第五章：数理统计的基本概念与抽样分布

4.1 三大分布

能求出抽样分布的确切而且具有简单表达式的情形并不多, 一般都较难. 所幸的是, 在总体分布为正态情形, 许多重要统计量的抽样分布可以求得, 这些多与下面讨论的三种分布有密切关系. 这三个分布在后面几章中有重要应用.

4.1.1 χ^2 分布

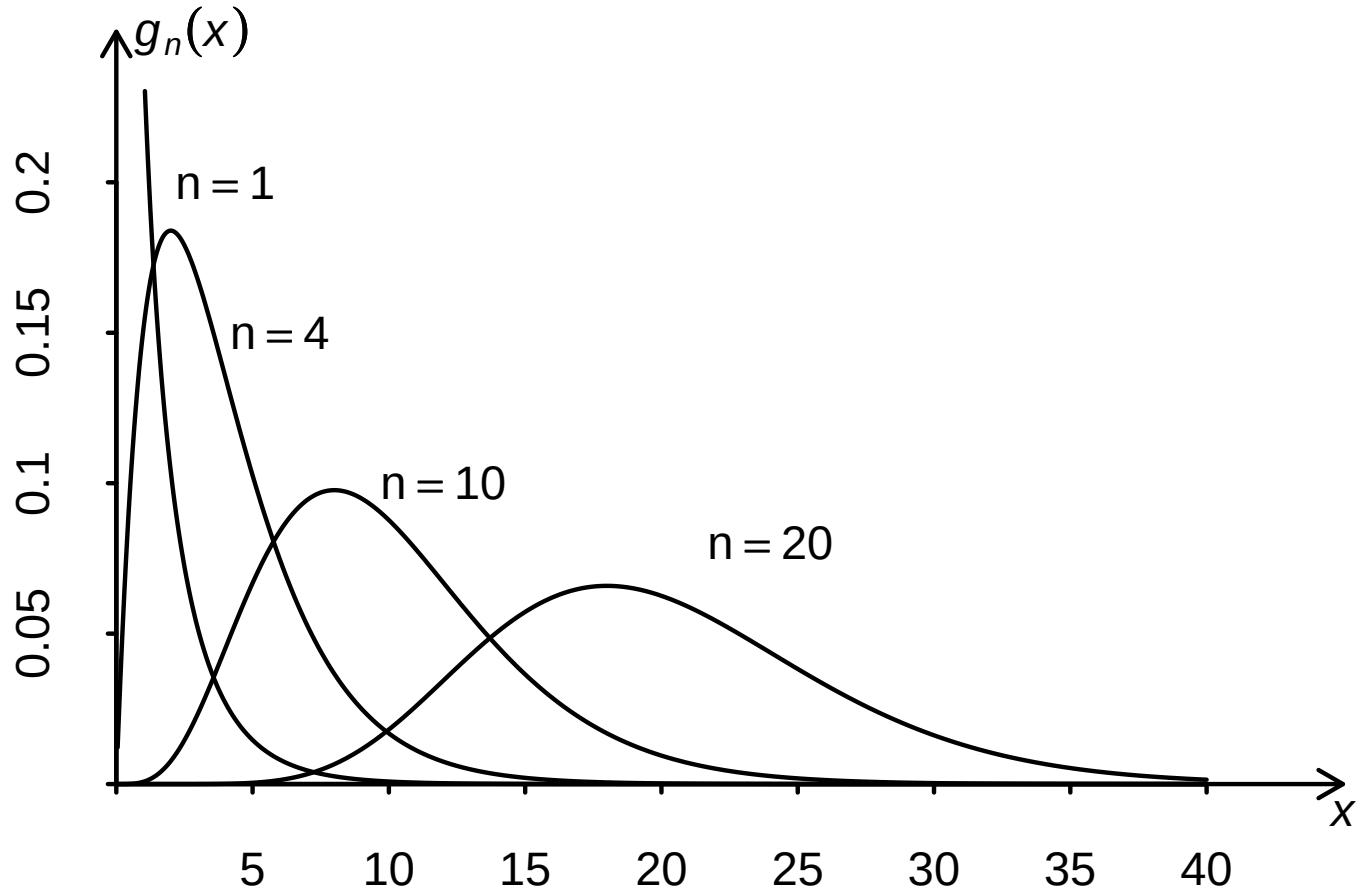
Definition

设 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim N(0,1)$, 令 $X = \sum_{i=1}^{n} X_i^2$, 则称 X 是自由度为 n 的 χ^2 变量，其分布称为自由度为 n 的χ^2 **分布**，记为 $X \sim \chi_n^2$.

设随机变量 X 是自由度为 n 的 χ^2 随机变量, 则其概率密度函数为

$$g_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{4.1}$$

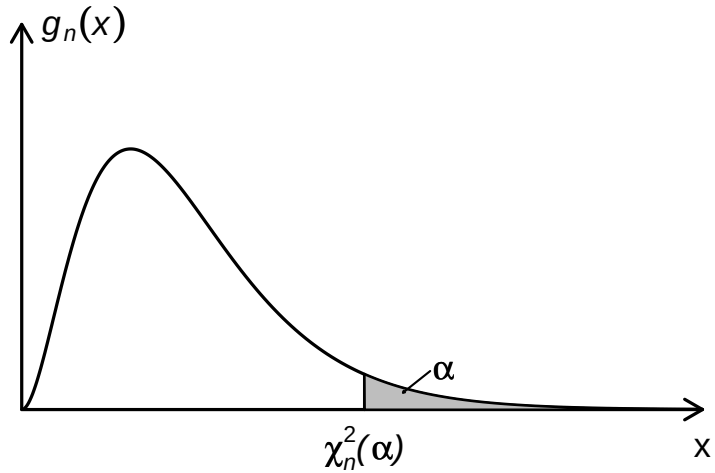
χ_n^2 的密度函数 $g_n(x)$ 形状如下图



χ^2_n 密度函数的支撑集 (即使密度函数为正的自变量的集合) 为 $(0, +\infty)$, 从上图可见当自由度 n 越大, χ^2_n 的密度曲线越趋于对称, n 越小, 曲线越不对称. 当 $n = 1, 2$ 时曲线是单调下降趋于 0. 当 $n \geq 3$

时曲线有单峰, 从 0 开始先单调上升, 在一定位置达到峰值, 然后单下降趋向于 0.

