第 6 章　数理统计基础

安徽财经大学

统计与应用数学学院

目录

- 数理统计是一门方法论学科，它以概率论为基础，研究如何有效地收集数据，如何对所获得的数据进行科学的整理和分析，从而做出有效的估计、推断和预测.
- 在许多实际问题中，广泛存在着随机性数据，因此，数理统计在实际生产中有着广泛的应用，并为生产实际决策和行动提供有力依据和有效建议.
- 本章介绍数理统计学中的一些基本概念、基本理论，并着重介绍几个常用统计量及其概率分布.

在统计中，我们通常把研究对象的全体构成的集合称为总体(population)，而把总体中的每一个元素称为个体(individual).

例 (6.1.1)

我们研究某芯片厂一批产品的质量时，该批产品组成总体，每件芯片产品就是个体.

例 (6.1.2)

我们研究安财 2020 级学生的身高情况时：安财 2020 级全体学生组成总体，每个学生就是个体.

- 在实际问题中，人们所关心的并非总体内个体的本身，而是关心个体的某一项 (或某几项) 数量指标，如芯片的寿命、学生的身高、体重.
- 因此，应该将总体理解为 “研究对象的某一 (或某些) 数量指标值的全体构成的集合”.
- 由于每个个体的出现是有随机性的，所以，相应的数量指标的出现也具有随机性. 从而可以把该种数量指标看作是一个随机变量 (或随机向量). 这样，总体就可以用一个随机变量 (或随机向量) 及其分布来描述.
- 我们用 $\xi, \eta, \cdots$ 或大写字母 $X, Y, \cdots$ 表示总体. 总体 ξ 根据其所包含的单位数目是否可数，可以分为有限总体或无限总体，也可以是离散型随机变量或连续型随机变量.

在统计问题中，总体的分布通常是全部或部分未知的. 为了对总体的分布情况进行各种研究，就需要对总体进行抽样观察. 按一定的规则抽取若干个体进行观察或试验，这种抽取过程称为抽样(sampling)，所抽出的个体称为样本(sample)，样本中个体的个数称为 样本容量(sample size). 由于任一抽样都具有随机性，所以容量为 n 的样本可以由这 n 个个体组成，也可以由另外 n 个个体组成，因而，容量为 n 的样本可以看作是 n 维随机变量 $(\xi_1, \xi_2, \cdots, \xi_n)$. 当对样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 进行了一次观察或试验后，得到了一组具体的数值 $(x_1, x_2, \cdots, x_n)$，此时称该组数值为样本的一组观察值，简称为样本值(sample value). 因此，样本既可看成具体的数值，又可看成随机变量 (或随机向量)，在实施抽样前被看成随机变量，在实施抽样后，它是具体的数值. 样本的这种既可看成数值又可看成随机变量的性质，称为样本的二重性.

- 我们通过对总体进行抽样所得到的样本来对总体分布中某些未知因素进行统计推断，因此，为了使抽取的样本能很好地反映总体的信息，最常用的是采取“简单随机抽样”的方法，它要求满足以下两点：
  (1) 代表性：样本的每个分量 ξ_i 与所考察的总体 ξ 具有相同的分布 F.
  (2) 独立性：$\xi_1, \xi_2, \cdots, \xi_n$ 为相互独立的随机变量，也就是每个观察结果既不影响其他观察结果，也不受其他观察结果的影响.
- 由简单随机抽样所得到的样本称为简单随机样本，它可用与总体同分布的 n 个相互独立的随机变量 $\xi_1, \xi_2, \cdots, \xi_n$ 来表示. 由于简单随机样本在实际中常被采用，所以本书以后所提到的样本在没有特别说明的情况下均指简单随机样本，又简称为样本.

- 由于样本彼此独立且与总体的分布相同，因此样本的分布可由总体的分布函数 F 完全决定，即样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合分布函数为

$$F(x_1, x_2, \cdots, x_n) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \cdots, \xi_n \leqslant x_n) = \prod_{i=1}^{n} F(x_i).$$

- 由于样本彼此独立且与总体的分布相同，因此样本的分布可由总体的分布函数 F 完全决定，即样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合分布函数为

$$F(x_1, x_2, \cdots, x_n) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \cdots, \xi_n \leqslant x_n) = \prod_{i=1}^{n} F(x_i).$$

- 若总体 ξ 是离散型随机变量，其分布列为 $p_i = P\{\xi = x\}$ $(i = 1, 2, \cdots)$，则 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合分布列为

$$P(\xi_1 = x_1, \xi_2 = x_2, \cdots, \xi_n = x_n) = \prod_{i=1}^{n} P\{\xi_i = x_i\} = \prod_{i=1}^{n} p_i.$$

- 由于样本彼此独立且与总体的分布相同，因此样本的分布可由总体的分布函数 F 完全决定，即样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合分布函数为

$$F(x_1, x_2, \cdots, x_n) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \cdots, \xi_n \leqslant x_n) = \prod_{i=1}^{n} F(x_i).$$

- 若总体 ξ 是离散型随机变量，其分布列为 $p_i = P\{\xi = x\}$ $(i = 1, 2, \cdots)$，则 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合分布列为

$$P(\xi_1 = x_1, \xi_2 = x_2, \cdots, \xi_n = x_n) = \prod_{i=1}^{n} P\{\xi_i = x_i\} = \prod_{i=1}^{n} p_i.$$

- 若总体 ξ 是连续型随机变量，其概率密度为 $f(x)$，则 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的联合概率密度为

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

例 (6.1.3)

设总体 ξ 服从正态分布 $N(\mu, \sigma^2)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体的样本, 求样本 $\xi_1, \xi_2, \cdots, \xi_n$ 的联合概率密度.

例 (6.1.3)

设总体 ξ 服从正态分布 $N(\mu, \sigma^2)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体的样本, 求样本 $\xi_1, \xi_2, \cdots, \xi_n$ 的联合概率密度.

解

$\xi \sim N(\mu, \sigma^2)$, 密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ (-\infty < x < +\infty)$$

则样本 $\xi_1, \xi_2, \cdots, \xi_n$ 的联合概率密度为:

例 (6.1.3)

设总体 ξ 服从正态分布 $N(\mu,\sigma^2)$, $\xi_1,\xi_2,\cdots,\xi_n$ 是来自总体的样本, 求样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的联合概率密度.

解

$\xi\sim N(\mu,\sigma^2)$, 密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ (-\infty<x<+\infty)$$

则样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的联合概率密度为:

$$\begin{aligned}f^*(x_1,x_2,\cdots,x_n)&=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right\}\\&=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.\end{aligned}$$

例 (6.1.4)

设总体 ξ 服从两点分布 $b(1,p)$, $\xi_1,\xi_2,\cdots,\xi_n$ 是来自总体 ξ 的样本, 求样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的联合分布列.

例 (6.1.4)

设总体 ξ 服从两点分布 $b(1,p)$, $\xi_1,\xi_2,\cdots,\xi_n$ 是来自总体 ξ 的样本, 求样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的联合分布列.

解

ξ 的分布列为

$$P\{\xi=x\}=p^x(1-p)^{1-x},\ x\in\{0,1\}.$$

因此, $\xi_1,\xi_2,\cdots,\xi_n$ 的联合分布列为

$$\begin{aligned}P(\xi_1=x_1,\xi_2=x_2,\cdots,\xi_n=x_n)&=\prod_{i=1}^{n}P\{\xi_i=x_i\}=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}\\&=p^{\sum\limits_{i=1}^{n}x_i}(1-p)^{n-\sum\limits_{i=1}^{n}x_i},\end{aligned}$$

$x_i\in\{0,1\},(i=1,\cdots,n)$.

为了从理论上进一步说明随机样本能够较好地反映总体 ξ 的情况, 我们引入经验分布函数的概念.

定义 (6.1.1)

$\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的容量为 n 的样本, $x_1, x_2, \cdots, x_n$ 是该样本的一个样本值, 把这些样本值按由小到大次序排列为 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(r)}$, 且 $x_{(i)}$ 出现的次数为 k_i, $i = 1, 2, \cdots, r$, $\sum_{i=1}^{r} k_i = n$, 对任意实数 x, 定义

$$F_n^*(x) = \begin{cases} 0, & x < x_{(1)}; \\ \dfrac{k_1 + \cdots + k_i}{n}, & x_{(i)} \leqslant x < x_{(i+1)} \ (i = 1, 2, \cdots, r-1); \\ 1, & x \geqslant x_{(r)}. \end{cases}$$

为经验分布函数(empirical distribution function), 或称样本分布函数.

显然, 经验分布函数 $F_n^*(x)$ 就是在抽取样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的试验中, 样本值 $x_1, x_2, \cdots, x_n$ 中小于 x 的频率. 它具有如下性质:

(1) $0 \leqslant F_n^*(x) \leqslant 1$.

(2) $F_n^*(x)$ 是单调不减函数.

(3) $F_n^*(-\infty) = 0$, $F_n^*(+\infty) = 1$.

(4) $F_n^*(x)$ 右连续, 即 $F_n^*(x) = F_n^*(x+0)$.

显然, 经验分布函数 $F_n^*(x)$ 就是在抽取样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的试验中, 样本值 $x_1, x_2, \cdots, x_n$ 中小于 x 的频率. 它具有如下性质:

(1) $0 \leqslant F_n^*(x) \leqslant 1$.

(2) $F_n^*(x)$ 是单调不减函数.

(3) $F_n^*(-\infty) = 0$, $F_n^*(+\infty) = 1$.

(4) $F_n^*(x)$ 右连续, 即 $F_n^*(x) = F_n^*(x+0)$.

值得注意的是, 对于不同的样本值 $x_1, x_2, \cdots, x_n$, 我们将得到不同的经验分布函数 $F_n^*(x)$. 所以当 x 固定时, $F_n^*(x)$ 是样本的函数, 因而它也是一个随机变量.

对于任意的实数 x, 总体 ξ 的分布函数 $F(x) = P\{\xi \leqslant x\}$ 是事件 $\{\xi \leqslant x\}$ 的概率, 样本分布函数 $F_n^*(x)$ 是事件 $\{\xi \leqslant x\}$ 发生的频率. 由伯努利大数定律可知, 当 $n \to \infty$ 时, 对于任意的正数 ε, 有

$$\lim_{n\to\infty} P\{|F_n^*(x) - F(x)| < \varepsilon\} = 1.$$

格里汶科 (W. Glivenko) 进一步证明了如下定理:

定理 (格里汶科定理 6.1.1)

对于任一实数 x, 当 $n \to \infty$ 时, $F_n^*(x)$ 以概率 1 一致收敛于总体的分布函数 $F(x)$, 即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n^*(x) - F(x)| = 0\right\} = 1.$$

定理告诉我们, 当 n 很大时, 对于任一实数 x, 经验分布函数 $F_n^*(x)$ 与总体 ξ 的分布函数 $F(x)$ 之差的绝对值依概率 1 一致地趋于零. 这就是我们可以用样本推断总体的基本理论依据.

例 (6.1.5)

设从总体 ξ 中抽取容量为 8 的样本, 得到的样本值为

$$-2.8, -1, 1.5, 2.1, 1.5, 0, 1.5, 3.4$$

试求样本的经验分布函数 $F_8^*(x)$.

例 (6.1.5)

设从总体 ξ 中抽取容量为 8 的样本, 得到的样本值为

$$-2.8, -1, 1.5, 2.1, 1.5, 0, 1.5, 3.4$$

试求样本的经验分布函数 $F_8^*(x)$.

解 (把样本值按从小到大的顺序排列为)

$$-2.8 < -1 < 0 < 1.5 = 1.5 = 1.5 < 2.1 < 3.4$$

由经验分布函数的定义可知

例 (6.1.5)

设从总体 ξ 中抽取容量为 8 的样本, 得到的样本值为

$$-2.8, -1, 1.5, 2.1, 1.5, 0, 1.5, 3.4$$

试求样本的经验分布函数 $F_8^*(x)$.

