

数理统计的基本任务之一是利用样本提供的信息对总体分布中的未知量进行推断, 简单来说, 就是通过样本推断总体. 总体、个体和简单随机样本是数理统计最基本的三个概念, 也与统计量的分布密切相关.

5.1 基本概念

5.1.1 总体与样本

总体是所研究对象的全体. 确切地说, 总体 X 是一个随机变量, 其分布函数 $F(x)$ 被称为总体 X 的分布函数, 而个体则是构成总体的每个单元.

n 个相互独立且与总体 X 具有相同分布的随机变量 $X_1, X_2, \dots, X_n$ 的集合称为来自总体 X 的一个简单随机样本, 简称为样本, 其中 n 称为样本容量. 为了能够通过样本推断总体, 样本必须能够代表总体. 因此, 简单随机样本是具有代表性和独立性的样本. 抽样后, 样本观测值记为 $x_1, x_2, \dots, x_n$. 在实际问题中, $x_1, x_2, \dots, x_n$ 即为给定的数据.

(1) 设 $X_1, X_2, \dots, X_n$ 为总体 X 的一个简单随机样本, 若 X 的分布函数为 $F(x)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合分布函数为

$$F^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F(x_i).$$

(2) 若连续型随机变量 X 的概率密度函数为 $f(x)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合概率密度函数为

$$f^*(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i).$$

(3) 若离散型随机变量 X 的分布律为 $P\{X_k = x_k\} = p_k (k = 1, 2, \dots)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合分布律为

$$P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = \prod_{i=1}^n P\{X_i = x_i\}.$$

5.1.2 统计量及其分布

设 $(X_1, X_2, \dots, X_n)$ 为取自总体 X 的一个样本, 则 $X_1, X_2, \dots, X_n$ 是独立同分布的随机变量, 且每一个 $X_i (i = 1, 2, \dots, n)$ 的分布都与总体 X 的分布相同. 样本 $(X_1, X_2, \dots, X_n)$ 的函数为 $g(X_1, X_2, \dots, X_n)$, 若函数 g 中不直接包含总体分布中的任何未知参数, 则称 $g(X_1, X_2, \dots, X_n)$ 为统计量.

在抽样前, 统计量是一个随机变量. 在抽样后, 若得到样本 $(X_1, X_2, \dots, X_n)$ 的一次观测值 $(x_1, x_2, \dots, x_n)$, 则 $g(x_1, x_2, \dots, x_n)$ 为统计量的一次观测值, 它可以由实际数据计算得到. 我们构造统计量的主要目的就是估计总体分布中的未知参数.

接下来, 我们给出两个常用的统计量: 样本均值和样本方差.

【知识点 5.1】 设 $(X_1, X_2, \dots, X_n)$ 为取自总体的一个样本, 则样本均值和样本方差分别为

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2. \quad (5.1.1)$$

按照我们的直觉均值应该是先求 n 个数之和, 再除以个数 n . 但是, 为什么样本方差的分母是 $n-1$, 而不是 n 呢?

一个直观的解释: 在 $\bar{X}$ 确定后, n 个偏差 $X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}$ 中只有 $n-1$ 个偏差可以自由变动, 而第 n 个则不能自由取值, 因为 $\sum_{i=1}^n (X_i - \bar{X}) = 0$. 这也解释了为什么本书后续章节经常用到的统计量 $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ 服从自由度为 $n-1$ 的 t 分布.

或许有同学会问, 为什么我们不用

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (5.1.2)$$

来定义样本方差呢? 这里 μ 是总体的均值. 如果用式 (5.1.2) 来定义方差: 设总体的方差为 σ^2 , 则根据方差的定义可得 $E[(X_i - \mu)^2] = \sigma^2, i = 1, 2, \dots, n$, 故 $E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right] = \sigma^2$. 这样定义不是很准确吗?