若 $X \sim \chi^2_n$, 记 $P(X > c) = \alpha$, 则 $c = \chi^2_n(\alpha)$ 称为 χ^2_n 分布的上侧 α 分位数, 如下图所示. 当 α 和 n 给定时可查表求出 $\chi^2_n(\alpha)$ 之值, 如 $\chi^2_{10}(0.01) = 23.209$, $\chi^2_5(0.05) = 12.592$ 等.



χ^2 变量具有下列性质:

(1) 设随机变量 $X \sim \chi^2_n$ 则有 $E(X) = n,\ Var(X) = 2n.$

(2) 设 $Z_1 \sim \chi^2_{n_1},\ Z_2 \sim \chi^2_{n_2}$, 且 Z_1 和 Z_2 独立, 则 $Z_1 + Z_2 \sim \chi^2_{n_1+n_2}.$

我们从 X^2 分布的定义出发给出一个简单证明: 由定义 $Z_1 = X_1^2 + \cdots + X_{n_1}^2$, 此处

$$X_1, X_2, \cdots, X_{n_1} \text{ i.i.d.} \sim N(0,1),$$

同理 $Z_2 = X_{n_1+1}^2 + \cdots + X_{n_1+n_2}^2$, 此处

$$X_{n_1+1}, X_{n_1+2}, \cdots, X_{n_1+n_2} \text{ i.i.d.} \sim N(0,1),$$

再由 Z_1 和 Z_2 的独立性可知

$$X_1, X_2, \cdots, X_{n_1}\ X_{n_1+1}, \cdots, X_{n_1+n_2} \text{ i.i.d.} \sim N(0,1).$$