解 (把样本值按从小到大的顺序排列为)

$$-2.8 < -1 < 0 < 1.5 = 1.5 = 1.5 < 2.1 < 3.4$$

由经验分布函数的定义可知

$$F_8^*(x) = \begin{cases} 0, & x < -2.8; \\ 0.125, & -2.8 \leqslant x < -1; \\ 0.25, & -1 \leqslant x < 0; \\ 0.375, & 0 \leqslant x < 1.5; \\ 0.75, & 1.5 \leqslant x < 2.1; \\ 0.875, & 2.1 \leqslant x < 3.4; \\ 1, & x \geqslant 3.4. \end{cases}$$

1. 统计量

定义

设 $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的一个样本，$x_1, x_2, \cdots, x_n$ 是样本值，$g(\xi_1, \xi_2, \cdots, \xi_n)$ 是 $\xi_1, \xi_2, \cdots, \xi_n$ 的函数. 如果 $g(\xi_1, \xi_2, \cdots, \xi_n)$ 中不含任何未知参数，则称 $g(\xi_1, \xi_2, \cdots, \xi_n)$ 为统计量(statistic)，而 $g(x_1, x_2, \cdots, x_n)$ 称为统计量的观测值.

例 (6.1.6)

设总体 $\xi \sim b(1,p)$, $P\{x=1\}=p$, $P\{x=0\}=1-p$, 其中 $p>0$ 为未知参数, $\xi_1,\xi_2,\cdots,\xi_n$ 为来自总体 ξ 的一个样本, 指出下列函数哪些是统计量, 哪些不是统计量.

(1) $\xi_1+\xi_2$; (2) $\max\limits_{1\leqslant i\leqslant n}\{\xi_i\}$; (3) ξ_n+2p; (4) $(\xi_n-\xi_1)^2$.

例 (6.1.6)

设总体 $\xi \sim b(1,p)$, $P\{x=1\}=p$, $P\{x=0\}=1-p$, 其中 $p>0$ 为未知参数, $\xi_1,\xi_2,\cdots,\xi_n$ 为来自总体 ξ 的一个样本, 指出下列函数哪些是统计量, 哪些不是统计量.

(1) $\xi_1+\xi_2$; (2) $\max\limits_{1\leqslant i\leqslant n}\{\xi_i\}$; (3) ξ_n+2p; (4) $(\xi_n-\xi_1)^2$.

解

根据统计量定义, 统计量必须满足两个条件: (a) 它是样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的函数, (b) 它不含任何未知参数.

在 $(1)-(4)$ 中, 它们都是样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的函数, 但 (3) 含未知参数 p, 所以 (1), (2) 及 (4) 中的样本函数是统计量, (3) 中的样本函数不是统计量.

例 (6.1.7)

ξ_1, ξ_2, ξ_3 为来自正态总体 $\xi \sim N(\mu, \sigma^2)$ 的一个样本, 其中 μ, σ^2 是未知参数, 则 $\frac{1}{3}(\xi_1+\xi_2+\xi_3)+\mu$, $\frac{1}{2}(\xi_1+\xi_2)-\mu$, $\frac{1}{\sigma}(\xi_1+\xi_2+\xi_3)$ 都不是统计量, 因为它们都含有未知参数, 而 $\frac{1}{2}(\xi_1+\xi_2+\xi_3)$, $\frac{1}{3}(\xi_1^2+\xi_2^2+\xi_3^2)$, $\xi_1+2\xi_2-3\xi_3$ 都是统计量.

2. 常用统计量

设 $\xi_1, \xi_2, \cdots, \xi_n$ 为来自总体的一个样本, $x_1, x_2, \cdots, x_n$ 是样本值, 数理统计中最常用的统计量及其观察值有

(1) 样本均值　称

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i,$$

为样本均值. 它反映了总体均值的信息.

2. 常用统计量

(2) 样本方差　称

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2 \text{ 或 } S_n^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2,$$

为样本方差. 它反映了总体方差的信息.

2. 常用统计量

(2) 样本方差　称

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2 \text{ 或 } S_n^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2,$$

为样本方差. 它反映了总体方差的信息.

(3) 样本标准差　称

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2} \text{ 或 } S_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2},$$

为样本标准差.

2. 常用统计量

(4) 样本 k 阶原点矩　称

$$A_k = \frac{1}{n}\sum_{i=1}^{n}\xi_i^k,\ k = 1, 2, \cdots,$$

为样本 (k 阶) 原点矩, 它的观测值为 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,\ k = 1, 2, \cdots.$

显然, 样本一阶原点矩就是样本均值. 即 $A_1 = \overline{\xi}$.

2. 常用统计量

(4) 样本 k 阶原点矩　称

$$A_k = \frac{1}{n}\sum_{i=1}^{n}\xi_i^k,\ k = 1, 2, \cdots,$$

为样本 (k 阶) 原点矩, 它的观测值为 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,\ k = 1, 2, \cdots.$

显然, 样本一阶原点矩就是样本均值. 即 $A_1 = \overline{\xi}$.

(5) 样本 k 阶中心矩　称

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^k,\ k = 1, 2, \cdots,$$

为样本 k 阶中心矩, 它的观测值为 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,\ k = 1, 2, \cdots.$

- 值得注意的是: 总体均值 $E\xi$ 是常数, 而样本均值 $\overline{\xi}$ 是随机变量, 这是两个不同的概念, 不能混淆. 当然, 这两者之间有一定的关系. 同样, 总体方差 $D\xi$ 与样本方差 S^2 (S_n^2)、总体矩与样本矩也是不同的概念.
- 若总体均值 $E\xi$ 、方差 $D\xi$ 都存在, 则由样本 $\xi_1, \xi_2, \cdots, \xi_n$ 的独立性及与总体 ξ 的同分布性, 有

$$\begin{aligned} E\xi_1 = E\xi_2 = \cdots = E\xi_n = E\xi; \\ D\xi_1 = D\xi_2 = \cdots = D\xi_n = D\xi. \end{aligned}$$

- 由于 $\xi_1^k, \xi_2^k, \cdots, \xi_n^k$ 也具有相互独立及与 ξ^k 同分布性, 于是

$$E\left(\xi_1^k\right) = E\left(\xi_2^k\right) = \cdots = E\left(\xi_n^k\right) = E\left(\xi^k\right).$$

例 (6.1.8)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的样本, 且总体均值 $E(\xi) = \mu$, 总体方差 $D(\xi) = \sigma^2$, 求 $E(\overline{\xi}), D(\overline{\xi}), E\left(S^2\right)$.

例 (6.1.8)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的样本, 且总体均值 $E(\xi) = \mu$, 总体方差 $D(\xi) = \sigma^2$, 求 $E(\overline{\xi}), D(\overline{\xi}), E\left(S^2\right)$.

解

由样本的独立性同分布性以及数学期望和方差的性质,

$$E(\xi_i) = \mu, D(\xi_i) = \sigma^2, \ i = 1, 2, \cdots, n.$$

$$E(\overline{\xi}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n}\sum_{i=1}^{n}E(\xi_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

$$D(\overline{\xi}) = D\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}D(\xi_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}.$$

因 $$\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2=\sum_{i=1}^{n}\left(\xi_i^2-2\xi_i\overline{\xi}+\overline{\xi}^2\right)=\sum_{i=1}^{n}\xi_i^2-2\overline{\xi}\left(\sum_{i=1}^{n}\xi_i\right)+\sum_{i=1}^{n}\overline{\xi}^2$$

$$=\sum_{i=1}^{n}\xi_i^2-2\overline{\xi}(n\overline{\xi})+n\overline{\xi}^2=\sum_{i=1}^{n}\xi_i^2-n\overline{\xi}^2.$$

因 $\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2=\sum_{i=1}^{n}\left(\xi_i^2-2\xi_i\overline{\xi}+\overline{\xi}^2\right)=\sum_{i=1}^{n}\xi_i^2-2\overline{\xi}\left(\sum_{i=1}^{n}\xi_i\right)+\sum_{i=1}^{n}\overline{\xi}^2$

$$=\sum_{i=1}^{n}\xi_i^2-2\overline{\xi}(n\overline{\xi})+n\overline{\xi}^2=\sum_{i=1}^{n}\xi_i^2-n\overline{\xi}^2.$$

而 $E\left(\xi_i^2\right)=D\left(\xi_i\right)+\left(E\xi_i\right)^2=\sigma^2+\mu^2, E\left(\overline{\xi}^2\right)=D(\overline{\xi})+(E\overline{\xi})^2=\frac{\sigma^2}{n}+\mu^2,$

所以 $E(S^2)=E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2\right]=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\xi_i^2-n\overline{\xi}^2\right]$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}E\left(\xi_i^2\right)-nE\left(\overline{\xi}^2\right)\right]$$

$$=\frac{1}{n-1}\left[n\left(\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]=\sigma^2.$$

3. 次序统计量

定义 (6.1.3)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 是取自总体 ξ 的一个样本, 称 $\xi_{(k)}$ 为第 k 个次序统计量. 它是样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 的满足如下条件的函数: 每当样本得到一组观测值 $x_1, x_2, \cdots, x_n$ 时, 将它们按由小到大重新排序

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(k)} \leqslant \cdots \leqslant x_{(n)},$$

第 k 个值 $x_{(k)}$ 就作为统计量 $\xi_{(k)}$ 的观察值. 而 $\xi_{(1)}, \xi_{(2)}, \cdots, \xi_{(k)}, \cdots, \xi_{(n)}$ 称为次序统计量或顺序统计量(order statistic). 其中 $\xi_{(1)} = \min(\xi_1, \xi_2, \cdots, \xi_n)$ 称为最小次序统计量 (minimum order statistic), $\xi_{(n)} = \max(\xi_1, \xi_2, \cdots, \xi_n)$ 称为最大次序统计量 (maximum order statistic).

设总体 ξ 的分布函数为 $F(x)$，记 $\xi_{(1)}$ 和 $\xi_{(n)}$ 的分布函数分别为 $F_{\xi_{(1)}}(x)$ 和 $F_{\xi_{(n)}}(x)$，则

$$
\begin{aligned}
F_{\xi_{(1)}}(x) &= P\{\min(\xi_1, \xi_2, \cdots, \xi_n\} \leqslant x\} \\
&= 1 - P\{\min(\xi_1, \xi_2, \cdots, \xi_n) > x\} \\
&= 1 - \prod_{i=1}^{n} P\{\xi_i > x\} = 1 - \prod_{i=1}^{n} [1 - P\{\xi_i \leqslant x\}] \\
&= 1 - [1 - F(x)]^n \\
F_{\xi(n)}(x) &= P\{\max(\xi_1, \xi_2, \cdots, \xi_n) \leqslant x\} \\
&= P\{\xi_1 \leqslant x, \xi_2 \leqslant x, \cdots, \xi_n \leqslant x\} \\
&= \prod_{i=1}^{n} P\{\xi_i \leqslant x\} = [F(x)]^n,
\end{aligned}
$$

当总体 ξ 为连续型随机变量且密度函数为 $f(x)$ 时, $\xi_{(1)}$ 和 $\xi_{(n)}$ 的密度函数分别为

$$f_{\xi_{(1)}}(x) = \frac{\mathrm{d}F_{\xi_{(1)}}(x)}{\mathrm{d}x} = n[1 - F(x)]^{n-1} f(x),$$

$$f_{\xi_{(n)}}(x) = \frac{\mathrm{d}F_{\xi_{(n)}}(x)}{\mathrm{d}x} = n[F(x)]^{n-1} f(x).$$

一般地, 有如下定理:

定理 (6.1.2)

设总体 ξ 的密度函数为 $f(x)$, 分布函数为 $F(x)$, $\xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的样本, 则第 k 个次序统计量 $\xi_{(k)}$ 的密度函数为

$$f_{\xi_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1 - F(x)]^{n-k} f(x).$$

例 (6.1.9)

$\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$ 是取自正态总体 $\xi \sim N(13, 4)$ 的容量为 5 的样本，求概率 $P\{\xi_{(5)} > 17\}$ 和 $P\{\xi_{(1)} < 11\}$.