我们之所以不用 S_n^2 作为样本方差, 一是因为它需要已知随机变量 X 的数学期望 μ ; 二是因为它与 $\bar{X}$ 不独立 (证明较为复杂). 因此, 我们将 S^2 作为样本方差.

样本均值和样本方差有如下重要性质.

【知识点 5.2】 设总体 X 的均值为 μ , 方差为 σ^2 , $(X_1, X_2, \dots, X_n)$ 为取自该总体的一个样本, 样本均值 $\bar{X}$ 和样本方差 S^2 的定义见式(5.1.1), 则

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

【证明】 由期望和方差的定义和性质 $X_1, X_2, \dots, X_n$ 相互独立可知

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu,$$

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

这证明了前两个等式, 对于第三个等式, 由 $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$ 可得

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

将 $E(X_i^2) = D(X_i) + [E(X_i)]^2$, $E(\bar{X}^2) = D(\bar{X}) + [E(\bar{X})]^2$ 代入上式, 可得

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

【例 5.1】 (2015 · 数三 8) 设总体 $X \sim b(m, \theta)$, $X_1, X_2, \dots, X_n$ 为来自该总体的简单随机样本, $\bar{X}$ 为样本均值, 则 $E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = (\quad)$.

A. $(m-1)n\theta(1-\theta)$

B. $m(n-1)\theta(1-\theta)$

C. $(m-1)(n-1)\theta(1-\theta)$

D. $mn\theta(1-\theta)$

【解析】 因为 $X \sim b(m, \theta)$, 所以 $D(X) = m\theta(1-\theta)$. 由 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 和 $E(S^2) = D(X)$ 可得

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

故选 B.

【例 5.2】 (2001 · 数一 23) 设总体 X 服从正态分布 $N(\mu, \sigma^2)$ ($\sigma > 0$), 从该总体中抽取简单随机样本 $X_1, X_2, \dots, X_{2n}$ ($n \geq 2$), 其样本均值为 $\bar{X} = \frac{1}{2n} \sum_{i=1}^{2n} X_i$, 求统计量 $Y = \sum_{i=1}^n (X_i + X_{n+i} - 2\bar{X})^2$ 的数学期望 $E(Y)$.

【解析】 设 $\bar{X}_1 = \frac{1}{n} \sum_{i=1}^n X_i$, $\bar{X}_2 = \frac{1}{n} \sum_{i=1}^n X_{n+i}$, 则 $2\bar{X} = \bar{X}_1 + \bar{X}_2$. 因此

$$\begin{aligned} E(Y) &= E\left[\sum_{i=1}^n (X_i + X_{n+i} - 2\bar{X})^2\right] = E\left\{\sum_{i=1}^n [(X_i - \bar{X}_1) + (X_{n+i} - \bar{X}_2)]^2\right\} \\ &= E\left\{\sum_{i=1}^n [(X_i - \bar{X}_1)^2 + 2(X_i - \bar{X}_1)(X_{n+i} - \bar{X}_2) + (X_{n+i} - \bar{X}_2)^2]\right\} \\ &= E\left[\sum_{i=1}^n (X_i - \bar{X}_1)^2\right] + 2E\left[\sum_{i=1}^n (X_i - \bar{X}_1)(X_{n+i} - \bar{X}_2)\right] + E\left[\sum_{i=1}^n (X_{n+i} - \bar{X}_2)^2\right]. \end{aligned}$$

因为 $X_1 - \bar{X}_1$ 是 $\{X_i\}_{i=1}^n$ 的线性组合, $X_2 - \bar{X}_2$ 是 $\{X_{n+i}\}_{i=1}^n$ 的线性组合, 且 $\{X_i\}_{i=1}^{2n}$ 相互独立, 所以 $X_i - \bar{X}_1$ 和 $X_{n+i} - \bar{X}_2$ 相互独立, 于是

$$E[(X_i - \bar{X}_1)(X_{n+i} - \bar{X}_2)] = E[(X_i - \bar{X}_1)]E[(X_{n+i} - \bar{X}_2)] = 0.$$