因此

$$Z_1 + Z_2 = X_1^2 + \cdots + X_{n_1}^2 + X_{n_1+1}^2 + \cdots + X_{n_1+n_2}^2.$$

按定义即有 $Z_1 + Z_2 \sim \chi^2_{n_1+n_2}$.

4.1.2 t 分布

设随机变量 $X \sim N(0,1), Y \sim \chi^2_n$, 且 X 和 Y 独立，则称

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

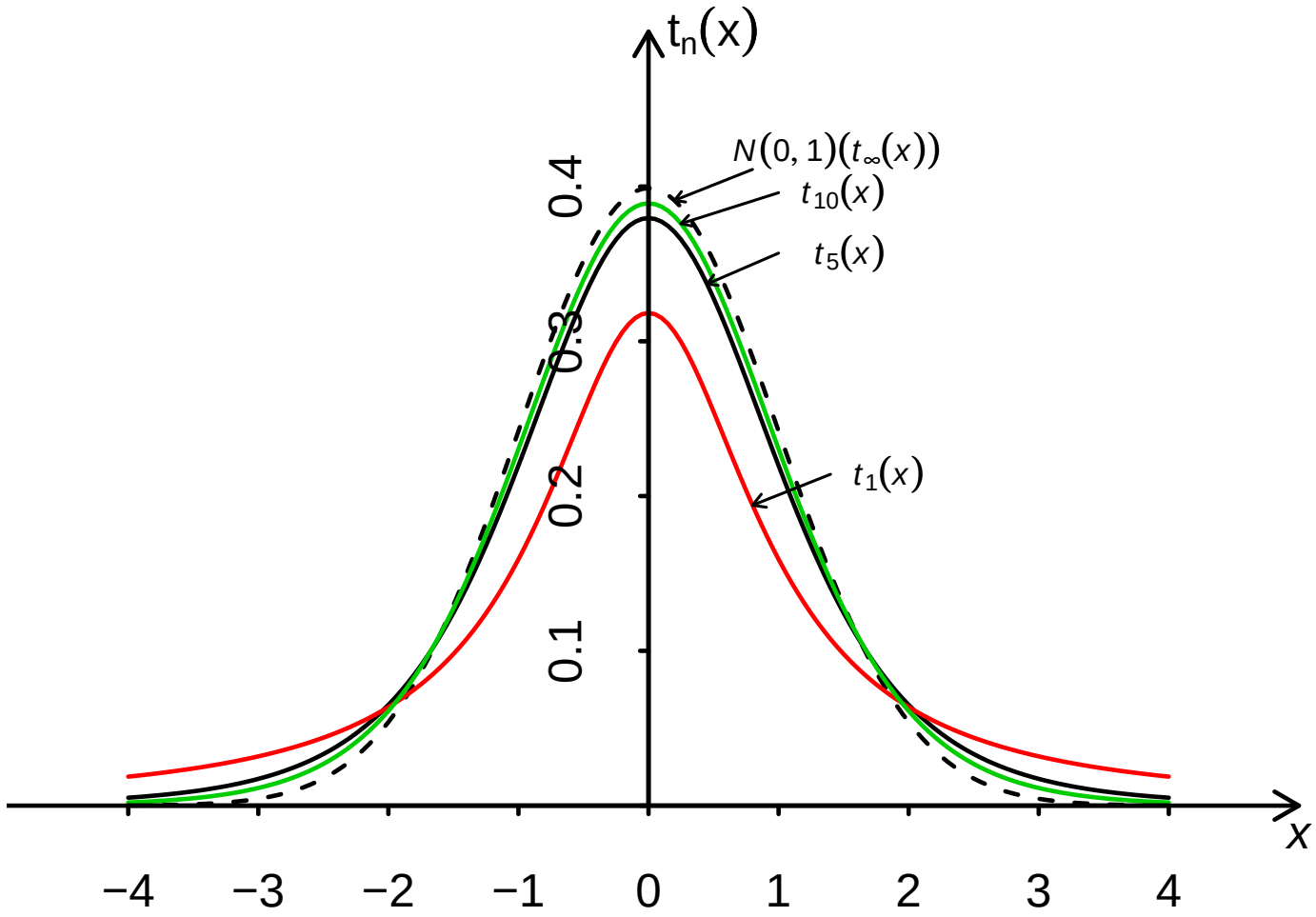
为自由为 n 的 t 变量，其分布称为由为 n 的t **分布**，记为 $T \sim t_n$.