例 (6.1.9)

$\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$ 是取自正态总体 $\xi \sim N(13, 4)$ 的容量为 5 的样本，求概率 $P\{\xi_{(5)} > 17\}$ 和 $P\{\xi_{(1)} < 11\}$.

解

设总体 ξ 的分布函数为 $F(x)$，则随机变量 $\xi_{(5)}$ 和 $\xi_{(1)}$ 的分布函数分别为

$$F_{\max}(x) = [F(x)]^5, F_{\min}(x) = 1 - [1 - F(x)]^5.$$

于是，

$$\begin{aligned} P\{\xi_{(5)} > 17\} &= 1 - P\{\xi_{(5)} \leqslant 17\} \\ &= 1 - [F(17)]^5 = 1 - [\Phi(2)]^5 = 0.1089, \\ P\{\xi_{(1)} < 11\} &= F_{\min}(11) = 1 - [1 - F(11)]^5 \\ &= 1 - [\Phi(1)]^5 \\ &= 1 - (0.8413)^5 = 0.5785. \end{aligned}$$

例 (6.1.10)

设总体 $\xi \sim U(0,1), \xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的样本, 求第 k 个次序统计量 $\xi_{(k)}$ 的密度函数.

例 (6.1.10)

设总体 $\xi \sim U(0,1), \xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的样本, 求第 k 个次序统计量 $\xi_{(k)}$ 的密度函数.

解

由于 $\xi \sim U(0,1)$, 则其分布函数 $F(x)$ 及密度函数 $f(x)$ 分别为

$$F(x)=\begin{cases}0, & x \leqslant 0;\\ x, & 0<x<1;\\ 1, & x \geqslant 1.\end{cases} \qquad f(x)=\begin{cases}1, & 0<x<1;\\ 0, & \text{其他}.\end{cases}$$

根据 (6.1.1) 式, 有

$$f_{\xi_{(k)}}(x)=\begin{cases}\dfrac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}, & 0<x<1;\\ 0, & \text{其他}.\end{cases}$$

为了能够从样本 $(\xi_1, \xi_2, \cdots, \xi_n)$ 大致确定总体 ξ 的概率分布，需要大量样本观测值. 若总体 ξ 为离散型随机变量，则可以通过计算样本观测值 $(x_1, x_2, \cdots, x_n)$ 中各个分量的重复次数，从而得到总体 ξ 取这些值的预率，由伯努利大数定律，当 n 较大时，ξ 取这些值的频率就可以近似地作为总体 ξ 取这些值的概率，这样就大致地得到总体 ξ 的概率分布.
对于总体 ξ 为连续型随机变量的情形，由于不能计算每个观测值的次数和频率，通常需要作出样本的预率分布直方图 (简称直方图)，作直方图的步骤如下:

设 $x_1, x_2, \cdots, x_n$ 是样本的 n 个观测值.

(1) 求 $x_1, x_2, \cdots, x_n$ 的最小值 $x_{(1)}$、最大值 $x_{(n)}$ 、极差 $R = x_{(n)} - x_{(1)}$.

(2) 选取常数 t_0 (略小于 $x_{(1)}$) 和 t_k (略大于 $x_{(n)}$), 并将区间 $[t_0, t_k]$ 等分成 k 个互不相交的子区间

$$[t_0, t_1), [t_1, t_2), \cdots, [t_{i-1}, t_i), \cdots, [t_{k-1}, t_k),$$

每个子区间的长度 (组距) $\Delta t = \dfrac{t_k - t_0}{k}$, 子区间的个数一般取 8 至 15 个, 太多则由于频率的随机摆动而使分布显得杂乱, 太少则难以显示分布的特征. 此外, 为了方便起见, 分点 t_i 应比样本 x_i 多取一位小数.

(3) 计算样本观察值 $x_1, x_2, \cdots, x_n$ 落入第 i 个小区间 $I_i = [t_{i-1}, t_i)$ 的组频数 n_i, 组频率 $f_i = \dfrac{n_i}{n}$ 以及 $h_i = \dfrac{f_i}{\Delta t}$ $(i = 1, 2, \cdots, k)$.

(4) 在 $[t_{i-1}, t_i)$ 上以 Δt 为底, 以 h_i 为高作小矩形, 各小矩形的面积恰为 f_i, 所有小矩形合在一起就构成了频率直方图.

例 (6.2.1)

以下是某班 60 位学生“概率论与数理统计”期末考试成绩

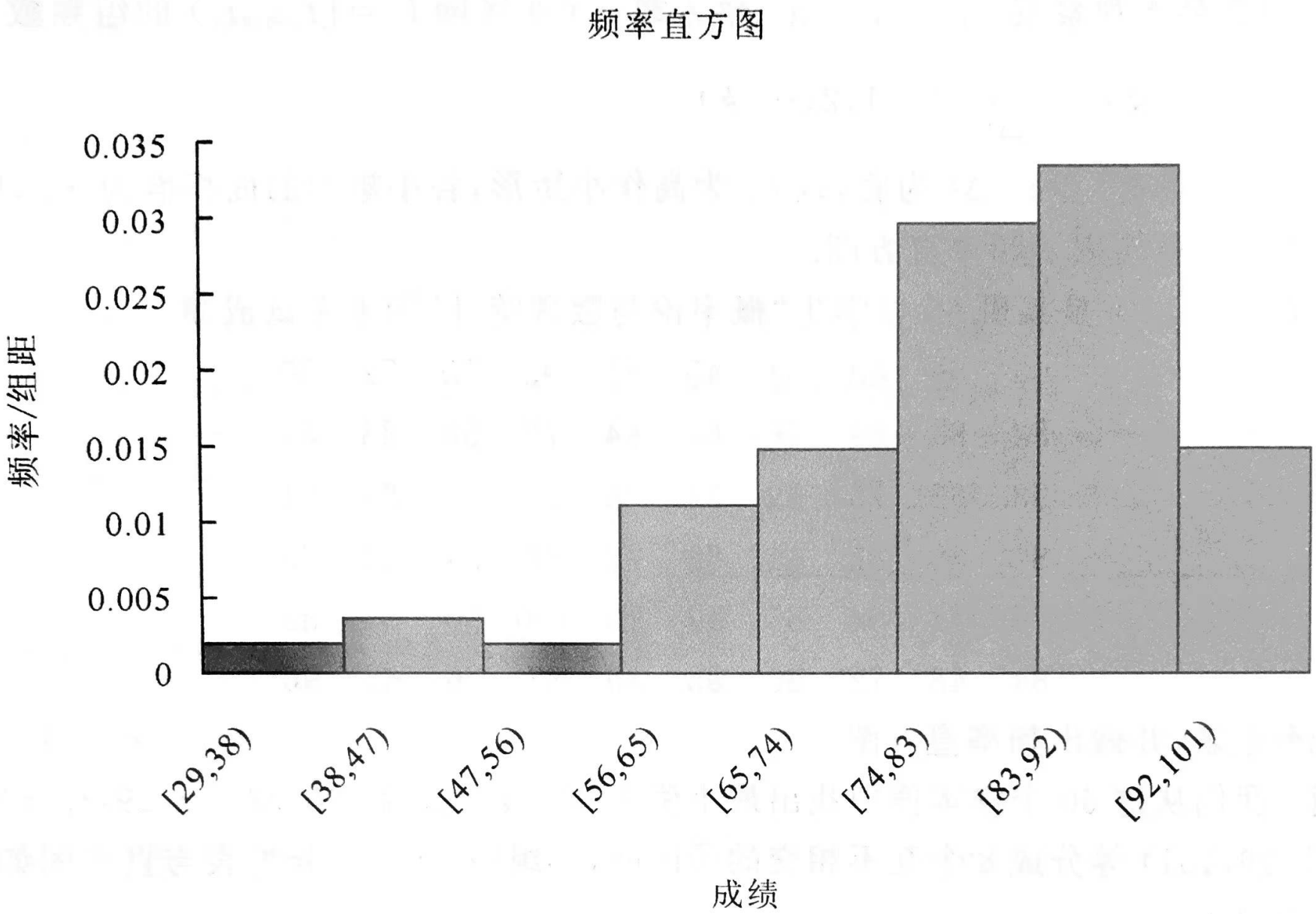
63	76	83	91	45	81	93	30	72	80
82	83	81	76	67	84	72	58	83	64
93	63	75	99	74	76	95	91	83	61
82	85	83	44	88	72	66	94	68	78
88	71	94	85	82	79	100	90	83	88
84	48	72	80	85	80	87	76	62	96

试列出分组表, 并做出频率直方图.

解

我们从这 60 个样本值中找出最小值为 30, 最大值为 100, 取 $t_0 = 29, t_k = 101$, 并将区间 $[29, 101)$ 等分成 8 个互不相交的子区间, 其组距 $\Delta t = 9$. 分组表与直方图如下.

区间	组频数 n_i	组频率 f_i	$h_i = f_i/\Delta t$
$[29, 38)$	1	0.02	0.0019
$[38, 47)$	2	0.03	0.0037
$[47, 56)$	1	0.02	0.0019
$[56, 65)$	6	0.10	0.0111
$[65, 74)$	8	0.13	0.0148
$[74, 83)$	16	0.27	0.0296
$[83, 92)$	18	0.30	0.0333
$[92, 101)$	8	0.13	0.0148
合计	60	1	


频率直方图
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
频率/组距
[29,38)
[38,47)
[47,56)
[56,65)
[65,74)
[74,83)
[83,92)
[92,101)
成绩

1. 算术平均值

算术平均值也称为均值 (mean), 是将一组数据的总和除以这组数据的个数所得的结果.

定义 (6.2.1)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的样本, $x_1, x_2, \cdots, x_n$ 是一组样本值, 其算术平均值称为样本均值, 一般用 $\overline{x}$ 表示, 即

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

例 (6.2.2)

随机抽取某班 11 名同学的 2014 年十月份的消费情况, 得到的数据如下 (单位: 元) :

$$890, 1300, 1196, 980, 998, 980, 1600, 1350, 1493, 1400, 2080,$$

计算这 11 人在十月份消费的平均值.

例 (6.2.2)

随机抽取某班 11 名同学的 2014 年十月份的消费情况, 得到的数据如下 (单位: 元) :

$$890, 1300, 1196, 980, 998, 980, 1600, 1350, 1493, 1400, 2080,$$

计算这 11 人在十月份消费的平均值.

解

$$\overline{x} = \frac{890 + 1300 + \cdots + 2080}{11} = 1297.$$

2. 分位数与中位数

定义 (6.2.2)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的一个样本, $\xi_{(1)}, \xi_{(2)}, \cdots, \xi_{(k)}, \cdots, \xi_{(n)}$ 是其次序统计量, 其观测值为 $x_{(1)}, x_{(2)}, \cdots, x_{(k)}, \cdots, x_{(n)}$. 对 $0 < p < 1$, 定义

$$m_p = \begin{cases} x_{([np+1])}, & np \text{ 不是整数;} \\ \dfrac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & np \text{ 是整数.} \end{cases}$$

为该样本的 p 分位数 (或 p 分位点) (quantile of order p).

例如, 若 $n = 10$, $p = 0.75$, 则 $m_{0.75} = x_{(8)}$, 若 $n = 20, p = 0.25$, 则 $m_{0.25} = \dfrac{1}{2}\left(x_{(5)} + x_{(6)}\right)$.

特别, 当 $p = 0.5$ 时, 样本的分位数 $m_{0.5}$ 称为样本中位数 (median), 并记为 m_e.

容易看出，样本中位数有一个简单的表示式

$$m_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ 为奇数;} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & n \text{ 为偶数.} \end{cases}$$

例如，若 $n = 9$，则 $m_e = x_{(5)}$，若 $n = 10$，则 $m_e = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right)$.

样本中位数是反映样本位置特征的一个量，它可以用于推断总体分布的中位数和总体的对称中心，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，此时样本中位数 m_e 也反映总体均值的信息. 与样本均值相比，样本中位数不受样本中的异常值的影响，有时比样本均值更有代表性.

例 (6.2.3)

根据例 6.2.2 中的数据, 计算该组数据的中位数.