故

$$\begin{aligned} E(Y) &= E\left[\sum_{i=1}^n (X_i - \bar{X}_1)^2\right] + E\left[\sum_{i=1}^n (X_{n+i} - \bar{X}_2)^2\right] \\ &= (n-1)\sigma^2 + (n-1)\sigma^2 = 2(n-1)\sigma^2. \end{aligned}$$

【变式 5.2.1】 (2009 · 数三 14) 设 $X_1, X_2, \dots, X_m$ 为来自二项分布总体 $b(n, p)$ 的简单随机样本, $\bar{X}$ 和 S^2 分别为样本均值和样本方差, 记统计量 $T = \bar{X} - S^2$, 则 $E(T) = \underline{\hspace{2cm}}$.

【变式 5.2.2】 (2011 · 数三 8) 设总体 X 服从参数为 $\lambda (\lambda > 0)$ 的泊松分布, $X_1, X_2, \dots, X_n (n \geq 2)$ 为来自总体的简单随机样本, 对应的统计量有 $T_1 = \frac{1}{n} \sum_{i=1}^n X_i, T_2 = \frac{1}{n-1} \sum_{i=1}^{n-1} X_i + \frac{1}{n} X_n$, 故 ().

A. $E(T_1) > E(T_2), D(T_1) > D(T_2)$

B. $E(T_1) > E(T_2), D(T_1) < D(T_2)$

C. $E(T_1) < E(T_2), D(T_1) > D(T_2)$

D. $E(T_1) < E(T_2), D(T_1) < D(T_2)$

【变式 5.2.3】 (2010 · 数三 14) 设 $X_1, X_2, \dots, X_n$ 为来自总体 $N(\mu, \sigma^2) (\sigma > 0)$ 的简单随机样本, 统计量 $T = \frac{1}{n} \sum_{i=1}^n X_i^2$, 则 $E(T) = \underline{\hspace{2cm}}$.

【变式 5.2.4】 (2025 · 数一 9、数三 9) 设 $X_1, X_2, \dots, X_{20}$ 是来自总体 $B(1, 0.1)$ 的简单随机样本, 令 $T = \sum_{i=1}^{20} X_i$, 利用泊松分布近似表示二项分布的方法可得 $P\{T \leq 1\} \approx (\quad)$.

A. $\frac{1}{e^2}$ B. $\frac{2}{e^2}$ C. $\frac{3}{e^2}$ D. $\frac{4}{e^2}$

5.1.3 经验分布函数 (仅数三作要求)

设 $x_1, x_2, \dots, x_n$ 是取自总体分布函数为 $F(x)$ 的样本, 若将样本观测值由小到大进行排列, 记为 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, 则 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ 称为有序样本, 用有序样本定义如下函数

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}, \\ \frac{k}{n}, & \text{当 } x_{(k)} \leq x < x_{(k+1)}, \quad k = 1, 2, \dots, n-1, \\ 1, & \text{当 } x \geq x_{(n)}, \end{cases}$$

则 $F_n(x)$ 是一非减右连续函数, 且满足 $F_n(-\infty) = 0$ 和 $F_n(\infty) = 1$. 由此可见, $F_n(x)$ 是一个分布函数且为 $F(x)$ 的近似函数, 称 $F_n(x)$ 为该样本的经验分布函数.

对每一固定的 x , $F_n(x)$ 是样本中事件 $\{x_i \leq x\}$ 发生的频率. 当 n 固定时, $F_n(x)$ 是样本的函数, 它是一个随机变量. 若对任意给定的实数 x , 定义

$$I_i(x) = \begin{cases} 1, & x_i \leq x, \\ 0, & x_i > x, \end{cases}$$

则由经验分布函数的定义可以看出, 对任意给定的实数 x

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

注意到 $I_i(x) (i = 1, 2, 3, \dots)$ 是独立同分布的随机变量, 其共同分布为 $B(1, F(x))$, 由伯努利大数定律可知, 只要 n 相当大, $F_n(x)$ 就依概率收敛于 $F(x)$. 更深刻的结果也是存在的, 这就是格利文科定理, 它表明, 当 n 相当大时, 经验分布函数是总体分布函数 $F(x)$ 的一个良好的近似. 经典统计学中一切统计推断都以样本为依据, 其理由就在于此.