Definition

设随机变量 $T \sim t_n$, 则其密度函数为

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty \qquad (4.2)$$

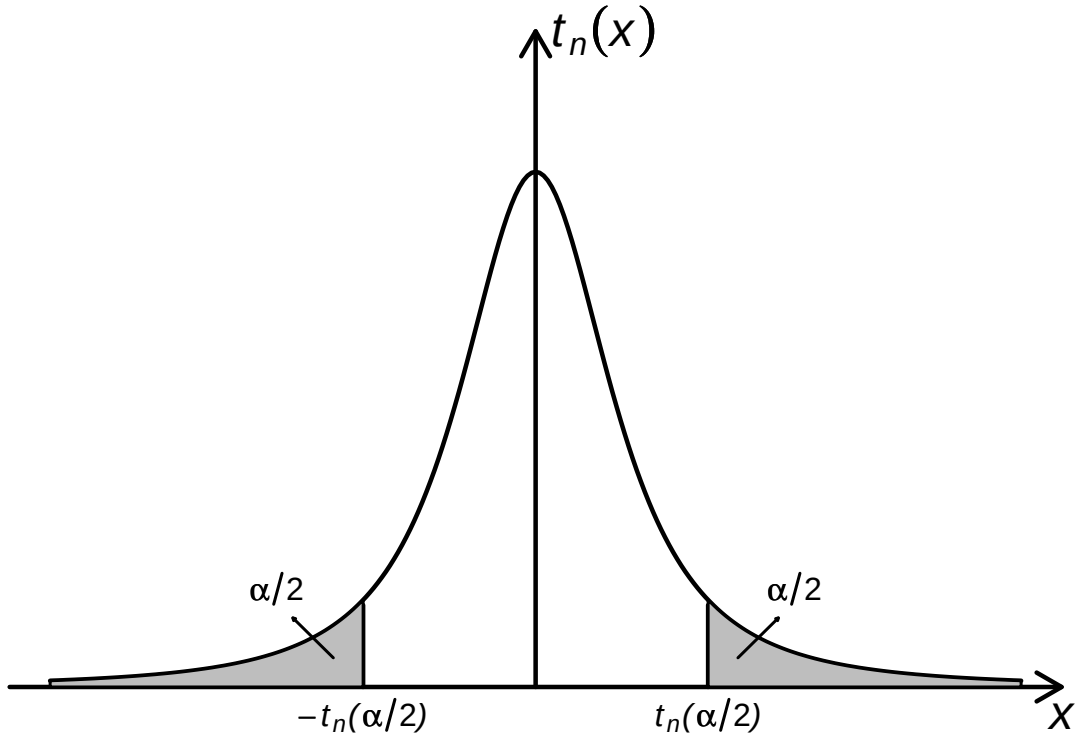
该密度函数的图形如下


$t_n(x)$
$N(0,1)(t_\infty(x))$
$t_{10}(x)$
$t_5(x)$
$t_1(x)$
0.4
0.3
0.2
0.1
−4
−3
−2
−1
0
1
2
3
4
x