解

把例 6.2.2 中数据按由小到大重新排序为

$$890, 980, 980, 998, 1196, 1300, 1350, 1400, 1493, 1600, 2080,$$

由于 $n = 11$ 为奇数, 故其中位数为

$$m_e = x_{\left(\frac{11+1}{2}\right)} = x_{(6)} = 1300.$$

3. 众数

定义 (6.2.3)

样本中出现次数最多的数据称为众数 (mode), 记作 m_o.

众数代表的是最常见、最普遍的状况, 是对现象集中趋势的度量, 它受数据中最大或最小值变化的影响较小, 从分布的角度看, 众数出现的频率最高. 如果样本数据中每个数出现的次数都相同, 它就没有众数. 如果样本数据中有两个或以上的数出现次数相同, 且出现次数超过其他数的出现次数, 这几个数都是众数.
例如我们有如下数据: $3, 4, 4, 5, 6, 6, 6, 8, 8, 8, 8, 10, 6$, 则众数为 6 和 8.

众数、中位数和平均值的关系

从分布的角度看, 众数始终是一组数据分布的最高峰值, 中位数是处于一组数据中间位置上的值, 而平均值则是全部数据的算术平均.

定义 (6.2.4)

设 $\xi_1, \xi_2, \cdots, \xi_n$ 为来自总体 ξ 的一个样本, $\xi_{(1)}, \xi_{(2)}, \cdots, \xi_{(k)}, \cdots, \xi_{(n)}$ 是其次序统计量, 称 $R = \xi_{(n)} - \xi_{(1)}$ 为样本极差, 它是反映总体分布散布程度的信息, 可以用于推断总体的标准差.

例如, 根据例 6.2.2 中的数据, 计算这 11 人十月份消费的极差为

$$R = 2080 - 890 = 1190 \text{ 元}.$$

描述样本数据分散程度的统计量还有如 6.1.3 节中定义的样本方差及样本标准差.

1. 样本偏度

定义 (6.2.5)

设 b_2, b_3 分别是样本的二阶、三阶中心矩, 称统计量

$$\gamma_1 = \frac{b_3}{b_2^{3/2}},$$

为样本偏度 (skewness).

样本偏度 γ_1 反映了总体分布密度曲线的对称信息. 如果数据完全对称, 显然 $b_3 = 0$. 数据不对称则 $b_3 \neq 0$. 这里, 用 b_3 除以 $b_2^{3/2}$ 是为了消除量纲的影响, γ_1 是个相对数, 它很好地刻画了数据分布的偏斜方向和程度. 若 $\gamma_1 = 0$ 表示样本对称; 若 $\gamma_1 > 0$ 表示样本的右尾长, 即样本中有几个较大的数, 这反映总体是正偏的 (或右偏的); 若 $\gamma_1 < 0$ 表示样本的左尾长, 即样本中有几个较小的数, 这反映总体是负偏的 (或左偏的).

2. 样本峰度

定义 (6.2.6)

设 b_2, b_4 分别是样本的二阶、四阶中心矩, 称统计量

$$\gamma_2 = \frac{b_4}{b_2^2} - 3,$$

为样本峰度 (kurtosis).

样本峰度 γ_2 反映了总体分布密度曲线在其峰值附近的陡峭程度. 当分布曲线为正态曲线时, $\gamma_2 = 0$; 当 $\gamma_2 > 0$ 时, 分布密度曲线在其峰值附近比正态分布来得陡, 称为尖峰型, γ_2 越大, 分布密度曲线的顶端越尖峭; 当 $\gamma_2 < 0$ 时, 分布密度曲线在其峰值附近比正态分布来平坦, 称为平坦型, γ_2 越小, 分布密度曲线的顶端越平坦.

2. 样本峰度

对于具有单峰分布的大多数数据而言，众数、中位数和平均值之间有以下关系:
如果数据的分布是对称的，众数 m_o 、中位数 m_e 和平均值 $\overline{x}$ 一定相等，三者合而为一，即 $m_o = m_e = \overline{x}$; 如果数据是左偏 (负偏) 的，数据中的极小值会使平均值偏向较小的一方，极小值的大小虽不影响中位数，但其所占项数会影响数据的中间位置从而使中位数偏小，众数则完全不受极小值大小和位置影响，因此，一般情况下，三者的关系表现为 $\overline{x} < m_e < m_o$; 反之，如果数据是右偏 (正偏) 的，一般有，$m_o < m_e < \overline{x}$.

统计量是样本的函数，它是一个随机变量. 统计量的分布称为抽样分布(sampling distribution). 在使用统计量进行统计推断时常需要知道它的分布. 当总体的分布已知时，抽样分布是确定的，但是要求出统计量的精确分布，一般来说是不容易的. 下面介绍来自正态总体的几个常用的统计量的分布.

1. χ^2- 分布的定义

定义 (6.3.1)

称 ξ 服从自由度为 n 的 χ^2 分布, 如果它的密度函数为

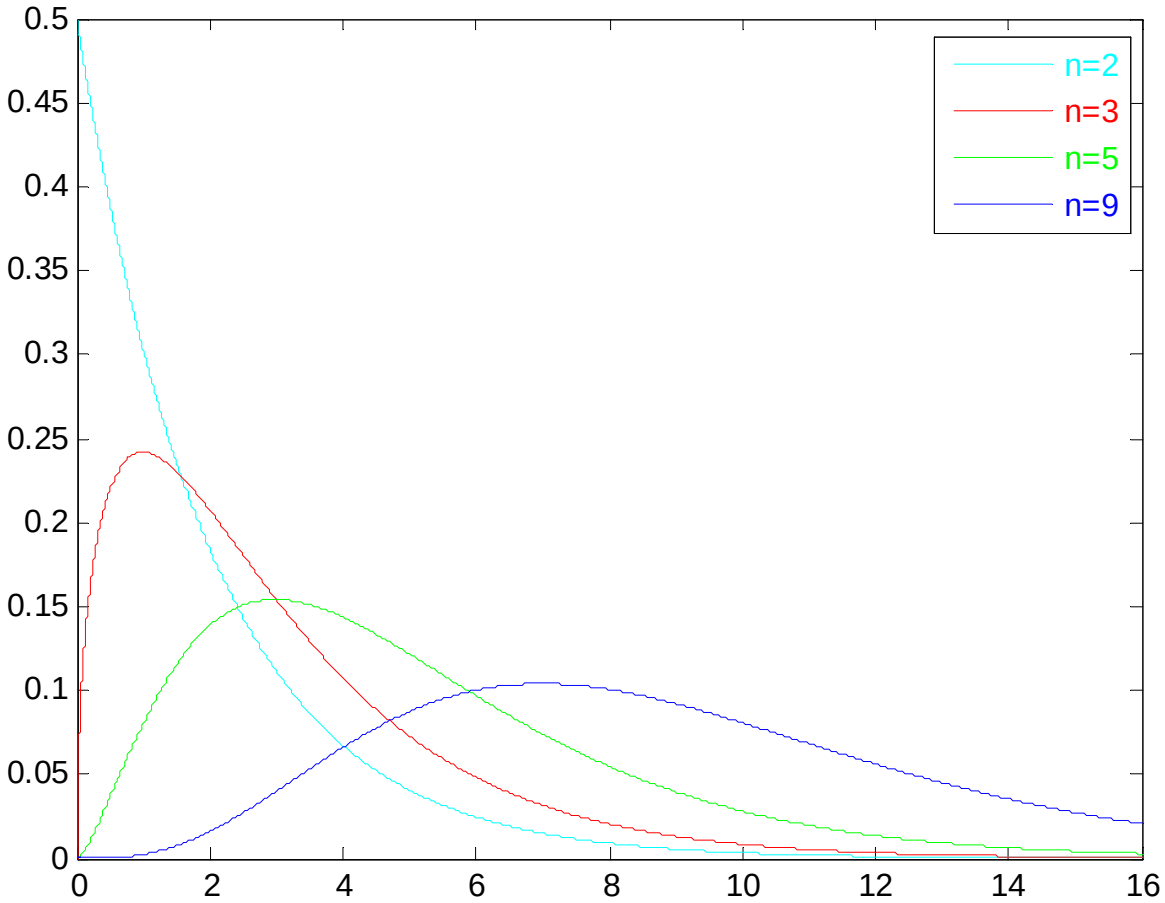
$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, & x\geqslant 0,\\ 0, & x<0,\end{cases}$$

其中 $\Gamma(x)$ 是 Gamma 函数, 记为$\xi\sim\chi^2(n)$.

χ^2 分布在数理统计中具有重要意义. χ^2 分布是由阿贝 (Abbe) 于 1863 年首先提出, 后来由海尔墨特 (Hermert) 和现代统计学的奠基人之一的卡 · 皮尔逊 (K. Pearson) 分别于 1875 年和 1900 年推导出来, 是统计学中的一个非常有用的分布.

χ^2- 分布的密度函数图像

χ^2 分布密度函数图像是一个取非负值的偏态分布，如下图所示.



2. χ^2- 分布的构造

定理 (6.3.1)

若 $\xi_1, \xi_2 \cdots, \xi_n$ 相互独立，且都服从标准正态分布 $N(0,1)$，则

$$\chi^2 = \sum_{i=1}^{n} \xi_i^2 \sim \chi^2(n).$$

在上述结论中，可以认为 $(\xi_1, \xi_2 \cdots, \xi_n)$ 是来自标准正态总体 $N(0,1)$ 的一个样本，从而 (6.3.1) 式定义的 χ^2 是一个统计量. 自由度 n 表示和式 $\sum_{i=1}^{n} \xi_i^2$ 中独立变量的个数.

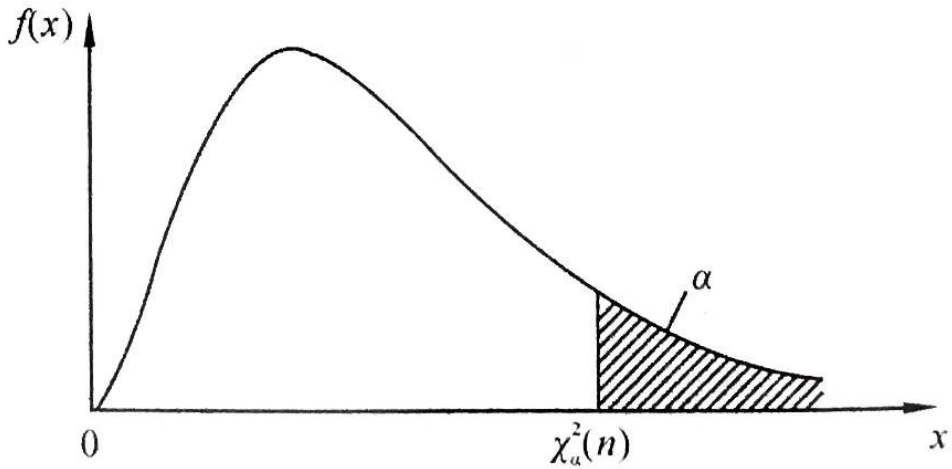
3. χ^2- 分布的分位数

定义 (6.3.2)

设 $\xi \sim \chi^2(n)$, 对给定的实数 α $(0 < \alpha < 1)$, 称满足条件

$$P\{\xi > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{+\infty} f(x)\mathrm{d}x = \alpha,$$

的点 $\chi^2_\alpha(n)$ 为 $\chi^2(n)$ 分布的上 α 分位数.



3. χ^2- 分布的分位数

对不同的 α 与 n, 分位数的值已经编制成表供查用 (见本书附录 1.3). 例如, 查表得:

$$\chi^2_{0.95}(45)=30.612, \chi^2_{0.1}(20)=28.412, \chi^2_{0.05}(15)=24.996.$$

表中只给出了自由度 $n=45$ 为止, 当 $n>45$ 时, 近似地有

$$\chi^2_\alpha(n)\approx\frac{1}{2}\left(u_\alpha+\sqrt{2n-1}\right)^2.$$

其中 u_α 是标准正态分布的上 α 分位数. 利用上式可对 $n>45$ 时的 $\chi^2(n)$ 分布的上 α 分位数进行近似计算.

