)$$

特别, 当 $c_1 = \cdots = c_n = 1/n$, 即 $T = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ 为样本均值时, 有

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

(2) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 为样本方差, 则

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1};$$

(3) $\overline{X}$ 和 S^2 相互独立.

(4) 进而,

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

推论 5.1

设 $X_1, X_2, \ldots, X_m$ i.i.d. $\sim N(a_1, \sigma_1^2)$, $Y_1, Y_2, \ldots, Y_n$ i.i.d. $\sim N(a_2, \sigma_2^2)$, 且假定 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 样本 $X_1, X_2, \ldots, X_m$ 与 $Y_1, Y_2, \ldots, Y_n$ 独立, 则

$$T = \frac{(\overline{X} - \overline{Y}) - (a_1 - a_2)}{S_T} \cdot \sqrt{\frac{mn}{n+m}} \sim t_{n+m-2},$$

此处 $(n+m-2)S_T^2 = (m-1)S_X^2 + (n-1)S_Y^2$, 其中

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

推论 5.2

设 $X_1, X_2, \ldots, X_m \ i.i.d. \sim N(a_1, \sigma_1^2)$, $Y_1, Y_2, \ldots, Y_n \ i.i.d. \sim N(a_2, \sigma_2^2)$, 且合样本 $X_1, X_2, \ldots, X_m$ 和 $Y_1, Y_2, \ldots, Y_n$ 相互独立, 则

$$F = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1},$$

此处 S_X^2 和 S_Y^2 定义如推论5.1所述.

推论 5.3

设 $X_1, X_2, \ldots, X_n$ *i.i.d.* 服从指数分布: $f(x,\lambda) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$, 则有

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

Thank you!