t_n 的密度函数与标准正态分布 $N(0,1)$ 密度很相似, 它们都是关于原点对称, 单峰偶函数, 在 $x=0$ 处达到极大. 但 t_n 的峰值低于 $N(0,1)$ 的峰值, t_n 的密度函数尾部都要比 $N(0,1)$ 的两侧尾部粗一些. 容易证明:

$$\lim_{n\to\infty} t_n(x) = \varphi(x)$$

此处 $\varphi(x)$ 是 $N(0,1)$ 变量的密度函数.



若 $T \sim t_n$, 记 $P(|T| > c) = \alpha$, 则 $c = t_n(\alpha/2)$ 为自由度为 n 的 t 分布的双侧 α 分位数 (如上图所示). 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 分布是英国统计学家 W.S. Gosset 在 1908 年以笔名 Student

发表的论文中提出的, 故后人称为 “学生氏 (Student) 分布” 或 “t **分布**”.

t 变量具有下列的性质:

(1) 若随机变量 $T \sim t_n$, 则当 $n \geq 2$ 时, $E(T) = 0$. 当 $n \geq 3$ 时, $Var(T) = \frac{n}{n-2}$.

(2) 当 $n \to \infty$ 时, t 变量的极限分布为 $N(0, 1)$.

4.1.3 F 分布