4. χ^2- 分布的性质

(1) 可加性.

若 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$, 且 ξ, η 相互独立, 则 $\xi+\eta \sim \chi^2(n_1+n_2)$.

4. χ^2- 分布的性质

(1) 可加性.

若 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$, 且 ξ, η 相互独立, 则 $\xi+\eta \sim \chi^2(n_1+n_2)$.

证明.

由 χ^2 分布的构造, 可设

$$\xi=\sum_{i=1}^{n_1} \xi_i^2, \ \eta=\sum_{i=n_1+1}^{n_1+n_2} \xi_i^2,$$

其中 $\xi_1, \xi_2, \cdots, \xi_{n_1}, \xi_{n_1+1}, \cdots, \xi_{n_1+n_2}$ 相互独立, 且都服从标准正态分布 $N(0,1)$, 于是, 由 χ^2 分布的构造, $\xi+\eta=\sum_{i=1}^{n_1+n_2} \xi_i^2$ 服从 $\chi^2(n_1+n_2)$, 即 $\xi+\eta \sim \chi^2(n_1+n_2)$. □

(2) 期望与方差.

若 $\xi \sim \chi^2(n)$, 则 $E(\xi) = n, D(\xi) = 2n$.

(2) 期望与方差.

若 $\xi \sim \chi^2(n)$, 则 $E(\xi)=n, D(\xi)=2n$.

证明.

因为 $\xi_i \sim N(0,1)$, 所以 $E\left(\xi_i^2\right)=D\left(\xi_i\right)=1$,

$$E\left(\xi_i^4\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^4 \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x=3,$$

$$D\left(\xi_i^2\right)=E\left(\xi_i^4\right)-\left[E\left(\xi_i^2\right)\right]^2=3-1=2,\ i=1,2,\cdots,n,$$

于是

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

$$D(\xi)=D\left(\sum_{i=1}^{n}\xi_i^2\right)=\sum_{i=1}^{n}D\left(\xi_i^2\right)=2n.$$

例 (6.3.1)

设 $\xi_1, \xi_2, \xi_3, \xi_4$ 是来自总体 $\xi \sim N(0,4)$ 的一个样本, 问当 a, b 为何值时, $\eta = a\left(\xi_1 - 2\xi_2\right)^2 + b\left(3\xi_3 - 4\xi_4\right)^2 \sim \chi^2(n)$, 并确定 n 的值.

例 (6.3.1)

设 $\xi_1, \xi_2, \xi_3, \xi_4$ 是来自总体 $\xi \sim N(0,4)$ 的一个样本, 问当 a, b 为何值时, $\eta = a(\xi_1 - 2\xi_2)^2 + b(3\xi_3 - 4\xi_4)^2 \sim \chi^2(n)$, 并确定 n 的值.

解

由于 $\xi_1, \xi_2, \xi_3, \xi_4$ 独立同分布于 $N(0,4)$, 由正态分布的性质

$$
\begin{gathered}
E(\xi_1 - 2\xi_2) = 0, E(3\xi_3 - 4\xi_4) = 0, \\
D(\xi_1 - 2\xi_2) = D(\xi_1) + (-2)^2 D(\xi_2) = 20, \\
D(3\xi_3 - 4\xi_4) = 3^2 D(\xi_3) + (-4)^2 D(\xi_4) = 100.
\end{gathered}
$$

于是

$$
\frac{\xi_1 - 2\xi_2}{\sqrt{20}} \sim N(0,1), \quad \frac{3\xi_3 - 4\xi_4}{10} \sim N(0,1).
$$

例 (6.3.1)

设 $\xi_1, \xi_2, \xi_3, \xi_4$ 是来自总体 $\xi \sim N(0,4)$ 的一个样本, 问当 a, b 为何值时, $\eta = a\left(\xi_1 - 2\xi_2\right)^2 + b\left(3\xi_3 - 4\xi_4\right)^2 \sim \chi^2(n)$, 并确定 n 的值.

解

而且 $\xi_1 - 2\xi_2$ 与 $3\xi_3 - 4\xi_4$ 相互独立, 所以

$$\frac{\left(\xi_1 - 2\xi_2\right)^2}{20} + \frac{\left(3\xi_3 - 4\xi_4\right)^2}{100} \sim \chi^2(2).$$

从而

$$a = \frac{1}{20}, b = \frac{1}{100}, n = 2.$$

1. $t-$ 分布的定义

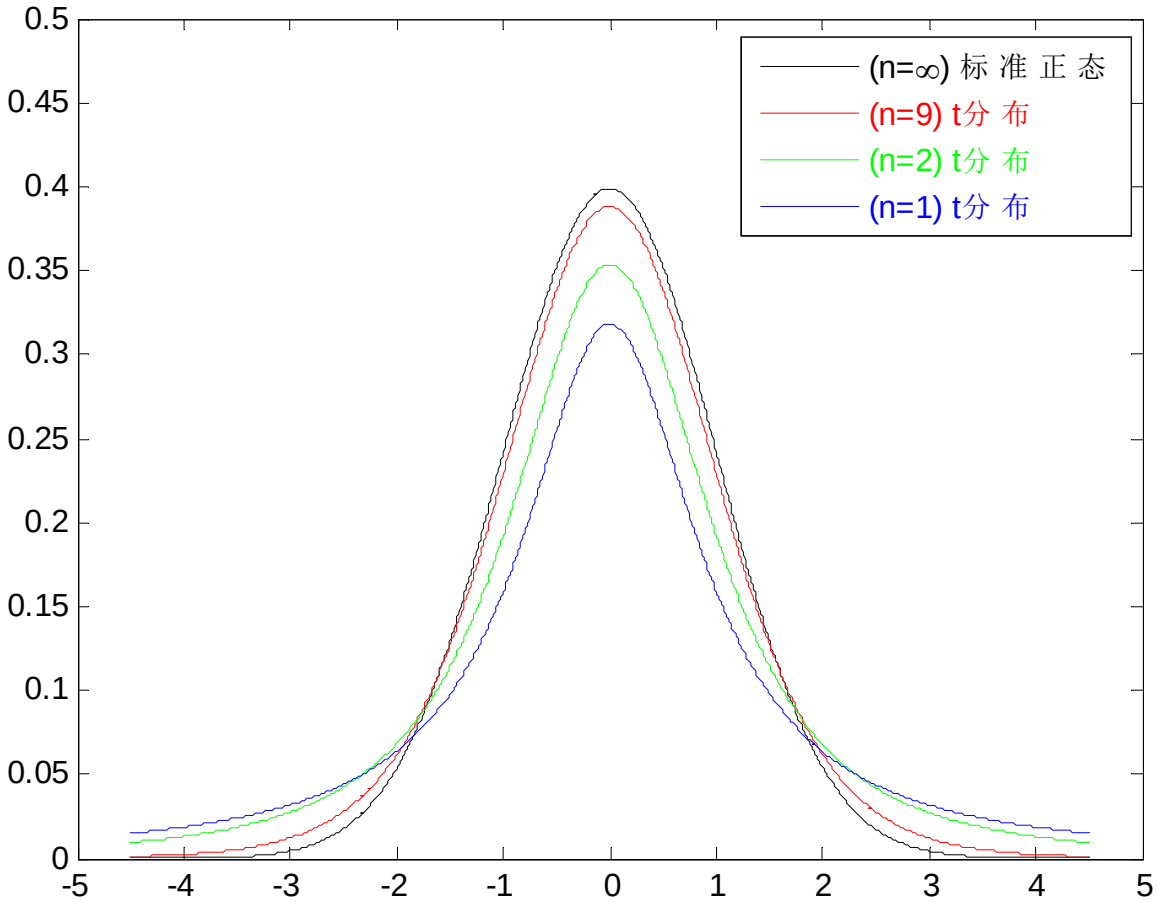
定义 (6.3.3)

称 ξ 服从自由度为 n 的 t 分布, 记为 $\xi \sim t(n)$, 如果它的密度函数为

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

$t-$ 分布的图像如下图所示.

$t-$ 分布的密度函数图像


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
(n=∞) 标 准 正 态
(n=9) t分 布
(n=2) t分 布
(n=1) t分 布

2. $t-$ 分布的构造

定理 (6.3.2)

若 $\xi \sim N(0,1), \eta \sim \chi^2(n)$, 且 ξ 与 η 相互独立, 则

$$t = \frac{\xi}{\sqrt{\eta/n}} \sim t(n).$$

$t-$ 分布是统计学中一类重要分布, 它与标准正态分布的微小差别是由英国统计学家戈赛特 (Gosset) 在其论文《均值的或然误差》中导出的, 该论文于 1908 年以 "Student" 的笔名发表在《生物统计》上, 这是统计量精确分布理论中一系列重要结果的开端. 后人也称 $t-$ 分布为学生氏分布, $t-$ 分布的发现在统计学上具有划时代的意义, 打破了原先正态分布一统天下的局面, 并开创了小样本统计推断的先河.

学生氏 t 分布的由来

定理 (6.3.2)

若 $\xi \sim N(0,1), \eta \sim \chi^2(n)$, 且 ξ 与 η 相互独立, 则

$$t = \frac{\xi}{\sqrt{\eta/n}} \sim t(n).$$

Gosset (1876 年-1937 年) 在爱尔兰首都都柏林的 Guinness 啤酒厂工作. 在对啤酒厂进行质量控制的研究中, Gosset 发现了 t 分布. 当时啤酒厂有规定, 禁止雇员将研究成果公开发表, 于是 Gosset 在 1908 年的论文中, 偷偷地以笔名 "Student" 发表了 t 分布的发现. 正是由于这个原因, t 分布也称学生氏分布.

3. $t-$ 分布的分位数

定义 (6.3.4)

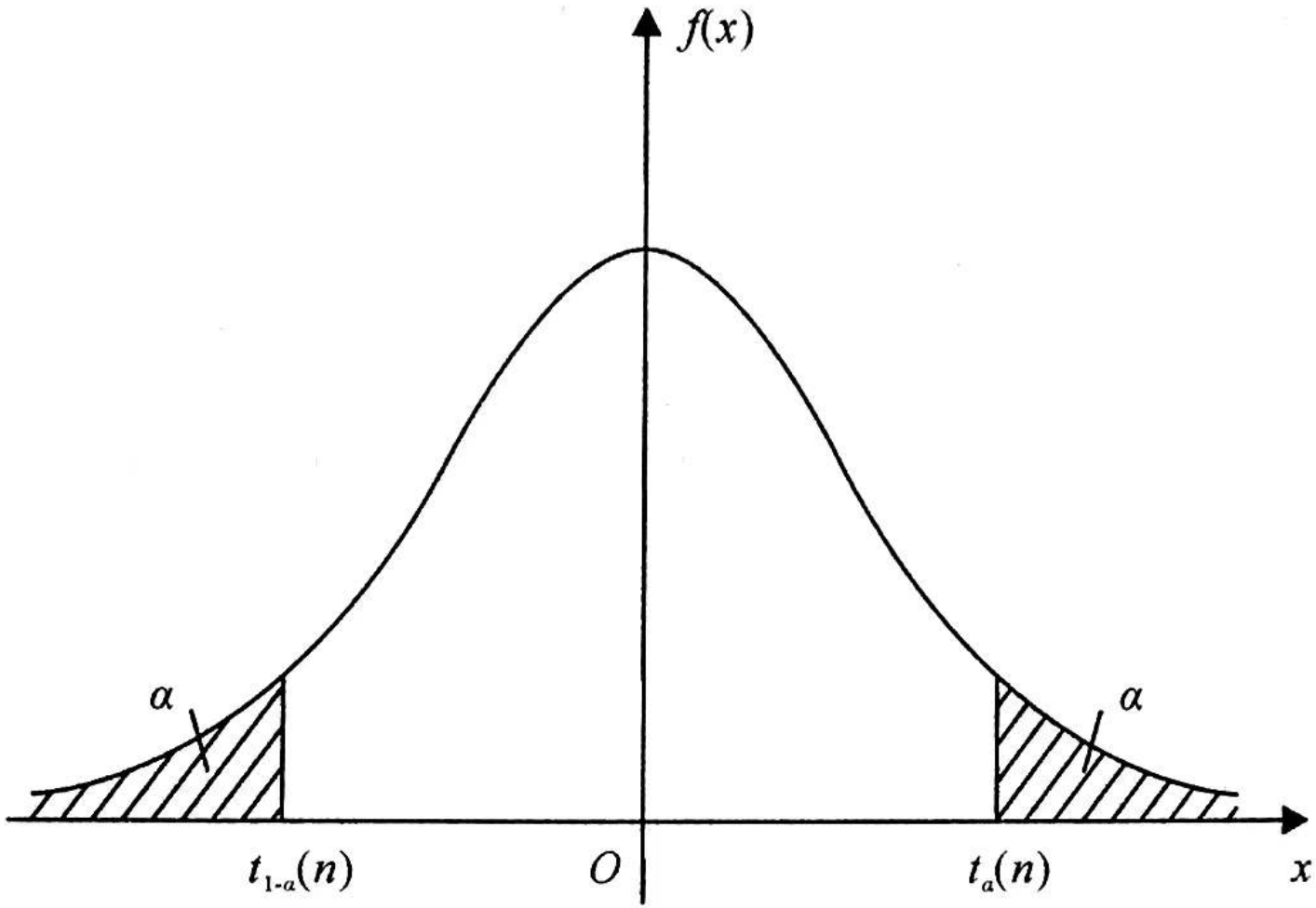
设 $t \sim t(n)$, 对于给定的正数 α $(0 < \alpha < 1)$, 称满足条件

$$P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} f(x)\mathrm{d}x = \alpha,$$

的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位数.