【例 5.3】 (2025 · 数三 10) 设总体 X 的均匀分布为 $F(x)$, $x_1, x_2, \dots, x_n$ 为来自总体 X 的简单随机样本, 样本经验分布函数为 $F_n(x)$, 对于给定的 $x (0 < F(x) < 1)$, $D(F_n(x)) = (\quad)$.

- A. $F(x)[1 - F(x)]$ B. $[F(x)]^2$ C. $\frac{1}{n}F(x)[1 - F(x)]$ D. $\frac{1}{n}[F(x)]^2$

【解析】 将样本观测值由小到大进行排列, 记为 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, 则经验分布函数为

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)}, \\ \frac{k}{n}, & \text{当 } x_{(k)} \leq x < x_{(k+1)}, \quad k = 1, 2, \dots, n-1, \\ 1, & \text{当 } x \geq x_{(n)}, \end{cases}$$

记 $I_i = \begin{cases} 1, & x \geq x_i, \\ 0, & x < x_i, \end{cases} i = 1, 2, \dots, n$; 此时 $I_i \sim B(1, p)$, $p = F(x)$, 即 $E(I_i) = p = F(x)$, $D(I_i) = p(1-p) =$

$F(x)[1 - F(x)]$, 而 $F_n(x) = \frac{1}{n} \sum_{i=1}^n I_i$, 所以

$$D(F_n(x)) = D\left(\frac{1}{n} \sum_{i=1}^n I_i\right) = \frac{1}{n} D(I_i) = \frac{1}{n} F(x)[1 - F(x)].$$

故选 C.

5.2 三大分布

在实践中, 我们很快会认识到, 许多统计推断都基于正态分布的假设. 有三个著名的分布, 它们以标准正态变量为基础, 并在实际应用中得到了广泛的运用. 这些分布之所以受到重视, 不仅因为它们有清晰的含义, 还因为它们抽样分布的密度函数具有明确的表达式. 这三个分布统称为统计学的“三大抽样分布”, 即 χ^2 分布、 t 分布和 F 分布. 同学们不必死记 χ^2 分布、 t 分布和 F 分布的概率密度函数的表达式, 只需了解这些分布的典型特征, 以及它们的分布曲线的示意图和分位数的概念, 并学会如何查阅相应的分位数数值表即可.

5.2.1 χ^2 分布

设 $X_1, X_2, \dots, X_n$ 为相互独立的、服从标准正态分布的随机变量, 称随机变量 $Y = X_1^2 + X_2^2 + \dots + X_n^2$ 服从自由度为 n 的 χ^2 分布, 记为 $Y \sim \chi^2(n)$. χ^2 分布的密度函数图像如图5-1所示.

【知识点 5.3】 当 $Y \sim \chi^2(n)$ 时, $E(Y) = n$, $D(Y) = 2n$.

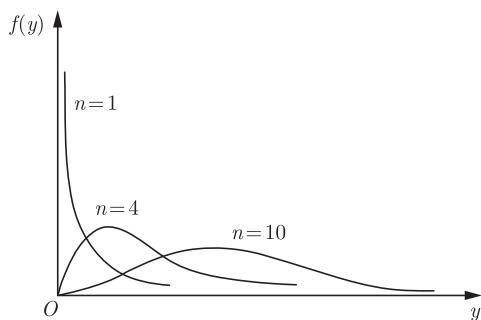


图 5-1 $\chi^2(n)$ 的密度函数

【证明】 设 $Y = \sum_{i=1}^n X_i^2$, 其中 $X_i \sim N(0, 1)$, 则

$$E(Y) = E\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n E(X_i^2) = n,$$

$$D(Y) = D\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n D(X_i^2)$$

$$= \sum_{i=1}^n [E(X_i^4) - (E(X_i^2))^2]$$

$$= \sum_{i=1}^n (3 - 1^2) = 2n.$$

【知识点 5.4】 (χ^2 分布具有可加性) 设 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则 $X + Y \sim \chi^2(m+n)$.