设随机变量 $X \sim \chi^2_m$, $Y \sim \chi^2_n$, 且 X 和 Y 独立，则称

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

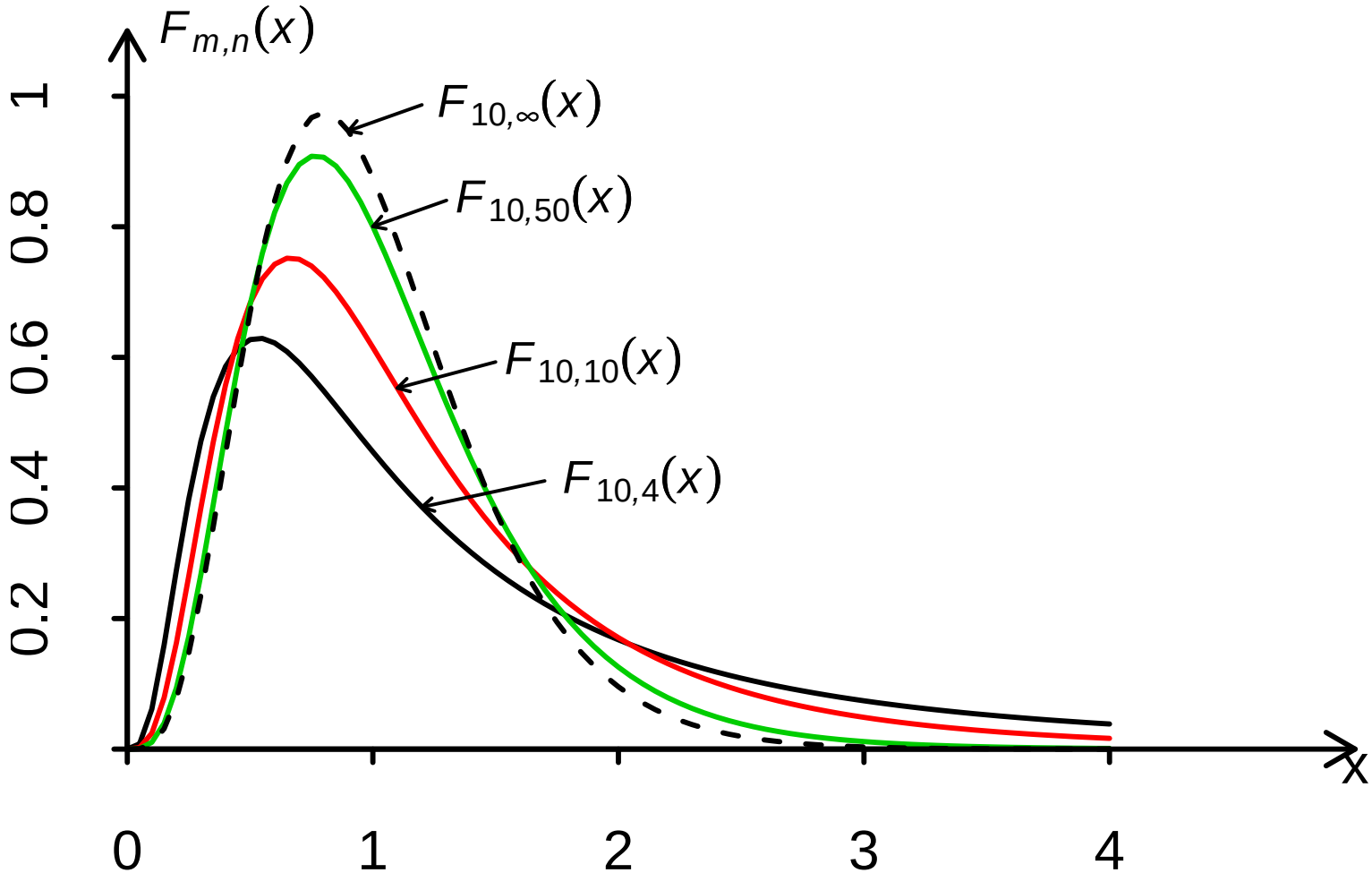
Definition

为自由度分别是 m 和 n 的 F 变量，其分布称为自由度分别是 m 和 n 的F **分布**，记为 $F \sim F_{m,n}$.

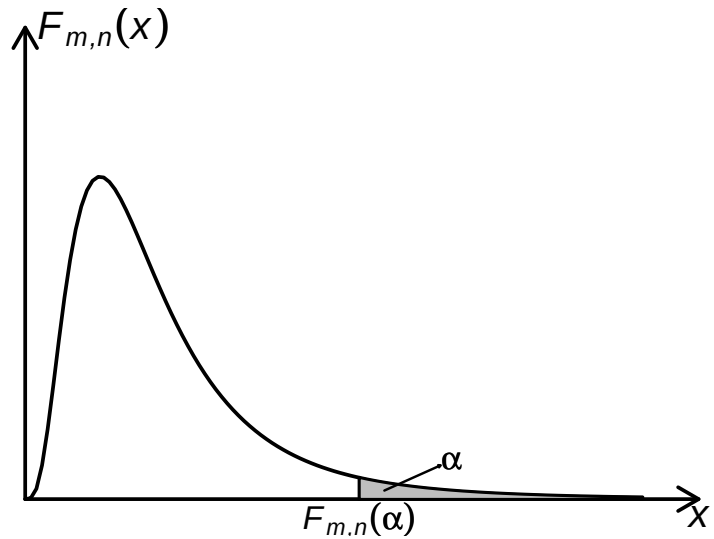
若随机变量 $Z \sim F_{m,n}$, 则其密度函数为

$$f_{m,n}(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (n+mx)^{-\frac{m+n}{2}}, & x > 0, \\ 0, & \text{其它}. \end{cases} \tag{4.3}$$

自由度为 m, n 的 F 分布的密度函数如下图:



注意 F 分布的自由度 m 和 n 是有顺序的, 当 $m \neq n$ 时若将自由度 m 和 n 的顺序颠倒一下, 得到的是两个不同的 F 分布. 从上图可见对给定 $m = 10$, n 取不同值时 $f_{m,n}(x)$ 的形状, 我们看到曲线是偏态的, n 越小偏态越严重.