下图给出了 $t(n)$ 分布的上 α 分位点 $t_\alpha(n)$.

$t(n)$ 分布的上 α 分位点 $t_\alpha(n)$



3. $t-$ 分布的分位数

由 t 分布概率密度 $f(x)$ 图形的对称性可知

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

书末附录 1.4 给出了 $t(n)$ 分布上 α 分位点 $t_{\alpha}(n)$ 的数值表, 例如

$$t_{0.025}(8) = 2.3060, t_{0.99}(12) = t_{1-0.01}(12) = -t_{0.01}(12) = -2.6810,$$

当 n 较大 (通常 $n > 45$) 时, $t_{\alpha}(n)$ 可由标准正态分布的上 α 分位点 u_{α} 来近似代替.

4. $t-$ 分布的性质

(1) $t-$ 分布的密度函数的图像是一个关于纵轴对称的分布. 且

$$\lim_{x\to\infty} f(x) = 0.$$

(2) $t-$ 分布的密度函数的图像与标准正态分布的密度函数形状类似, 只是峰比标准正态分布低一些, 尾部的概率比标准正态分布的厚一些. 实际上, 当 n 充分大时 $t-$ 分布近似于标准正态分布. 即

$$\lim_{n\to+\infty} f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x2}{2}}.$$

(3) 若 $\xi \sim t(n)$, 则 $E(\xi) = 0$, $(n > 1)$; $D(\xi) = \dfrac{n}{n-2}(n > 2)$. 当 $n = 1$ 时, $t-$ 分布就是标准柯西分布, 它的期望和方差不存在.

例 (6.3.2)

设总体 ξ 和 η 相互独立且都服从 $N\left(0,3^2\right)$ 分布, 而样本 $\xi_1,\xi_2,\cdots,\xi_9$ 和 $\eta_1,\eta_2,\cdots,\eta_9$ 分别来自 ξ 和 η, 求统计量

$$T=\frac{\xi_1+\xi_2+\cdots+\xi_9}{\sqrt{\eta_1^2+\eta_2^2+\cdots+\eta_9^2}}$$ 的分布.

例 (6.3.2)

设总体 ξ 和 η 相互独立且都服从 $N\left(0,3^2\right)$ 分布，而样本 $\xi_1,\xi_2,\cdots,\xi_9$ 和 $\eta_1,\eta_2,\cdots,\eta_9$ 分别来自 ξ 和 η，求统计量

$$T=\frac{\xi_1+\xi_2+\cdots+\xi_9}{\sqrt{\eta_1^2+\eta_2^2+\cdots+\eta_9^2}}$$ 的分布.

解

由于 $\overline{\xi}=\frac{1}{9}\sum_{i=1}^{9}\xi_i\sim N(0,1),\ \eta_i/3\sim N(0,1),\quad i=1,2,\cdots,9,$

$\zeta=\sum_{i=1}^{9}\left(\frac{\eta_i}{3}\right)^2=\frac{1}{9}\sum_{i=1}^{9}\eta_i^2\sim\chi^2(9)$，并且 $\overline{\xi}$ 和 ζ 相互独立，由 $t-$ 分布的构造知

$$T=\frac{\overline{\xi}}{\sqrt{\zeta/9}}=\sum_{i=1}^{9}\xi_i\Big/\sqrt{\sum_{i=1}^{9}\eta_i^2}\sim t(9).$$

例 (6.3.3)

设 $\xi_1, \xi_2, \xi_3, \xi_4$ 为来自总体 $N\left(1, \sigma^2\right)(\sigma > 0)$ 的简单随机样本，求统计量 $\dfrac{\xi_1 - \xi_2}{\left|\xi_3 + \xi_4 - 2\right|}$ 的分布.

解

由正态分布的性质知 $\xi_1-\xi_2\sim N\left(0,2\sigma^2\right)$, $\xi_3+\xi_4\sim N\left(2,2\sigma^2\right)$. 所以,

$$\frac{\xi_1-\xi_2}{\sqrt{2}\sigma}\sim N(0,1),\ \frac{\xi_3+\xi_4-2}{\sqrt{2}\sigma}\sim N(0,1),$$

而

$$\left(\frac{\xi_3+\xi_4-2}{\sqrt{2}\sigma}\right)^2\sim\chi^2(1),$$

由于 $\dfrac{\xi_1-\xi_2}{\sqrt{2}\sigma}$ 与 $\left(\dfrac{\xi_3+\xi_4-2}{\sqrt{2}\sigma}\right)^2$ 相互独立, 从而由 $t-$ 分布构造, 有

$$\frac{\xi_1-\xi_2}{|\xi_3+\xi_4-2|}=\frac{\xi_1-\xi_2}{\sqrt{2}\sigma}\Bigg/\sqrt{\left(\frac{\xi_3+\xi_4-2}{\sqrt{2}\sigma}\right)^2}\sim t(1).$$

1. $F-$ 分布的定义

定义 (6.3.5)

称 ξ 服从自由度为 (n_1, n_2) 的 F 分布, 记为 $\xi \sim F(n_1, n_2)$, 如果它的密度函数为

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} \left(\dfrac{n_1}{n_2}\right)^{\frac{n_2}{2}} x^{\frac{n_1}{2}-1} \left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x > 0; \\ 0, & x \leqslant 0. \end{cases}$$

其中 n_1, n_2 是两个正整数, n_1 是分子的自由度, 称为第一自由度, n_2 是分母的自由度, 称为第二自由度.

1. $F-$ 分布的定义

- $F-$ 分布最早见于英国统计学家费歇尔 (R.A. Fisher) 1922 年发表的论文.
- $F-$ 分布的名称由美国统计学家斯纳德柯 (G.W. Snedecor) 在 1932 年引进, 以纪念费歇尔的功绩.
- $F-$ 分布经常被用来对两个样本方差进行比较. 它是方差分析的一个基本分布, 也被用于回归分析中的显著性检验.

2. $F-$ 分布的构造

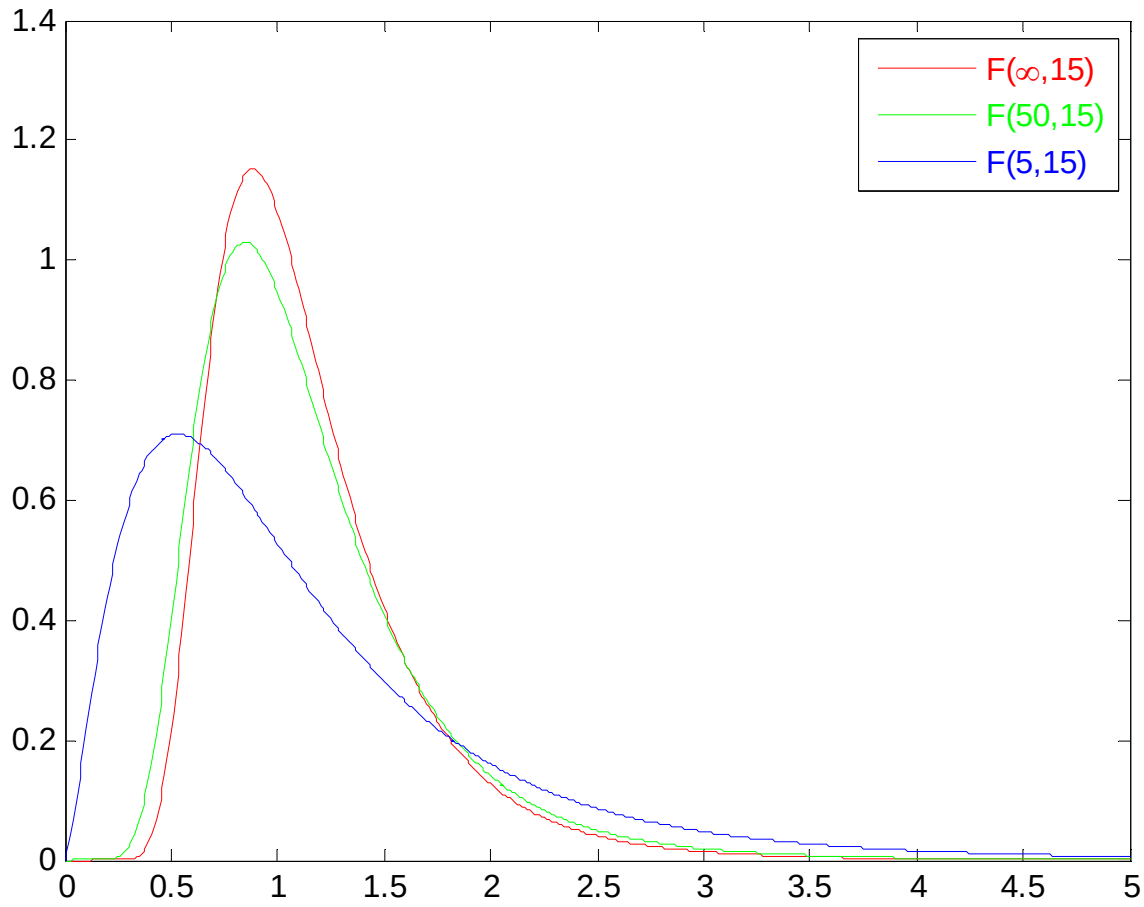
定理 (6.3.3)

设 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$, 且 ξ 和 η 相互独立, 则

$$F = \frac{\xi/n_1}{\eta/n_2} \sim F(n_1, n_2).$$

$F(n_1, n_2)$ 的密度函数的图像如下图所示, 它是一个只取非负值的偏态分布.