【证明】 不妨记 $X = \sum_{i=1}^m X_i^2 \sim \chi^2(m)$, $Y = \sum_{i=1}^n Y_i^2 \sim \chi^2(n)$, 其中 $X_1, X_2, \dots, X_m$ 独立同分布于 $N(0, 1)$, $Y_1, Y_2, \dots, Y_n$ 相互独立同分布于 $N(0, 1)$. 又因为 X 与 Y 相互独立, 故 $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$ 相互独立, 则

$$X + Y = \sum_{i=1}^m X_i^2 + \sum_{i=1}^n Y_i^2 \sim \chi^2(m+n).$$

【注】 类似具有可加性的分布还有二项分布、泊松分布和正态分布.

【例 5.4】 (1998 • 数三 5) 设 X_1, X_2, X_3, X_4 是来自正态总体 $N(0, 2^2)$ 的简单随机样本, $X = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$, 其中 $a, b \neq 0$, 则当 $a = \underline{\hspace{2cm}}$, $b = \underline{\hspace{2cm}}$ 时, 统计量 X 服从 χ^2 分布, 其自由度为 $\underline{\hspace{2cm}}$.

【解析】 因为 X_1, X_2, X_3, X_4 相互独立且均服从 $N(0, 2^2)$, 则 $X_1 - 2X_2 \sim N(0, 20)$, $3X_3 - 4X_4 \sim N(0, 100)$, 且两者相互独立, 于是

$$\frac{1}{\sqrt{20}}(X_1 - 2X_2) \sim N(0, 1), \quad \frac{1}{\sqrt{100}}(3X_3 - 4X_4) \sim N(0, 1).$$

由 χ^2 的定义可知, 当 $a = \frac{1}{20}$, $b = \frac{1}{100}$ 时, 有

$$X = \frac{1}{20}(X_1 - 2X_2)^2 + \frac{1}{100}(3X_3 - 4X_4)^2 \sim \chi^2(2),$$

即 $a = \frac{1}{20}$, $b = \frac{1}{100}$ 时, X 服从 χ^2 分布, 且自由度为 $n = 2$.

χ^2 分布之所以实用的一个重要原因是如下定理的存在.

【知识点 5.5】 设 $(X_1, X_2, \dots, X_n)$ 是取自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 样本均值和样本方差分别为 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$, 则

- (1) $\bar{X}$ 与 S^2 相互独立; (2) $\bar{X}$ 与 S_n^2 不独立; (3) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;
 (4) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$; (5) $\frac{nS_n^2}{\sigma^2} \sim \chi^2(n)$.

【变式 5.4.1】 (2017 • 数一 8、数三 8) 设 $X_1, X_2, \dots, X_n (n \geq 2)$ 来自总体 $N(\mu, 1)$ 的简单随机样本, 记 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 则下列结论中不正确的是 ().

A. $\sum_{i=1}^n (X_i - \mu)^2$ 服从 χ^2 分布

B. $2(X_n - X_1)^2$ 服从 χ^2 分布

C. $\sum_{i=1}^n (X_i - \bar{X})^2$ 服从 χ^2 分布

D. $n(\bar{X} - \mu)^2$ 服从 χ^2 分布

【变式 5.4.2】 (2008 · 数一 23、数三 23) 设 $X_1, X_2, \dots, X_n$ 是总体为 $N(0, 1)$ 的简单随机样本. 记 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, T = \bar{X}^2 - \frac{1}{n} S^2$, 求 $D(T)$.