若 $F \sim F_{m,n}$, 记 $P(F > c) = \alpha$, 则 $c = F_{m,n}(\alpha)$ 称为 F 分布的上侧 α 分位数 (见上图). 当 m, n 和 α 给定时, 可以通过查表求出 $F_{m,n}(\alpha)$ 之值, 例如 $F_{4,10}(0.05) = 3.48$, $F_{10,15}(0.01) = 3.80$ 等. 这

在区间估计和假设检验问题中常常用到.

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

以上性质中 (1) 和 (2) 是显然的, (3) 的证明不难. 尤其性质 (3) 在求区间估计和假设检验问题时会常常用到. 因为当 α 为较小的数, 如 $\alpha = 0.05$ 或 $\alpha = 0.01$, m, n 给定时, 从已有的 F 分布表上查不到 $F_{m,n}(1-0.05)$ 和 $F_{m,n}(1-0.01)$ 之值, 但它们的值可利用性质 (3) 求得, 因为 $F_{n,m}(0.05)$ 和 $F_{n,m}(0.01)$ 是可以通过查 F 分布表求得的.

4.1.4 正态总体样本均值和样本方差的分布

为方便讨论正态总体样本均值和样本方差的分布, 我们先给出正态随机变量的线性函数的分布.

1. 正态变量线性函数的分布

设随机变量 $X_1, \cdots, X_n$ $i.i.d.$ $\sim N(a, \sigma^2)$, $c_1, c_2, \cdots, c_n$ 为常数, 则有

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

特别, 当 $c_1 = \cdots = c_n = 1/n$, 即 $T = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$ 时, 有

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

2. 正态变量样本均值和样本方差的分布

下述定理给出了正态变量样本均值和样本方差的分布和它们的独立性.

定理 1. 设 $X_1, X_2, \cdots, X_n$ *i.i.d.* $\sim N(a, \sigma^2)$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 和 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 分别为样本均值和样本方差, 则有

(1) $\bar{X} \sim N(a, \frac{1}{n}\sigma^2)$;

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

(3) $\bar{X}$ 和 S^2 独立.

4.1.5 几个重要推论

下面几个推论在正态总体区间估计和假设检验问题中有着重要应用.

推论 1. 设 $X_1, X_2, \cdots, X_n$ 相互独立相同分布 *(i.i.d.)* $\sim N(a, \sigma^2)$, 则

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

证: 由注 5.4.3 可知 $\bar{X} \sim N(a, \sigma^2/n)$, 将其标准化得 $\sqrt{n}(\bar{X} - a)/\sigma \sim N(0, 1)$. 又 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, 即 $S^2/\sigma^2 \sim \chi^2_{n-1}/(n-1)$, 且 $\bar{X}$ 和 S^2 独立, 按定义有

$$T = \frac{\sqrt{n}(\bar{X} - a)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X} - a)}{S} \sim t_{n-1}.$$

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

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

此处 $(n+m-2)S_w^2 = (m-1)S_1^2 + (n-1)S_2^2$, 其中

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2.$$

证: 由注 5.4.3 可知 $\bar{X} \sim N(a, \sigma^2/m), \bar{Y} \sim N(a_2, \sigma^2/n)$, 故有 $\bar{X} - \bar{Y} \sim N\big(a_1 - a_2, (\frac{1}{m} + \frac{1}{n})\sigma^2\big) = N\big(a_1 - a_2, \frac{n+m}{mn}\sigma^2\big)$. 将其标准化得

$$\frac{\bar{X} - \bar{Y} - (a_1 - a_2)}{\sigma}\sqrt{\frac{mn}{m+n}} \sim N(0,1). \tag{4.4}$$