$F(n_1, n_2)$ 的密度函数图像



3. $F-$ 分布的分位数

定义 (6.3.6)

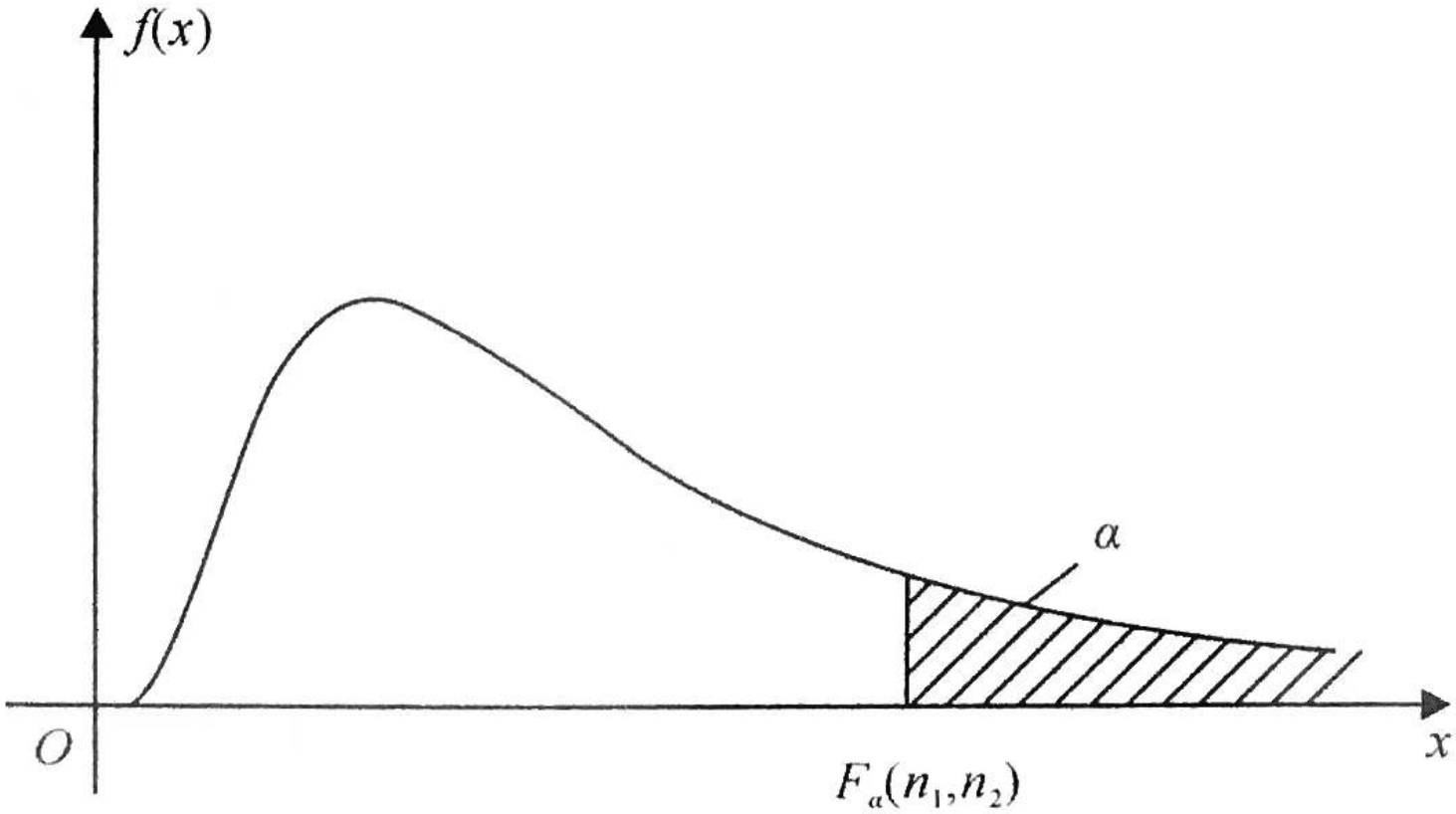
设 $F \sim F(n_1, n_2)$, 对于给定的正数 α $(0 < \alpha < 1)$, 称满足条件

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1,n_2)}^{+\infty} f(x)\mathrm{d}x = \alpha$$

的点 $F_\alpha(n_1, n_2)$ 为 $F(n_1, n_2)$ 分布的上 α 分位数.

下图给出了 $F(n_1, n_2)$ 分布的上 α 分位点 $F_\alpha(n_1, n_2)$.

$F(n_1, n_2)$ 分布的上 α 分位点 $F_\alpha(n_1, n_2)$



4. $F-$ 分布的性质

(1) 若 $\xi \sim F(n_1, n_2)$, 则

$$E(\xi) = \frac{n_2}{n_2 - 2}, \quad (n_2 > 2),$$
$$D(\xi) = \frac{n_2^2 (2n_1 + 2n_2 - 4)}{n_1 (n_2 - 2)^2 (n_2 - 4)}, \quad (n_2 > 4).$$

(2) 若 $X \sim t(n)$, 则 $X^2 \sim F(1, n)$.

(3) 若 $F \sim F(n_1, n_2)$, 则 $\frac{1}{F} \sim F(n_2, n_1)$.

(4) $F-$ 分布的上 α 分位数有如下的性质:

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}.$$

4. $F-$ 分布的性质

(2) 若 $X \sim t(n)$, 则 $X^2 \sim F(1,n)$.

证明.

由 $t-$ 分布构造可知, 随机变量 ξ 与 η 相互独立, 使得

$$X = \xi / \sqrt{\eta/n},$$

其中 $\xi \sim N(0,1), \eta \sim \chi^2(n)$. 而 $X^2 = \dfrac{\xi^2}{\eta/n} = \dfrac{\xi^2/1}{\eta/n}$, 并且 $\xi^2 \sim \chi^2(1)$, 所以由 F 分布的定义知

$$X^2 = \frac{\xi^2}{\eta/n} \sim F(1,n),$$

即 $X^2 \sim F(1,n)$. □

(4) $F_{1-\alpha}(n_1, n_2) = \dfrac{1}{F_{\alpha}(n_2, n_1)}$.

证明. 事实上, 对于给定的 α $(0 < \alpha < 1)$, 有

$$1-\alpha = P\{F > F_{1-\alpha}(n_1, n_2)\} = P\left\{\frac{1}{F} < \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\}$$
$$= 1 - P\left\{\frac{1}{F} \geqslant \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\},$$

于是

$$P\left\{\frac{1}{F} \geqslant \frac{1}{F_{1-\alpha}(n_1, n_2)}\right\} = \alpha,$$

由于 $\dfrac{1}{F} \sim F(n_2, n_1)$, 因此 $\dfrac{1}{F_{1-\alpha}(n_1, n_2)}$ 就是 $F(n_2, n_1)$ 的上 α 分位点 $F_{\alpha}(n_2, n_1)$, 即

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_{\alpha}(n_2, n_1)}.$$

例 (6.3.4)

设总体 ξ 服从标准正态分布 $N(0,1)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的一个简单随机样本, 试问统计量

$$\eta = \left(\frac{n}{5} - 1\right) \sum_{i=1}^{5} \xi_i^2 \Big/ \sum_{i=6}^{n} \xi_i^2, \quad n > 5.$$

服从何种分布?

例 (6.3.4)

设总体 ξ 服从标准正态分布 $N(0,1)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自总体 ξ 的一个简单随机样本, 试问统计量

$$\eta = \left(\frac{n}{5} - 1\right) \sum_{i=1}^{5} \xi_i^2 \Big/ \sum_{i=6}^{n} \xi_i^2, \quad n > 5.$$

服从何种分布?

解

因为 $\xi_i \sim N(0,1)$, $\sum_{i=1}^{5} \xi_i^2 \sim \chi^2(5)$, $\sum_{i=6}^{n} \xi_i^2 \sim \chi^2(n-5)$, 且 $\sum_{i=1}^{5} \xi_i^2$ 与 $\sum_{i=6}^{n} \xi_i^2$ 相互独立, 所以

$$\eta = \frac{\sum_{i=1}^{5} \xi_i/5}{\sum_{i=6}^{n} \xi_i^2/(n-5)} \sim F(5, n-5).$$

定理 (6.3.1)

若 $\xi_1, \xi_2 \cdots, \xi_n$ 相互独立，且都服从标准正态分布 $N(0,1)$，则

$$\chi^2 = \sum_{i=1}^{n} \xi_i^2 \sim \chi^2(n).$$

定理 (6.3.2)

若 $\xi \sim N(0,1), \eta \sim \chi^2(n)$，且 ξ 与 η 相互独立，则

$$t = \frac{\xi}{\sqrt{\eta/n}} \sim t(n).$$

定理 (6.3.3)

设 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$，且 ξ 和 η 相互独立，则

$$F = \frac{\xi/n_1}{\eta/n_2} \sim F(n_1, n_2).$$

分位数

定义 (分位数)

(1) χ^2- 分布的上 α 分位数 $\chi^2_\alpha(n)$:

$$P\{\xi > \chi^2_\alpha(n)\} = \alpha.$$

(2) $t-$ 分布的上 α 分位数 $t_\alpha(n)$:

$$P\{t > t_\alpha(n)\} = \alpha.$$

(3) $F-$ 分布的上 α 分位数 $F_\alpha(n_1, n_2)$:

$$P\{F > F_\alpha(n_1, n_2)\} = \alpha.$$

$$t_{1-\alpha}(n) = -t_\alpha(n), \quad F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_1, n_2)}.$$

χ^2- 分布的性质

(1) 若 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$, 且 ξ, η 相互独立, 则

$$\xi + \eta \sim \chi^2(n_1 + n_2).$$

(2) 若 $\xi \sim \chi^2(n)$, 则 $E(\xi) = n, D(\xi) = 2n.$

$t-$ 分布的性质

(1) $t-$ 分布的密度函数的图像关于纵轴对称且 $\lim_{x\to\infty} f(x) = 0.$

(2) 当 n 充分大时 $t-$ 分布近似于标准正态分布.

(3) 若 $\xi \sim t(n)$, 则 $E(\xi) = 0, (n > 1); D(\xi) = \dfrac{n}{n-2} (n > 2).$

$F-$ 分布的性质

(1) 若 $\xi \sim t(n)$, 则 $\xi^2 \sim F(1, n).$

(2) 若 $F \sim F(n_1, n_2)$, 则 $\dfrac{1}{F} \sim F(n_2, n_1).$

1. 单正态总体的样本均值与样本方差的分布

设总体 ξ 的均值为 μ, 方差为 σ^2, $\xi_1, \xi_2, \cdots, \xi_n$ 是取自总体 ξ 的一个样本, $\overline{\xi} = \dfrac{1}{n}\sum\limits_{i=1}^{n}\xi_i$ 与 $S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$ 分别为该样本的样本均值与样本方差.

定理 (6.3.4)

设总体 $\xi \sim N\left(\mu, \sigma^2\right), \xi_1, \xi_2, \cdots, \xi_n$ 是来自正态总体 ξ 的一个简单样本, 则

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

该定理的证明可由 χ^2 分布的构造直接得到.

定理 (Fisher 定理 6.3.5)

设总体 $\xi \sim N\left(\mu, \sigma^2\right)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自正态总体 ξ 的一个简单样本, $\overline{\xi}$, S^2 分别为该样本的样本均值与样本方差, 则有

$(1)\ \overline{\xi} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$; $(2)\ \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$; $(3)\ \overline{\xi}$ 与 S^2 相互独立.

定理 (Fisher 定理 6.3.5)

设总体 $\xi \sim N(\mu, \sigma^2)$, $\xi_1, \xi_2, \cdots, \xi_n$ 是来自正态总体 ξ 的一个简单样本, $\overline{\xi}$, S^2 分别为该样本的样本均值与样本方差, 则有

(1) $\overline{\xi} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$; (2) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$; (3) $\overline{\xi}$ 与 S^2 相互独立.

证明.

由

$$E(\overline{\xi}) = \mu, \quad D(\overline{\xi}) = \sigma^2/n,$$

又 $\xi_1, \xi_2, \cdots, \xi_2$ 相互独立, 并且与总体 ξ 同服从 $N(\mu, \sigma^2)$, 由正态分布的性质知, $\overline{\xi} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$, 这就证明了结论 (1).

记 $\xi = (\xi_1, \xi_2, \cdots, \xi_n)^{\mathrm{T}}, \eta = (\eta_1, \eta_2, \cdots, \eta_n)^{\mathrm{T}}$, 则 $E(\xi) = (\mu, \mu, \cdots, \mu)^{\mathrm{T}}$, $D(\xi) = \sigma^2 I$, 其中 I 为单位矩阵. □

证明.