【分析】 注意到 $\bar{X}$ 与 S^2 相互独立. 如果用公式

$$D(T) = E(T^2) - (E(T))^2 = E(T^2) = E\left(\bar{X}^4 - \frac{2}{n} \bar{X}^2 \cdot S^2 + \frac{S^4}{n^2}\right),$$

计算会很繁杂. 如果利用 $\bar{X}, S^2$ 的独立性, 那么

$$D(T) = D\left(\bar{X}^2 - \frac{1}{n} S^2\right) = D(\bar{X}^2) + \frac{1}{n^2} D(S^2),$$

再计算

$$D(\bar{X}^2) = D\left(\frac{1}{n^2} \sum_{i=1}^n X_i^2 + \frac{1}{n^2} \sum_{i \neq j} X_i X_j\right)$$

和 $D(S^2)$ 更困难.

注意到 $\sqrt{n}\bar{X} \sim N(0, 1)$, 即 $n\bar{X}^2 \sim \chi^2(1)$, $(n-1)S^2 \sim \chi^2(n-1)$, 而 $D(\chi^2(n)) = 2n$. 再来计算 $D(T) = D(\bar{X}^2) + \frac{1}{n^2} D(S^2)$ 就不难了.

5.2.2 t 分布

设随机变量 X 与 Y 相互独立, 且 $X \sim N(0, 1), Y \sim \chi^2(n)$, 则称 $T = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 记为 $T \sim t(n)$, 其密度函数图像 (见图 5-2) 关于直线 $t = 0$ 对称, 当 n 充分大时, 其图形类似于标准正态分布的密度函数图像, 如图 5-3 所示. 事实上, 当 n 充分大时, $t(n)$ 分布近似于 $N(0, 1)$.

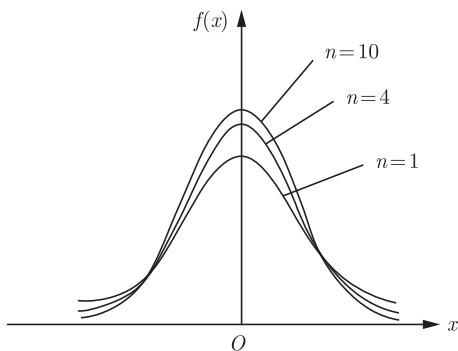


图 5-2 $t(n)$ 的密度函数

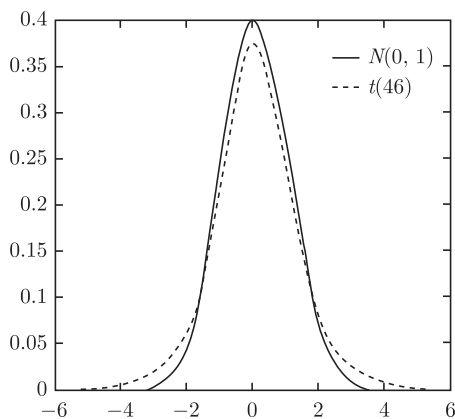


图 5-3 标准正态分布和 t 分布的密度函数

【例 5.5】 (2014·数三 8) 设 X_1, X_2, X_3 为来自正态总体 $N(0, \sigma^2)$ 的简单随机样本, 则统计量 $S = \frac{X_1 - X_2}{\sqrt{2}|X_3|}$ 服从的分布为 ().

A. $F(1, 1)$ B. $F(2, 1)$ C. $t(1)$ D. $t(2)$

【解析】 因为 X_1, X_2, X_3 为来自正态总体 $N(0, \sigma^2)$ 的简单随机样本, 所以 $X_1 - X_2 \sim N(0, 2\sigma^2)$, 于是 $U = \frac{X_1 - X_2}{\sqrt{2}\sigma} \sim N(0, 1)$. 又因为 $X_3 \sim N(0, \sigma^2)$, 所以 $\frac{X_3}{\sigma} \sim N(0, 1)$, 于是 $V = \frac{X_3^2}{\sigma^2} \sim \chi^2(1)$. 因为 $\frac{X_1 - X_2}{\sqrt{2}\sigma}$ 与 $\frac{X_3^2}{\sigma^2}$ 相互独立, 所以由 t 分布可得

$$\frac{X_1 - X_2}{\sqrt{2}|X_3|} = \frac{\frac{X_1 - X_2}{\sqrt{2}\sigma}}{\sqrt{\frac{X_3^2}{\sigma^2}/1}} \sim t(1),$$

故选 C.