又 $(m-1)S_1^2/\sigma^2 \sim \chi_{m-1}^2$, $(n-1)S_2^2/\sigma^2 \sim \chi_{n-1}^2$, 再利用 χ^2 分布的性质可知

$$\frac{(m-1)S_1^2 + (n-1)S_2^2}{\sigma^2} \sim \chi_{n+m-2}^2. \tag{4.5}$$

再由 (4.4) 和 (4.5) 中 $(\bar{X}, \bar{Y})$ 与 (S_1^2, S_2^2) 相互独立, 由定义可知

$$\begin{aligned} T &= \frac{(\bar{X}-\bar{Y})-(a_1-a_2)}{\sigma}\sqrt{\frac{mn}{n+m}}\Big/\sqrt{\frac{(m-1)S_1^2+(n-1)S_2^2}{\sigma^2(n+m-2)}} \\ &= \frac{(\bar{X}-\bar{Y})-(a_1-a_2)}{S_w}\sqrt{\frac{nm}{n+m}} \sim t_{n+m-2}. \end{aligned}$$

推论 3. 设 $X_1, X_2, \cdots, X_m$ *i.i.d.* $\sim N(a_1, \sigma_1^2)$, $Y_1, Y_2, \cdots, Y_n$ *i.i.d.* $\sim N(a_2, \sigma_2^2)$, 且合样本 $X_1, X_2, \cdots, X_m$ 和 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 则

$$F = \frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1},$$

此处 S_1^2 和 S_2^2 定义如推论2所述.

证: 由注 5.4.3 可知 $(m-1)S_X^2/\sigma_1^2 \sim \chi_{m-1}^2$, $(n-1)S_Y^2/\sigma_2^2 \sim \chi_{n-1}^2$,

且二者独立, 由 F 分布的定义可知

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

证毕.

下列这一推论给出了服从指数分布随机变量的线性函数的分布与 χ^2 分布的关系. 这在指数分布总体的区间估计和假设检验问题中有重要应用.

推论 4. 设 $X_1, X_2, \cdots, X_n$ *i.i.d.* 服从指数分布: $f(x,\lambda) = \lambda e^{-\lambda x} I_{[x>0]}$, 则有

$$2\lambda n\bar{X} = 2\lambda \sum_{i=1}^{n} X_i \sim \chi_{2n}^2.$$

证: 首先证明 $2\lambda X_1 \sim \chi^2_2$. 因为

$$F(y) = P(2\lambda X_1 < y) = P\Big(X_1 < \frac{y}{2\lambda}\Big) = \int_0^{\frac{y}{2\lambda}} \lambda e^{-\lambda x} dx,$$

所以

$$f(y) = F^{'}(y) = \begin{cases} \frac{1}{2} e^{-\frac{y}{2}} & \text{当 } y > 0 \\ 0 & \text{当} y \leq 0. \end{cases}$$

因此 $f(y)$ 即为自由度为 2 的 χ^2 密度, 即 $2\lambda X_1 \sim \chi^2_2$.

再利用 χ^2 分布的性质 (3), $2\lambda X_i \sim \chi^2_2,\ i = 1, 2, \cdots, n$; 又它们相互独立, 故有 $2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{2n}$.

4.2 总结

数据在使用前要注意采用有效的方法收集数据, 如设计好抽样方案, 安排好试验等等. 只有有效的收集了数据, 才能有效地使用数据, 开展统计推断工作.

获得数据后, 根据问题的特点和抽样方式确定抽样分布, 即统计模型. 基于统计模型, 统计推断问题可以按照如下的步骤进行:

1. 确定用于统计推断的合适统计量;
2. 寻求统计量的精确分布; 在统计量的精确分布难以求出的情形, 可考虑利用中心极限定理或其它极限定理找出统计量的极限分布.
3. 基于该统计量的精确分布或极限分布, 求出统计推断问题的精确解或近似解.
4. 根据统计推断结果对问题作出解释.

其中第二步是最重要, 但也是最困难的一步. 统计三大分布及正态总体下样本均值和样本方差的分布, 在寻求与正态变量有关的统计量精确分布时, 起着十分重要作用. 尤其在后面两章中求区间估计和假设检验问题时可以看得十分清楚.