取一个正交矩阵 A 为

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & -\frac{1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & -\frac{2}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \frac{1}{\sqrt{n\cdot(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n\cdot(n-1)}} \end{pmatrix}$$

令 $\eta = A\xi$, 由多维正态分布的性质知 η 仍服从 n 维正态分布. 经过正交变换, 有 $\eta_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i = \sqrt{n}\,\overline{\xi}$, 于是 $\eta_1^2 = n\overline{\xi}^2$. □

证明.

由于 A 的正交性,

$$\sum_{i=1}^{n}\eta_i^2=\eta^T\eta=\xi^TA^TA\xi=\xi^T\xi=\sum_{i=1}^{k}\xi_i^2=\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2+n\overline{\xi}^2,$$

故

$$(n-1)S^2=\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2=\sum_{i=1}^{n}\eta_i^2-\eta_1^2=\sum_{i=2}^{n}\eta_i^2,$$

且有

$$E(\eta)=\begin{pmatrix}E(\eta_1)\\E(\eta_2)\\\vdots\\E(\eta_n)\end{pmatrix}=AE(\xi)=\begin{pmatrix}\sqrt{n}\mu\\0\\\vdots\\0\end{pmatrix},$$

□

证明.

$$D(\eta) = A(D(\xi))A^T = A\sigma^2 IA^T = \sigma^2 I.$$

即 $\eta = (\eta_1, \eta_2, \cdots, \eta_n)^T$ 的各个分量相互独立, 且都服从正态分布, 其方差均为 σ^2, η_1 的均值为 $\sqrt{n}\mu$, 其余 $\eta_2, \cdots, \eta_n$ 的均值均为 0. 故它们相互独立, 从而 $\eta_i^2 = n\overline{\xi}^2$ 与 $\sum\limits_{i=2}^{n} \eta_i^2 = (n-1)S^2$ 也相互独立, 这就证明了 (3) 成立.

由于 $\eta_2, \cdots, \eta_n$ 相互独立且同分布于 $N\left(0, \sigma^2\right)$. 于是

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} \left(\frac{\eta_i}{\sigma}\right)^2 \sim \chi^2(n-1),$$

即 (2) 成立. □

推论 (6.3.1)

设总体 $\xi \sim N\left(\mu, \sigma^2\right), \xi_1, \xi_2, \cdots, \xi_2$ 是取自 ξ 的一个样本, $\overline{\xi}$, S^2 分别为该样本的样本均值与样本方差, 则有

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

推论 (6.3.1)

设总体 $\xi \sim N\left(\mu, \sigma^2\right), \xi_1, \xi_2, \cdots, \xi_2$ 是取自 ξ 的一个样本, $\overline{\xi}$, S^2 分别为该样本的样本均值与样本方差, 则有

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

证明.

由定理 6.3.5 及 $t-$ 分布的构造知

$$\frac{\dfrac{\overline{\xi}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} \sim t(n-1),$$

化简上式即得 $\dfrac{\overline{\xi}-\mu}{S/\sqrt{n}} \sim t(n-1).$ □

例 (6.3.5)

设总体 $\xi \sim N\left(20, 4^2\right), \xi_1, \xi_2, \cdots, \xi_{16}$ 为来自总体 ξ 的一个样本, 求:
(1) 样本均值 $\overline{\xi}$ 的数学期望与方差; (2) $P\{|\overline{\xi} - 20| \leqslant 0.6\}$.

例 (6.3.5)

设总体 $\xi \sim N\left(20, 4^2\right), \xi_1, \xi_2, \cdots, \xi_{16}$ 为来自总体 ξ 的一个样本, 求:
(1) 样本均值 $\overline{\xi}$ 的数学期望与方差; (2) $P\{|\overline{\xi} - 20| \leqslant 0.6\}$.

解

(1) 由于 $\xi \sim N\left(20, 4^2\right)$, 样本容量 $n = 16$. 由 *Fisher* 定理, 有

$$\overline{\xi} \sim N\left(20, \frac{4^2}{16}\right),$$

于是 $E(\overline{\xi}) = 20, D(\overline{\xi}) = \dfrac{4^2}{16} = 1.$

(2) 由 $\overline{\xi} \sim N\left(20, \dfrac{4^2}{16}\right)$, 得 $\overline{\xi} - 20 \sim N(0, 1)$, 故

$$P\{|\overline{\xi} - 20| \leqslant 0.6\} = 2\Phi(0.6) - 1 = 0.4514.$$

例 (6.3.6)

从正态总体 $\xi \sim N\left(\mu, \sigma^2\right)$ 中抽取容量为 n 的一个样本, S^2 为其样本方差, 求 S^2 的方差 $D\left(S^2\right)$.

例 (6.3.6)

从正态总体 $\xi \sim N\left(\mu, \sigma^2\right)$ 中抽取容量为 n 的一个样本, S^2 为其样本方差, 求 S^2 的方差 $D\left(S^2\right)$.

解

因为 $\xi \sim N\left(\mu, \sigma^2\right)$, 由 *Fisher* 定理得 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$,
由 χ^2 分布的方差, 得

$$D\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1).$$

又由方差的性质 $\dfrac{(n-1)^2}{\sigma^4} D\left(S^2\right) = 2(n-1)$, 于是

$$D\left(S^2\right) = \frac{2\sigma^4}{n-1}.$$

表: 单个正态总体 $\xi \sim N(\mu, \sigma^2)$

样本函数	分布
$U = \dfrac{\overline{\xi} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$
$t = \dfrac{\overline{\xi} - \mu}{S/\sqrt{n}}$	$t(n-1)$
$\chi^2 = \dfrac{\sum\limits_{i=1}^{n}(\xi_i - \mu)^2}{\sigma^2}$	$\chi^2(n)$
$\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum\limits_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2}{\sigma^2}$	$\chi^2(n-1)$

2. 双正态总体的样本均值与样本方差的分布

推论 (6.3.2)

$\xi \sim N\left(\mu_1, \sigma_1^2\right)$ 与 $\eta \sim N\left(\mu_2, \sigma_2^2\right)$ 为两个独立的正态总体，分别从 ξ 和 η 中抽取样本 $\xi_1, \xi_2, \cdots, \xi_{n_1}$ 和 $\eta_1, \eta_2, \cdots, \eta_{n_2}$，则随机变量

$$F = \frac{n_2}{n_1} \cdot \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{\sum\limits_{i=1}^{n_1} \left(\xi_i - \mu_1\right)^2}{\sum\limits_{i=1}^{n_2} \left(\eta_i - \mu_2\right)^2} \sim F\left(n_1, n_2\right).$$

证明.

由定理 6.3.4 知

$$\chi_1^2=\sum_{i=1}^{n_1}\left(\frac{\xi_i-\mu_1}{\sigma_1}\right)^2\sim\chi^2(n_1),\quad \chi_2^2=\sum_{i=1}^{n_2}\left(\frac{\eta_i-\mu_2}{\sigma_2}\right)^2\sim\chi^2(n_2),$$

且 χ_1^2 和 χ_2^2 相互独立，由 $F-$ 分布的构造知

$$F=\frac{n_2}{n_1}\cdot\frac{\sigma_2^2}{\sigma_1^2}\cdot\frac{\sum\limits_{i=1}^{n_1}(\xi_i-\mu_1)^2}{\sum\limits_{i=1}^{n_2}(\eta_i-\mu_2)^2}=\frac{\chi_1^2/n_1}{\chi_2^2/n_2}\sim F(n_1,n_2).$$

□

推论 (6.3.3)

设 $\xi \sim N\left(\mu_1, \sigma_1^2\right)$ 与 $\eta \sim N\left(\mu_2, \sigma_2^2\right)$ 是两个相互独立的正态总体, 又设 $\xi_1, \xi_2, \cdots, \xi_{n_1}$ 是取自总体 ξ 的样本, $\overline{\xi}$ 与 S_1^2 分别为该样本的样本均值与样本方差. $\eta_1, \eta_2, \cdots, \eta_{n_2}$ 是取自总体 η 的样本, $\overline{\eta}$ 与 S_2^2 分别为此样本的样本均值与样本方差. 记 S_w^2 是 S_1^2 与 S_2^2 的加权平均, 即

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

则 (1) $\dfrac{(\overline{\xi} - \overline{\eta}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$.

(2) $F = \left(\dfrac{\sigma_2}{\sigma_1}\right)^2 \dfrac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$.

(3) 当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时, $t = \dfrac{(\overline{\xi} - \overline{\eta}) - (\mu_1 - \mu_2)}{S_w \sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$.

证明.

(1) 由定理的条件知 $\overline{\xi}$ 和 $\overline{\eta}$ 相互独立, 且

$$\overline{\xi} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \overline{\eta} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right),$$

所以

$$\overline{\xi} - \overline{\eta} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

即

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

□

证明.

(2) 由定理 6.3.5 知

$$\chi_1^2=\frac{(n_1-1)\,S_1^2}{\sigma_1^2}\sim\chi^2\,(n_1-1)\,,\chi_2^2=\frac{(n_2-1)\,S_2^2}{\sigma_2^2}\sim\chi^2\,(n_2-1)\,,$$

且 χ_1^2 和 χ_2^2 相互独立, 由 $F-$ 分布的构造得

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

□

证明.

(3) 由定理 6.3.5 及 χ^2- 分布的可加性得

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

由 (1) 知, $\dfrac{(\overline{\xi}-\overline{\eta})-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$, 根据 $t-$ 分布的构造可得

$$t=\frac{\overline{\xi}-\overline{\eta}-(\mu_1-\mu_2)}{S_w\sqrt{1/n_1+1/n_2}}\sim t\left(n_1+n_2-2\right).$$

□

表: 双正态总体, $\xi \sim N\left(\mu_1, \sigma_1^2\right)$, $\eta \sim N\left(\mu_2, \sigma_2^2\right)$, ξ, η 独立

样本函数	分布
$U=\dfrac{\overline{\xi}-\overline{\eta}-\left(\mu_1-\mu_2\right)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	$N(0,1)$
$T=\dfrac{\overline{\xi}-\overline{\eta}-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{\left(n_1-1\right)S_1^2+\left(n_2-1\right)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ (当 $\sigma_1^2=\sigma_2^2$ 时)	$t(n_1+n_2-2)$
$F=\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	$F(n_1-1, n_2-1)$

本章小结

- 本章第一节主要介绍了数理统计中的一些基本概念，如总体、样本、样本容量、简单随机抽样、经验分布函数、统计量等概念，引入了两个重要统计量样本均值 $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ 与样本方差

  $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2.$$

- 第二节介绍了描述性统计中一些特征值，如描述数据集中趋势的统计量：平均数、中位数、众数的概念；描述数据离散趋势的统计量：极差、方差、标准差的概念以及峰度与偏度的概念和频率直方图的做法.
- 第三节重点探讨了数理统计中常用三大抽样分布：χ^2- 分布、$t-$ 分布、$F-$ 分布的定义、构造、密度函数的图像、性质、查分布表确定上 α 分位点.
- 对于正态总体统计量样本均值与样本方差的分布总结如下表：

表: 单个正态总体 $\xi \sim N(\mu, \sigma^2)$

样本函数	分布
$U = \dfrac{\overline{\xi} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$
$t = \dfrac{\overline{\xi} - \mu}{S/\sqrt{n}}$	$t(n-1)$
$\chi^2 = \dfrac{\sum\limits_{i=1}^{n}(\xi_i - \mu)^2}{\sigma^2}$	$\chi^2(n)$
$\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum\limits_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2}{\sigma^2}$	$\chi^2(n-1)$

表: 双正态总体, $\xi \sim N\left(\mu_1, \sigma_1^2\right)$, $\eta \sim N\left(\mu_2, \sigma_2^2\right)$, ξ, η 独立

样本函数	分布
$U=\dfrac{\overline{\xi}-\overline{\eta}-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$T=\dfrac{\overline{\xi}-\overline{\eta}-\left(\mu_1-\mu_2\right)}{\sqrt{\dfrac{\left(n_1-1\right) S_1^2+\left(n_2-1\right) S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ (当 $\sigma_1^2=\sigma_2^2$ 时)	$t(n_1+n_2-2)$
$F=\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	$F(n_1-1, n_2-1)$