【变式 5.5.1】 (2012·数三 8) 设 X_1, X_2, X_3, X_4 为来自总体 $N(1, \sigma^2)$ ($\sigma > 0$) 的简单随机样本, 则统计量 $\frac{X_1 - X_2}{|X_3 + X_4 - 2|}$ 的分布为 ().

A. $N(0, 1)$ B. $t(1)$ C. $\chi^2(1)$ D. $F(1, 1)$

【变式 5.5.2】 (1997·数三 5) 设随机变量 X 和 Y 相互独立且都服从正态分布 $N(0, 3^2)$, 而 $X_1, X_2, \dots, X_9$ 和 $Y_1, Y_2, \dots, Y_9$ 分别是来自总体 X 和 Y 的简单随机样本, 则统计量 $Z = \frac{X_1 + X_2 + \dots + X_9}{\sqrt{Y_1^2 + Y_2^2 + \dots + Y_9^2}}$ 服从 _____ 分布, 参数为 _____.

【知识点 5.6】 设 $(X_1, X_2, \dots, X_n)$ 是取自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$,

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

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

【证明】 因为 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, 且 $\bar{X}$ 与 S^2 相互独立, 根据 t 分布的定义可得

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

【注】 若记 $S_n = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$, 则 $\frac{nS_n^2}{\sigma^2} \sim \chi^2(n)$. 能否由 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, 得到

$$\frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS_n^2}{\sigma^2}/n}} = \frac{\bar{X} - \mu}{S_n/\sqrt{n}} \sim t(n)?$$

答案是不能, 因为 $\bar{X}$ 与 S_n^2 不独立.

【例 5.6】 (2018 • 数三 8) 已知 $X_1, X_2, \dots, X_n$ 为来自总体 $X \sim N(\mu, \sigma^2)$ 的简单随机样本, 令

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}, S^* = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}, \text{ 则 } ().$$

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

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

C. $\frac{\sqrt{n}(\bar{X} - \mu)}{S^*} \sim t(n)$

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

【解析】 由知识点5.6的分析可知只有 B 正确. 故选 B.

【例 5.7】 (1999 • 数三 23) 设 $X_1, X_2, \dots, X_9$ 是来自正态总体 X 的简单随机样本, $Y_1 = \frac{1}{6}(X_1 + X_2 + \dots + X_6)$, $Y_2 = \frac{1}{3}(X_7 + X_8 + X_9)$, $S^2 = \frac{1}{2} \sum_{i=7}^9 (X_i - Y_2)^2$, $Z = \frac{\sqrt{2}(Y_1 - Y_2)}{S}$, 证明统计量 Z 服从自由度为 2 的 t 分布.

【解析】 因为服从正态分布的独立随机变量其线性组合也服从正态分布, 所以

$$Y_1 = \frac{1}{6}(X_1 + X_2 + \dots + X_6) \sim N\left(\mu, \frac{\sigma^2}{6}\right),$$

$$Y_2 = \frac{1}{3}(X_7 + X_8 + X_9) \sim N\left(\mu, \frac{\sigma^2}{3}\right).$$

由于 Y_1, Y_2 独立且都服从正态分布, 故 $Y_1 - Y_2$ 也服从正态分布, 且 $Y_1 - Y_2 \sim N\left(0, \frac{\sigma^2}{2}\right)$. 标准化得

$$U = \frac{Y_1 - Y_2}{\sigma/\sqrt{2}} \sim N(0, 1).$$